WHITE PAPER AUGUST 2026

Meaningful to Whom?

How minimal clinically important differences are misused in clinical trials — and what defensible thresholds look like

IN THIS PAPER

NEKODA HEALTH PTY LTD · AUSTRALIA

NEKODAHEALTH.COM.AU

ABSTRACT

Every clinical outcome assessment (COA) endpoint eventually collides with the same question: how much change matters? The industry's standard answer — the minimal clinically important difference (MCID) — is routinely transplanted between populations, instruments, administration modes and decades as though it were a physical constant. It is not. MCIDs are estimates, produced by methods of highly variable credibility, anchored to specific populations and contexts, and frequently reported so poorly that their quality cannot even be evaluated. This paper reviews the peer-reviewed evidence of the past decade on how meaningful-change thresholds are derived, how credible the published estimates actually are, why group-level MCIDs are misapplied to within-patient responder analyses, and why a threshold derived on a paper instrument in one population does not automatically govern a digital instrument in another. It closes with a practical framework for selecting and defending thresholds. The audience is clinical trial and pharmaceutical professionals who use MCIDs without being COA specialists.

01 The number everyone borrows

MCIDs are among the most-quoted and least-examined numbers in clinical research. They appear in sample-size justifications, responder definitions, payer dossiers and journal discussion sections, usually as a bare figure with a citation — three points on this scale, five on that one. The figure is treated as a property of the instrument. In reality it is a property of an estimation exercise: a particular method, applied to a particular population, at a particular time, against a particular anchor. Change any of those and the estimate can change with it.

The consequences of borrowing carelessly run in both directions. A threshold set too low lets measurement noise masquerade as benefit; a threshold set too high condemns genuinely useful treatments. Either way, the error is invisible in the trial report, because the threshold arrived with the authority of a citation.

02 How thresholds are made — and how credible they are

Methodologists distinguish two families of approach. Anchor-based methods tie score change to an external judgement of importance — typically the patient's own global rating of change — and are the only family that connects a threshold to meaning as patients experience it. Distribution-based methods derive thresholds from statistical properties such as the standard error of measurement; they describe what an instrument can reliably detect, not what patients consider important, and are best treated as lower-bound checks on score precision within a triangulation exercise rather than as meaningfulness estimates in their own right[1].

Even within the preferred anchor-based family, credibility varies enormously. A formal credibility instrument, published in The BMJ, sets out five core criteria: the anchor must be patient-rated, interpretable and relevant to patients; the estimate precise; the anchor–instrument correlation satisfactory; and the anchor threshold genuinely reflect a small-but-important difference[2]. When this lens was applied at scale, the results were sobering: a systematic survey of 585 studies reporting 5,324 MID estimates for 526 instruments found serious reporting deficiencies throughout, including failure to report the anchor–instrument correlation in 66% of estimates and inadequate information to judge the precision of the estimate or the anchor threshold in a further tranche[3]. The methodological community has since had to build a workaround — a construct-proximity item allowing credibility assessment when the correlation is unreported — an accommodation that exists only because so much of the MCID literature cannot meet the original standard[4].

The practical upshot for trial teams: the citation attached to a borrowed MCID frequently resolves to an estimate whose credibility cannot be established from the publication. The number is not wrong, necessarily. It is unverifiable — which, for a figure that decides whether a treatment worked, amounts to much the same thing.

03 Group thresholds, individual claims

A second, subtler misuse concerns what kind of question the threshold answers. The MCID tradition largely concerns between-group differences — how big a mean difference between arms is worth caring about. Responder analyses ask a different question: how much change within an individual patient constitutes benefit. Regulatory science has moved decisively toward this within-patient framing, and the methodological literature is explicit that responder definitions require their own estimation logic — patient-rated anchors, appropriate study designs, multiple anchors and analytic methods, and

triangulation across them — rather than the recycling of group-level MCIDs[1]. Applying a group-difference threshold to classify individual responders conflates two statistically and conceptually distinct quantities, and it does so at the exact analytic step regulators scrutinise most closely.

Within-patient interpretation also collides with instrument noise. Registry analyses of the Mini-Mental State Examination in dementia populations formalised this as a distinction between the minimal clinically important difference and a 'real-world reassessment threshold' — the smallest change exceeding expected measurement variability plus expected decline. Anchor-based MCIDs in those cohorts clustered around 0.7 to 3.8 points depending on diagnosis, while the reassessment threshold was 5 to 7 points: on this instrument, a change patients would consider meaningful is smaller than a change the instrument can reliably distinguish from noise over a year[5]. A responder definition that ignores that gap classifies measurement error as response.

04 Thresholds do not travel

The deepest misuse is transplantation. Three findings from the past decade explain why a threshold minted in one context does not automatically govern another. First, thresholds are population- and context-dependent by construction: the same MMSE analysis found anchor-based thresholds differing several-fold between diagnoses and by baseline severity[5]. Second, instruments claiming to measure the same construct differ substantially in content — seven common depression scales share a mean content overlap of only 0.36[6], and 450 recent depression trials used 388 different outcome measures[7] — so a threshold is welded to its instrument, not to the disease. Third, administration mode is part of the estimation context. The good-practice framework for electronic migration requires evidence that a change in data-collection mode has not affected the instrument's measurement properties[8]; a paper-era threshold applied to a migrated electronic instrument silently assumes exactly the equivalence that framework exists to verify. In most well-executed migrations that assumption will hold — but it is an assumption, and for a decision-making threshold it should be an audited one.

The interpretive stakes are visible in the evidence-synthesis literature. A Cochrane review of omega-3 fatty acids for depression translated its pooled effect — roughly 2.5 points on the 17-item Hamilton scale — against a minimal clinically important change of 3.0 points, concluding the effect was unlikely to be clinically meaningful despite statistical significance[9]. The entire clinical conclusion of a major review pivoted on a threshold of half a scale point. Numbers doing that much work deserve provenance checks.

05 A defensible-threshold framework

The evidence supports a five-step discipline for any programme that will live or die by a meaningful-change claim. First, audit the provenance of every candidate threshold against formal credibility criteria — patient-rated anchor, relevance, precision, correlation, appropriate anchor threshold — rather than accepting the citation at face value[2,3,4]. Second, match the threshold's estimation context to the trial's context: population, disease stage, instrument version and administration mode all need to correspond, or the mismatch needs to be justified[5,8]. Third, distinguish between-group interpretation from within-patient responder definitions, and derive the latter with patient-rated anchors and triangulation across multiple methods rather than by recycling the former[1]. Fourth, quantify the instrument's noise floor in the trial population and confirm the proposed threshold exceeds it — a responder definition below test–retest variability is an error-classification machine[5]. Fifth, where no credible, context-matched estimate exists, generate one: embedding anchor questions in phase 2 costs little and converts the phase 3 interpretation from borrowed authority to owned evidence[1].

For sponsors, the reframe is simple. An MCID is not a fact about a disease; it is a piece of evidence about a measurement, with all the quality gradations evidence carries. It should be selected, appraised and defended the way any other pivotal evidence is.

06 Conclusion

Meaningful-change thresholds sit at the point where measurement becomes decision: they convert score movements into claims about patients' lives, approvals and reimbursement. The past decade of methodological work has delivered both the tools to do this well — credibility instruments, responder-definition methodology, noise-floor quantification — and the uncomfortable finding that most published estimates cannot demonstrate they meet the standard[2,3]. The industry's habit of transplanting thresholds across populations, instruments, modes and decades persists because it is convenient and because its failures are silent. The fix is not exotic. It is to ask of every threshold the question its name already implies: meaningful to whom, on what instrument, measured how, and shown by what evidence?

NOTE ON LITERATURE SOURCING

Literature for this white paper was identified through structured searches of PubMed (National Library of Medicine), restricted to peer-reviewed publications from 2016 onwards. Digital object identifiers (DOIs) are provided for all cited works.

References

01 Coon CD, Cook KF. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Quality of Life Research 2018;27(1):33–40. https://doi.org/10.1007/s11136-017-1616-3

02 Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020;369:m1714. https://doi.org/10.1136/bmj.m1714

03 Carrasco-Labra A, Devji T, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Serious reporting deficiencies exist in minimal important difference studies: current state and suggestions for improvement. Journal of Clinical Epidemiology 2022;150:25–32. https://doi.org/10.1016/j.jclinepi.2022.06.010

04 Wang Y, Devji T, Carrasco-Labra A, Qasim A, Hao Q, Kum E, et al. An extension minimal important difference credibility item addressing construct proximity is a reliable alternative to the correlation item. Journal of Clinical Epidemiology 2023;157:46–52. https://doi.org/10.1016/j.jclinepi.2023.03.001

05 Abzhandadze T, Hoang MT, Bao X, Åkerman M, Norgren J, García Pascual B, et al. Thresholds for meaningful change in Mini-Mental State Examination scores in rare dementias. Alzheimer's Research & Therapy 2026;18(1). https://doi.org/10.1186/s13195-026-02136-y

06 Fried EI. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. Journal of Affective Disorders 2016;208:191–197. https://doi.org/10.1016/j.jad.2016.10.019

07 Veal C, Tomlinson A, Cipriani A, Bulteau S, Henry C, Müh C, et al. Heterogeneity of outcome measures in depression trials and the relevance of the content of outcome measures to patients: a systematic review. The Lancet Psychiatry 2024;11(4):285–294. https://doi.org/10.1016/S2215-0366(23)00438-8

08 O'Donohoe P, Reasner DS, Kovacs SM, Byrom B, Eremenco S, Barsdorf AI, et al. Updated recommendations on evidence needed to support measurement comparability among modes of data collection for patient-reported outcome measures: a good practices report of an ISPOR task force. Value in Health 2023;26(5):623–633. https://doi.org/10.1016/j.jval.2023.01.001

09 Appleton KM, Voyias PD, Sallis HM, Dawson S, Ness AR, Churchill R, Perry R. Omega-3 fatty acids for depression in adults. Cochrane Database of Systematic Reviews 2021;11:CD004692. https://doi.org/10.1002/14651858.CD004692.pub5