The Lancet Psychiatry

Alzheimer's disease clinical trials need to demonstrate minimum clinically important differences

--Manuscript Draft--

Manuscript Number:	thelancetpsych-D-21-00380R1
Article Type:	Personal View (Unsolicited)
Keywords:	Alzheimer disease; dementia; clinical trials; minimal clinically important difference
Corresponding Author:	Kathy Y. Liu University College London UNITED KINGDOM
First Author:	Kathy Y. Liu
Order of Authors:	Kathy Y. Liu
	Lon S. Schneider
	Robert Howard
Manuscript Region of Origin:	UNITED KINGDOM
Abstract:	Summary Deciding on the smallest change in an outcome that constitutes a clinically meaningful treatment effect, i.e. the minimum clinically important difference (MCID), is fundamental to interpreting clinical trial outcomes, making clinical decisions, and designing studies with sufficient statistical power to detect any such effect. There is no consensus on MCIDs for outcomes in Alzheimer's disease (AD) trials, but the US Food and Drug Administration (FDA)'s consideration of aducanumab clinical trials data has exposed the uncertainty of the clinical meaning of statistically significant but small improvements. Although MCIDs for outcomes, including Clinical Dementia Rating – Sum of Boxes (CDR-SB) and Mini-Mental State Examination (MMSE) in AD have been reported, the FDA guidelines, drafted in 1989 to facilitate regulatory approval of "substantially effective" antidementia drugs, do not specify quantified minimum differences. While it is important that regulatory requirements encourage drug development and approval, without MCIDs, sponsors are motivated to power trials to detect statistical significance for only small and potentially inconsequential effects on clinical outcomes. MCIDs benefit patients, family, caregivers and healthcare systems and should be incorporated into clinical trials and AD drug development guidance.

Alzheimer's disease clinical trials need to demonstrate minimum clinically important differences

Authors: Kathy Y. Liu*, Lon S. Schneider, Robert Howard

Affiliations:

Division of Psychiatry, University College London, London, UK (K. Liu, MRCPsych; *Professor* R. Howard M.D.)

Department of Psychiatry and the Behavioral Sciences, and Department of Neurology, Keck School of Medicine, and Leonard Davis School of Gerontology of the University of Southern California, Los Angeles, CA, USA (*Professor* L. Schneider, M.D.)

*Corresponding author. Address: UCL Division of Psychiatry, 6th Floor Maple House, 149 Tottenham Court Road, W1T 7NF, UK. Email: kathy.liu@ucl.ac.uk

Summary:

Deciding on the smallest change in an outcome that constitutes a clinically meaningful treatment effect, i.e. the minimum clinically important difference (MCID), is fundamental to interpreting clinical trial outcomes, making clinical decisions, and designing studies with sufficient statistical power to detect any such effect. There is no consensus on MCIDs for outcomes in Alzheimer's disease (AD) trials, but the US Food and Drug Administration (FDA)'s consideration of aducanumab clinical trials data has exposed the uncertainty of the clinical meaning of statistically significant but small improvements. Although MCIDs

for outcomes, including Clinical Dementia Rating – Sum of Boxes (CDR-SB) and Mini-Mental State Examination (MMSE) in AD have been reported, the FDA guidelines, drafted in 1989 to facilitate regulatory approval of “substantially effective” antidementia drugs, do not specify quantified minimum differences. While it is important that regulatory requirements encourage drug development and approval, without MCIDs, sponsors are motivated to power trials to detect statistical significance for only small and potentially inconsequential effects on clinical outcomes. MCIDs benefit patients, family, caregivers and healthcare systems and should be incorporated into clinical trials and AD drug development guidance.

Main text:

Clinical trial outcomes for neuropsychiatric conditions, which form the basis for drug marketing decisions, are generally presented as quantitative differences between treatment groups on relevant symptom scales. Deciding on the smallest change in an outcome that constitutes a clinically meaningful treatment effect, i.e. the minimum clinically important difference (MCID), is fundamental to interpreting trial outcomes, making clinical decisions, and designing studies with sufficient statistical power to detect such an effect. This has become particularly important in the interpretation of data from drugs currently under investigation for treatment of dementia. In this paper we ask if it is now time to include agreed MCIDs in the design, analysis and interpretation of Alzheimer’s disease (AD) clinical trials?

Since the cholinesterase inhibitors (donepezil, rivastigmine, galantamine) and memantine obtained regulatory approval in the late 1990s and early 2000s, no further approved treatments for AD have materialized. Recent excitement surrounding Biogen's Biologics License Application for aducanumab, an anti-amyloid antibody, has been dampened by uncertainty over its effectiveness and controversy over the US Food and Drug Administration's (FDA) and Biogen's interpretation of the clinical trials data [1–3]. The FDA Peripheral and Central Nervous System Drugs Advisory Committee voted nearly unanimously on November 6, 2020 against approval (10 against with 1 abstention) [4]. Nevertheless, the possibility that the FDA might reject this recommendation and approve aducanumab for marketing raises again the important question: how should we objectively define whether a dementia treatment is clinically effective?

The same FDA Advisory Committee addressed this issue in 1989 [5], recommending, in part, that a 3-point difference between drug and placebo groups on the 11-item Alzheimer's Disease-Cognitive Subscale (ADAS-Cog11) represented a clinically meaningful difference. A statistically significant difference on a prespecified neuropsychological outcome alone, however, was considered insufficient to indicate that an intervention makes a clinically meaningful difference as the p-value is the likelihood that such a difference is attributable to random chance [6]. The p-value itself is not a measure of effect size. Indeed, any effect larger than no (i.e. zero) effect can be demonstrated to be statistically significant with a large enough sample size[7]. Yet, there remains no consensus or agreement on minimum clinically important differences (MCID) for outcomes in AD trials.

The FDA has long considered a treatment for mild to moderate AD dementia (corresponding to stages 4 and 5 in their 2018 draft guidance[8]) to be substantially effective if there is improvement on a 'core' symptom (e.g. a measure of cognition) and a global clinical measure (e.g. a clinician's judgement of change) or a functional measure (e.g. activities of daily living) [9]. For studies including mild cognitive impairment (MCI) patients, or stage 3[8], the FDA requires only statistically significant change on a pre-specified composite measure that includes cognition and daily function combined, as demonstration of substantial effectiveness. Moreover, in 2013 the agency specifically recommended the Clinical Dementia Rating – Sum of Boxes (CDR-SB) [10] as a composite measure that had demonstrated validity and reliability for this purpose [11,12]. No quantified minimum differences were specified, but the rationale is that such a composite measure serves as an indicator of change in both the 'core' or cognitive outcome, i.e., memory, orientation, judgment and problem solving, and in global or daily function, i.e., community affairs, home and hobbies, so that a treatment that delivers statistical significance on the CDR-SB as a pre-specified primary outcome is, by definition, clinically meaningful [9, 13]. The European Medicines Agency (EMA) has a similar approach [14]. But, has this bar for effectiveness been set too low, and have the – at best – modest effects of existing treatments served to downgrade our expectations for future treatments?

The current Biologics License Application for aducanumab provides some insight. Two identically designed and simultaneously executed trials were undertaken in early-stage AD (i.e., MCI and mild dementia due to AD) but were stopped early due to perceived

futility before half the participants had an opportunity to complete the trials. On post hoc analysis, the higher dose group of the EMERGE trial was judged to have shown statistically significant reduced decline at the 78-week endpoint, with the following differences between placebo and aducanumab groups favouring aducanumab: -0.39 points on the CDR-SB, 0.6 points on the Mini-Mental State Examination (MMSE), -1.4 points on the ADAS-Cog13, and 1.7 points on the Alzheimer's Disease Cooperative Study - Activities of Daily Living (ADCS-ADL-MCI)[15]. The companion trial, ENGAGE, did not show statistical significance on any of the outcomes: 0.03 on the CDR-SB, -0.1 points on the MMSE, -0.59 points on the ADAS-Cog13 and 0.7 points on the ADCS-ADL-MCI [15]. The negligible effects in ENGAGE clearly indicate no clinically meaningful effect for aducanumab. The very small mean differences favoring aducanumab in EMERGE, however, raise the question of whether these statistically significant outcomes were clinically meaningful.

A recent study estimated the MCID for clinical outcomes using anchor-based (change in outcome linked to clinical opinion) and distribution-based (MCID calibration based on the variation across participants) approaches, stratified by severity of cognitive impairment [16]. It found that MCIDs increased with disease severity. For MCI and mild AD, differences of 0.98 and 1.63 points for CDR-SB and 1.26 and 2.32 points for MMSE represented clinically meaningful change. Another study, also using anchor-based methodology that linked scores to clinicians' assessment of clinically meaningful change in cognitive, functional and behavioral domains, reported that the MCID for the ADAS-Cog11 in mild AD patients was 3 points[17]. Meeting MCID thresholds, however, are not requisites for the

FDA concluding that a trial shows substantial effectiveness or authorizing marketing approval [18]. Notably, with the aducanumab trials, neither the CDR-SB difference at -0.39 nor the MMSE at 0.6 points reached MCID thresholds (Table 1).

The FDA, however, per its own guidance, can consider a single positive well-controlled trial that is supported by 'confirmatory evidence' to be substantial evidence of effectiveness without considering mean difference or effect size [18]. Applying this interpretation to the aducanumab trials created considerable controversy. This was especially so given that the identical ENGAGE trial failed to show any statistically significant benefits over placebo, and, indeed, mean effects numerically favored placebo on the CDR-SB and MMSE. The FDA's decision on approval is expected by June 7, 2021.

There is no gold-standard method for determining MCIDs, and each approach has limitations. In anchor-based approaches, the external measure of change (or anchor) is usually subjective, and definitions of what constitutes meaningful change may differ between clinicians and patients or caregivers. The MCIDs in Table 1 represent what constitutes a clinically meaningful decline, as patients' longitudinal change from baseline were anchored to clinicians' assessment of meaningful change [16,17]. However, clinical trial outcomes are used to detect a treatment benefit, and the threshold for worsening may not equate to the threshold for improvement [19]. Relatively larger changes may be interpreted as clinically important at the individual level, whereas relatively smaller changes may be considered important at the group level, so the application of MCIDs to group means might set the bar high. The FDA supports anchor-based methods to

establish what constitutes meaningful individual-level change, which defines a 'responder' in adjunctive analyses, [19,20] where a treatment can be considered to have clinically important effects at the individual level if the proportion of responders is greater in the treatment versus the comparator group. However, this is ineffectual if the clinically relevant response is undefined and clinical trials are not powered to detect this. It is notable that neither Biogen's nor FDA's analyses of the aducanumab trials included response at the individual level.

Despite their limitations, MCIDs are important, and problems arise if we don't use them when considering potential AD treatments. As "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" [21], the MCID is a model that attempts to evaluate whether the efficacy of a therapy reflects clinical effectiveness experienced by clinicians and patients in the real world. To make informed decisions, physicians, patients and caregivers need to understand the benefits any treatment is likely to provide and the period over which the benefits may persist, and to weigh this knowledge against information about potential side effects and other risks. Of course, clinicians will differ in how they make these decisions, but if aducanumab is approved, what could clinicians tell patients and caregivers about what they should expect based on the data from two conflicting trials?

For comparison, donepezil has shown modest benefits over placebo across several trials [22,23]. On average, 10mg per day of donepezil for 24-26 weeks was associated with

improvements of -2.67 points on ADAS-Cog, 1.05 points on MMSE, and -0.53 points on CDR-SB, compared to placebo. Importantly, in terms of safety, donepezil is well-tolerated, whereas high dose aducanumab is associated with a 35% rate of potentially non-trivial brain oedema and 20% rate of brain microhemorrhages (compared to 2% and 7% in placebo, respectively)[15]. This would have increased the risk of unblinding in the high dose aducanumab group, which would have subjected outcomes to reporting bias, particularly with caregiver-informed scales such as the ADCS-ADL-MCI and CDR-SB. As the small effect sizes seen with aducanumab also apply to other amyloid-targeting agents trialed so far, it seems that amyloid reduction alone does not produce clinically meaningful improvements in cognition. Indeed, a recent meta-analysis found that the cognitive effect of reducing amyloid levels by 0.1 standardized uptake value ratio units was an improvement of 0.03 points on the MMSE [24].

While it is important that regulatory requirements encourage drug development and approval, an alternate view would be that regulatory requirements for effectiveness set at a low bar encourage sponsors to substantially increase sample sizes of trials in order to raise their chances of detecting statistical significance for small or inconsequential effects on clinical outcomes. For example, the EMERGE and ENGAGE aducanumab trials were initially powered at 90% (or 10% β error) to detect a CDR-SB difference of -0.5 over 78 weeks with a planned sample of 1350, but this was increased to 1605 (and achieved 1638 and 1647) midway through the trials [15] as it appeared that statistically significant outcomes might not be obtained. Thus, EMERGE achieved a smaller than expected -0.39 difference on the CDR-SB, which was statistically significant when the trial was stopped,

while ENGAGE resulted in a 0.03 point CDR-SB difference slightly favoring placebo that could not have been made statistically significant favoring aducanumab by increasing the sample size. For comparison, the MCID for CDR-SB for MCI and mild AD has been considered to be 0.98 and 1.63 points respectively (Table 1) [16].

Considerations related to the instruments' psychometric properties (i.e. reliability, validity and responsiveness) are relevant when deciding to use MCIDs to judge the clinical meaningfulness of treatments. For example, the MMSE may be prone to unstable inter-rater reliability [25] and the ADAS-Cog has low sensitivity to detect change in MCI/prodromal AD [26,27], where relatively little longitudinal change will occur over the course of a trial as currently conducted. Although composite outcomes, which aggregate cognitive and functional outcomes into a single summary score, are suggested to be more sensitive instruments for these early disease stages [12,14,28], the evidence for their superiority over single test/domain measures is unclear [13]. These psychometric issues emphasise the contribution that MCIDs could offer in distinguishing between clinically meaningful changes and small changes in score due to measurement error. As clinically meaningful change needs to be statistically reliable, methods to assess individual-level reliable change, e.g. using Reliable Change Index (RCI) methods [29,30], may complement group and individual-level MCID approaches. It is also important to account for the effect of baseline disease severity on MCIDs, which will influence trials' statistical power requirements.

It is clear that regulatory approval decisions made primarily on the basis of statistically significant differences in cognitive composite and global outcomes from AD dementia trials are unsatisfactory. The FDA's 'dual' outcome criteria approach (i.e. requiring statistical significance on both 'core' and global or functional measures) to determine the substantial effectiveness of antidementia drugs originated from the first FDA draft guidelines in 1989 [9] to facilitate a pathway for regulatory approval, in response to pharmaceutical industry concerns about a lack of regulatory guidance. The criteria were intended to provide "the lowest standard a sponsor must achieve" to establish effectiveness; but three decades on, the aducanumab data has re-exposed the uncertainty of clinical outcomes and the clinical meaning of statistically significant but small improvements.

We need to strike a better balance between regulators, sponsors, and patients' needs to achieve a common goal. Clinical trials for cognitive impairment should be appropriately powered to reflect clinically meaningful differences in outcomes. Drug development guidance for AD needs to incorporate definitions of clinically meaningful responses for at least the CDR-SB and MMSE, and studies of treatments should determine and report the MCID for other trial outcomes [31] and functional measures such as the ADCS-ADL-MCI. The use of MCIDs would increase the clarity of and confidence in the outcomes of Alzheimer trials, substantially benefiting patients, family, caregivers and healthcare systems.

AUTHORS' CONTRIBUTIONS

R.H. discussed the initial idea with K.L. and L.S. to develop the concept. K.L. wrote the initial draft, under supervision from R.H, which she revised after receiving comments and edits from L.S. and R.H.

CONFLICTS OF INTERESTS

Outside the submitted work, L.S. reports grants and personal fees from Eli Lilly, Merck, and Roche/Genentech; personal fees from Boehringer Ingelheim, Neurim, Ltd, Neuronix, Ltd, Cognition, Eisai, Takeda, vTv, IBC, Abbot, and Samus; and grants from Biogen, Novartis, Biohaven, and Washington Univ/ NIA DIAN-TU.

ACKNOWLEDGMENTS

K.L. is supported by the Medical Research Council (MR/S021418/1). R.H. is supported by the UCLH NIHR Biomedical Research Centre. The authors' funding sources were not involved in the writing of the manuscript or the decision to submit it for publication.

References

1 Howard R, Liu KY. Questions EMERGE as Biogen claims aducanumab turnaround. *Nat Rev Neurol* 2019; published online Nov 29. DOI:10.1038/s41582-019-0295-9.

2 Schneider L. A resurrection of aducanumab for Alzheimer's disease. *Lancet Neurol* 2020; **19**: 111–2.

3 Knopman DS, Jones DT, Greicius MD. Failure to demonstrate efficacy of aducanumab: An analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. *Alzheimers Dement* 2020; published online Nov 1. DOI:10.1002/alz.12213.

4 U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Final Summary Minutes of the Peripheral and Central Nervous System Drugs Advisory Committee Meeting (aducanumab). 2020; published online Nov 6. https://www.fda.gov/media/145690/download (accessed April 19, 2021).

5 U.S. Food and Drug Administration Peripheral and Central Nervous System Drugs Advisory Committee Meeting, July 7, 1989, Rockville, MD. *Department of Health and Human Services, Public Health Service, Food and Drug Administration* 1989.

6 Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. *Am Stat* 2016; **70**: 129–33.

7 Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ* 2012; **4**: 279–82.

8 U.S. Food and Drug Administration. Early Alzheimer's Disease: Developing Drugs for Treatment Guidance for Industry. 2018; published online Feb. https://www.fda.gov/media/110903/download (accessed April 19 2021).

9 Leber P, Millson D, Jolley D, Ward H, Qizilbash N, Schneider LS. What is the Evidence that a Dementia Treatment Works? *In: Qizilbash N SL, Chui H, Tariot P, Brodaty H, Kaye J, Erkinjuntti T, ed Evidence-based Dementia Practice* 2003; **Oxford, UK: Blackwell Science Ltd**: 376–427.

10 Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 1993; **43**: 2412–4.

11 Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer's disease. *N Engl J Med* 2013; **368**: 1169–71.

12 U.S. Food and Drug Administration. Draft Guidance For Industry On Alzheimer's Disease: Developing Drugs For The Treatment Of Early Stage Disease. 2013; published online March 28. https://www.fda.gov/media/85666/download (accessed April 19, 2021).

13 Schneider LS, Goldberg TE. Composite cognitive and functional measures for early stage Alzheimer's disease trials. *Alzheimers Dement* 2020; **12**: e12017.

14 European Medicines Agency. CHMP Guideline on the clinical investigation of medicines for the treatment of Alzheimer's disease. 2018; published online Feb 22. http://www.amge.it/wp/wp-content/uploads/2018/03/Alzheimer1.pdf (accessed April 19, 2021).

15 U.S. Food and Drug Administration. Combined FDA and Biogen Briefing Information for the November 6, 2020 Meeting of the Peripheral and Central Nervous System Drugs Advisory Committee. 2020; published online Nov 6. https://www.fda.gov/media/143502/download (accessed Apr 19, 2021).

16 Andrews JS, Desai U, Kirson NY, Zichlin ML, Ball DE, Matthews BR. Disease severity and minimal clinically important differences in clinical outcome assessments for Alzheimer's disease clinical trials. *Alzheimers Dement* 2019; **5**: 354–63.

17 Schrag A, Schott JM, Alzheimer's Disease Neuroimaging Initiative. What is the clinically relevant change on the ADAS-Cog? *J Neurol Neurosurg Psychiatry* 2012; **83**: 171–3.

18 U.S. Food and Drug Administration. Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products Guidance for Industry (Draft Guidance). 2019; published online Dec. https://www.fda.gov/media/133660/download (accessed April 19, 2021).

19 Coon CD, Cappelleri JC. Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Ther Innov Regul Sci* 2016; **50**: 22–9.

20 U.S. Food and Drug Administration. Patient-Focused Drug Development Public Workshop Guidance 4 Discussion Document. 2019; published online Dec 6. https://www.fda.gov/media/132505/download (accessed April 19, 2021).

21 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; **10**: 407–15.

22 Takeda A, Loveman E, Clegg A, *et al.* A systematic review of the clinical effectiveness of donepezil, rivastigmine and galantamine on cognition, quality of life and adverse events in Alzheimer's disease. *Int J Geriatr Psychiatry* 2006; **21**: 17–28.

23 Birks JS, Harvey RJ. Donepezil for dementia due to Alzheimer's disease. *Cochrane Database Syst Rev* 2018; **6**: CD001190.

24 Ackley SF, Zimmerman SC, Brenowitz WD, *et al.* Effect of reductions in amyloid levels on cognitive change in randomized trials: instrumental variable meta-analysis. *BMJ* 2021; **372**: n156.

25 Nieuwenhuis-Mark RE. The death knoll for the MMSE: has it outlived its purpose? *J Geriatr Psychiatry Neurol* 2010; **23**: 151–7.

26 McDougall F, Edgar C, Mertes M, *et al.* Psychometric Properties of the Clinical Dementia Rating - Sum of Boxes and Other Cognitive and Functional Outcomes in a Prodromal Alzheimer's Disease Population. *J Prev Alzheimers Dis* 2021; **8**: 151–60.

27 Dowling NM, Bolt DM, Deng S. An approach for estimating item sensitivity to within-person change over time: An illustration using the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog). *Psychol Assess* 2016; **28**: 1576–85.

28 Vellas B, Bateman R, Blennow K, *et al.* Endpoints for Pre-Dementia AD Trials: A Report from the EU/US/CTAD Task Force. *J Prev Alzheimers Dis* 2015; **2**: 128–35.

29 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991; **59**: 12–9.

30 Murray AL, Vollmer M, Deary IJ, Muniz-Terrera G, Booth T. Assessing individual-level change in dementia research: a review of methodologies. *Alzheimers Res Ther* 2021; **13**: 26.

31 Howard R, Phillips P, Johnson T, *et al.* Determining the minimum clinically important differences for outcomes in the DOMINO trial. *Int J Geriatr Psychiatry* 2011; **26**: 812–7.

32 Doody RS, Thomas RG, Farlow M, *et al.* Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N Engl J Med* 2014; **370**: 311–21.

33 Honig LS, Vellas B, Woodward M, *et al.* Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. *N Engl J Med* 2018; **378**: 321–30.

34 Mintun MA, Lo AC, Duggan Evans C, *et al.* Donanemab in Early Alzheimer's Disease. *N Engl J Med* 2021; published online March 13. DOI:10.1056/NEJMoa2100708.

Table

Study	AD population	Endpoint (weeks)	CDR-SB	MMSE	ADAS-Cog11	ADAS-Cog13	ADAS-Cog14
MCID[16,17]	**MCI**	-	**-0.98**	**1.26**	-	-	-
	Mild	-	**-1.63**	**2.32**	**-3**	-	-
	Moderate-severe	-	**-2.3**	**3.22**	-	-	-
10mg donepezil[23]	Mild-severe	24-26	-0.53	1.05	-2.67	-	-
High dose aducanumab (EMERGE)[4]	MCI-mild	78	-0.39	0.6	-	-1.4	-
High dose aducanumab (ENGAGE)[4]	MCI-mild	78	0.03	-0.1	-	-0.59	-
Solanezumab (EXPEDITION-1)[32]	Mild-moderate	80	0.1	0.6	-0.8	-	-1.4
Solanezumab (EXPEDITION-2)[32]	Mild	80	-0.3	0.7	-1.5	-	-1.7
Solanezumab (EXPEDITION-3)[33]	Mild	80	-0.34	0.49	-	-	-0.80
Donanemab (TRAILBLAZER-ALZ)[34]	MCI-mild	76	-0.36	0.64	-	-1.86	-

Table 1: Comparison of reported MCIDs and placebo-controlled outcomes for 10mg donepezil, high dose aducanumab, solanezumab and donanemab. All the listed outcomes for ENGAGE and EXPEDITION 1-3 trials, and outcomes except for ADAS-Cog13 for TRAILBLAZER-ALZ, were not statistically significant ($p>0.05$). Abbreviations: MCID = minimal clinically important difference; AD = Alzheimer's disease; MCI = mild cognitive impairment; CDR-SB = Clinical Dementia Rating - Sum

of Boxes; MMSE = Mini-mental State Examination; ADAS-Cog11/13/14 = 11/13/14-item Alzheimer’s Disease-Cognitive Subscale.

Alzheimer's disease clinical trials need to demonstrate minimum clinically important differences

Authors: Kathy Y. Liu[1*], Lon S. Schneider[2], Robert Howard[1]

Affiliations:

[1] Division of Psychiatry, University College London, London, UK

[2] Department of Psychiatry and the Behavioral Sciences, and Department of Neurology, Keck School of Medicine, and Leonard Davis School of Gerontology of the University of Southern California, Los Angeles, CA, USA

*Corresponding author email: kathy.liu@ucl.ac.uk

Summary:

Deciding on the smallest change in an outcome that constitutes a clinically meaningful treatment effect, i.e. the minimum clinically important difference (MCID), is fundamental to interpreting clinical trial outcomes, making clinical decisions, and designing studies with sufficient statistical power to detect any such effect. There is no consensus on MCIDs~~minimum clinically important differences (MCID)~~ for outcomes in Alzheimer's disease (AD) trials, but the US Food and Drug Administration (FDA)~~FDA~~'s consideration of aducanumab ~~study~~ clinical trials data has exposed the uncertainty of the clinical meaning of statistically significant but small improvements. Although~~MCIDs provide a model to evaluate whether the efficacy of a therapy reflects clinical effectiveness experienced by clinicians and patients, and~~ MCIDs for outcomes, ~~(~~including Clinical

Dementia Rating – Sum of Boxes (CDR-SB) and Mini-Mental State Examination (MMSE) in AD have been reported, . The FDA guidelines, drafted in 1989 to facilitate regulatory approval of "substantially effective" antidementia drugs, do not specify quantified minimum differences. While it is important that regulatory requirements encourage drug development and approval, without MCIDs, sponsors are motivated to power trials to detect statistical significance for only small and potentially inconsequential ~~trivial~~ effects on clinical outcomes. MCIDs benefit patients, family, caregivers and healthcare systems and should be incorporated into clinical trials and AD drug development guidance.

Main text:

Clinical trial outcomes for neuropsychiatric conditions, which form the basis for drug marketing decisions, are generally presented as quantitative differences between treatment groups on relevant symptom scales. Deciding on the smallest change in an outcome that constitutes a clinically meaningful treatment effect, i.e. the minimum clinically important difference (MCID), is fundamental to interpreting trial outcomes, making clinical decisions, and designing studies with sufficient statistical power to detect such an effect. This has become particularly important in the interpretation of data from drugs currently under investigation for treatment of dementia. In this paper we ask if it is now time to include agreed MCIDs in the design, analysis and interpretation of Alzheimer's disease (AD) clinical trials?

Since the cholinesterase inhibitors (donepezil, rivastigmine, galantamine) and memantine obtained regulatory approval in the late 1990s and early 2000s, no further approved

treatments for AD~~Alzheimer's disease (AD)~~ have materialized. Recent excitement surrounding ~~the~~ Biogen's Biologics License Application for aducanumab, an anti-amyloid antibody, has been dampened by uncertainty over its effectiveness and controversy over the US Food and Drug Administration's (FDA) and Biogen's interpretation of the clinical trials data [1–3]. The FDA Peripheral and Central Nervous System Drugs Advisory Committee voted nearly unanimously on November 6, 2020 against approval (10 against with 1 abstention) [4]. Nevertheless, the possibility that the FDA might reject this recommendation and approve aducanumab for marketing raises again the important question: how should we objectively define whether a dementia treatment is clinically effective?

The same FDA Advisory Committee addressed this issue in 1989 [5], recommending, in part, that a 3-point difference between drug and placebo groups on the 11-item Alzheimer's Disease-Cognitive Subscale (ADAS-Cog11) represented a clinically meaningful difference. A statistically significant difference on a prespecified neuropsychological outcome alone, however, was considered insufficient to indicate that an intervention makes a clinically meaningful difference as the p-value is the likelihood that such a difference is ~~not~~ attributable to random chance (Wasserstein and Lazar 2016). The p-value itself is not a measure of effect size. Indeed, any effect larger than no (i.e. zero) effect can be demonstrated to be statistically significant with a large enough sample size[6]. Yet, there remains no consensus or agreement on minimum clinically important differences (MCID) for outcomes in AD trials.

The FDA has long considered a treatment for mild to moderate AD dementia (corresponding to stages 4 and 5 in their 2018 draft guidance[7]) to be substantially effective if there is improvement on a 'core' symptom (e.g. a measure of cognition) and a global clinical measure (e.g. a clinician's judgement of change) or a functional measure (e.g. activities of daily living) [8]. For studies including ~~early-stage patients (i.e.,~~ mild cognitive impairment (MCI)~~ and mild dementia~~ patients, or stage~~s~~ 3~~ and 4)~~[7], the FDA requires only statistically significant change on a pre-specified composite measure that includes cognition and daily function combined, as demonstration of substantial effectiveness. Moreover, in 2013 the agency specifically recommended the Clinical Dementia Rating – Sum of Boxes (CDR-SB) [9] as a composite measure that had demonstrated validity and reliability for this purpose [10,11]. No quantified minimum differences were specified, but the rationale is that such a composite measure serves as an indicator of change in both the 'core' or cognitive outcome, i.e., memory, orientation, judgment and problem solving, and in global or daily function, i.e., community affairs, home and hobbies, so that a treatment that delivers statistical significance on the CDR-SB as a pre-specified primary outcome is, by definition, clinically meaningful [8 12]. The European Medicines Agency (EMA) has a similar approach [13]. But, has this bar for effectiveness been set too low, and have the – at best – modest effects of existing treatments served to downgrade our expectations for future treatments?

The current Biologics License Application for aducanumab provides some insight. Two identically designed and simultaneously executed trials were undertaken in early-stage AD (i.e., MCI and mild dementia due to AD) but were stopped early due to perceived

futility before half the participants had an opportunity to complete the trials. On post hoc analysis, the higher dose group of the EMERGE trial was judged to have shown ~~nominally~~ statistically significant reduced decline at the 78-week endpoint, with the following differences between placebo and aducanumab groups favouring aducanumab: -0.39 points on the CDR-SB, 0.6 points on the Mini-Mental State Examination (MMSE), -1.4 points on the ADAS-Cog13, and 1.7 points on the Alzheimer's Disease Cooperative Study - Activities of Daily Living (ADCS-ADL-MCI)[14]. The companion trial, ENGAGE, did not show ~~nominal~~ statistical significance on any of the outcomes: 0.03 on the CDR-SB, -0.1 points on the MMSE, -0.59 points on the ADAS-Cog13 and 0.7 points on the ADCS-ADL-MCI [14]. The negligible effects in ENGAGE clearly indicate no clinically meaningful effect for aducanumab. The very small mean differences favoring aducanumab in EMERGE, however, raise the question of whether these ~~nominally,~~ statistically significant outcomes were clinically meaningful.

A recent study estimated the MCID for clinical outcomes using anchor-based (change in outcome linked to clinical opinion) and distribution-based (MCID calibration based on the variation across participants) approaches, stratified by severity of cognitive impairment [15]. It found that MCIDs increased with disease severity. For MCI and mild AD, differences of 0.98 and 1.63 points for CDR-SB and 1.26 and 2.32 points for MMSE ~~over one year~~ represented clinically meaningful change. Another study, also using anchor-based methodology that linked scores to clinicians' assessment of clinically meaningful change in cognitive, functional and behavioral domains, reported that the MCID for the ADAS-Cog11 in mild AD patients was 3 points[16]. Meeting MCID thresholds, however, are not

requisites for the FDA concluding that a trial shows substantial effectiveness or authorizing marketing approval [17]. Notably, with the aducanumab trials, neither the CDR-SB difference at -0.39 nor the MMSE at 0.6 points reached MCID thresholds (Table 1).

The FDA, however, per its own guidance, can consider a single positive well-controlled trial that is supported by 'confirmatory evidence' to be substantial evidence of effectiveness without considering mean difference or effect size [17]. Applying this interpretation to the aducanumab trials created considerable controversy. This was especially so given that the identical ENGAGE trial failed to show any ~~nominally~~ statistically significant benefits over placebo, and, indeed, mean effects numerically favored placebo on the CDR-SB and MMSE. The FDA's decision on approval is expected by June 7, 2021.

~~Why is the MCID important, and what problems arise if we don't use it when considering potential AD treatments? As "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" [18], the MCID is a model that attempts to evaluate whether the efficacy of a therapy reflects clinical effectiveness experienced by clinicians and patients in the real world.~~ There is no gold-standard method for determining MCIDs, and each approach has limitations. In anchor-based approaches, the external measure of change (or anchor) is usually subjective, and definitions of what constitutes meaningful change may differ between clinicians and patients or caregivers. The MCIDs in Table 1 represent what constitutes a

clinically meaningful decline, as patients' longitudinal change from baseline were anchored to clinicians' assessment of meaningful change (Andrews et al. 2019; Schrag et al. 2012). However, clinical trial outcomes are used to detect a treatment benefit, and the threshold for worsening may not equate to the threshold for improvement (Coon and Cappelleri 2016). Relatively larger changes may be interpreted as clinically important at the individual level, whereas relatively smaller changes may be considered important at the group level, so the application of MCIDs to group means might set the efficacy bar high. The FDA supports anchor-based methods to establish what constitutes meaningful individual-level change, which defines a 'responder' in adjunctive analyses, (Coon and Cappelleri 2016; Center for Drug Evaluation 2019), where a treatment can be considered to have clinically important effects at the individual level if the proportion of responders is greater in the treatment versus the comparator group. However, this is ineffectual if the clinically relevant response is undefined and clinical trials are not powered to detect this. It is notable that neither Biogen's nor the FDA's analyses of the aducanumab trials included response at the individual level.

Despite their limitations, MCIDs are important, and problems arise if we don't use them when considering potential AD treatments. As "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" [18], the MCID is a model that attempts to evaluate whether the efficacy of a therapy reflects clinical effectiveness experienced by clinicians and patients in the real

world. To make informed decisions, physicians, patients and caregivers need to understand the benefits any treatment is likely to provide and the period over which the benefits may persist, and to weigh this knowledge against information about potential side effects and other risks. Of course, clinicians will differ in how they make these decisions, but if aducanumab is approved, what could clinicians tell patients and caregivers about what they should expect based on the data from two conflicting trials?

For comparison, donepezil has shown modest benefits over placebo across several trials [19,20]. On average, 10mg per day of donepezil for 24-26 weeks was associated with improvements of -2.67 points on ADAS-Cog, 1.05 points on MMSE, and -0.53 points on CDR-SB, compared to placebo. Importantly, in terms of safety, donepezil is well-tolerated, whereas high dose aducanumab is associated with a 35% rate of potentially non-trivial brain oedema and 20% rate of brain microhemorrhages (compared to 2% and 7% in placebo, respectively)[14]. This would have increased the risk of unblinding in the high dose aducanumab group, which would have subjected outcomes to reporting bias, particularly with caregiver-informed scales such as the ADCS-ADL-MCI and CDR-SB. As the small effect sizes seen with aducanumab also apply to other amyloid-targeting agents trialed so far, it seems that amyloid reduction alone does not produce clinically meaningful improvements in cognition. Indeed, a recent meta-analysis found that the cognitive effect of reducing amyloid levels by 0.1 standardized uptake value ratio units was an improvement of 0.03 points on the MMSE [21].

While it is important that regulatory requirements encourage drug development and approval, an alternate view would be that regulatory requirements for effectiveness set at a low bar encourage sponsors to substantially increase sample sizes of trials in order to raise their chances of detecting statistical significance for small or inconsequential ~~trivial~~ effects on clinical outcomes. For example, the EMERGE and ENGAGE aducanumab trials were initially powered at 90% (or 10% β error) to detect a CDR-SB difference of -0.5 over 78 weeks with a planned sample of 1350, but this was increased to 1605 (and achieved 1638 and 1647) midway through the trials (FDA 2020) as it appeared that statistically significant outcomes might not be obtained. Thus, EMERGE achieved a smaller than expected -0.39 difference on the CDR-SB, which was ~~nominally~~ statistically significant when the trial was stopped, while ENGAGE resulted in a 0.03 point CDR-SB difference slightly favoring placebo that could not have been made statistically significant favoring aducanumab by increasing the sample size. For comparison, the MCID for CDR-SB for MCI and mild AD has been considered to be 0.98 and 1.63 points respectively ~~over one year~~ (Table 1) [15].

Considerations related to the instruments' psychometric properties (i.e. reliability, validity and responsiveness) are relevant when deciding to use MCIDs to judge the clinical meaningfulness of treatments. For example, the MMSE may be prone to unstable inter-rater reliability (Nieuwenhuis-Mark 2010) and the ADAS-Cog has low sensitivity to detect change in MCI/prodromal AD (McDougall et al. 2021; Dowling et al. 2016), where relatively little longitudinal change will occur over the course of a trial as currently conducted. Although composite outcomes, which aggregate cognitive and functional

outcomes into a single summary score, are suggested to be more sensitive instruments for these early disease stages (Vellas et al. 2015; FDA: Draft Guidance For Industry On A...; EMA 2018), the evidence for their superiority over single test/domain measures is unclear (Schneider and Goldberg 2020). These psychometric issues emphasise the contribution that MCIDs could offer in distinguishing between clinically meaningful changes and small changes in score due to measurement error. As clinically meaningful change needs to be statistically reliable, methods to assess individual-level reliable change, e.g. using Reliable Change Index (RCI) methods (Jacobson and Truax 1991; Murray et al. 2021), may complement group and individual-level MCID approaches. It is also important to account for the effect of baseline disease severity on MCIDs, which will influence trials' statistical power requirements.

It is clear that ~~Should we be satisfied with~~ regulatory approval decisions made primarily on the basis of statistically significant differences in cognitive composite and global outcomes from AD dementia trials are unsatisfactory.~~, even if those effects are not clinically meaningful?~~ The FDA's 'dual' outcome criteria approach (i.e. requiring statistical significance on both 'core' and global or functional measures) to determine the substantial effectiveness of antidementia drugs originated from the first FDA draft guidelines in 1989 [8] to facilitate a pathway for regulatory approval, in response to pharmaceutical industry concerns about a lack of regulatory guidance. The criteria were intended to provide "the lowest standard a sponsor must achieve" to establish effectiveness; but three decades on, the aducanumab data has re-exposed the uncertainty of clinical outcomes and the clinical meaning of ~~nominally~~ statistically significant but small improvements.

We need to strike a better balance between regulators, sponsors, and patients' needs to achieve a common goal. Clinical trials for cognitive impairment should be appropriately powered to reflect clinically meaningful ~~, and not trivial~~ differences in outcomes. Drug development guidance for AD needs to incorporate definitions of clinically meaningful responses for ~~n MCIDs (~~at least ~~for~~ the CDR-SB and MMSE~~)~~, and studies of treatments should determine and report the MCID for other trial outcomes [22] ~~e.g. ADAS-Cog13 and 14~~ and functional measures such as the ADCS-ADL-MCI. The use of MCIDs would increase the clarity of and confidence in the outcomes of Alzheimer trials, substantially benefiting patients, family, caregivers and healthcare systems.

Table 1: Comparison of reported MCIDs and placebo-controlled outcomes for 10mg donepezil, high dose aducanumab, solanezumab and donanemab. All the listed outcomes for ENGAGE and EXPEDITION 1-3 trials, and outcomes except for ADAS-Cog13 for TRAILBLAZER-ALZ, were not statistically significant (p≥~~<~~0.05). Abbreviations: MCID = minimal clinically important difference; AD = Alzheimer's disease; MCI = mild cognitive impairment; CDR-SB = Clinical Dementia Rating - Sum of Boxes; MMSE = Mini-mental State Examination; ADAS-Cog11/13/14 = 11/13/14-item Alzheimer's Disease-Cognitive Subscale.

Study	AD population	Endpoint (weeks)	CDR-SB	MMSE	ADAS-Cog11	ADAS-Cog13	ADAS-Cog14
MCID[15,16]	**MCI**	- ~~5 6~~	**-0.98**	**1.26**	-	-	-
	Mild	- ~~2 6~~	**-1.63**	**2.32**	**-3**	-	-

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

		~~5~~ ~~6~~					
	Moderate-severe	- ~~5~~ ~~6~~	**-2.3**	**3.22**	-	-	-
10mg donepezil[20]	Mild-severe	24-26	-0.53	1.05	-2.67	-	-
High dose aducanumab (EMERGE)[4]	MCI-mild	78	-0.39	0.6	-	-1.4	-
High dose aducanumab (ENGAGE)[4]	MCI-mild	78	0.03	-0.1	-	-0.59	-
Solanezumab (EXPEDITION-1)[23]	Mild-moderate	80	0.1	0.6	-0.8	-	-1.4
Solanezumab (EXPEDITION-2)[23]	Mild	80	-0.3	0.7	-1.5	-	-1.7
Solanezumab (EXPEDITION-3)[24]	Mild	80	-0.34	0.49	-	-	-0.80
Donanemab (TRAILBLAZER-ALZ)[25]	MCI-mild	76	-0.36	0.64	-	-1.86	-

Formatted: Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0.25" + Indent at: 0.5"

AUTHORS' CONTRIBUTIONS

R.H. discussed the initial idea with K.L. and L.S. to develop the concept. K.L. wrote the initial draft, under supervision from R.H, which she revised after receiving comments and edits from L.S. and R.H.

CONFLICTS OF INTERESTS

Outside the submitted work, L.S. reports grants and personal fees from Eli Lilly, Merck, and Roche/Genentech; personal fees from Boehringer Ingelheim, Neurim, Ltd, Neuronix, Ltd, Cognition, Eisai, Takeda, vTv, IBC, Abbot, and Samus; and grants from Biogen, Novartis, Biohaven, and Washington Univ/ NIA DIAN-TU.

ACKNOWLEDGMENTS

K.L. is supported by the Medical Research Council (MR/S021418/1). R.H. is supported by the UCLH NIHR Biomedical Research Centre. The authors' funding sources were not involved in the writing of the manuscript or the decision to submit it for publication.

References

1 Howard R, Liu KY. Questions EMERGE as Biogen claims aducanumab turnaround. *Nat Rev Neurol* 2019; published online Nov 29. DOI:10.1038/s41582-019-0295-9.

2 Schneider L. A resurrection of aducanumab for Alzheimer's disease. *Lancet Neurol* 2020; **19**: 111–2.

3 Knopman DS, Jones DT, Greicius MD. Failure to demonstrate efficacy of aducanumab: An analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. *Alzheimers Dement* 2020; published online Nov 1. DOI:10.1002/alz.12213.

4 FDA: Final Summary Minutes of the Peripheral and Central Nervous System Drugs Advisory Committee Meeting (aducanumab). *Food and Drug Administration Center for Drug Evaluation and Research* 2020; published online Nov 6. DOI:https://www.fda.gov/media/145690/download.

5 FDA: United States Food and Drug Administration Peripheral and Central Nervous System Drugs Advisory Committee Meeting, July 7, 1989, Rockville, MD. *Department of Health and Human Services, Public Health Service, Food and Drug Administration* 1989.

6 Sullivan GM, Feinn R. Using Effect Size-or Why the P Value Is Not Enough. *J Grad Med Educ* 2012; **4**: 279–82.

7 FDA: United States Food and Drug Administration. Early Alzheimer's Disease: Developing Drugs for Treatment Guidance for Industry. *Revision 1 ed: US Food and Drug Administration, Center for Drug Evaluation and Research* 2018. DOI:https://www.fda.gov/media/110903/download.

8 Leber P, Millson D, Jolley D, Ward H, Qizilbash N, Schneider LS. What is the Evidence that a Dementia Treatment Works? *In: Qizilbash N SL, Chui H, Tariot P, Brodaty H, Kaye J, Erkinjuntti T, ed Evidence‐based Dementia Practice* 2003; **Oxford, UK: Blackwell Science Ltd**: 376–427.

9 Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 1993; **43**: 2412–4.

10 Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer's disease. *N Engl J Med* 2013; **368**: 1169–71.

11 FDA: Draft Guidance For Industry On Alzheimer's Disease: Developing Drugs For The Treatment Of Early Stage Disease. 2013. https://www.fda.gov/media/85666/download (accessed Feb 23, 2021).

12 Schneider LS, Goldberg TE. Composite cognitive and functional measures for early stage Alzheimer's disease trials. *Alzheimers Dement* 2020; **12**: e12017.

13 EMA. CHMP Guideline on the clinical investigation of medicines for the treatment of Alzheimer's disease. 2018. http://www.amge.it/wp/wp-content/uploads/2018/03/Alzheimer1.pdf.

14 FDA. Combined FDA and Biogen Briefing Information for the November 6, 2020 Meeting of the Peripheral and Central Nervous System Drugs Advisory Committee. 2020. https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-6-2020-meeting-peripheral-and-central-nervous-system-drugs-advisory-committee-meeting#event-materials.

15 Andrews JS, Desai U, Kirson NY, Zichlin ML, Ball DE, Matthews BR. Disease severity and minimal clinically important differences in clinical outcome assessments for Alzheimer's disease clinical trials. *Alzheimers Dement* 2019; **5**: 354–63.

16 Schrag A, Schott JM, Alzheimer's Disease Neuroimaging Initiative. What is the clinically relevant change on the ADAS-Cog? *J Neurol Neurosurg Psychiatry* 2012; **83**: 171–3.

17 FDA: Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products Guidance for Industry (Draft Guidance). *Center for Biologics Evaluation and Research (CBER)* 2019. DOI:https://www.fda.gov/media/133660/download.

18 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; **10**: 407–15.

19 Takeda A, Loveman E, Clegg A, *et al.* A systematic review of the clinical effectiveness of donepezil, rivastigmine and galantamine on cognition, quality of life and adverse events in Alzheimer's disease. *Int J Geriatr Psychiatry* 2006; **21**: 17–28.

20 Birks JS, Harvey RJ. Donepezil for dementia due to Alzheimer's disease. *Cochrane Database Syst Rev* 2018; **6**: CD001190.

21 Ackley SF, Zimmerman SC, Brenowitz WD, *et al.* Effect of reductions in amyloid levels on cognitive change in randomized trials: instrumental variable meta-analysis. *BMJ* 2021; **372**: n156.

22 Howard R, Phillips P, Johnson T, *et al.* Determining the minimum clinically important differences for outcomes in the DOMINO trial. *Int J Geriatr Psychiatry* 2011; **26**: 812–7.

23 Doody RS, Thomas RG, Farlow M, *et al.* Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *N Engl J Med* 2014; **370**: 311–21.

24 Honig LS, Vellas B, Woodward M, *et al.* Trial of Solanezumab for Mild Dementia Due to Alzheimer's Disease. *N Engl J Med* 2018; **378**: 321–30.

25 Mintun MA, Lo AC, Duggan Evans C, *et al.* Donanemab in Early Alzheimer's Disease. *N Engl J Med* 2021; published online March 13. DOI:10.1056/NEJMoa2100708.

Reviewer comments	Author response and changes made	Page number in revised paper
Editors specific Comments: Although the piece is focused on Alzheimer's disease clinical trials, would it be possible to start the piece off with a paragraph that puts the concept of minimum clinically important differences into a broader perspective of mental health disorders more generally?	We have now included the following paragraph at the start of the piece: "**Clinical trial outcomes for neuropsychiatric conditions, which form the basis for drug marketing decisions, are generally presented as quantitative differences between treatment groups on relevant symptom scales. Deciding on the smallest change in an outcome that constitutes a clinically meaningful treatment effect, i.e. the minimum clinically important difference (MCID), is fundamental to interpreting trial outcomes, making clinical decisions, and designing studies with sufficient statistical power to detect any such effect. This has become particularly important in the interpretation of data from drugs currently under investigation for treatment of dementia. In this paper we ask if it is now time to include agreed MCIDs in the design, analysis and interpretation of Alzheimer's disease (AD) clinical trials?"**	2
Reviewers' Comments: **Reviewer #1:** This manuscript is a personal view paper on a highly relevant and topical issue, i.e. the minimum clinically important difference for AD trials. It adds to the scientific and clinical discussion on the recent aducanumab trial and the FDA's recommendation. The writing is clear, and the call for establishing valid MCIDs that are informative in	We thank the Reviewer for their helpful comments. We agree that there are instrument-related issues related to MCIDs, which would be important to mention in the article. We have now discussed these issues, along with the potential benefit of reliable change indices, in more detail in a new paragraph on Page 9 (below).	6, 9, 10

addition to statistically significant changes is very urgent, in RCTs in general but especially for the field of dementia trials. I have some general comments that the authors may want to address in a revised version of their statement: The question how we define whether a dementia treatment is clinically effective is highly important, and I fully support the authors' statement that only a statistically significant improvement (or difference with control arm) is insufficient or a treatment to be considered effective. However, the notion of a 'clinically meaningful difference' based on just a specific instrument or rating scale (eg a 3-point difference on the ADAS-Cog) is also arbritary in nature, as effects may not be linear (as stated in the manuscript), that is, potential benefits may differ for patients in the early stages of AD (incl MCI) compared to later stages of AD. Also, the MCID should have good criterion validity, not just face validity (ie expert opinion). That is how large does the difference on (say) the ADAS-Cog have to be, in order to reflect a *meaningful* change in everyday memory performance (eg not forgetting appointments, or remembering to check one's agenda)? What also should be taken into account are the psychometric properties of the outcome scale, which is often ignored in the RCT field. A lot of effort is often put into the trial analysis part and the trial's statistical power, a field that is dominated by biostatisticians.	We also thank the Reviewer for drawing our attention to the relevant references, which we now cite in the article. We note that the McDougall paper on the psychometric limitations of the ADAS-Cog (and MMSE) was specific to prodromal AD, and although Dowling et al. found that the ADAS-Cog had limited sensitivity to detect change in the study group overall, this was driven by the normal control and MCI subgroups and not the AD subgroup. Nieuwenhuis-Mark reported good psychometric properties for the MMSE but had concerns about unstable inter-rater reliability, so we have focused on this specific property in our discussion of the MMSE. We consider that demonstration of reliable change is necessary but not sufficient for clinically meaningful change, so we have described individual-level reliable change indices as complementary to MCID approaches. **“Considerations related to the instruments’ psychometric properties (i.e. reliability, validity and responsiveness) are relevant when deciding to use MCIDs to judge the clinical meaningfulness of treatments. For example, the MMSE may be prone to unstable inter-rater reliability (Nieuwenhuis-Mark 2010) and the ADAS-Cog has low sensitivity to detect change in MCI/prodromal AD (McDougall et al. 2021; Dowling et al. 2016), where relatively little longitudinal change will occur over the course of a trial as currently conducted. Although composite outcomes, which aggregate cognitive and functional outcomes into a single summary**	

Psychometricians, however, who are experts on the construction and validity of rating scales, are often not involved. The outcome instrument's validity is consequently often taken for granted (because a measure is widely used, researchers have published with it before, or they are even part of consensus guidelines in a specific field), which is problematic. For instance, cognitive instruments such as the MMSE or ADAS-Cog have been repeatedly criticized with respect to their psychometric properties (eg. McDougall et al. https://doi.org/10.14283/jpad.2020.73; Dowling et al. https://dx.doi.org/10.1037/pas0000285; Nieuwenhuis-Mark https://doi.org/10.1177/0891988710363714) but are continued to be used (or even recommended) as outcome measures for dementia trials. Also, their use as an outcome measure is not always in line with the intruments construct validity. I.e. the MMSE was developed as a diagnostic instrument, ie a cognitive screen for detecting Alzheimer's dementia, not as a monitoring tool for cognitive decline over time or disease/symptom progression. This is too often ignored in RCT outcome discussions, although an excellent construct validity, test-retest and interrater reliability are essential for drawing valid conclusions about statistically reliable change and clinically meaningful improvement. While I fully agree that a clinician's or other informant's rating of change in a patient is highly informative (for disorders, such as dementia, in which the use of PROMs is complicated), these validity and reliability challenges are not trivial. Furthermore, current recommendations still emphasize	**score, are suggested to be more sensitive instruments for these early disease stages (Vellas et al. 2015; FDA: Draft Guidance For Industry On A...; EMA 2018), the evidence for their superiority over single test/domain measures is unclear (Schneider and Goldberg 2020). These psychometric issues emphasise the contribution that MCIDs could offer in distinguishing between clinically meaningful changes and small changes in score due to measurement error. As clinically meaningful change needs to be statistically reliable, methods to assess individual-level reliable change, e.g. using Reliable Change Index (RCI) methods (Jacobson and Truax 1991; Murray et al. 2021), may complement group and individual-level MCID approaches. It is also important to account for the effect of baseline disease severity on MCIDs, which will influence trials' statistical power requirements."** We accept that the ADAS-Cog is unlikely to fully align with a person's functional status in AD. As a specific instrument for measuring cognition, we have described in the article how the ADAS-Cog fulfils a requirement set by regulators, such as the FDA, to provide a measure of change in a 'core' AD symptom, i.e. cognition, alongside change in at least another measure of function or a global clinical measure (or both in a composite measure). It is therefore equally valid to ask how large the difference on a functional measure (such as ADCS-ADL-MCI) needs to be to reflect a meaningful clinical change in function. We have focused on MCID data for ADAS-Cog, MMSE	

the use of group statistics (ie, a statistically significant difference between two arms and an overall effect size for such an analysis). However, measures of individual change are far more relevant as an outcome measure, e.g. the use of reliable change index (RCI) analyses, which also take psychometric limitations of a given outcome scale into account. I would argue that the use of group statistics in RCTs alone (even including effect sizes expressing the overall magnitude that may even reflect an MCID) is insufficient, but that analyses of individual change are essential too (in the end, a treatment should benefit individual patients as well), but --as far as I know-- not part of routine RCTs or evaluation agencies. In my opinion, MCID should always incorporate some type of responder analysis.	and CDR-SB because values have been published, but MCIDs for functional outcomes, as raised by the Reviewer, are also important to interpret clinical trial outcomes. We now include this recommendation in the last paragraph on Page 10: "...studies of treatments should determine and report the MCID for other outcomes [22] **and functional measures such as the ADCS-ADL-MCI**." We have also expanded on the limitations of MCIDs and distinguish more clearly between group and individual level MCIDs on Page 6.	
Reviewer #2: This manuscript addresses an important issue that is particularly pertinent in the context of recent results in AD clinical trials. Approaches to establishing the clinical meaningfulness of a treatment effect in AD, and on the CDR-SB in particular, are not well established and deserve further attention. While the authors propose applying MCIDs to the delta between mean changes in treatment groups, some of the methods described (e.g. the anchor based methods) are more commonly used for individual level change. As the authors state, the FDA does not supply MCIDs in their AD guidance but the FDA's draft Patient Focused Drug Development guidances/workshop documents express a	We thank the Reviewer for their helpful comments and discussion on this issue. We have now distinguished more clearly between individual-level and group-level MCID, and refer to the FDA PFDD guidance on Page 6. We also describe some limitations of the MCID, including that applying them to group means that this may represent the setting of a high bar for efficacy. **"Relatively larger changes may be interpreted as clinically important at the individual level, whereas relatively smaller changes may be considered important at the group level, so the application of MCIDs to group means might set the efficacy bar**	6

clear preference for application of meaningful change thresholds at the individual level in order to compare the proportion of patients with a meaningful improvement/deterioration. Such supportive analyses are viewed as useful in contextualising the clinical meaningfulness of a continuous primary endpoint. This approach is consistent with previous presentations/guidances for example the PRO guidance (2009), the principles of which, the FDA have stated, apply to all types of COAs. The application of MCIDs to between group deltas typically sets a very high bar. Based on the mean decline in the ENGAGE and EMERGE pbo groups, and the stated .98 MCID on the CDR-SB, there would need to be a 63% and 56% reduction in decline in Aducanumab arms to deem it clinically meaningful.This seems to be overwhelming efficacy rather than a threshold for clinical meaningfulness. Based on the above, I do not agree with the authors' proposed approach to using MCIDs, but do agree that something more than statistical significance is needed.	**high. The FDA supports anchor-based methods to establish what constitutes meaningful individual-level change, which defines a 'responder' in adjunctive analyses, (Coon and Cappelleri 2016; Center for Drug Evaluation 2019), where a treatment can be considered to have clinically important effects at the individual level if the proportion of responders is greater in the treatment versus the comparator group. However, this is ineffectual if the clinically relevant response is undefined and clinical trials are not powered to detect this. It is notable, that neither Biogen's nor the FDA's analyses of the aducanumab trials included response at the individual level."**	
Specific suggestions: • Add detail on the various ways MCIDs can be calculated and used. Coon and Cappelleri (2016, Therapeutic Innovation & Reg Science, 50, 1,22-29) provide a useful summary.	We are grateful for the Reviewer's specific suggestions. In the last paragraph on Page 5, we referred to "anchor-based (change in outcome linked to clinical opinion) and distribution-based (MCID calibration based on the variation across participants) approaches".	6

	We have now cited the Coon and Cappelleri paper and have included a statement that there is no “gold standard” method for determining the MCID, and have included a consideration of limitations related to this. Whilst the Coon and Capelleri article provides a useful summary and proposes use of regression techniques to ascertain the clinically important thresholds at a group and individual level, we are also aware of a conflicting view that a regression analysis may not be appropriate to link corresponding points on different, but correlated, instruments, due to a conceptual mismatch (Leucht et al., 2005). **“There is no gold-standard method for determining MCIDs, and each approach has limitations. In anchor-based approaches, the external measure of change (or anchor) is usually subjective, and definitions of what constitutes meaningful change may differ between clinicians and patients or caregivers. The MCIDs in Table 1 represent what constitutes a clinically meaningful decline, as patients' longitudinal change from baseline were anchored to clinicians’ assessment of meaningful change (Andrews et al. 2019; Schrag et al. 2012). However, clinical trial outcomes are used to detect a treatment benefit, and the threshold for worsening may not equate to the threshold for improvement (Coon and Cappelleri 2016).”**	
• With reference to the comment that FDA are supportive of the CDR-SB as a primary endpoint - I	In the last paragraph on Page 4, we stated that “For studies including mild cognitive impairment (MCI) patients, or stage 3[7], the FDA requires only statistically	9

think it is worth noting that their 2018 guidance encourages the development of novel endpoints for stage 3, suggesting that they are not entirely satisfied with the options available.	significant change on a pre-specified composite measure that includes cognition and daily function combined, as demonstration of substantial effectiveness." The FDA 2018 guidance "encourages the development of novel approaches to the integrated evaluation of subtle functional deficits arising from early cognitive impairment". We have now added a sentence referring to the composite outcomes on Page 9: **"Although composite outcomes, which aggregate cognitive and functional outcomes into a single summary score, are proposed to be more sensitive instruments for these early disease stages (Vellas et al. 2015; FDA: Draft Guidance For Industry On A...; EMA 2018), the evidence for their superiority over single test/domain measures is unclear (Schneider and Goldberg 2020)."**	
• I think this manuscript is incomplete without a discussion of using anchor based methods to determine thresholds for individual change given that FDA promote their use elsewhere. I would also encourage the authors to consider discussing other approaches to determining clinical meaningfulness, e.g. would it be sufficient to show that a statistically significant CDR-SB translates to a large proportion of patients retaining ADLs on a secondary outcome?	The discussion on anchor-based methods for individual level change with reference to the FDA guidance is addressed in a previous comment above. The question of whether a statistically significant but sub-MCID change in CDR-SB, combined with a large and significant change on a secondary and functional outcome, would represent clinical meaningfulness is relevant. We consider that MCIDs for individual outcomes are needed to make clear conclusions about the treatment effect in trials. For example, for aducanumab, the higher rate of adverse outcomes would be associated with a higher risk of unblinding. As the ADL scale is rated by or in consultation with a	8

	caregiver, it is possible that this scale is more susceptible to the effects of unblinding compared to a more objective cognitive test completed with the patient alone. We have now added on Page 8, **"This would have increased the risk of unblinding in the aducanumab group, which may have subjected outcomes to reporting bias, particularly with caregiver-informed scales such as the ADCS-ADL-MCI and CDR-SB."**	
• The authors make a number of statements regarding the size of the Aducanumab delta either directly or indirectly, e.g.: ○ "The very small mean differences favoring aducanumab in EMERGE,however, raise the question of whether these nominally, statistically significant outcomes were clinically meaningful." ○ "Clinical trials for cognitive impairment should be appropriately powered to reflect clinically meaningful, and not trivial differences in outcomes" While I think it absolutely appropriate to question the meaningfulness of these treatment effects, I would stop short of labelling them as nominal or trivial. If the 22% reduction is believed, it would take patients in the Aducanumb group approx 23 months to reach the mean decline that was reached by the pbo group at 18 months. This 5	We have now removed the words "nominal" and "trivial" from these and similar sentences.	Throughout

month delay may or may not be considered of consequence by the community.		
Reviewer #3: The article by Liu et. al. argued the importance of demonstrating minimum clinically importance difference (MCID) in the Alzheimer's disease (AD) clinical trials. This is a well thought-out article - the controversy surrounding Biogen's anti-amyloid agent Aducanumab for the treatment of AD has been widely discussed recently. The authors addressed the issue from clinicians' and patients' perspectives and pointed out why the demonstration of MCID is essential in patient care and decision making. This is a welcoming view as the authors raised an important but often neglected issue: a statistical significance does not necessarily imply a clinical significance. Indeed, a p-value of <0.05 alone is not enough to judge the effectiveness of an intervention vs. its comparator, and so the effect size should also be considered. Some minor points: 1. The authors compared the data from ENGAGE and EMERGE trials, which were based on 78 weeks of follow-up time to the MCIDs estimated from the Andrews' study, which was derived over a 1-year study period. Please comment on if it is necessary to calibrate the MCIDs values from 1-year to 78-week study period for the comparison.	We thank the Reviewer for their helpful comments. The Andrews et al. and Schrag et al. studies obtained MCID estimates by linking change in outcomes from baseline (i.e. decline in cognition and function) to clinicians' assessment of clinically meaningful change (Yes/No). So the MCID should not have to be calibrated, as what constitutes a clinically important change can be applied to outcomes over any duration. Inclusion of the time to endpoint column in Table 1 was intended to compare the study durations, especially for anti-amyloid compounds versus donepezil, however we can see how including the observation period for the MCID values at the top of the table was confusing. We have removed these data as they are not applicable and are not study endpoints. We have also removed the term "over one year" in reference to the MCIDs from Andrews et al. on pages 5 and 9.	Table 1; Pages 5, 9
2. The authors may consider citing a statement issued by The American Statistical Association in 2016 regarding	We are grateful for this reference, which we have now cited on Page 3.	3

the use of p-values (The American statistician, 2016-04-02, Vol.70 (2), p.129-133).		
3. Please remove 'not' from the statement ".. as the p-value is the likelihood that such a difference is not attributable to random chance." (Page 2 last line). A small p-value suggests a low probability that the observed difference between groups is due to chance.	Thank you, we have now done this.	3
4. Please provide a reference for the statement that the sample size of the EMERGE and ENGAGE studies was increased from the planned N=1350 to an N=1650 (page 7).	This has been added on Page 8.	8.
5. Table 1: The time period listed for MCID among mild AD population (i.e. 26-56 weeks) was confusing. Does it mean that all the MCIDs were estimated based on 26-56 weeks of follow up? Or, does it mean some were from 26 weeks and some were from 56 weeks of study? Please clarify.	Please see response to Comment 1 above. We have now removed these data for the MCID values because they were confusing and were not relevant to the Column ‘Study Endpoint’.	Table 1
6. Table 1 legend: Should the 'p<0.05' be 'p>0.05' for not statistically significant?	We have now changed this to p>0.05 as suggested.	Table 1

Editors' General Comments: (not all will apply; please check carefully and respond as appropriate): **All completed in the clean version.**

- Your revised paper should have fewer than 4500 words (not including references) and a maximum of 100 references. The abstract should be unstructured and less than 250 words long. **Yes**
- References
  References should be in Vancouver style. For references with six authors or fewer, all authors should be listed. For those with seven or more authors, only the first three authors and 'et al' should be listed. Please ensure that reference numbering throughout the manuscript is not inserted with electronic referencing software, such as Endnote. All web references should have the date they were last accessed. With your revised submission please enclose copies of any papers cited as being 'in-press', along with a copy of the acceptance letter from the journal. **Yes**
- Figures **NA**
  Please supply editable files (eg, EPS files) for all figures such as diagrams and graphs. If figures are composed of photographs or other images, high-resolution files (300 dpi or greater) should be provided. If your figures are annotated, please supply two copies of each of these figures as separate jpgs (one copy annotated and one non-annotated). Our in-house illustrators will annotate according to journal style using the annotated figures as a guide. For multi-part figures, please supply the individual parts as well as a combined version to be used as a guide for our illustrators to recreate the files. Generally, figure legends should be a maximum of 30 words long. For artwork guidelines please see https://www.thelancet.com/for-authors/forms?section=artwork
  Images that have been published previously should be accompanied by a statement indicating permission to reproduce the image. If required, further assistance can be obtained from the editorial team. If you have used previously published images, you must obtain permission from the copyright holder of the paper, which might be the authors or the publisher. If all the figures are your own and have not been published before, then this requirement does not apply.
- Tables should be supplied in a separate Word file (not Excel or fdf/pdf). Each row of data should be in a separate line. Please ensure that rows and columns are not tabbed; data should be entered in cell form. **Yes**
- Approval of the revised manuscript
  As corresponding author, please confirm with your submission that all authors have seen and approved the revised version. **Yes**
- Author contribution statement
  All authors are required to provide a signed author contribution statement form, available to download here (completed forms should be uploaded with your revised manuscript).
  At the end of the manuscript include a "Contributors" section stating the contribution each author has made to the paper. These statements should match exactly those given on the Authors' signed forms. Authors should be referred to by their initials in this section. **Yes**
  Please note, if an author is added to or deleted from the author list, we need a signed statement from the other authors to agree to the amendment (including the author who was removed).
- Conflict-of-interest statements for individual authors
  All authors should complete and return an ICMJE conflict of interest form, available to download here.

A conflict of interest exists if authors or their institutions have financial or personal relationships with other people or organisations that could inappropriately influence (bias) their actions. Financial relationships are easily identifiable, but conflicts can also occur because of personal relationships, academic competition, or intellectual passion. A conflict can be actual or potential, and full disclosure to The Editor is the safest course. Failure to disclose conflicts might lead to publication of an Erratum or even to retraction.
All submissions to *The Lancet Psychiatry* must include disclosure of all relationships that could be viewed as presenting a potential conflict of interest (see Lancet 2001; 358: 854-56 and Lancet 2003; 361: 8-9). The Editor may use such information as a basis for editorial decisions, and will publish such disclosures if they are believed to be important to readers in judging the manuscript.
At the end of the text, under a subheading "Declaration of interest", all authors must disclose any financial and personal relationships with other people or organisations that could inappropriately influence (bias) their work. Examples of financial conflicts include employment, consultancies, stock ownership, honoraria, paid expert testimony, patents or patent applications, and travel grants, all within 3 years of beginning the work submitted. If there are no conflicts of interest, authors should state that none exist ("We declare no competing interests."). Authors should be referred to by their initials in this section. The statement should match the information on the supplied ICMJE forms. **Yes**
For Comments, Personal Views, and Reviews, *The Lancet Psychiatry* will not usually publish if an author, within the past 3 years, and with a relevant company or competitor, has any stocks or shares, equity, a contract of employment, or a named position on a company board; or has been asked by any organisation other than *The Lancet Psychiatry*, be named on, or to submit the paper (see Lancet 2004; 363: 2-3).

- Author affiliations and qualifications
  Please format the author affiliation list to *Lancet* style. Please list authors by full first name and last name. Then for affiliations, please list each institution (department, institute, city, state, country) with the author initial and full last name, followed by one degree, in brackets following the author institution. Please indicate any full professors. Please provide full institutional correspondence address for corresponding author. **Yes**
- Acknowledgments
  We need written confirmation, including a signature, from everyone who is mentioned in the Acknowledgments section to confirm that they are happy to be quoted in your paper. The following format can be used:
  "I permit <corresponding author> et al to list my name in the acknowledgments section of their manuscript and I have seen a copy of the paper <full article title>" **NA**
- Role of medical writer or editor **NA**
  Was a medical writer or editor involved in the creation of your manuscript? If yes, we need a signed statement from the corresponding author to include the name and information on funding of this person. This information should be added to the Acknowledgment section. In addition, you will need to send us a signed statement from this person declaring that he or she has given you permission to name him or her in the Acknowledgment section.
- Personal communication **NA**
  For each personal communication, please provide a letter confirming that the person agrees to their name being used. Signed paper

originals are required. The following format can be used:
"I permit <corresponding author> et al to cite a personal communication from me in their manuscript <full article title>"

- Appendix **NA**
  It is not *Lancet* policy to edit or style supplementary material for the web; however, this material will be hosted on our website as a pdf conversion of the author-supplied Word file. Please style your supplementary material as per the guidelines below. Please note that we will be unable to correct any errors in the online appendix following publication; as such, please check carefully when submitting.
  Please supply the appendix as a single Word file, with numbered pages. Please do not include a cover page with details of the paper, as we add a cover page to the appendix file before publication with this information.
  Required formatting of data in the appendix:
  - SI units are required, if appropriate
  - Numbers in text and tables should always be provided if % is shown.
  - Means should be accompanied by SDs, and medians by interquartile range.
  - Exact p values should be provided, unless p<0·0001
  - For drug names, recommended international nomenclature (rINN) is required
  - References should be in Vancouver style (eg, Smith A, Jones, B, Clements S. Clinical transplantation of tissue-engineered airway. Lancet 2008; 372: 1201-09. Hourigan P. Ankle injuries. In: Sports medicine. Chan D, ed. London: Elsevier, 2008: 230-47.) They should be numbered in order of mention in appendix and numbered separately from references in the full paper
  - For figures in the appendix, all images must have a minimum resolution of 300 dpi at a width of 107 mm