Journal of Clinical Epidemiology 111 (2019) 11−22

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# The number needed to treat in pairwise and network meta-analysis and its graphical representation

Areti Angeliki Veroniki[a,b,c,*], Ralf Bender[d], Paul Glasziou[e], Sharon E. Straus[a,f], Andrea C. Tricco[a,g]

[a]*Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, East Building, Toronto, Ontario, M5B 1T8, Canada*
[b]*Department of Primary Education, School of Education,University of Ioannina, Ioannina, Greece*
[c]*Institute of Reproductive and Developmental Biology, Department of Surgery & Cancer, Faculty of Medicine, Imperial College, London W12 0NN, UK*
[d]*Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Im Mediapark 8, 50670 Cologne, Germany*
[e]*Centre for Research on Evidence Based Practice, Bond University, Gold Coast, Australia*
[f]*Department of Geriatric Medicine, University of Toronto, Toronto, Ontario, Canada*
[g]*Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, 155 College Street, 6th floor, Toronto, Ontario, M5T 3M7, Canada*

Accepted 6 March 2019; Published online 21 March 2019

## Abstract

**Objective:** The objective of this study was to present ways to graphically represent a number needed to treat (NNT) in (network) meta-analysis (NMA).

**Study Design and Setting:** A barrier to using NNT in NMA when an odds ratio (OR) or risk ratio (RR) is used is the determination of a single control event rate (*CER*). We discuss approaches to calculate a *CER*, and illustrate six graphical methods for NNT from NMA. We illustrate the graphical approaches using an NMA of cognitive enhancers for Alzheimer's dementia.

**Results:** The NNT calculation using a relative effect measure, such as OR and RR, requires a *CER* value, but different *CER*s, including mean *CER* across studies, pooled *CER* in meta-analysis, and expert opinion-based *CER* may result in different NNTs. An NNT from NMA can be presented in a bar plot, Cates plot, or forest plot for a single outcome, and a bubble plot, scatterplot, or rank-heat plot for ≥2 outcomes. Each plot is associated with different properties and can serve different needs.

**Conclusion:** Caution is needed in NNT interpretation, as considerations such as selection of effect size and *CER*, and *CER* assumption across multiple comparisons, may impact NNT and decision-making. The proposed graphs are helpful to interpret NNTs calculated from (network) meta-analyses. 

*Keywords:* Multiple treatment meta-analysis; Multiple outcomes; Number needed to harm; Rank-heat plot; Graphical displays; Presentation results

## 1. Introduction

The number needed to treat (NNT) is an absolute measure of effect used to communicate the effectiveness or safety of an intervention [1]. The NNT was first introduced to describe the absolute effect of a certain intervention vs. a standard treatment or control in randomized clinical trials [2] and then was adopted in systematic reviews and meta-analyses [3]. The NNT provides insight into the clinical relevance of an effect size because it is defined as the average number of patients who need to be treated to prevent one extra person from having a bad outcome compared with another treatment. For positive outcomes, the NNT can be equivalently defined as the number of people that need to be treated to have one person with a good outcome. Similarly, the number needed to harm (NNH) indicates how many people need to be treated in order for one patient to have a particular adverse effect. To avoid the unfavorable NNH term, Altman [4] suggested the terms ''number needed to treat for an additional beneficial outcome'' (NNTB) and ''number needed to treat for an additional harmful outcome'' (NNTH), respectively, instead of using

Conflict of interest: A.C.T. and S.E.S. are on the editorial board of the Journal of Clinical Epidemiology, but were not involved with the peer review process or decision for publication and not involved in any way in the journal management of this manuscript. The other authors have nothing to declare.

* Corresponding Author: Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece. Tel.: +30 26510 05712; fax: +30 26510 05854.

*E-mail address:* averonik@cc.uoi.gr (A.A. Veroniki).

https://doi.org/10.1016/j.jclinepi.2019.03.007
0895-4356/

**What is new?**

**Key findings**

- The number needed to treat (NNT) is an absolute measure of effect used to communicate the effectiveness or safety of an intervention and is frequently used in the meta-analytical literature.

**What this adds to what was known?**

- Different considerations of calculating an NNT in both pairwise and network meta-analysis (NMA), including effect size and assumptions for the control event rate across multiple comparisons, may impact NNT results. We present potential ways of calculating NNT in (network) meta-analysis, such as mean control event rate (*CER*) across studies, pooled *CER* in meta-analysis, expert opinion-based *CER*, and range of possible *CER*.

**What is the implication and what should change now?**

- The graphical representation of NNTs from NMA is crucial to ease interpretation of results. We present six graphical approaches for NNT from NMA and discuss their properties. We suggest the NNT graphical representation in a bar plot, Cates plot, or forest plot for a single outcome, and in a bubble plot, scatterplot, or rank-heat plot for at least two outcomes.
- Different plots can be used for different needs. For example, if uncertainty around NNT should be considered in decision-making, then a bar plot or a forest plot can be used. When multiple outcomes need to be considered, then a rank-heat plot is suggested. For communication purposes, the Cates plot is suggested if the corresponding effect estimate is statistically significant and the confidence interval is not too wide.

NNT and NNH to show direction of effect. In this article, we use the terms NNTB and NNTH.

The NNTB and NNTH are calculated by taking the inverse of the risk difference (RD) [2], yet can also be calculated using other effect measures, such as the odds ratio (OR) and risk ratio (RR) [5]. The higher the NNTB value, the less effective the treatment will be; and, the higher the NNTH, the more safe a treatment is. For example, intervention A with an NNTB of 20 whereby one patient is saved for every 20 patients treated with A is better than a competing intervention B (with an NNTB of 80) that saves one patient for every 80 patients treated with B. The use and interpretation of NNT requires understanding of several factors [1,6], such as 1) clinical insight and patient values and circumstances, as it may depend on how difficult it is to implement the intervention and how accessible and cost-effective the intervention is, 2) follow-up period, as NNTs in studies with different follow-up times are not directly comparable [7], 3) baseline risk of the event, 4) statistical properties of NNT, 5) alternative treatment to which the intervention is being compared, 6) outcome, 7) direction and size of the effect measure, 8) NNTB (and NNTH) scale, and 9) confidence interval (CI) surrounding NNTB or NNTH [4,8]. CIs for NNTs can be calculated to inform us about the range of NNT values we may expect. However, CIs for the estimated NNTs are usually given for statistically significant results [8], and this is mainly because of a complication of the NNT calculation when dealing with nonsignificant results (i.e., there is discontinuity when RD is 0).

The NNT measure has been particularly useful in systematic reviews and meta-analyses [3]. However, caution is needed in the NNT calculation as differences in baseline risks, lengths of follow-up, outcome definitions, and clinical settings across the studies included in a meta-analysis can impact the magnitude and direction of NNT [1]. In the meta-analysis context, it is recommended to calculate NNT using an overall treatment effect that remains constant in baseline risk variations. For example, is has been shown that OR and RR effect measures appear to be relatively constant for differences in *CERs* across studies [9]. Caution is also needed when the between-study heterogeneity in the included studies' results is substantial. When the study-specific effect measures vary substantially (e.g., owing to notable differences in baseline risks or in patient characteristics or in study-designs), it may not be advisable to combine the study results into a single overall effect estimate or calculate the respective NNT.

Overall, the NNT is a clinically useful measure for expressing binary and survival outcome results [10], and is frequently used in the published literature [11−13]. Several attempts have also been made to extend NNT for continuous outcomes [14,15], as well as to graphically represent NNTs [4,16,17]. However, knowledge users (such as patients, health care providers, and policy-makers) are faced with a multitude of intervention options and the need to compare several treatments for a clinical condition are required to make informed health care decisions. As such, more complex statistical approaches, such as network meta-analysis (NMA) are required. NMA combines the results of trials that undertake different treatment comparisons [18−20] and is being conducted with increasing frequency in the health care literature [21,22]. The aim of this article is to present graphical approaches of NNTs from NMA to facilitate interpretation of results.

## 2. Number needed to treat in pairwise and network meta-analysis

An NNT can be calculated from the overall RD, RR, and OR effect measures using the following formulas:

$$NNT = \frac{1}{|RD|}$$

$$NNT = \frac{1}{(1 - RR) \cdot CER}$$

$$NNT = \frac{1 - CER + OR \cdot CER}{(1 - RR) \cdot CER \cdot (1 - CER)}$$

where *CER* is the control (or placebo or usual care) event rate, defined as the observed risk of having an event in the control group (ranges between 0 and 1). A barrier to expanding the use of NNT in meta-analysis when an OR or RR is used is the determination of a single *CER* value, as the *CER* will vary for each study included in the meta-analysis. In the following we present potential ways of analyzing *CER* in meta-analysis:

1. The naïve approach, where the sum of events in the control group is divided by the sum of patients in the control group;
2. Median/mean *CER* across all studies containing the control group;
3. Pooled *CER* from a meta-analysis across all studies containing the control group;
4. Expert opinion based, for example, on the patient population included across the studies or on local data from the researcher's own patient population;
5. A range of possible *CER* values, which can be used to compare potential NNT differences.

Of all approaches, the naïve should be avoided, as it ignores study randomization and between-study variability. To derive a pooled *CER* across studies in a meta-analysis, we may need to use transformations [23]. The *CER* follows a binomial distribution and its variance, which is a function of the mean, reaches a maximum value at $CER = 0.5$. While this works well for *CER* around 0.5, when *CER* is closer to 0 or 1, its variance declines to 0, and hence an inverse-variance meta-analysis assigns a very large weight to these studies [23]. Variance-stabilizing transformations help not only to correct this problem in binomial data, but also to obtain a sampling distribution closer to a normal distribution. Two of the most common variance-stabilizing transformations are the logit or double arcsine transformations [23]. Although the logit transformation helps better approximate a normal distribution, the transformed sampling variance can be quite inaccurate. For a *CER* close to 0 or 1, its variance becomes extremely large, whereas for a *CER* close to 0.5, its variance becomes extremely low [23]. Hence, an inverse-variance meta-analysis assigns small weights to studies with small or large *CERs* and large weights to *CERs* around 0.5, irrespective of the sample size. To improve normalizing and variance-stabilizing the *CER* sampling distribution, Freeman and Tukey [24] suggested the double arcsine transformation. A back-transformation on the original *CER* scale can be performed using the approach suggested by Miller [25]. However, it has been suggested not to use the double arcsine transformation for meta-analysis of single proportions because of potential problems with the back-transformation [26]. As alternative, the application of generalized linear mixed models is proposed [26].

Extending the calculation of NNT in NMA, additional considerations to the aforementioned should be made. First, the order of treatments when these are compared in an NMA should be presented in a meaningful and consistent way, to facilitate the *CER* choice and the NNT interpretation. A consistent way could be ordering treatments within comparisons referring to active treatment vs. placebo/usual care or referring to new pharmacological treatment vs. old pharmacological treatment (alternative strategies are needed for nonpharmacological interventions). Let us consider the fictional example of six studies comparing treatments A, B, and C, as shown in Appendix 1. If treatment C is newer than treatment B, which is newer than treatment A, then a presentation of the treatment comparisons evaluated in the six different studies could be B vs. A, B vs. A, B vs. A, C vs. B, C vs. B, and C vs. A. This facilitates the determination of the comparator (e.g., control) group in each case, so as to calculate the study-specific *CER*. For example, in study 1 that compares B and A, a *CER* is defined using evidence from treatment A. A *CER* is defined in a similar way in studies 2 to 6. Second, a *CER* should be defined across multiple treatment comparisons that share the same control (or comparator) group. This may include choosing between a common and comparison-specific *CER*. Selection of the most appropriate assumption will depend on the patient population. However, different assumptions may impact the NNT results. For example, for an $OR = 0.80$, a common $CER = 0.05$ across all treatments vs. control and equal to 0.50 gives an $NNT = 18$, whereas a comparison-specific *CER* gives a notably different $NNT = 104$. In our fictional example in Appendix 1, a common *CER* across treatment comparisons sharing the same comparator group was assumed. Hence, the *CER* for treatment A for the treatment comparison B vs. A ($CER^{A}_{Bvs.A}$) is equal to the $CER^{A}_{Cvs.A}$. In particular, under the common *CER* assumption, we need to estimate 2 *CERs*: $CER^{A}_{Bvs.A} = CER^{A}_{Cvs.A}$ and $CER^{B}_{Cvs.B}$. The *CER* for treatment A was estimated from studies 1, 2, 3 (comparing B vs. A treatments), and 6 (comparing C vs. A treatments), whereas the *CER* for treatment B was estimated from studies 1, 2, 3 (comparing B vs. A treatments), and 4 and 5 (comparing C vs. B treatments). However, under the comparison-specific assumption, we need to estimate 3 *CERs*: $CER^{A}_{Bvs.A}$,$CER^{A}_{Cvs.A}$, and $CER^{B}_{Cvs.B}$. In Appendix 1, we used approach 2 and calculated a mean *CER* across studies comparing the same control group.

Once a meaningful *CER* is calculated, the NNT can be determined using the formulae presented earlier. In Appendix 1, we present NNT using both the OR effect measure ($NNT_{OR}$) and the RD ($NNT_{RD}$) effect measure. To

account for uncertainty around NNT, a CI can be calculated for the NNT values obtained in an NMA. Several approaches have been suggested to calculate CIs for NNTs for results from RCTs [8,27], among which the Daly (or substitution method), the method of variance estimates recovery (MOVER), and the propagating imprecision (PropImp) CIs can also be used for results from meta-analyses [28−30]. For the NNT CI calculation, an appropriate method should be chosen to calculate CIs for the selected effect measure. For a review of methods to obtain CIs for the estimated overall effect from a random-effects meta-analysis see the study by Veroniki et al. [31]. If the chosen effect measure is the RD, then the NNT CI is simply obtained by inverting and exchanging the corresponding RD confidence limits. If the chosen effect measure is the RR or OR, additionally, a meaningful *CER* is required. The Daly CIs start with a CI for the estimated overall treatment effect, and then calculate a CI based on a transformed scale. Although the method is simple to apply, it does not account for the estimation uncertainty of *CER*, when OR and RR effect sizes are used. On the contrary, the MOVER and PropImp approaches allow for a degree of imprecision of both treatment effect and *CER* estimation, and can be used when the estimates of *CER* and treatment effect are independent (e.g., derived from separate studies). This means that the average *CER* derived across the same eligible studies should not be used to calculate an MOVER or PropImp CI for NNT [32].

In NNT, values between −1 and 1 are impossible, and the domain of NNT uses two regions: a) the NNTB region, including the union of 1 (where is the largest possible beneficial treatment effect) to ∞ (no treatment effect), and b) the NNTH region, −∞ (no treatment effect) to −1 (where is the largest possible harmful treatment effect). For example, a nonstatistically significant NNT = 5 with CI −40 and 2 is a combination of the two regions (−∞,−40] and [2, ∞). The suggested presentation of a nonstatistically significant NNT is NNTB = 5 (NNTH 40 to ∞ NNTB 2) [4]. The presentation indicates that an NNTB = 5 is estimated implying that on average five people should receive the treatment for an additional beneficial outcome compared with the control group. However, the uncertainty of this estimation is large with a harmful effect up to NNTH = 40, a less harmful effect up to NNTH = ∞ (no effect; need to treat an infinite number of people to cause or avoid an event), and a more beneficial effect up to NNTB = 2. Because of this limitation and the difficulty in interpreting a nonstatistically significant NNT, many authors do not report a CI for nonstatistically significant NNTs.

In Appendix 1, we present a 95% CI for each NNT using the Daly approach [30]. It should be highlighted that the resulting CIs for $NNT_{OR}$ contain only the OR uncertainty and do not account for *CER* uncertainty. Hence, the NNT estimation is conditional on the assumed *CER*. However, in case of large between-study heterogeneity, we recommend the use of various clinically meaningful *CER* values (e.g., for low- and high-risk patients) to explore differences in NNT. In case a *CER* estimate with CI is available, which is independent of the (network) meta-analysis, the PropImp approach can be used to calculate CIs for NNTs taking into account uncertainty of the *CER* and the OR estimation [29]. In Appendix 1, the NNT values slightly differ when calculated from OR and from RD. This is because different properties are associated with different effect measures, which can affect the NNT value. It should also be considered that small changes in RD close to zero may result in large changes in the estimation of NNT, as RD=0 corresponds to NNT = ±∞.

## 3. Illustrative example

To illustrate different approaches for the graphical representation of NNT (see section 4), we use a published systematic review and NMA on the comparative effectiveness and safety of cognitive enhancers for treating Alzheimer's dementia [33]. The example includes eight dichotomous outcomes and 10 treatments. The network representation of each outcome is presented in Appendix 2. The treatment comparisons including placebo, the estimated ORs, RRs, and RDs in a frequentist NMA (using the *mvmeta* command is Stata) [34], the mean *CERs*, and the estimated NNTs for each outcome and effect measure are provided in Appendix 3. In this example, we used the OR, which was transformed to NNT for the graphical approaches in section 4. To ease interpretation in plots, we present the results on the RD scale after converting the transformed NNTs to RDs. For illustration purposes, we also present in Appendices 4-10 the same graphical approaches using the RD effect measure as estimated in NMA and its conversion to the NNT scale. However, it should be noted that the use of the RD effect measure is not appropriate in this example. Nevertheless, we used this approach here to illustrate the different graphs for RD-based NNTs without switching to another data example. We ordered treatments from oldest to newest by year of availability in Canada [33], used approach 2, and calculated a mean *CER* across studies comparing the control group, and considered a common *CER* value across treatment comparisons including the same, older treatment. Because the aim of this article is to present different ways of depicting NNT, we used a single *CER* value to calculate NNT under the OR (and RR in Cates plot in Appendix 5) effect measure. We calculated a 95% CI for each NNT using the Daly approach [30]. In the following, we infer on whether a treatment is harmful or beneficial based on the NNT scale, that is, a treatment is harmful when NNTH ranges between 1 and ∞, and a treatment is beneficial when NNTB ranges between 1 and ∞.

## 4. Graphical approaches for NNT based on absolute measures

Several graphical ways can be used to present the NNT in an NMA. In this article, we discuss six potential approaches to graphically represent NNT. We also categorize the plots

when a single outcome or multiple outcomes are available in an NMA. The uncertainty around NNT can graphically be depicted in a bar plot and a forest plot. A scatterplot can also be extended to include CIs for NNTs as ellipse regions across treatment comparisons and outcomes [35]. For a comparison of the graphical approaches, see Appendix 11.

### 4.1. Plots for a single outcome

#### 4.1.1. Bar plot

A bar plot can graphically depict the NNT for each treatment comparison or the NNT for the treatments compared with a common comparator (e.g., placebo), as presented in Figure 1.

In Figure 1, we graphically represent the NNT values of six treatments for Alzheimer's dementia vs. placebo for the vomiting outcome [33] of 42 RCTs and 12,997 patients, in a bar plot. According to the NNT point estimates, all treatments but one are suggested as harmful treatments compared with placebo. The only beneficial treatment is memantine (NNTB = 27, 95% CI [NNTH 3, ∞, NNTB 15]), suggesting that 27 patients need to be treated with memantine compared with placebo to prevent one patient from vomiting. However, the point estimate is associated with large uncertainty, where its CI goes from large harm to large benefit. Donepezil, galantamine, and oral rivastigmine are statistically significantly more harmful against placebo. The most harmful treatment among the six treatments evaluated in an NMA vs. placebo is oral rivastigmine (NNTH = 7, 95% CI [5,14]), suggesting that seven patients need to be treated with oral rivastigmine in order for one patient to vomit.

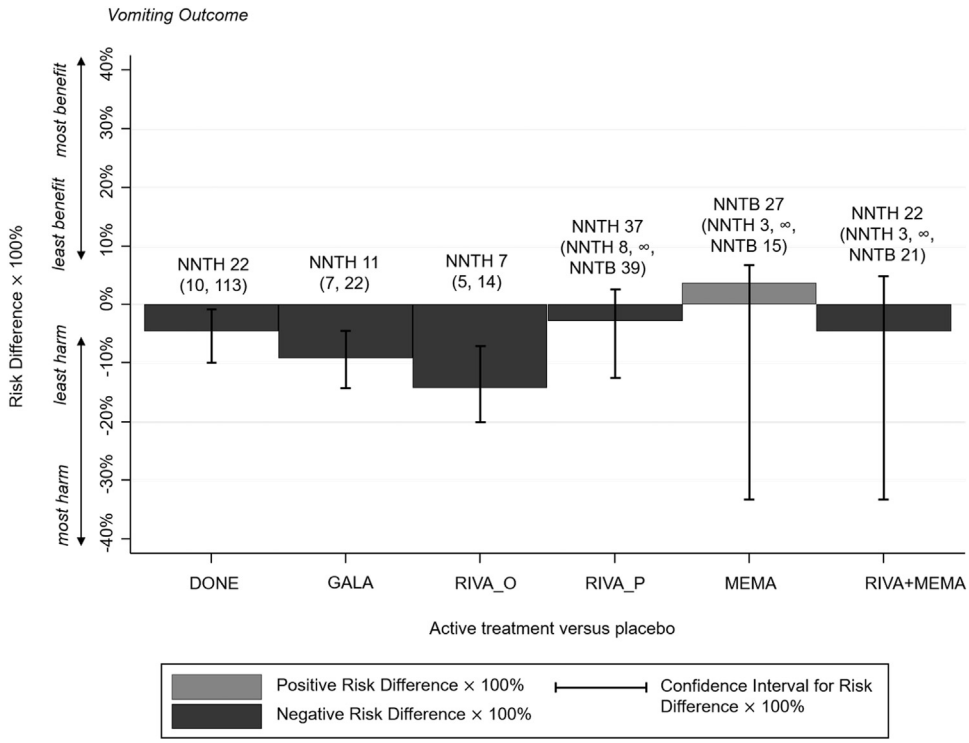


**Fig. 1.** Bar plot of six treatments against placebo for the vomiting outcome. On the x-axis, the treatment comparisons are presented, and on the y-axis, the RD × 100% scale is shown, whereas the NNT values along with their CIs are depicted at the top of each bar. Each vertical bar depicts the RD × 100% value that corresponds to the specific treatment comparison, and each vertical line crossing a bar correspond to the error bars depicting the 95% CI for RD × 100%. Error bars crossing the line of no effect (RD × 100% = 0%) suggest a non-statistically significant result. The greater the area in a bar, the larger the absolute RD, that is, the smaller the NNTB (or NNTH) value, and hence the most beneficial (or harmful) the treatment. We can distinguish between harmful (below the horizontal line of no effect) and beneficial (above the horizontal line of no effect) treatments using dark grey and light grey colored bars, respectively. On the horizontal axis the six treatment comparisons are presented, whereas on the vertical axis the RD × 100% value of each treatment comparison is presented. Each bar represents one of the six (i.e., total number of treatments in vomiting outcome-1) possible comparisons against a common comparator (i.e., placebo). The error bars represent the 95% CI for NNT. Light grey bars represent the number of patients need to be treated to prevent one patient from experiencing the event; dark grey bars represent the number of patients need to be treated in order for one patient to experience the event. Although the NNT scale suggests that memantine is the only beneficial treatment compared to its alternatives in the network, given that memantine does not treat vomiting clinically memantine would be described as the least harmful treatment. *Abbreviations*: CI, confidence interval; DONE, donepezil; GALA, galantamine; MEMA, memantine; NNT, number needed to treat; NNTB, number needed to treat for an additional beneficial outcome; NNTH, number needed to treat for an additional harmful outcome; RD, risk difference; RIVA, rivastigmine; RIVA_O, oral rivastigmine; RIVA_P, transdermal patch rivastigmine.

*4.1.2. Cates plot*

A Cates plot can be used to graphically present the NNT values derived from NMA evidence. The Cates plot shows the average rate of having a good outcome with treatment (green faces), a bad outcome with treatment (red faces), a better outcome with control (crossed green faces), and a change in outcome category if a patient is treated (yellow faces) per treatment comparison. 100 faces are depicted in a Cates plot representing patients treated with the underlying treatment. The more green faces in a section referring to a certain treatment comparison indicate the most beneficial the treatment against its comparator, whereas the more red faces in a section referring to a certain treatment comparison indicate the most harmful the treatment against its comparator.

Assuming a common *CER* across comparisons with mean *CER*=7% across studies, we plotted a Cates plot for each treatment against placebo at http://www.nntonline.net/visualrx/. Figure 2 demonstrates the Cates plot for the vomiting outcome and six treatments against placebo: donepezil, galantamine, oral rivastigmine, transdermal patch rivastigmine, memantine, and rivastigmine + memantine. This plot suggests that the most beneficial treatment is memantine, where 93 patients treated with memantine had a good outcome of not vomiting, four patients, who would vomit without memantine, had a change in outcome and did not vomit after receiving treatment, and three patients had a bad outcome of vomiting even if they were treated with memantine (NNTB = 27, 95% CI [NNTH 3, ∞, NNTB 15]). The least harmful

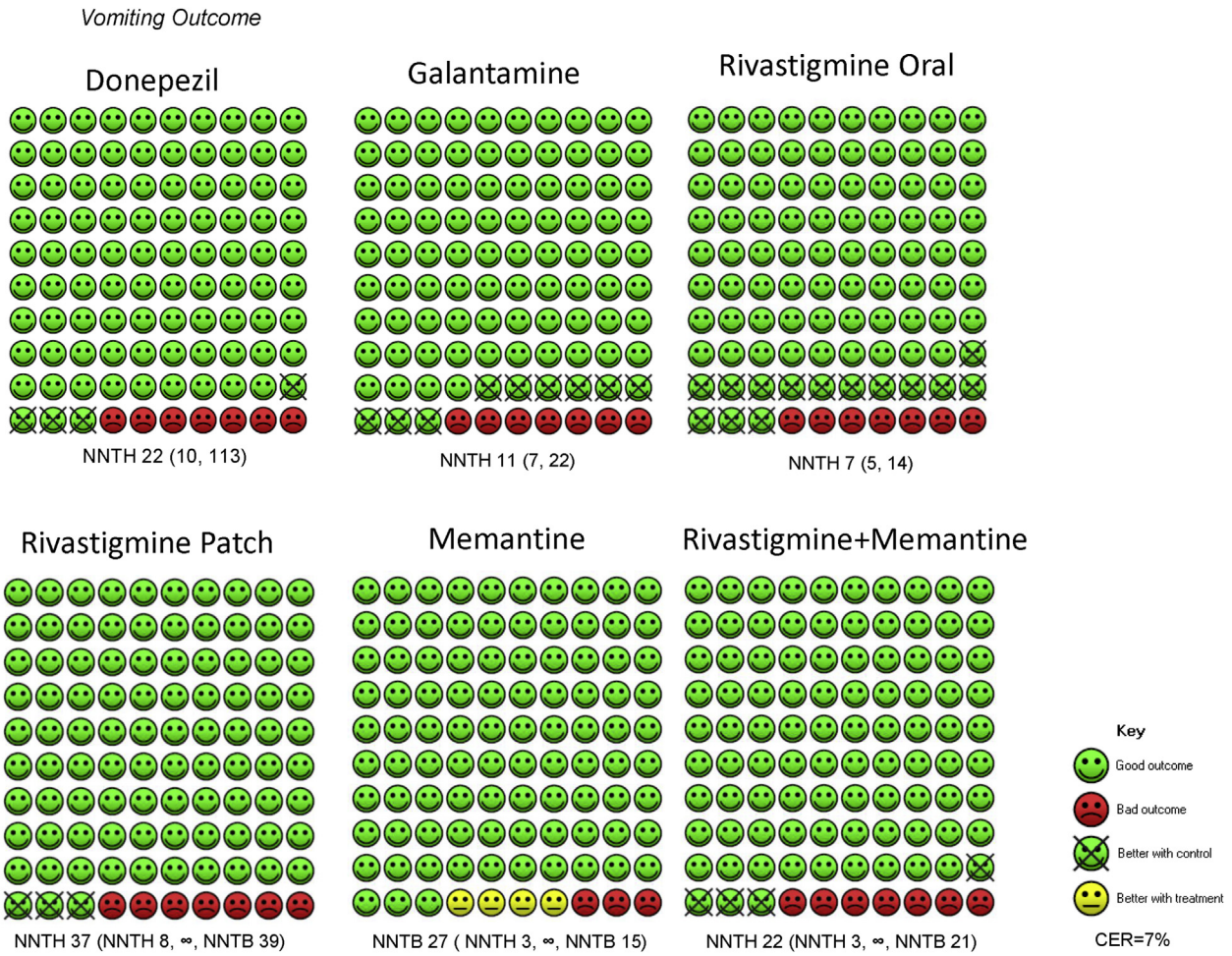


**Fig. 2.** Cates plot representing six treatments against placebo in the vomiting network. Each region corresponds to a different treatment (vs. placebo) and includes 100 faces corresponding to the patients treated with the underlying treatment. Green faces represent patients not vomiting with the underlying treatment; red faces represent patients vomiting with the underlying treatment; crossed green faces represent patients not vomiting with control; yellow faces represent patients that would not vomit if they would be treated with the underlying treatment. The NNT values have been recalculated in http://www.nntonline.net/visualrx/ using the odds ratios estimated in an NMA model and a mean CER = 7%. Although the NNT scale suggests that memantine is the only beneficial treatment compared with its alternatives in the network, given that memantine does not treat vomiting clinically memantine would be described as the least harmful treatment. *Abbreviations*: CER, control event rate; NMA, network meta-analysis; NNT, number needed to treat; NNTB, number needed to treat for an additional beneficial outcome; NNTH, number needed to treat for an additional harmful outcome. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

treatment is transdermal patch rivastigmine, where 38 patients need to be treated with transdermal patch rivastigmine in order for one patient to vomit (NNTH = 37, 95% CI [NNTH 8, ∞, NNTB 39]). The Cates plot suggests that 90 patients treated with transdermal patch rivastigmine had a good outcome of not vomiting, three patients had an adverse event with transdermal patch rivastigmine and their category from a good outcome changed to a bad outcome of vomiting, and seven patients had a bad outcome of vomiting even if they were treated with transdermal patch rivastigmine.

#### 4.1.3. Forest plot

A forest plot can graphically depict the estimated NNT for each treatment comparison. On the x-axis, the $RD \times 100\%$ scale is shown with 0% corresponding to the line of no treatment difference, and on the y-axis, the treatment comparisons are presented. The NNT values along with their CIs are depicted on the left-hand side of the plot next to the RD values. In the forest plot, each treatment comparison is presented by a diamond on the $RD \times 100\%$ scale and a horizontal line extending either side of the diamond depicts a CI for $RD \times 100\%$. The treatment comparisons may be divided into subsets for presentation in a forest plot, such as according to the common comparator in the NMA treatment comparisons.

The forest plot of six treatments against placebo assessed for vomiting in an NMA of patients with Alzheimer's dementia is shown in Figure 3. This plot shows that donepezil, galantamine, and oral rivastigmine are statistically significantly harmful when compared with placebo, and among the three treatments, the highest uncertainty around NNT is observed for donepezil (NNTH = 22, 95% CI [10,113]). Memantine, transdermal patch rivastigmine, and rivastigmine + memantine are associated with nonstatistically significant NNTs and very wide 95% CIs.

### 4.2. Plots for multiple outcomes

#### 4.2.1. Bubble plot

A bubble plot shows the NNT values for all treatment comparisons assessed in an NMA for two outcomes. The plot arranges the presentation of NNT for all treatments on a certain outcome on the x-axis against the treatments included in a second outcome on the y-axis. Let us define that the outcome presented at the lower diagonal part of the plot is outcome 1 (e.g., headache in Figure 4) and the outcome presented at the upper diagonal part of the plot is outcome 2 (e.g., nausea in Figure 4). The direction of the treatment comparisons in outcome 1 (e.g., headache) is defined as treatment at the relevant row (e.g., donepezil) vs. treatment at the relevant column (e.g., placebo).

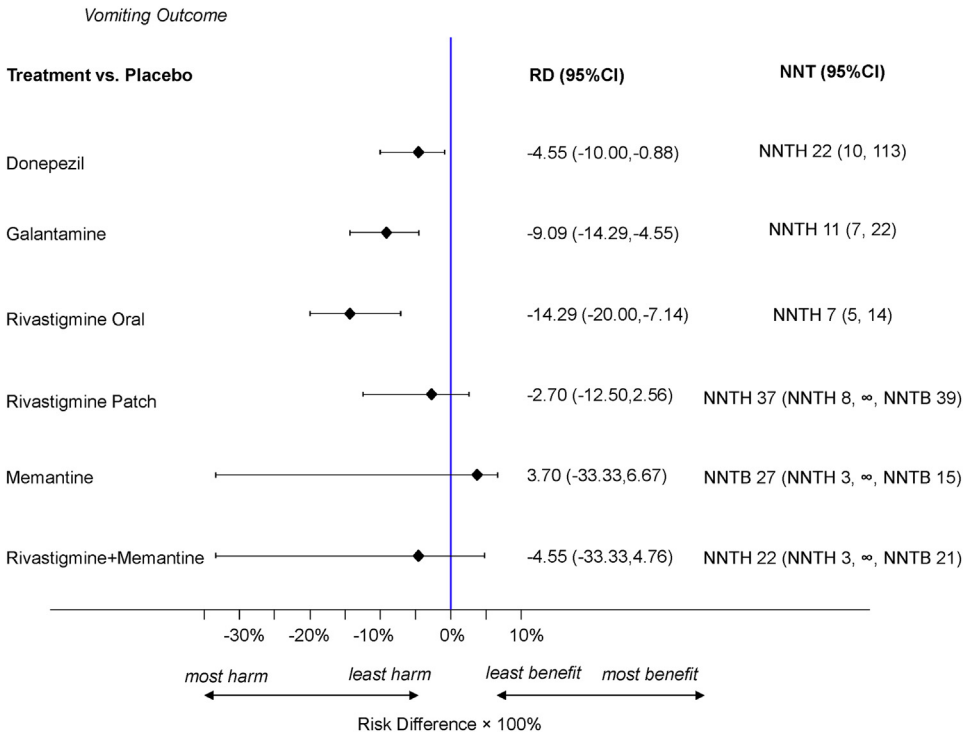


**Fig. 3.** Forest plot for six NMA treatment comparisons against placebo to assess vomiting in Alzheimer's dementia. The RD × 100%, its 95% CI, the corresponding NNT and its 95% CI for each comparison are shown. Note that the pooled or effect measure estimated in NMA has been transformed to NNT, and the NNT has been converted to an RD effect measure, which is presented in this plot. *Abbreviations*: CI, confidence interval; NNT, number needed to treat; NNTB, number needed to treat for an additional beneficial outcome; NNTH, number needed to treat for an additional harmful outcome; RD, risk difference.

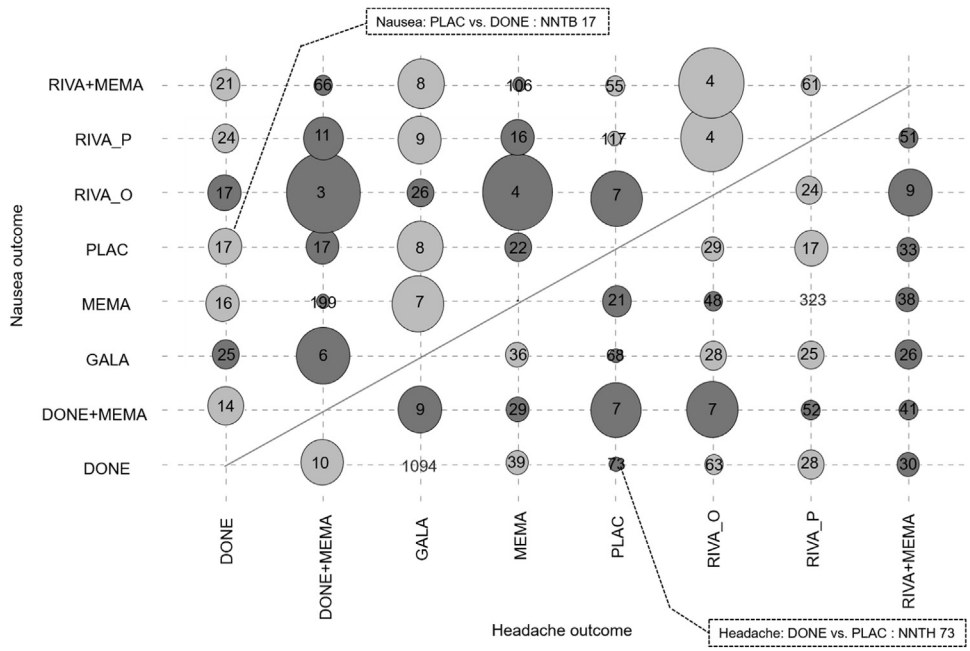


**Fig. 4.** Bubble plot for the eight NMA treatments (including placebo) included in headache (x-axis) and nausea (y-axis) outcomes. The NNT values for each pair of the Alzheimer's dementia treatments according to headache (lower diagonal) and nausea (upper diagonal) are presented. The area of each circle is proportional to the RD × 100% value of each treatment comparison, and the NNT value is presented in the center of each circle. Light grey circles represent the number of patients need to be treated to prevent one patient from experiencing the event; dark grey circles represent the number of patients need to be treated in order for one patient to experience the event. The direction of the treatment comparisons is defined as row treatment vs. column treatment. Although the NNT scale suggests that there are beneficial treatments in the network, given that these treatments do not treat nausea or headache clinically they would be described as the less harmful treatments. *Abbreviations*: DONE, donepezil; GALA, galantamine; MEMA, memantine; NMA, network meta-analysis; NNT, number needed to treat; NNTB, number needed to treat for an additional beneficial outcome; NNTH, number needed to treat for an additional harmful outcome; PLAC, placebo; RD, risk difference; RIVA, rivastigmine; RIVA_O, oral rivastigmine; RIVA_P, transdermal patch rivastigmine.

Similarly, the direction of the treatment comparisons in outcome 2 (e.g., nausea) is defined as treatment at the relevant row (e.g., placebo) vs. treatment at the relevant column (e.g., donepezil). The diagonal of the plot gives no information about NNT. The area in each circle is proportional to the absolute $RD \times 100\%$, and the number in each circle represents the NNT value for the corresponding treatment against the comparator for the specific outcome. However, a challenge with bubble plots is that the smaller the circle, the harder it is to read the NNT value. Each point corresponds to four pieces of information: treatment comparison, magnitude of RD, benefit/harm of treatment, and NNT value. Light grey circles represent the number of patients need to be treated to prevent one patient from having an event, whereas dark grey circles show the number of patients need to be treated in order for one patient to experience a harmful event.

Figure 4 demonstrates the bubble plot for the NNT of eight NMA treatments included in headache (x-axis) and nausea (y-axis) outcomes. In this plot, the NNT values for all NMA treatment comparisons are presented according to headache (lower diagonal) and nausea (upper diagonal) outcomes. For example, for the headache outcome, the comparison row treatment vs. column treatment donepezil vs. placebo has an NNTH = 73, whereas in nausea, the comparison row treatment vs. column treatment placebo vs. donepezil has an NNTB = 17 (equivalent to donepezil vs. placebo: NNTH = 17). The plot suggests that the most beneficial treatment for nausea is donepezil + memantine against placebo (donepezil + memantine vs. placebo: NNTB = 17), but the same treatment is the most harmful treatment for headache against placebo (donepezil + memantine vs. placebo: NNTH = 7). The only beneficial treatment against placebo in headache is rivastigmine + memantine, which is also beneficial in nausea (rivastigmine + memantine vs. placebo, headache: NNTB = 33, nausea: NNTB = 55). The least beneficial treatment in nausea is transdermal patch rivastigmine, which is one of the most harmful treatments against placebo in headache (transdermal patch rivastigmine vs. placebo, headache: NNTH = 17, nausea: NNTB = 117).

### 4.2.2. Scatterplot

For the case of two outcomes, two-dimensional scatterplots can be used, which can be extended to the case of three outcomes with three-dimensional plots. The plot presents both the $RD \times 100\%$ and NNT values for treatments against a common comparator across two (or a maximum

of three) outcomes. Clustering methods can be used to group the NNT performance of treatments according to their efficacy and/or safety [36].

Figure 5 depicts the *RD* × 100% values of seven treatments against placebo for headache and nausea in a scatterplot. The NNT scale is presented on the right and upper scales of the plot. The plot suggests that the only beneficial treatment vs. placebo in both outcomes is rivastigmine + memantine (headache: NNTB = 33, nausea: NNTB = 55). The least beneficial treatment in nausea is transdermal patch rivastigmine against placebo (NNTB = 117), whereas the only beneficial treatment in headache is rivastigmine + memantine. The most harmful treatment in the headache outcome is donepezil + memantine vs. placebo (NNTH = 7), but the same treatment is the most beneficial treatment in the nausea outcome vs. placebo (NNTB = 17).

### 4.2.3. Rank-heat plot

The rank-heat plot can be used for the visual presentation of the NNT values across multiple treatments and outcomes [37]. A rank-heat plot includes N circles with the same center corresponding to the N outcomes assessed in an NMA. The radii included in each concentric circle correspond to T treatment comparisons as assessed in NMA. Instead of presenting all available treatment comparisons, one can present the NMA treatments against a common comparator (e.g., placebo). Each section in the rank-heat plot is colored according to the NNT value of the particular treatment at the corresponding outcome. The NNT scale ranges from NNTH = 1 to ∞ to NNTB = 1 and is transformed using three colors: red (NNTH = 1), yellow (NNTH/NNTB = ∞), and green (NNTB = 1). Although the color of the section is interpretable, the section area does not convey any information. Statistically significant NNT results are depicted in the rank-heat plot by highlighting the borders in the corresponding section. Uncolored sections refer to NMA treatments without data on the outcome within the circle. A star symbol can be used to highlight these cases (see https://rh.ktss.ca/site/nnt).

Figure 6 displays the hierarchy of 10 treatments for Alzheimer's dementia against placebo across eight outcomes according to their NNT values in a rank-heat plot. The NNT scale is presented at the top of the graph. The plot suggests that donepezil + memantine lies among the most beneficial treatments when compared with placebo across most outcomes except for headache with NNTH = 7. However, owing to lack of evidence, we cannot infer the treatment's harm or benefit in bradycardia and diarrhea outcomes. Across all outcomes and treatment comparisons, 10 NNTs are statistically significant and these refer to donepezil, galantamine, and oral rivastigmine treatments.

All plots can present NNTs from NMA, but the bar plot, Cates plot, and forest plot can be cumbersome when a large

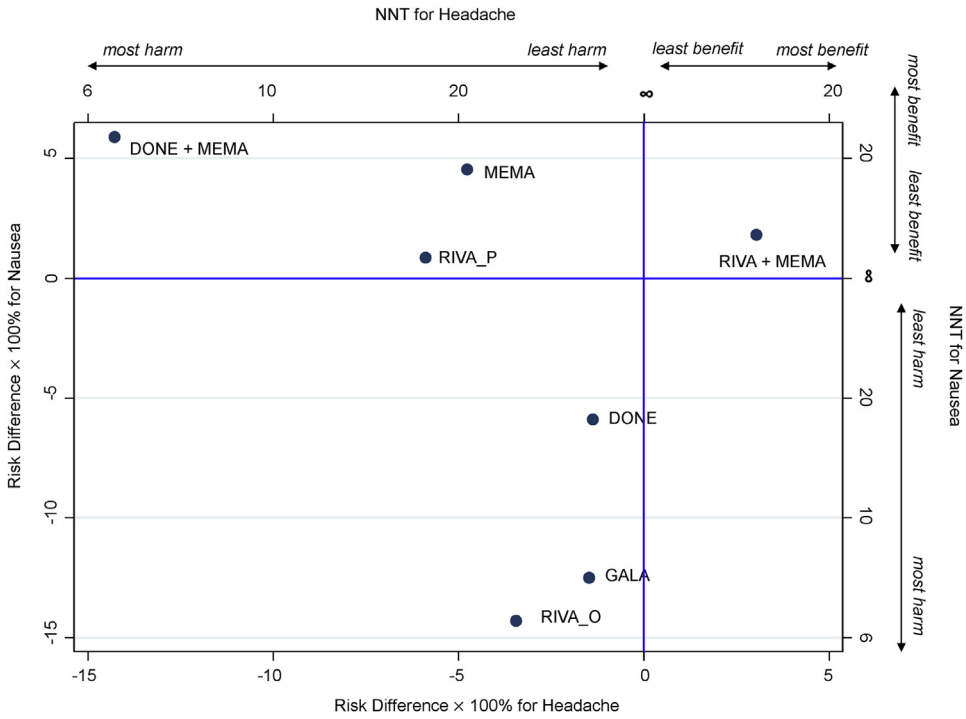


**Fig. 5.** Scatterplot for the NNT of seven treatments vs. placebo for headache (x-axis) and nausea (y-axis) outcomes for patients with Alzheimer's dementia. Treatments lying on the upper right-hand side quarter are more beneficial against placebo for both outcomes. Although the NNT scale suggests that there are beneficial treatments in the network, given that these treatments do not treat nausea or headache clinically they would be described as the less harmful treatments. *Abbreviations*: DONE, donepezil; GALA, galantamine; MEMA, memantine; NNT; number needed to treat; RIVA, rivastigmine; RIVA_O, oral rivastigmine; RIVA_P, transdermal patch rivastigmine.

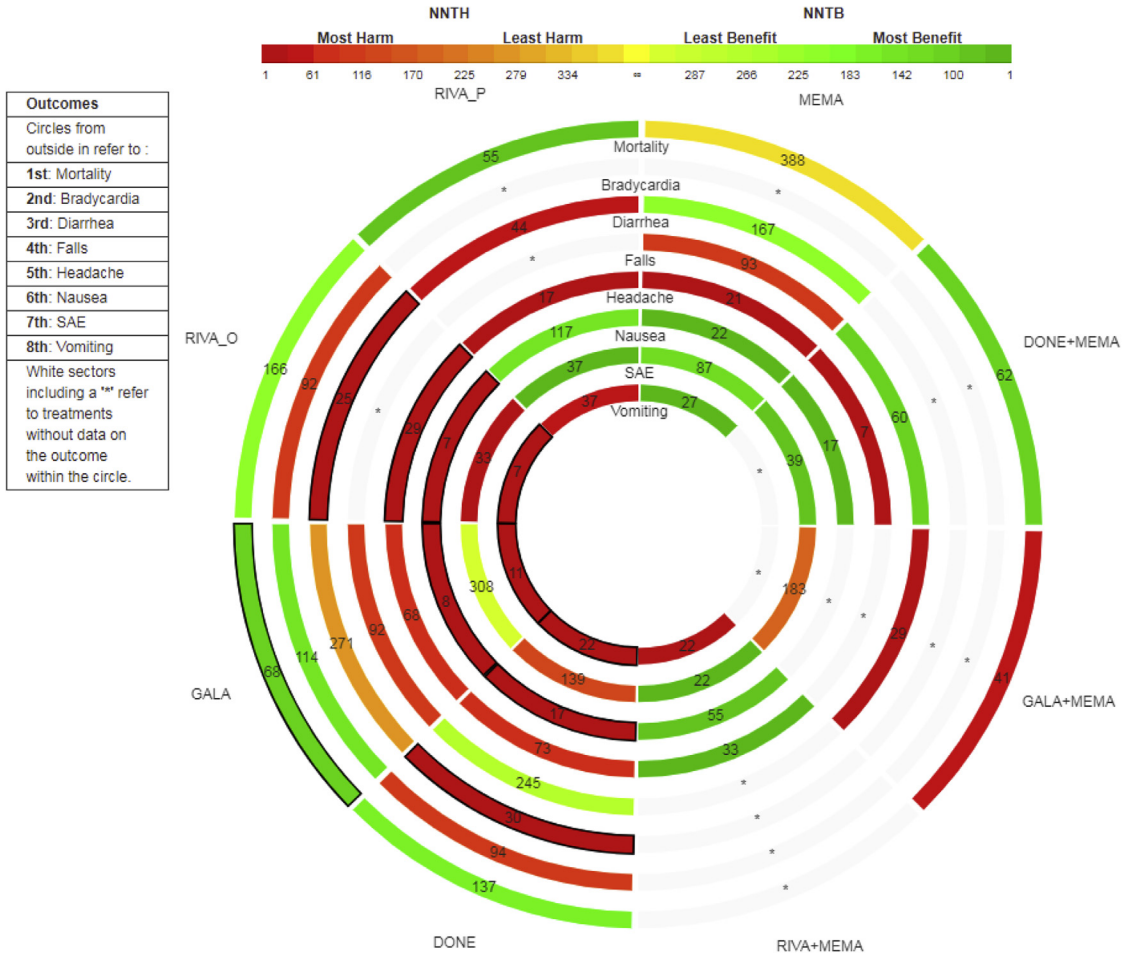


**Fig. 6.** Rank-heat plot of the NNT values of eight treatments in eight different outcomes. Each section is colored according to the NNT value of the corresponding treatment and outcome. The scale consists of the transformation of three colors red (NNTH = 1), yellow (∞), and green (NNTB = 1). Each section includes also the NNT value corresponding to the specific treatment and outcome. Highlighted borders in a section correspond to statistically significant NNT results. Uncolored sections show that the underlying treatment was not included in the NMA for the particular outcome. Although the NNT scale suggests that there are beneficial treatments in the network, given that these treatments do not treat nausea or headache clinically they would be described as the less harmful treatments. *Abbreviations*: DONE, donepezil; GALA, galantamine; MEMA, memantine; NMA, network meta-analysis; NNT, number needed to treat; NNTB, number needed to treat for an additional beneficial outcome; NNTH, number needed to treat for an additional harmful outcome; RIVA, rivastigmine; RIVA_O, oral rivastigmine; RIVA_P, transdermal patch rivastigmine; SAE, serious adverse events. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

number of NNTs is available, especially when more than one outcome is available. An NNT can be particularly helpful when different outcomes with widely different *CER* values (e.g., harmful and beneficial outcomes) are compared, to reflect the different likelihood of each outcome. Similarly, an NNT is helpful for comparing different interventions. A disadvantage of the bubble plot is that the smaller the circle the harder it is to read the NNT value, whereas a scatterplot cannot be produced when different treatments (or treatment comparisons) are included in the studied outcomes. A challenge associated with the Cates plot, bubble plot, and rank-heat plot is that they do not depict the NNT uncertainty, which can impact interpretation. Although the interpretation of nonstatistically significant NNTs is challenging, when we only consider an NNT point estimate or direction of effect that shows benefit (or harm) and do not account for the huge estimation uncertainty (CI is going from large harm to large benefit), our conclusions can be misleading. For example, based on the NNT point estimate, rivastigmine + memantine is suggested as a beneficial treatment when compared with placebo in both headache and nausea outcomes (headache: NNTB = 33, nausea: NNTB = 55). Interpreting only the point estimates rather than the combination of point estimates and uncertainty around them can lead to an erroneous decision-making considering also that nausea and headache are adverse events related to these medications (headache: NNTB 33 [NNTH 2, ∞, NNTB 13], nausea: NNTB 55 [NNTH 7, ∞, NNTB 15]). The Forest plot on the RD scale is probably one of the easiest ways

to present uncertainty around each result, followed by the bar plot.

## 5. Discussion

We recommend the presentation of NNT along with the relevant effect measure and its CI when it is useful to describe the treatment effects in an absolute scale. The NNT values can be presented for all available or selected treatment comparisons (e.g., active treatments vs. placebo) from an NMA. We suggest the presentation of all results using the main effect measure used in the analysis (e.g., OR), and of selected, interesting for the considered research question, results using NNT. An important consideration when calculating NNT is that it may vary according to the effect measure used in meta-analysis or NMA. Therefore, it is important to choose the appropriate main effect measure in meta-analysis or NMA before NNTs are calculated.

When the OR or RR effect measures are used in meta-analysis or NMA, a useful *CER* should be assumed to estimate NNT. As discussed in section 2, several ways exist to select a *CER* value for the NNT calculation, including mean *CER* across studies, pooled *CER* in meta-analysis, expert opinion-based *CER*, and range of possible *CERs*. In the presence of small to moderate heterogeneity, we suggest the use of several *CER* values (e.g., for low- and high-risk patients) to estimate an NNT and assess robustness of results. By means of these assumed *CERs*, we can calculate NNTs and their CIs (e.g., based on the Daly approach [30], but this neglects the estimation uncertainty of *CER*). In case external estimation of *CER* is available (e.g., from registry data), which is independent on the data used in the underlying meta-analysis or NMA, and in case the RR is the main effect measure, we can use the MOVER approach to estimate NNT and its CI taking into account the *CER* estimation uncertainty [28]. In case the OR is the main effect measure, the PropImp approach can be used [29]. In NMA, additional considerations are required to calculate NNTs, which include a meaningful and consistent order of treatments to facilitate the *CER* choice for NNT calculation and interpretation, and a selection among different *CER* assumptions (i.e., common or comparison-specific *CER*). Because NNT is dependent on *CER* and study duration, the comparison of multiple treatments for a specific outcome through NNT may be difficult. The selection among different *CER* assumptions depends on the clinical field and the nature of the treatments assessed in an NMA. If different *CER* values (e.g., derived from control arms of the included studies) influence the NNT calculation, then this should be considered when interpreting the NNT. We suggest that NNTs always be interpreted along with the relevant treatment effects and their CIs estimated in a meta-analysis or NMA. Ranking statistics derived from an NMA model can also be used as complementary information to NNTs to compare treatments within each outcome of interest [38].

In this article, we discussed the NNT calculation for dichotomous data. However, the estimation of NNT can also be helpful for continuous outcomes. A way to calculate NNT for continuous data can be by converting a standardized mean difference to OR and then calculate NNT [5] or by dichotomizing the continuous data and then calculate an effect size for dichotomous data. In any case, we suggest to graphically represent NNT to ease interpretation. However, it should be considered that the NNT interpretation may differ according to the NMA considerations, including selection of effect size and *CER*, as well as *CER* assumption across multiple comparisons.

## Availability of data and materials

The data sets used and/or analyzed during this study are available from the corresponding author on reasonable request.

## Acknowledgments

The authors would like to thank Alexandros Fyraridis, Andrea Jimenez, and Megan Mak for building the rank-heat plot web-tool (https://rh.ktss.ca/site/nnt). They thank Shazia Siddiqui, Myanca Rodrigues, Krystle Amog, and Sinit Michael for helping format the manuscript.

Authors' contributions: A.A.V., R.B., P.G., S.E.S., and A.C.T. conceived and designed the study. A.A.V. conducted the analysis. A.A.V. wrote the first draft manuscript and the other authors edited the manuscript.

S.E.S. is funded by a Tier 1 Canada Research Chair in Knowledge Translation. A.C.T. is funded by a Tier 2 Canada Research Chair in Knowledge Synthesis.

## Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.03.007.

## References

[1] Bender R. Number needed to treat: overview. In: Balakrishnan N, Brandimarte P, Everitt BS, Molenberghs G, Ruggeri F, Piegorsch W, editors. Wiley StatsRef: Statistics Reference Online. Chichester: John Wiley & Sons, Ltd; 2017:1–7.

[2] Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988;318:1728–33.

[3] McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. Ann Intern Med 1997;126:712–20.

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

[5] Version 5.1.0 [Updated March 2011]. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions.

The Cochrane Collaboration; 2011. Available at. http://handbook.cochrane.org. Accessed April 12, 2019.
[6] Stang A, Poole C, Bender R. Common problems related to the use of number needed to treat. J Clin Epidemiol 2010;63:820−5.
[7] Christensen PM, Kristiansen IS. Number-needed-to-treat (NNT)–needs treatment with care. Basic Clin Pharmacol Toxicol 2006;99:12−6.
[8] Bender R. Calculating confidence intervals for the number needed to treat. Control Clin Trials 2001;22:102−10.
[9] Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol 2002;31:72−6.
[10] 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.
[11] Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995;310:452−4.
[12] Sackett DL. On some clinically useful measures of the effects of treatment. Evid Based Med 1996;1:37−8.
[13] Nuovo J, Melnikow J, Chang D. Reporting number needed to treat and absolute risk reduction in randomized controlled trials. JAMA 2002;287:2813−4.
[14] da Costa BR, Rutjes AW, Johnston BC, Reichenbach S, Nuesch E, Tonia T, et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. Int J Epidemiol 2012;41:1445−59.
[15] Furukawa TA, Leucht S. How to obtain NNT from Cohen's d: comparison of two methods. PLoS One 2011;6:e19070.
[16] Lesaffre E, Boon P, Pledger GW. The value of the number-needed-to-treat method in antiepileptic drug trials. Epilepsia 2000;41:440−6.
[17] Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. BMJ 2002;324:827−30.
[18] Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005;331:897−900.
[19] Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med 1996;15:2733−49.
[20] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105−24.
[21] Petropoulou M, Nikolakopoulou A, Veroniki AA, Rios P, Vafaei A, Zarin W, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. J Clin Epidemiol 2017;82:20−8.
[22] Zarin W, Veroniki AA, Nincic V, Vafaei A, Reynen E, Motiwala SS, et al. Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. BMC Med 2017;15:3.
[23] Barendregt JJ, Doi SA, Lee YY, Norman RE, Vos T. Meta-analysis of prevalence. J Epidemiol Community Health 2013;67:974−8.
[24] Freeman MF, Tukey JW. Transformations related to the angular and the square root. Ann Math Stat 1950;21:607−11.
[25] Miller JJ. The inverse of the Freeman-Tukey double arcsine transformation. Am Stat 1978;32:138.
[26] Schwarzer G, Rücker G. Freeman-Tukey Double Arcsine Transformation in Meta-Analysis of Single Proportions. Biometrisches Kolloquium Biometrie: Gelebte Vielfalt. Frankfurt: University of Frankfurt; 2018.
[27] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 1998;17:873−90.
[28] Newcombe RG. MOVER-R confidence intervals for ratios and products of two independently estimated quantities. Stat Methods Med Res 2016;25:1774−8.
[29] Newcombe RG. Propagating imprecision: combining confidence intervals from independent sources. Commun Stat Theory Methods 2011;40:3154−80.
[30] Daly LE. Confidence limits made easy: interval estimation using a substitution method. Am J Epidemiol 1998;147:783−90.
[31] Veroniki A, Jackson D, Bender R, Kuss O, Langan D, Higgins J, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis, 2018. Res Syn Meth 2019;10:23−43.
[32] Newcombe RG, Bender R. Implementing GRADE: calculating the risk difference from the baseline risk and the relative risk. Evid Based Med 2014;19:6−8.
[33] Tricco AC, Ashoor HM, Soobiah C, Rios P, Veroniki AA, Hamid JS, et al. Comparative effectiveness and safety of cognitive enhancers for treating Alzheimer's disease: systematic review and network metaanalysis. J Am Geriatr Soc 2017;66:170−8.
[34] White IR. Multivariate random-effects meta-analysis. Stata J 2009;9:40−56.
[35] Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. Res Synth Methods 2015;6:157−74.
[36] Romesburg HC. Cluster analysis for researchers. Belmont, CA: Lifetime Learning Publications; 1984. xiii, 334.
[37] Veroniki AA, Straus SE, Fyraridis A, Tricco AC. The rank-heat plot is a novel way to present the results from a network meta-analysis including multiple outcomes. J Clin Epidemiol 2016;76:193−9.
[38] Veroniki AA, Straus SE, Rucker G, Tricco AC. Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? J Clin Epidemiol 2018;100:122−9.