Original research

Comparison of relative change with effect size metrics in Alzheimer’s disease clinical trials

Terry E Goldberg,1 Seonjoo Lee,2,3 Davangere P Devanand,1 Lon S Schneider4

ABSTRACT

Background Per cent slowing of decline is frequently used as a metric of outcome in Alzheimer’s disease (AD) clinical trials, but it may be misleading. Our objective was to determine whether per cent slowing of decline or Cohen’s d is the more valid and informative measure of efficacy.

Methods Outcome measures of interest were per cent slowing of decline; Cohen’s d effect size and number-needed-to-treat (NNT). Data from a graphic were used to model the inter-relationships among Cohen’s d, placebo decline in raw score units and per cent slowing of decline with active treatment. NNTs were computed based on different magnitudes of d. Last, we tabulated recent AD anti-amyloid clinical trials that reported per cent slowing and for which we computed their respective d’s and NNTs.

Results We demonstrated that d and per cent slowing were potentially independent. While per cent slowing of decline was dependent on placebo decline and did not include variance in its computation, d was dependent on both group mean difference and pooled SD. We next showed that d was a critical determinant of NNT, such that NNT was uniformly smaller when d was larger. In recent AD associated trials including those focused on anti-amyloid biologics, d’s were below 0.23 and thus considered small, while per cent slowing was in the 22–29% range and NNTs ranged from 14 to 18.

Conclusions Standardised effect size is a more meaningful outcome than per cent slowing of decline because it determines group overlap, which can directly influence NNT computations, and yield information on the likelihood of minimum clinically important differences. In AD, greater use of effect sizes, NNTs, rather than relative per cent slowing, will improve the ability to interpret clinical trial results and evaluate the clinical meaningfulness of statistically significant results.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Per cent slowing of decline has become a widely used metric in describing the results of Alzheimer’s disease clinical trials, including those related to anti-amyloid immunotherapies.

WHAT THIS STUDY ADDS

⇒ However, per cent slowing can be largely independent of a standardised effect size, such as Cohen’s d. Recent trials have claimed seemingly impressive per cent slowing in the range of 22–29%, but d’s have been small and under 0.24 and numbers-needed-to-treat (NNT) have ranged from 14 to 18. D is the more informative metric because it directly indicates group separation, NNT and clinically important differences.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ We propose that greater use of standardised effect sizes, less or no use of per cent differences in decline and use of NNTs, as well as determining consensus minimum clinically important differences when they become well-defined, will improve the field’s interpretation of clinical trial results and clinical meaningfulness for patients who are affected by these disabling cognitive disorders.

AD clinical trial outcomes may be presented as a percentage slowing of decline, for example, the drug slowed decline by x% at the 18-month endpoint compared with placebo. This type of relative measure is often reported as a primary outcome in abstracts, scientific lectures and public presentations. Such a comparison may be attractive to clinicians and investors because it implies a slower disease progression rate for the active treatment compared with placebo. However, relative change can be misleading, and cannot be used alone for assessing outcomes and effectiveness. Here we also consider standard statistical metrics: (1) The mean difference between treatments and (2) the standardised effect size (ES), relative outcomes and their relationships to number-needed-to-treat (NNT) and area under the receiver operator characteristics curve (AUC), a measure of group separation. We also discuss the need to express outcomes as standardised effects, as well as to consider the minimum clinically important difference (MCID) between treatment and control to make inferences about effectiveness.

A key reason that per cent difference is not informative is that it is a relative measure from which a magnitude of effect cannot be determined. For example, a 20% less decline at the endpoint could mean scores of 40 versus 50 for placebo with an absolute difference of 10; or mean scores of 8 versus 10, with a difference of 2. Relative change cannot consider the magnitude or variance of the outcome measures as would a widely established statistic (eg, t-test, F test, beta coefficient) or a standardised ES. Standardised ESs express an effect such as a mean difference in terms of its variance. Examples include Cohen’s d, Hedges’ g and z scores, all of which take
Comparisons of per cent slowing, d and NNT in four recent clinical trials

We computed respective ESs and NNTs from three recent amyloid antibody secondary prevention trials: The aducanumab and lecanemab trials that resulted in accelerated Food and Drug Administration (FDA) approvals for these biologics, and a phase 3 trial of donanemab4 9 will likely receive FDA approval.

These latter three studies reported estimates of the regression models without relevant statistics (eg, t-statistics or F-statistics with a df) and standardised ESs such as Cohen’s d, partial eta-squared and/or Cohen’s f2. Thus, based on the reported information and the statistical analysis models, we estimated the range of the df using the Kenward-Roger approximation and Satterthwaite’s method. Since the variance of the follow-ups was not reported, we simulated data sets using the reported model estimates for the parameters (ie, changes in the placebo group and changes in the treatment group, SD at baseline); the residuals’ SD at the follow-up varied from 0.25 to 1. To be conservative, we only simulated the data set including baseline and final time points, while the reported analyses included all intermediate time points, which would increase the df, resulting in smaller ESs. Thus, we emphasise that the estimated range of Cohen’s d we derived is possibly larger than the direct estimates of the model and hence is liberal. Given the t-statistics we estimated, we converted these to Cohen’s d using the following conversion formula: Cohen’s $d = 2t / \sqrt{df}$.

To provide context, we also included a large recent multimodal lifestyle intervention trial, FINGER, that reported per cent slowing of decline.9 The FINGER trial reported d using a modified intention-to-treat analyses with df specified. Our method yielded near identical results to that in the publication.

For NNTs we used the d for each trial and assumed a response rate of 0.68 or 0.66 in the placebo group for the three anti-amyloid trials (see above) as their AD samples, which included Mild Cognitive Impairment (MCI) with positive biomarkers for AD or mild AD, were similar at baseline.

The study conforms to Standards for Quality Improvement Reporting Excellence guidelines.10

RESULTS

Per cent slowing of decline

Figure 1 shows the mean difference between treatment and placebo groups fixed at $-0.5$ in keeping with observed CDR-Sum of Boxes (SB) differences. Placebo decline varied from $-0.5$ to $-3$ points, and pooled SD from 1 to 8 in whole numbers. These SDs generate a wide range of Cohen’s d’s from 0.06 to 0.5. We graphically display along three dimensions the results for d (dependent on pooled SD), per cent slowing of decline (based on the formula in the Methods) and placebo mean change from baseline. For any given d, per cent slowing of decline can vary widely based on the magnitude of placebo decline, and the two measures are largely independent of each other.

Nnnumber-needed-to-treat

We show NNT across a range of Cohen’s d’s from 0.10 to 1.0 and three placebo response rates (0.20, 0.35 and 0.68) in figure 2. It can be observed that the smaller the d, the larger the NNT for any given placebo response rate.

Results for an AUC-dependent formula demonstrate the same trend in online supplemental figure 1. Larger d’s were associated with smaller NNTs. AUC is dependent on d because the latter has the property of identifying degree of group separation:

\[
NNT = \frac{1}{\Phi(d/\sqrt{2})} - CER, \quad \Phi = \text{cumulative distribution, } \Psi = \text{inverse of cumulative distribution and } CER = \text{response rate in control group.}
\]

We estimated NNT using a placebo event rate of 0.20, 0.35 (frequently viewed as the frequency of placebo response rate in a wide variety of psychiatric and neurologic conditions) and 0.68 in the placebo group across a wide range of d’s. The latter value was based on the Kaplan-Meier curve presented for the CLARITY-AD trial for lecanemab that showed that approximately 68% of patients in the placebo group did not decline by 0.5 or more points on the Clinical Dementia Rating Scale (CDR) over the 18-month trial (ie, had an advantageous outcome).4 We determined NNT using a web-based application.5

We also determined the NNT developed for continuous data: NNT = 1/2AUC-1.6 This formula is dependent on AUC, a measure of group separation derived from the receiver operating characteristic curve, and not on response rate. AUC represents the probability that an individual selected at random in the active treatment group has a better score (ie, less decline) than an individual selected at random from the control group. It can be derived from d using the following formula:

\[
AUC = \Phi(d/\sqrt{2}) \quad \text{where } d \text{ is Cohen’s d statistic and } \Phi = \text{cumulative distribution function.}
\]
Neurodegeneration

larger d’s are uniformly associated with greater group separation as can be seen in online supplemental figure 2.

Comparison of results from major anti-amyloid secondary prevention trials in MCI and AD

As shown in table 1, the significance levels vary markedly among the FINGER trial and amyloid antibody trials with aducanumab (EMERGE), donanemab (TRAILBLAZER ALZ2) and lecanemab (CLARITY). As expected, increased sample size was associated with greater significance. The Cohen’s d’s for CDR-SB, however, ranged from 0.16 to 0.23, all in the small to small medium ES range. The range of d, df’s and t’s for these studies are in online supplemental table 1. The NNTs ranged from 14 to 18 across these three studies. The FINGER trial was also associated with a small effect size (d=0.13), but with a comparatively large per cent slowing of decline.

DISCUSSION

We have shown that per cent difference in decline and ES can be independent using graphic derivations. Thus, the same per cent difference in decline value may be associated with a wide range of ESs and conversely, the same ES can be associated with a wide range of per cent slowing of decline. This is because per cent slowing does not consider the variances of the treatment and control groups nor their pooled SD. Indeed, the magnitude of placebo decline can be a critical determinant of the difference in per cent slowing of decline.

We go on to show that a standardised ES is more meaningful clinically because it directly influences NNT computations along with response rates. NNT cannot be derived from per cent slowing of decline. Larger d’s were associated uniformly with smaller NNTs, as depicted in figure 2. Online supplemental figure 2 using an AUC-based formula demonstrated the same trend. Irrespective of whether formulae for deriving NNT are

Figure 1  The figure has three axes: Cohen’s d, per cent slowing of decline and placebo change in raw scores. For any given d (curved lines) holding the difference between groups constant and varying the pooled SDs, yielded a wide range of per cent slowing values. Two examples make this clear. For 25 per cent slowing of decline (and a placebo change of 2 units) multiple Cohens d exist as shown in the blue vertical line. See also the second blue line representing a 50 per cent slowing of decline. Conversely, for a given Cohen’s d=0.25, per cent slowing of decline can range from 0 to near 100 per cent, as based on placebo change in raw scores.
from continuous variables that include (or do not include) response rates, the basic trend is clear: the larger the $d$, the less group overlap and larger the AUC, the smaller the NNT. NNT is a clinically important measure and the figure and table show that ES is key for determining NNT.

With respect to real-world clinical trial examples, the small ESs and large NNTs in three completed anti-amyloid secondary prevention trials considered to be ‘positive’ for drug versus placebo raises questions about the clinical meaningfulness of these treatments for older adults with MCI or AD. In these trials, all $d$’s were below 0.23 and NNTs 14 or above, despite seemingly encouraging slowing of decline greater than 22% compared with placebo. These results, as well as multiple earlier negative anti-amyloid trials, have led some investigators to question amyloid as a target or single target.11

We emphasised that ES is more informative than per cent difference in decline as a metric. Another metric that we did not discuss is also relevant, namely MCID in which the magnitude of changes in cognition and function are meaningful, relevant and observable to clinicians, patients and/or caregivers. While it is often reported in the health science literature that $d$’s between 0.35 and 0.50 will be associated with clinically relevant change,12 13 there is a lack of consensus in the field on what constitutes an MCID for various measures of cognition when treated

Figure 2  Numbers-needed-to-treat (NNTs) examined as a function of Cohen’s $d$ and event rate in the placebo group. Here we examined NNT at multiple $d$’s and event rates. CER (control event rate, that is, response rate in the placebo group) represents the proportion of advantageous outcomes in the placebo group. For any given CER, the larger the $d$, the smaller the NNT.

Table 1  Population characteristics, trial methods and outcomes of selected recent AD primary and secondary prevention clinical trials with positive results

<table>
<thead>
<tr>
<th>Study</th>
<th>Population</th>
<th>N treatment/N control</th>
<th>Duration (months)</th>
<th>Outcome</th>
<th>Difference vs control</th>
<th>P value</th>
<th>Per cent slowing</th>
<th>d</th>
<th>NNT*</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINGER</td>
<td>Cognitively unimpaired</td>
<td>554/565</td>
<td>24</td>
<td>NTB†</td>
<td>0.04</td>
<td>0.03</td>
<td>25</td>
<td>0.13</td>
<td>#</td>
</tr>
<tr>
<td>Aducanumab EMERGE high dose</td>
<td>Early AD‡</td>
<td>547/548</td>
<td>18</td>
<td>CDR-SB</td>
<td>−0.39</td>
<td>0.01</td>
<td>22</td>
<td>0.16</td>
<td>18</td>
</tr>
<tr>
<td>Donanemab TRAILBLAZER-ALZ 2 combined group</td>
<td>Early AD‡</td>
<td>794/838</td>
<td>18</td>
<td>CDR-SB</td>
<td>−0.70</td>
<td>&gt;0.001</td>
<td>29</td>
<td>0.23</td>
<td>14</td>
</tr>
<tr>
<td>Lecanemab CLARITY</td>
<td>Early AD‡</td>
<td>898/897</td>
<td>18</td>
<td>CDR-SB</td>
<td>−0.45</td>
<td>0.0005</td>
<td>27</td>
<td>0.21</td>
<td>15</td>
</tr>
</tbody>
</table>

# unknown response rate.
* Number-needed-to-treat.
† Neuropsychological Test Battery reported in z score units.
‡ Early AD=prodromal AD and mild AD dementia.
AD, Alzheimer’s disease; NNT, number-needed-to-treat.
as a change in raw values for cognition and function and how it should be applied (eg, at the group level and/or individual level where the proportions of advantageous MCID case outcomes could be contrasted between the active and placebo groups). Liu et al14 showed that recent findings in clinical trials, including use of anti-amyloid antibodies, that demonstrated significant group differences did not approach the threshold necessary for potentially clinically relevant difference. For example for CDR-SB, the MCID was found to be 0.98 for prodromal AD and thus current effects in anti-amyloid trials would not be ‘difference makers’ to patients or caregivers.14 15 Nevertheless, it should be acknowledged that there are no consensus metrics for an acceptable MCID for commonly used measures in AD trials despite suggestions from several groups.15–18 This is why we propose ES and NNT as essential and MCID, if established, as added validation of a therapeutic effect.

Further consideration of NNTs in the anti-amyloid trials is also warranted. An NNT of 14 as found in the donanemab trial indicates that in a hypothetical population of 1000 treated individuals, 71 will have more favourable outcomes than those in an untreated group of 1000. Thus, 929 individuals will have had exposure to the drug with benefits no greater than that observed in the untreated group’s individuals. They would also be subject to any adverse events associated with the drug (eg, amyloid-related imaging abnormalities (ARIA)) and the potential financial duress imposed by treatment costs.

There have also been proposals to use other relative metrics, including per cent slowing based on individual cases or a metric involving delay in progression in units of time (‘time saved from decline’). However, these suffer from the same problem as per cent slowing in that they do not account for variance.

Another implication of our work is that trials for early AD are planned to detect small ES differences between treatment and control, and this contributes to the frequency of uncertain outcomes and apparent ‘trends’ in outcomes. These can result in failures to replicate, as with the phase 3 aducanumab, solanezumab and gantenerumab trials, with outcomes less than the lower limits of clinician or observer resolution. Moreover, as they are planned (powered) for small effect sizes they also require large sample sizes, and allow for few dropouts and implicitly recognise that only a small minority of participants will likely benefit.

Statistical significance alone is insufficient to indicate that an intervention makes a clinically meaningful difference as the p value is partly a function of sample size and reflects only the likelihood that the distributions of the outcomes is not attributable to random chance, that is, a p value is not a measure of ES. For instance, the lecanemab CLARITY trial had very small p values, generated by a large N (approximately 900 each in the treatment and placebo group), as its Cohen’s d was similar to the other trials. Indeed, any effect larger than null (ie, zero) effect can be demonstrated to be statistically significant with a large enough sample size.

One potential ‘medico-sociologic’ criticism of our study is the concern that patients ‘need something’ and the pharmaceutical industry needs incentives to continue work in this disease. We take issue with this position because of the aforementioned small treatment effect sizes, unknown long-term outcomes and serious adverse event rates, though we appreciate that there is also some merit in this view as some individuals may experience large positive outcomes.19 19 Investigators may also say that the study met its specified endpoint at p<0.05 without consideration of effect size and NNT, which are clinically meaningful metrics. But if use of relative statistics (ie, differences in per cent decline between groups) masks an otherwise trivial ES, does not lead to clinically meaningful functional gains, exposes people to unfavourable side effects and at the societal level has large economic costs, the results can be misleading to patients and clinicians and lead to general frustration with implementation of new treatments approved by regulatory authorities.20 21 We have focused on efficacy only but recognise that the risk of adverse effects such as ARIA and cerebral volumetric reductions may further lower the benefit-to-risk ratio for new treatments.

Summary and recommendations for quality improvement

Parametric statistics are conducted on mean differences, accounting for variance. Statistical power analyses consider both mean differences and variance using ES. These traditional, well-established statistical measures are not conducted on per cent slowing of decline. In this paper we demonstrated through derivations, real-world examples and thought experiments that the use of per cent difference in decline can be largely unrelated to effect size. ES directly influences NNT and is therefore a more informative metric with respect to clinical meaningfulness. We propose that greater use of standardised ESs, less or no use of per cent differences in decline and use of NNTs, as well as determining consensus MCIDs when they become well-defined, will improve the field’s interpretation of clinical trial results and clinical meaningfulness for patients who are affected by these disabling cognitive disorders.

Contributors TG initiated the study and wrote the first draft of the manuscript. SL made the graphic, devised the statistical approach for determining Cohen’s d and made revisions in drafts. DPD and LSS substantially revised all drafts of the manuscript. TG acted as guarantor of this work.

Funding This work was funded by the following National Institute on Aging grants: P30AG066530, R01AG051346, R01AG052440, R01AG055422, R01AG062578 and R01AG062687.

Competing interests LSS reports personal fees from AC Immune, Alpha-cognition, Athira, Corium, Cortelyse, BioVie, Eli Lilly, GW Research, Lundbeck, Merck, Neurim, Novo-Nordisk, Otsuka, Roche/Genentech, Cognition Therapeutics, Takeda; grants from Biohaven, Biogen, Eisai, Eli Lilly and Novartis. DPD reports research support from the National Institute on Aging, Alzheimer’s Association, is a scientific adviser to Acadia, TauRx, Corium, Genentech, and is a member of the Data and Safety Monitoring Board of BioXcel. TG and SL have no conflicts of interest to disclose.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD
Terry E Goldberg http://orcid.org/0000-0002-2455-7247

REFERENCES
2 Funakawa TA. From effect size into number needed to treat. Lancet 1999;353:1680.
20 Knopman DS, Jones DT, Greicius MD. Failure to demonstrate efficacy of aducanumab: an analysis of the EMERGE and ENGAGE trials as reported by Biogen. Alzheimers Dement 2021;17:696–701.