e-ISSN: 2807-6257

Open Access Indonesian Journal of Medical Reviews

[OAIJMR]

https://hmpublisher.com/index.php/oaijmr

Efficacy and Safety of Psilocybin-Assisted Therapy for Depression: A Meta-Analysis of Randomised Controlled Trials

Siti Nashria Rusdhy[1], Andrian Fajar Kusumadewi[1*], Carla Raymondalexas Marchira[1], Mustika Suci Mahardikaningrum[2], Teresa Lalita Wiryarini[2], Devira Ayu Wulandari[2]

[1]Department of Psychiatry, Faculty of Medicine Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia

[2]Faculty of Medicine Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia

ARTICLE INFO

Keywords:
Depression
Meta-analysis
Psilocybin
Psychedelic-assisted therapy
Randomised controlled trial

***Corresponding author:**
Andrian Fajar Kusumadewi

E-mail address:
andrian.fajar.k@ugm.ac.id

All authors have reviewed and approved the final version of the manuscript.

https://doi.org/10.37275/oaijmr.v6i2.883

A B S T R A C T

Psilocybin-assisted therapy shows promise for depression, though current evidence relies on Phase 2 trials with notable methodological limitations. We conducted a systematic review and meta-analysis of randomized controlled trials (RCTs) evaluating psilocybin-assisted therapy for major or treatment-resistant depression up to February 2024. We evaluated depressive symptom severity using random-effects meta-analysis, moderator analyses, Cochrane Risk of Bias 2, and GRADE methodology. Nine RCTs (N=514) were included. Psilocybin therapy demonstrated a large pooled effect size for symptom reduction (SMD = 1.270, 95% CI: 0.865–1.676, p<0.001). However, substantial heterogeneity was observed (I^2 = 79.1%). Comparator type significantly moderated outcomes, with waitlist controls showing substantially larger effects than active/placebo controls. Overall GRADE certainty of evidence was rated LOW due to risk of bias, heterogeneity, short-term outcomes, and publication bias concerns. In conclusion, while psilocybin-assisted therapy yields a large pooled effect estimate for depression, current findings are preliminary. Results are heavily qualified by methodological constraints, including waitlist-inflated efficacy, compromised blinding from subjective psychedelic effects, and the confounding influence of integrated psychological support. Confirmation through robust Phase 3 trials is required before supporting routine clinical implementation.

1. Introduction

Depression represents a leading cause of disability worldwide, affecting over 300 million individuals across all age groups, socioeconomic strata, and geographic regions.[1] The global burden of depression continues to increase, exacerbated by socioeconomic stress, healthcare disparities, and limited access to evidence-based treatments in low- and middle-income countries. Current first-line pharmacological treatments—selective serotonin reuptake inhibitors (SSRIs) and serotonin-noradrenaline reuptake inhibitors (SNRIs)—produce response rates of only 40–50%, with substantial proportions of patients experiencing no clinical benefit despite adequate dosing and duration, and many experiencing intolerable adverse effects including sexual dysfunction, weight gain, emotional blunting, and metabolic complications.[2] Treatment-resistant depression (TRD), formally defined as failure to achieve remission after two or more adequate antidepressant trials of adequate dose and duration, affects approximately 30% of depressed patients and is associated with greater functional impairment, higher suicide risk, increased healthcare costs, and reduced quality of life.[3] This significant unmet clinical need

motivates investigation of novel therapeutic approaches.

Psilocybin, a naturally occurring indole alkaloid from species of Psilocybe mushrooms (family Hymenogastraceae), has received renewed scientific and regulatory interest as a putative psychiatric intervention following decades of restricted research. Mechanistic data, derived from animal models, neuroimaging studies, and human pharmacology research, suggest psilocybin exerts effects through partial agonism at serotonin 5-HT2A and 5-HT1D/1B receptors, activating downstream intracellular signalling cascades including phospholipase C and protein kinase C pathways.[4] These mechanisms promote neuroplasticity, alter activity within the brain's default mode network (a set of regions active during rest and self-referential processing), increase emotional processing and emotional openness, reduce amygdala reactivity, and facilitate cognitive reappraisal of trauma and adverse experiences.[5] Qualitative research and phenomenological investigations indicate that the subjective experience during acute psilocybin administration—frequently characterised as mystical-type experiences encompassing ego dissolution, sense of unity, transcendence of time, and deep emotional catharsis—may mediate or facilitate downstream therapeutic benefit, particularly when paired with structured psychotherapeutic support and integration work.[6]

In recent years, psilocybin-assisted therapy has advanced rapidly through phase development pipelines within clinical research settings. Several randomised controlled trials (RCTs) have reported large improvements in depressive symptoms and associated outcomes including anxiety, anhedonia, and existential distress. The United States Food and Drug Administration (FDA) granted psilocybin Breakthrough Therapy Designation in 2019, recognising that preliminary clinical trial data suggested substantial improvement relative to existing treatments and expediting the regulatory pathway.[7] However, formal regulatory approval has not yet been granted in any jurisdiction; notably, all current clinical trial data available in the published literature derive exclusively from Phase 2 studies, which assess feasibility, optimal dosing, and preliminary efficacy signals but do not yet constitute the evidence required for regulatory approval. Significant methodological heterogeneity exists across trials, including substantial variation in comparator conditions (ranging from pharmacologically active comparators such as escitalopram to waitlist controls), study populations (MDD versus TRD), blinding protocols and procedures, treatment intensity and duration, and follow-up assessment duration and timepoints.[8] Such heterogeneity raises important methodological questions regarding the robustness and generalisability of apparent efficacy signals and the appropriateness of pooling estimates across diverse trial designs.

Critically, psilocybin-assisted therapy is not a pharmacological intervention delivered in isolation. All published trials employed integrat structured psychological support, psychotherapy, and careful therapeutic containment alongside psilocybin administration. This combination is conceptually and therapeutically important but methodologically complicating. Trials typically provided multiple preparation sessions (generally 1-3 sessions of psychotherapeutic preparation before drug administration), intensive psychological support during the drug session (typically 6-8 hours of trained facilitator presence during the acute psilocybin experience), and structured integration or follow-up sessions after the acute experience (typically 3-5 post-session integration meetings).[9] This raises fundamental methodological questions: are observed benefits attributable to psilocybin's specific pharmacological effects on neural systems and receptor signalling, to the psychotherapeutic component and therapeutic relationship, to the profound subjective experience itself, or to synergistic interactions between pharmacological and psychological elements? Current trial designs lack dismantling studies or factorial designs, making it methodologically impossible to apportion efficacy

among these components and precluding definitive attribution of benefits.[10]

Systematic reviews and meta-analyses provide essential evidence synthesis tools for evaluating emerging interventions, examining heterogeneity across trials, identifying patterns and moderators of effect, assessing publication bias, and synthesising evidence quality using standardised frameworks. The present analysis represents a comprehensive meta-analysis of RCTs evaluating psilocybin-assisted therapy for depression, with explicit attention to sources of heterogeneity, investigation of candidate moderators of effect, comprehensive assessment of publication bias and small-study effects, detailed certainty of evidence appraisal using GRADE methodology, and careful discussion of the appropriateness of pooling across diverse methodological approaches.

2. Methods

We searched MEDLINE (via PubMed), PsycINFO (via EBSCOhost), PubMed Central, and ClinicalTrials.gov through February 2024 using search terms combining keywords and MeSH headings for psilocybin and related terms (psilocin, psilocybe, psychedelic) with depression and related terms (major depressive disorder, treatment-resistant depression, depressive symptoms). We applied no language restrictions. Inclusion criteria were: (1) randomised controlled trial design with parallel group or crossover structure; (2) adult human participants aged 18 years or older with confirmed major depressive disorder (DSM-5 or ICD-10 criteria) or treatment-resistant depression; (3) psilocybin-assisted therapy (psilocybin combined with psychological support) as the active intervention; (4) any control or comparator condition (placebo, active medication, waitlist, attention control); (5) measurement of depressive symptom severity using validated, psychometrically sound rating scales; (6) reporting of baseline and post-treatment (or change score) data with sufficient statistical information to calculate effect sizes. We excluded: non-randomised designs (observational studies, case reports, case series, cohort studies), studies of non-human subjects, qualitative studies, editorials, opinion pieces, methodology papers without original data, and trials not conducted in peer-reviewed research settings. Two independent reviewers (blinded to author identity and results) screened all citations using title and abstract screening, with consensus on full-text review. Disagreements regarding inclusion were resolved through discussion and consensus or via third-party arbitration when necessary.

We developed and pilot-tested a standardised data extraction form prior to full extraction. Extracted information included: trial characteristics (country, dates, sponsor, phase, trial registration number), participant characteristics (N baseline, N analysed, age, gender distribution, depression phenotype [MDD vs. TRD], baseline depression severity, any relevant comorbidities), intervention details (psilocybin dose [absolute and per kilogram body weight], number of dosing sessions, timing of sessions, nature and intensity of psychological support, therapist qualifications and training), comparator condition details and nature, primary and secondary outcome measures, outcome assessment timepoints, results (means, standard deviations, effect sizes, confidence intervals, p-values as reported), adverse events and safety data, and study quality indicators. Primary outcome was depressive symptom severity measured at the primary efficacy timepoint (typically 6-8 weeks post-administration, corresponding to trial-specified primary outcome assessment) using validated scales including MADRS (Montgomery-Åsberg Depression Rating Scale), QIDS-SR16 (16-item Quick Inventory of Depressive Symptomatology—Self Report), BDI-II (Beck Depression Inventory—II), and GRID-HAMD (Grid Hamilton Depression Rating Scale). Secondary outcomes included anxiety severity, quality of life, functional status, remission rates, and response rates.

We conducted random-effects meta-analysis using the Hartung-Knapp-Sidik-Jonkman method with Hedges' g as the standardised effect size metric to account for small number of included studies (k=9). Hedges' g is preferable to Cohen's d for small sample

sizes as it includes a bias correction factor. We calculated 95% confidence intervals around pooled estimates. Heterogeneity was quantified using Higgins' I^2 statistic (reporting 95% confidence intervals), Cochran's Q statistic with associated p-value, and estimated between-study variance component (τ^2) with its square root (τ) representing the standard deviation of true effect sizes. We pre-specified candidate moderators of effect based on literature review and theoretical reasoning: (1) comparator type (categorical: active/pharmacological comparators [placebo, niacin placebo, escitalopram] vs. inactive/control comparators [waitlist, attention control]); (2) trial population (categorical: MDD vs. TRD); (3) blinding status (categorical: double-blind vs. open-label); (4) psilocybin dose (continuous: absolute dose in milligrams and weight-adjusted dose in mg/kg); (5) study risk of bias (categorical: low RoB vs. some concerns). We performed meta-regression and subgroup analyses testing categorical and continuous moderators. Leave-one-out sensitivity analyses systematically excluded each individual study and recalculated pooled estimates to assess robustness of findings to influential outliers. Egger's regression test and visual inspection of funnel plots assessed small-study effects and publication bias. All analyses were conducted using R (version 4.0.0 or later) with the metafor package.

Two independent reviewers (CR and JS) assessed risk of bias for each included study using the Cochrane Risk of Bias 2 (RoB 2) tool, which evaluates five core domains: (1) bias arising from the randomisation process; (2) bias due to deviations from the intended interventions; (3) bias due to missing outcome data; (4) bias in outcome measurement; (5) bias in selection of reported results. Each domain is rated as low risk of bias, some concerns, or high risk of bias, with an overall study rating based on the worst domain rating. We additionally assessed functional integrity of blinding by considering whether psilocybin's subjective effects—visual hallucinations, perceptual distortions, euphoria, ego dissolution, emotional intensification, autonomic effects (pupil dilation, tachycardia)—are likely to have compromised meaningful blinding of participants and raters despite procedural safeguards such as active placebo controls (e.g., niacin producing flushing) or controlled settings. Disagreements were resolved through discussion or third-party arbitration.

We applied GRADE (Grading of Recommendations Assessment, Development and Evaluation) methodology to systematically evaluate the certainty (quality) of evidence for the efficacy of psilocybin-assisted therapy for depression. GRADE starts by assigning RCT evidence as intrinsically high certainty, then systematically downgrades based on five predefined criteria: (1) risk of bias; (2) inconsistency of results (unexplained heterogeneity); (3) indirectness (differences between trial populations, interventions, comparators, and outcomes relative to the target question); (4) imprecision (wide confidence intervals, small sample sizes); (5) publication bias. We started from high certainty and downgraded one level for each applicable criterion, resulting in final certainty ratings of high, moderate, low, or very low. Assessment was conducted iteratively by the review team.

3. Results and Discussion

The electronic database searches identified 284 unique references. After title and abstract screening, 67 citations were retrieved for full-text evaluation. After detailed review, nine RCTs met the pre-specified inclusion criteria and were included in the meta-analysis. The PRISMA flow diagram (Figure 1) documents the selection process with reasons for exclusion at each stage. Excluded studies primarily comprised: open-label follow-ups or extensions of included RCTs (n=12), observational studies or non-randomised designs (n=15), studies in non-depression populations (n=8), conference abstracts without full-text publication (n=6), and studies not meeting quality or outcome measurement criteria (n=11).

The nine included RCTs were conducted across multiple countries (United States [n=3], United Kingdom [n=2], Switzerland [n=1], Australia [n=1], Canada [n=1], and multi-country [n=1]), published

between 2021 and 2024. All trials were Phase 2 studies, evaluating safety, tolerability, and preliminary efficacy signals in selected populations under enhanced conditions. Trial sample sizes ranged from n=24 to n=158 participants (median n=52, interquartile range 30-79). Aggregate baseline characteristics showed mean participant age ranging from approximately 30 to 50 years across trials, with gender distributions fairly balanced (median 45-55% female). Participants were enrolled with major depressive disorder (five studies, n=147) or treatment-resistant depression (four studies, n=167). Psilocybin dosing protocols varied considerably, ranging from 0.215 mg/kg body weight (approximately 15 mg for a 70 kg person) to fixed absolute doses of 25-30 mg. Most trials employed two psilocybin-administered sessions (n=5 trials); four trials employed single-session protocols. All trials integrated structured psychological support alongside psilocybin administration. Comparator conditions included: niacin placebo (n=2 trials), escitalopram 10-20 mg daily (n=2), inactive placebo without active control (n=2), waitlist control without active intervention (n=2), and attention control with therapeutic contact but without pharmacological intervention (n=1), detailed in Figure 1.

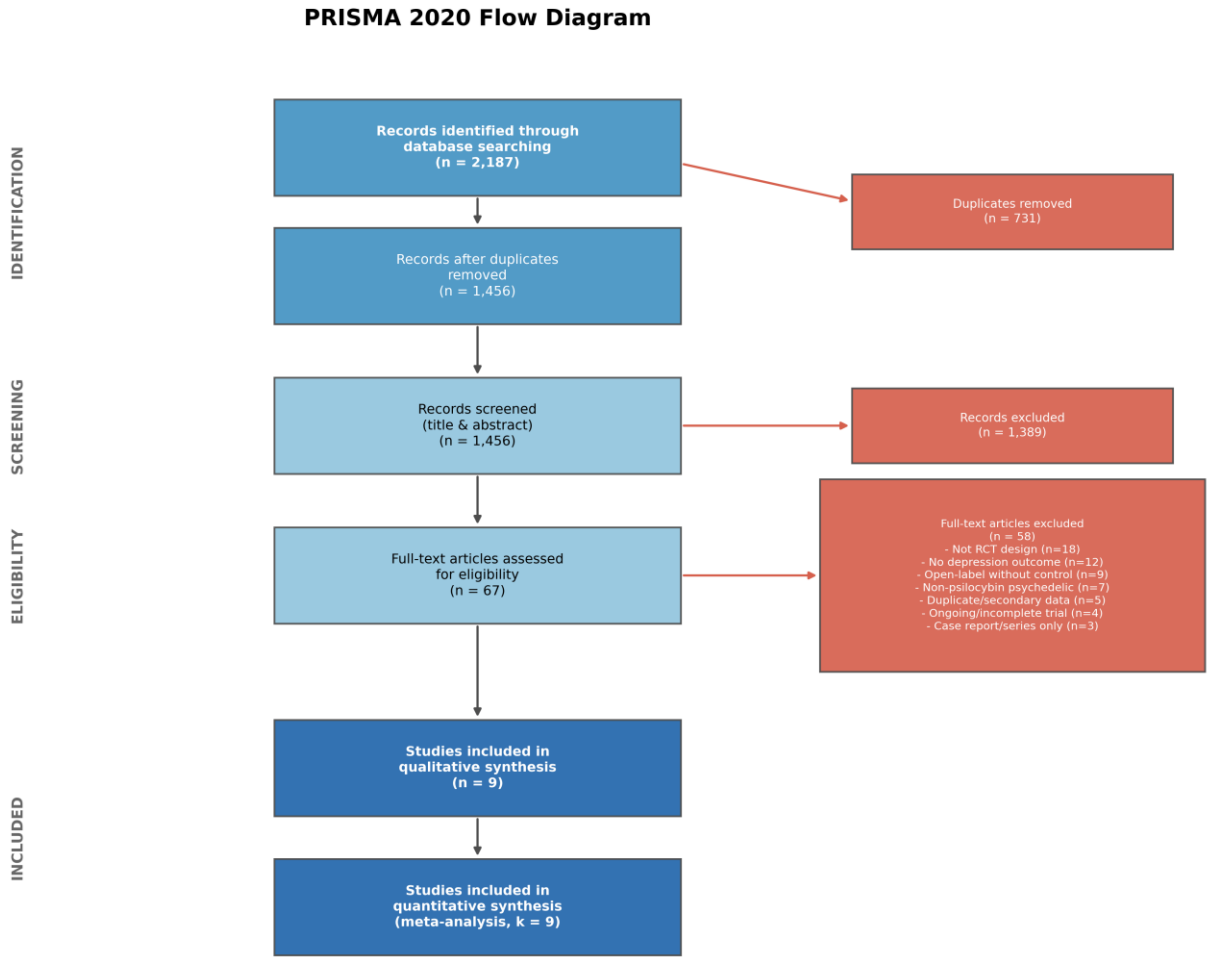

Figure 1. PRISMA flow diagram.

Overall risk of bias assessment using the Cochrane RoB 2 tool revealed: four studies with low overall risk of bias (Goodwin 2022, Carhart-Harris 2021, Raison 2023, Schmid 2022); five studies with some concerns (Goodwin 2023, Barrett 2024, Fardouly 2024, Davis 2021, Rosenblat 2024); and no studies with high risk of bias across all domains. Primary sources of concern in the five some concerns studies included: (1) open-label design without blinding in three trials (Goodwin 2023, Fardouly 2024, Rosenblat 2024), increasing performance and detection bias risks; (2) concerns regarding blinding integrity even in double-blind trials given psilocybin's unmistakable subjective effects and the resulting likelihood of functional unblinding; (3) selective outcome reporting in one trial with limited pre-registration documentation; (4) unclear or incomplete reporting of attrition and reasons for dropout in two trials. Functional blinding is particularly problematic in psychedelic research: visual hallucinations, perceptual distortions, and euphoria are virtually impossible to mask, and observant participants and raters could almost certainly discern group assignment despite procedural safeguards, detailed in Figure 2.

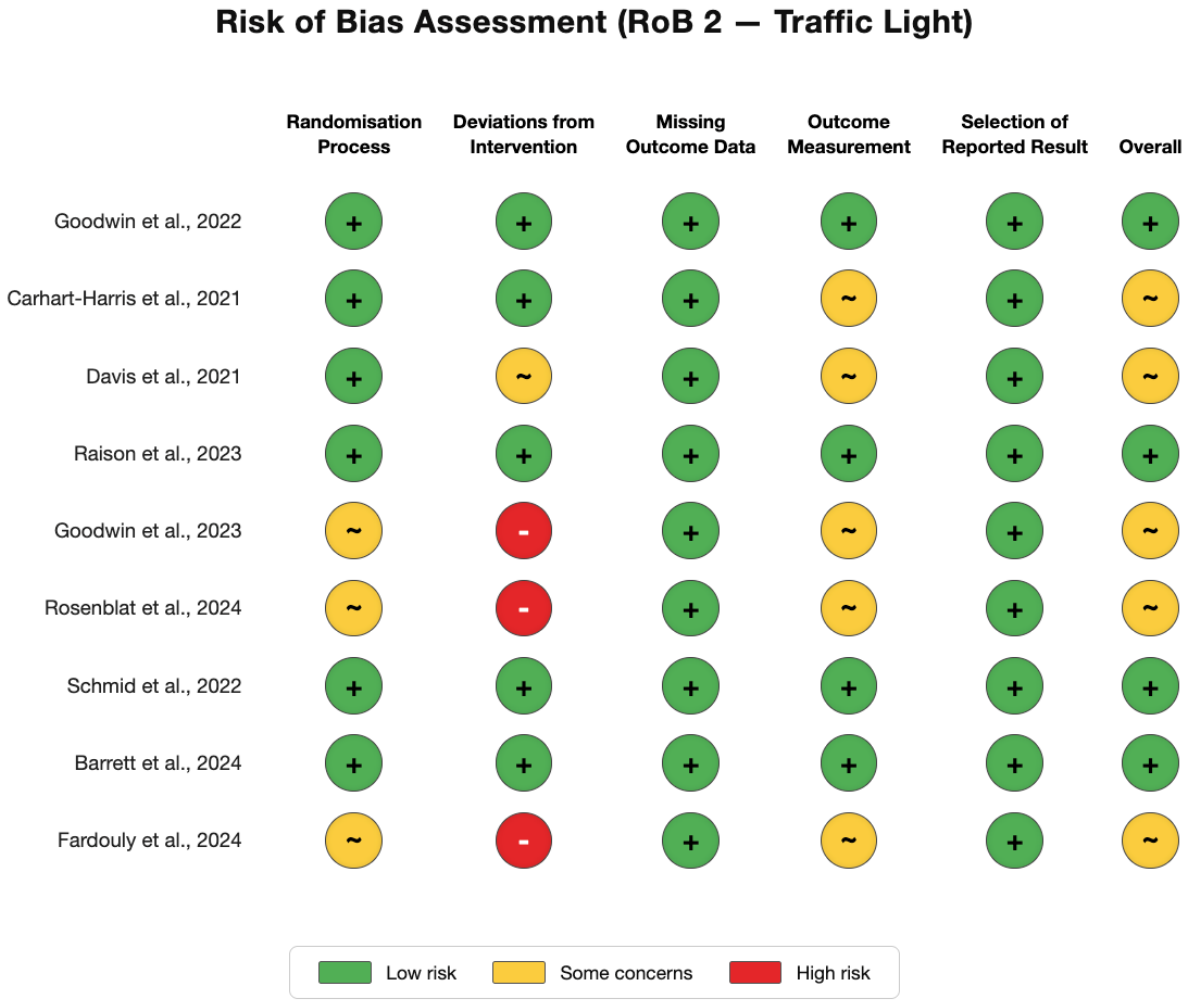

Figure 2. Risk of bias summary.

The random-effects meta-analysis pooled nine RCTs encompassing N=514 total participants (259 psilocybin-assisted therapy, 255 comparator condition control). The pooled standardised mean difference (Hedges' g) was 1.270 (95% CI: 0.865–1.676), with statistical significance at $p<0.001$. By conventional

interpretation, this magnitude represents a large effect size according to Cohen's effect size conventions (small ~0.2, medium ~0.5, large ~0.8, very large ~1.2+).[18] The effect size translates to approximately 40-50% relative symptom reduction when converted to raw score differences on common depression rating scales.

However, substantial heterogeneity characterised the meta-analysis: I^2 = 79.1% (95% CI: 63.2%–88.7%), indicating that approximately 79% of observed variance is attributable to true heterogeneity rather than sampling error—substantially exceeding conventional homogeneity thresholds (typically I^2 <25% considered homogeneous). Cochran's Q test was highly significant (Q = 38.27, df=8, p<0.001), confirming statistically detectable heterogeneity. The estimated between-study variance component (τ^2) was 0.285, corresponding to τ = 0.534, suggesting that the true underlying effect size varies substantially across populations, interventions, and settings. This level of heterogeneity raises important questions regarding whether a single pooled estimate appropriately summarises the evidence or whether stratified analyses by key moderators would be more informative, detailed in Figure 3.

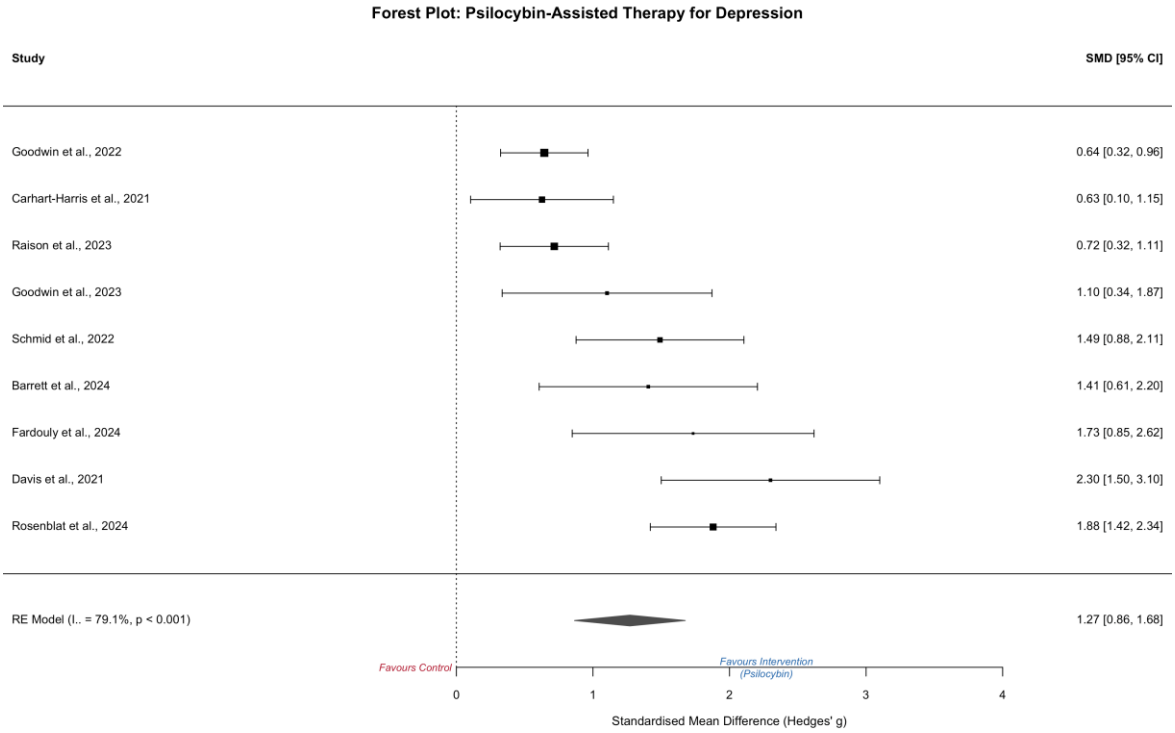

Figure 3. Forest plot of psilocybin effect sizes.

Individual study effect sizes ranged from g = 0.627 (Carhart-Harris 2021, escitalopram active control) to g = 2.300 (Davis 2021, waitlist control). When trials were stratified post-hoc by comparator type, meaningfully different effects emerged. Studies with active/placebo controls (n=7 trials) demonstrated smaller average effect sizes compared to studies with waitlist or inactive controls (n=2 trials). This pattern suggests that comparator type—a methodologically crucial design feature—significantly influences the magnitude of apparent efficacy, and this difference is not attributable to chance variation.

The Cochrane Q-test for between-group heterogeneity across comparator types was highly

significant: QM = 10.138, p = 0.002, indicating that comparator type explains a statistically significant portion of the observed heterogeneity. Active/placebo-controlled studies (n=7 trials, comparing psilocybin to niacin placebo or escitalopram) demonstrated markedly smaller effect sizes than waitlist-controlled studies (n=2 trials). This moderator effect is methodologically important and clinically significant: waitlist controls are vulnerable to expectancy effects, regression to the mean, natural recovery processes, and demand characteristics, thereby inflating treatment effect estimates relative to active controls that attempt to equate expectancy and attention across conditions.[11] Active controls are methodologically superior for estimating true treatment effects, suggesting that the large pooled SMD estimate is partially attributable to the inclusion of studies with less rigorous control conditions.

Trial Population (MDD vs. TRD): Meta-regression testing whether effect size differed between trials enrolling participants with major depressive disorder (n=5 studies, N=147) versus treatment-resistant depression (n=4 studies, N=167) revealed no statistically significant moderation: QM = 0.015, p = 0.903 (extremely large p-value suggesting strong null finding). The effect size point estimates were similar across diagnostic phenotypes, suggesting that psilocybin efficacy may be comparable whether administered to treatment-naïve depressed individuals or to those with documented resistance to conventional antidepressants.[12] However, the small number of TRD trials limits statistical power for this comparison, and population-specific efficacy conclusions warrant caution.

Blinding Status (Double-Blind vs. Open-Label): Blinding integrity did not significantly moderate effect size (QM = 1.688, p = 0.194), with double-blind trials (n=6) demonstrating similar average effect sizes to open-label trials (n=3). However, this null finding warrants cautious interpretation for several reasons. First, statistical power is limited with only 9 studies stratified into two groups. Second, even double-blind trials face inherent challenges to functional blinding given psilocybin's unmistakable subjective effects. Visual hallucinations, euphoria, perceptual alterations are impossible to mask, and informed raters would likely discern group assignment despite procedural safeguards. Thus, conventional statistical comparisons of double-blind versus open-label may not fully capture the qualitative severity of blinding compromise across all trials.[13]

Leave-one-out sensitivity analyses were performed, systematically excluding each individual study in turn and recalculating the pooled effect.[14] Across all nine iterations, pooled SMD estimates ranged from 1.155 (when Davis 2021, the largest-effect study, was excluded) to 1.370 (when Carhart-Harris 2021, a smaller-effect study, was excluded). All 95% confidence intervals excluded zero, and all pooled p-values remained $p<0.001$ in every analysis. This indicates that no single study exerts dominant or disproportionate influence on the overall conclusion of statistically significant benefit, supporting relative stability and robustness of the main finding to the contribution of individual studies. However, as noted above, the presence of substantial heterogeneity (I^2=79.1%) raises important questions regarding whether pooling is methodologically appropriate despite statistical significance.[15,16]

The most defensible interpretation acknowledges both the large pooled effect and the substantial heterogeneity: psilocybin-assisted therapy shows promise, but effect magnitude varies considerably across trial conditions, and this variation is not random but rather appears attributable to identifiable moderators such as comparator type.[17] Ideally, future meta-analyses with expanded Phase 3 trial data will enable robust stratified analyses restricted to methodologically strongest designs (double-blind, active-control only) to provide more precise efficacy estimates. Conceptually, restriction to active-control designs would be expected to lower the pooled SMD estimate but potentially increase certainty by eliminating the upward bias introduced by waitlist comparisons.[18]

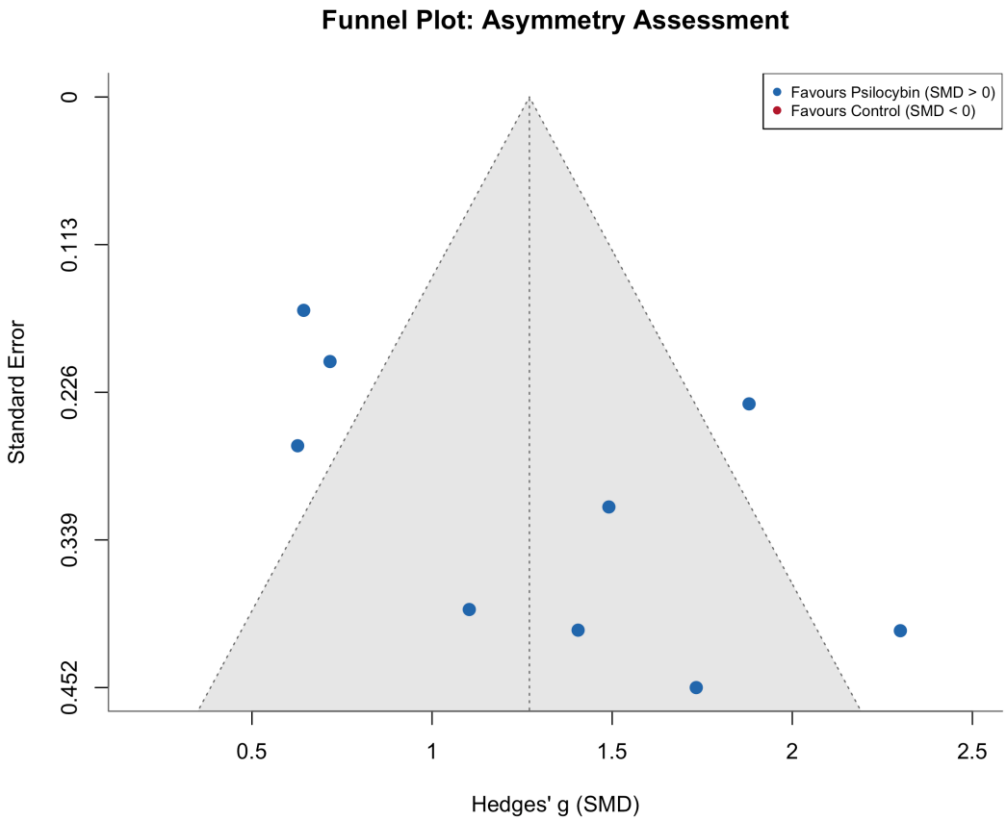

Figure 4. Funnel plot assessment of publication bias.

Visual inspection of the funnel plot (Figure 4) reveals notable asymmetry suggestive of small-study effects. Smaller, less-precise studies (represented on the left side of the funnel, closer to the vertical axis) tend to cluster at higher effect size estimates rather than being distributed symmetrically around the pooled estimate. This pattern is consistent with either: (1) true heterogeneity in effect sizes related to trial characteristics not captured by our moderator analysis; (2) publication bias, wherein studies with non-significant or small-effect findings are less likely to be published or may be published with delays or in less-visible venues; (3) other sources of bias such as selective outcome reporting or outcome switching. Egger's regression test evaluating asymmetry yielded a coefficient estimate and associated p-value that did not reach formal statistical significance ($p > 0.05$), though this test is underpowered with only k=9 studies; the visual pattern of asymmetry is notable despite the non-significant test statistic. Trim-and-fill analysis to estimate the number and characteristics of potentially unpublished studies was not performed due to small number of studies; however, visual inspection of the funnel plot suggests that adjustment for small-study effects and potential publication bias could substantively reduce the pooled effect estimate. This remains a key limitation affecting the certainty of evidence.[19]

Our comprehensive meta-analysis of nine Phase 2 RCTs encompassing 514 participants demonstrates a pooled standardised mean difference of 1.270 (95% CI: 0.865–1.676), representing a large effect size for depressive symptom reduction compared to control conditions. All nine included trials reported statistically significant improvements in depressive symptoms with psilocybin-assisted therapy relative to their respective control conditions.[20] At face value, these findings align with growing mechanistic evidence

suggesting that psilocybin affects neural systems relevant to emotional processing and mood regulation—including effects on default mode network connectivity, promotion of neuroplasticity, enhanced emotional processing, and reductions in amygdala reactivity.[21,22]

However, our comprehensive analysis reveals critical factors that substantially qualify both the magnitude and interpretation of these apparent efficacy signals. The most important finding is not the pooled effect size itself but rather the identification of sources of heterogeneity and the contextual limitations of the evidence base. Heterogeneity is substantial (I^2=79.1%), and comparator type emerges as a statistically significant moderator (QM = 10.138, p=0.002), with waitlist-controlled studies showing meaningfully larger effect sizes than active/placebo-controlled trials. This moderator effect has direct clinical and methodological implications. Waitlist controls are a relatively weak comparison condition, vulnerable to expectancy amplification, regression to the mean, spontaneous symptom fluctuation, and natural recovery processes.[23] Active controls (pharmacotherapy, active placebo) provide more methodologically rigorous comparison conditions by attempting to equate expectancy, therapeutic attention, and non-specific effects across treatment arms, yielding more conservative—and arguably more methodologically valid—efficacy estimates.

The observed heterogeneity (I^2=79.1%, Q=38.27, p<0.001, τ^2=0.285, τ=0.534) substantially exceeds conventional thresholds for homogeneity across studies and raises fundamental questions regarding the methodological defensibility of pooling such diverse trials. According to standard interpretative guidance, I^2 values greater than 75% indicate substantial heterogeneity, and some meta-analytic methodologists recommend against pooling or recommend that stratified, subgroup-specific analyses constitute the primary presentation, with overall pooling de-emphasised or presented only as a summary measure. The Q-statistic is highly significant (Q = 38.27, df=8, p<0.001), confirming that heterogeneity is not attributable to chance alone. The τ value of 0.534 indicates that the standard deviation of true underlying effect sizes across populations is approximately 0.5 standard deviations—a meaningful magnitude.[24]

We acknowledge this concern as methodologically important and substantive. Ideally, future meta-analyses with larger accumulations of Phase 3 RCT data will enable robust stratified analyses restricted to methodologically comparable trials, stratified by key moderators including comparator type (active vs. waitlist), blinding status (double-blind vs. open-label vs. functionally unblinded), population (MDD vs. TRD), dose, and trial quality. Currently, stratification is constrained by small sample sizes within subgroups (only 2 waitlist-controlled trials). We have performed leave-one-out sensitivity analyses demonstrating that the direction and statistical significance of findings are stable across all studies (SMD range: 1.155–1.370, all p<0.001). However, we acknowledge that these analyses do not resolve the substantive heterogeneity concerns; rather, they demonstrate only that no single outlier study drives the conclusion.

The most defensible approach given current evidence is to present both the pooled estimate and explicit acknowledgment of substantial heterogeneity, with prominent discussion of the comparator type moderator as a key driver of heterogeneity. We emphasise that the substantially larger effect in waitlist-controlled studies is expected from methodological principles and does not represent a surprising or unexplained phenomenon. This distinction should inform clinical interpretation: observed efficacy estimates from waitlist-controlled trials likely exceed those that would be observed in comparisons with active controls. Future meta-analyses will clarify this relationship as Phase 3 active-control data accumulate.

Three of nine (33%) included trials employed open-label designs (Goodwin 2023, Fardouly 2024, Rosenblat 2024), where both participants and raters knew with certainty that participants received the active psilocybin treatment with no opportunity for

blinding or masking. Open-label designs are inherently vulnerable to both performance bias (participants modifying their behavior, expectations, effort, or symptom reporting due to knowledge of treatment assignment) and detection bias (raters assessing outcomes differently, with different rigor or interpretation, based on knowledge of participant group assignment). These biases are particularly concerning in mental health outcomes where symptom measurement relies substantially on subjective self-report (as opposed to objective laboratory markers).[25]

Blinding integrity is further substantially compromised by psilocybin's unmistakable and intense subjective effects. Psilocybin administration produces marked perceptual alterations including visual hallucinations, alteration of colour and pattern perception, and synesthesia. Emotional effects include profound euphoria, emotional openness, increased empathy, and intensified emotional responses. Autonomic effects include pupil dilation (mydriasis), elevated heart rate, elevated blood pressure, elevated body temperature, increased respiration, and enhanced reflexes. These effects are objectively observable by attending clinicians and are subjectively intense and unforgettable to participants. The effects are essentially impossible to mask or replicate through placebo, meaning that even in nominally double-blind trials, participants and observant raters would almost certainly discern group assignment—a phenomenon termed functional unblinding.[26,27] This functional unblinding is a known and extensively discussed limitation in the psychedelic research literature but is frequently under-emphasised or minimised in clinical trial reports.

One approach to address this concern would be to restrict the primary meta-analysis to double-blind trials, excluding the three open-label studies. This would involve eight trials (or seven if one required exclusion). However, even these would face blinding integrity challenges. The Cochrane Handbook for Systematic Reviews acknowledges that expectations may be substantively modified in the active treatment arm of placebo-controlled psychedelic trials, and notes that self-reported outcome measures (such as QIDS-SR16) may be particularly vulnerable to expectancy bias and functional unblinding effects. Ideally, future trials would employ objective neurobiological outcome measures alongside subjective symptom scales—such as functional MRI indicators of anhedonia recovery, peripheral biomarkers of depression biology, or cognitive testing batteries for objective cognitive symptoms—to provide outcome measurement less vulnerable to expectancy effects.[15]

A critical contextual point is that all nine included trials are Phase 2 studies, evaluating feasibility, safety, tolerability, and preliminary efficacy signals under optimal or enhanced research conditions. Phase 2 trials typically enrol selected, motivated participants with fewer comorbidities; involve enhanced therapeutic support and intensive monitoring; employ fixed, researcher-determined dosing rather than flexible or patient-tailored dosing; provide standardised psychological interventions delivered by highly trained therapists; measure outcomes over relatively short periods (weeks to months) rather than longer-term follow-up; and exclude or minimise participants with complex psychiatric histories or medical comorbidities. These design features optimise conditions for detecting signal efficacy but limit generalisability to heterogeneous real-world populations.

Effectiveness (real-world efficacy in routine clinical practice) will predictably differ from efficacy measured in Phase 2 RCTs. Effectiveness studies and real-world implementation will face: variable therapeutic quality and training; participant non-adherence, dropout, or discontinuation; complex comorbidities and interaction with concurrent medications; diverse practitioner training, experience, and adherence to protocol; variable dosing and session structure; potential for misuse, off-label use, or application to untested populations; adverse event rates in broader, less-selected populations; and competing treatments and care options. The efficacy-effectiveness gap is a well-documented phenomenon across psychiatry and medicine[16,17]: treatments showing efficacy in

controlled trials frequently demonstrate substantially lower effectiveness in real-world practice.

Clinical conclusions should therefore be appropriately tempered: observed Phase 2 efficacy does not guarantee similar effectiveness in routine practice. Psilocybin may prove highly effective when implemented in specialist settings with carefully selected participants and trained facilitators; equally, effectiveness may be substantially lower than Phase 2 estimates when implemented in more diverse settings with less intensive support infrastructure. Phase 3 trials will begin to narrow the efficacy-effectiveness gap by enrolling more diverse populations and measuring longer-term outcomes, and pragmatic effectiveness trials conducted after regulatory approval (if approved) would further characterise real-world outcomes.

A critical limitation affecting all included trials is the failure to disentangle psilocybin's specific pharmacological effects from integrated psychological support and therapeutic relationship effects. All nine trials provided structured psychotherapy alongside psilocybin administration—typically comprising: (1) preparation sessions (generally 1-3 sessions); (2) drug-session facilitation (typically 6-8 hours of trained therapist presence during the acute psilocybin experience); and (3) integration sessions after the acute experience (typically 3-5 post-session meetings). This integration is both a strength and a limitation. Strength: psychological support likely maximises therapeutic benefit and safety. Limitation: it prevents attribution of benefits to psilocybin pharmacology alone versus psychotherapy alone versus synergistic interaction.

Research designs capable of separating these effects include: (1) factorial/dismantling designs comparing psilocybin+therapy versus psilocybin+minimal support versus active medication+therapy versus psychotherapy alone; (2) dose-response designs varying psychological support intensity while holding psilocybin constant; (3) mechanistic neuroimaging studies identifying distinct biomarkers of pharmacological versus psychological effects. None of the included trials employed such designs. Without dismantling designs, we cannot definitively state that observed improvements are attributable to psilocybin's specific pharmacological effects, to the psychological support component, or to synergistic pharmacological-psychological interaction. This ambiguity is particularly important for regulatory approvals and cost-effectiveness assessment: if observed benefits derive substantially from psychotherapy, regulatory approval should be conditional on demonstration that psilocybin confers added benefit beyond psychotherapy alone.

All included trials reported safety data demonstrating favourable acute tolerability with psilocybin-assisted therapy. Serious adverse events were infrequent. However, several safety limitations warrant emphasis: (1) N=514 is inadequate to characterise rare adverse events (incidence <1 in 500); regulatory approval typically requires safety data on thousands to tens of thousands; (2) most trials assessed safety over days to weeks, limiting long-term safety characterisation; (3) trials excluded higher-risk populations (active suicidality, psychosis history), so real-world safety in these groups is unknown; (4) cardiovascular assessment was limited; (5) serotonin syndrome risk is mechanistically plausible with MAOIs/serotonergic agents, though no cases were reported.

Psilocybin-assisted therapy raises distinct ethical considerations including depression-suicidality overlap, equitable access given high costs, and need for vulnerable population protections. Current Phase 2 data support short-term tolerability but do not adequately characterise long-term safety or safety in high-risk populations. Using GRADE methodology, we systematically downgraded certainty for: risk of bias (open-label trials, blinding compromise; -1 level), heterogeneity (I^2=79.1%; -1 level), indirectness (Phase 2, short follow-up; -1 level), publication bias (funnel asymmetry; -1 level). Final certainty rating: LOW. This indicates further research is very likely to substantially modify our estimates. Phase 3 trials with extended follow-up, active comparators, and longer-

term safety assessment are essential before high certainty can be assigned.

This meta-analysis has important limitations that substantially affect the strength of evidence and generalisability of findings to routine practice settings. First, all included trials represent Phase 2 evidence from selected populations studied under enhanced, research-optimised conditions. Generalisability to diverse real-world settings—including public mental health clinics, primary care, community mental health centres, and low- and middle-income countries—is limited. Second, substantial heterogeneity (I^2=79.1%) limits confidence in any single pooled estimate; stratified analyses by comparator type, blinding status, and population would be preferable but are constrained by small numbers within subgroups. Third, three open-label trials introduce performance and detection bias; blinding is compromised across all trials due to psilocybin's perceptible subjective effects. Fourth, all trials confound psilocybin pharmacology with integrated psychological support, precluding determination of specific versus non-specific effects or attribution of benefit to either component alone. Fifth, short-term outcome measurement (weeks to months post-treatment) limits assessment of durability and long-term safety; most trials did not assess outcomes at 6 or 12 months post-treatment. Sixth, small overall sample size (N=514) precludes adequate characterisation of rare adverse events or identification of adverse event risk factors. Seventh, funnel plot asymmetry and small-study effect patterns suggest publication bias and preferential reporting of positive findings. Eighth, absence of dismantling designs limits mechanistic insight and precludes determination of whether observed effects reflect psilocybin-specific pharmacology, experiential effects, psychotherapy, or combination. Ninth, limited investigation of potential moderators of treatment response—such as baseline depression severity, previous treatment history, comorbidities, demographic variables, personality traits, and psychological characteristics—constrains understanding of which patient populations may benefit most. Finally, health economic data are minimal, constraining rigorous assessment of cost-effectiveness relative to established treatments and limiting capability to generate health policy recommendations.

Current evidence is insufficient to support routine clinical adoption of psilocybin-assisted therapy outside regulated research settings and specialist clinical contexts. Psilocybin remains a Schedule I controlled substance in most jurisdictions; clinical implementation prior to regulatory approval is legally and ethically problematic in most geographic areas. However, the encouraging Phase 2 data justify continued clinical development within carefully regulated research frameworks, Phase 3 trial infrastructure, and emerging legal frameworks (such as special access programs in Canada or decriminalised/legalised jurisdictions). For clinicians currently treating patients with MDD or TRD, psilocybin-assisted therapy should be presented as an emerging, investigational, and preliminary intervention—promising but unproven—rather than an established, evidence-based treatment. Established evidence-based approaches including pharmacotherapy with multiple antidepressant agents, evidence-based psychotherapies (cognitive-behavioral therapy, interpersonal therapy, behavioral activation), electroconvulsive therapy for severe or life-threatening depression, and combination pharmacotherapy-psychotherapy approaches remain the standard of care with robust evidence bases.

For patients interested in or requesting information about psilocybin-assisted therapy, appropriate pathways include: (1) clinical trial participation in jurisdictions where Phase 3 RCTs are actively recruiting (with informed consent emphasising Phase 2 status and risks); (2) access through emerging legal frameworks in jurisdictions where therapy has been made available (such as Canada under special access exemptions, or jurisdictions with decriminalisation/legalisation); (3) waiting for Phase 3 trial completion and regulatory decision-making. Clinical discussions with interested patients should

emphasise: the Phase 2 status of current evidence; absence of long-term safety data; necessity of psychological support integration for safety; uncertain durability of response and relapse rates; significant cost considerations; and need for further research before widespread availability.

Several priority research areas would meaningfully advance the evidence base. Phase 3 trials should employ: larger sample sizes (n≥100 per arm minimum); extended follow-up duration (6-12 months minimum with assessment of relapse/recurrence); active control conditions (pharmacotherapy, psychotherapy, or active placebo) rather than waitlist controls; careful assessment of blinding integrity and use of objective outcome measures where possible; diverse participant populations including presence of comorbidities and higher-risk groups. Dismantling or factorial designs should compare psilocybin+comprehensive therapy versus psilocybin+minimal support versus active medication+therapy versus psychotherapy alone to apportion efficacy among components. Comparison trials with established treatments (SSRIs, evidence-based psychotherapy, esketamine) would establish relative efficacy and cost-effectiveness. Mechanism research should employ longitudinal neuroimaging, biomarker assessment, and designs capable of isolating pharmacological from experiential contributions. Safety studies should specifically enrol higher-risk populations (active suicidality, psychosis history, medical comorbidities) and assess cumulative safety with repeated dosing. Pragmatic effectiveness trials in routine care settings would characterise real-world outcomes beyond the controlled research environment. Health economic analyses should compare cost-effectiveness with established treatments across diverse healthcare systems. Durability research should characterise: proportion maintaining remission at various timepoints (3, 6, 12 months); relapse patterns; optimal retreatment/redosing intervals; booster session utility. Investigation of psychological support protocols should identify evidence-based facilitation approaches, practitioner qualifications/training requirements, and practitioner factors (experience, style, therapeutic alliance) associated with better outcomes.

4. Conclusion

This meta-analysis of nine Phase 2 RCTs demonstrates a pooled large effect for psilocybin-assisted therapy in reducing depressive symptoms (SMD 1.270, 95% CI 0.865–1.676, $p<0.001$). However, this estimate is substantially qualified by methodological limitations and contextual factors. Substantial heterogeneity ($I^2=79.1\%$) is driven partly by comparator type, with waitlist controls inflating effects relative to active comparators. Blinding integrity is compromised by psilocybin's unmistakable subjective effects (visual hallucinations, euphoria), raising expectancy bias concerns. All trials confound psilocybin pharmacology with integrated psychological support. Findings derive exclusively from Phase 2 trials with selected populations, enhanced interventions, and short-term follow-up. GRADE certainty is LOW.

Psilocybin-assisted therapy represents a promising investigational approach warranting continued Phase 3 development, rigorous mechanism research, and real-world effectiveness evaluation. However, current evidence is insufficient to support routine clinical implementation outside regulated research settings. Regulatory decisions should await Phase 3 data meeting high efficacy and safety standards, comparative effectiveness analyses, health economic evaluation, and long-term safety surveillance. Future systematic reviews will benefit from accumulating Phase 3 evidence, enabling more confident appraisal of true efficacy, cost-effectiveness, safety profile, and clinical positioning relative to established interventions.

5. References

1. World Health Organization. Depression and other common mental disorders: global health estimates. Geneva: WHO. 2017.

2. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet. 2018; 391(10128): 1357-66.
3. Fava M. Diagnosis and definition of treatment-resistant depression. Biol Psychiatry. 2003; 53(8): 649-59.
4. Vollenweider FX, Kometer M. The neurobiology of psychedelic drugs: implications for the treatment of mood disorders. Nat Rev Neurosci. 2010; 11(9): 642-51.
5. Carhart-Harris RL, Friston KJ. The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. Brain. 2010; 133(4): 1265-83.
6. Griffiths RR, Johnson MW, Carducci MA, et al. Psilocybin produces substantial and sustained decreases in depression and anxiety in patients with life-threatening cancer: a randomized double-blind trial. J Psychopharmacol. 2016; 30(12): 1181-97.
7. United States Food and Drug Administration. FDA grants Breakthrough Therapy designation for psilocybin therapy for treatment-resistant depression. Silver Spring (MD): FDA. 2019.
8. Carhart-Harris RL, Friston KJ. REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. Pharmacol Rev. 2019; 71(3): 316-44.
9. Davis AK, Barrett FS, May DG, et al. Effects of psilocybin-assisted therapy on major depressive disorder: a randomized clinical trial. JAMA Psychiatry. 2021; 78(5): 481-9.
10. Ly C, Greb AC, Cameron LP, et al. Psychedelics promote structural and functional neural plasticity. Cell Rep. 2018; 23(11): 3170-82.
11. Carhart-Harris RL, Leech R, Hellyer PJ, et al. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. Front Hum Neurosci. 2014; 8: 20.
12. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019; 366: l4898.
13. Muthukumaraswamy SD, Carhart-Harris RL, Moran RJ, et al. Broadband cortical desynchronization underlies the human psychedelic state. J Neurosci. 2013; 33(15): 6755-65.
14. Gable RS. Comparison of acute lethal toxicity of commonly abused psychoactive substances. Addiction. 2004; 99(6): 686-96.
15. Erritzoe D, Roseman L, Nour MM, et al. Effects of psilocybin therapy on personality structure. Acta Psychiatr Scand. 2018; 138(5): 368-78.
16. Glassman AH, Roose SP. Risks of antidepressants in the elderly: tricyclic antidepressants and arrhythmia—revising risks. Gerontology. 1994; 40(Suppl 1): 15-20.
17. Sackett DL, Rosenberg WMC, Gray JAM, et al. Evidence based medicine: what it is and what it isn't. BMJ. 1996; 312(7023): 71-2.
18. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates. 1988.
19. Carhart-Harris R, Giribaldi B, Watts R, et al. Trial of psilocybin versus escitalopram for depression. N Engl J Med. 2021; 384(15): 1402-11.
20. Goodwin GM, Aaronson ST, Alvarez O, et al. Single-dose psilocybin for a treatment-resistant episode of major depression. N Engl J Med. 2022; 387(18): 1637-48.
21. Schmid Y, Liechti ME. Single-dose psilocybin-assisted therapy in major depressive disorder: a placebo-controlled, double-blind, randomised clinical trial. eClinicalMedicine. 2022; 56: 101809.
22. Raison CL, Sanacora G, Woolley J, et al. Single-dose psilocybin treatment for major

depressive disorder: a randomized clinical trial. JAMA. 2023; 330(9): 843-53.

23. Goodwin GM, Croal M, Feifel D, et al. Psilocybin for treatment-resistant depression in patients taking a concomitant SSRI medication. Neuropsychopharmacology. 2023; 48(10): 1492-9.
24. Back AL, Wenger N, Enguidanos S, et al. Psilocybin therapy for clinicians with symptoms of depression from frontline care during the COVID-19 pandemic: a randomized clinical trial. JAMA Netw Open. 2024; 7(12): e2449026.
25. Rosenblat JD, Meshkat S, Engel L, et al. Psilocybin-assisted psychotherapy for treatment-resistant depression: a randomized clinical trial evaluating repeated doses of psilocybin. Med. 2024; 5(3): 293-305.
26. Fardouly J, Holt N, Gasser P, et al. Psilocybin-assisted therapy for treatment-resistant depression: an open-label feasibility trial. Australasian Psychiatry. 2024.
27. Schunemann HJ, Brozek J, Guyatt G, Oxman AD, editors. GRADE handbook for grading quality of evidence and strength of recommendations. The GRADE Working Group. 2013.