Minimal detectable change of the Patient Health Questionnaire-9, Patient Health Questionnaire-8, and Patient Health Questionnaire-2: individual patient data meta-analysis

Yutong Wang,[1,2] Yin Wu,[3] Nadia P González-Domínguez,[2] Jill T Boruff,[4] Brooke Levis,[1] Pim Cuijpers,[5] Simon Gilbody,[6] Daphna Harel,[7] John P A Ioannidis,[8] Sarah Markham,[9] Scott B Patten,[10] Simone N Vigod,[11] Roy C Ziegelstein,[12] Andrea Benedetti,[1,13,14] Brett D Thombs[1,2,13,15]; on behalf of the DEPRESSion Screening Data (DEPRESSD) PHQ Collaboration

For numbered affiliations see end of the article

Correspondence to: B D Thombs, Jewish General Hospital, Centre for Clinical Epidemiology, Montréal, QC, H3T 1E2, Canada brett.thombs@mcgill.ca; (ORCID 0000-0002-5644-8432)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2026;394:e100119
http://dx.doi.org/10.1136/bmj-2026-100119

Accepted: 23 June 2026

ABSTRACT

OBJECTIVE

To estimate the minimal detectable change (MDC) for the Patient Health Questionnaire-9 (PHQ-9) and its eight item (PHQ-8) and two item (PHQ-2) versions including differences by participant and study characteristics.

DESIGN

Individual participant data meta-analysis.

DATA SOURCES

Medline, Medline In-Process and other non-indexed citations, PsycInfo, and Web of Science, 1 January 2000 to 9 May 2018.

ELIGIBILITY CRITERIA FOR SELECTING STUDIES

Datasets from articles in any language if participants were aged ≥18 years, were recruited from any non-psychiatric setting, and were not recruited because they were seeking mental healthcare. Eligible datasets had a classification for major depressive disorder or major depressive episode based on a validated semi-structured or fully structured interview conducted within two weeks of administering the PHQ-9, PHQ-8, or PHQ-2.

RESULTS

Pooled MDCs across studies were estimated for the PHQ-9, PHQ-8, and PHQ-2 with random effects meta-analysis for 95% (MDC95), 90% (MDC90), and 67% (MDC67) confidence that change beyond measurement error occurred. PHQ-9, PHQ-8, and PHQ-2 analyses included 42 548 participants (94 studies), 42 592 participants (94 studies), and 44 085 participants (98 studies), respectively. Mean participant age was 49 years (standard deviation 17), and 60% of participants were women. Overall, 10% of participants had major depression (range 1-57% across studies). MDC95 was 5.72 points (95% confidence interval (CI) 5.54 to 5.90, 95% prediction interval (PI) 4.00 to 7.44) for the PHQ-9, 5.51 points (95% CI 5.33 to 5.68, 95% PI 3.87 to 7.15) for the PHQ-8, and 2.26 points (95% CI 2.15 to 2.37, 95% PI 1.20 to 3.32) for the PHQ-2. For the PHQ-9, MDC95 was highest in inpatient healthcare settings at 6.48 (95% CI 6.05 to 6.92) points. MDC95 for the PHQ-9 increased by 0.40 (95% CI 0.25 to 0.55) points for each 10% increase in the proportion of participants with major depression. Sex and age had minimal or no association. Subgroup and meta-regression findings were similar for the PHQ-8 and PHQ-2.

CONCLUSIONS

Based on the pooled estimate, a six point difference on the PHQ-9, the PHQ version most used in clinical practice, could be an appropriate MDC threshold in general practice. A higher threshold may be preferred in specialty mental healthcare. MDC67 or MDC90 thresholds would provide less certainty that change has occurred. Alternative strategies, such as using the upper end of a prediction interval, would provide more certainty but a greater likelihood of not recognising change.

STUDY REGISTRATION

PROSPERO CRD42014010673.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The minimal detectable change (MDC) of the Patient Health Questionnaire-9 (PHQ-9) and the shorter PHQ-8, and PHQ-2, has not been assessed in a systematic review

Only one primary study in systemic sclerosis has assessed the MDC for the PHQ-8, and none has evaluated PHQ-9 or PHQ-2 MDCs

The UK NHS's Talking Therapies programme uses a six point PHQ-9 reliable change index (equivalent to MDC with 95% confidence, MDC95) to assess change in individuals in treatment, but no supporting analyses have been published

WHAT THIS STUDY ADDS

This study synthesised evidence from 98 studies (44 085 participants) and estimated MDC95, MDC90, and MDC67 for the PHQ-9, PHQ-8, and PHQ-2

A six point threshold for the PHQ-9 could be considered for evaluating individual changes in depression symptoms and as a threshold for assessing estimates for a minimal important difference

A higher threshold for the PHQ-9 may be more appropriate for people in specialty mental health treatment settings

Introduction

Patient reported outcome measures assess people's health and wellbeing from their perspective. Measurement based care involves administering patient reported outcome measures, with scores to evaluate change, and discussing the results collaboratively with patients to inform treatment decisions. Clinical decision making must consider whether changes in scores are clinically important and represent reliable change beyond random fluctuation.[1]

Minimal important differences, also known as minimal clinically important differences, represent

the smallest change likely perceived as worthwhile by patients.[2 3] Minimal important differences are estimated by comparing changes in scores to an anchor that reflects a patient's perspective of whether or how much change has occurred.[4-6] Sometimes, however, estimated minimal important differences are too small to be reliably detected.

The minimal detectable change (MDC) is the smallest change in score that can be detected beyond measurement error.[7 8] Minimal important differences that are smaller than the MDC cannot distinguish meaningful change from chance fluctuations. Determining that meaningful change has occurred requires that a change in score exceeds the minimal important difference threshold and is detectable beyond error fluctuation based on the MDC.

The Patient Health Questionnaire (PHQ)-9, PHQ-8, and PHQ-2 were originally developed as screening tools and are now the most commonly used mental health outcome measures in clinical practice.[2 9-16] The PHQ-9 has nine questions, with each question scored 0-3. Possible total scores range from 0 (least depressed) to 27 (most depressed). The eight item PHQ-8 and two item PHQ-2 are subsets of the PHQ-9 with score ranges of 0-24 and 0-6, respectively. The PHQ-8 differs from the PHQ-9 by excluding an item on suicidal ideation. The PHQ-2 includes the first two items of the PHQ-9 and is mainly used as a first step in screening for depression.

No systematic review has assessed the minimal important difference or MDC of the PHQ-9, PHQ-8, or PHQ-2. Two primary studies, both from the UK, estimated the minimal important difference for the PHQ-9 with patient reported anchors for improvement. One study assessed 1122 participants from two trials on treatments for depression at three month intervals and reported a minimal important difference of 3.7 points.[10] The other study assessed 400 patients in primary care at two week intervals and reported minimal important difference thresholds of 2.0, 1.7, and 2.4 points among participants with low, medium, and high baseline symptoms.[11] Neither study provided estimates of precision, an important consideration in evaluating the credibility of the minimal important difference estimates.[17]

The PHQ-9 is commonly used in measurement based care. The UK NHS's Talking Therapies programme (formerly known as the Improving Access to Psychological Therapies programme) uses a six point PHQ-9 threshold for reliable change, equivalent to 95% confidence that change beyond measurement error has occurred (MDC95). No analysis of estimation of this threshold, however, has been published.[18-20] This threshold has also been adopted in Canadian mental healthcare.[21-23] In the US, the National Committee for Quality Assurance defines a response to treatment for depression as a reduction in PHQ-9 score of ≥50% and remission as a PHQ-9 score of < 5, but has not identified metrics for assessing meaningful or reliable change.[24 25]

The only study that has evaluated the MDC for the PHQ-9, PHQ-8, or PHQ-2 reported an MDC of 5.2 points (95% confidence interval (CI) 5.1 to 5.3) for the PHQ-8 in 2571 people with the rare autoimmune disease systemic sclerosis.[26] This finding suggests that existing estimates of minimal important difference (range 1.7-3.7 points) may be too small to be distinguished from measurement error. Also, the MDC for people with systemic sclerosis, a specific population with a rare disease, may not be generalisable to other populations. No studies have estimated MDCs for the PHQ-9, PHQ-8, or PHQ-2 in large representative samples or evaluated whether MDCs vary by factors such as age, sex, recruitment setting, and major depression status.

Individual participant data meta-analyses (IPDMAs) can be used to generate estimates from data collected in many primary studies, even when those studies did not publish the outcome of interest. We have assembled a large database to conduct IPDMAs on measurement properties of the PHQ-9, PHQ-8, and PHQ-2.[27-30] The objectives of our study were to evaluate the MDC for the PHQ-9, PHQ-8, and PHQ-2, including MDC95, MDC90 (90% confidence), and MDC67 (67% confidence), and examine whether MDCs differ by age, sex, recruitment setting, or major depression status.

Methods

The MDC can be calculated with test-retest reliability, which requires multiple observations and is often not available, or with a cross sectional reliability estimate. We calculated MDC based on Cronbach's alpha, a cross sectional estimation method.

Our main PHQ-9, PHQ-8, and PHQ-2 screening accuracy IPDMAs were registered in the International Prospective Register of Systematic Reviews (PROSPERO CRD42014010673), and the results have been published.[27-30] The protocol for this study was uploaded to the Open Science Framework before the start of the study (https://osf.io/xsyw8). We have reported our study consistent with relevant Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) items.[31] Because data collection methods were similar to our main PHQ-9, PHQ-8, and PHQ-2 IPDMAs,[27-30] and data analysis methods were similar to our study of Geriatric Depression Scale MDCs,[32] we followed the reporting guidance of the Text Recycling Research Project.[33]

Eligible studies and participants

For our main PHQ-9, PHQ-8, and PHQ-2 screening accuracy IPDMAs,[27-30 34] datasets from articles in any language were eligible if participants were aged ≥18 years, were recruited from any non-psychiatric setting, and were not recruited because they were seeking mental healthcare, because screening is done to identify previously unrecognised cases. Examples of eligible studies included studies of participants from the community, general practice, or non-psychiatric specialty inpatient or outpatient settings (eg, oncology or cardiology). Eligible datasets had to include a classification for major depressive

doi: 10.1136/bmj-2026-100119 | BMJ 2026;394:e100119 | the bmj

disorder or major depressive episode based on a validated semi-structured or fully structured interview conducted within two weeks of administering the PHQ-9, PHQ-8, or PHQ-2 because diagnostic criteria for major depression are for symptoms in the past two weeks. Datasets where not all participants were eligible were included if the primary data allowed us to select eligible participants. In defining major depression, we considered major depressive disorder or major depressive episode based on any version of the Diagnostic and Statistical Manual of Mental Disorders or the International Classification of Diseases. If more than one was reported, we prioritised major depressive episode over major depressive disorder and the Diagnostic and Statistical Manual of Mental Disorders over the International Classification of Diseases. For this study, we included only primary study datasets with PHQ-9, PHQ-8, or PHQ-2 item scores (ie, not if only total scores were reported), because item scores are needed to calculate Cronbach's alpha reliability, which was used to calculate the standard error of measurement and thus MDC.[35]

Identification and selection of eligible studies

We searched Medline, Medline In-Process and other non-indexed citations, PsycInfo, and Web of Science from 1 January 2000 to 9 May 2018 with a peer reviewed[36] search strategy (supplementary methods 1). We did not perform searches before 2000 because the PHQ-9 was originally published in 2001.[37] We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, ON, Canada) to store and track search results, conduct eligibility screening, record correspondence with primary study authors, and extract study characteristics.

Two investigators independently reviewed titles and abstracts for eligibility. If either investigator considered a study potentially eligible, a full text review was done by two investigators, independently (supplementary methods 2). Any disagreements after full text review were resolved by consensus, consulting a third investigator when necessary. Translators were consulted to evaluate titles, abstracts, and full text articles for languages other than those for which team members were fluent.

Data extraction, contribution, and synthesis

We invited authors of eligible datasets to contribute de-identified primary datasets. We emailed corresponding authors of eligible studies at least three times, as necessary. If no response was received, we emailed co-authors and attempted telephone contact.

We compared published participant characteristics and diagnostic accuracy results with results that we generated from raw datasets. We resolved any discrepancies with the primary study investigators. We then converted individual participant data to a standard format and entered the data, along with study level data, into one harmonised dataset. When datasets included statistical weights to reflect sampling procedures, we used the provided weights. For studies where sampling procedures merited weighting, but the original study did not weight, we constructed weights with inverse selection probabilities. For our screening accuracy IPDMAs, we assessed the risk of bias with the Quality Assessment of Diagnostic Accuracy Studies-2 tool.[38] No risk of bias tool exists for studies assessing MDCs, and risk of bias for screening accuracy would not necessarily reflect risk of bias for MDCs, so we did not incorporate a risk of bias assessment into our study.

Data analysis

We used a two stage IPDMA approach. In the first stage, individual participant data were analysed separately within each study to obtain study level MDC estimates. In the second stage, study level estimates were pooled with restricted maximum likelihood random effects meta-analysis and mixed effects meta-regression models.

We calculated MDC95 as $1.96\times\sqrt{2}\times$SEM, MDC90 as $1.64\times\sqrt{2}\times$SEM, and MDC67 as $\sqrt{2}\times$SEM,[39-41] which reflect CIs for a difference. SEM (standard error of measurement) is the standard deviation of test scores multiplied by the square root of one minus the reliability coefficient.[12 35 41 42] We used Cronbach's alpha as the reliability coefficient.[12 35] Ideally, test-retest reliability would be used to reflect reliability in scores over time. Test-retest reliability estimates, however, must be done over a period that is long enough to avoid memory effects but short enough to ensure that symptom status has not changed; few studies provide such estimates.[43 44] Thus Cronbach's alpha is commonly used in estimating MDC[43 45] (supplementary methods 3 has more details).[43 45-47]

For each study, we estimated MDC and its standard deviation (SD) by a bootstrap approach with resampling at the participant level and 1000 bootstrap iterations. We used restricted maximum likelihood random effects meta-analysis to obtain pooled MDC95, MDC90, and MDC67 estimates with 95% CIs and 95% prediction intervals (PIs). CIs were derived with normal based Wald methods and PIs were calculated as the pooled estimate $\pm 1.96\times\sqrt{(\tau^2+\text{Var(pooled estimate)})}$ under a normal predictive distribution. We examined the distributions of the random effects for each model, which were found to be reasonably normal. Because CIs can be sensitive to distributional assumptions, particularly when between study heterogeneity is substantial, we conducted sensitivity analyses with the Hartung-Knapp approach to derive CIs and PIs, with the corresponding t distribution and degrees of freedom based on the number of studies contributing to each analysis minus one. We excluded participants with missing item scores on the PHQ-9, PHQ-8, or PHQ-2. Given the small proportion of missing data (PHQ-9, n=566 (1%); PHQ-8, n=522 (1%); PHQ-2, n=203 (< 1%); and only one study with >15% of participants with missing items), we used complete case analyses.

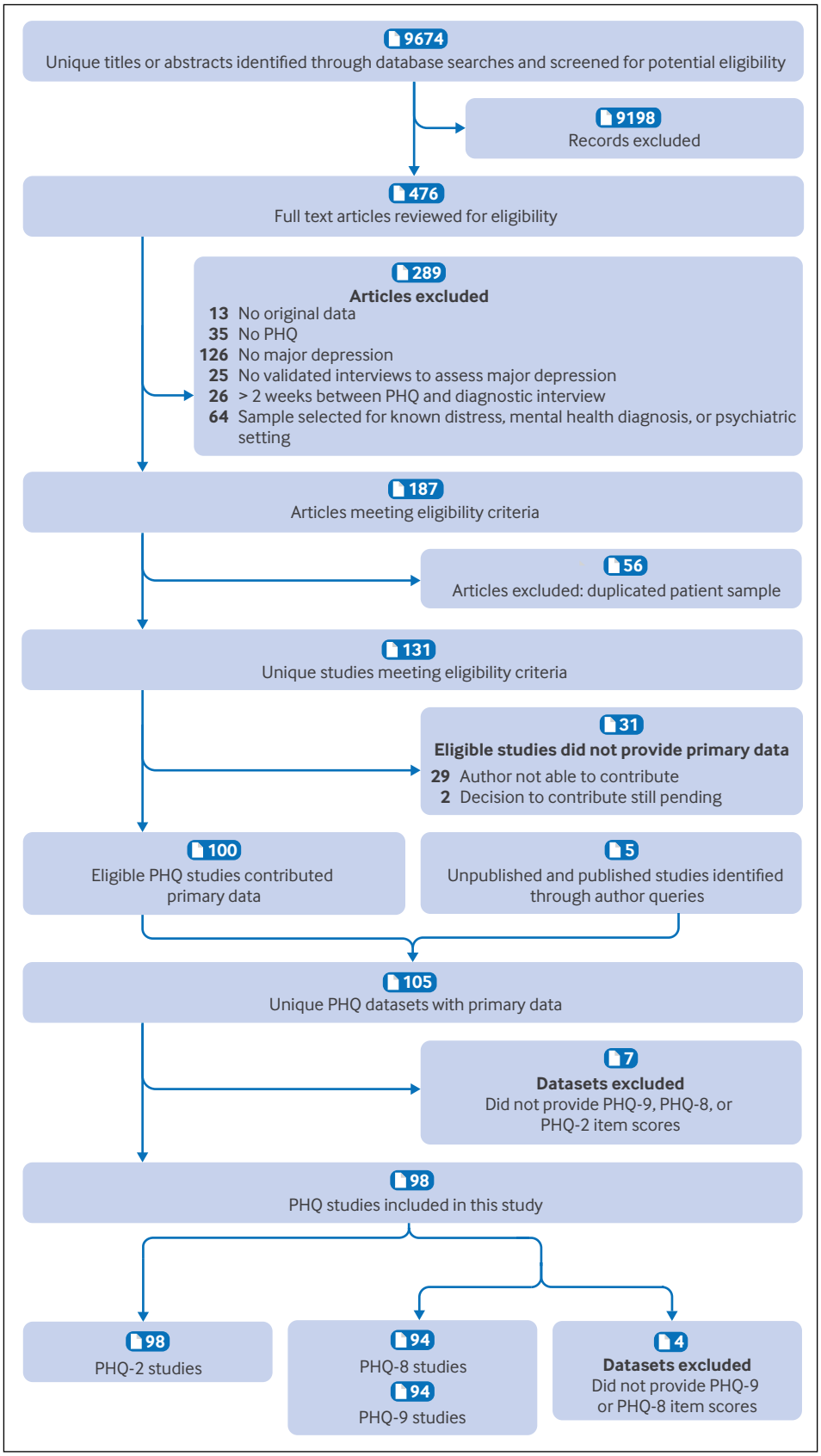


Fig 1 | Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram. PHQ=Patient Health Questionnaire; PHQ-2=Patient Health Questionnaire-2; PHQ-8=Patient Health Questionnaire-8, and PHQ-9=Patient Health Questionnaire-9

Table 1 | Unweighted characteristics of 42 548 participants from 94 studies included in the Patient Health Questionnaire-9 (PHQ-9) analysis

Subgroup	No of participants*	Mean (SD) proportion across studies*	Mean (SD) PHQ-9 score*	No (%) of participants with major depression positive classification status*
Overall sample	42 548	—	4.96 (5.24)	4255 (10)
Age (years)†:				
18-29	6684	16 (18)	5.66 (5.28)	664 (10)
30-44	11 273	25 (17)	5.18 (5.41)	1205 (11)
45-64	16 161	36 (19)	4.82 (5.24)	1616 (10)
≥65	8163	22 (25)	4.35 (4.84)	739 (9)
Missing	267	1 (10)	6.20 (5.78)	31 (12)
Country‡:				
US	4028 (15)	—	5.99 (5.64)	619 (15)
Australia or New Zealand	4299 (9)	—	4.55 (5.08)	315 (7)
UK	613 (3)	—	7.99 (6.88)	108 (18)
Germany	1605 (6)	—	7.13 (5.76)	147 (9)
Canada	5788 (9)	—	4.19 (4.81)	269 (5)
Brazil	543 (4)	—	7.52 (6.61)	132 (24)
Netherlands	7049 (7)	—	3.62 (4.37)	494 (7)
Other European countries§	3817 (7)	—	6.18 (5.37)	532 (14)
Middle Eastern or African countries¶	4817 (14)	—	6.21 (5.67)	833 (17)
Asian countries or regions**	9452 (18)	—	4.24 (4.75)	708 (7)
Central or South American countries††	537 (2)	—	7.05 (5.71)	98 (18)
Sex‡‡:				
Women	25 013	60 (23)	5.38 (5.34)	2797 (11)
Men	17 512	40 (22)	4.37 (5.03)	1455 (8)
Missing	23	0 (1)	7.30 (6.24)	3 (13)
Setting§§:				
Non-medical	15 552	—	3.48 (4.32)	944 (6)
Primary care	13 983	—	5.46 (5.35)	1580 (11)
Outpatient specialty	9822	—	5.99 (5.58)	1369 (14)
Inpatient	3061	—	7.03 (5.94)	357 (12)
Mixed outpatient specialty and inpatient	130	—	3.80 (3.78)	5 (4)

SD=standard deviation.
*Characteristics are based on observed participants and calculated without participant level sampling weights.
†267 participants missing for age across 23 studies.
‡Number of participants with number of studies in parentheses.
§Includes Spain (n=1), Italy (n=1), Greece (n=3), Belgium (n=1), and Latvia (n=1).
¶Includes South Africa (n=4), Israel (n=2), Iran (n=2), Ethiopia (n=1), Kenya (n=1), Zimbabwe (n=1), Cameroon (n=1), and Uganda (n=2).
**Includes China (n=2), Japan (n=4), Hong Kong (n=1), Malaysia (N=3), Singapore (n=2), South Korea (n=1), Taiwan (n=1), India (n=1), Thailand (n=1), Nepal (n=1), and Myanmar (n=1).
††Includes Mexico (n=1) and Argentina (n=1).
‡‡23 participants missing for sex across eight studies.
§§130 participants from one study with mixed outpatient specialty and inpatient settings, without individual classification.

We estimated MDCs for subgroups based on age (18-29, 30-44, 45-64, and ≥65 years),[26] sex (women and men), and recruitment setting (non-medical, primary care, outpatient specialty, and inpatient) with the same analysis method as the main MDC analyses. To estimate MDC differences between age and sex subgroups, we estimated the difference in each study and its SD followed by restricted maximum likelihood random effects meta-analysis. In each subgroup analysis, the subgroup with the largest number of included studies was used as the reference group to maximise study coverage and sample size (age reference subgroup was 45-64 years and sex reference subgroup was women). To compare the age groups 18-29 years and 45-64 years, for example, we performed bootstrap resampling within each study that included participants from both age groups to estimate separate MDCs for each age group. We then calculated within study subgroup differences by subtracting the MDC of the reference age group (45-64 years) from that of the comparison age group. Within study differences were then pooled with restricted maximum likelihood random effects meta-analysis. Participants with missing data on a subgroup defining variable were excluded from that subgroup analysis.

We used meta-regression, rather than meta-analysis of differences within studies, to estimate MDC differences between recruitment sites because all participants in each study were recruited from the same site. We also used meta-regression, rather than a subgroup analysis, for major depression classification status. We used this approach because of the indirect range restriction of PHQ-9, PHQ-8, and PHQ-2 scores, which would have occurred if we had grouped participants based on depression classification status. This grouping would have altered variance within classification groups, thus making the estimation of reliability untenable.[48] We did not conduct a subgroup analysis on symptom severity scores for the same reason. We fitted two stage linear mixed effects meta-regression models in which the MDC95 of each study was used as the outcome. Study recruitment setting

Table 2 | Synthesised estimates of minimal detectable change with 95% (MDC95), 90% (MDC90) and 67% (MDC67) confidence for Patient Health Questionnaire-9 (PHQ-9), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-2 (PHQ-2)

Scale	No of participants (studies)	Mean (SD) score*	MDC95, MDC90, or MDC67	MDC estimate (95% CI)	95% PI	τ^2†
PHQ-9	42 548 (94)	6.14 (5.29)	MDC95	5.72 (5.54 to 5.90)	4.00 to 7.44	0.76
			MDC90	4.79 (4.64 to 4.94)	3.35 to 6.23	0.54
			MDC67	2.92 (2.83 to 3.01)	2.04 to 3.80	0.20
PHQ-8	42 592 (94)	5.95 (5.02)	MDC95	5.51 (5.33 to 5.68)	3.87 to 7.15	0.69
			MDC90	4.61 (4.47 to 4.76)	3.24 to 5.99	0.49
			MDC67	2.81 (2.72 to 2.90)	1.97 to 3.65	0.18
PHQ-2	44 085 (98)	1.45 (1.59)	MDC95	2.26 (2.15 to 2.37)	1.20 to 3.32	0.29
			MDC90	1.89 (1.80 to 1.99)	1.01 to 2.78	0.20
			MDC67	1.15 (1.10 to 1.21)	0.61 to 1.69	0.08

CI=confidence interval; PI= prediction interval; SD=standard deviation.
*Pooled estimates with restricted maximum likelihood random effects meta-analysis. The correlation between study sample size and mean score was -0.38. Although study weights were relatively similar because of high between study variability, this negative correlation may have influenced the results or interpretation of the weighted means.
†Estimated variance of between study heterogeneity.

(reference category was non-medical) and study level proportion of participants with major depression were included as predictors in separate models.

To assess the heterogeneity of MDC across studies, we generated forest plots for the MDC95 estimate for each study and the random effects summary MDC95 estimate. We also reported the estimated between study variance (τ^2), percentage of variability caused by between study heterogeneity (I^2), and PIs.[48 49] All analyses were done in R (R version R 4.4.1[50] and R Studio version 2024.04.2+764[51]) with the lme4 and meta packages.[52 53]

Patient and public involvement

Sarah Markham, an experienced patient advisor and member of the BMJ's international Patient Advisory Panel, was a research team member from project inception. Sarah provided comments on the study protocol and content of this article and is a coauthor.

Results

Search results and selection of eligible studies

Database searches identified 9674 unique titles and abstracts, of which 9198 were excluded at the title and abstract stage and 289 at the full text stage. After removing 56 articles that used the same dataset as another included article, 131 unique primary studies remained, of which 100 (76%) contributed datasets. Five additional unpublished studies were contributed by authors. Seven datasets provided only total PHQ-9, PHQ-8, or PHQ-2 scores, without individual item scores, and were therefore excluded, and four datasets had individual item scores for the PHQ-2 but not the PHQ-9 or PHQ-8. Thus our analyses included 94 studies for the PHQ-9 and PHQ-8, and 98 studies for the PHQ-2 (fig 1).

Included studies and sample characteristics

Mean PHQ-9 (n=42 548 participants), PHQ-8 (n=42 592), and PHQ-2 (n=44 085) scores were 4.96 (SD 5.24), 4.82 (4.99), and 1.18 (1.56), respectively. Table 1 shows unweighted characteristics of participants included in the PHQ-9 analyses (PHQ-8 and PHQ-2 analyses are in supplementary tables 1 and 2, respectively). Across PHQ-9, PHQ-8, and PHQ-2 samples, mean age was 49 years (SD 17) and 60% of participants were women. Overall, 10% of participants had a positive major depression classification status (range 1-57% across studies). Participants were recruited from non-medical (n=14), primary care (n=28), outpatient specialty (n=38), inpatient (n=13), and mixed (n=5) settings. Supplementary Table 3a shows participant characteristics for the primary studies included in our study, and supplementary table 3b shows the characteristics of the eligible studies that did not provide data or provided total scores but not item level data.

Estimated overall and subgroup MDC

Pooled Cronbach's alpha was 0.84 for the PHQ-9 and PHQ-8, and 0.74 for the PHQ-2. Supplementary table 3a shows study specific Cronbach's alpha values. Table 2 shows the pooled estimates of MDC95, MDC90, and MDC67 for the PHQ-9, PHQ-8, and PHQ-2. Overall MDC95 estimates were 5.72 points (95% CI 5.54 to 5.90, 95% PI 4.00 to 7.44, I^2=98%, τ^2=0.76) for the PHQ-9, 5.51 points (95% CI 5.33 to 5.68, 95% PI 3.87 to 7.15, I^2=98%, τ^2=0.69) for the PHQ-8, and 2.26 points (95% CI 2.15 to 2.37, 95% PI 1.20 to 3.32, I^2=96%, τ^2=0.29) for the PHQ-2. Figure 2 shows MDC95 estimates for the PHQ-9 for each study, and pooled MDC95 values. Supplementary figures 1 and 2 show corresponding forest plots for the PHQ-8 and PHQ-2, respectively.

Table 3 shows subgroup estimates for MDCs for the PHQ-9, PHQ-8, and PHQ-2 by participant age, sex, and recruitment setting. Supplementary figures 3a-j, figures 4a-j, and figures 5a-j show forest plots of subgroup estimates for the PHQ-9, PHQ-8, and PHQ-2, respectively. MDC95 estimates for the PHQ-9 were similar across age groups. Table 4 shows a slightly higher MDC95 value for women (5.82 points, 95% CI 5.62 to 6.03, 95% PI 4.02 to 7.64) than men (5.53 points, 95% CI 5.33 to 5.73, 95% PI 3.84 to 7.22; difference 0.29 points, 95% CI 0.21 to 0.38). Subgroup results were similar for the PHQ-8 and PHQ-2.

Table 5 shows meta-regression analyses, based on recruitment setting and proportion of participants

Table 3 | Estimates of minimal detectable change with 95% (MDC95), 90% (MDC90), and 67% (MDC67) confidence for Patient Health Questionnaire-9 (PHQ-9), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-2 (PHQ-2) by age, sex, and recruitment setting subgroups

Scale and subgroup	No of participants (studies)*	Mean (SD) score †	MDC95		MDC90		MDC67	
			Estimate (95% CI)	95% PI	Estimate (95% CI)	95% PI	Estimate (95% CI)	95% PI
PHQ-9								
Age (years):								
18-29	6672 (67)	6.45 (4.83)	5.59 (5.34 to 5.84)	3.72 to 7.46	4.68 (4.47 to 4.89)	3.11 to 6.25	2.85 (2.73 to 2.98)	1.90 to 3.81
30-44	11 268 (85)	6.54 (5.26)	5.60 (5.40 to 5.80)	3.92 to 7.28	4.69 (4.52 to 4.86)	3.28 to 6.10	2.86 (2.76 to 2.96)	2.00 to 3.72
45-64	16 157 (87)	6.15 (5.22)	5.67 (5.47 to 5.88)	3.88 to 7.46	4.75 (4.58 to 4.92)	3.25 to 6.25	2.90 (2.79 to 3.00)	1.98 to 3.81
≥65	8155 (72)	5.12 (4.61)	5.63 (5.37 to 5.88)	3.66 to 7.59	4.71 (4.50 to 4.93)	3.06 to 6.36	2.87 (2.74 to 3.00)	1.87 to 3.88
Sex:								
Women	25 013 (94)	6.54 (5.28)	5.83 (5.64 to 6.03)	4.02 to 7.64	4.89 (4.72 to 5.05)	3.37 to 6.40	2.98 (2.88 to 3.08)	2.05 to 3.90
Men	17 512 (83)	5.48 (5.05)	5.53 (5.33 to 5.73)	3.84 to 7.22	4.63 (4.47 to 4.80)	3.22 to 6.05	2.82 (2.72 to 2.92)	1.96 to 3.69
Setting:								
Non-medical	15 552 (14)	5.64 (5.07)	5.39 (4.82 to 5.96)	3.22 to 7.56	4.51 (4.04 to 4.99)	2.70 to 6.33	2.75 (2.46 to 3.04)	1.64 to 3.86
Primary care	13 983 (28)	6.17 (5.24)	5.58 (5.30 to 5.86)	4.11 to 7.04	4.67 (4.44 to 4.91)	3.45 to 5.90	2.85 (2.70 to 2.99)	2.10 to 3.59
Outpatient specialty	9822 (39)	6.10 (5.31)	5.72 (5.44 to 5.99)	4.05 to 7.39	4.79 (4.56 to 5.02)	3.39 to 6.19	2.92 (2.78 to 3.06)	2.07 to 3.77
Inpatient	3061 (15)	7.01 (5.66)	6.48 (6.05 to 6.92)	4.82 to 8.15	5.43 (5.07 to 5.79)	4.04 to 6.82	3.31 (3.09 to 3.53)	2.46 to 4.16
PHQ-8								
Age (years):								
18-29	6676 (67)	6.24 (4.56)	5.37 (5.14 to 5.61)	3.60 to 7.15	4.50 (4.30 to 4.70)	3.02 to 5.99	2.74 (2.62 to 2.87)	1.84 to 3.65
30-44	11 280 (85)	6.35 (5.00)	5.36 (5.16 to 5.55)	3.75 to 6.97	4.49 (4.33 to 4.65)	3.14 to 5.83	2.74 (2.64 to 2.83)	1.91 to 3.56
45-64	16 181 (87)	5.99 (4.98)	5.46 (5.27 to 5.66)	3.74 to 7.19	4.58 (4.41 to 4.74)	3.13 to 6.02	2.79 (2.69 to 2.89)	1.91 to 3.67
≥65	8158 (72)	4.97 (4.38)	5.45 (5.20 to 5.69)	3.54 to 7.35	4.56 (4.36 to 4.77)	2.97 to 6.16	2.78 (2.65 to 2.91)	1.81 to 3.75
Sex:								
Women	25 039 (94)	6.34 (5.02)	5.61 (5.42 to 5.79)	3.90 to 7.32	4.70 (4.54 to 4.85)	3.26 to 6.13	2.86 (2.77 to 2.96)	1.99 to 3.74
Men	17 530 (83)	5.32 (4.79)	5.33 (5.14 to 5.52)	3.70 to 6.96	4.47 (4.31 to 4.63)	3.10 to 5.83	2.72 (2.63 to 2.82)	1.89 to 3.55
Setting:								
Non-medical	15 568 (14)	5.46 (4.83)	5.17 (4.63 to 5.71)	3.11 to 7.23	4.33 (3.88 to 4.78)	2.60 to 6.06	2.64 (2.36 to 2.91)	1.59 to 3.69
Primary care	13 998 (28)	5.96 (4.97)	5.36 (5.10 to 5.62)	4.00 to 6.72	4.49 (4.27 to 4.71)	3.35 to 5.63	2.74 (2.60 to 2.87)	2.04 to 3.43
Outpatient specialty	9831 (39)	5.92 (5.04)	5.51 (5.25 to 5.77)	3.92 to 7.10	4.62 (4.40 to 4.83)	3.28 to 5.95	2.81 (2.68 to 2.95)	2.00 to 3.63
Inpatient	3065 (15)	6.82 (5.39)	6.25 (5.85 to 6.65)	4.71 to 7.79	5.23 (4.90 to 5.57)	3.94 to 6.52	3.19 (2.99 to 3.40)	2.40 to 3.98
PHQ-2								
Age (years):								
18-29	6765 (68)	1.50 (1.47)	2.24 (2.09 to 2.39)	1.19 to 3.28	1.87 (1.75 to 2.00)	1.00 to 2.75	1.14 (1.06 to 1.22)	0.61 to 1.68
30-44	11 688 (89)	1.52 (1.58)	2.11 (1.98 to 2.23)	1.10 to 3.12	1.76 (1.66 to 1.87)	0.92 to 2.61	1.08 (1.01 to 1.14)	0.56 to 1.59
45-64	16 919 (91)	1.47 (1.58)	2.21 (2.08 to 2.34)	1.10 to 3.32	1.85 (1.74 to 1.96)	0.92 to 2.78	1.13 (1.06 to 1.19)	0.56 to 1.70
≥65	8396 (73)	1.20 (1.51)	2.18 (2.01 to 2.35)	0.95 to 3.42	1.83 (1.69 to 1.97)	0.79 to 2.86	1.11 (1.03 to 1.20)	0.48 to 1.74
Sex:								
Women	25 914 (98)	1.55 (1.60)	2.24 (2.12 to 2.35)	1.18 to 3.29	1.87 (1.78 to 1.97)	0.99 to 2.75	1.14 (1.08 to 1.20)	0.60 to 1.68
Men	18 147 (87)	1.29 (1.52)	2.21 (2.08 to 2.33)	1.16 to 3.25	1.85 (1.74 to 1.95)	0.97 to 2.72	1.13 (1.06 to 1.19)	0.59 to 1.66
Setting:								
Non-medical	15 675 (14)	1.33 (1.50)	2.05 (1.74 to 2.35)	0.90 to 3.19	1.71 (1.46 to 1.97)	0.76 to 2.67	1.04 (0.89 to 1.20)	0.46 to 1.63
Primary care	14 126 (29)	1.59 (1.60)	2.31 (2.11 to 2.52)	1.22 to 3.40	1.94 (1.76 to 2.11)	1.03 to 2.85	1.18 (1.07 to 1.29)	0.63 to 1.74
Outpatient specialty	10 923 (42)	1.37 (1.59)	2.19 (2.02 to 2.35)	1.18 to 3.20	1.83 (1.70 to 1.97)	0.99 to 2.68	1.12 (1.03 to 1.20)	0.60 to 1.63
Inpatient	3231 (16)	1.61 (1.74)	2.52 (2.24 to 2.81)	1.45 to 3.59	2.11 (1.87 to 2.35)	1.21 to 3.01	1.29 (1.14 to 1.43)	0.74 to 1.83

CI=confidence interval; PI= prediction interval; SD=standard deviation.
*Subgroups with only one or two participants or zero variance were excluded to ensure valid estimation of pooled effects (PHQ-8 and PHQ-9: age 18-29 years, n=1 (6 studies), n=2 (3 studies); age 30-44 years, n=1 (1 study), n=2 (2 studies); age 45-64 years, n=2 (2 studies); age ≥65 years, n=1 (3 studies), n=2 (1 study), n=3 (1 study, var=0); PHQ-2: age 18-29 years, n=1 (7 studies), n=2 (3 studies), n=3 (1 study, var=0), n=4 (1 study, var=0); age 30-44 years, n=1 (1 study), n=2 (2 studies); age 45-64 years, n=2 (2 studies); age ≥65 years, n=1 (3 studies), n=2 (1 study), n=3 (2 studies, var=0), n=6 (1 study, var=0)).
†Pooled estimates with restricted maximum likelihood random effects meta-analysis.

with major depression. Compared with participants from non-medical settings (5.39 points, 95% CI 4.82 to 5.96, 95% PI 3.22 to 7.56), the MDC95 for PHQ-9 was higher among participants from inpatient settings (6.48 points, 95% CI 6.05 to 6.92, 95% PI 4.82 to 8.15; difference 1.10 points, 95% CI 0.47 to 1.74). For every 10% increase in the proportion of participants with major depression, the MDC95 increased on average by 0.40 points for the PHQ-9 (95% CI 0.25 to 0.55), 0.35 points for the PHQ-8 (0.20 to 0.49), and 0.17 points for the PHQ-2 (0.07 to 0.27). Sensitivity analyses based on the Hartung-Knapp approach for estimating CIs and PIs were consistent with the main analyses (supplementary tables 4 and 5).

Discussion

Principal findings

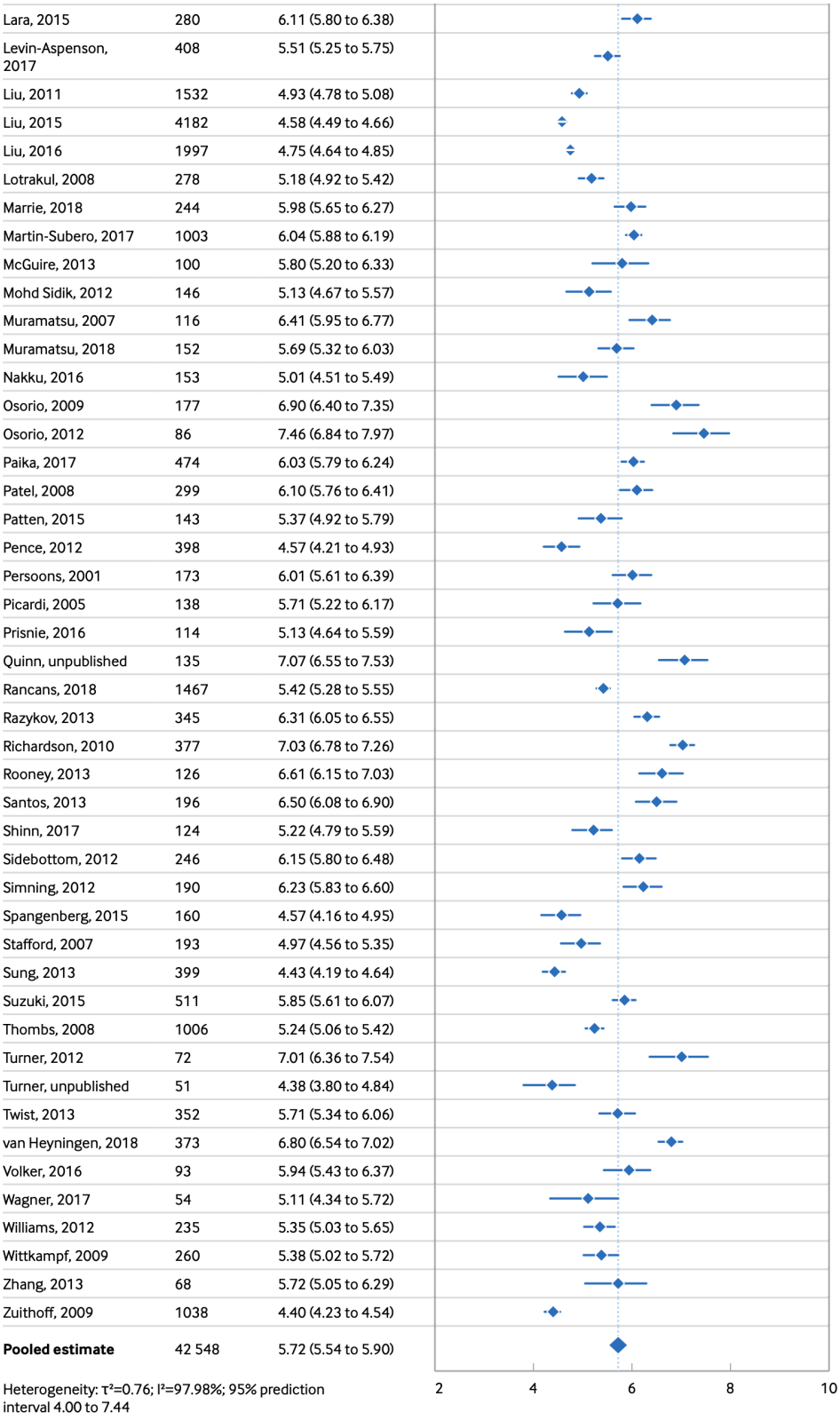
We estimated the MDC95, MDC90, and MDC67 values for the PHQ-9, PHQ-8, and PHQ-2. MDC95 estimates were 5.7 points for the PHQ-9, 5.5 points for the PHQ-8, and 2.3 points for the PHQ-2. Findings for subgroup analyses were consistent across PHQ-9, PHQ-8, and PHQ-2 analyses. MDCs were slightly lower for men than women, and about one point higher for medical inpatients compared with non-medical study participants for the PHQ-9 and PHQ-8 (and about 0.5 points for the PHQ-2). In our meta-regression analysis, the estimated MDC95 for the PHQ-9 was 0.40 points higher for each 10% increase in the proportion of study

Meta-analysis results for MDC95 of Patient Health Questionnaire-9 (PHQ-9)

Pooled MDC95 for PHQ-9 scores based on individual participant data meta-analysis

MDC95 = 1.96 × √2 × SEM, where SEM (standard error of measurement) was calculated as SD (standard deviation) × √(1 − reliability)

Study	No of participants	MDC95 (95% CI)
Akena, 2013	91	7.09 (6.60 to 7.50)
Amoozegar, 2017	203	5.77 (5.45 to 6.05)
Amtmann, 2015	164	5.67 (5.23 to 6.07)
Arroll, 2010	2528	4.80 (4.69 to 4.91)
Ayalon, 2010	151	5.40 (4.81 to 5.97)
Azah, 2005	180	5.74 (5.33 to 6.11)
Baron, 2017	851	5.67 (5.49 to 5.85)
Beraldi, 2014	116	5.42 (5.02 to 5.76)
Bernstein, 2018	240	5.33 (5.02 to 5.62)
Bhana, 2015	679	5.37 (5.13 to 5.60)
Bombardier, 2012	160	6.62 (6.19 to 7.02)
Buji, 2018	130	5.36 (4.84 to 5.83)
Chagas, 2013	84	6.90 (6.30 to 7.37)
Chibanda, 2016	264	7.15 (6.88 to 7.38)
Cholera, 2014	397	7.00 (6.73 to 7.26)
Conway, 2016	26	5.17 (4.08 to 6.01)
de la Torre, 2016	257	6.76 (6.39 to 7.08)
de Man-van Ginkel, 2012	382	7.35 (7.05 to 7.60)
Eack, 2006	48	5.97 (5.36 to 6.44)
Fiest, 2014	169	5.27 (4.88 to 5.61)
Fischer, 2014	194	4.82 (4.47 to 5.15)
Fisher, 2016	357	4.45 (4.16 to 4.71)
Garabiles, unpublished	99	5.71 (5.05 to 6.25)
Gelaye, 2014	923	6.72 (6.54 to 6.91)
Gholizadeh, 2019	79	6.97 (6.17 to 7.68)
Gjerdingen, 2009	419	4.90 (4.63 to 5.15)
Grafe, 2004	494	5.80 (5.60 to 5.99)
Green, 2017	176	5.38 (5.07 to 5.64)
Green, 2018	192	7.72 (7.36 to 8.05)
Grool, 2011	477	4.02 (3.75 to 4.27)
Hahn, 2006	211	6.03 (5.74 to 6.31)
Hantsoo, 2017	321	4.64 (4.33 to 4.92)
Haroz, 2017	132	6.28 (5.78 to 6.73)
Henkel, 2004	430	5.18 (4.97 to 5.38)
Hides, 2007	103	6.76 (6.28 to 7.17)
Hitchon, 2019	148	5.41 (5.03 to 5.76)
Hobfoll, 2011	144	5.16 (4.69 to 5.58)
Hyphantis, 2011	213	6.03 (5.70 to 6.33)
Hyphantis, 2014	349	6.79 (6.51 to 7.05)
Inagaki, 2013	104	4.84 (4.27 to 5.39)
Janssen, 2016	4695	3.83 (3.76 to 3.90)
Khamseh, 2011	122	7.37 (6.95 to 7.70)
Kiely, 2014	822	4.55 (4.37 to 4.72)
Kim, 2017	3071	3.51 (3.41 to 3.62)
Kohrt, 2016	125	6.06 (5.66 to 6.41)
Kwan, 2012	113	6.66 (6.01 to 7.26)
Lambert, 2015	147	5.58 (5.21 to 5.91)
Lamers, 2008	104	5.76 (4.57 to 6.95)



Article DOI: 10.1136/bmj-2026-100119 • Download data
CI=confidence interval; MDC95=minimal detectable change based on 95% confidence of change beyond measurement error

Fig 2 | Meta-analysis results for MDC95 of Patient Health Questionnaire-9 (PHQ-9). An interactive version of this graphic and downloadable data are available at https://public.flourish.studio/visualisation/29605896/

Table 4 | Age and sex subgroup differences in estimates of minimal detectable change with 95% confidence (MDC95) for Patient Health Questionnaire-9 (PHQ-9), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-2 (PHQ-2)

Subgroup*	PHQ-9		PHQ-8		PHQ-2	
	No of studies (participants)	MDC95 (95% CI)	No of studies (participants)	MDC95 (95% CI)	No of studies (participants)	MDC95 (95% CI)
Age (years)						
45-64 (reference)	61 (11 140)	5.72 (5.50 to 5.94)	61 (11 161)	5.50 (5.29 to 5.71)	62 (11 692)	2.27 (2.12 to 2.42)
18-29	61 (5604)	5.54 (5.28 to 5.79)	61 (5608)	5.31 (5.07 to 5.56)	62 (5692)	2.20 (2.04 to 2.35)
Difference†	—	-0.12 (-0.34 to 0.10)	—	-0.13 (-0.35 to 0.09)	—	-0.06 (-0.19 to 0.08)
45-64 (reference)	79 (15 904)	5.65 (5.45 to 5.86)	79 (15 927)	5.44 (5.24 to 5.63)	83 (16 658)	2.19 (2.06 to 2.32)
30-44	79 (10 481)	5.56 (5.36 to 5.77)	79 (10 492)	5.32 (5.12 to 5.52)	83 (10 894)	2.10 (1.97 to 2.23)
Difference†	—	-0.07 (-0.18 to 0.04)	—	-0.09 (-0.21 to 0.02)	—	-0.09 (-0.17 to -0.02)
45-64 (reference)	72 (14 626)	5.69 (5.47 to 5.90)	72 (14 649)	5.48 (5.27 to 5.69)	73 (15 208)	2.22 (2.07 to 2.36)
≥65	72 (8155)	5.63 (5.37 to 5.88)	72 (8158)	5.45 (5.20 to 5.69)	73 (8396)	2.18 (2.01 to 2.35)
Difference†	—	-0.06 (-0.21 to 0.09)	—	-0.03 (-0.17 to 0.11)	—	-0.04 (-0.16 to 0.07)
Sex						
Women (reference)	83 (22 552)	5.82 (5.62 to 6.03)	83 (22 575)	5.60 (5.40 to 5.80)	87 (23 433)	2.21 (2.09 to 2.33)
Men	83 (17 512)	5.53 (5.33 to 5.73)	83 (17 530)	5.33 (5.14 to 5.52)	87 (18 147)	2.21 (2.08 to 2.33)
Difference†	—	-0.29 (-0.38 to -0.21)	—	-0.27 (-0.35 to -0.18)	—	-0.06 (-0.11 to -0.02)

CI=confidence interval.
*Subgroup comparisons and differences were calculated based on data from studies that included participants in both comparison groups.
†Estimated differences diverge somewhat in some cases from what is obtained by subtracting the estimate for each subgroup from the reference group. This result is because of different relative weighting of some studies in models that pooled MDC estimates across studies versus models that pooled the difference between subgroups from each study.

participants with major depression (0.35 points for the PHQ-8 and 0.17 points for the PHQ-2), starting at the overall mean of 10%. Thus in a setting where 50% of patients meet the criteria for major depression, for example, an MDC for the PHQ-9 of seven points might be used (5.7 points + 4 × 0.4 = 7.3 points). We found substantial heterogeneity across all main analyses and most subgroup analyses.

Comparison with other studies

To the best of our knowledge, our study is the first to publish analyses of the MDC for the PHQ-9 and PHQ-2. Our estimated MDC for the PHQ-8 (5.5 points) was similar to the only other study that has evaluated the MDC for the PHQ-8 (5.2 points, 95% CI 5.1 to 5.3; 2571 individuals with systemic sclerosis).[26] Only two primary studies have used anchor based approaches to estimate minimal important differences for the PHQ. Bauer-Staeb et al[10] reported an average minimal important difference of 3.7 points, with the minimal important difference defined as the threshold for a 50% probability of reporting "feeling better". Credible minimal important differences should be based on change ratings with correlations of 0.50 or higher

Table 5 | Simple meta-regressions of minimal detectable change with 95% confidence (MDC95) on study level variables for Patient Health Questionnaire-9 (PHQ-9), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-2 (PHQ-2)

Scale and variable	Raw regression coefficient (95% CI)	τ^2*	I^2 (%)†
PHQ-9			
10% increase in proportion of participants with major depression	0.40 (0.25 to 0.55)	0.58	97.3
Setting‡:			
Non-medical	—	—	—
Primary care	0.20 (-0.36 to 0.75)	0.72	97.7
Outpatient specialty	0.33 (-0.19 to 0.86)	—	—
Inpatient	1.10 (0.47 to 1.74)	—	—
PHQ-8			
10% increase in proportion of participants with major depression	0.35 (0.20 to 0.49)	0.55	97.3
Setting‡:			
Non-medical	—	—	—
Primary care	0.20 (-0.33 to 0.72)	0.64	97.5
Outpatient specialty	0.35 (-0.15 to 0.85)	—	—
Inpatient	1.09 (0.49 to 1.69)	—	—
PHQ-2			
10% increase in proportion of participants with major depression	0.17 (0.07 to 0.27)	0.26	95.6
Setting‡:			
Non-medical	—	—	—
Primary care	0.27 (-0.08 to 0.62)	0.28	95.6
Outpatient specialty	0.15 (-0.19 to 0.48)	—	—
Inpatient	0.48 (0.07 to 0.89)	—	—

CI=confidence interval.
*Estimated amount of (residual) heterogeneity.
†Total variability in the effect size estimates that can be attributed to heterogeneity among the true effects.
‡Estimate represents the increase in MDC95 compared with the non-medical setting.

Table 6 | Thresholds for minimal detectable change with 95% (MDC95), 90% (MDC90), and 67% (MDC67) confidence for Patient Health Questionnaire-9 (PHQ-9), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-2 (PHQ-2) based on standard and alternative approaches

Estimated MDC thresholds	PHQ-9	PHQ-8	PHQ-2
Standard			
Pooled MDC95	6	6	2 to 3
Alternatives			
Pooled MDC90	5	5	2
Pooled MDC67	3	3	1 to 2
MDC95 prediction interval upper limit	7 to 8	7 to 8	3 to 4

The pooled MDC95 is typically used in practice. Alternatively, other methods for setting thresholds could be selected and would increase the likelihood of identifying change but reduce certainty (MDC67 and MDC90) or increase certainty but decrease the likelihood of classifying change when it has occurred (upper limit of prediction interval). Thresholds can be adjusted for different settings; a higher threshold may be more appropriate among people in specialty mental health treatment settings.

with the measure being assessed,[17 54] however, and this value was somewhat lower in this study. Also, the reported minimal important difference range across PHQ-9 scores was wide (0-14 points) and no estimates of precision were provided. Kounali et al[11] reported that the minimal important difference for PHQ-9 ranged from 1.7 to 2.4 points, a score change threshold that best distinguished participants who reported "feeling better" from those who reported "feeling the same". No correlation coefficients, however, were reported to show that changes in PHQ-9 scores were meaningfully correlated with the anchor. Also, the precision of the minimal important difference estimate was not reported, but each analysis included only 83-149 participants. Estimates of minimal important difference from both studies were substantially below our estimated MDC values, which suggests that the values may be too small to be reliably distinguished from measurement error.

Clinical and research implications

The minimal important difference, which represents the smallest change that a patient would likely perceive as important, is used along with patient perception of change to guide treatment decisions. As highlighted in guidelines for measurement based care,[1] however, when the MDC is greater than the minimal important difference, change that exceeds the minimal important difference might be because of measurement error rather than real change. In this case, the MDC should be used as the change threshold. In contrast, if the MDC is smaller than the minimal important difference, then the minimal important difference value can guide interpretation.[55]

Jacobson and Truax introduced the reliable change index,[47] equivalent to MDC95, over 30 years ago. Since then, MDC95 has been the standard for evaluating reliable change. Less stringent thresholds (eg, MDC90 and MDC67) could also be used to inform the likelihood that change has occurred. Historically, the pooled MDC has been used to guide decisions. Alternatively, when MDC is calculated in primary studies, the upper end of a confidence interval around the mean would provide greater certainty that change has occurred. Given the heterogeneity across studies that we identified in our IPDMA, using the upper end of a PI, rather than the mean, would provide greater certainty of change. The PI provides the range of estimates that might be generated in one future study.[56]

Table 6 shows possible MDC thresholds that could be used in practice. Among the three questionnaires, the PHQ-9 is the most used in clinical settings. The UK NHS's Talking Therapies programme[18] and other programmes[21-23] define reliable change, equivalent to MDC95, for the PHQ-9 as a change in score of ≥6 points. We found a pooled MDC95 of 5.72 points on the PHQ-9, which is consistent. The mean baseline PHQ-9 score of patients in the UK NHS's Talking Therapies programme was 14-15,[19 57] however, suggesting that about 50% of participants may have major depression.[58] Considering the possible influence of baseline severity, our analysis suggests that an MDC95 of seven points could be used. Our finding that MDC estimates were similar in primary care and outpatient specialty medical settings suggests that these estimates apply to people with and without co-existing medical conditions, although a higher threshold might be considered in inpatient medical settings.

The PHQ-2 is typically used as a screening tool but can sometimes be used as a brief tool to monitor symptom changes in clinical settings.[14-16] The PHQ-8 is often used in research because item 9 of the PHQ-9 generates high numbers of false positive signals about the risk of suicide. Evaluations of status change in research could incorporate the minimal important difference and MDC to classify participants as changed if both thresholds were exceeded, not changed if neither was exceeded, or uncertain if the minimal important difference was exceeded but not the MDC. Recommended values for PHQ-8 and PHQ-2 MDCs have not been established, and our findings can inform recommendations.

When research studies assess whether group differences or changes, rather than individual differences, are clinically meaningful, the minimal important difference is used. The precision of these magnitudes for groups is assessed with 95% CIs or PIs, which depend chiefly on the adequacy of the sample size,[7 59 60] rather than MDCs.

Strengths and limitations of this study

In this meta-analysis, we estimated MDCs for the PHQ-9, PHQ-8, and PHQ-2. We used a large and diverse individual participant data sample, many times larger than previous studies. Furthermore, the use of individual level data allowed us to conduct subgroup analyses that are not typically possible in individual primary studies or when using study level data in aggregate data meta-analysis.

Our study had some limitations. Primary data from 31 of 131 published eligible datasets (24%) were not obtained. Of the studies that contributed datasets, seven did not provide item level scores for the PHQ-2 and 11 did not provide item level scores for the PHQ-8 and PHQ-9 and therefore could not be included. In our previous IPDMAs on depression screening tool accuracy, however, we found that published sensitivity

and specificity from studies that did not contribute data were similar to the results from those that provided data.[30 61 62]

Standard error of measurement was estimated with Cronbach's alpha, a measure of internal consistency, rather than by test-retest reliability.[63] A meta-analysis on PHQ-9 reliability reported a pooled Cronbach's alpha of 0.86 from 60 studies, similar to our estimate of 0.84, and a pooled test-retest reliability of 0.82 from eight studies.[44] Our estimates for PHQ-8 (0.84) and PHQ-2 (0.74) Cronbach's alpha were also similar to estimates of Cronbach's alpha and test-retest reliability reported in previous studies for the PHQ-8 (Cronbach's alpha 0.85; test-retest over a mean of 11 days, 0.83) and PHQ-2 (Cronbach's alpha 0.76; test-retest over two weeks, 0.70).[64 65] Thus, although possible, it seems unlikely that using Cronbach's alpha for PHQ-9 and PHQ-8 resulted in MDC estimates that were substantively different than if we had obtained test-retest reliability estimates. It is more likely that using Cronbach's alpha could have influenced MDC estimates for the two item PHQ-2 because Cronbach's alpha depends on the number of items in a scale and could have been lower than test-retest estimates.[66]

We found substantial heterogeneity. Heterogeneity is almost always present in the evaluations of tests, but the test results reflect real differences that affect scale properties across studies.[67] Also, because we did not have data on score changes, we could not compare our estimated MDCs with minimal important differences from the same dataset. We also did not incorporate a risk of bias assessment, given the absence of a validated tool for MDC studies. The COSMIN (Consensus Based Standards for the Selection of Health Measurement Instruments) Risk of Bias Checklist,[8] for example, only determines whether internal consistency reliability of a measure was evaluated with Cronbach's alpha but does not establish factors associated with the risk of bias for Cronbach's alpha.

Our study was a secondary analysis of an existing dataset and only included studies published up to May 2018. MDC is considered a stable metric, however, and therefore its estimated value would not likely change meaningfully with the inclusion of data from more recent studies. The main change in assessment methods in recent years has been the increased use of remote assessments; we previously analysed >50 000 participants with scores on three depression screening tools and found no differences in scores by administration modality.[68] Also, we conducted subgroup analyses by age, sex, and setting, but we could not analyse additional factors, such as race or ethnic group or socioeconomic status, because these data were not included in our dataset.

Because our meta-regression analysis was based on study level data, observed associations may be influenced by ecological confounding from other study level characteristics. Also, although many participants in the UK NHS's Talking Therapies programme are referred by general practitioners, others are self-referred and may differ somewhat from our sample, which could limit generalisability. Finally, we used a classical test theory approach because the results of a simulation study that compared classical test with item response theory approaches found that classical test theory more accurately identified change for measures with <20 items.[69] This approach, however, may have limitations because of assumptions such as unidimensionality and approximate tau equivalence.

Conclusions

MDC thresholds can be used in conjunction with patient perspectives to evaluate treatment progress. We found that the pooled MDC95 was 5.7 points for the PHQ-9, 5.5 points for the PHQ-8, and 2.3 points for the PHQ-2. Across the three scales, the effects of individual level and study level factors were similar. MDC was higher in inpatient medical settings than in other healthcare settings. Sex and age had minimal or no association with MDC. About 10% of participants across the included studies were classified as having major depression, and we estimated that MDC increased by about 0.4 points for every 10% increase in the proportion of participants with major depression. Our pooled estimate of MDC95 is consistent with the six point change in PHQ-9 scores that is used in the UK to evaluate improvement in symptoms, although using seven points might be considered, given the likely proportion of patients with major depression in clinical settings. Other methods for setting thresholds could be selected and would increase the likelihood of identifying change but reduce certainty (MDC67, MDC90) or increase certainty but increase the likelihood of not recognising change (upper end of 95% PI).

AUTHOR AFFILIATIONS

[1]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, QC, Canada

[2]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, QC, Canada

[3]School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]McGill University Libraries, Montréal, QC, Canada

[5]Department of Clinical, Neuro, and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[6]Hull York Medical School and the Department of Health Sciences, University of York, Heslington, York, UK

[7]Department of Applied Statistics, Social Science, and Humanities, New York University, New York, NY, USA

[8]Department of Medicine, Department of Epidemiology and Population Health, Department of Biomedical Data Science, Department of Statistics, Stanford University, Stanford, CA, USA

[9]Department of Biostatistics and Health Informatics, King's College London, London, UK

[10]Department of Community Health Sciences, University of Calgary, Calgary, AB, Canada

[11]Women's College Hospital and Research Institute, University of Toronto, Toronto, ON, Canada

[12]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[13]Department of Medicine, McGill University, Montréal, QC, Canada

[14]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, QC, Canada

[15]Department of Psychiatry, McGill University, Montréal, QC, Canada

Contributors: YWa, YWu, NPG-D, BLe, PC, SG, DH, JPAI, SM, SBP, SNV, RCZ, AB, and BDT were responsible for the study conception and design. JTB was responsible for the database searches. YWa, YWu, NPG-D, BLe, and BDT contributed to data extraction, coding, evaluation of included studies, and data synthesis. YWa, YWu, AB, and BDT contributed to data analysis and interpretation. YWa, YWu, AB, and BDT drafted the manuscript. YWa, YWu, NPG-D, BLe, PC, SG, DH, JPAI, SM, SBP, SNV, RCZ, AB, and BDT provided a critical review and approved the final manuscript. BDT had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analyses and is the guarantor. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted. Contributions of members of the DEPRESsion Screening Data (DEPRESSD) PHQ Collaboration: PMB, DN, YS, CH, AK, MIm, DBR, MA, MJC, LAK, and YT contributed to data extraction and coding. DA, BA, LA, HRB, CNB, CHB, RIB, PB, GC, MHC, LFC, JCNC, DC, KC, AC, YC, FMD, JMMG, JRF, SF, FHF, JRWF, DSSF, BG, LG, FG-S, EPG, CGG, BJH, PH, LHa, EEH, MH, UH, LHi, SEH, TH, MI-G, MIn, KI, HJJ, NJ, MEK, KMK, BAK, YK, MAL, HFL-A, S-IL, ML, BLö, SRL, NPL, CL, RAM, LM, BPM, AM, SMS, TNM, KM, JEMN, LN, FLO, BWP, IP, AP, SLP, TJQ, ER, SDR, KR, AGR, ISS, MTS, JS, EHS, AS, LSp, LSt, SCS, KS, RHS, PLLT, TDT, AT, TVH, CMvdFC, LIW, JLW, JW, MAW, KWi, KWy, MY, QZZ, and YZ contributed primary datasets that were included in this study.

Funding: The DEPRESsion Screening Data (DEPRESSD) PHQ study was funded by the Canadian Institutes of Health Research (CIHR) (KRS-134297, PCG-155468, PJT-162206, PJT-183746). BDT was supported by a Tier 1 Canada Research Chair. SBP was supported by the Cuthbertson and Fischer Chair in Paediatric Mental Health. The primary studies (supplementary material references lists all of the primary studies) by Fiest et al, Patten et al, Amoozegar et al, and Prisnie et al were supported by the Cumming School of Medicine, University of Calgary, and Alberta Health Services through the Calgary Health Trust, as well as the Hotchkiss Brain Institute. NJ was supported by a Canada Research Chair in Neurological Health Services Research and an Alberta Innovates-Health Solutions (AIHS) Population Health Investigator Award. The primary study by Amtmann et al was supported by a grant from the Department of Education (NIDRR grant No H133B080025) and by the National Multiple Sclerosis Society (MB 0008). Collection of data for the study by Arroll et al was supported by a project grant from the Health Research Council of New Zealand. Data collection for the study by Ayalon et al was supported by a grant from Lundbeck International. JS was supported by funding from Universiti Sains Malaysia. The primary study by Baron et al was supported by the National Income Dynamics Study (NIDS) at the University of Cape Town, South Africa. NIDS is implemented by the Southern Africa Labour and Development Research Unit, and is funded by the Department of Planning, Monitoring, and Evaluation. The funding body was involved in the design of the primary study. The primary studies by Marrie et al and Bernstein et al were supported by CIHR (THC-135234) and Crohn's and Colitis Canada. CHB was supported in part by the Bingham Chair in Gastroenterology. The primary study by Bombardier et al was supported by the Department of Education, National Institute on Disability and Rehabilitation Research, Spinal Cord Injury Model Systems: University of Washington (grant No H133N060033), Baylor College of Medicine (grant No H133N060003), and University of Michigan (grant No H133N060032). The primary study by Buji et al was supported by grants from the Universiti Kebangsaan Malaysia Medical Centre (UKMMC) Fundamental Research Fund (FF-2015-051) and the Fundamental Research Grant Scheme by the Malaysian Ministry of Higher Education (FRGS/2/2014/SKK09/UKM/02/1). The primary study by Conway et al was supported by the Institute of Health and Biomedical Innovation at Queensland University of Technology and the Sigma Theta Tau International Honour Society of Nursing (ID 8580). The primary study by Chibanda et al was supported by a grant from Grand Challenges Canada (0087-04). The primary study by Eack et al was funded by the National Institute of Mental Health (NIMH) (R24 MH56858). The primary study by Fischer et al was funded by the German Federal Ministry of Education and Research (01GY1150). The primary study by Fisher et al was supported by grants from the National Health and Medical Research Council (APP1026550), the Australian Government Department of Social Services-Families, Housing, Community Services, and Indigenous Affairs, and the Victorian Department of Education and Early Childhood Development. JRWF was supported by a Monash Professorial Fellowship and the Jean Hailes Professorial Fellowship, which is supported by a grant to the Jean Hailes Foundation from the H and L Hecht Trust managed by perpetual trustees. TDT was supported by a Monash Bridging Postdoctoral Fellowship. BJH received support from a grant awarded by the Research and Development Administration Office, University of Macau (MYRG2015-00109-FSS). The primary study by Garabiles et al was supported by the Macao (SAR) Government, through the University of Macau RSKTO grants (MYRG-2014-111). The primary study by Gholizadeh et al was supported by University of Technology Sydney under UTS Research Reestablishment Grants. Data for the primary study by Gelaye et al were supported by a grant from the National Institutes of Health (NIH) (T37 MD001449). Collection of data for the primary study by Gjerdingen et al was supported by grants from the NIMH (R34 MH072925, K02 MH65919, P30 DK50456). BLö received research grants from Pfizer, Germany, and from the Medical Faculty of the University of Heidelberg, Germany (project 121/2000) for the study by Gräfe et al. BPM was supported by the Department of Defence (W81XWH-08-2- 0100/W81XWH-08-2-0102 and W81XWH-12- 2-0117/W81XWH-12-2-0121). The primary study by Green et al (2018) was supported by a grant from the Duke Global Health Institute (453-0751). The primary study by Grool et al was supported by a programme grant from the Netherlands Heart Foundation (2007B027). Collection of data provided by MH and KR was supported by the Federal Ministry of Education and Research (grants Nos 01 GD 9802/4 and 01 GD 0101) and by the Federation of German Pension Insurance Institute. The primary study by Hantsoo et al was supported by K23 MH107831-02, Brain and Behaviour Research Foundation NARSAD Young Investigator Award. The primary study by Haroz et al was supported by the United States Agency for International Development Victims of Torture Fund (AID-DFD A-00-08-00308). EEH was supported by a NIMH T32 predoctoral training grant (MH014592-38) and postdoctoral training grant (MH103210) during the conduct of the primary study. The primary study by Henkel et al was funded by the German Ministry of Research and Education. The primary study by Hides et al was funded by the perpetual trustees, Flora and Frank Leith Charitable Trust, Jack Brockhoff Foundation, Grosvenor Settlement, Sunshine Foundation, and Danks Trust. Collection of data for the primary study by Hobfoll et al was made possible in part from grants from NIMH (RO1 MH073687) and the Ohio Board of Regents. Collection of data for the primary study by Hyphantis et al (2014) was supported by a grant from the National Strategic Reference Framework, European Union, and the Greek Ministry of Education, Lifelong Learning, and Religious Affairs (ARISTEIA-ABREVIATE, 1259). The primary study by Inagaki et al was supported by the Ministry of Health, Labour, and Welfare, Japan. The primary study by Janssen et al was supported by the European Regional Development Fund as part of OP-ZUID, the Province of Limburg, Department of Economic Affairs of the Netherlands (310.041), Stichting the Weijerhorst, the Pearl String Initiative Diabetes, the Cardiovascular Centre Maastricht Cardiovasculair Research Institute Maastricht, School of Nutrition, Toxicology, and Metabolism, Stichting Annadal, and Health Foundation Limburg. The primary study by Khamseh et al was supported by a grant (M-288) from Tehran University of Medical Sciences. Collection of data for the primary study by Kiely et al was supported by the National Health and Medical Research Council (grant No 1002160) and Safe Work Australia. PB was supported by Australian Research Council Future Fellowship FT130101444. KMK was supported by funding from an Australian National Health and Medical Research Council fellowship (grant No 1088313). The primary study by Kim et al was supported by a grant from the Korean Mental Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (HM14C2567), an Institute for Information and Communications Technology Promotion grant funded by the Korean government (MSIP) (B0132- 15-1003: the development of skin adhesive patches for the monitoring and prediction of mental disorders), and by the Original Technology Research Programme for Brain Science through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (NRF-2016M3C7A1947307). The primary studies by Bhana et al, Kohrt et al, and Nakku et al were output of the PRogramme for Improving Mental health carE (PRIME) and were supported by the UK Department for International Development (201446). The views expressed do not necessarily reflect the UK Government's official policies. The primary study by Lambert et al was supported by Calvary Mater Newcastle (grant No 11-09). The primary study by Lamers et al was funded by the Netherlands Organisation for Health Research and Development (grant No 945-03-047). The primary study by Lara et al was supported by the Consejo Nacional de Ciencia y Tecnología/National Council for Science and Technology (CB-2009-133923-H). The primary study by Liu et al (2011) was funded by a grant from the National Health Research Institute, Republic of China (NHRI-EX97-9706PI). The primary study by Liu et al (2015) was supported by CIHR (MOP-114970). The primary study by Liu et al (2016) was supported by Shanghai Municipal Health and

Family Planning Commission Bureau-level Project-Preliminary Exploration of Depression Risk Prediction Model for Outpatients in General Hospitals (project No 2010105). The primary study by Lotrakul et al was supported by the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (grant No 49086). The primary study by Martin-Subero et al was supported in part by a grant from the Spanish Ministry of Health's Health Research Fund (Fondo de Investigaciones Sanitarias, project 97/1184). The primary study by Mohd Sidik et al was funded under the Research University Grant Scheme from Universiti Putra Malaysia, Malaysia and the Postgraduate Research Student Support Accounts of the University of Auckland, New Zealand. The primary study by Muramatsu et al (2007) was supported by an educational grant from Pfizer US Pharmaceutical. The primary study by Muramatsu et al (2018) was supported by grants from Niigata Seiryo University. FLO was supported by Productivity Grants (PQ-CNPq-2- No 301321/2016-7). The primary studies by Osório et al (2012) were funded by Reitoria de Pesquisa da Universidade de São Paulo (grant No 09.1.01689.17.7) and Banco Santander (grant No 10.1.01232.17.9). The primary study by Paika et al was supported by the European Economic Area (EEA) Financial Mechanism 2009-14 (EEA GR07/3767) and national funds as part of the programme Dissimilarity, Inequality, and Social Integration (132324/I4-25/8/2015). Collection of primary data for the study by Pence et al was provided by NIMH (R34MH084673). The primary study by Persoons et al was partly funded by a grant from the Belgian Ministry of Public Health and Social Affairs and supported by a limited grant from Pfizer Belgium. The primary study by Rancans et al was supported by the National Research Programme BIOMEDICINE 2014-17 (5.8.1.). Data for the study by Razykov et al were collected by the Canadian Scleroderma Research Group, which was funded by CIHR (FRN 83518), the Scleroderma Society of Canada, the Scleroderma Society of Ontario, the Scleroderma Society of Saskatchewan, Sclérodermie Québec, the Cure Scleroderma Foundation, Inova Diagnostics, Euroimmun, Fonds de recherche du Québec-Santé (FRQS), the Canadian Arthritis Network, and the Lady Davis Institute of Medical Research of the Jewish General Hospital, Montréal, QC. YC received support from NIMH (R24MH071604) and the Centers for Disease Control and Prevention (R49 CE002093). The primary study by Roch et al was supported by grants from German Pension Insurance Company (Bund). The primary study by Rooney et al was funded by the UK NHS Lothian Neuro-Oncology Endowment Fund. The primary study by Santos et al was funded by the National Programme for Centres of Excellence (PRONEX/FAPERGS/CNPq, Brazil). The primary study by Shinn et al was supported by grant NCI K07 CA 093512 and the Lance Armstrong Foundation. The primary study by Sidebottom et al was funded by a grant from the US Department of Health and Human Services, Health Resources and Services Administration (grant No R40MC07840). The research of Simning et al was supported in part by grants from the NIH (T32 GM07356), Agency for Healthcare Research and Quality (R36 HS018246), NIMH (R24 MH071604), and the National Centre for Research Resources (TL1 RR024135). The primary study by Spangenberg et al was supported by a junior research grant from the Medical Faculty, University of Leipzig. LS received PhD scholarship funding from the University of Melbourne. The primary study by Swartz et al was supported by grants from HSF Ontario (000392) and CIHR (1012404). RHS received salary support from a New Investigator Award and the HJ Barnett Award from the Heart and Stroke Foundation (HSF) Canada, the Canadian Partnership for Stroke Recovery, the Department of Medicine, Sunnybrook Health Sciences Centre (HSC) and University of Toronto, the Brill Chair in Neurology Sunnybrook HSC, and advisory board Hoffman LaRoche stock ownership FollowMD. The primary study by Thombs et al was done with data from the Heart and Soul Study. The Heart and Soul Study was funded by the Department of Veterans Epidemiology Merit Review Programme, the Department of Veterans Affairs Health Services Research and Development service, the National Heart Lung and Blood Institute (R01 HL079235), the American Federation for Ageing Research, the Robert Wood Johnson Foundation, and the Ischaemia Research and Education Foundation. Collection of data for the study by Turner et al (2012) was funded by a bequest from Jennie Thomas through the Hunter Medical Research Institute. The primary study by Twist et al was funded by the UK National Institute for Health Research under its Programme Grants for Applied Research Programme (grant reference No RP-PG-0606-1142). The primary study by van Heyningen et al was supported by the Medical Research Council of South Africa (415865), Cordaid Netherlands (project 103/10002 G Sub 7), and the Truworths Community Foundation Trust, South Africa. The primary study by Volker et al was supported by the Netherlands Organisation for Health Research and Development (ZonMw) and from Achmea SZ, a Dutch insurance company. The primary study by Wagner et al was supported by grants U10CA21661, U10CA180868, U10CA180822, and U10CA37422 from the National Cancer Institute. The study was also funded in part by a grant from the Pennsylvania Department of Health. The department specifically declaims responsibility for any analyses, interpretations, or conclusions of the primary study. Collection of data for the primary study by Williams et al was supported by an NIMH grant to LM (RO1-MH069666). The study by Wittkampf et al was funded by the Netherlands Organisation for Health Research and Development (ZonMw) Mental Health Programme (Nos 100.003.005 and 100.002.021) and the Academic Medical Centre, University of Amsterdam. Collection of data for the primary study by Zhang et al was supported by the European Foundation for Study of Diabetes, the Chinese Diabetes Society, Lilly Foundation, Asia Diabetes Foundation, and Liao Wun Yuk Diabetes Memorial Fund. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare no financial relationships with any organisations that might have an interest in the submitted work in the previous three years except: SNV reports royalties from UpToDate for authorship of materials related to depression and pregnancy. CNB declares that he has consulted to Abbvie Canada, Amgen Canada, Bristol Myers Squibb Canada, Celltrion, Eli Lilly Canada, Fresenius Kabi, JAMP Pharmaceuticals, Janssen Canada, Pendopharm Canada, Pfizer Canada, Sandoz Canada, and Takeda Canada. CNB has received grants from Abbvie Canada, Pfizer Canada, and Takeda Canada and also received unrestricted educational grants from Abbvie Canada, Boston Scientific, Bristol Myers Squibb, Eli Lilly, Ferring Canada, Fresenius Kabi, Janssen Canada, Organon Canada, Pfizer Canada, and Takeda Canada, as well as been on speaker's bureau of Abbvie Canada, Eli Lilly, Fresenius Canada, Janssen Canada, Pfizer Canada, and Takeda Canada, all outside the submitted work. MIn declares that he has received personal fees from Janssen, Viatris, Mochida, Yoshitomi, Otsuka, MSD, Sumitomo, Meiji, Takeda, Eisai, Nippon Kayaku, Kyowa Kirin, Sanwa Kagaku, and Shionogi. MTS declares that the primary study by Janssen et al was supported by unrestricted grants from Janssen, Novo Nordisk, and Sanofi. ER declares that he received grants, personal fees, and non-financial support from Gedeon Richter; personal fees and non-financial support from Lundbeck, Servier, and Janssen Cilag; and personal fees from Zentiva and Abbvie outside the submitted work. KI declares that she has received honorarium for speaker fees for educational lectures for Sanofi, Sunovion, Janssen, and Novo Nordisk. LIW declares that she receives personal fees from Celgene, outside the submitted work. SLP declares that she received salary support from Pfizer-Astella and Millennium, outside the submitted work. JCNC is a steering committee member or consultant, or both, of Astra Zeneca, Bayer, Lilly, MSD, and Pfizer. JCNC has received sponsorships and honorarium for giving lectures and providing consultancy, and her affiliated institution has received research grants from these companies. The contributions of BG were made as part of his official duties as a National Institutes of Health (NIH) federal employee, are in compliance with agency policy requirements, and are considered Works of the US Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or US Department of Health and Human Services. All authors declare no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: The research ethics committee of the Jewish General Hospital declared that research ethics approval was not required because the study involved IPDMAs of anonymised previously collected data. For each included dataset, however, we confirmed that the original study received ethics approval and that participants provided informed consent.

Data sharing: Requests to access data should be made to the corresponding authors.

Transparency: The manuscript's guarantor affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned and registered have been explained.

Dissemination to study participants and related patient and public communities: This study was a synthesis of previously collected data, and no specific dissemination to study participants or patient communities is planned.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 American Psychological Association. APA Guidelines on Measurement-Based Care. https://www.apa.org/about/policy/guidelines-measurement-based-care
2 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407-15. doi:10.1016/0197-2456(89)90005-6
3 Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;40:593-7. doi:10.1111/j.1475-6773.2005.00374.x
4 Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;7:541-6. doi:10.1016/j.spinee.2007.01.008
5 Mouelhi Y, Jouve E, Castelli C, et al. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health Qual Life Outcomes* 2020;18:136. doi:10.1186/s12955-020-01344-w
6 Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102-9. doi:10.1016/j.jclinepi.2007.03.012
7 van Kampen DA, Willems WJ, van Beers LWAH, et al. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *J Orthop Surg Res* 2013;8:40. doi:10.1186/1749-799X-8-40
8 Mokkink LB, Elsman EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Qual Life Res* 2024;33:2929-39. doi:10.1007/s11136-024-03761-6
9 Lynch CP, Cha EDK, Jenkins NW, et al. The Minimum Clinically Important Difference for Patient Health Questionnaire-9 in Minimally Invasive Transforaminal Interbody Fusion. *Spine (Phila Pa 1976)* 2021;46:603-9. doi:10.1097/BRS.0000000000003853
10 Bauer-Staeb C, Kounali DZ, Welton NJ, et al. Effective dose 50 method as the minimal clinically important difference: Evidence from depression trials. *J Clin Epidemiol* 2021;137:200-8. doi:10.1016/j.jclinepi.2021.04.002
11 Kounali D, Button KS, Lewis G, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. *Psychol Med* 2022;52:1875-82. doi:10.1017/S0033291720003700
12 Kroenke K, Wu J, Yu Z, et al. The Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS): initial validation in three clinical trials. *Psychosom Med* 2016;78:716-27. doi:10.1097/PSY.0000000000000322
13 Ayilara OF, White A, Dunleavy G, et al. Sensitivity of the Patient-Health Questionnaire (PHQ-9) items to change in three clinical trials in low- and middle-income countries. *J Affect Disord* 2026;393:120226. doi:10.1016/j.jad.2025.120226
14 Löwe B, Kroenke K, Gräfe K. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J Psychosom Res* 2005;58:163-71. doi:10.1016/j.jpsychores.2004.09.006
15 Staples LG, Dear BF, Gandy M, et al. Psychometric properties and clinical utility of brief measures of depression, anxiety, and general distress: The PHQ-2, GAD-2, and K-6. *Gen Hosp Psychiatry* 2019;56:13-8. doi:10.1016/j.genhosppsych.2018.11.003
16 Hlynsson JI, Carlbring P. Diagnostic accuracy and clinical utility of the PHQ-2 and GAD-2: a comparison with long-format measures for depression and anxiety. *Front Psychol* 2024;15:1259997. doi:10.3389/fpsyg.2024.1259997
17 Devji T, Carrasco-Labra A, Qasim A, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ* 2020;369:m1714. doi:10.1136/bmj.m1714
18 National Collaborating Centre for Mental Health. The Improving Access to Psychological Therapies Manual. 2018. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/
19 Porter A, Franklin M, De Vocht F, et al. Estimating the effectiveness of an enhanced 'Improving Access to Psychological Therapies' (IAPT) service addressing the wider determinants of mental health: a real-world evaluation. *BMJ Open* 2024;14:e077220. doi:10.1136/bmjopen-2023-077220
20 Gyani A, Shafran R, Layard R, et al. Enhancing recovery rates: lessons from year one of IAPT. *Behav Res Ther* 2013;51:597-606. doi:10.1016/j.brat.2013.06.004
21 Ministère de la Santé et des Services sociaux. Document d'information à l'intention des établissements - Programme québécois pour les troubles mentaux : des autosoins à la psychothérapie (PQPTM) [Information document for establishments - Quebec program for mental disorders: from self-care to psychotherapy (PQPTM)]. Québec: Gouvernement du Québec. 2020. https://publications.msss.gouv.qc.ca/msss/document-002841/?&type=guide&date=DESC&sujet=pqptm&critere=sujet
22 Antony M. *The Ontario structured psychotherapy program: evidence-based, publicly funded, stepped-care treatment for depression and anxiety-related Disorders* . Richard Swinson Innovations in Anxiety Memorial Lecture. 2023
23 Greenspace Health. Reliable change index & caseness threshold. https://admin.help.greenspacehealth.com/article/647-reliable-change
24 Coley RY, Boggs JM, Beck A, et al. Defining Success in Measurement-Based Care for Depression: A Comparison of Common Metrics. *Psychiatr Serv* 2020;71:312-8. doi:10.1176/appi.ps.201900295
25 National Committee for Quality Assurance. HEDIS depression measures specified for electronic clinical data. https://www.ncqa.org/hedis/the-future-of-hedis/hedis-depression-measures-specified-for-electronic-clinical-data/
26 Alkan A, Carrier M, Henry RS, et al. Minimal Detectable Changes of the Health Assessment Questionnaire-Disability Index, Patient-Reported Outcomes Measurement Information System-29 Profile Version 2.0 Domains, and Patient Health Questionnaire-8 in People With Systemic Sclerosis: A Scleroderma Patient-Centered Intervention Network Cohort Cross-Sectional Study. *Arthritis Care Res* 2024;76:1549-57. doi:10.1002/acr.25397
27 Levis B, Benedetti A, Thombs BD, et al. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:l1476. doi:10.1136/bmj.l1476
28 Levis B, Sun Y, He C, et al. Accuracy of the PHQ-2 Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-analysis. *JAMA* 2020;323:2290-300. doi:10.1001/jama.2020.6504
29 Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol Med* 2020;50:1368-80. doi:10.1017/S0033291719001314
30 Negeri ZF, Levis B, Sun Y, et al. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. *BMJ* 2021;375:n2183. doi:10.1136/bmj.n2183
31 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi:10.1136/bmj.n71
32 González-Domínguez NP, Wu Y, Fan S, et al. Individual patient data meta-analysis estimates the minimal detectable change of the Geriatric Depression Scale-15. *J Clin Epidemiol* 2024;173:111443. doi:10.1016/j.jclinepi.2024.111443
33 Text Recycling Research Project. Best practices for researchers. https://textrecycling.org/resources/best-practices-for-researchers/
34 Thombs BD, Benedetti A, Kloda LA, et al. The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Syst Rev* 2014;3:124. doi:10.1186/2046-4053-3-124
35 Löwe B, Unützer J, Callahan CM, et al. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care* 2004;42:1194-201. doi:10.1097/00005650-200412000-00006
36 McGowan J, Sampson M, Salzwedel DM, et al. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol* 2016;75:40-6. doi:10.1016/j.jclinepi.2016.01.021
37 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606-13. doi:10.1046/j.1525-1497.2001.016009606.x
38 Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009
39 de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;4:54. doi:10.1186/1477-7525-4-54
40 Turner D, Schünemann HJ, Griffith LE, et al. The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol* 2010;63:28-36. doi:10.1016/j.jclinepi.2009.01.024

41 Toussaint A, Hüsing P, Gumz A, et al. Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7). *J Affect Disord* 2020;265:395-401. doi:10.1016/j.jad.2020.01.032
42 Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 1999;52:861-73. doi:10.1016/s0895-4356(99)00071-2
43 Blampied NM. Reliable change and the reliable change index: still useful after all these years? *Cogn Behav Therap* 2022;15:e50. doi:10.1017/S1754470X22000484
44 Ajele KW, Idemudia ES. Charting the course of depression care: a meta-analysis of reliability generalization of the patient health questionnaire (PHQ- 9) as the measure. *Discov Ment Health* 2025;5:50. doi:10.1007/s44192-025-00181-x
45 Trigg A, Ratitch B, Kruesmann F, et al. Interpretation of Change in Novel Digital Measures: A Statistical Review and Tutorial. *Digit Biomark* 2025;9:52-66. doi:10.1159/000543899
46 Jacobson NS, Follette WC, Revenstorf D. Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behav Ther* 1984;15:336-52. doi:10.1016/S0005-7894(84)80002-7
47 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;59:12-9. doi:10.1037//0022-006x.59.1.12
48 Fagerland MW, Lydersen S, Laake P. Recommended tests and confidence intervals for paired binomial proportions. *Stat Med* 2014;33:2850-75. doi:10.1002/sim.6148
49 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58. doi:10.1002/sim.1186
50 R: The R Project for Statistical Computing. https://www.r-project.org/
51 Posit. Posit. https://www.posit.co/
52 Bates D, Mächler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2014;67:1-48. doi:10.18637/jss.v067.i01
53 Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *BMJ Ment Health* 2019;22:153-60. doi:10.1136/ebmental-2019-300117
54 Wang Y, Devji T, Carrasco-Labra A, et al. A step-by-step approach for selecting an optimal minimal important difference. *BMJ* 2023;381:e073822. doi:10.1136/bmj-2022-073822
55 Kemmler G, Zabernigg A, Gattringer K, et al. A new approach to combining clinical relevance and statistical significance for evaluation of quality of life changes in the individual patient. *J Clin Epidemiol* 2010;63:171-9. doi:10.1016/j.jclinepi.2009.03.016
56 Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549. doi:10.1136/bmj.d549
57 Saunders R, Moinian D, Stott J, et al. Measurement invariance of the PHQ-9 and GAD-7 across males and females seeking treatment for common mental health disorders. *BMC Psychiatry* 2023;23:298. doi:10.1186/s12888-023-04804-x
58 Levis B, Benedetti A, Ioannidis JPA, et al. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. *J Clin Epidemiol* 2020;122:115-128.e1. doi:10.1016/j.jclinepi.2020.02.002
59 Hays RD, Peipert JD. Between-group minimally important change versus individual treatment responders. *Qual Life Res* 2021;30:2765-72. doi:10.1007/s11136-021-02897-z
60 Peipert JD, Cella D, Hays RD. Significant individual change should be used as a lower bound for anchor based estimates of meaningful change on patient-reported outcome scores. *Qual Life Res* 2024;33:3223-8. doi:10.1007/s11136-024-03788-9
61 Levis B, Negeri Z, Sun Y, et al. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. *BMJ* 2020;371:m4022. doi:10.1136/bmj.m4022
62 Wu Y, Levis B, Sun Y, et al. Accuracy of the Hospital Anxiety and Depression Scale Depression subscale (HADS-D) to screen for major depression: systematic review and individual participant data meta-analysis. *BMJ* 2021;373:n972. doi:10.1136/bmj.n972
63 Fife DA, Mendoza JL, Terry R. The assessment of reliability under range restriction: a comparison of α, ω, and test-retest reliability for dichotomous data. *Educ Psychol Meas* 2012;72:862-88. doi:10.1177/0013164411430225
64 Mattsson M, Sandqvist G, Hesselstrand R, et al. Validity and reliability of the Patient Health Questionnaire-8 in Swedish for individuals with systemic sclerosis. *Rheumatol Int* 2020;40:1675-87. doi:10.1007/s00296-020-04641-1
65 Yu X, Stewart SM, Wong PTK, et al. Screening for depression with the Patient Health Questionnaire-2 (PHQ-2) among the general population in Hong Kong. *J Affect Disord* 2011;134:444-7. doi:10.1016/j.jad.2011.05.007
66 Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ* 2011;2:53-5. doi:10.5116/ijme.4dfb.8dfd
67 White SJ, Phua QS, Lu L, et al. Heterogeneity in Systematic Reviews of Medical Imaging Diagnostic Test Accuracy Studies: A Systematic Review. *JAMA Netw Open* 2024;7:e240649. doi:10.1001/jamanetworkopen.2024.0649
68 Harel D, Wu Y, Levis B, et al. Comparison of Patient Health Questionnaire-9, Edinburgh Postnatal Depression Scale and Hospital Anxiety and Depression - Depression subscale scores by administration mode: An individual participant data differential item functioning meta-analysis. *J Affect Disord* 2024;361:674-83. doi:10.1016/j.jad.2024.06.033
69 Jabrayilov R, Emons WHM, Sijtsma K. Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Appl Psychol Meas* 2016;40:559-72. doi:10.1177/0146621616664046