



# UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale - DPG  
Corso di laurea in Psicologia Clinica

Elaborato finale

## **Psychotherapies for Personality Disorders: A Systematic Review of RCTs with Focus on Psychosocial Functioning**

*Relatore*

Prof. Gentili Claudio

*Correlatrice esterna*

Prof.ssa Karyotaki Eirini

*Laureanda* Radio Marta

*Matricola* 2122087

Anno Accademico 2024/2025



**OUTLINE:**

- ABSTRACT:** ..... 1
- 1. INTRODUCTION** ..... 3
- 2. METHODS:** ..... 10
  - 2.1 Database and preliminary outcome analysis ..... 10
  - 2.2 Eligibility criteria..... 12
  - 2.3 Selection process ..... 14
  - 2.4 Data collection process ..... 15
  - 2.5 Study Risk of Bias Assessment..... 16
- 3. RESULTS:** ..... 17
  - 3.1 Preliminary outcome analysis ..... 17
  - 3.2 Study Selection ..... 17
  - 3.3 Description of included studies ..... 18
  - 3.4 Quality of the included studies ..... 19
  - 3.5 Functioning Outcomes ..... 20
- 4. DISCUSSION** ..... 22
  - 4.1 Preliminary outcome analysis ..... 22
  - 4.2 Study characteristics ..... 22
  - 4.3 Comparator issues ..... 24
  - 4.4 Risk of Bias ..... 26
  - 4.5 Strengths and limitations ..... 27
- 5. CONCLUSIONS** ..... 29
- REFERENCES**..... 32
- TABLES** ..... 37



# ABSTRACT:

**Background.** Psychotherapy is central to the care of personality disorders, yet trials often emphasize symptom change rather than psychosocial functioning, which is the conceptual core of contemporary AMPD and ICD-11 models. The field lacks an up-to-date review of how functioning is measured in RCTs for adult personality disorders.

**Objective.** To systematically identify randomized controlled trials of psychotherapies for adult personality disorders that used inactive comparators and reported functioning at the end of treatment, and to classify functioning outcomes into global, interpersonal, and self domains.

**Methods.** Reporting followed PRISMA 2020. Searches of PubMed, Embase, and APA PsycInfo to November 2024 informed a larger database from which eligible trials were selected. Inclusion criteria were adults with DSM or ICD personality disorder, a manualized or structured psychotherapy, an inactive comparator, and at least one functioning measure assessed at treatment end. Study characteristics were extracted with a standardized template. Risk of bias was appraised with RoB 2. Synthesis was qualitative.

**Results.** The search yielded 4,914 records; 13 trials met the inclusion criteria. Studies were conducted in Canada, the Netherlands, Italy, the United Kingdom, and the United States, with a total randomized sample of 2,008 participants. Diagnoses included borderline personality disorder in six trials, any personality disorder in four, avoidant in one, antisocial in one, and mixed non-BPD in one. Interventions comprised dialectical behavior therapy, schema therapy, cognitive-behavioral and interpersonal approaches,

short-term dynamic psychotherapy, psychoeducation/structured support, and a web-based program. Comparators were treatment as usual (n=9), wait-list (n=3), or monitoring-only (n=1). Functioning outcomes were reported as global in seven studies, interpersonal in eight, and self in one; no trial used AMPD Level of Personality Functioning measures. Overall risk of bias was high in 9/13 trials (69%) and raised some concerns in 4/13 (31%); none were low risk.

**Conclusions.** Across randomized studies, functioning was included but operationalized heterogeneously; assessments of self-functioning were sparse, and no study employed AMPD-consistent measures. The overall certainty is weakened by methodological issues. Next trials should prospectively specify functioning outcomes, use validated AMPD or ICD-11 instruments, register protocols in advance, blind outcome assessment where feasible, and report functioning together with symptoms at post-treatment and follow-up.

# 1. INTRODUCTION

In DSM-5 and DSM-5-TR, personality disorder is defined as a stable and pervasive pattern of maladaptive cognition, affect, interpersonal functioning, or impulse control that begins by early adulthood and leads to clinically significant impairment or distress (American Psychiatric Association, 2013; American Psychiatric Association, 2022).

Despite increasing recognition, personality disorders remain under-identified and exert a disproportionate burden on affected individuals and on health systems. They are strongly associated with functional impairment, elevated risk of suicidality, and high service utilization (Tyrer, 2010; Skodol, 2018; McClelland, 2023; Botham, 2024). A global meta-analysis estimates a prevalence of 7.8% in the general population (95% CI 6.1–9.5), with higher rates in high-income countries than in low- and middle-income settings (Winsper, 2020). Prevalence is considerably higher in secondary-care contexts, underscoring the demand for mental-health services (Tyrer, 2010). Beyond prevalence, impairments in psychosocial functioning exceed those observed in major depressive disorder and persist even when comorbid conditions are controlled (Skodol, 2018; Tyrer, 2010). Individuals with personality disorders face a markedly elevated risk of suicide attempts and deaths compared with those without personality disorder or with other psychiatric disorders (McClelland, 2023). From a health-system perspective, personality disorder care is costly. Specialist services tend to shift spending from hospital stays to community care, and costs rise mainly when serious mental illness or substance use are also present (Botham, 2024). Taken together, the combination of high prevalence, disproportionate functional impairment, heightened suicide risk, and

substantial health-system costs underscores the need for improved recognition, diagnosis, and effective treatment.

Historically, categorical approaches to personality disorders have been criticized for high comorbidity, heterogeneity within diagnoses, and limited validity. The categorical approach in DSM-IV and ICD-10 mixed trait descriptors with severity indices, undermining diagnostic specificity (Widiger, 2005). These limitations motivated the development of dimensional models. The concept of personality functioning has long existed in the field, but it was explicitly formalized as Criterion A in the DSM-5 Alternative Model for Personality Disorders (AMPD) and paralleled in ICD-11 (Hopwood, 2024). Criterion A, the Level of Personality Functioning, operationalizes impairment in identity and self-direction on the self-side and in empathy and intimacy on the interpersonal side, establishing both the presence and the severity of personality disorder. Criterion B specifies maladaptive personality traits across five broad domains: negative affectivity, detachment, antagonism, disinhibition, and psychoticism, with facet-level detail that describes stylistic expression. Within this model, six personality disorder types are retained and reformulated in terms of Criteria A and B, while other cases are diagnosed as personality disorder, trait specified; this allows clinicians to report severity and trait configuration instead of forcing assignment to a mismatched category (American Psychiatric Association, 2013). ICD-11 mirrors this architecture by diagnosing personality disorder primarily according to overall severity of impairment in self and interpersonal functioning, rated as mild, moderate, or severe, and then using trait qualifiers, Negative Affectivity, Detachment, Dissociality, Disinhibition, and

Anankastia, with an optional borderline pattern descriptor (World Health Organization, 2019). Longitudinal meta-analytic evidence indicates only moderate stability of personality disorders over time. Across 40 studies and 38,432 participants, 56.7% retained any PD and 45.2% retained borderline PD at follow-up (d'Huart, 2023). Dimensional mean levels of most PD criteria declined from baseline to the last assessment, with antisocial, obsessive–compulsive, and schizoid criteria showing no decrease. According to D'Huart et al. (2023), estimates varied widely, and between-study heterogeneity was substantial; few studies used AMPD or ICD-11 measures. These patterns reinforce the centrality of self and interpersonal functioning in contemporary nosology and motivate explicit assessment of functioning in outcome research (d'Huart, 2023).

This reorientation has important clinical implications. Interpersonal functioning improves following psychotherapy in BPD, although effects are not yet robust and measurement remains heterogeneous (Sinnaeve et al., 2015). Longitudinal studies indicate that functioning tends to improve earlier and more strongly than trait profiles, supporting a function-first treatment strategy (Kiel, 2024). Change is also domain-specific. Improvements in personality functioning may occur alongside slower gains in occupational or academic participation, underscoring the need to track functioning explicitly and across multiple domains (Kvarstein, 2023). For treatment planning, the Level of Personality Functioning offers a common indicator that applies to all patients and can be complemented by trait assessment to capture stylistic variations (Natoli and Hopwood, 2025). Personality functioning is distinct from trait style and shows stronger

links to reductions in personality disorder symptoms over time, reinforcing its value as a central clinical and research outcome (Hopwood et al., in press). In routine clinical contexts, brief psychotherapy produces larger changes in functioning than in trait profiles, and lower baseline functioning predicts dropout, which reinforces its relevance outside specialist settings as well (Kiel, 2024). In sum, the AMPD and ICD-11 represent a shift away from categorical classification and toward a dimensional understanding of personality disorder that emphasizes functioning as the conceptual and clinical core.

Recent debate around the AMPD concerns how to conceptualize the relations among personality functioning, maladaptive traits, and real-world problems in living (Hopwood, 2024). Hopwood proposes a tripartite framework that treats these as conceptually distinct, with personality functioning as a latent severity dimension in self and interpersonal domains, traits as dispositional style, and problems in living as the clinical sequelae; clearer distinctions can reorient research toward clinical utility (Hopwood, 2024). In a paired commentary, Fonagy and colleagues endorse the tripartite view but argue that the field remains overly descriptive; they call for explanatory and causal models, that specify how self and other processes give rise to trait expression and consequently life difficulties. This perspective clarifies when functioning operates as a mechanism rather than mere covariance (Zavlis & Fonagy, 2024). At the same time, Hutsebaut and Sharp remind us that a diagnostic system should ultimately serve patients and clinicians. The promise of the AMPD lies not only in its conceptual structure but also in its ability to guide treatment planning, provide less stigmatizing language for personality pathology, and connect assessment with processes of change.

Because clinicians and patients value transparency, collaboration, and hope in assessment, formulations based on functioning and traits are best presented as tools for empowerment rather than labels. Progress in the field depends on bridging research and practice while keeping sight of the shared goal of clinically useful and theoretically coherent diagnosis and treatment. To this end, the thesis adopts a function-first orientation. Functioning indexes severity and helps identify targets for change; traits specify stylistic expression; problems in living anchor clinical priorities. Together, these link classification, mechanism, and practice (Hopwood, 2024; Zavlis and Fonagy, 2024; Hutsebaut and Sharp, 2024).

The literature lacks a recent, cross-diagnostic systematic review of psychotherapy trials for adult personality disorders that prioritizes functioning rather than focusing only on borderline presentations, and existing syntheses are either out of date or confined to single diagnoses, which leaves uncertainty about how functioning is operationalized across the field. Twelve years after the introduction of the AMPD and in parallel with ICD-11, it is timely to examine how contemporary trials are designed, which comparators are used, which functioning domains are measured, and where methodological effort should be directed. The present review addresses this gap by mapping functioning outcomes in randomized controlled trials that used inactive comparators and reported functioning at post-treatment. Functioning outcomes are classified into three prespecified domains: global, interpersonal, and self-functioning; and the instruments used within each domain are catalogued. Trial characteristics and risk of bias are summarized to contextualize the evidence. By clarifying how functioning

is measured and reported, the review is conceptually consistent with frameworks that define personality disorder by impairment in self and interpersonal functioning and identifies priorities for future trials, including consistent use of validated measures of personality functioning.

A systematic review is a structured and transparent synthesis of research that follows a predefined protocol. The protocol states a focused question and specifies eligibility using the PICO framework, where Population defines who is enrolled, Intervention specifies the treatment under evaluation, Comparator identifies what it is measured against, and Outcomes state what is measured. Searches are run across multiple bibliographic databases and other sources, with strategies documented and yields recorded. Titles and abstracts are screened against the eligibility criteria, full texts are assessed, and all decisions are documented. Data are extracted with a standardized template and checked for accuracy. Systematic reviews are defined by exhaustive, documented searching, explicit eligibility criteria, formal quality appraisal, and transparent synthesis and analysis; these methods are chosen to minimize bias and enable replication (Grant, 2009). In contrast with narrative reviews, which may emphasize selected studies and are vulnerable to quality bias and selective emphasis, systematic methods make each step reproducible and evaluate study quality explicitly. When the included studies are sufficiently comparable, a meta-analysis can be used to estimate an average effect across trials and to examine variability in results; when comparability is limited or confidence in the evidence is low, statistical pooling is not appropriate, and a qualitative synthesis is preferable.

Quality of evidence matters because even a well-planned systematic review or meta-analysis cannot compensate for weak primary studies; poor inputs can degrade a synthesis and mislead conclusions (Ahn & Kang, 2018). To limit this risk, study limitations are appraised prospectively and used in interpreting results. For randomized trials, the Cochrane Risk of Bias 2 tool is used. RoB 2 evaluates five domains: the randomization process, deviations from intended interventions and analysis, missing outcome data, measurement of outcomes, and selection of the reported result, yielding result level judgments of low risk, some concerns, or high risk (Sterne et al., 2019). RoB 2 provides tailored variants for parallel, cluster, and crossover designs and supports standardized visualization. It is more complex and time-intensive, and inter-rater reliability can be modest. The removal of the former “other bias” domain motivates additional tracking of contextual features such as funding and registration, baseline imbalance, sample size methods, data availability, and follow-up (Nejadghaderi, 2024). Tracking these features can reduce residual quality bias and improve interpretability. For bodies of evidence across studies, the GRADE approach rates certainty by considering risk of bias, inconsistency, imprecision, indirectness, and publication bias, and links these judgments to the strength of recommendations (Ahn & Kang, 2018).

## 2. METHODS:

The reporting of this thesis was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021). PRISMA 2020 replaces the 2009 statement and provides clearer, more implementable reporting advice. It includes a 27-item checklist, an expanded checklist with detailed recommendations, an abstract checklist, and revised flow diagrams (templates for new/updated reviews). Adhering to PRISMA 2020 improves transparency and completeness, enabling readers to appraise methods, applicability, and credibility, while supporting authors, editors, guideline developers, and policy makers in using the findings (Page et al., 2021).

### 2.1 Database and preliminary outcome analysis

The studies analysed in this systematic review originate from a larger database developed within a collaborative project with other researchers. This database was designed to collect randomized controlled trials (RCTs) evaluating psychotherapeutic interventions for personality disorders, using deliberately broad eligibility criteria. At this initial stage, the aim was to map the literature comprehensively and to assess whether the heterogeneity of available trials would allow for a systematic review or a meta-analysis, and, if so, what type of meta-analysis could be conducted, and which

outcomes could be extracted for analysis. From this larger pool, a targeted selection of studies was carried out for the present review, according to specific inclusion and exclusion criteria (described in Section 2.2).

The broader database was built through systematic searches across major electronic databases: *PubMed*, *EMBASE*, and *PsycInfo*. Searches covered all records up to November 2024.

In the screening phase, titles and abstracts were assessed in Rayyan to facilitate collaboration between multiple reviewers, while full-text records were managed in EndNote 21 (Clarivate Analytics). Within this process, RCTs published in English that evaluated psychotherapeutic interventions for individuals diagnosed with personality disorders according to DSM or ICD criteria were included. Systematic reviews and meta-analyses synthesizing RCTs of this type were also retained, as they were considered relevant for mapping the existing evidence base, whereas non-systematic sources such as narrative reviews, dissertations, protocols, and other non-peer-reviewed documents were excluded. To facilitate later synthesis, included records were further organized in Excel according to population (adolescents vs adults), comparator type (active vs inactive), outcomes, etc. Companion papers were managed in a separate Excel file, developed to classify them and cross-reference each with the corresponding main trial report. This organizational framework ensured that the database could be flexibly applied to different forms of evidence synthesis.

A preliminary outcome analysis was conducted to describe the range of outcome domains and instruments used across psychotherapy trials for personality disorders. Given the heterogeneity of this research area, the analysis was not designed as a formal synthesis but as an exploratory step to classify outcomes into broad domains (e.g., symptom severity, functioning, quality of life). It was performed on the consecutively coded set of eligible trials (n = 32) present in the extraction file during database building. This approach served two main purposes: first, to provide an overview of which outcome domains were most consistently reported; and second, to inform the feasibility of future evidence syntheses, particularly with respect to functioning outcomes, which constitute the primary focus of this review. Findings of this analysis are reported in the results section.

## 2.2 Eligibility criteria

Eligibility criteria were structured according to the PICO framework (Population, Intervention, Comparator, Outcomes), which is widely used to guide research questions in evidence synthesis (Eriksen & Frandsen, 2018).

Population (P): Adults ( $\geq 18$  years) diagnosed with a personality disorder according to DSM or ICD criteria. Studies including mixed or partially adolescent samples were excluded. Comorbid psychiatric diagnoses did not constitute an exclusion criterion, provided that a personality disorder diagnosis was the primary inclusion requirement of the trial. All settings, including outpatient, community, and forensic settings, were included

Intervention (I): Manualized or structured psychotherapies. Interventions could be delivered as standalone treatments or as adjuncts to treatment as usual (TAU) or pharmacotherapy, provided that the psychotherapy component was clearly defined and constituted the primary intervention under evaluation. Concomitant medication use by participants was not considered an exclusion criterion, unless medication was administered according to a systematic protocol forming part of the randomized intervention (e.g., psychotherapy combined with pharmacotherapy versus placebo), in which case the independent effect of psychotherapy could not be disentangled.

Comparator (C): Inactive control conditions, namely treatment as usual (TAU) or wait-list (WL), were eligible, whereas trials comparing two active psychotherapies were excluded. Trials were eligible when psychotherapies were compared with TAU, WL, or other control conditions that did not constitute *a bona fide* manualized psychotherapy. TAU was defined broadly as routine clinical care for personality disorders, which could include medication management, clinical monitoring, supportive visits, or structured pharmacotherapy, provided that such elements were applied equally across arms. Add-on designs were eligible when both groups received the same background treatment (e.g., identical pharmacotherapy protocol or clinical management), and randomization concerned the psychotherapy component. TAU was considered an “inactive” comparator for this review, although it is recognized that its content and implementation are highly heterogeneous across settings and trials (Duggan et al., 2007).

Outcomes (O): The primary focus was on measures of functioning, in line with the predefined domains of global, interpersonal, and self-functioning. Trials that reported only symptom outcomes without any functioning measure were excluded. Symptom

outcomes were extracted as secondary information when available, but they were not considered in the synthesis. A detailed mapping of functioning outcomes and instruments across included trials is provided in **Table 2**.

## 2.3 Selection process

When this review was initiated, the broader database, described in Section 2.1, was still under development, and full-text cross-checking between reviewers had not yet been completed. The present review, therefore, relied on the version of the database available at that time.

Based on this database version, records were managed in EndNote 21. Candidate studies were identified by filtering the database's internal tags (e.g., comparator status: inactive; presence of functioning outcomes), which had also been tracked in an auxiliary Excel file. These indicators were used only to locate potentially relevant records; eligibility was reassessed against the predefined criteria. All records were screened by a single reviewer under supervision; independent duplicate screening was not performed.

No automation or machine-learning tools were used in study selection, and no translations or author contacts were required at this stage. A PRISMA flow diagram summarizes the number of records at each stage and the documented reasons for exclusion at full text.

## 2.4 Data collection process

A customized Excel template was developed specifically for this review to standardize data extraction. Data extraction was performed by a single reviewer under supervision.

The primary outcomes of interest were measures of functioning. In line with the review aims, all functioning outcomes were classified into three domains: global (role/overall functioning), interpersonal (relationships/social participation), and self-functioning (identity/self-direction). Any validated instrument, or in older trials, trial-specific composite measures with clear face validity for global, interpersonal, or self-functioning, were considered eligible. A detailed mapping of functioning measures by study and domain is presented in **Table 2**. Secondary outcomes included symptom severity and general psychological distress, but these were not synthesized. Data were collected at post-treatment only, defined as the assessment closest to the end of the intervention. Follow-up timepoints were recorded during screening but were not extracted or synthesized.

In addition to outcomes, the following variables were extracted for each trial: authorship and year of publication, country of study, and sample size. Participant characteristics included mean age, percentage of female participants, and diagnostic criteria for personality disorder. Intervention data included treatment type, number of sessions, and duration of treatment. Comparator conditions were described according to how they were defined in the original reports (e.g., treatment as usual, waitlist). Information on dropout and attrition rates was also collected.

## 2.5 Study Risk of Bias Assessment

The methodological quality of the included studies was assessed using the Cochrane Risk of Bias tool, version 2.0 (RoB 2; Sterne et al., 2019), which evaluates potential sources of bias in RCTs across five domains: the randomization process, deviations from the intended interventions, completeness of outcome data, measurement of the outcomes, and selection of the reported results. The assessment was conducted by a single reviewer under supervision. Independent duplicate assessment was not performed.

For each domain, signalling questions were answered in the official RoB 2 Excel workbook, and judgements of low risk, some concerns, or high risk were derived using the Metapsy Risk of Bias automated scoring procedure (Miguel et al., 2025), which implements the RoB 2 decision rules. The Excel sheet was uploaded to the Metapsy interface, which computed domain-level and overall judgements. RoB was assessed at the study level (one assessment per trial rather than outcome-specific ratings) and was based primarily on the main report of each trial. No adaptations were made to the RoB 2 tool.

## 3. RESULTS:

### 3.1 Preliminary outcome analysis

A preliminary mapping of outcomes across the 32 included trials shows that the most frequently assessed domains were functioning (14 studies), psychiatric symptoms (12), behavioral symptoms (10), personality-disorder (PD) symptoms (10), cognitive patterns (9), psychological distress (7), self-directed risk behaviors (5), and substance use (5); less common were family functioning, school functioning, aggression, and antisocial behavior (each 3 studies), with other categories appearing only sporadically. Within functioning, subdomains most often targeted were social/role functioning (6 studies), interpersonal functioning (4), and global functioning (2). Across domains, the most frequently used instruments were SCL-90-R and BDI (each in 5 trials) and AUDIT (3 trials). Follow-up assessments were reported in 20/32 trials (~63%); considering the longest interval per trial, follow-ups clustered at >6–12 months (5 trials) and >24 months (5), with smaller numbers at >3–6 months (4), ≤3 months (3), >12–24 months (1), and two trials specifying follow-up without a precise interval.

### 3.2 Study Selection

The electronic database search yielded a total of 4,914 records (PubMed = 1,787; Embase = 2,324; APA PsycInfo = 803). After removal of 1,840 duplicates, 3,074 titles

and abstracts were screened, of which 2,284 were excluded. We sought full texts for 790 reports, but 58 could not be retrieved, leaving 732 reports assessed for eligibility. Of these, 719 were excluded for the following reasons: not a randomized controlled trial (n = 49), ineligible comparator (n = 91), ineligible population (n = 96), ineligible intervention (n = 76), full text not available in English (n = 9), not a primary trial report (n = 367), functioning outcomes not assessed (n = 16), or other reasons (n = 15). Ultimately, 13 trials met the inclusion criteria and were included in the review. The study selection process is presented in the PRISMA 2020 flow diagram (**Figure 1**)

### 3.3 Description of included studies

Thirteen randomized controlled trials were included. Studies were conducted in Canada (n=3), the Netherlands (n=2), Italy (n=1), the United Kingdom (n=5), and the United States (n=2), with a total randomized sample of N=2,008 (range 27–495 per trial). Reported mean ages spanned 21.4–40.3 years, and the percentage of women ranged from 0–100%. Diagnoses comprised borderline personality disorder (BPD; 6 trials), any personality disorder (4), avoidant personality disorder (1), antisocial personality disorder (1), and mixed non-BPD personality disorders (1). Diagnosis was established using DSM III and DSM IV instruments, primarily SCID II, IPDE, DIPD IV, and MCMI, across all 13 trials. No trial used the DSM-5 Alternative Model for Personality Disorders for eligibility or outcome measurement (see **Table 1**).

Experimental interventions included dialectical behavior therapy (comprehensive and skills-only), schema therapy (individual; combined individual+group; predominantly group), cognitive behavioral therapy, interpersonal psychotherapy adapted for BPD, short-term dynamic psychotherapy (ISTDP), structured psychological support, psychoeducation with problem solving (PEPS), STEPPS, and a web-based psychoeducation program. Formats were individual, group, mixed, or web-based. Comparators were wait-listed in three trials and treatment as usual (TAU) in nine, typically unstructured; one trial specified optimal TAU (Arntz, 2022). In addition, one study used monitoring-only rather than TAU/WL (Zanarini, 2018). Planned dose ranged from 10 weeks of group programs to 12 months (e.g., CBT/DBT) and 2 years (schema therapy); frequency was most often weekly, with one trial scheduled two times per week (Arntz, 2022).

The included trials were heterogeneous with respect to population (BPD-only vs mixed PD vs specific PDs such as AvPD or ASPD), interventions (comprehensive vs skills-only DBT, schema therapy variants, CBT, interpersonal or psychodynamic approaches, psychoeducation), format (individual, group, mixed, or web-based), dose (from 2 sessions to 2 years), and comparators (wait-list, unstructured TAU, optimal TAU, monitoring-only). These results are presented in **Table 1**.

### 3.4 Quality of the included studies

Risk of bias was assessed at the trial level for all 13 randomized controlled trials. Overall, 9/13 (69%) were rated high risk and 4/13 (31%) some concerns; none were low risk. The main drivers were non-ITT or post-randomization exclusions (D2: deviations/analysis) and incomplete outcome data with limited handling (D3). The randomization process (D1) was generally acceptable, although allocation concealment was often unreported. Outcome measurement (D4) was largely appropriate; higher risk arose when assessor-rated outcomes were collected without blinding. Selection of the reported result (D5) was typically judged with some concerns, reflecting limited prospective registration or prespecified protocol/SAP; registration was reported in recent trials (Arntz, 2022; Bamelis, 2014; Crawford, 2020; McMain, 2017; McMurrin, 2017; Zanarini, 2018). Overall RoB judgments per study are reported in **Table 1**; domain-level ratings are provided in **Table 4**.

### 3.5 Functioning Outcomes

Across the 13 trials, functioning was assessed in three prespecified domains. Global functioning was assessed in seven studies and was most often measured with GAF/GAS or SOFAS/WSAS. Interpersonal functioning was assessed in eight studies, typically using SAS/SAS-SR or SFQ, with IIP-32 in one study. Self-functioning was assessed in one study via SPSI-R (Huband, 2007). Several trials measured more than one domain; the most common pairing was global and interpersonal (Alden, 1989; Blum, 2008). One study (Alden, 1989) used a trial-specific interpersonal composite (social-contact frequency and satisfaction, weekly social-task progress, interview

ratings). Quality-of-life measures were excluded by design. No trial assessed the Level of Personality Functioning Scale (LPFS) (Bender, 2011). The study-by-study mapping of domains and instruments is reported in **Table 2**.

Symptom outcomes at post-treatment were typically assessed with borderline-specific scales (e.g., BPDSI, BSL-23, BEST, ZAN-BPD), measures of general psychopathology or distress (e.g., SCL-90/-R, BSI, OQ-45.2, CORE-OM symptoms), and clinician global ratings (CGI). Given the heterogeneity of instruments and scaling, no quantitative synthesis was performed.

## 4. DISCUSSION

### 4.1 Preliminary outcome analysis

The preliminary outcome mapping was a scoping exercise to describe which outcome domains and instruments were used and to help set the focus on functioning. Counts were descriptive and based on the consecutively coded trials available at that time (n = 32); they do not indicate effect size or importance. Functioning appeared in 14 of 32 trials, which supported centering the synthesis on functioning. Because outcome labels differed across studies and several instruments occurred only once, subdomains were classified conservatively. Constructs were merged only when instrument content clearly matched, for example, IIP or IIP-C as interpersonal, SAS, SAS-SR, SFQ, or UPSA as social or role, and SOFAS or GAF as global, to avoid unnecessary fragmentation.

### 4.2 Study characteristics

The current evidence base is limited in scope and generalizability. Thirteen randomized controlled trials (total N = 2,008; range 27–495 per trial) were conducted exclusively in Western settings (Canada, the Netherlands, Italy, the United Kingdom, and the United States), constraining external validity across other health systems and cultural contexts. Samples were predominantly young to mid-adulthood (mean ages 21.4–40.3 years), and the proportion of women varied widely (0–100%), limiting sex- and age-specific inferences. Diagnostic coverage was uneven: borderline personality disorder (BPD)

dominated (6/13), followed by any personality disorder samples (4/13), with single-trial coverage for avoidant PD (1/13), antisocial PD (1/13), and mixed non-BPD PDs (1/13). Consequently, conclusions are driven primarily by BPD evidence.

Diagnoses relied on older classification systems. Across all 13 trials, eligibility was established using DSM-III or DSM-IV instruments (for example, SCID-II, IPDE, DIPD-IV, MCMI). No trial used the DSM-5 Alternative Model for Personality Disorders, such as the Level of Personality Functioning or SCID-AMPD. More than a decade after DSM-5's publication, this gap limits alignment between inclusion criteria and the functioning-centered outcomes emphasized in this review, and it prevents tracking remission within a unified AMPD framework of self and interpersonal functioning. Future randomized trials may consider specifying eligibility using AMPD (or ICD-11 severity) and include validated LPFS measures at post-treatment and follow-up.

Clinical and methodological heterogeneity was substantial. Interventions differed in model (comprehensive vs skills-only dialectical behavior therapy; multiple schema-therapy formats; cognitive-behavioral therapy; interpersonal psychotherapy adapted for BPD; short-term psychodynamic approaches such as ISTDP; structured psychological support; psychoeducation, including a web-based program). Delivery formats varied (individual, group, mixed, and web-based), and settings were mostly outpatient, with therapist training and fidelity procedures inconsistently reported. Planned dose ranged widely: from brief psychoeducation/skills programs over ~8–12 weeks to year-long CBT/DBT and two-year schema therapy.

## 4.3 Comparator issues

Comparator selection is a key design decision in psychosocial randomized controlled trials and shapes both the size and the meaning of observed effects. Across personality-disorder trials, treatment as usual is highly heterogeneous and often poorly specified, ranging from minimal clinical monitoring to intensive multidisciplinary care. Prior reviews, for example, Duggan (2007), note that such variability can bias comparative effects in either direction. Usual care may include techniques clinicians are particularly practiced in, which can favour the control condition, or it may reflect low engagement and sparse contact, which can exaggerate effects for the experimental intervention. A recent commentary in *World Psychiatry* synthesizes recurring problems with control conditions in psychosocial trials (Cuijpers, 2025). Wait-list designs often show larger effects than usual-care comparators and can overestimate intervention benefits, while usual care varies substantially across settings and countries, which limits generalizability. Psychological placebos may function as credible and therefore active treatments, or, if non-credible, may demoralize participants. Because each control option carries distinct risks, the comparator is best justified in context, with clear reporting of its content and of how potential biases were mitigated.

In this review, inactive comparators were defined as wait-list or treatment-as-usual that did not constitute a bona fide manualized psychotherapy. Treatment as usual was understood as routine clinical care that could include medication management, clinical monitoring, supportive visits, or structured pharmacotherapy algorithms, provided these elements were applied equally across arms. Add-on designs were eligible only when

both groups received the same background treatment, and randomization concerned the psychotherapy component. Under this definition, treatment as usual was treated as an inactive comparator for analytic purposes while acknowledging its heterogeneity across settings. The components and delivery of treatment as usual and wait-list comparators are summarized in **Table 3**. Applying this stricter definition produced a focused evidence base of 13 trials and led to the exclusion of 13 otherwise eligible reports because the control condition delivered psychotherapy or structured psychosocial programmes despite being labelled treatment as usual or inactive.

Illustrative examples:

- **Weinberg, 2006:** participants were already in ongoing individual and or group psychotherapy within treatment as usual when randomized to MACT or no MACT, which makes the comparator psychotherapy-as-usual.
- **Farrell, 2009:** controls received ongoing weekly individual psychotherapy from community clinicians, a bona fide psychotherapy rather than inactive treatment as usual.
- **Feigenbaum, 2012:** treatment as usual included psychoanalytic psychotherapy, CBT, and supportive counselling, indicating an active psychotherapy control.
- **Hilden, 2021:** treatment as usual included regular individual therapy and, for some, DBT group, with higher therapy exposure than the experimental arm; hence ineligible as inactive treatment as usual.

In fields with very large randomized evidence bases, such as depression, the added value of new control-group trials may be limited, and emphasis may shift toward direct comparative evaluations or work that isolates active components (Cuijpers, 2025). In personality-disorder psychotherapy, where trials are fewer and more heterogeneous, these directions remain future-oriented at this stage. Accordingly, the present review confines synthesis to inactive comparators to preserve a more homogeneous reference condition and to support clearer attribution of effects to the psychotherapy component. Trials with active comparators identified during full-text screening are retained in the broader database for future analyses but are not included in the current synthesis.

## 4.4 Risk of Bias

The risk-of-bias profile across trials was generally negative, with most studies judged at high risk. The most frequent problems were post-randomization exclusions and incomplete handling of missing data, both of which can systematically distort effect estimates. Selective reporting was another recurring issue, as only a minority of studies had a prospectively registered protocol or prespecified analysis plan. The randomization process itself appeared broadly adequate, although reporting was often incomplete, particularly regarding allocation concealment. Measurement of outcomes was usually appropriate, but the lack of consistent assessor blinding in some trials left residual risk of detection bias. Taken together, these weaknesses reduce confidence in the reliability of trial findings and may contribute to an overestimation of treatment effects.

Nonetheless, the fact that more recent trials adopted registration and clearer

methodological safeguards points to a gradual improvement in study quality, highlighting the direction in which future research should continue.

## 4.5 Strengths and limitations

Screening, extraction, and risk-of-bias procedures were not fully duplicated. Titles and abstracts were screened in duplicate with consensus resolution. Full-text eligibility assessment was conducted by a single reviewer (with ad-hoc consultation in unclear cases) and was not independently duplicated; therefore, occasional missed or misclassified reports cannot be excluded. Data extraction and RoB 2 assessments were completed by a single reviewer using a piloted template; internal consistency checks (cross-checking companion reports and verifying sample sizes, time points, and comparator labels) were applied, but the absence of independent duplicate extraction and blinded risk-of-bias rating may introduce error and subjective judgment. RoB 2 was conducted at the trial level rather than at the outcome-specific level; domain judgements therefore reflect study-level risk and may over- or underestimate bias for particular outcomes (e.g., functioning vs symptom severity) and time points. Future updates should incorporate independent duplicate full-text screening, paired extraction with verification of key variables, outcome-level RoB 2 assessment where feasible, and consensus RoB assessment.

Follow-up outcomes were not extracted or analyzed in the main synthesis of 13 randomized trials. This choice prioritized comparability across studies, given heterogeneous follow-up schedules and instruments, and kept the scope focused on post-treatment functioning. Nevertheless, longitudinal evidence indicates why follow-ups matter: a ten-year CLPS analysis found a general change trajectory interpretable as personality functioning that tracked reductions in personality disorder severity more strongly than trait-specific change factors, underscoring the value of repeated follow-up assessments for capturing clinically meaningful remission (Hopwood et al., 2025). Future updates of this review will therefore include standardized follow-up endpoints (for example, 6 and 12 months) and functioning measures to characterize longer-term trajectories.

Symptom outcomes (e.g., BPDSI, BSL-23/95, SCL-90-R/BSI, CGI) were extracted at post-treatment but were not reported in the main synthesis, which centers on functioning. This is a limitation because it prevents a direct appraisal of whether symptomatic improvement aligns with functional gains within trials. Future updates will report symptom outcomes alongside functioning and explicitly compare their concordance.

## 5. CONCLUSIONS

This systematic review mapped how functioning is operationalized and measured in randomized trials of psychotherapies for adult personality disorders under inactive comparators. The review addresses a gap in the literature by focusing explicitly on functioning within contemporary dimensional frameworks that place impairment in self and interpersonal domains at the core of diagnosis.

The evidence base is modest and uneven. Thirteen trials were eligible, conducted exclusively in Western health systems, with samples dominated by borderline personality disorder and far fewer studies in other diagnoses. Interventions and formats varied widely in model, intensity, and delivery, and comparators included wait-list and heterogeneous forms of treatment as usual. This heterogeneity limits generalizability and complicates cross-trial comparison.

Functioning outcomes are common but inconsistently defined. Trials classified outcomes as global, interpersonal, or self-functioning, yet instruments and timing varied, and alignment with the Alternative Model for Personality Disorders or ICD-11 was rare. No included trial used AMPD Level of Personality Functioning measures, and most trials established eligibility with DSM III or DSM IV instruments. These choices constrain inferences about change in personality functioning as conceived by current nosology. Methodological quality tempers confidence in the findings. Most trials were judged at high risk of bias, with recurrent problems in post-randomization handling and missing data, and with limited prospective registration. None met a low-risk profile. Under these

conditions, a qualitative synthesis was appropriate and statistical pooling was not pursued.

Comparator selection remains a central challenge. The review applied a strict definition of inactive comparators to avoid conflating usual care with bona fide psychotherapy. It excluded several studies where controls received structured psychosocial treatments despite inactive labels. Clearer specification and justification of usual care, as well as careful reporting of its intensity and content, are required to interpret effects on functioning.

These findings support several priorities for future trials. Eligibility and outcomes should be anchored in AMPD or ICD-11, with routine inclusion of validated measures of personality functioning at post-treatment and follow-up. Trials should preregister protocols with prespecified functioning endpoints, use blinded outcome assessment when feasible, and plan sample sizes to detect change in functioning. Usual care comparators should be described in sufficient detail to judge credibility and dose, and wait-list designs should be used sparingly, given concerns about inflated effects. Multi-site collaboration and standardized follow-up time points would improve precision and external validity.

The review has limitations that should guide interpretation. Full text eligibility, data extraction, and risk of bias assessment were completed by a single reviewer; risk of bias was assessed at the study level rather than outcome level, and synthesis was confined to post-treatment endpoints. These constraints may introduce error and preclude conclusions about the durability of change.

In conclusion, psychotherapy trials for personality disorders frequently assess functioning, but operationalization is heterogeneous and seldom aligned with contemporary dimensional models. Paired with a notable risk of bias and reliance on variably specified usual care, this limits confident inference. A next generation of trials that adopt AMPD or ICD-11-based eligibility and outcomes, preregistered analyses, transparent comparators, and standardized follow-up will be able to test whether improvements in self and interpersonal functioning are consistent, clinically meaningful, and durable across diagnoses.

# REFERENCES

- Abbass, A., Sheldon, A., Gyra, J., & Kalpin, A. (2008). Intensive short-term dynamic psychotherapy for DSM-IV personality disorders: A randomized controlled trial. *The Journal of Nervous and Mental Disease*, 196(3), 211–216. <https://doi.org/10.1097/NMD.0b013e3181662ff0>
- Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of Anesthesiology*, 71(2), 103–112. <https://doi.org/10.4097/kjae.2018.71.2.103>
- Alden, L. (1989). Short-term structured treatment for avoidant personality disorder. *Journal of Consulting and Clinical Psychology*, 57(6), 756–764. <https://doi.org/10.1037/0022-006X.57.6.756>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.
- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.; DSM-5-TR). <https://doi.org/10.1176/appi.books.9780890425787>
- Arntz, A., Jacob, G. A., Lee, C. W., Brand-de Wilde, O. M., Fassbinder, E., Harper, R. P., Lavender, A., Lockwood, G., Malogiannis, I. A., Ruths, F. A., Schweiger, U., Shaw, I. A., Zarbock, G., & Farrell, J. M. (2022). Effectiveness of predominantly group schema therapy and combined individual and group schema therapy for borderline personality disorder: A randomized clinical trial. *JAMA Psychiatry*, 79(4), 287–299. <https://doi.org/10.1001/jamapsychiatry.2022.0010>
- Bamelis, L. L., Evers, S. M., Spinhoven, P., & Arntz, A. (2014). Results of a multicenter randomized controlled trial of the clinical effectiveness of schema therapy for personality disorders. *The American Journal of Psychiatry*, 171(3), 305–322. <https://doi.org/10.1176/appi.ajp.2013.12040518>
- Bellino, S., Rinaldi, C., & Bogetto, F. (2010). Adaptation of interpersonal psychotherapy to borderline personality disorder: A comparison of combined therapy and single pharmacotherapy. *Canadian Journal of Psychiatry*, 55(2), 74–81. <https://doi.org/10.1177/070674371005500203>
- Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in DSM-5, Part I: A review of theory and methods. *Journal of Personality Assessment*, 93(4), 332–346. <https://doi.org/10.1080/00223891.2011.583808>
- Blum, N., St John, D., Pfohl, B., Stuart, S., McCormick, B., Allen, J., Arndt, S., & Black, D. W. (2008). Systems training for emotional predictability and problem solving (STEPPS) for outpatients with borderline personality disorder: A randomized controlled trial and one-year follow-up. *The American Journal of Psychiatry*, 165(4), 468–478. <https://doi.org/10.1176/appi.ajp.2007.07071079>
- Botham, J., Simpson, A., & McCrone, P. (2024). Mental health service use and costs associated with complex emotional needs and a diagnosis of personality disorder: Analysis of routine data. *BJPsych Bulletin*, 48(2), 85–92. <https://doi.org/10.1192/bjb.2023.41>

- Crawford, M. J., Thana, L., Parker, J., Turner, O., Carney, A., McMurrin, M., Moran, P., Weaver, T., Barrett, B., Roberts, S., Claringbold, A., Bassett, P., Sanatinia, R., & Spong, A. (2020). Structured psychological support for people with personality disorder: Feasibility randomised controlled trial of a low-intensity intervention. *BJPsych Open*, 6(2), e25. <https://doi.org/10.1192/bjo.2020.7>
- Cuijpers, P. (2025). Has the time come to stop using control groups in trials of psychosocial interventions? *World Psychiatry*, 24(3), 436–437. <https://doi.org/10.1002/wps.21359>
- Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology*, 76(6), 909–922. <https://doi.org/10.1037/a0013075>
- Cuijpers, P., van Straten, A., Warmerdam, L., & Andersson, G. (2010). Psychological treatment of depression: A meta-analytic database of randomized studies. *BMC Psychiatry*, 10(1), 23. <https://doi.org/10.1186/1471-244X-10-23>
- Davidson, K., Tyrer, P., Gumley, A., Tata, P., Norrie, J., Palmer, S., Millar, H., Drummond, L., Seivewright, H., Murray, H., & Macaulay, F. (2006). A randomized controlled trial of cognitive behavior therapy for borderline personality disorder: Rationale for trial, method, and description of sample. *Journal of Personality Disorders*, 20(5), 431–449. <https://doi.org/10.1521/pedi.2006.20.5.431>
- Davidson, K. M., Tyrer, P., Tata, P., Cooke, D., Gumley, A., Ford, I., Walker, A., Bezlyak, V., Seivewright, H., Robertson, H., & Crawford, M. J. (2009). Cognitive behaviour therapy for violent men with antisocial personality disorder in the community: An exploratory randomized controlled trial. *Psychological Medicine*, 39(4), 569–577. <https://doi.org/10.1017/S0033291708004066>
- d'Huart, D., Seker, S., Bürgin, D., Birkhölzer, M., Boonmann, C., Schmid, M., & Schmeck, K. (2023). The stability of personality disorders and personality disorder criteria: A systematic review and meta-analysis. *Clinical Psychology Review*, 102, 102284. <https://doi.org/10.1016/j.cpr.2023.102284>
- Duggan, C., Huband, N., Smailagic, N., Ferriter, M., & Adams, C. (2007). The use of psychological treatments for people with personality disorder: A systematic review of randomized controlled trials. *Personality and Mental Health*, 1(2), 95–125. <https://doi.org/10.1002/pmh.22>
- Eriksen, M. B., & Frandsen, T. F. (2018). The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: A systematic review. *Journal of the Medical Library Association*, 106(4), 420–431. <https://doi.org/10.5195/jmla.2018.345>
- Farrell, J. M., Shaw, I. A., & Webber, M. A. (2009). A schema-focused approach to group psychotherapy for outpatients with borderline personality disorder: A randomized controlled trial. *Journal of Behavior Therapy and Experimental Psychiatry*, 40(2), 317–328. <https://doi.org/10.1016/j.jbtep.2009.01.002>
- Feigenbaum, J. D., Fonagy, P., Pilling, S., Jones, A., Wildgoose, A., & Bebbington, P. E. (2012). A real-world study of the effectiveness of DBT in the UK National Health Service. *The British Journal of Clinical Psychology*, 51(2), 121–141. <https://doi.org/10.1111/j.2044-8260.2011.02017.x>

- Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
- Hilden, H. M., Rosenström, T., Karila, I., Elokorpi, A., Torpo, M., Arajärvi, R., & Isometsä, E. (2021). Effectiveness of brief schema group therapy for borderline personality disorder symptoms: A randomized pilot study. *Nordic Journal of Psychiatry*, 75(3), 176–185. <https://doi.org/10.1080/08039488.2020.1826050>
- Hopwood, C. J. (2024). Personality functioning, problems in living, and personality traits. *Journal of Personality Assessment*, 107(2), 143–158. <https://doi.org/10.1080/00223891.2024.2345880>
- Hopwood, C. J., Driver, C. A., Morey, L. C., & Skodol, A. E. (2025). Personality functioning as generalized correlated changes in personality traits [Preprint]. *PsyArXiv*. <https://doi.org/10.31234/osf.io/n63cg>
- Huband, N., McMurrin, M., Evans, C., & Duggan, C. (2007). Social problem-solving plus psychoeducation for adults with personality disorder: Pragmatic randomised controlled trial. *The British Journal of Psychiatry*, 190(1), 307–313. <https://doi.org/10.1192/bjp.bp.106.023341>
- Hutsebaut, J., & Sharp, C. (2024). Opportunities for the AMPD: Commentary on Hopwood, 2024. *Journal of Personality Assessment*, 107(2), 159–163. <https://doi.org/10.1080/00223891.2024.2430321>
- Kiel, L., Lind, M., Bo, S., Jørgensen, C. R., Bøye, R., Frederiksen, C. K., & Spindler, H. (2025). Associations between pathological personality traits, functional impairment, and personality disorder: Controlling for basic personality traits and identity disturbance. *Personality Disorders: Theory, Research, and Treatment*. Advance online publication. <https://doi.org/10.1037/per0000731>
- Kvarstein, E. H., Antonsen, B. T., Pedersen, G., Folmo, E., Urnes, Ø., Schlüter, C., Hummelen, B., Wilberg, T., & Johansen, M. S. (2023). Improvement of personality functioning among people with personality disorders receiving treatment. *Frontiers in Psychiatry*, 14, 1163347. <https://doi.org/10.3389/fpsy.2023.1163347>
- McClelland, H., Cleare, S., & O'Connor, R. C. (2023). Suicide risk in personality disorders: A systematic review. *Current Psychiatry Reports*, 25(9), 405–417. <https://doi.org/10.1007/s11920-023-01440-w>
- McMain, S. F., Guimond, T., Barnhart, R., Habinski, L., & Streiner, D. L. (2017). A randomized trial of brief dialectical behaviour therapy skills training in suicidal patients suffering from borderline disorder. *Acta Psychiatrica Scandinavica*, 135(2), 138–148. <https://doi.org/10.1111/acps.12664>
- McMurrin, M., Day, F., Reilly, J., Delport, J., McCrone, P., Whitham, D., Tan, W., Duggan, C., Montgomery, A. A., Williams, H. C., Adams, C. E., Jin, H., Moran, P., & Crawford, M. J. (2017). Psychoeducation and problem solving (PEPS) therapy for adults with personality disorder: A pragmatic randomized controlled trial. *Journal of Personality Disorders*, 31(6), 810–826. <https://doi.org/10.1521/pedi.2017.31.286>
- Miguel, C., Harrer, M., Karyotaki, E., Sahker, E., Sakata, M., Furukawa, T., & Cuijpers, P. (2025). Operationalization of Cochrane's risk of bias 2 tool (RoB 2) in the context of psychotherapy trials. *medRxiv*. <https://doi.org/10.1101/2025.06.26.25330349>

Natoli, A. P., Murdock, J. G., Merguie, J. L., & Hopwood, C. J. (2025). Dimensional models of personality and a multidimensional framework for treating personality pathology. *BJPsych Advances*, 31(3), 137–146. <https://doi.org/10.1192/bja.2024.55>

Nejadghaderi, S. A., Balibegloo, M., & Rezaei, N. (2024). The Cochrane risk of bias assessment tool 2 (RoB 2) versus the original RoB: A perspective on the pros and cons. *Health Science Reports*, 7(6), e2165. <https://doi.org/10.1002/hsr2.2165>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>

Sinnaeve, R., van den Bosch, L. M. C., & van Steenbergen-Weijenburg, K. M. (2015). Change in interpersonal functioning during psychological interventions for borderline personality disorder: A systematic review of measures and efficacy. *Personality and Mental Health*, 9(3), 173–194. <https://doi.org/10.1002/pmh.1290>

Skodol, A. E. (2018). Impact of personality pathology on psychosocial functioning. *Current Opinion in Psychology*, 21, 33–38. <https://doi.org/10.1016/j.copsyc.2017.09.006>

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., Reeves, B. C., Shepperd, S., Shrier, I., Stewart, L. A., Tilling, K., White, I. R., Whiting, P. F., & Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, l4898. <https://doi.org/10.1136/bmj.l4898>

Tyrer, P., Mulder, R., Crawford, M., Newton-Howes, G., Simonsen, E., Ndeti, D., Koldobsky, N., Fossati, A., Mbatia, J., & Barrett, B. (2010). Personality disorder: A new global perspective. *World Psychiatry*, 9(1), 56–60. <https://doi.org/10.1002/j.2051-5545.2010.tb00270.x>

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions? A question for DSM-V. *Journal of Abnormal Psychology*, 114(4), 494–504. <https://doi.org/10.1037/0021-843X.114.4.494>

Winsper, C., Bilgin, A., Thompson, A., Marwaha, S., Chanen, A. M., Singh, S. P., Wang, A., & Furtado, V. (2020). The prevalence of personality disorders in the community: A global systematic review and meta-analysis. *The British Journal of Psychiatry*, 216(2), 69–78. <https://doi.org/10.1192/bjp.2019.166>

World Health Organization. (2019). *International classification of diseases for mortality and morbidity statistics* (11th revision). World Health Organization. <https://icd.who.int/>

Zanarini, M. C., Conkey, L. C., Temes, C. M., & Fitzmaurice, G. M. (2018). Randomized controlled trial of web-based psychoeducation for women with borderline personality disorder. *The Journal of Clinical Psychiatry*, 79(3), 16m11153. <https://doi.org/10.4088/JCP.16m11153>

Zavlis, O., & Fonagy, P. (2024). Beyond descriptive models of personality problems. *Journal of Personality Assessment*, 107(2), 164–167. <https://doi.org/10.1080/00223891.2024.2430322>

Weinberg, I., Gunderson, J. G., Hennen, J., & Cutter, C. J., Jr. (2006). Manual-assisted cognitive treatment for deliberate self-harm in borderline personality disorder patients. *Journal of Personality Disorders*, 20(5), 482–492. <https://doi.org/10.1521/pedi.2006.20.5.482>

# TABLES

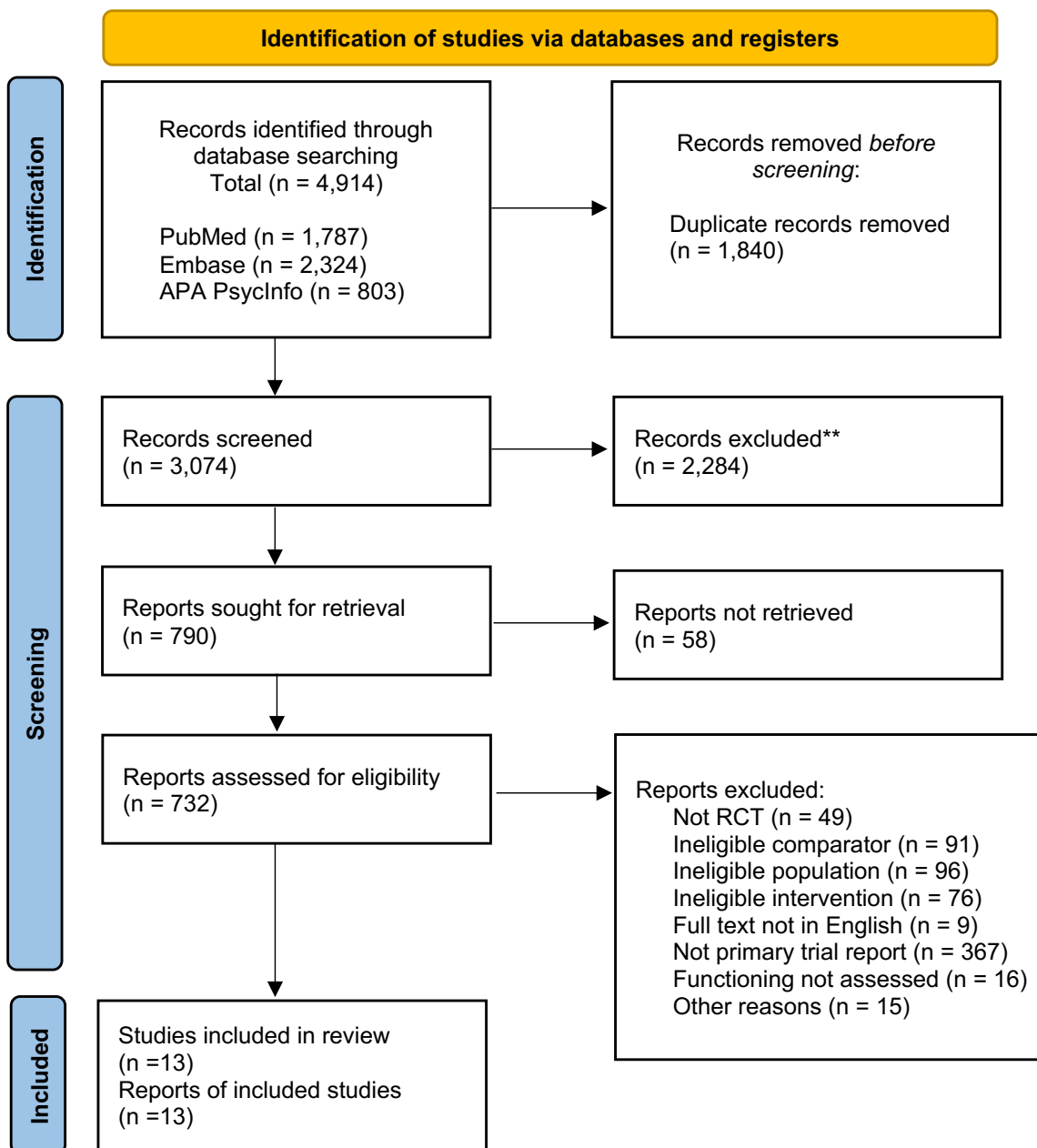


Figure 1: PRISMA 2020 flow diagram for study selection

Study ID	Country	PD diagnosis	Comorbidity criteria	Diagnostic Criteria / Assessment Tool	Sample (N; M_age; %F)	Intervention Type	Dose (duration / sessions; frequency)	Comparator	RoB rating
Abbass, 2008	Canada	Any PD	–	DSM-IV diagnosis (SCIP-PQ and SCID II PQ)	N=27; M_age=40,3; %F=59	Short-term dynamic psychotherapy (ISTDP), individual	open-ended; flexible	WL	High
Alden, 1989	Canada	AvPD	–	DSM-III diagnosis (psychometric testing + MCMI)	N=76; M_age=27,5; %F=45	Graduated exposure, group; Interpersonal skills training, group; Intimacy-focused therapy, group	10 wk; weekly	WL	High
Arntz, 2022	Netherlands	BPD	–	DSM-IV diagnosis (SCID-II)	N=495; M_age=37,3; %F=56	Schema therapy (PGST: predominantly group), mixed; Schema therapy (IGST: individual + group), mixed	2 y; 2x/wk	TAU	Some concerns
Bamelis, 2014	Netherlands	Cluster C (AvPD, DPD, OCPD) + PPD/HPD/NPD	–	DSM-IV diagnosis (SCID-II)	N=323; M_age=38,0; %F=55	Schema therapy (ST), individual; Clarification-oriented psychotherapy (COP), individual	50 sessions; weekly (Y1), monthly boosters (Y2)	TAU	Some concerns
Bellino, 2010	Italy	BPD	–	DSM-IV diagnosis (SCID-II)	N=55; M_age=26,0; %F=67	Interpersonal psychotherapy adapted for BPD (IPT-BPD), individual + fluoxetine	32 wk; weekly	TAU	High
Blum, 2008	USA	BPD	–	DSM-IV diagnosis (SCID-II)	N=165; M_age=31,5; %F=83	Systems training for emotional predictability and problem solving (STEPPS), group + TAU	20 wk; weekly	TAU	High
Crawford, 2020	UK	Any PD	–	ICD-11 diagnosis (SASPD)	N=63; M_age=36,3; %F=68	Structured psychological support (SPS), individual	6–10 sessions; weekly	TAU	High
Davidson, 2006	UK	BPD	–	DSM-IV diagnosis (SCID-II)	N=106; M_age=31,9; %F=89	Cognitive behavioral therapy (CBT), individual + TAU	12 mo; flexible	TAU	High
Davidson, 2009	UK	ASPD	–	DSM-IV diagnosis (SCID-II)	N=52; M_age=37,9; %F=0	Cognitive behavioral therapy (CBT), individual + TAU	12 mo; weekly	TAU	High
Huband, 2007	UK	Any PD	–	DSM-IV diagnosis (IPDE)	N=176; M_age=36,2; %F=51	Problem-solving therapy plus psychoeducation, mixed	19 sessions; weekly	TAU	Some concerns
McMain, 2017	Canada	BPD	Two suicidal and/ or NSSI episodes in the past 5 years, with one occurring within 10 weeks prior to enrolment	DSM-IV diagnosis (SCID-II)	N=84; M_age=29,7; %F=66	Dialectical behavior therapy skills training (DBT skills), group	20 wk; weekly	WL	Some concerns
McMurrin, 2017	UK	Any PD	–	DSM-IV diagnosis (IPDE)	N=306; M_age=38,2; %F=75	Psychoeducation with problem solving (PEPS), mixed	16 sessions; weekly	TAU	High
Zanarini, 2018	USA	BPD	–	DSM-IV diagnosis (DIPD-IV, BPD module)	N=80; M_age=21,4; %F=100	Web-based psychoeducation for borderline personality disorder, web (self-guided)	12 wk; flexible	Monitoring-only	High

**Table 1: Characteristics of included randomized trials.** Abbreviations: AvPD = avoidant personality disorder; BPD = borderline personality disorder; ASPD = antisocial personality disorder; SCID-II = Structured Clinical Interview for DSM-IV Axis II; IPDE = International Personality Disorder Examination; DIPD-IV = Diagnostic Interview for DSM-IV Personality Disorders; SASPD = Standardized Assessment of Severity of Personality Disorder; WL = waiting list; TAU = treatment as usual; wk = weeks; y = years.

Study ID	Global (measures)	Interpersonal (measures)	Self (measures)
Abbass, 2008	(GAF-SO)		
Alden, 1989	(GAF; SQ-Work)	(SQ–Social; Freq/Sat Social; Social Targets; Self-Monitoring; Int. Ratings)	
Arntz, 2022	(SOFAS; WSAS)		
Bamelis, 2014	(GAF; SOFAS; WSAS)		
Bellino, 2010	(SOFAS)		
Blum, 2008	(GAS)	(SAS)	
Crawford, 2020	(WSAS)		
Davidson, 2006		(IIP-32; SFQ)	
Davidson, 2009		(SFQ)	
Huband, 2007		(SFQ)	(SPSI-R)
McMain, 2017		(SAS)	
McMurrin, 2017		(SFQ)	
Zanarini, 2018		(SAS)	

**Table 2: Functioning outcome instruments by domain in included randomized trials.** Abbreviations: GAF = Global Assessment of Functioning; SOFAS = Social and Occupational Functioning Assessment Scale; WSAS = Work and Social Adjustment Scale; GAS = Global Assessment Scale; SAS = Social Adjustment Scale; SFQ = Social Functioning Questionnaire; IIP-32 = Inventory of Interpersonal Problems-32; SPSI-R = Social Problem-Solving Inventory-Revised.

Study ID	Control label	Control description
Abbass, 2008	WL	Wait-list; monthly supportive meetings with site coordinator (psychiatric monitoring permitted); no trial manual or training.
Alden, 1989	WL	No-treatment waiting-list control; participants continued usual external care if any; no trial-delivered psychotherapy.
Arntz, 2022	Optimal treatment	Optimal psychological treatment
Bamelis, 2014	Optimal treatment	Local intake chose treatment by patient needs; no study protocol/trial training; mix of insight-oriented/supportive/CBT/EMDR as usual care.
Bellino, 2010	Pharmacotherapy + clinical management	Fluoxetine (20–40 mg/day) with clinical management; no trial psychotherapy provided to control.
Blum, 2008	TAU (routine)	STEPPS + TAU vs TAU alone; usual care could include individual psychotherapy/medication/case management; no trial protocol for TAU.
Crawford, 2020	TAU delivered by community mental health teams	Routine community mental-health team care: assessment, care-planning, review; referrals as usual; no trial protocol/training.
Davidson, 2006	Standard treatment within NHS	NHS usual care (GP/CMHT contacts; crisis input as needed); control barred from CBT; otherwise services as usual.
Davidson, 2009	Routine care documented after trial exit	NHS routine care documented with CSRI; background services accessible equally across arms; no trial protocol/training.
Huband, 2007	WL plus TAU	Wait-list while continuing usual services (psychiatrist/resident monthly nurse sessions etc.); no STEPPS-like program in control.
McMain, 2017	WL plus TAU	Wait-list for ~5 months; participants remained in usual services; no DBT-skills group until after WL period.
McMurrin, 2017	Standard treatment per NICE guidelines	Standard treatment per NICE via CMHT; time-limited service with referrals as usual; no trial protocol/training.
Zanarini, 2018	Monitoring-only control	Assessment-only control; no treatment at baseline by design; scheduled online assessments only.




**Table 3: Control conditions in included randomized trials.** Abbreviations: WL = waiting list; TAU = treatment as usual; NHS = National Health Service; NICE = National Institute for Health and Care Excellence; CMHT = community mental health team; CSRI = Client Service Receipt Inventory

		Risk of bias domains					
		D1	D2	D3	D4	D5	Overall
Study	Abbass, 2008	+	X	+	X	-	X
	Alden, 1989	-	-	-	+	-	X
	Arntz, 2022	-	+	+	+	+	-
	Bamelis, 2014	-	+	+	+	+	-
	Bellino, 2010	-	X	X	+	-	X
	Blum, 2008	-	+	+	X	-	X
	Crawford, 2020	+	-	X	+	-	X
	Davidson, 2006	+	+	X	+	-	X
	Davidson, 2009	+	+	X	+	-	X
	Huband, 2007	+	+	-	+	-	-
	McMain, 2017	+	+	-	+	-	-
	McMurrin, 2017	+	X	+	+	+	X
	Zanarini, 2018	X	+	+	X	-	X

Domains:

- D1: Bias arising from the randomization process.
- D2: Bias due to deviations from intended intervention.
- D3: Bias due to missing outcome data.
- D4: Bias in measurement of the outcome.
- D5: Bias in selection of the reported result.

Judgement

-  High
-  Some concerns
-  Low

**Table 4: Risk of bias judgments (RoB 2) across included randomized trials.** Note. D1 randomization process; D2 deviations from intended interventions and analysis; D3 missing outcome data; D4 measurement of outcomes; D5 selection of the reported result. Green = low risk; yellow = some concerns; red = high risk.