Heterogeneity of outcome measures in depression trials and the relevance of the content of outcome measures to patients: a systematic review

Christopher Veal, Anneka Tomlinson, Andrea Cipriani, Samuel Bulteau, Chantal Henry, Chlöé Müh, Suzanne Touboul, Nikki De Waal, Hana Levy-Soussan, Toshi A Furukawa, Eiko I Fried, Viet-Thi Tran, Astrid Chevance

Research waste occurs when randomised controlled trial (RCT) outcomes are heterogeneous or overlook domains that matter to patients (eg, relating to symptoms or functions). In this systematic review, we reviewed the outcome measures used in 450 RCTs of adult unipolar and bipolar depression registered between 2018 and 2022 and identified 388 different measures. 40% of the RCTs used the same measure (Hamilton Depression Rating Scale [HAMD]). Patients and clinicians matched each item within the 25 most frequently used measures with 80 previously identified domains of depression that matter to patients. Seven (9%) domains were not covered by the 25 most frequently used outcome measures (eg, mental pain and irritability). The HAMD covered a maximum of 47 (59%) of the 80 domains that matter to patients. An interim solution to facilitate evidence synthesis before a core outcome set is developed would be to use the most common measures and choose complementary scales to optimise domain coverage.

Lancet Psychiatry 2024; 11: 285–94

For the French translation of the abstract see Online for appendix 1

For the Dutch translation of the abstract see Online for appendix 2

Université Paris Cité and Université Sorbonne Paris Nord, INSERM INRAE, Centre for Research in Epidemiology and Statistics, Paris, France (C Veal MSc, Prof V-T Tran MD PhD, A Chevance MD PhD)**; Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, Paris, France** (C Veal, Prof V-T Tran, A Chevance)**; Department of Psychiatry, University of Oxford, Oxford, UK** (A Tomlinson MD PhD, Prof A Cipriani MD PhD)**; Oxford Precision Psychiatry Lab, NIHR Oxford Health Biomedical Research Centre, Oxford, UK** (Prof A Cipriani)**; Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, UK** (Prof A Cipriani MD PhD)**; UMR INSERM 1246, SPHERE, University of Nantes and University of Tours, Nantes, France** (S Bulteau MD PhD)**; CHU Nantes, Department of Addictology, Psychiatry and Old Age Psychiatry, Nantes, France** (S Bulteau)**; Université Paris Cité, Paris, France** (Prof C Henry MD PhD)**; Department of Psychiatry, Service Hospitalo-Universitaire, GHU Paris Psychiatrie and Neurosciences, Paris, France** (Prof C Henry)**; Perception and Memory Unit, Institut Pasteur, UMR3571, CNRS, Paris, France** (C Müh MSc)**; Université Paris Cité, Collège Doctoral, Paris, France** (C Müh)**; Paris, France** (S Touboul PhD)**; Amsterdam, Netherlands** (N De Waal MSc)**; La Maison Perchée, Paris, France** (H Levy-Soussan MSc)**; Department of Health Promotion and Human Behavior, Kyoto University**

Introduction

Depression affects approximately 300 million adults worldwide and is one of the most common causes of years lived with disability.[1] The clinical effects of hundreds of interventions, such as psychotherapies, drugs, neurostimulation, physical activity, and complementary and alternative medicines, are assessed via randomised controlled trials (RCTs).[2] The efficacy of interventions is assessed by measuring outcomes, also known as endpoints. Outcomes are defined by a domain (what is measured) and a corresponding measure (how an outcome is measured).[3] For example, the outcome of response to treatment based on the depression severity domain is assessed with the Montgomery-Åsberg Depression Rating Scale (MADRS).[3] The selection of outcomes in RCTs is important for four reasons. First, it is crucial to understand whether a patient's condition improves specifically because of the efficacy of the intervention. Second, it is not feasible, for patients and trialists, to assess a large number of outcomes. Third, the efficacy of the intervention will be inferred from the primary outcome, which determines the statistical hypothesis to be tested and the sample size of patients to be recruited. Finally, the choice of outcomes establishes the possibility of the trial being included in evidence synthesis, and thus, will support decisions for market access, reimbursement, or recommendation in clinical guidelines.

Using heterogeneous outcomes across RCTs is a potential source of research waste because it limits the comparison and combination of their results in meta-analyses.[4,5] We hypothesise that research waste might occur in depression RCTs, given that 280 depression measures have been identified, many of which differ substantially in content.[6,7] Outcome heterogeneity, both in domains and measures, was also found in RCTs of adolescent depression (118 measures in 32 trials) and depression in older adults (135 measures in 49 trials).[8,9] A second problem arises from domains that matter to patients in RCTs being overlooked, which limits the usefulness of their results to support therapeutic and regulatory decisions.[2,4,5,10–14]

In a previous study, PROCEED,[14] we obtained free-text responses about different aspects of depression from 1912 people with depression, 624 clinicians, and 464 informal caregivers from 52 countries. PROCEED identified a list of 80 outcome domains that matter to patients, including 64 clinical symptoms (eg, fatigue, insomnia, and anxiety) and 16 functioning-related dimensions (eg, sick leave and difficulties in parenting).[14] Some of the domains identified by PROCEED are not currently assessed by the measures commonly used in RCTs. For example, mental pain was the third most common domain mentioned,[14] yet it is not evaluated by some common depression measures.[6] In this systematic review, we aimed to identify outcome measures reported in phase 3 and 4 RCTs of adults with depression registered between 2018 and 2022 and to evaluate to what extent the content of these outcome measures matters to patients. The study recognises that descriptions surrounding people with a lived experience of a mental disorder often uses insensitive terminology that might have a stigmatising connotation. For consistency with the biomedical literature and for brevity, in this systematic review we use the word patients to refer to those with lived experience of depression, including the co-researchers who are also co-authors of this study.

Methods

Search strategy and selection criteria

We investigated the heterogeneity of the outcome measures used to evaluate the efficacy of therapeutic interventions for depression by conducting a systematic review of RCT protocols published in literature databases and RCT registries. We then evaluated whether the content of the retrieved measures overlapped with domains that matter to patients, as identified in the PROCEED study.[14]

Graduate School of Medicine/ School of Public Health, Kyoto, Japan (Prof T A Furukawa MD); **Clinical Psychology Unit, Psychology Department, Leiden University, Leiden, Netherlands** (E I Fried PhD)

Correspondence to: Dr Astrid Chevance, Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, Paris 75004, France
astrid.chevance@gmail.com

For the **detailed methods** see https://osf.io/8bde3/

See **Online** for appendix 3

We focused on efficacy-related outcomes, the choice of which is crucial when planning an RCT. We did not consider harms-related outcomes (eg, adverse events of drugs) because they are systematically collected in the course of an RCT and, thus, there is no selection to make during planning. Moreover, we focused on clinical and functional outcomes—ie, outcomes related to death, psychiatric symptoms, life impact, and resource use—because they are directly related to patients' experiences of depression.[15] We excluded biological outcomes (eg, electroencephalograms, arterial pressure, and blood samples) because they are surrogate outcomes.[16,17]

The methods are described in detail in the protocol. The overall design of this two-step study is presented in appendix 3 (pp 1–2). The reporting of this systematic review follows the PRISMA guidelines[18] and the Guidance for Reporting Involvement of Patients and the Public[19] (short form).

We first aimed to systematically review the outcomes measured in RCTs evaluating interventions for adult depression. The full procedure for the search strategy, eligibility criteria of RCTs, and procedures for study selection and data extraction are available in appendix 3 (pp 3–5). In brief, we searched for protocols in two databases (PubMed and Embase) and for records in three registries (the International Clinical Trials Registry Platform, ClinicalTrials.gov, and EU Clinical Trials Register) and 18 pharmaceutical websites.[20] To focus on recent, ongoing, and future trials, we included protocols and entries of phase 3 and 4 RCTs published or recorded between Jan 1, 2018, and Oct 26, 2022. Search terms are given on p 3 of appendix 3. RCTs were eligible if they evaluated the clinical efficacy or effectiveness of therapeutic interventions for depression. In compliance with the pragmatic approach of therapeutic evaluation,[21] we chose broad inclusion criteria with regard to population characteristics, thus reflecting the diversity of people included in RCTs.[21,22] We included all RCTs with adult populations (ie, participants >18 years) with depressive disorders or a major depressive episode within a bipolar disorder as identified by either a formal diagnosis, scores exceeding any threshold on a depression outcome measure, or patient-reported diagnosis of depression. There were no eligibility criteria for studies regarding the outcomes or measures used.

Measures were categorised as objective measures (eg, duration of hospitalisation or death) or subjective measures that rely on human judgement or interpretation (eg, clinical symptoms or satisfaction).[23] Using the US Food and Drug Administration taxonomy, subjective outcomes were categorised as patient-reported outcome measures (PROMs), clinician-reported outcome measures (ClinROMs), observer-reported outcome measures (ObsROMs), or performance outcome measures (PerfOMs; appendix 3 p 6). All descriptive analyses were conducted using R statistical software, version 4.3.1.

Data analysis

We focused on the most frequently used outcome measures in each category (PROMs, ClinROMs, PerfOMs,

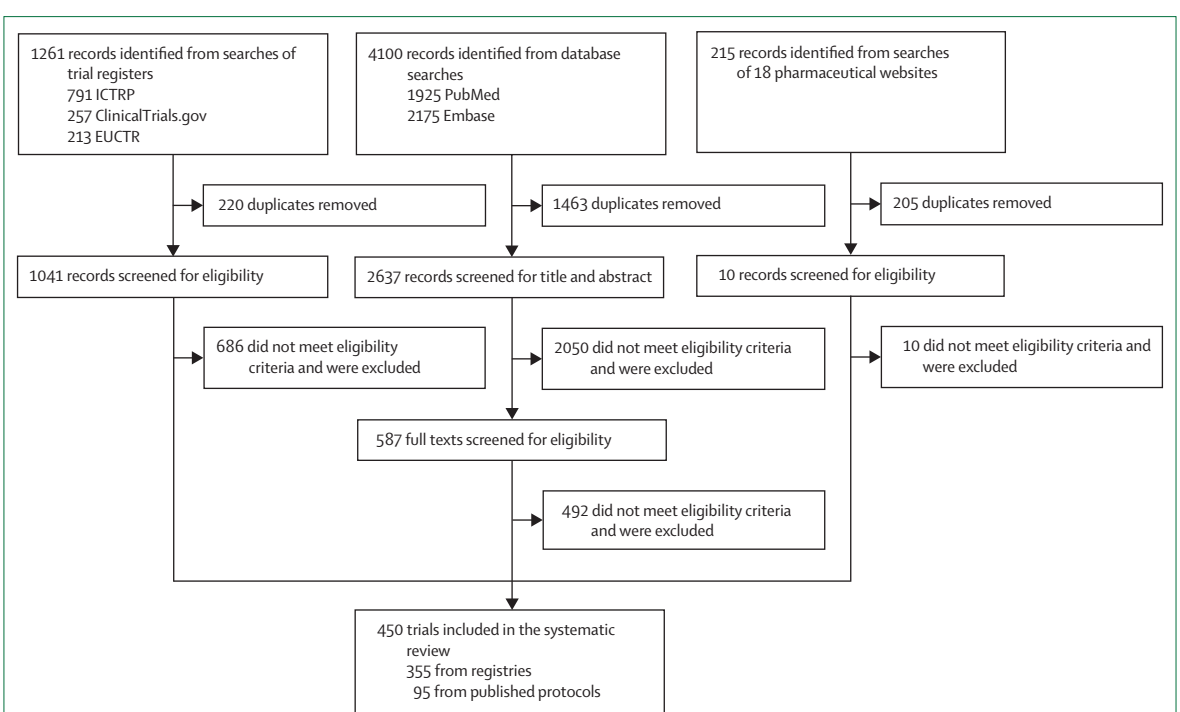

Figure 1: **Study selection**

ObsROMs, and objective outcomes used by more than 1% of trials), up to a maximum of ten per category. We evaluated whether the content of these measures overlapped with the 80 domains that matter to patients,[14] with a matching task involving people with lived experiences of depression, and clinicians. The 80 domains reflect the diversity of symptoms and effects on functioning for patients with unipolar and bipolar depression. As the PROMs and ClinROMs used in depression trials are usually multi-item measures composed of several domains (eg, sad mood, suicidal ideation, and sleep problems),[6] we planned to identify which of the 80 domains, if any, were represented across the individual items of the different measures. Several multi-item measures are based on a latent construct conceptualisation and were developed using item reduction tools to minimise the number of included items. Hence, by design, such measures narrow the scope of symptoms of depression and effects on functioning. Our approach did not consider the latent

	Trials (N=450)
Sex	
Women	44 (10%)
Men	3 (1%)
Both	403 (90%)
Age, categories*	
Adults and adolescents (13–65 years)	11 (2%)
Adults (18–65 years)	269 (60%)
Adults and older adults (>18 years)	149 (33%)
Older adults (>65 years)	10 (2%)
All ages	9 (2%)
Missing data	2 (<1%)
COVID-19	
Before COVID-19 (2018–19)	250 (56%)
During or after COVID-19 (2020–22)	200 (44%)
Primary sponsor, type†	
Universities and research institutes	242 (54%)
Hospitals and private clinics	79 (18%)
Industry	55 (12%)
Foundations and private individuals	36 (8%)
Other	41 (9%)
Missing data	2 (<1%)
Primary sponsor, geographical region†	
Africa	2 (<1%)
Americas	100 (22%)
Asia	241 (54%)
Europe	94 (21)
Oceania	12 (3%)
Missing data	2 (<1%)
Countries, n	
1	410 (91%)
2–5	12 (3%)
6–10	6 (1%)
11–20	9 (2%)
>20	1 (<1%)
Mean (SD)	1·47 (2·44)
Missing data	12 (3%)
Continents, n†	
1	417 (93%)
2	9 (2%)
3–4	12 (3%)
Mean (SD)	1·08 (0·41)
Missing data	12 (3%)

(Table 1 continues in next column)

	Trials (N=450)
(Continued from previous column)	
LMICs	
At least one LMIC included	127 (28%)
No LMIC included	312 (69%)
Missing data	11 (2%)
Sample size, categories	
0–100	202 (45%)
101–250	128 (28%)
251–500	69 (15%)
501–1000	39 (9%)
>1000	10 (2%)
Mean (SD)	213·8 (261·7)
Missing data	2 (<1%)
Trial phase	
Phase 3	326 (72%)
Phase 4	124 (28%)
Depression, type	
Major depressive disorder	316 (70%)
Depressive disorder due to another condition	65 (14%)
Postpartum or perinatal depression	26 (6%)
Bipolar depression	23 (5%)
All types of depression combined	8 (2%)
Depression and anxiety	6 (1%)
Dysthymia	6 (1%)
Intervention, type†	
Drugs	317 (70%)
Other‡	58 (13%)
Neurostimulation	40 (9%)
Psychotherapy	40 (9%)
Psychosocial	37 (8%)
Physical activity	12 (3%)
Dietary	3 (1%)
Placebo or sham	
Used	230 (51%)
Not used	220 (49%)
Groups, number	
2	353 (78%)
3	68 (15%)
4	25 (6%)
5	3 (1%)
14	1 (<1%)

(Table 1 continues in next column)

	Trials (N=450)
(Continued from previous column)	
Control, type†	
Placebo or sham	230 (51%)
Drugs	171 (38%)
Treatment as usual or standard care	40 (9%)
Other‡	31 (7%)
Neurostimulation	27 (6%)
Waiting list or no treatment	28 (6%)
Psychosocial	24 (5%)
Therapy	20 (4%)
Physical activity	6 (1%)
Dietary	2 (<1%)
Conventional treatments, type	
Conventional	403 (90%)
Complementary and alternative medicine	47 (10%)
Protocols available	
No protocol made available in registry	340 (76%)
Protocol made available in registry	110 (24%)
Protocols checked	
Unable to check	340 (76%)
Not checked	108 (24%)
Checked	2 (<1%)

Data are n (%) unless otherwise stated. LMIC=low-income and middle-income countries. *Total percentage at 99·9% due to rounding calculations. †Total exceeds 100% because some studies have more than one primary sponsor type, geographical region, intervention, or control. ‡Examples of other interventions or controls include acupuncture, pharmacogenomic-informed prescribing, stress reduction programmes, gamified smartphone apps based on cognitive behavioural therapy techniques and methods, and music therapy.

Table 1: **Study characteristics**

construct but rather the manifest content of each item of the scales—ie, the semantic significance of the label and of the answers' modalities in the context of the scale. We chose this approach because some measures of severity of depression might have good content validity while overlooking some symptoms and functioning dimensions that might be important to patients who seek treatments and which should, therefore, be measured in RCTs.

The content of each PROM was evaluated by three co-researchers (ST, HL-S, and NDW, from France and the Netherlands) who have lived experience of major depressive disorder (MDD) or bipolar depression. The content of each ClinROM was evaluated by three psychiatrists (AT, CH, and SB, from the UK and France) with expertise in either bipolar depression or depressive disorders and the use of these measures in clinical practice, and by two researchers (ACi and TAF, from the UK and Japan) with experience of trials and meta-analysis of depression. PerfOMs were evaluated by a clinical neuropsychologist (CM, from France). ACh and CV assessed the objective outcomes and facilitated the matching task for all working groups. As ObsROMs were not used in more than five trials, they were not assessed.

To assist the matching task, ACh, CV, and two of the co-researchers with lived experience of depression (ST, who was also a co-researcher for the PROCEED study,[14] and HL-S) developed a depression dictionary (appendix 3 pp 7–23). On the basis of the dataset of the free-speech responses of PROCEED,[14] we developed a definition and a vignette illustrating the experience of a person with depression for each of the 80 domains. First, researchers worked independently to match the content of each PROM or ClinROM at the item level with the 80 domains described in the depression dictionary. To mimic the use of these measures in trials, patients assessed the overlap between the 157 items from the ten most-used PROMs and the 80 domains, and five clinicians assessed the overlap of 94 items from the nine most-used ClinROMs with the 80 domains. Patients and clinicians worked with the full scales of each outcome measure (in paper format or digitised) so that the items were not decontextualised. For the Hamilton Depression Rating Scale (HAMD), the 21-item version was used to obtain the maximum possible coverage. For each item, patients and clinicians were asked to report the different domains of the depression dictionary they considered to be reflected by the label of the item in the general context of the scale. Discrepancies in individual matches for the PROMs and ClinROMs were discussed during a consensus meeting (one for PROMs, involving patients, and one for ClinROMs, involving clinicians). The matches for each item were finalised and approved by all participants in their respective consensus meetings. For instance, after consensus was reached, item 3 of the MADRS (inner tension) was considered as matching three domains of PROCEED, namely anxiety, irritability, and restlessness. A complete description of the matching task is available in the protocol. In addition, the neuropsychologist assessed the content of the six most frequently used PerfOMs for their content overlap with the 80 domains.

Results

The search retrieved 5576 records (figure 1). After removing duplicates, we screened 1041 entries from registries, 2637 records from literature databases, and ten records from pharmaceutical websites. This process led to the final inclusion of 450 unique RCTs, of which 110 (24%) had a published protocol.

Of the 450 RCTs, MDD RCTs were the most common (316 [70%]), whereas 23 (5%) investigated bipolar depression (table 1). Most RCTs were conducted in a single country (410 [91%]), with low-income and middle-income countries included in 127 (28%) studies. The mean total sample size was 214 (SD 262), with 330 (73%) trials including fewer than 250 participants.

Phase 3 trials accounted for 326 (72%) of the 450 RCTs. In these trials, drugs were the most commonly evaluated intervention (200 [61%]), followed by psychotherapy

interventions (38 [12%]), psychosocial interventions (35 [11%]), and neurostimulation (33 [10%]). The primary sponsors of the studies were predominantly universities and research institutes (242 [54%] of 450), followed by hospitals and private clinics (79 [18%]) and industry (55 [12%]).

	Acronym	Measure type	Total use (N=450 RCTs)	Primary outcome (N=450 RCTs)	Secondary outcome (N=450 RCTs)
Hamilton Rating Scale for Depression	HAMD	ClinROM	180 (40%)	156 (35%)	64 (14%)
Montgomery-Åsberg Depression Rating Scale	MADRS	ClinROM	138 (31%)	106 (24%)	81 (18%)
Beck Depression Inventory-II	BDI-II	PROM	81 (18%)	60 (13%)	26 (6%)
Clinical Global Impression-Severity	CGI-S	ClinROM	72 (16%)	6 (1%)	67 (15%)
Patient Health Questionnaire-9	PHQ-9	PROM	63 (14%)	29 (6%)	43 (10%)
Clinical Global Impression-Improvement	CGI-I	ClinROM	43 (10%)	4 (1%)	39 (9%)
General Anxiety Disorder-7	GAD-7	PROM	37 (8%)	3 (1%)	34 (8%)
EuroQol-5 Dimensions-5 Levels	EQ-5D-5L	PROM	35 (8%)	4 (1%)	33 (7%)
Hamilton Anxiety Rating Scale	HAMA	ClinROM	34 (8%)	10 (2%)	25 (6%)
SF-36 Health Survey	SF-36	PROM	34 (8%)	4 (1%)	30 (7%)
Quick Inventory Of Depressive Symptomatology, self-reported	QIDS-SR	PROM	29 (6%)	14 (3%)	23 (5%)
Pittsburg Sleep Quality Index	PSQI	PROM	22 (5%)	7 (2%)	16 (4%)
Sheehan Disability Score	SDS	PROM	19 (4%)	1 (<1%)	18 (4%)
Edinburgh Postnatal Depression Scale	EPDS	PROM	18 (4%)	17 (4%)	3 (1%)
WHOQOL-BREF Questionnaire	WHOQOL	PROM	16 (4%)	3 (1%)	13 (3%)
Columbia Suicide Severity Rating Scale	C-SSRS	ClinROM	16 (4%)	4 (1%)	12 (3%)
Depression Anxiety Stress Scale	DASS	PROM	14 (3%)	6 (1%)	9 (2%)
Beck Scale for Suicide Ideation	BSSI	PROM	13 (3%)	7 (2%)	7 (2%)
Trail Making Task	TMT	PerfOM	13 (3%)	2 (<1%)	12 (3%)
Quick Inventory Of Depressive Symptomatology, clinician-rated	QIDS-C	ClinROM	13 (3%)	5 (1%)	8 (2%)
Beck Anxiety Inventory	BAI	PROM	12 (3%)	9 (2%)	3 (1%)
Montreal Cognitive Assessment	MoCA	PerfOM	12 (3%)	0	12 (3%)
Geriatric Depression Scale	GDS	PROM	11 (2%)	8 (2%)	5 (1%)
Insomnia Severity Index	ISI	PROM	11 (2%)	1 (<1%)	10 (2%)
Work and Social Adjustment Scale	WSAS	PROM	10 (2%)	1 (<1%)	9 (2%)
Snaith-Hamilton Pleasure Scale	SHAPS	PROM	10 (2%)	2 (<1%)	8 (2%)
Health service use	HSU	OOM	9 (2%)	4 (1%)	7 (2%)
Time in hospital	TiH	OOM	9 (2%)	2 (<1%)	8 (2%)
Medication changes in medical record	..	OOM	9 (2%)	0	9 (2%)
Quality of Life Enjoyment and Satisfaction	Q-LES-Q	PROM	9 (2%)	1 (<1%)	8 (2%)
Stroop Color Word Tests	SCWT	PerfOM	8 (2%)	1 (<1%)	8 (2%)
Hospital Anxiety Depression Scale	HADS	PROM	8 (2%)	3 (1%)	5 (1%)
Clinical Global Impression-Bipolar	CGI-BP	ClinROM	8 (2%)	0	8 (2%)
Mini Mental State Examination Scale	MMSE	PerfOM	7 (2%)	3 (1%)	4 (1%)
Centre for Epidemiological Studies Depression Scale	CES-D	PROM	7 (2%)	5 (1%)	2 (<1%)
Numerical pain scale for pain after childbirth	PAIN CBIRTH	PROM	7 (2%)	2 (<1%)	5 (1%)
Patient Global Impressions-Improvement	PGI-I	PROM	7 (2%)	1 (<1%)	6 (1%)
Client Service Receipt Inventory	CSRI	OOM	7 (2%)	0	7 (2%)
Assessment of Quality of Life	AQoL-4D	PROM	7 (2%)	0	7 (2%)
WHO-Five Well-Being Index	WHO-5	PROM	7 (2%)	0	7 (2%)
WHO Disability Assessment Schedule	WHODAS	PROM	6 (1%)	1 (<1%)	6 (1%)
Forward and Backward Digit Span	FBDS	PerfOM	6 (1%)	1 (<1%)	6 (1%)
International Physical Activity Questionnaire	IPAQ	PROM	6 (1%)	0	6 (1%)
Epworth Sleepiness Scale	ESS	PROM	6 (1%)	0	6 (1%)
Digital Symbol Substitution Test	DSST	PerfOM	6 (1%)	0	6 (1%)
Functioning Assessment Short Test	FAST	ClinROM	6 (1%)	2 (<1%)	4 (1%)
Treatment dropout measured by patient interview	..	OOM	6 (1%)	1 (<1%)	5 (1%)

(Table 2 continues on next page)

	Acronym	Measure type	Total use (N=450 RCTs)	Primary outcome (N=450 RCTs)	Secondary outcome (N=450 RCTs)
(Continued from previous page)					
Behavioural Activation for Depression Scale	BADS	PROM	6 (1%)	1 (<1%)	5 (1%)
Rumination Response Scale	RRS	PROM	6 (1%)	1 (<1%)	5 (1%)
Perceived Stress Scale	PSS	PROM	6 (1%)	1 (<1%)	5 (1%)

Data are n (%). Total use indicates the number of unique outcomes used across the 450 RCTs; trials using the same measure as both primary and secondary outcome are included only once under total use. Primary outcome reports the number of RCTs using a measure as a primary outcome; in the case of co-primary outcome measures (90 RCTs), each unique measure is included under primary outcomes. Secondary outcome reports the number of RCTs using a measure as a secondary outcome. ClinROM=clinician-reported outcome measure. PROM=patient-reported outcome measure. PerfOM=performance outcome measure. OOM=objective outcome measure. RCT=randomised controlled trial.

Table 2: **Outcome measures used by more than five trials of adult depression**

Among the 450 RCTs, we identified 388 different measures of efficacy outcomes. Of these 388 measures, 259 (67%) were PROMs, 63 (16%) were PerfOMs, 45 (12%) were ClinROMs, one (<1%) was an ObsROM, and eight (2%) were objective outcomes. Two (<1%) measures, which assessed the carer's health or burden, fell outside our taxonomy. Ten (3%) measures were not categorised because information was missing.

Overall, depression trials used a mean of 3·77 (SD 3·58) measures related to death, psychiatric symptoms, life impact, or resource use. The most common measures were PROMs (mean 1·96 per study [SD 2·61]), followed by ClinROMs (1·29 [1·28]) and PerfOMs (0·32 [1·14]). For primary outcome assessment, 176 (39%) of the 450 RCTs used a PROM, and 276 (61%) used a ClinROM. Only 62 (14%) included at least one PerfOM. All-cause mortality was used as a secondary outcome in four (1%) trials. Death by suicide was measured in one RCT (<1%).

50 measures were used in more than five RCTs (table 2). Overall, the most commonly used measures were the HAMD (ClinROM), used in 180 (40%) of the 450 RCTs, followed by the MADRS (ClinROM), used in 138 (31%) trials, and the Beck Depression Inventory-II (BDI-II; PROM), used in 81 (18%) trials. The most-used PerfOM was the Trail Making Task, used in 13 (3%) trials.

MDD trials used 296 different measures, of which the most commonly used was the HAMD (147 [47%] of 316 trials; figure 2). Bipolar depression trials showed lower heterogeneity, using 68 different measures in total, of which the most commonly used was the MADRS (18 [78%] of 23 trials). Postpartum depression trials had similarly low heterogeneity, using 52 different measures in total, with 18 (69%) of 26 trials using the Edinburgh Postnatal Depression Scale (EPDS). Drug and psychotherapy trials most commonly used the HAMD as an outcome measure (136 [43%] of 317 and 16 [40%] of 40, respectively). The MADRS (ClinROM) was the second most common measure in drug trials (118 [37%] of 317), whereas in psychotherapy trials, the BDI-II (PROM) was the second most commonly used measure (13 [33%] of 40). In addition, fewer industry-sponsored trials (40 [73%] of 55) had outcome measure heterogeneity, compared with academic-sponsored trials (91 [37%] of 243). Further results regarding the ranking of the measures according to other trial characteristics are presented in appendix 3 (p 24).

Figure 3 presents the overlap between the 80 outcome domains that matter to patients and the ten most frequently used PROMs and nine ClinROMs used by more than five trials. The overlap of the 80 domains for PROMS, ClinROMs and PerfOMs, ordered by use of the PROCEED taxonomy,[14] is presented in appendix 3 (pp 25–27). Of the 450 trials, only 45 (10%) did not use any of the ten most-used PROMs or nine most-used ClinROMs. Half of the trials (224 [50%]) used at least one of the ten most-used PROMs, and two-thirds of the trials (296 [66%]) used at least one of the nine most-used ClinROMs.

All ten PROMs were multi-item measures assessing four latent constructs: depression, sleep quality, quality of life, and disability or impairment. The matching task showed that these ten PROMS overlapped with between five (6%) and 31 (39%) of the 80 domains (median 18 [23%], IQR 14·5). In total, 24 (30%) of the 80 domains that matter to patients were not covered by any of the ten most-used PROMs, of which 22 (34%) of 64 were symptoms (eg, helplessness, memory loss, or dissociation) and two (13%) of 16 related to functioning (eg, capacity to get out of bed or communicating feelings; figure 3).

Among the nine most-used ClinROMs, six multi-item measures assessed four different latent constructs: depression, suicidality, anxiety, and functioning. There were also three Clinical Global Impression scales, assessing severity of depression, patient improvement, and bipolar severity. The nine ClinROMs contained items overlapping with between zero and 47 (59%) of the 80 domains (median 21 [26%], IQR 35). Of the 18 (23%) domains not covered by these nine ClinROMs, 13 (20%) of 64 were symptoms (eg, emotional blunting, incurability, or feeling alone) and five (31%) of 16 related to functioning (eg, social isolation or ability to cope with a life event; figure 3).

Measures in the six PerfOMs overlapped with between four (5%) and ten (13%) of the 80 domains (median 7·5 [9%], IQR 2·5; appendix 3 p 27). Five objective outcomes were

	Drug trials (N=317)	Psychotherapy trials (N=40)	Neurostimulation trials (N=40)	MDD trials (N=316)	Bipolar trials (N=23)	Postpartum depression trials (N=26)	Industry trials (N=55)	Academic trials (N=242)
1	HAMD	HAMD	HAMD	HAMD	MADRS	EPDS	MADRS	HAMD
	136 (43%)	16 (40%)	25 (63%)	147 (47%)	18 (78%)	18 (69%)	40 (73%)	92 (38%)
2	MADRS	BDI-II	MADRS	MADRS	CGI-BP and C-SSRS	PAIN CBIRTH	CGI-S	BDI-II
	118 (37%)	13 (33%)	16 (40%)	98 (31%)	7 (30%)	7 (27%)	29 (53%)	58 (24%)
3	CGI-S	PHQ-9 and SF-36	BDI-II	CGI-S	HAMD, QIDS-SR	GAD-7 and PHQ-9	CGI-I	MADRS
	64 (20%)	8 (20%)	12 (30%)	58 (18%)	5 (22%)	4 (15%)	18 (33%)	42 (17%)
4	BDI-II	WHOQOL	CGI-S	BDI-II	Q-LES-Q	BDI-II	HAMD	PHQ-9
	49 (15%)	5 (13%)	9 (23%)	55 (17%)	4 (17%)	3 (12%)	15 (27%)	24 (10%)
5	CGI-I	ISI and MSPSS, HSU	BSSI and EQ-5D-5L	PHQ-9	CGI-S and QIDS-C	CGI-I and HAM-D, PSS and PSQI, TiH	PHQ-9	EQ-5D-5L and GAD-7
	37 (12%)	4 (10%)	5 (13%)	45 (14%)	3 (13%)	2 (8%)	13 (24%)	19 (8%)

Outcome assessment type
- Patient-reported outcome measures
- Clinician-reported outcome measures
- Objective outcome

Figure 2: **Ranking of the most frequently used measures in trials by intervention type (drug, psychotherapy, or neurostimulation), depression type (major depressive disorder, bipolar depression, or postpartum depression), and sponsor type (industry or academic)**

The measures are ranked by frequency of use, calculated from combined primary and secondary outcome use by trials. The numbers and proportions of trials using each measure are presented beneath the acronyms. BDI-II=Beck Depression Inventory-II. BSSI=Beck Scale for Suicide Ideation. CGI-BP=Clinical Global Impression scale-Bipolar. CGI-I=Clinical Global Impression-Improvement. CGI-S=Clinical Global Impression-Severity. C-SSRS=Colombia Suicide Severity Rating Scale. EPDS=Edinburgh Postnatal Depression Scale. EQ-5D-5L=EuroQoL-5 Dimensions-5 Levels. GAD-7=Generalized Anxiety Disorder-7. HAMD=Hamilton Rating Scale for Depression. HSU=health service use. ISI=Insomnia Severity Index. MADRS=Montgomery-Åsberg Depression Rating Scale. MDD=major depressive disorder. MSPSS=Multidimensional Scale of Perceived Social Support. PAIN CBIRTH=numerical rating scale for pain after childbirth. PHQ-9=Patient Health Questionnaire-9. PSQI=Pittsburgh Sleep Quality Index. PSS=Perceived Stress Scale. QIDS-C=Quick Inventory of Depressive Symptomatology, clinician-rated. QIDS-SR=Quick Inventory of Depressive Symptomatology, self-reported. Q-LES-Q=Quality of Life Enjoyment and Satisfaction Questionnaire. SF-36=Short Form-36. TiH=time in hospital. WHOQOL=World Health Organization Quality of Life Scale.

used by more than five trials, of which only one (the Client Service Receipt Inventory) overlapped with one domain, sick leave.

Among the PROMs and ClinROMs, the most-measured domains that matter to patients were feeling bad and sadness (assessed by 11 of 19 measures); anhedonia, interest, disturbed sleep, insomnia, and weakness (assessed by ten of 19 measures); concentration, restlessness, and coping with daily tasks (assessed by nine of 19 measures); and suicidal ideation, anxiety, energy, and fatigue (assessed by eight of 19 measures). Seven domains were not covered by any of the PROMs, ClinROMs, or PerfOMs, namely, mental pain, incurability, emotion regulation, mood reactivity, feeling misunderstood, self-recognition, and capacity to get out of bed.

The HAMD, which was the most frequently used measure across all trials (180 [40%] of 450), particularly MDD trials (147 [47%] of 316), contains 47 (59%) of the 80 domains that matter to patients (42 of 64 symptom domains and five of 16 functioning domains). The MADRS, the second most-used measure across all trials (138 [31%] of 450) and the most-used measure in bipolar disorder trials (18 [78%] of 23), contains 42 (53%) of the domains that matter to patients (41 of the 64 symptom domains and one of the 16 functioning domains). The BDI-II, the third most-used measure overall (81 [18%] of 450 RCTs) and the most frequently used PROM, contains 31 (39%) of the 80 domains that matter to patients (28 of the 64 symptom domains and three of the 16 functioning domains). The Patient Health Questionnaire-9 (PHQ-9), the second most-used PROM (63 [14%] of 450 RCTs), contains 22 (28%) of the 80 domains that matter to patients (21 of the 64 symptom domains and one of the 16 functioning domains). Finally, the EPDS, the most-used measure in postpartum depression trials (18 [69%] of 26 RCTs), contains 14 (18%) of the 80 domains that matter to patients (13 of the 64 symptom domains and one of the 16 functioning domains).

Discussion

This systematic review evaluated to what extent the measures used in ongoing and future depression trials cover the outcomes that matter to patients. We identified

388 unique measures used across 450 RCTs. We found that commonly used measures only partly reflect the domains that matter to patients (with a maximum of 59% overlap by the HAMD, a ClinROM) and that several important domains such as mental pain, irritability, or emotion regulation are not assessed by these measures.[14] The most inclusive PROM, the BDI-II, only covers 39% of the domains that matter to patients. Notably, the most-used measures in depression RCTs were screening tools (eg, the PHQ-9 and EPDS), which were not developed to track treatment progress, consequently raising questions about the validity of these measures in the context of RCTs.[12,24]

Additionally, we identified heterogeneity in the outcome measures used, with only 40% of RCTs using the same measure (HAMD). Heterogeneity in outcomes is a known source of research waste that prevents the comparability and combination of RCT results in meta-analyses. We have identified patterns of measure usage by grouping RCTS according to particular characteristics (figure 2), but considering further subgroups of RCTs—for instance, type of drug (eg, SSRI or ketamine), psychotherapy (eg, cognitive behavioural therapy or psychodynamic therapy), or inclusion criteria (eg, depression severity or comorbidities)—might reveal additional patterns.

Our study results need to be interpreted considering several limitations. First, we did not find any standardised or universally accepted method for quantifying outcome heterogeneity. Second, it was unfeasible to conduct the matching task on all 388 measures identified across the 450 trials. Rather, we focused on the most commonly used PROMs, ClinROMs, and PerfOMs; nonetheless, 90% of the trials used at least one of these measures. Third, as PROCEED[14] aimed to identify the largest diversity of domains rather than ranking them by importance, the frequencies of domains identified by spontaneous reports from the participants should not be used as indicators of their relative importance. As a consequence, we do not know whether the domains we have identified as not covered by the measures are as important as those covered. In addition, the importance

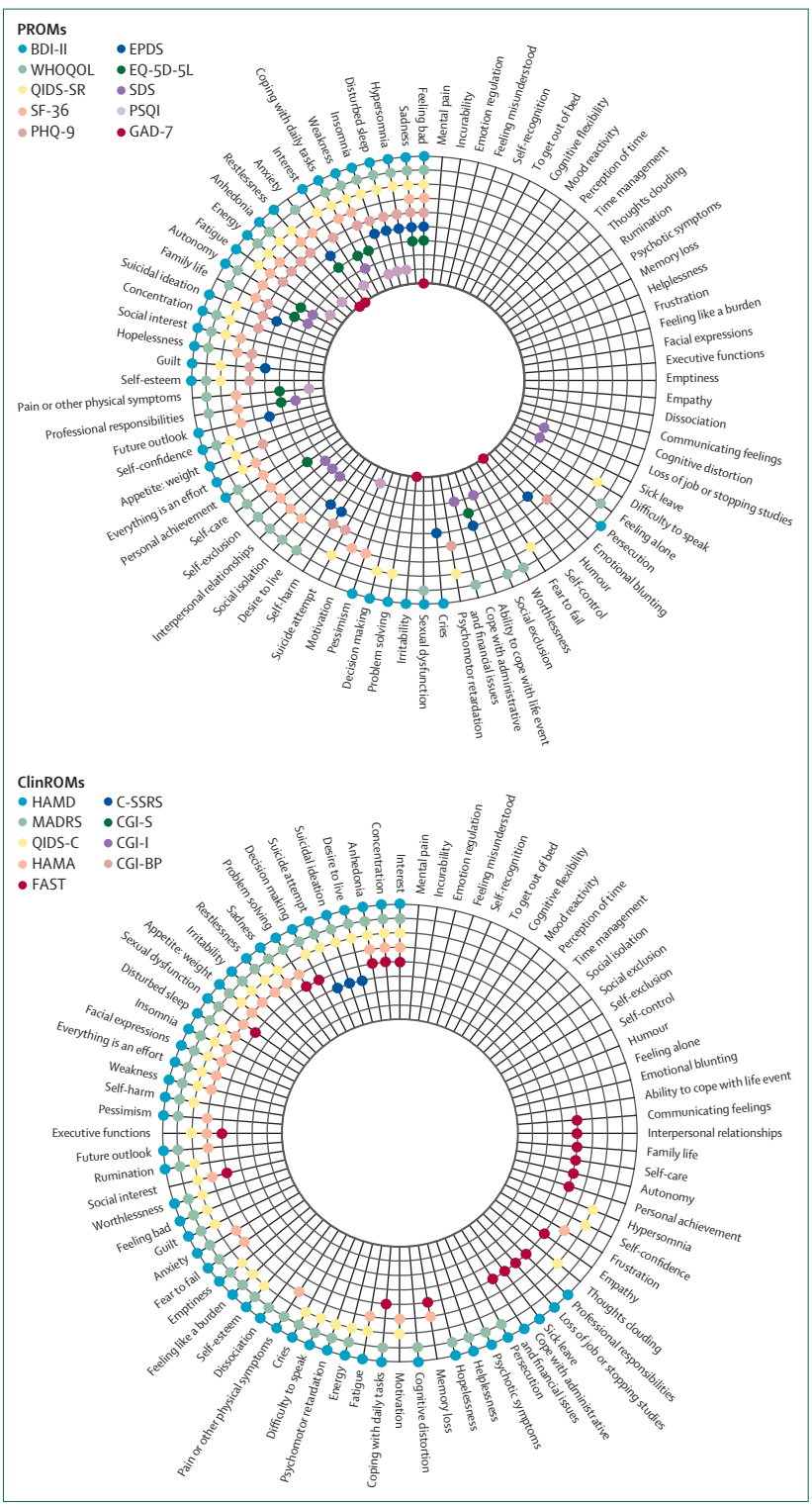

Figure 3: **Content overlap of the most frequently used measures of depression and the 80 domains that matter to patients**
Measures (PROM or ClinROM) are represented by rings of concentric circles; the lines intersecting the rings indicate the 80 domains that matter to patients identified in the PROCEED study.[14] A coloured dot indicates overlap of the intersecting domain with at least one item from the PROM or ClinROM. Moving anticlockwise, PROMs and ClinROMs are ordered from most overlap (outer circle) to least overlap (inner circle) for each measure. The 80 domains are ordered from the most covered to the least (or not) covered. BDI-II=Beck Depression Inventory-II. CGI-BP=Clinical Global Impression-Bipolar. CGI-I=Clinical Global Impression-Improvement. CGI-S=Clinical Global Impression-Severity. ClinROMs=clinician-reported outcome measures. CSSRS=Colombia Suicide Severity Rating Scale. EPDS=Edinburgh Postnatal Depression Scale. EQ-5D-5L=EuroQoL-5 Dimensions-5 Levels. FAST=Functional Assessment Short Test. GAD-7=Generalized Anxiety Disorder-7. HAMA=Hamilton Anxiety. HAMD=Hamilton Rating Scale for Depression. MADRS=Montgomery-Åsberg Depression Rating Scale. PHQ-9=Patient Health Questionnaire-9. PROMs=patient-reported outcome measures. PSQI=Pittsburg Sleep Quality Index. QIDS-C=Quick Inventory of Depressive Symptomatology, clinician-rated. QIDS-SR=Quick Inventory of Depressive Symptomatology, self-rated. SDS=Sheehan Disability Scale. SF-36=Short Form-36. WHOQOL=World Health Organization Quality of Life Scale.

of domains might vary according to demographic group (eg, gender and cultural groups).

A promising solution to reduce heterogeneity and improve the coverage of domains that matter to patients is to develop a core outcome set (COS) for depression. A COS is a minimum set of outcomes agreed on by all relevant stakeholders to be measured in all trials of a given condition.[5,25] For other conditions, such as rheumatoid arthritis, COS endorsement by regulatory authorities and trials has proved efficient in reducing heterogeneity;[26] for example, patients' involvement in the development process allowed for the selection of previously overlooked outcomes, such as fatigue.[27] The ongoing development of a COS for depression includes the elicitation of patients' preferences regarding the list of 80 domains to identify the most important, also taking into account potential subgroup specificities (as done in PROCEED),[14,27] and finalising the COS as a reduced set of outcomes with their corresponding measures. The final set of measures should be validated for the evaluation of the efficacy of treatments for depression and maximisation of the coverage of the most important domains, alongside minimisation of the burden for patients and trialists. We recommend restraint in standardising outcomes in trials without a rigorous and inclusive COS development and update process.[28]

In the interim, two temporary and complementary solutions could be considered. The first solution is to continue the use of frequently used measures, such as the HAMD, MADRS, and BDI-II, to ensure comparability with previous and contemporary studies. In addition, we suggest the use of complementary measures to ensure that a larger number of domains that matter to patients are captured. For example, complementary use of the HAMD, BDI-II, and Functional Assessment Short Test would result in 59 (74%) of the 80 patient-relevant domains being assessed. Additionally, we consider the possibility that some instruments might more comprehensively assess the domains that matter to patients with depression than the most used measures we evaluated in this study.

Contributors

CV devised the project, designed the study, extracted and analysed the data, and wrote the first draft. EIF and V-TT designed the study and edited the protocol and manuscript. ST and HL-S developed the depression dictionary. ST, HL-S, and NdeW edited the protocol, performed the matching task for the PROMs, and edited the manuscript. AT, ACi, and CH edited the protocol, edited the depression dictionary, performed the matching task for the ClinROMs, and edited the manuscript. TAF and SB performed the matching task for the ClinROMs and edited the manuscript. CM performed the matching task for the PerfOMs and edited the manuscript. ACh devised the project, formed the main conceptual ideas, designed the study, extracted and analysed the data, and critically edited the manuscript. ACh is the guarantor of the study, accepts full responsibility for the work, has access to the data, and took the decision to publish.

Declaration of interests

AT has received research, educational and consultancy fees from Italian Network for Paediatric Trials (INCiPiT), Angelini Pharma, and Takeda and acted as a clinical advisor for Akrivia Health. ACi has received research, educational and consultancy fees from INCiPiT, the Cariplo Foundation, Lundbeck, and Angelini Pharma; he is also the chief investigator and principal investigator of a trial about seltorexant for adolescents with depression, sponsored by Janssen. SB has received grants from CHU Nantes and the French Ministry of Health and educational fees from Lundbeck and Janssen, outside the submitted work. CH has received educational fees from Lundbeck and Sanofi Aventis. TAF reports personal fees from Boehringer-Ingelheim, DT Axis, Kyoto University Original, Shionogi, SONY, and UpToDate, and a grant from Shionogi, outside the submitted work; in addition, TAF has patents 2020-548587 and 2022-082495 pending and intellectual properties for Kokoro-app, licensed to Mitsubishi-Tanabe.

ACi is supported by the National Institute for Health Research (NIHR) Oxford Cognitive Health Clinical Research Facility, an NIHR Research Professorship (grant RP-2017-08-ST2-006), the NIHR Oxford and Thames Valley Applied Research Collaboration, the NIHR Oxford Health Biomedical Research Centre (grant NIHR203316), and Wellcome (GALENOS project). All other authors declare no competing interests.

Acknowledgments

ST, NDW, and HL-S are co-researchers with lived experience. HL-S has a role as a peer support community organiser at La Maison Perchée, Paris, France. The views expressed in this Review are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health. The data that support the findings of this study are available from the corresponding author, ACh, upon reasonable request.

References

1 GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Psychiatry* 2022; **9:** 137–50.

2 Leichsenring F, Steinert C, Rabung S, Ioannidis JPA. The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. *World Psychiatry* 2022; **21:** 133–45.

3 Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database—update and key issues. *N Engl J Med* 2011; **364:** 852–60.

4 Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *Lancet* 2014; **383:** 156–65.

5 Williamson P, Altman D, Blazeby J, Clarke M, Gargon E. Driving up the quality and relevance of research through the use of agreed core outcomes. *J Health Serv Res Policy* 2012; **17:** 1–2.

6 Fried EI. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *J Affect Disord* 2017; **208:** 191–97.

7 Santor DA, Gregus M, Welch A. Eight decades of measurement in depression. *Measurement* 2006; **4:** 135–55.

8 Rodrigues M, Syed Z, Dufort A, et al. Heterogeneity across outcomes reported in clinical trials for older adults with depression: a systematic survey. *J Clin Epidemiol* 2023; **157:** 59–73.

9 Mew EJ, Monsour A, Saeed L, et al. Systematic scoping review identifies heterogeneity in outcomes measured in adolescent depression clinical trials. *J Clin Epidemiol* 2020; **126:** 71–79.

10 Chevance A, Ravaud P, Cornelius V, Mayo-Wilson E, Furukawa TA. Designing clinically useful psychopharmacological trials: challenges and ways forward. *Lancet Psychiatry* 2022; **9:** 584–94.

11 Hieronymus F, Lisinski A, Nilsson S, Eriksson E. Influence of baseline severity on the effects of SSRIs in depression: an item-based, patient-level post-hoc analysis. *Lancet Psychiatry* 2019; **6:** 745–52.

12 Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. *Nat Rev Psychol* 2022; **1:** 358–68.

13 Hengartner MP, Plöderl M. Statistically significant antidepressant-placebo differences on subjective symptom-rating scales do not prove that the drugs work: effect size and method bias matter! *Front Psychiatry* 2018; **9:** 517.

14 Chevance A, Ravaud P, Tomlinson A, et al. Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. *Lancet Psychiatry* 2020; **7:** 692–702.

15 Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. *J Clin Epidemiol* 2018; **96:** 84–92.

16 COMET Initiative. Physiological or impact? 2022. https://www.comet-initiative.org/Resources/Physiological (accessed Dec 9, 2022).

17 Walton MK, Powers JH 3rd, Hobart J, et al. Clinical outcome assessments: conceptual foundation-report of the ISPOR clinical outcomes assessment—emerging good practices for outcomes research task force. *Value Health* 2015; **18:** 741–52.

18 Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; **4:** 1.

19 Staniszewska S, Brett J, Simera I, et al. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *BMJ* 2017; **358:** j3453.

20 WHO. ICRTP Registry Network. 2022. https://www.who.int/clinical-trials-registry-platform/network (accessed Oct 26, 2022)

21 Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015 ; **350:** h2147.

22 Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Clin Epidemiol* 2009; **62:** 499–505.

23 Mobbs RJ. From the subjective to the objective era of outcomes analysis: how the tools we use to measure outcomes must change to be reflective of the pathologies we treat in spinal surgery. *J Spine Surg* 2021; **7:** 456–57.

24 McPherson S, Armstrong D. Psychometric origins of depression. *Hist Human Sci* 2022; **35:** 127–43.

25 Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. *Trials* 2017; **18** (suppl 3)**:** 280.

26 Kirkham JJ, Boers M, Tugwell P, Clarke M, Williamson PR. Outcome measures in rheumatoid arthritis randomised trials over the last 50 years. *Trials* 2013; **14:** 324.

27 Chevance A, Tran VT, Ravaud P. Controversy and debate series on core outcome sets. Paper 1: Improving the generalizability and credibility of core outcome sets (COS) by a large and international participation of diverse stakeholders. *J Clin Epidemiol* 2020; **125:** 206–12.

28 Patalay P, Fried EI. Editorial perspective: prescribing measures: unintended negative consequences of mandating standardized mental health measurement. *J Child Psychol Psychiatry* 2021; **62:** 1032–36.