WHITE PAPER AUGUST 2026

When the Measure Moves

Rater bias and inconsistent scoring as threats to clinical outcome assessment data in CNS trials

IN THIS PAPER

ECOALA PTY LTD · AUSTRALIA
ECOALA.COM.AU

ABSTRACT

Central nervous system (CNS) trials depend on clinical outcome assessments (COAs) in a way that few other therapeutic areas do: the endpoint is not a laboratory value or an imaging finding, but a human judgement about another human's symptoms. That dependence creates a distinctive vulnerability. Rater bias, inconsistent scoring, expectation effects and inflated baseline ratings inject measurement noise and systematic error into the very data on which go/no-go decisions rest. This paper reviews the peer-reviewed evidence from the past decade on how and where COA data in CNS trials become unreliable or unusable, what that costs sponsors in statistical power and failed studies, and which mitigations — calibrated rater training, centralised and site-independent surveillance, remote and electronic administration, objective digital measurement, and analytical adjustment — have demonstrated measurable benefit. It is written for clinical trial and pharmaceutical professionals who work alongside COA data rather than within COA science.

01 Introduction: why CNS trials live and die by their COAs

In oncology, a tumour can be measured. In cardiology, blood pressure can be read from a cuff. In most CNS indications — depression, schizophrenia, Alzheimer's disease, Parkinson's disease — the primary endpoint is a score generated when a trained clinician interviews or observes a participant and converts what they see and hear into numbers on a rating scale. Instruments such as the Montgomery–Åsberg Depression Rating Scale (MADRS), the Positive and Negative Syndrome Scale (PANSS) and the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) are the effective currency of CNS drug development.

These clinician-reported outcomes (ClinROs) are, by design, structured subjective judgements. Their reliability is therefore not a property of the instrument alone: it is a property of the instrument, the rater, the training programme, the visit conditions and the incentives operating at the site. When any of those elements drifts, the data drift with them. Recent work on functional unblinding notes that CNS studies are particularly susceptible precisely because they rely on relatively subjective symptom assessments[10].

The consequences are not abstract. Measurement noise shrinks observable effect sizes, inflates required sample sizes, and — in the presence of systematic bias — can produce false-negative trials of genuinely effective drugs or false-positive signals that collapse in phase 3[8].

This paper examines the two intertwined problems named in its title: rater bias (systematic error introduced by the person doing the scoring) and inconsistent scoring (random or semi-random variability between raters, between visits and between sites). It then surveys the mitigations with published evidence behind them.

02 The instruments themselves are part of the problem

Before considering the rater, it is worth acknowledging that the measurement foundations in CNS research are less solid than they appear. A content analysis of seven commonly used depression scales found that, collectively, they encompass 52 distinct symptoms, with a mean content overlap between scales of only 0.36; roughly 40% of symptoms appear in just a single scale[1]. Two trials that both claim to measure “depression severity” may in fact be measuring substantially different constructs, which threatens replicability and the generalisability of any single trial's result[1].

The problem has not resolved with time. A systematic review of 450 randomised controlled trials in unipolar and bipolar depression registered between 2018 and 2022 identified 388 different outcome measures in use. The single most common instrument, the Hamilton Depression Rating Scale, was used in 40% of trials — yet covered at most 59% of the 80 depression domains that patients and clinicians identified as mattering to patients[2].

Even where a scale is well established, its inherent measurement variability can be large relative to the treatment effects being sought. Registry analyses of the Mini-Mental State Examination in dementia populations estimated that score changes of less than approximately five points over a year may simply reflect expected test variability combined with expected decline, with test–retest reliability coefficients between 0.70 and 0.90[3]. When the signal a sponsor hopes to detect is smaller than the instrument's noise floor, everything downstream — rater performance included — must be managed with unusual rigour.

03 Rater bias: where it comes from

3.1 Inter-rater and intra-rater variability

The most basic failure mode is that two competent clinicians, watching the same patient, produce different scores. In Parkinson's disease, a study in which eight clinicians independently rated video-recorded MDS-UPDRS hand movements found poor agreement for most items — an intraclass correlation coefficient (ICC) as low as 0.14 for postural tremor of the left hand and 0.34 for wrist pronation–supination — with agreement improving, though not to uniformly high levels, only after a dedicated training and calibration session[4]. These are not exotic research measures; they are items from the field's standard clinical rating scale, scored by practising clinicians.

Depression ratings show the same pattern at the level of total scores. In a quality-assurance programme in which every site-based MADRS interview was audio-recorded and independently re-scored by a blinded reviewer within 24–48 hours, 94.5% of ratings fell within a three-point concordance window — but the lowest concordance in the entire study, 87.9%, occurred at the baseline visit, exactly the timepoint at which eligibility and stratification decisions are made[5].

3.2 Expectation bias and the engine of placebo response

Raters and participants alike bring expectations into the interview room, and those expectations move scores. Experimental work manipulating patient outcome expectancy within an antidepressant trial demonstrated that expectancy alone — before any medication was given — changed neural activity in the amygdala, and that this change predicted subsequent symptom improvement, confirming expectancy as a primary mediator of placebo effects in antidepressant trials[6].

At portfolio scale, the fingerprints of expectation are visible in regulatory datasets. An analysis of US Food and Drug Administration reviews covering 85 antidepressant trials approved between 1987 and 2013 found that the magnitude of placebo response has grown steadily for three decades[7]. In schizophrenia, a meta-regression of 167 placebo-controlled antipsychotic trials involving more than 28,000 participants likewise documented placebo response rising over the decades and identified design- and patient-related moderators — including larger samples, more sites and shorter illness duration — associated with larger placebo responses[8]. Placebo conditions in modern CNS trials are not therapeutically inert; they deliver structured clinical attention that produces real measured improvement, complicating the interpretation of treatment differences and of familiar statistics such as the number needed to treat[9].

The practical implication for data quality is direct: any rater tendency to score in the direction of expected improvement is amplified across every arm of the study, eroding drug–placebo separation.

3.3 Baseline score inflation

A specific and well-documented form of rater bias occurs at screening and baseline, where entry criteria create an incentive — conscious or not — to score participants as severe enough to qualify. In a Japanese phase 3 trial of vortioxetine, central monitoring that compared physician-rated and patient-rated symptom scores at baseline and over time was credited with reducing baseline score inflation, by drawing site attention to discrepancies between the two rating sources; the analysis further showed that patient subgroups defined by concordant baseline ratings exhibited the smallest placebo response[11]. The MADRS quality-assurance programme described above independently corroborates the baseline problem: rating concordance was at its worst at the eligibility visit and improved steadily at every subsequent visit once monitoring was in place[5].

Baseline inflation is doubly corrosive. It admits participants whose true severity is below the intended threshold — participants with more room for apparent spontaneous "improvement" — and it guarantees regression toward the mean that manifests as placebo response.

3.4 Functional unblinding

Even a well-calibrated, well-intentioned rater can be biased by information leakage. Treatment-emergent adverse events with a recognisable signature — the cholinergic side effects of a muscarinic agonist, for example — can functionally unblind site raters, who may then score in line with their inference about treatment allocation[10]. Investigations of this problem in three placebo-controlled schizophrenia trials used site-independent raters, blinded to adverse events, who re-scored audio-recorded PANSS interviews; the remote ratings closely replicated site-based scores (ICCs of 0.88 at baseline and 0.93 at endpoint), and in that programme no impact of adverse-event-driven unblinding was detected[10]. The reassuring result should not obscure the method's significance: the sponsors judged the risk serious enough to build a dedicated verification apparatus, and the published framework now exists precisely because functional unblinding is a credible threat to CNS trial integrity[10].

3.5 Severity-dependent scoring drift

Paired-ratings surveillance has also revealed a subtler pattern. In a negative-symptom schizophrenia study in which more than 1,000 site-based PANSS and Brief Negative Symptom Scale interviews were video-recorded and independently re-scored, site raters scored higher than remote raters when symptom severity was high, and lower than remote raters when severity was low — a bidirectional "score expansion" observed previously in acute schizophrenia and major depression samples as well[12]. Whatever its cause, the effect stretches the apparent range of change over a trial and is invisible without an independent second set of eyes on the same interviews.

04 What unreliable COA data costs

The costs of the failure modes above arrive through two channels. The first is statistical. Random scoring inconsistency is noise; it widens confidence intervals and shrinks standardised effect sizes, so that a real drug effect becomes harder to distinguish from placebo at any given sample size. The second channel is systematic. Expectation bias, baseline inflation and functional unblinding do not merely add noise — they move both arms of the trial in ways that specifically compress drug–placebo separation.

Re-analytic work makes the stakes concrete. Conventional statistical analyses implicitly assume that every participant carries the same propensity to show a placebo effect; when that assumption fails — as it demonstrably does — trials with a high proportion of placebo-prone participants suffer inflated false-negative risk, and trials with the opposite composition suffer inflated false-positive risk[13]. A propensity-weighted re-analysis of one failed and one negative antidepressant trial recovered an enhanced drug signal in the failed trial while correctly confirming the absence of effect in the negative one — evidence that measurement- and expectancy-driven distortion can plausibly flip a trial's headline outcome[13].

For a sponsor, the arithmetic is unforgiving. A false-negative phase 2 study can terminate a viable asset; a noisy phase 3 programme can demand hundreds of additional participants; and any of these failure modes discovered late — during regulatory review or data audit — can render an entire COA dataset unusable as evidence.

05 What demonstrably helps

5.1 Calibrated training, not one-off certification

Training works, but only the right kind. In the Parkinson's inter-rater study, a single structured calibration session — reviewing high-variance cases against scale instructions as a group — measurably improved agreement on most items[4]. In depression, the combination of rigorous initial training with continuous performance feedback achieved an intraclass correlation of 0.984 between site raters and independent reviewers across 236 ratings, with a mean absolute score difference of under two MADRS points[5]. The common thread is that calibration is treated as an ongoing process anchored to real recorded assessments, not a slide deck completed at the investigator meeting.

5.2 Central and site-independent surveillance

The strongest published results come from programmes that put a second, blinded scorer on the same primary data. Audio-based independent review of every MADRS interview, with a tight (≤3-point) concordance standard and remediation triggered within days, sustained concordance above 94%[5]. Video-based review extends the method to instruments where visual observation matters, and delivered ICCs of 0.84–0.87 across more than 1,000 paired schizophrenia ratings while catching and remediating outlier raters mid-study[12]. Central statistical monitoring of rating patterns — comparing clinician and patient ratings for discordance — has been credited with reducing baseline inflation and placebo response in a phase 3 programme[11]. Site-independent re-scoring by raters deliberately blinded to adverse events adds a defence against functional unblinding that no amount of site-level training can provide[10].

5.3 Remote administration and electronic capture

Remote assessment, once viewed as a compromise, now has direct equivalence evidence in CNS populations: a multicentre randomised comparison of remote versus in-person MADRS interviews in major depressive disorder found strong agreement (overall ICC 0.886), supporting remote interviews as a feasible alternative for severity assessment[14]. Electronic capture also changes the data-quality equation in ways paper never could — enforced completeness, timestamps, audit trails and structured administration — but the industry's own best-practice work is candid that these benefits are only realised when eCOA datasets are built to consistent standards, with defined structures, validation and quality control from the outset rather than retrofitted at analysis[15].

5.4 Objective and computational measurement

The longer-term direction of travel is toward measurement that does not depend on a human scorer at all for its consistency. A computer-vision system assessing MDS-UPDRS finger-tapping from webcam video achieved a mean absolute error of 0.58 points against expert consensus — outperforming two certified human raters (0.83 points), though still slightly behind expert neurologists (0.53 points)[16]. Such tools will not replace clinical judgement in the near term, but as adjuncts they offer something raters cannot: perfect intra-rater reliability, indifference to expectation, and scalability across sites and time zones[16].

5.5 Design and analysis that respect the placebo problem

Finally, measurement quality interacts with design and analysis choices. Meta-regression evidence identifies modifiable design features — among them trial size, site count and population characteristics — associated with larger placebo responses[8], and analytic frameworks now exist to adjust for individual placebo-response propensity rather than assuming it away[13]. None of these substitutes for reliable ratings, but together they reduce the probability that residual measurement error decides the fate of a development programme.

06 Conclusion

CNS trials will remain dependent on human judgement for as long as the diseases they study are defined by experience and behaviour rather than biomarkers. That dependence is manageable — but only if sponsors treat COA reliability as an operational discipline with the same seriousness applied to pharmacovigilance or data management. The published evidence of the past decade supports a clear, layered playbook: choose instruments with honest awareness of their content and noise characteristics; train and continuously calibrate raters against recorded assessments; put independent, blinded eyes on primary rating data throughout the study, with special vigilance at baseline; capture data electronically to standards designed before first patient in; and adopt objective digital measures and placebo-aware analytics as they mature.

Every one of these measures exists because a trial somewhere generated data that could not be trusted. The organisations that internalise that lesson before their pivotal study — rather than in its post-mortem — will run smaller, faster, more decisive CNS programmes.

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 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

02 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

03 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

04 Kenny L, Azizi Z, Moore K, Alcock M, Heywood S, Johnson A, et al. Inter-rater reliability of hand motor function assessment in Parkinson's disease: impact of clinician training. Clinical Parkinsonism & Related Disorders 2024;11:100278. https://doi.org/10.1016/j.prdoa.2024.100278

05 Sapko MT, Kolesar C, Sharp IR, Javitt JC. Quality assurance of depression ratings in psychiatric clinical trials. Journal of Clinical Psychopharmacology 2025;45(1):28–31. https://doi.org/10.1097/JCP.0000000000001936

06 Zilcha-Mano S, Wang Z, Peterson BS, Wall MM, Chen Y, Wager TD, et al. Neural mechanisms of expectancy-based placebo effects in antidepressant clinical trials. Journal of Psychiatric Research 2019;116:19–25. https://doi.org/10.1016/j.jpsychires.2019.05.023

07 Khan A, Fahl Mar K, Faucett J, Khan Schilling S, Brown WA. Has the rising placebo response impacted antidepressant clinical trial outcome? Data from the US Food and Drug Administration 1987–2013. World Psychiatry 2017;16(2):181–192. https://doi.org/10.1002/wps.20421

08 Leucht S, Chaimani A, Leucht C, Huhn M, Mavridis D, Helfer B, et al. 60 years of placebo-controlled antipsychotic drug trials in acute schizophrenia: meta-regression of predictors of placebo response. Schizophrenia Research 2018;201:315–323. https://doi.org/10.1016/j.schres.2018.05.009

09 Roose SP, Rutherford BR, Wall MM, Thase ME. Practising evidence-based medicine in an era of high placebo response: number needed to treat reconsidered. British Journal of Psychiatry 2016;208(5):416–420. https://doi.org/10.1192/bjp.bp.115.163261

10 Targum SD, Horan WP, Davis VG, Breier A, Brannan SK. Methods to address functional unblinding of raters in CNS trials. Translational Psychiatry 2025;15(1):47. https://doi.org/10.1038/s41398-025-03262-1

11 Watanabe Y, Nishimura A, Kikuchi T, Sawada N, Imazaki M, Inada I, Watanabe K. Central monitoring of depression and anxiety symptoms reduces placebo responses in depression clinical trials: a post hoc exploratory analysis of data from the phase III CCT-004 trial of vortioxetine. Neuropsychopharmacology Reports 2022;42(4):468–477. https://doi.org/10.1002/npr2.12288

12 Targum SD, Ge T, Asgharnejad M, Reksoprodjo P, Singh JB, Murthy V. Use of video-recordings of site-based interviews for quality assurance in a study of subjects with schizophrenia and persistent negative symptoms. Schizophrenia Research 2024;272:61–68. https://doi.org/10.1016/j.schres.2024.08.014

13 Gomeni R, Hopkins S, Bressolle-Gomeni F, Fava M. Interpreting clinical trial outcomes complicated by placebo response with an assessment of false-negative and true-negative clinical trials in depression using propensity-weighting. Translational Psychiatry 2023;13(1):388. https://doi.org/10.1038/s41398-023-02685-y

14 Sumiyoshi T, Morio Y, Kawashima T, Tachimori H, Hongo S, Kishimoto T, et al. Feasibility of remote interviews in assessing disease severity in patients with major depressive disorder: a pilot study. Neuropsychopharmacology Reports 2024;44(1):149–157. https://doi.org/10.1002/npr2.12411

15 Hudgens S, Kern S, Barsdorf AI, Cassells S, Rowe A, King-Kallimanis BL, et al. Best practice recommendations for electronic patient-reported outcome dataset structure and standardization to support drug development. Value in Health 2023;26(8):1242–1248. https://doi.org/10.1016/j.jval.2023.02.011

16 Islam MS, Rahman W, Abdelkader A, Lee S, Yang PT, Purks JL, et al. Using AI to measure Parkinson's disease severity at home. npj Digital Medicine 2023;6(1):156. https://doi.org/10.1038/s41746-023-00905-9