Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Narrative Methodological Review of Discordance Between Statistical and Clinical Significance in Research on Digital Interventions for Adolescent Depression

Melika Mirzabeygi
University of Social Welfare and Rehabilitation Sciences
Fatemeh Khosravi Shadmani
khosravishadman@gmail.com

University of Social Welfare and Rehabilitation Sciences

Systematic Review

Keywords: adolescent digital depression, clinical and Statistical significance, minimal clinically important difference, p-value, Multidimensional Clinical Interpretation Framework

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11178844/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background and Objective: Across trials of digital intervention for adolescent depression, statistically significant findings often coexist with only marginal clinical improvement whereas underpowered studies may fail to detect clinically meaningful benefits. This study aimed to apply the Multidimensional Clinical Interpretation Framework (MCIF), a seven-domain sequential framework for evaluate the clinical relevance of evidence from adolescent's Digital depression interventions trials.

Methods: A narrative methodological review with conceptual framework development was conducted. PubMed, PsycINFO, Scopus and Web of science were searched from January 2005 to March 2026, with targeted citation tracking of key methodological publications. Two independent reviewers screened titles and abstracts, with disagreements resolved by consensus. Of 3,847 identified records, 42 studies met eligibility criteria.

Results: Review of the 42 included trials showed that clinically meaningful outcomes were inconsistently reported with statistical significance frequently prioritized over interpretable effect magnitudes. To address these shortcomings, the Multidimensional Clinical Interpretation Framework (MCIF) was developed. It integrates seven interpretive domains (statistical outcome, effect size, measurement reliability, clinical meaningfulness, functional improvement, implementation feasibility, and reporting transparency) into a single sequential evaluative structure. The MCIF provides a practical and step-by-step approach to evaluating the real-world relevance of digital depression interventions for adolescents.

Conclusion: reliance of statistical significance alone is insufficient for interpreting effectiveness. The MCIF is a tool that provides a structured, multidimensional approach to improve interpretation, strengthen reporting, and facilitate translation of research evidence into clinical practice.

Introduction

Depression is highly prevalent among adolescents, with lifetime prevalence estimates of major depressive disorder reaching approximately 20% by late adolescence [1]. Despite the high prevalence of depression among adolescents, a substantial treatment gap exists, with many young people never receiving any form of care [2]. Although effective psychological and pharmacological treatments are available a substantial proportion of adolescents with depression do not receive appropriate treatment due to limited access to mental health services, workforce shortages, financial constraints, and persistent stigma surrounding mental illness [3]. To address these barriers, researchers have been developing tools such as smartphone programs, websites, and virtual reality platforms to help depressed adolescents [4]. These tools include online therapy apps that adolescents can use on their phones [5, 6], and blended programs that combine online and in-person therapy [7].

A wide range of digital interventions have been developed for adolescents with depression [4, 5], including internet-based cognitive behavioral therapy (iCBT), apps that track symptoms, virtual reality platforms, applications and chatbots [5]. The effectiveness of these interventions has been evaluated in randomized controlled trials and systematic reviews. Meta-analysis estimates generally suggest small to moderate effects, with reported effect sizes typically ranging between approximately 0.20 and 0.50, depending on intervention type and study characteristics, values that fall within the small-to-medium range under conventional benchmarks for interpreting standardized effect sizes [8, 9] and that are typically synthesized using standard meta-analytic pooling methods [10, 11, 12].

A major methodological challenge lies not only in determining whether digital interventions are statistically effective, but in establishing whether the observed effects are also clinically meaningful [13, 19]. In large meta-analyses of digital interventions, pooled effects may reach statistical significance even when the effect sizes remain small [11, 12]. Conversely, small or underpowered studies may detect real within-participant improvements yet still fail to achieve $p < 0.05$ [15], a problem that is particularly common in adolescent digital psychiatry research [5]. At the same time, the absence of a statistically significant difference does not necessarily indicate that an intervention lacks practical value [16]. Discrepancies between statistical significance and clinical relevance have long been recognized across healthcare, underscoring the need for more comprehensive approaches to evidence interpretation [14, 41]. This concern reflects a broader shift toward emphasizing effect sizes and confidence intervals rather than dichotomous significance testing [17], and echoes longstanding warnings that overreliance on statistical significance can distort the published research record [13, 18].

The gap between what the numbers indicate and what genuinely makes a difference in adolescents' lives requires careful attention in research on digital interventions for depression. Although increasing attention has been given to effect sizes, minimal clinically important differences, and other indicators of clinical relevance, these approaches are typically considered in isolation and do not provide an integrated framework for interpreting intervention effectiveness. To our knowledge, no previous methodological review has systematically examined statistical–clinical discordance or proposed a comprehensive interpretive framework specifically for adolescent digital depression interventions. Accordingly, this narrative methodological review proposes the Multidimensional Clinical Interpretation Framework (MCIF). The MCIF integrates seven interrelated domains: statistical outcome, effect size, measurement reliability, clinical meaningfulness, functional improvement, implementation feasibility, and reporting transparency. By moving beyond an exclusive focus on statistical significance, the framework aims to support more nuanced evidence interpretation, improve research reporting, and enhance the translation of digital mental health research into clinical practice.

Methods

Review Design

This study was conducted a narrative methodological review with conceptual framework development, drawing on general systematic review conduct standards where applicable [20]. The review aimed to identify methodological challenges in the evaluation of digital interventions for adolescent depression and develop an integrated interpretive framework for the clinical interpretation of research finding. Throughout this review, the term depression refers to clinically significant depressive symptoms, as measured by validated self-report or clinician-rated instruments such as the PHQ-9, rather than a formal diagnosis of major depressive disorder, unless a specific study explicitly used diagnostic criteria; this distinction matters because most digital intervention trials in this literature enroll participants on the basis of elevated symptom scores rather than a structured clinical diagnosis. Rather than synthesizing treatment effects systematically, this review focused on methodological approaches to interpreting finding from existing studies. This review was reported in accordance with the SANRA (Scale for the Assessment of Narrative Review Articles) guidance [21]. Methodological decisions regarding the scope of evidence, eligibility criteria, and conceptual domains were specified a priori to improve transparency and minimized selection bias during framework development. A narrative review approach was considered appropriate because the object was to synthesize methodological

concepts and interpretive framework rather than estimate intervention effectiveness or quantitatively summarize treatment outcomes.

Search Strategy

A comprehensive literature search was conducted in PubMed, PsycINFO, Scopus and Web of science to identify studies on digital mental health interventions for adolescent depression as well as methodological approaches addressing the interpretation of clinical significance. The search period spanned January 2005 to March 2026. Search terms were combined using Boolean operators and were developed to capture three main concepts: adolescent depression, digital interventions, and method for evaluating clinical significance. Key terms included: "digital depression," "adolescent digital mental health," "internet-based CBT adolescent depression," "app-based depression youth," "smartphone intervention adolescent," "clinical significance," "effect size," "minimal clinically important difference," "reliable change index," "p-value," and "null hypothesis significance testing.

Key methodological publications published before 2005, including seminal work by Cohen, Jacobson and Truax, and Hedges and Olkin, were identified through citation tracking and reference list searches to ensure comprehensive coverage of fundamental methodologic concepts. The search was limited to English-language publications.

Eligibility Criteria

Eligible sources included origin studies evaluating digital interventions for adolescent depression involving participants aged 12–18 years. Studies were required to report quantitative clinical outcomes and inferential statistical results and provided stuffiest methodological and outcome information to allow interpretation of both statistical and clinical significance. Eligible outcomes included measure of depressive symptom severity, functional status or other clinically relevant indicators

We also included key methodological and statistical publications that were directly relevant to the MCIF framework, such as the ASA statements on p-values, papers explaining effect sizes, and publications on clinical significance.

Sources were excluded if they did not report quantitative clinically outcomes, were unrelated to adolescent depression, digital mental health interventions, or methodological approaches to statistical and clinical statistical, or presented duplicate publications based on the same dataset. Case reports and participant studies were also excluded because their findings have limited generalizability and were not suitable for developing the interpretive framework.

Study Selection and Coding

Following removal of duplicate records, 3847 unique publications were identified through the database search. Two independent reviewers screened titles and abstracts for eligibility with disagreements resolved through discussion and consensus. A total of 3,621 records were excluded because they were unrelated to digital interventions for adolescent depression or did not address methodological aspects of statistical and clinical interpretation.

The full texts of the remaining 226 articles were assessed for eligibility. Of these, 184 publications were excluded for not meeting the inclusion criteria, resulting in a final sample of 42 studies. Each included source was assessed using the seven predefined methodological domains of the MCIF framework (Table 1).

Table 1
MCIF Coding Framework: Seven Predefined Methodological Domains

Coding Domain	Operational Definition
Statistical significance	Presence and reporting quality of primary inferential outcome (exact p-values, test statistics, degrees of freedom)
Effect magnitude	Reporting and interpretation of standardized effect-size estimates (Cohen d, Hedges g, odds ratio with 95% CI, rate ratio or risk ratio with %% CI)
Clinical relevance	Presence of clinically meaningful improvement defined against MCID thresholds or established benchmarks for adolescent depression scales
Functional recovery	Reporting of real-world functioning outcomes (social, occupational, educational, quality of life)
Reliable improvement	Reporting of reliable change metrics (Jacobson–Truax RCI or equivalent psychometric criterion)
Reporting transparency	Adequacy of effect-size reporting, confidence intervals, and covariance parameters for secondary synthesis
Implementation relevance	Inclusion of scalability, accessibility, adherence, cost, or feasibility considerations

Discordance Classification

Pattern of statistical and clinical discordance identified across the included literature were synthesized and classified into four conceptually distinct categories based on their underline methodological characteristics (Table2). This classification was developed through iterative comparison of the included intervention and methodological studies and was intended to capture the principal sources of discrepancy in the interpretation of intervention outcomes.

The first two categories describe discordance rising from the relationship between statistical significance and clinical significance.

Specifically, they distinguish situations in which statistically significant finding corresponds to only limited critical benefit from those in which clinically meaningful improvements fail to achieve statistical significance because of methodological constrains, such as inadequate statistical power.

The remaining two categories address discordance related to outcue selection and reporting practices. Unlike first two categories, these form of discordance reflect how study finding are measured, reported, and interpreted, irrespective of whether the observed results are statistically and clinically positive. Together, these four categories provided the conceptual foundation for the development of the Multidimensional Clinical Interpretation Framework (MCIF).

Table 2
Four Principal Discordance Types in Adolescent Digital Depression Research

Discordance Type	Description	Clinical Implication
Type I	Statistically significant but clinically marginal effects	Large-sample amplification of trivial differences; may mislead treatment guidelines
Type II	Clinically meaningful improvement despite non-significant inferential testing	Underpowered studies obscure genuine benefit; patient-level improvement is masked
Type III	Symptom reduction without functional recovery	Scale improvement does not translate to real-world benefit; outcome selection is inadequate
Type IV	Methodologically under-reported but potentially clinically relevant outcomes	Insufficient reporting prevents secondary synthesis and clinical interpretability

Results

Overview of Identified Discordance Patterns

Analysis of the 42 included sources identified four recurring patterns of statistical-clinical discordance in studies evaluating digital interventions for adolescent depression (Table 2). These patterns reflected inconsistencies between statistical findings, clinical relevance, outcome selection, and reporting practices. Type I discordance (statistically significant but clinically marginal effects) was identified in 1 article [25]: studies reporting statistically significant intervention effect that were accompanied by only modest or clinically limited improvements. This pattern was most commonly identified in trials evaluating internet based and smartphone-delivered interventions.

Type II discordance (clinically meaningful improvement despite non-significant testing) was identified in 1 article [26]: smaller studies that reported improvements in clinical outcomes but failed to demonstrate statistical significance. These studies were generally characterized by limited sample size and insufficient statistical power. Type III discordance (symptom reduction without functional recovery) was identified in 1 article [42]: the study reported changes in depressive symptom severity without evaluating broader indicators of clinical recovery, such as functional outcomes or quality of life. Type IV discordance (methodologically under-reported but potentially clinically relevant outcomes) was identified in 7 articles [4, 5, 23, 24, 25, 27, 42]: some studies differed in how they reported their results. In particular, clinically meaningful measures, such as reliable change indices, responder analyses, and functional recovery indicators, were not used consistently across studies. These counts are drawn from 8 included sources reporting quantitative efficacy findings for adolescent digital depression interventions; several sources are dual-classified because they exhibit more than one discordance pattern simultaneously. Seven of these 8 sources are systematic reviews or meta-analyses that pool findings across multiple primary trials, while one (Merry et al.) is a single randomized controlled trial; where a source reports a pooled effect, the discordance described reflects the pooled estimate rather than any one constituent trial (Table 3).

Overall, most studies emphasized change in symptom severity while providing limited information on clinically meaningful. Only a small proportion incorporated clinically anchored outcome measure, such as reliable change

indices or responder analysis, highlighting substantial variability in the interpretation and reporting of intervention effectiveness. [21].

Representative Empirical Examples

Examples that represent each of the four patterns of statistical-clinical discordance found are provided in Table 3. Examples chosen aim at representing methodological features of each pattern and possible implications they can have when interpreting the effectiveness of interventions. As a whole, the discussed examples demonstrate the drawbacks of using statistical significance only and form the empirical basis for further creation of the MCIF. The examples serve the illustrative purpose and are not to be considered as a quantitative representation of the literature reviewed.

Table 3
Representative Examples of Statistical–Clinical Discordance in Adolescent Digital Depression Intervention Research

First author(s)	Intervention domain	Statistical finding	Clinical interpretation	Discordance type	Key methodological consideration
Garrido et al. (2019; pooled)	Digital mental health interventions (DMHIs) for depression in young people (Garrido et al. [25])	Significant vs. no-intervention control (d = 0.33, 95% CI 0.11–0.55); not significant vs. active control (d = 0.14, 95% CI −0.04–0.31)	Authors note effect "may not always reflect a clinically significant level of change"; moderate effect (d = 0.52) only when highly supervised; attrition ranged 0%–74.3% across included trials (retention ≈ 25.7%–100%); no MCID threshold reported	**Type I / IV**	Low-risk-of-bias subset showed near-null effect (d = 0.01) vs. unclear/high-risk subset (d = 0.44), suggesting bias-driven amplification
Buttazzoni et al. (2021; pooled)	Smartphone-based interventions for internalizing disorders, including depression, in youth (Buttazzoni et al. [42])	Small but significant pooled effect (Hedges g = 0.20, 95% CI 0.02–0.38; k = 12 studies, N = 1370)	Outcome of interest restricted to symptom severity; functional/quality-of-life outcomes and remission rates not part of the pooled analysis	**Type III / IV**	Symptom scales only; no validated functional outcome instrument included in the pooled analysis
Merry et al. (2012; single trial)	SPARX computerised intervention for adolescents (Merry et al. [26])	Non-inferiority shown vs. treatment-as-usual (p > 0.05 for superiority)	Clinically relevant within-group improvement (Cohen's d = 0.3, ITT); remission 43.7% (per-protocol) vs. 26.4% (TAU); functioning improved at follow-up	**Type II**	Superiority framing obscures genuine benefit; adequate power only for non-inferiority
Lehtimaki et al. (2021; pooled)	Digital mental health interventions for adolescents: systematic overview (Lehtimaki et al. [5])	Mixed significance across included trials	Evidence quality rated low to moderate; functional outcomes rarely primary endpoint	**Type IV**	MCID benchmarks absent; RCI not computed in majority of included studies

Note. *These examples are analytical illustrations rather than an exhaustive inventory. CBT, cognitive behavioural therapy; CI, confidence interval; MCID, minimal clinically important difference; PHQ-9, Patient Health Questionnaire-9; RCI, Reliable Change Index; ITT, intention-to-treat. Effect sizes are reported as Hedges g or Cohen d as in the original sources.*

First author(s)	Intervention domain	Statistical finding	Clinical interpretation	Discordance type	Key methodological consideration
Fleming et al. (2018; pooled)	Real-world uptake and engagement with digital self-help for depression (Fleming et al. [24])	Not an efficacy study — reports uptake/engagement data only, not statistical significance of symptom change	Engagement highly variable across interventions (21%–88% minimal use); sustained use/completion only 0.5%–28.6%; underlying trial efficacy not assessed by this review	**Type IV**	Attrition bias; implementation relevance absent from reporting

Note. *These examples are analytical illustrations rather than an exhaustive inventory. CBT, cognitive behavioural therapy; CI, confidence interval; MCID, minimal clinically important difference; PHQ-9, Patient Health Questionnaire-9; RCI, Reliable Change Index; ITT, intention-to-treat. Effect sizes are reported as Hedges g or Cohen d as in the original sources.*

From the analysis of the representative studies, it is evident that patterns related to statistical-clinical disparity exist in all four categories. Larger randomized trials and meta-analyses more often revealed statistically significant differences in intervention effect, while smaller studies more often mentioned clinical benefit without obtaining statistical significance due to small sample sizes [11, 12, 15]

Furthermore, in the reviewed studies, it could be observed that decrease in depressive symptom severity was not always associated with improved recovery indicators. In several studies, positive results in depression scores were found without any benefits in functional areas, including academic, social, or health-related quality of life.

On the whole, representative examples reveal that statistical significance, reduction of symptoms, and clinically meaningful improvement are different dimensions of the same concept.

Seven-Domain Taxonomy for Clinical Interpretation

Considering the results obtained through this literature review and methodological literature on how to interpret the effectiveness of the treatments, a framework was created to interpret the research outcomes about digital interventions for adolescent depression.

The framework contains seven domains which should be taken into account in order to conclude about the effectiveness of the treatment. These domains are presented in the logical order, with each domain based on the previous step (Table 4).

Domains in the beginning of the framework concern the results which are obtained statistically, while domains later in the list take into account if these results mean any benefit for the patient. Together, seven domains form a structured approach to interpreting the results of the treatment and not just considering statistical results. [14, 19, 28].

Table 4
Seven-Domain Multidimensional Clinical Interpretation Framework (MCIF) for Adolescent Digital Depression Interventions

Domain	Primary Metric	Clinical Relevance	Relationship to Existing Approaches
Statistical significance	p-value; exact test statistic	Does not quantify patient benefit; sensitive to sample size	Present in all frameworks; MCIF treats it as one layer only
Effect magnitude	Cohen d / Hedges g / OR with 95% CI	Estimates relative improvement; benchmarkable against clinical norms	GRADE uses similar metrics; MCIF operationalises domain-specific benchmarks
Precision	95% Confidence interval width and bounds	Stability of estimated benefit across plausible sampling scenarios	CONSORT requires CIs; MCIF interprets them clinically
Clinical significance	MCID / responder rate / remission rate	Patient-centred relevance; threshold-anchored interpretation	Jacobson–Truax provides RCI; MCIF integrates RCI with MCID and remission jointly
Reliable improvement	Reliable Change Index (RCI)	Probable genuine change beyond scale unreliability	Jacobson–Truax (1991) source [28]; MCIF embeds within larger architecture
Functional recovery	SF-36; WHODAS; GAF; QoL instruments	Social, educational, and occupational benefit; most clinically proximal	GRADE downrates evidence lacking patient-important outcomes; MCIF prioritizes functional outcomes when available
Implementation significance	Reach; adherence; cost; scalability; accessibility	Public health relevance beyond efficacy under trial conditions	Absent from GRADE/CONSORT/Jacobson–Truax; unique MCIF contribution

The Multidimensional Clinical Interpretation Framework (MCIF)

The multidimensional Clinical Interpretation Framework (MCIF) integrate seven complementary domains into a sequential interpretive process for evaluating the effectiveness of digital interventions for adolescent depression. Rather than relying solely on statistical significance, the framework incorporates statistical findings, clinical relevance, functional outcomes, implementation feasibility, and reporting quality to support a more comprehensive interpretation of intervention effectiveness. By integrating these dimensions within a single structured approach, the MCIF facilitates a more balanced assessment of both the methodological robustness and the real-world clinical value of digital mental health interventions. The conceptual structure, sequential workflow, and practical application of the framework are presented in Figs. 1–3. The

Figure 1. Conceptual Structure of the MCIF for Adolescent Digital Depression Intervention Research

The MCIF conceptualizes intervention interpretation as a seven-domain evaluation framework. Each domain contributes complementary information to the overall assessment of therapeutic benefit

► Layer 1 — Statistical Inference: p-value and test statistic (hypothesis test; null compatibility)

► Layer 2 — Effect Magnitude: Standardised effect size with 95% CI (Cohen d, Hedges g)

► Layer 3 — Confidence interval width, uncertainty assessment, and robustness of effect estimates

► Layer 4 — Clinical Meaningfulness: MCID anchoring, responder rate, remission probability

► Layer 5 — Reliable Improvement: Jacobson–Truax RCI; proportion of reliably changed individuals

► Layer 6 — Functional Recovery: Quality of life, social, educational, and age-appropriate functional outcomes

► Layer 7 — Implementation Relevance: Reach, engagement, adherence, scalability, accessibility, and cost-effectiveness

Key principle: A finding may satisfy upper-layer criteria (statistical significance) without satisfying lower-layer criteria (functional recovery, implementation relevance). Interpretation is strengthened when evidence is considered across all seven domains rather than relying on any single metric in isolation.

Figure 2. Statistical–Clinical Discordance Zones in Adolescent Digital Depression Research

Four principal discordance zones are identified in the two-dimensional interpretive space defined by statistical significance (x-axis) and clinical meaningfulness (y-axis):

Clinical meaningfulness refers to clinically anchored outcomes such as MCID attainment, remission rates, or responder status.

► Zone A (Type I) — HIGH statistical significance / LOW clinical significance

Large-sample amplification; effect below MCID threshold; may overstate the practical importance of treatment effects

► Zone B (Type II) — LOW statistical significance / HIGH clinical significance

Underpowered design; clinically relevant effects remain plausible despite non-significant findings; CI spans clinically meaningful effects

► Zone C — HIGH statistical significance / HIGH clinical significance

Concordant findings; concordant statistical and clinical findings; still requires functional and implementation layers

► Zone D — LOW statistical significance / LOW clinical significance

No convincing evidence of clinically meaningful benefit based on available data; interpretation depends on study precision and confidence interval bounds.

Types III and IV are cross-cutting patterns that can occur in any zone, identified through Layers 5–7.

Figure 3. Recommended Sequential Reporting Pathway for Adolescent Digital Depression Trials

Stage 1 — Statistical Inference Reporting

Recommended reporting elements: Exact p-values | Test statistic and degrees of freedom | Model specification

Avoid: exclusive reliance on binary significant/non-significant interpretations

Stage 2 — Effect Size and Precision Reporting

Recommended reporting elements: Cohen d or Hedges g with 95% CI | Covariance or correlation parameters for repeated-measures analyses when applicable

Stage 3 — Clinical Significance Reporting

Recommended reporting elements: MCID-anchored responder rates | Remission rates | Number Needed to Treat

Stage 4 — Psychometric Reliability Assessment

Recommended reporting elements: Reliable Change Index | Scale reliability (Cronbach alpha) | Proportion reliably improved vs. deteriorated

Stage 5 — Functional Outcome Reporting

Recommended reporting elements: Quality of life | Social functioning | Educational outcomes

Preferred practice: inclusion of at least one validated functional outcome measure in addition to symptom-based outcomes.

Stage 6 — Heterogeneity and Subgroup Reporting

Recommended reporting elements: Moderation analyses | Deterioration rates | Engagement variability

Stage 7 — Implementation and Scalability Reporting

Recommended reporting elements: Adherence rates | Dropout patterns | Cost-effectiveness indicators | Scalability potential

Recommended Reporting Elements for Adolescent Digital Depression Trials

Based on the evidentiary findings from methodology gathered in this review and based on the foundational framework of the Multidimensional Clinical Interpretation Framework (MCIF), a list of recommended reporting elements is proposed to enhance evidence-based interpretation of adolescent digital depression studies (see Table 5 below). The purpose of this list of recommendations is to supplement the current reporting guidelines, which encourage the reporting of statistical, clinical, functional, and implementation outcomes for evidence interpretation.

Table 5
MCIF-Informed Reporting Recommendations for Adolescent Digital Depression Intervention Studies

Reporting Domain	Recommended Minimum Elements	Rationale
Statistical inference	Exact p-values (not p < threshold); test statistic; model specification; multiple comparison correction	Enables secondary reanalysis; Reduces overreliance on dichotomous significance thresholds
Effect estimation	Cohen d or Hedges g with 95% CI; omega-squared for ANOVA; separate within- and between-group effect sizes	Quantifies magnitude independently of sample size
Clinical relevance	MCID-anchored responder rates; remission rates; NNT; comparison against established clinical benchmarks	Connects statistical to patient-centred outcomes
Measurement transparency	Scale reliability (Cronbach alpha, omega); covariance parameters for repeated-measures models	Enables RCI computation; facilitates individual participant data meta-analysis
Functional outcomes	At least one validated measure of quality of life, social, educational, and age-appropriate functional outcomes	Captures broader recovery beyond symptom reduction
Heterogeneity reporting	Subgroup moderation analyses; individual-level deterioration rates; engagement variability	Addresses treatment response heterogeneity critical in adolescent digital trials
Implementation context	Adherence rates; dropout patterns with timing; cost per improved case; scalability assessment	Translates efficacy evidence to real-world deployment feasibility

Discussion

Principal Findings

The evaluation of digital interventions in treating adolescent depression is a complicated process which goes far beyond the assessment of the statistical significance of an intervention [5]. This narrative methodological review revealed the presence of a clear separation between statistical findings and clinical significance in the literature dealing with digital interventions in the treatment of adolescent depression. While the results of statistics are important as an indicator of the effectiveness of an intervention, they do not necessarily show whether the interventions bring any clinically significant effects to the functioning of adolescents in their everyday lives [19, 28]. In the included studies, several dimensions were found to be critical for assessing intervention efficacy, such as statistical efficacy, clinical efficacy, functional benefits, and feasibility in practical settings. These dimensions, however, rarely aligned, indicating that efficacy demonstrated in one dimension does not always translate into efficacy in other dimensions. A comparable re-evaluation has occurred in the broader adult psychotherapy literature, where accounting for publication bias and study quality reduced earlier pooled estimates of treatment benefit [29].

In order to facilitate a more comprehensive interpretation of evidence, the Multidimensional Clinical Interpretation Framework (MCIF) was developed, which is a seven-dimensional framework incorporating statistics, clinical, psychometrics, functional, and implementation dimensions. The purpose of this framework is

to help both researchers and clinicians interpret evidence related to digital interventions for depression in adolescents.

Positioning the MCIF in Relation to Other Methods

Given the searches we have undertaken in the four databases (PubMed, PsycINFO, Scopus and Web of science) it seems clear that there is no other framework that incorporates statistical inference, effect size, accuracy, clinical significance, reliable improvement, functional recovery, and implementation relevance in one framework to be used for evaluating the efficacy of digital interventions in adolescent depression. While there may be other well-established frameworks that deal with some of these issues, the problem lies in the fact that they deal with specific aspects of evidence gathering, reporting, or interpretation in isolation from each other.

It is important to note that the MCIF is not meant to compete with any of these well-established methods but rather complement them by combining elements of evidence evaluation that are usually considered in isolation from each other in separate frameworks. The GRADE approach focuses on the quality of the evidence available in order to generate guidelines [30].

In the same way, CONSORT offers recommendations on reporting of RCTs, promoting transparency via comprehensive reporting of methods and results [31]. It does not offer an approach for interpretation of the clinical, psychometric, functional, and implementation significance of findings [31]. The Reliable Change Index proposed by Jacobson and Truax offers a useful tool for the estimation of individual change [28]. Nevertheless, it is focused mainly on the estimation of reliable improvement, but not on integration of several aspects concerning intervention effectiveness. Related methodological work has also emphasized that trials framed around non-inferiority or equivalence, rather than superiority, require dedicated statistical approaches and should not be interpreted using conventional significance-testing logic [32].

These approaches play different roles in evidence generation, reporting and interpretation. The MCIF attempts to unite all these perspectives within one framework, allowing statistical results, clinical effects, functional recovery, and implementation aspects to be considered together during evaluation of adolescent digital depression interventions. Selection of these dimensions has been guided by previous literature on effectiveness, usability, and engagement of interventions [6, 24].

Due to the heterogeneous methodology of adolescent digital mental health literature and its tendency to focus on symptoms-based outcome measures, multidimensional framework for interpretation based on previous evidence about intervention effectiveness, usability, and engagement [6, 19, 23, 24] may promote a better evaluation of intervention benefits [5, 28]. This divergence between statistical and clinically meaningful outcomes is not unique to depression; comparable gaps between absolute and relative treatment effects have been documented across multiple psychiatric disorders [33].

Clinical and Research Implications

The identified trends in statistical–clinical discordance in studies on digital depression in adolescents have significant implications for interpreting clinical evidence, synthesizing research results, and presenting them in publications. For example, meta-analytic findings of the small but statistically significant effect (e.g., $d \approx 0.33$, as in Garrido et al. [25]), pooled using standard meta-analytic synthesis methods [34], might be used as the proof of the effectiveness of interventions, although the size of the effect could be lower than those that can be

considered clinically meaningful in terms of alleviation of symptoms; more precise approaches to calculating clinical significance have been proposed to help address this gap [35]. This problem might be especially topical in digital interventions because of high attrition rates and variation in user engagement, which affect intent-to-treat analyses. At the same time, pilot or feasibility studies might yield non-significant results despite the potential significance of changes in the group under investigation (in case of a small sample size). As a consequence, promising interventions could be deemed as ineffective and discontinued [15, 16]. On the methodology front, the MCIF reporting scheme could assist in a more complete interpretation of adolescent digital mental health studies through the encouragement of the evaluation of several aspects of the intervention efficacy. In particular, reporting of covariance parameters in repeated measures analysis can aid in secondary analysis, allow more accurate estimates of the effect on an individual level and facilitate the conversion of group results into clinically significant probabilities of improvement [25, 35]. Also, additional information on adherence to the therapy, engagement patterns and attrition could offer additional insights into the real-life application of digital therapies [24].

Strengths

This review advances the current methodological knowledge about adolescent digital depression intervention research. Firstly, the review combines seven distinct domains from various fields of investigation into an organized and comprehensible structure for understanding the results of intervention. Secondly, the concept of MCIF is thoroughly defined and its contribution to the existing approaches, such as GRADE, CONSORT, and the Jacobson–Truax Reliable Change Index, is discussed. Thirdly, this review presents a four-category classification of discordance for describing discrepancies among various aspects of intervention results. Fourthly, this review makes suggestions concerning reporting limitations encountered in adolescent digital depression studies. Building on previous findings regarding limited reporting of cost-effectiveness and real-world outcomes in digital mental health interventions for young people [5], researchers are encouraged to provide a broader representation of intervention results by taking into consideration statistical, clinical, functional, and implementation aspects of outcomes.

Limitations

There are several issues that need to be considered when evaluating the results of the review. Firstly, the choice of a narrative methodology, despite its appropriateness for creating and improving conceptual frameworks, makes any kind of quantitative analysis impossible. Furthermore, the current review was not registered in PROSPERO, as it is usually done with systematic reviews, but this decision was made following the planned methodology of the study. Additionally, the discordance patterns in Table 3 are illustrated using 10 representative sources; although this number is modest in isolation, 9 of these 10 sources are themselves systematic reviews or meta-analyses that collectively synthesize evidence from dozens of primary trials, so the patterns identified draw on a considerably larger underlying evidence base than the source count alone suggests. Relatedly, because most of these source's report pooled, meta-analytic estimates rather than single-trial findings, the discordance identified at the level of a pooled effect size is not always equivalent to discordance within any individual trial; where possible, this distinction is noted alongside each example in Table 3.

Secondly, the inter-rater reliability of domain coding has not been assessed using statistical methods, such as Cohen’s kappa. This factor can influence the reliability of the classification. Any future applications of the MCIF

need to use the methodology of assessing the reliability of the classification.

Thirdly, the MCIF is an innovative instrument that has not been validated empirically by comparison with any external criterion (interpretation or decision-making validity), hence it can currently only be viewed as conceptual and further validation studies should be carried out.

Lastly, differences in the availability of benchmark values for MCID for the assessment tools used to measure adolescent depression such as PHQ-A and SMFQ may have impacted the clinical significance determination. The application of the MCIF for other adolescent mental disorders will likely need calibration to specific conditions.

Future Directions

Future research should analyze the empirical efficacy and interpretability of the MCIF by conducting meta-research through large-scale IPD analysis of existing adolescent digital depression studies. This approach would contribute toward getting a clearer picture of how valid, applicable, and valuable the framework would be for clinical interpretation purposes.

It is necessary to conduct research to understand the relationship between the domains of the MCIF in different studies. The application of advanced statistical methodologies, such as Bayesian modeling [39], might enable to measure uncertainty and robustness of the interpretations made within each particular domain of the framework and, at the same time, to estimate its performance more accurately. Machine learning methods could also become a way to explore the heterogeneity of treatment effects and to define particular subgroups of adolescents benefiting from digital interventions. Nonetheless, such an approach must be validated properly to make the results clinically meaningful and interpretable.

Long-term studies should be conducted to ascertain the durability of benefits in terms of reduction of depressive symptoms and recovery of functioning.

Lastly, the effectiveness of implementing the MCIF is going to rely on cooperation between researchers, trialists, clinicians, and journal editors to ensure standard application and reporting in the research process. It is going to take some time and additional assessment to develop common standards in methodology, but these processes could facilitate better comparison of future research studies and contribute to the framework’s utility.

Proposed Validation Pathway

In order to validate the MCIF as an effective interpretive instrument, a systematic four-stage validation pathway is suggested (Fig. 4). At the first stage, a Delphi expert consultation will have to be organized among researchers specializing in digital mental health care, clinicians, and methodologists to improve the seven MCIF domains, define their structure, and develop initial interpretive thresholds. At the second stage, the reliability of the MCIF should be tested to find out whether the interpretation provided by independent raters using the MCIF is consistent. The inter-rater agreement can be tested through statistical analysis tools, such as Cohen's kappa coefficient. At the third stage, the MCIF should be tested on already conducted randomized control trials dealing with the use of digital interventions to help adolescents suffering from depression. The individual participant data (IPD) should be used wherever possible; adaptive trial designs, which permit pre-specified modifications based on interim data, may also offer a practical way to test MCIF-informed refinements prospectively [40].

At the fourth stage, the MCIF needs to be assessed for the possibility to make predictions based on its use. It means testing the relationship between MCIF domain profiles and clinical outcomes and functional recovery. Thus, the suggested four-stage validation path will help to ensure the MCIF validity.

Conclusion

Statistical significance and p-values alone provide an incomplete basis for evaluating treatment effectiveness in adolescent digital depression research. Although these measures offer important information about intervention effects, they do not fully capture effect magnitude, clinical relevance, functional recovery, or implementation feasibility. Consequently, statistically significant findings may not always translate into meaningful clinical benefits, while potentially valuable improvements may remain undetected due to limited statistical power, participant attrition, or measurement constraints commonly observed in digital intervention trials.[14,15,41]

To address these challenges, we developed the Multidimensional Clinical Interpretation Framework (MCIF), a seven-domain sequential framework that integrates statistical, clinical, psychometric, functional, and implementation perspectives to support a more comprehensive interpretation of treatment outcomes. The MCIF is not intended to replace established approaches such as GRADE, CONSORT, or the Jacobson–Truax Reliable Change Index; rather, it provides an additional interpretive layer that facilitates the integration of diverse sources of evidence when evaluating individual studies.

Based on structured searches of PubMed, PsycINFO, Scopus and Web of science we did not identify an existing framework that integrates all seven MCIF domains within a single model specifically for adolescent digital depression research. The MCIF therefore represents a potential contribution toward addressing an important gap in the organization and interpretation of evidence in this rapidly evolving field.

Declarations

Data Availability Statement

This is a narrative methodological review with conceptual framework development. No original datasets were generated or analyzed. All cited studies are publicly available through the referenced databases. Full domain-level coding of included sources is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest relevant to this work. No external funding was received for this review.

Declarations

In accordance with the journal's guidelines on AI-assisted writing, during the preparation of this manuscript, the authors used ChatGPT (Version 5-5) to assist with [1]translation, [2]paraphrasing, [3]editing, and [4]improving the language, [5]grammar, [6]clarity, and [7]readability of the manuscript, the primary draft of which was written by the authors. All AI-assisted content was critically reviewed, verified, and revised by the authors, who take full responsibility for the accuracy, originality, and integrity of the manuscript.

Ethics approval and consent to participate

Not applicable. This article is a narrative review and does not report on or involve the use of any human participants, human data, human tissue, or animals.

Consent to participate

Not applicable.

Consent for publication

Not applicable. This article does not contain any individual person's data in any form (including individual details, images, or videos).

Clinical trial number: not applicable

References

1. Lewinsohn PM, Rohde P, Seeley JR. Major depressive disorder in older adolescents: prevalence, risk factors, and clinical implications. Clin Psychol Rev. 1998;18(7):765-94. doi:10.1016/S0272-7358(98)00010-5
2. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet. 2018;392(10157):1553-98. doi:10.1016/S0140-6736(18)31612-X
3. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther. 2017;88:7-18. doi:10.1016/j.brat.2016.06.004
4. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual research review: digital health interventions for children and young people with mental health problems – a systematic and meta-review. J Child Psychol Psychiatry. 2017;58(4):474-503. doi:10.1111/jcpp.12663
5. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health. 2021;8(4): e25847. doi:10.2196/25847
6. Torous J, Bucci S, Bell IH, Kueider-Paisley A, Firth J, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry. 2021;20(3):318-35. doi:10.1002/wps.20883
7. Erbe D, Eichert HC, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res. 2017;19(9):e306. doi:10.2196/jmir.6588
8. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
9. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 2013; 4:863. doi:10.3389/fpsyg.2013.00863
10. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando: Academic Press; 1985.

11. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry. 2017;16(3):287-98. doi:10.1002/wps.20472
12. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry. 2019;18(3):325-36. doi:10.1002/wps.20673
13. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-50. doi:10.1007/s10654-016-0149-3
14. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “p < 0.05”. Am Stat. 2019;73(sup1):1-19. doi:10.1080/00031305.2019.1583913
15. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76. doi:10.1038/nrn3475
16. Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. doi:10.1136/bmj.311.7003.485
17. Cumming G. The new statistics: why and how. Psychol Sci. 2014;25(1):7-29. doi:10.1177/0956797613504966
18. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
19. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry. 2006;59(11):990-6. doi: 10.1016/j.biopsych.2005.09.014
20. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane handbook for systematic reviews of interventions. 2nd ed. Chichester: Wiley; 2019.
21. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019; 4:5. doi:10.1186/s41073-019-0064-8
22. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry. 2021;78(4):361-71. doi:10.1001/jamapsychiatry.2020.4364
23. Liverpool S, Mota CP, Sales CMD, Čuš A, Carletto S, Hancheva C, et al. engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res. 2020;22(6): e16317. doi:10.2196/16317
24. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res. 2018;20(6): e199. doi:10.2196/jmir.9275
25. Garrido S, Millington C, Cheers D, Boydell K, Schubert E, Meade T, et al. What works and what doesn’t work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry. 2019; 10:759. doi:10.3389/fpsyt.2019.00759
26. Merry SN, Stasiak K, Shepherd M, Frampton C, Fleming T, Lucassen MFG. The effectiveness of SPARX, a computerised self-help intervention for adolescents seeking help for depression: randomised controlled

non-inferiority trial. BMJ. 2012;344: e2598. doi:10.1136/bmj. e2598

27. Ebert DD, Zarski AC, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One. 2015;10(3):e0119895. doi: 10.1371/journal.pone.0119895
28. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991;59(1):12-9. doi:10.1037/0022-006X.59.1.12
29. Cuijpers P, Karyotaki E, Reijnders M, Ebert DD. Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. Epidemiol Psychiatr Sci. 2019;28(1):21-30. doi:10.1017/S2045796018000057
30. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924-6. doi:10.1136/bmj.39489.470347.AD
31. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869. doi:10.1136/bmj.c869
32. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Adv Methods Pract Psychol Sci. 2018;1(2):259-69. doi:10.1177/2515245918770963
33. Cuijpers P, Miguel C, Ciharova M, Harrer M, Karyotaki E. Absolute and relative outcomes of psychotherapies for eight mental disorders: a systematic review and meta-analysis. World Psychiatry. 2024;23(2):267-75. doi:10.1002/wps.21203
34. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2009.
35. Kendall PC, Marrs-Garcia A, Nath SR, Sheldrick RC. Normative comparisons for the evaluation of clinical significance. J Consult Clin Psychol. 1999;67(3):285-99. doi:10.1037/0022-006X.67.3.285
36. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41(5):582-92. doi: 10.1097/01.MLR.0000062554.74615.4C
37. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychol Methods. 2002;7(1):105-25. doi:10.1037/1082-989X.7.1.105
38. Richardson LP, McCauley E, Grossman DC, McCarty CA, Richards J, Russo JE, et al. Evaluation of the Patient Health Questionnaire-9 Item for detecting major depression among adolescents. Pediatrics. 2010;126(6):1117-23. doi:10.1542/peds.2010-0852
39. Kruschke JK, Liddell TM. The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev. 2018;25(1):178-206. doi:10.3758/s13423-016-1221-4
40. Bhatt DL, Mehta C. Adaptive designs for clinical trials. N Engl J Med. 2016;375(1):65-74. doi:10.1056/NEJMra1510061
41. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129-33. doi:10.1080/00031305.2016.1154108

42. Buttazzoni A, Brar K, Minaker L. Smartphone-based interventions and internalizing disorders in youth: systematic review and meta-analysis. J Med Internet Res. 2021;23(1):e16490. doi:10.2196/16490

Figures

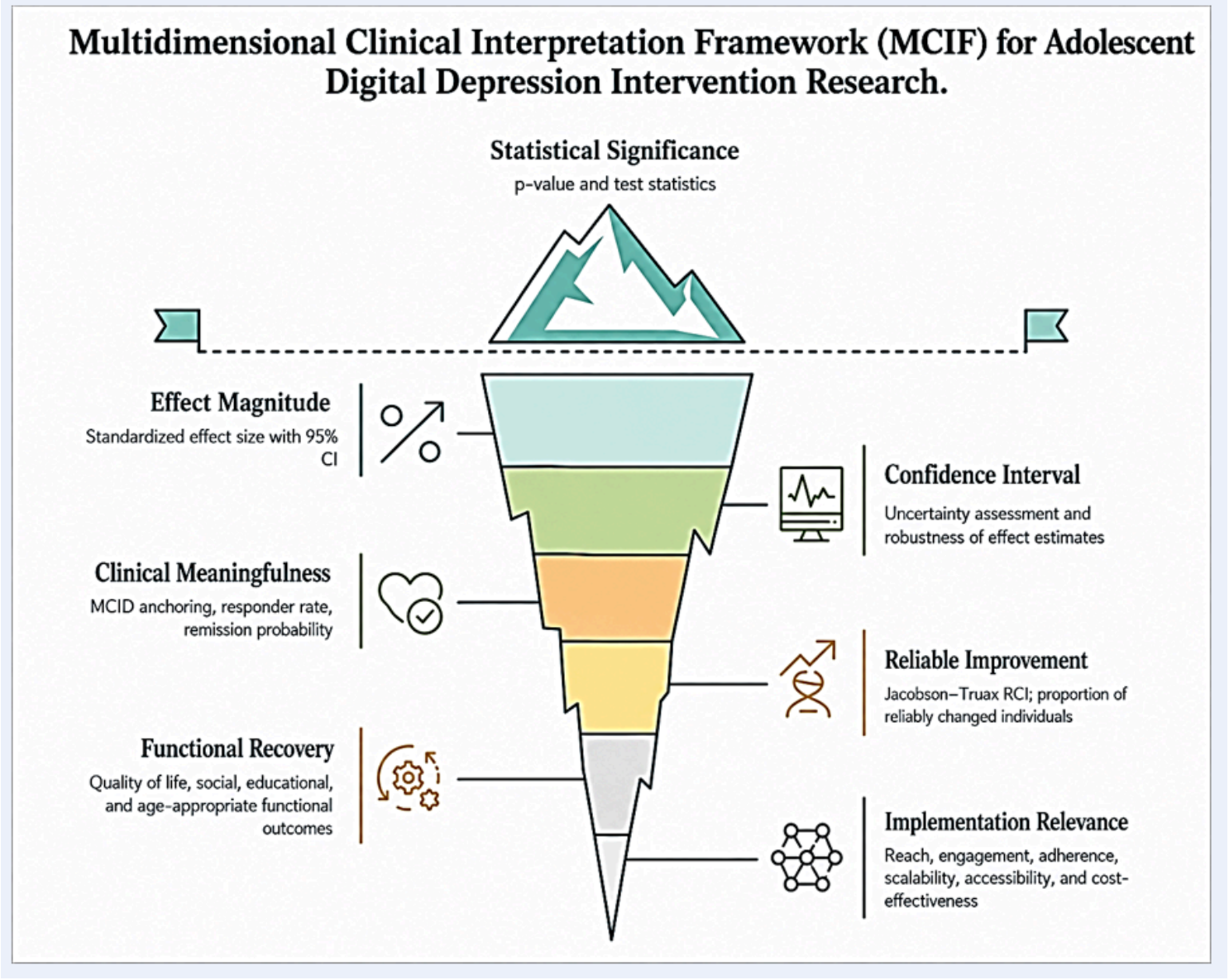


Figure 1

Conceptual Structure of the MCIF for Adolescent Digital Depression Intervention Research

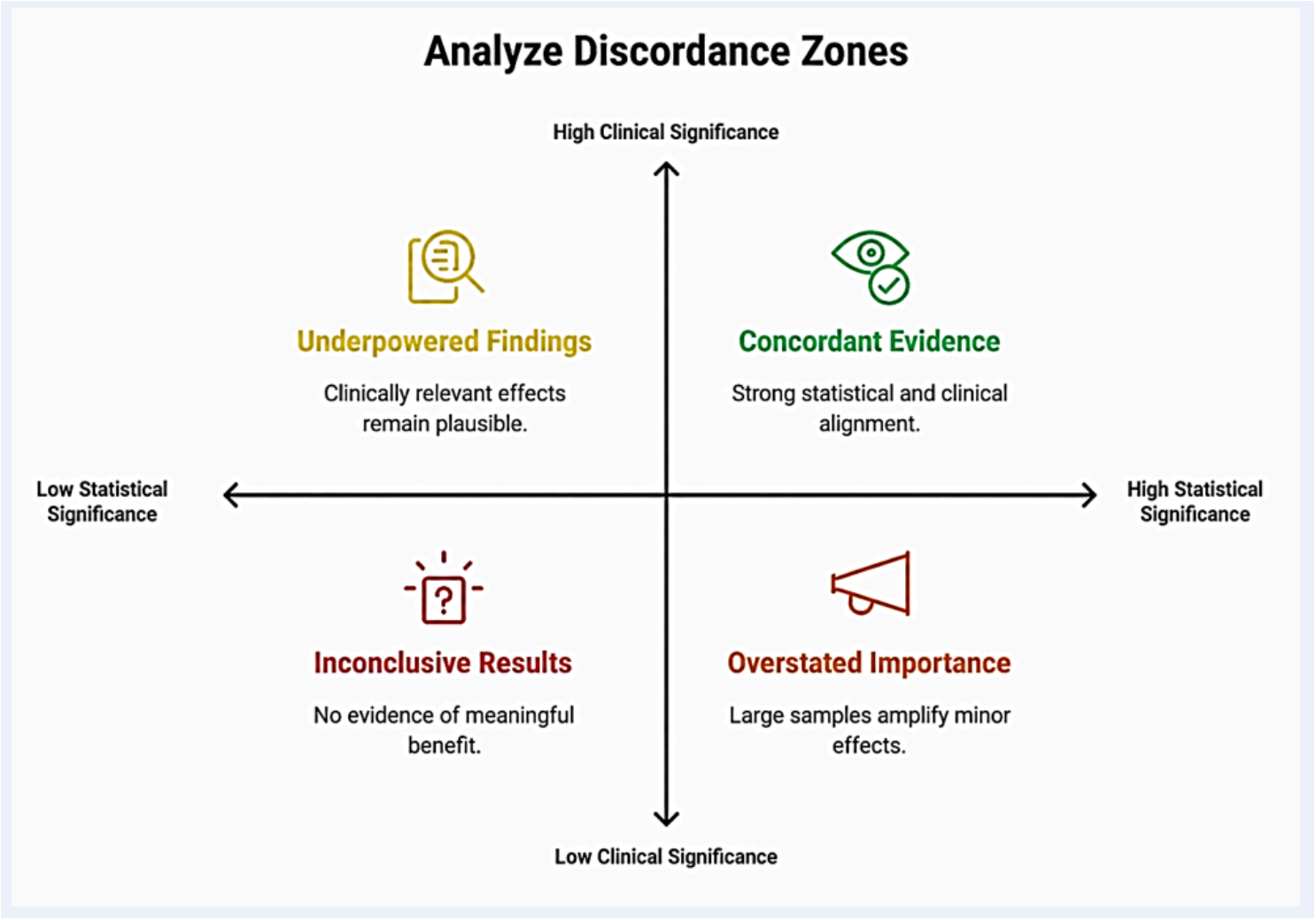


Figure 2

Statistical–Clinical Discordance Zones in Adolescent Digital Depression Research

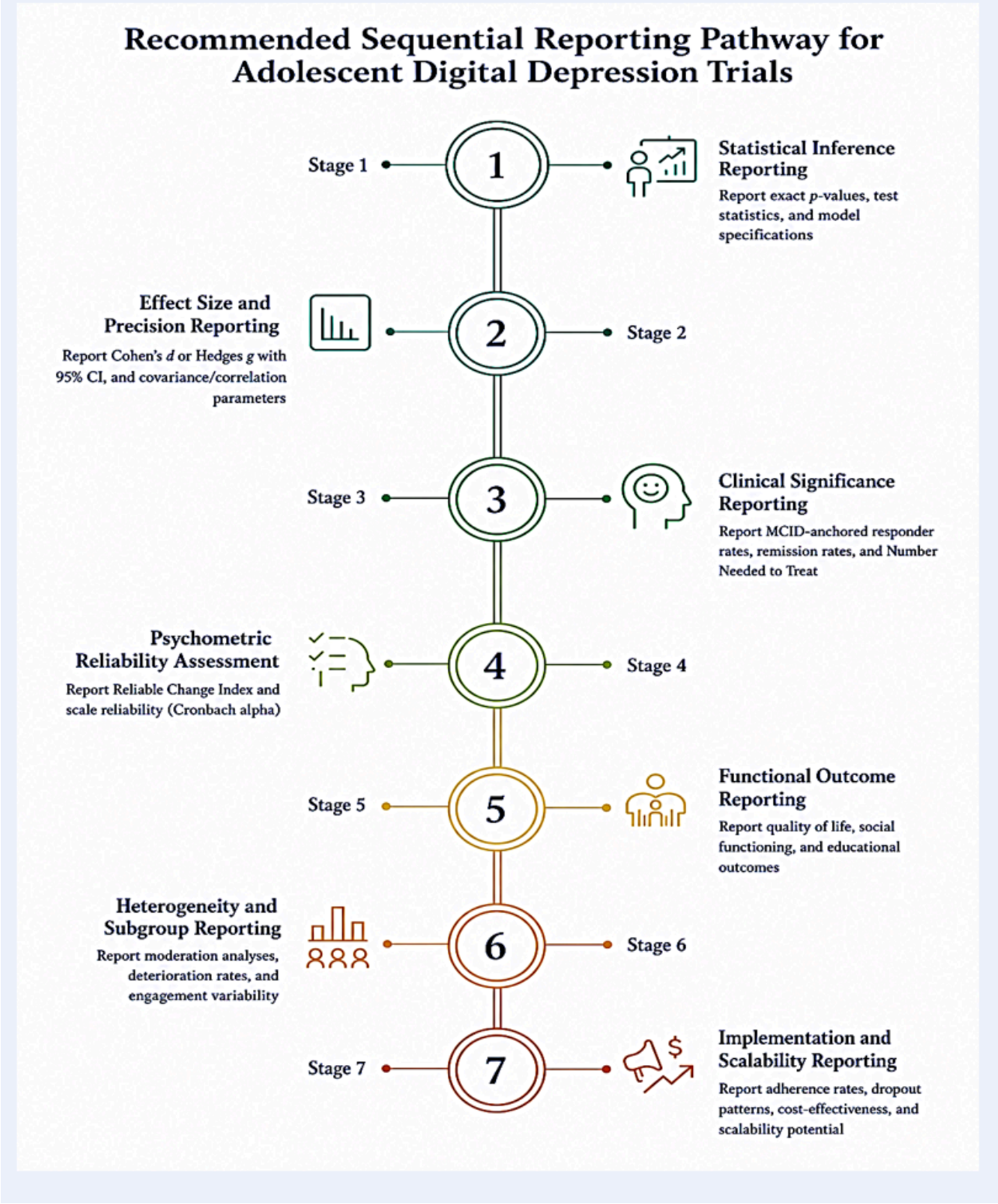


Figure 3

Recommended Sequential Reporting Pathway for Adolescent Digital Depression Trials

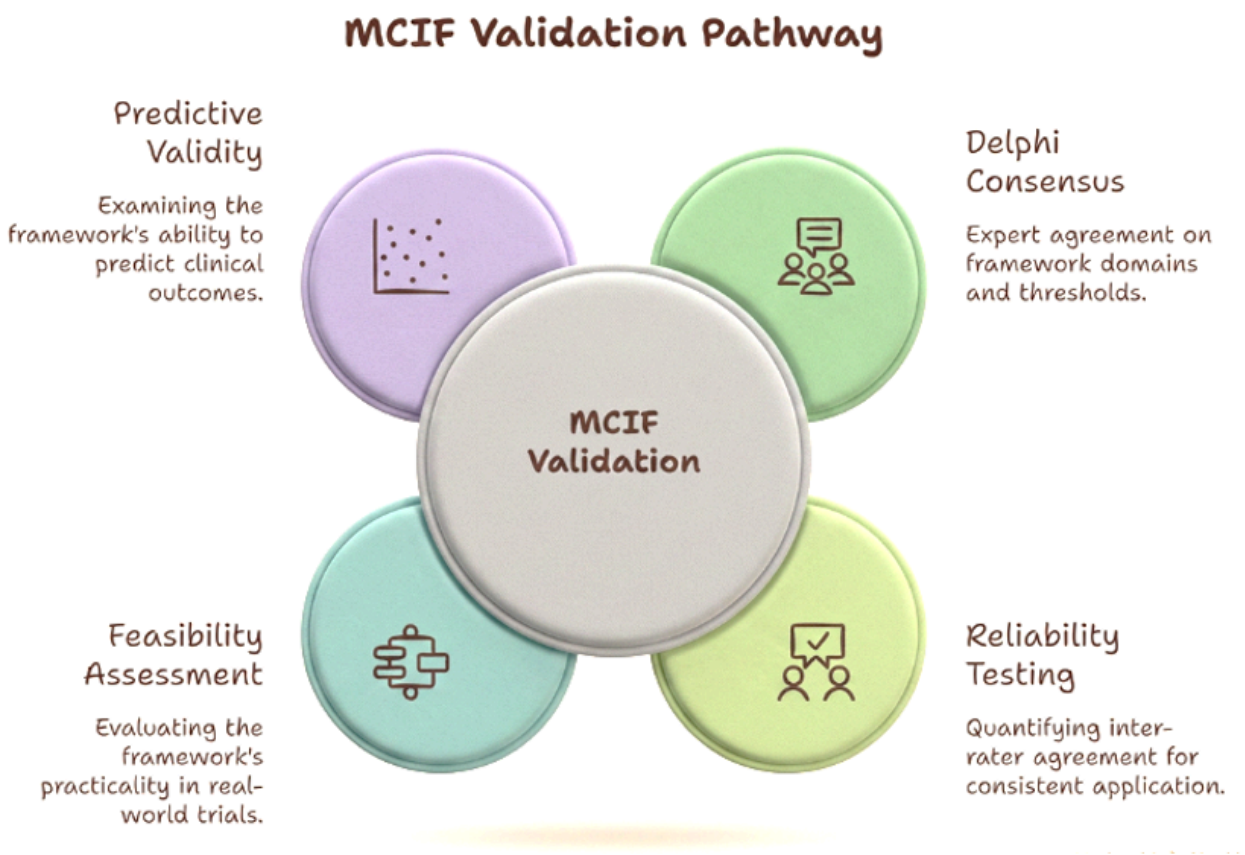


Figure 4

Proposed Validation Pathway for the Multidimensional Clinical Interpretation Framework (MCIF). Sequential stages for future validation of the framework, including Delphi consensus, inter-rater reliability testing, application to existing trials, and predictive validity assessment.