DELRICHT RESEARCH

WHITE PAPER

Managing Diagnostic Ambiguity in Psychiatric Research

September 2025

Why Psychiatric Trials Fail More Often

Psychiatric trials are among the most complex in all of clinical research, not because treatments lack potential, but because reliably measuring their effects remains one of the field's greatest challenges. The stakes are enormous: an estimated 1 in every 8 people (nearly 970 million worldwide) lives with a mental disorder,[1] with anxiety and depression being the most common. Yet despite this global burden, the success rate in Phase II trials for psychiatric drugs is only 24%, the lowest among 14 major disease areas.[2]

1 in 8 people
live with a mental disorder[1]

Only 24%
of phase II psychiatric trials succeed[2]

At the heart of this challenge is not only diagnostic complexity, but also measurability. Unlike areas such as diabetes or dyslipidemia, where lab values like A1c or LDL provide clear benchmarks, psychiatric conditions rarely present as neatly defined, singular disorders. Instead, patients often live with overlapping symptoms, comorbidities, and fluctuating severity, making consistent diagnosis and measuring efficacy a challenge from the very first screening visit. Still, the FDA requires evidence from at least two well-controlled clinical trials, showing statistically significant and clinically meaningful improvements on validated psychiatric rating scales, before a therapy can be considered effective.

Evolving diagnostic definitions, subjective rating scales, and rigid protocol criteria compound the problem. Eligibility often hinges on scoring thresholds and structured assessments that are less familiar in routine practice, leading to the exclusion of otherwise appropriate patients and a participant pool that may differ from the real-world population the drug is intended to treat. Added to this are the operational realities of psychiatric research: long, complex visits; high patient sensitivity; and potential disconnect between local and central raters, which can contribute to slower recruitment, inconsistent data, and weakened signal detection.

While sponsors and CROs cannot simplify the nature of psychiatric illness, they can design protocols and treatment plans better equipped to manage it. Approaches that integrate everyday clinical practice with research requirements, where raters apply scales in context, maintain continuity across visits, and evaluate outcomes with clinical as well as protocol-driven insight, offer a way to strengthen study integrity and improve the likelihood of detecting true treatment effects.

Diagnostic Ambiguity & the Enrollment Bottleneck

The first, and often most difficult, phase of a psychiatric trial is getting the right patients enrolled. Unlike many medical conditions, psychiatric diagnoses rarely follow a clear-cut template. Patients often present with multiple overlapping symptoms, major depression with anxiety, PTSD with depressive episodes, ADHD with social anxiety, making it difficult to fit them neatly into rigid protocol-defined categories. Large epidemiological studies highlight just how rare singular psychiatric diagnoses are. For example, in Kessler et al.'s DSM-IV analysis of U.S. adults with a past-year psychiatric disorder, only 55% had a single diagnosis, while 45% had two or more.[3] As Dr. Alison Baker, a board-certified psychiatrist in Atlanta, GA, explains, "Psychiatric trials are uniquely challenging because they rely on patient recall and insight, are affected by strong placebo responses, and often require parsing overlapping symptoms and comorbidities. These factors make screening and monitoring outcomes much more complex."

> ***"Psychiatric trials are uniquely challenging because they rely on patient recall and insight, are affected by strong placebo responses, and often require parsing overlapping symptoms and comorbidities. These factors make screening and monitoring outcomes much more complex.***

Dr. Alison Baker
Atlanta, GA
Alison Baker, MD Clinic

In clinical practice, psychiatrists often rely on a combination of clinical judgment, patient history, and selected tools rather than structured rating scales alone. While research instruments such as the Brief Psychiatric Rating Scale (BPRS), Hamilton Rating Scales for Anxiety and Depression (HAM-A/HAM-D), and the Positive and Negative Syndrome Scale (PANSS) are standard in research, they are not routinely used in many encounters. A national survey of over 300 psychiatrists found that many cite barriers to routinely using standardized rating scales in practice, including time burden, limited clinical utility in busy settings, and lack of training.[4] In real-world care, physicians may draw on multiple tools and clinical expertise to confirm a diagnosis; in trials, however, eligibility may hinge on a single score threshold, excluding patients who would otherwise benefit and be representative of the population the therapy is intended to serve. This highlights not a lack of rigor, but the difference between real-world care and research requirements, a gap trials must thoughtfully bridge when defining eligibility and outcomes.

Centralized raters play an important role in many psychiatric trials, helping to standardize assessments, reduce site-level bias, and lower placebo effects while promoting consistency across sites. At the same time, when rating scales are administered by raters without a prior relationship with the patient, certain nuances of symptom presentation can be harder to capture. Patients may respond differently in the absence of established rapport, feel more anxious with a new evaluator, or present in ways that do not fully reflect their clinical status. Both approaches bring important strengths, but balancing standardization with clinical context is essential for reliable outcomes and becomes even more critical as patients move through the trial.

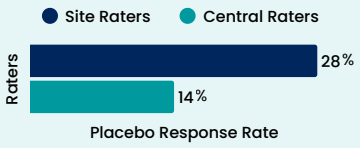
Variability in Evaluation and the Breakdown of Continuity

Once a participant is enrolled, the challenge shifts from eligibility to accurately tracking their progress over time. In psychiatric trials, this is seldom straightforward. Patients are often evaluated by multiple raters, some local, some central, each bringing a slightly different perspective to the assessment. While this structure is designed to enhance consistency, it can also introduce variation when assessments occur at different timepoints or without the continuity of an established patient–clinician relationship.

For example, a patient might complete an in-person interview with a site rater on Monday, then answer the same questions for a central rater by phone on Friday, yielding different results influenced by mood changes, memory recall, or even differences in how questions are toned or phrased. The absence of visual cues in phone-based assessments may further limit accuracy, with important nonverbal signals, such as restlessness, flat affect, and tearfulness, going unnoticed. Even video assessments can miss subtle movements or gestures outside the camera frame.

Research comparing site-based and centralized raters in major depressive disorder trials underscores how impactful these differences can be. In one study, more than a third of patients who qualified for enrollment under site ratings would have been excluded if centralized ratings were applied at baseline, and placebo response rates were markedly higher for site raters than central raters (28% vs. 14%).[5] These findings suggest that centralized raters strengthen standardization and may help reduce placebo effects, while site raters, drawing on rapport and clinical context, may capture a broader and more representative patient population. Both approaches are valid, but together they highlight the broader challenge of ensuring reliable, clinically meaningful assessments in psychiatric research.

Placebo Response Rate with Site vs. Central Rater



Because psychiatric endpoints are subjective and symptoms can fluctuate daily, variability across raters can compound noise in the data. Without continuity, it becomes harder to determine whether changes reflect a true treatment effect or differences in how the assessment was conducted. Addressing this variability is critical to strengthening inter-rater reliability, improving signal detection, and ultimately ensuring that study outcomes reflect a therapy's true potential.

Measuring Improvement in a Complex Landscape

By the end of a psychiatric trial, the core question appears simple: did the patient improve? In practice, answering that question is more nuanced. Standardized rating scales remain the foundation for measuring efficacy, providing the structure and comparability needed across sites and studies. At the same time, these tools may not fully reflect the day-to-day realities of patients or the subtleties clinicians observe in practice. For example, a patient may report feeling more capable at work or in relationships even if their rating scale score shows only modest change, while another

may achieve measurable symptom reduction that is less immediately visible in daily life.

Capturing improvement is further complicated by variability introduced earlier in the trial. Differences in baseline assessments, rater approaches, and natural symptom fluctuations all contribute to the difficulty of establishing a stable reference point. Psychiatric symptoms are rarely linear; patients have good days and bad days, and timing alone can influence how change is recorded. Ecological momentary assessment (EMA) studies highlight this complexity, showing that depressed individuals experience more frequent dysphoric days and greater variability in negative affect than non-depressed controls, and often recall symptoms less accurately when asked retrospectively.[6] As a result, assessments conducted on different days or under different circumstances can present very different pictures of the same patient. As Dr. Caroline Campion, a board-certified psychiatrist in New Orleans, LA, points out, "Psychiatric clinical trials rely on subjective reports, which can hinder reproducibility as an individual's daily symptoms may not reflect their global experience. In striving to obtain accurate information, additional layers of assessments are added, and trial protocols become more complex."

Protocol-defined measures of improvement are indispensable for demonstrating efficacy, but integrating clinical context alongside those metrics can create a more complete understanding of therapeutic impact. Balancing the consistency of structured endpoints with the richness of patient and clinician-level insights represents a clear opportunity to strengthen reliability in psychiatric research.

> ***"Psychiatric clinical trials rely on subjective reports, which can hinder reproducibility as an individual's daily symptoms may not reflect their global experience. In striving to obtain accurate information, additional layers of assessments are added, and trial protocols become more complex.***

Dr. Caroline Campion
New Orleans, LA
Touro Medical Center

Reliability Starts with Raters and the Right Site Model

Diagnostic ambiguity, rater variability, and differing interpretations of endpoints are persistent challenges in psychiatric trials, but they can be partially mitigated with a site model designed for consistency and context. Dedicated research operations within clinical practice, where investigators are both clinically active and research-trained, offer a practical framework for overcoming these issues.

1. Addressing Diagnostic Ambiguity at Screening

Overlapping symptoms and comorbidities make psychiatric eligibility complex. Site models that engage physician-investigators allow those actively treating psychiatric patients to apply research rating scales directly with them. This combination of everyday psychiatric expertise and protocol familiarity enables nuanced diagnostic judgment, helping distinguish primary from secondary conditions, clarify borderline cases, reduce unnecessary screen failures, and maintain protocol adherence.

2. Ensuring Reliable Outcomes Through Rater Continuity

In traditional trial designs, participants may be assessed by multiple raters, local, central, or rotating, which can introduce variation in how changes are observed over time. While central raters play an important role in standardization, consistency across onsite visits is equally critical. Embedded research site models address this by ensuring that the same investigator or trained rater follows the participant from baseline to endpoint. This continuity provides familiarity with the patient's baseline presentation, enabling raters to detect subtle changes over time and apply rating scales more consistently, while reducing discrepancies that may arise when assessments are spread across different raters or settings.

3. Aligning Efficacy Measurement with Clinical Relevance

Psychiatric trial endpoints are typically defined by standardized scale scores, which provide essential structure and comparability across studies. At the same time, they may not always capture the full range of functional or quality-of-life improvements that matter to patients and clinicians. By retaining the same clinical perspective from screening through endpoint, investigators who can help align scale-based measurements with real-world changes will be most effective in progress reporting. This continuity ensures that efficacy is evaluated with both protocol-defined rigor and clinically meaningful interpretation, offering a more complete picture of therapeutic benefit.

By integrating care and research in this way, sites can offer a structured path to cleaner data, stronger inter-rater reliability, and trial outcomes that better capture a treatment's true potential.

Psychiatric research will always contend with diagnostic ambiguity, evolving definitions, and the inherent subjectivity of rating scales, but these challenges do not have to limit trial success. When investigators bring both clinical expertise and research discipline to the same encounter, they can apply rating scales with continuity, interpret symptom changes in context, and distinguish signal from noise more effectively. This alignment supports more accurate eligibility decisions, reduces inconsistencies in assessment, and helps ensure that trial populations reflect the patients who will ultimately use the therapy.

In a therapeutic area where progress is hard-won, strengthening measurement reliability is one of the clearest opportunities available to sponsors and CROs. With the right site design, even complex endpoints can be measured in ways that are consistent, clinically relevant, and actionable. By prioritizing consistency, context, and continuity at every stage, psychiatric trials can generate outcomes that both meet regulatory standards and capture a treatment's ability to improve patients' lives.

Learn how DelRicht Research's embedded model delivers cleaner data and stronger outcomes in psychiatry trials.

Visit www.DelRichtResearch.com

References

1. World Health Organization. Mental Disorders. World Health Organization. Published June 8, 2022. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

2. Yatham LN. All levels of the translational spectrum must be targeted to advance psychopharmacology and improve patient outcomes. World Psychiatry. 2023;22(1):75-76. doi:https://doi.org/10.1002/wps.21060

3. Kessler RC, Chiu WT, Demler O, Walters EE. Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry. 2005;62(6):617-627. doi:https://doi.org/10.1001/archpsyc.62.6.617

4. Zimmerman M, McGlinchey JB. Why Don't Psychiatrists Use Scales to Measure Outcome When Treating Depressed Patients? The Journal of Clinical Psychiatry. 2008;69(12):1916-1919. doi:https://doi.org/10.4088/jcp.v69n1209

5. Kobak KA, Leuchter A, DeBrota D, et al. Site Versus Centralized Raters in a Clinical Depression Trial. Journal of Clinical Psychopharmacology. 2010;30(2):193-197. doi:https://doi.org/10.1097/jcp.0b013e3181d20912

6. Armey MF, Schatten HT, Haradhvala N, Miller IW. Ecological momentary assessment (EMA) of depression-related phenomena. Current Opinion in Psychology. 2015;4:21-25. doi:https://doi.org/10.1016/j.copsyc.2015.01.002