When does a difference make a difference? Interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed

L. Citrome,¹ T. A. Ketter²

SUMMARY

Although great effort is made in clinical trials to demonstrate statistical superiority of one intervention vs. another, insufficient attention is paid regarding the clinical relevance or clinical significance of the observed outcomes. Effect sizes are not always reported. Available absolute effect size measures include Cohen’s d, area under the curve, success rate difference, attributable risk and number needed to treat (NNT). Of all of these measures, NNT is arguably the most clinically intuitive and helps relate effect size difference back to real-world concerns of clinical practice. This commentary reviews the formula for NNT, and proposes acceptable values for NNT and its analogue, number needed to harm (NNH), using examples from the medical literature. The concept of likelihood to be helped or harmed (LHH), calculated as the ratio of NNH to NNT, is used to illustrate trade-offs between benefits and harms. Additional considerations in interpreting NNT are discussed, including the importance of defining acceptable response, adverse outcomes of interest, the effect of time, and the importance of individual baseline characteristics.

The savvy practitioner is well aware that a clinical trial result with a low ‘p-value’, even as low as \( p < 0.00001 \), does not necessarily mean that the result is clinically relevant. For appraisal of clinical relevance, or clinical significance, we require measures of effect size. A small effect size difference between two interventions is not usually clinically compelling, even though the result may be statistically significant. To be relevant for providers and patients, an observed difference in a trial must make a meaningful clinical difference in real-world practice, whether it is comparing two antihypertensive medications for decreases in blood pressure, or diagnostic accuracy when comparing two test procedures. One well-known effect size metric is the number needed to treat (NNT), and its counterpart, number needed to harm (NNH) (1–3). Less well-known is the ratio of NNH to NNT, or likelihood to be helped or harmed (LHH) (4). In this review, we offer a set of suggestions on how to interpret NNT, NNH and LHH, using various examples of previously published work. Key concepts include the need to carefully scrutinise the definitions of potential benefits (such as response) being measured, as well as the accompanying potential harms (such as side effects). Absolute (as opposed to relative) rates of outcomes and the effect of time must also be explicitly considered.

Available absolute effect size measures include Cohen’s d, area under the curve, success rate difference, attributable risk and NNT (2). Of all of these measures, NNT is arguably the most clinically intuitive and helps relate effect size difference back to real-world concerns of clinical practice. NNT answers the question ‘How many patients would you need to treat with Intervention A instead of Intervention B before you would expect to encounter one additional positive outcome of interest?’ Complementing NNT is NNH; NNH answers the question ‘How many patients would you need to treat with Intervention A instead of Intervention B before you would expect to encounter one additional outcome of interest that you would like to avoid?’

Calculating NNT is straightforward. To calculate the NNT for an outcome for Intervention A vs. Intervention B, the process is as follows:

Review criteria

This is a narrative review based on the authors’ prior work in this area.

Message for the clinic

Number needed to treat (NNT) and number needed to harm (NNH) can be useful in medical decision-making. Of importance are defining acceptable response and adverse outcomes of interest. The concept of likelihood to be helped or harmed (LHH), calculated as the ratio of NNH to NNT, can be used to illustrate trade-offs between benefits and harms.
Interpretation of number needed to treat

$$f_A = \text{frequency of outcome for Intervention A}$$

$$f_B = \text{frequency of outcome for Intervention B}$$

Attributable Risk (AR) = $$f_A - f_B$$

NNT = 1/AR

By convention, when not presenting fractions, NNT is rounded upwards to the next higher whole number. This serves two purposes: the first is that NNT is a clinician’s tool and clinicians treat whole (as opposed to fractional) patients and secondly, rounding downwards may exaggerate the difference and render the NNT estimate as too optimistic. A useful fictional example is that of Drug A vs. Drug B for the outcome of remission where remission rates for each intervention are 50% and 20%, respectively. The relative advantage of the first over the second intervention is 2.5, but stating that Drug A is 2.5 times better than Drug B does not help place this difference into clinical perspective. To answer the question ‘How many patients would you need to treat with Drug A instead of Drug B before you would expect to encounter one additional remitted patient?’ requires calculation of NNT, where

$$\text{NNT} = \frac{1}{0.50 - 0.20} = \frac{1}{0.30} = 3.33,$$

and rounded up to 4. As it turns out this is a reasonable NNT and Drug A (depending on clinical context, including relative tolerability compared with Drug B) may be preferred because of the greater likelihood of benefit as a therapeutic advantage may be expected to be seen on a frequent (every fourth patient) basis in a clinic where this decision point is confronted frequently.

Reporting the NNT without the rates that were used to calculate it can be misleading. Knowledge of the rates of the outcomes used to calculate NNT will help distinguish between a NNT of 10 when it is calculated from rates of 20% vs. 10% as opposed to when it is calculated from rates of 80% vs. 70% – two very different clinical treatment scenarios.

Calculating the 95% confidence interval for NNT is more complicated, but can be easily done using a spreadsheet program or on-line calculator. The formula can be found elsewhere (3). Calculating NNH uses the same procedures as calculating NNT. Alternate nomenclature is NNT for benefit (NNTB) and NNT for harm (NNTH).

Acceptable values for NNT and NNH

When appraising new medications, the only systematic clinical studies commonly available are those conducted for registration or regulatory purposes. In psychiatry, these are usually multicentre, randomised, double-blind, placebo-controlled trials. For dichotomous outcomes such as response/non-response, a ‘good enough’ NNT is usually a single digit (i.e. at least 10% better than placebo, resulting in a NNT less than 10) and the lower the number the better. For example, United States Food and Drug Administration approved agents for Bipolar Disorder generally have single-digit NNTs (5). It is commonly expected that successful interventions for ‘very treatable’ acute conditions (such as acute agitation) would have NNTs vs. placebo in the range of 2–3. Interventions for ‘somewhat treatable’ acute or chronic conditions (such as osteoarthritic pain) commonly have NNTs vs. placebo in the range of 4–6. Higher NNTs, including those greater than 9, may be acceptable under certain conditions, such as less clinically urgent situations with a sufficiently high need for tolerability that it mitigates efficacy limitations. For example, in mild acute bipolar depression, lamotrigine has a suboptimal low double-digit NNT that may be mitigated by a desirable even higher double-digit NNH, which can make it preferable to quetiapine, which has a desirable single-digit NNT for response, but a problematic single-digit NNH for sedation (6). Thus, in some instances, a low NNT does not guarantee that a choice is a clinically useful one; as a low NNH for an adverse outcome of concern can be so burdensome that it may negate an otherwise promising low NNT. Another example of where interventions with a higher NNT would be considered would be in more difficult to treat conditions where other interventions have already failed and the remaining choices are limited, such as treatment-resistant major depressive disorder where certain adjunctive second-generation antipsychotics may be used to augment inadequate antidepressant response (7).

An acceptable NNH for drug vs. placebo depends on the outcome in question. To be clear, not all outcomes are equivalent in clinical importance. An error that is sometimes made is to pair an important beneficial outcome, such as remission, with a common but relatively minor adverse outcome such as transient mild nausea – this leads to under appreciation of the likelihood benefit compared with harm. A single-digit NNH may be acceptable if the adverse event is mild or moderate, does not lead to discontinuation, and is temporary or causes little distress, and does not pose a serious health risk (e.g. transient mild nausea or sedation), or if a treatment has good (single-digit NNT) efficacy and there is a compelling need for efficacy that mitigates the low NNH tolerability limitation. For example, the need for efficacy may mitigate the tolerability limitations of quetiapine in patients with acute severe bipolar depression (5,6). Higher (double-
(single-digit) NNHs are desired for more significant adverse events. A NNH in the range of 10–100 may be acceptable for adverse events that may lead to discontinuation, but are not associated with serious immediate health risks, or when alternatives do not have a better profile; an example can be moderate weight gain. Even higher (triple-digit) NNHs are usually required for adverse events that pose a significant health risk; for very severe adverse outcomes, NNH values as high as 1000 may be necessary to be acceptable; examples include acute haemorrhage or serious rash.

Thus, for psychiatric practice, clinically meaningful differences between treatments may be anticipated when there is at least a 10% advantage with respect to potential benefits (single-digit NNT) and no more than a 10% disadvantage with respect to potential harms (double-digit NNH), although details of the clinical context will commonly yield personalised exceptions to such a general guideline.

As noted, it is desirable for the NNH to be greater than the NNT, so that benefits are encountered more frequently than harms. In some instances, NNH needs to be very much greater than NNT when comparing a desired beneficial outcome with a very severe adverse event, for example agranulocytosis vs. stabilising treatment-resistant schizophrenia. There are many situations, however, that NNH is only modestly greater than NNT, for example comparing a desired beneficial outcome with an adverse event that is usually mild or moderate, but that may still lead to discontinuation. Examples include mild or transient sedation vs. response to an antidepressant medication. For some treatments, NNH may actually be lower than NNT when comparing a desired beneficial outcome with an adverse event that is usually mild or moderate, but that is usually temporary and does not lead to discontinuation, such as a mild dry mouth vs. response to an antidepressant medication. Indeed, in some depressed or manic patients with prominent insomnia, for a limited time, a degree of sedation may even be considered a benefit more than a harm; although later on when such patients have improved and are attempting to return to normal function, the same degree of sedation could be considered a harm more than a benefit.

In summary, although we generally want NNT to be low and NNH to be high, in some specific instances, high NNH advantages may mitigate some high NNT disadvantages, whereas in some clinical contexts, low NNH disadvantages may make some low NNTs unacceptable in spite of their therapeutic effect advantages; on the other hand, low NNT advantages may mitigate some low NNHs, while high NNT disadvantages may make some high NNHs unacceptable. It can be useful to calculate LHH, the ratio of NNH to NNT. A LHH much greater than 1 is the norm when comparing a desired outcome, for example remission, with a very severe adverse event. A LHH a little greater than 1 is usually observed for acceptable interventions when comparing a desired outcome with an adverse event that is usually mild or moderate, but that may still lead to discontinuation. LHH less than or equal to 1 is usually only acceptable when comparing a desired outcome with an adverse event that is usually mild or moderate, but that is usually temporary and does not lead to discontinuation, or there is a particularly urgent need for benefit (efficacy) that mitigates an otherwise prohibitive risk of harm (side effects).

**Definition for response matters**

For a NNT to be clinically meaningful, it must relate to an outcome that is clinically important to the patient. For example, in the treatment of schizophrenia, it has become customary to categorise response as achieving a 20% reduction in symptoms as measured on a standardised rating scale. Unfortunately, a 20% reduction may be considered a degree of improvement that is barely perceptible to clinicians and patients. It would be of interest then to calculate NNH for higher thresholds for response in patients with schizophrenia, such as 30%, 40% and 50% reductions in symptoms from baseline (the response threshold in acute mania and acute depression is commonly a 50% reduction in symptoms). This was done with lurasidone, a new second-generation antipsychotic medication, by pooling together the results of five 6-week randomised controlled schizophrenia trials (8). Using a 160 mg/day dose, the NNTs vs. placebo for response were 3, 4, 5 and 8, for the response thresholds of 20%, 30%, 40% and 50%, respectively. As expected, it is more difficult to achieve the higher thresholds of symptom reduction. However, it can be easily argued that a 50% reduction in symptoms matters a great deal more clinically than a 20% decrease. Missing from the discussion so far are the absolute rates of response. These must also be considered. In our example, the placebo response rates were 45%, 32%, 24% and 15% for the response thresholds of 20%, 30%, 40% and 50%, respectively. Response rates for lurasidone 160 mg/day were 78%, 63%, 46% and 29%, respectively. In all instances, lurasidone 160 mg/day was superior to placebo, no matter the response threshold, but absolute response rates were lower as response threshold increased.
What adverse outcomes are of interest?

For a NNH to be meaningful, the adverse outcome must be clinically relevant. Using our prior example of lurasidone (8), the NNH vs. placebo for Parkinsonism was 20. If a patient is particularly sensitive to this adverse effect, then it becomes more important. If a patient desires to avoid that outcome as much as possible, then the occurrence may lead to discontinuation. LHH can be useful here to illustrate trade-offs. If a 30% decrease in symptoms is selected as a relevant threshold for response, the NNT is 4, giving a LHH of 20:4 or 5.0. This LHH of 5.0 for lurasidone 160 mg/day for response vs. Parkinsonism can be interpreted that 'lurasidone treatment at 160 mg/day was five times more likely to help (≥ 30% decrease in symptoms) than to harm (Parkinsonism) the patient.' Other alternative medications may have different harms that may be of interest to the clinician or the patient. In the same report, data were available for olanzapine 15 mg/day. For olanzapine 15 mg/day, NNH vs. placebo for ≥ 30% decrease in symptoms from baseline was four and NNH for at least 7% weight gain at the study endpoint was four, yielding a LHH of 1.0, which can be interpreted as ‘olanzapine treatment at 15 mg/day was equally likely to help (≥ 30% decrease in PANSS) as to harm (at least 7% weight gain) the patient.’

Another example is that of vilazodone, a new antidepressant medication (9). NNT vs. placebo for response for vilazodone 40 mg/day in the 8-week pivotal trials was eight, as measured by an improvement of at least 50% in depressive symptoms. The NNH vs. placebo for discontinuation because of an adverse event was 27. LHH = NNH/NNT = 27/8 = 3.4. Thus, vilazodone 40 mg/day was 3.4 times more likely to result in therapeutic response than discontinuation because of an adverse event. The rates used to calculate the NNT and NNH can be found in the report (8). For the desired outcome of remission (symptoms lower than an absolute threshold), the applicable NNT was 14. LHH in this case was 27 / 14 = 1.9; so, vilazodone 40 mg/day was 1.9 times more likely to achieve remission than result in a discontinuation because of an adverse event. One potentially relevant example of LHH was for the outcome of achieving response vs. encountering nausea, LHH = 6/8 = 0.75. This means that the likelihood of achieving a response was actually less than the likelihood of encountering nausea; in other words, nausea was 1.3 times more likely to be encountered than a therapeutic response. This may not be clinically prohibitive if the nausea is time-limited and easily managed.

Time

The effect of time on benefits such as treatment response can be profound. Moreover, the longer the clinical trial, the greater the opportunity for harms such as adverse events to occur or resolve. Thus, the time points when NNT and NNH are calculated need to be noted. An example of how NNT values can change can be found in a review of inhaled loxapine for the treatment of agitation (10). In patients with schizophrenia administered inhaled loxapine 10 mg, response (at least 40% reduction in symptoms) at 10 min was observed in 19% of those receiving the active medication vs. 6% receiving placebo, yielding a NNT of eight, whereas at 120 min, response rates were 70% and 38% for inhaled loxapine and placebo, respectively, yielding a lower NNT of 4.

Another example is that of duloxetine for the treatment of major depressive disorder (11). NNT for remission (depression rating scale score no higher than a prespecified threshold) based on weeks of treatment in patients receiving 60 mg/day of duloxetine decreased steadily from a NNT of 79 after 1 week to a NNT of 6 after 9 weeks.

The importance of individual baseline risk

Although challenging to ascertain in studies of medications for psychiatric disorders, baseline severity of illness may have profound effects on calculations of NNT. A more clear-cut example comes from the treatment of hypertension where the NNT to prevent one event of death, stroke, or myocardial infarction in a 5-year period is 141 for the use of antihypertensives for mild or moderate diastolic hypertension (90–109 mmHg) (12). The NNT for this outcome is three when using this intervention for severe diastolic hypertension (110–129 mmHg).

Additional caveats

As noted earlier, there are other effect size measures of importance, but these are beyond the scope of this article. However, one of these, area under the curve deserves special mention and the reader is referred to Kraemer and Kupfer’s excellent review of the size of treatment effects and their importance to clinical research and practice (13).

Although rating scale scores are often used to assess clinically important differences, there may be discords with patient perceptions; for example, some depressed patients who are categorised as being in remission according to a rating scale score may not consider themselves to be in remission (14). A
review of the concepts and methods of determining minimum clinically important differences that take into account different perspectives of what is important can be found in (15).

Conclusions

NNT and NNH can be used to quantify the effect sizes of clinically relevant benefits and harms in clinical trials. Moreover, LHH can quantify benefit: harm trade-offs that are crucial in real-world clinical decision-making. An important limitation of NNT and NNH is that these metrics are limited to dichotomous (rather than continuous) outcomes. Nevertheless, NNT and NNH, measured in ‘patient units’ compared with commonly reported continuous metrics (such as changes in symptom rating scale scores) are more clinically relevant. As demonstrated in the case of lurasidone, it is important to select the most clinically meaningful improvement threshold on a continuous metric (such as a psychosis rating scale) when converting to a dichotomous metric (such as antipsychotic response) for NNT analysis. Analysis of vilazodone data illustrates an important potential challenge when selecting the appropriate outcomes for the determination of LHH – specifically, the risk of transient nausea, although common, may not be necessarily prohibitive considering the increased likelihood of achieving response, unless discontinuation because of nausea is relevant for that particular patient. The importance of time when calculating NNT is evident for inhaled loxapine for agitation in schizophrenia and duloxetine for treatment of major depressive disorder. Although not well characterised for psychiatric interventions, baseline risk can impact NNT calculations in a profound way as demonstrated in the treatment of persons with mild or moderate vs. severe diastolic hypertension.

As described above, there are clearly specific instances in which NNT, NNH and LHH appear very useful. Nevertheless, additional research is needed to further our understanding of optimal statistical approach(s) for the diverse and complex benefit vs. harm assessments encountered in real-world clinical practice.

References

7 Citrome L. Adjunctive aripiprazole, olanzapine, or quetiapine for major depressive disorder: an analysis of number needed to treat, number needed to harm, and likelihood to be helped or harmed. Postgrad Med 2010; 122: 39–48.
8 Citrome L. Lurasidone for the acute treatment of adults with schizophrenia: what is the number needed to treat, number needed to harm, and likelihood to be helped or harmed? Clin Schizophr Relat Psychoses 2012; 6: 76–85.
9 Citrome L. Vilazodone for major depressive disorder: a systematic review of the efficacy and safety profile for this newly approved antidepressant – what is the number needed to treat, number needed to harm and likelihood to be helped or harmed? Int J Clin Pract 2012; 66: 356–68.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Teaching slides.

Paper received December 2012, accepted January 2013