Gough, A., Marshall, T., Ferris, E. and Stitch, A. (2023) *Within-individual variation of measured depression symptoms: A systematic review and meta-analysis.* Journal of Affective Disorders Reports, 14 . ISSN 2666-9153.

Journal of Affective Disorders Reports 14 (2023) 100675

Contents lists available at ScienceDirect

Journal of Affective Disorders Reports

journal homepage: www.sciencedirect.com/journal/journal-of-affective-disorders-reports

Research Paper

Within-individual variation of measured depression symptoms: A systematic review and meta-analysis

Alex Gough [*], Tom Marshall, Erica Ferris, Alice Sitch

Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

ARTICLE INFO

Keywords:
Depression
Within-individual variation
Patient health questionnaire-9
Patient health questionnaire-2
Beck's depression inventory
Hospital anxiety and depression scale

ABSTRACT

Background: Depression screening instruments are commonly used to assess the presence and severity of depression symptoms. However, there is little information on variation in depression screening scores over time within the same individual.

Methods: A systematic review and meta-analysis was performed of studies reporting the within-subject variability of the Beck's Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS), Patient Health Questionnaire-9 (PHQ-9) and Patient Health Questionnaire-2 (PHQ-2). Multiple databases were searched from inception to 14th July 2023. Title and abstract screening was performed in duplicate, full text screening and data extraction by one reviewer and verified by a second. Risk of bias was assessed with a modified COSMIN tool.

Results: Of 2798 titles and abstracts and 157 full text articles screened, 41 met the inclusion criteria. No studies were on patients with depression, most had only two measurements, less than two weeks apart. The pooled estimates of ICCs (Intraclass Correlations) and 95 % confidence intervals for BDI, HADS, PHQ-9 and PHQ-2 were 0.89 (0.84–0.93), 0.89 (0.85–0.93), 0.84 (0.81–0.88) and 0.75 (0.56–0.93), respectively.Studies in healthy subgroups showed lower ICC than those with physical illness.

Limitations: Assessment of variability is not the main aim of most papers assessing depression measurement instruments, so it is possible that some relevant papers have been missed.

Conclusions: Within-subject scores for BDI, HADS and PHQ-9, show generally excellent agreement over periods of up to two weeks. However published data on within-subject variability is lacking over longer time periods and for patients with depression.

Abbreviations

HADS	Hospital Anxiety and Depression Scale
BDI	Beck's Depression Inventory
PHQ-9	Patient Health Questionnaire-9
PHQ-2	Patient Health Questionnaire-2
CV	Coefficient of variation
CV_i	Coefficient of variation within an individual
ICC	Intraclass correlation
SD	Standard Deviation
NICE	National Institute for Health and Care Excellence
QOF	Quality Outcomes Framework

1. Introduction

Depression is a mental disorder characterised by loss of positive affect, low mood and behavioural, cognitive, emotional and physical symptoms (Anonymous 2019). Since there is no objective diagnostic test for depression the "gold standard" diagnosis is a structured clinical interview (Lakkis and Mahmassani, 2015; Anonymous 2013; Osório et al., 2019). Unfortunately, this takes around ninety minutes to complete and requires a large degree of clinician training (Brodey et al., 2018).

However, measurement-based care is becoming standard practice and involves ongoing collection of outcome measures including depression measures. Depression measurement instruments have been developed for use in non-psychiatric settings to screen for depression symptoms, to assess severity and for monitoring (e.g. response to treatment). These are less time-consuming and do not require clinician training.

Validated depression measurement instruments, in particular the PHQ-9, BDI and HADS were recommended by NICE. Their use was

* Corresponding author.
E-mail address: Agg932@student.bham.ac.uk (A. Gough).

https://doi.org/10.1016/j.jadr.2023.100675
Received 27 September 2023; Accepted 16 October 2023
Available online 17 October 2023

incentivised under the Quality Outcomes Framework (QOF) for both initial management and monitoring of depression in the UK between 2006 and 2013 (Anonymous 2006). Current NICE guidelines still recommend using a validated measure to inform and evaluate treatment, but this is no longer incentivised under QOF (Anonymous 2009).

Problems with diagnosis, assessment of severity and monitoring of depression symptoms arise due to factors such as the imperfect sensitivity and specificity of depression measurement instruments, practitioners' concerns about the reliability and practicality of depression measurement instruments, and problems with reliance on clinical judgement. Another problem is biological variation, the natural variation within a subject of the severity of the condition over time. This is due to ageing, season, diurnal variation and reproductive cycles, disease progression or a response to treatment (Matheson et al., 2015). However, most within-subject biological variation is chance variation (Fraser Callum, 2001). There is biological variation in mood and measurement variation in its assessment, which lead to within-subject variation of measured depression symptoms. Because mental health measurement instruments produce categorical scores, within-subject variation can impact the interpretation of measurement findings. This has the potential to lead to inaccurate conclusions about symptom progress and treatment outcomes (Sitch, 2019; Kendrick et al., 2009).

The depression measurement instruments most commonly used when screening for depression in a non-psychiatric setting (Lakkis and Mahmassani, 2015) are: Patient Health Questionnaire-9 (PHQ-9); Patient Health Questionnaire-2 (PHQ-2); Hospital Anxiety and Depression Scale (HADS); Beck Depression Inventory first and second editions (BDI and BDI-II). This review was therefore restricted to these measurement instruments, and other measurement instruments were not included in order that the scope of the review not be too broad. Details on these instruments such as structure, scoring method and sensitivity/specificity can be found in Appendix 1.

The problem of within-subject variation in measurement using these instruments has received little consideration. An unpublished scoping review identified only one paper on within-subject variation of measurement scales for depression and no systematic review.

The purpose of this review therefore is to review the literature on within-subject variation of commonly used depression measurement instruments and estimate the extent of within-subject variation in measured depression symptoms.

1.1. *Aim*

To describe and evaluate the current literature on the within-subject variability of measured depression symptom scores in healthy subjects and patients with physical and psychological disorders and to estimate within-subject variability of these measures.

2. Methods

PRISMA and PRISMA-S guidelines were followed in the preparation of this review.

Searches were devised to identify studies with multiple depression scores measured within the same subject. These included cohort studies, clinical trials or any studies in which a depression measurement was repeated more than once in the same individual. Search strategies were developed in collaboration with co-authors and colleagues.

Any study was included if it recorded primary research data on the variability of at least two scores from the same depression measurement instrument (PHQ-9, PHQ-2, BDI or HADS) within the same subject. Variability could be reported as coefficient of variation (CV_i), standard deviation (SD), variability independent of the mean (VIM), index of individuality (II), Reference Change Value (RCV), index of heterogeneity, validity coefficient (VC), ICC agreement, ICC consistency, Cronbach's alpha or Cohen's kappa. There was no restriction on time of publication, population, setting or sample size.

Studies were excluded if participants were not in a steady state (measurements were before and after an intervention or had an acute or rapidly changing illness) or data were secondary (systematic and narrative reviews).

Medline, Embase, APA Psychinfo, Cochrane Central, Epistemonikos and Open Grey were searched from inception to 9th October 2020 and an updated search was performed up to 14th July 2023 (full details in Appendix 2). Search terms were adapted for each database searched. Subject experts were contacted for suggestions for further papers. The references of included papers were checked by hand for further relevant papers. PROSPERO was checked for ongoing reviews, and the protocol was registered with PROSPERO (Ref CRD42020213398) . See Appendix 3 for sample search strategy used in searching Medline and Embase.

Titles and abstracts were screened independently by two reviewers in Abstrackr systematic review software (Byron et al., 2012), and those identified of interest underwent full text searches. Full texts were screened by AG and exclusions confirmed by EF. Differences were resolved by discussion. Foreign language papers were translated by Google translate software. Data was extracted from full texts that met the inclusion criteria.

Data were extracted into Excel from the full text of papers and from abstracts where full texts were not available. All data extraction was by a single reviewer (AG) but verified independently by a second reviewer (EF). Where the same study was reported in multiple papers, the full text paper was preferred over an abstract, English language preferred over non-English, and the earliest English version over the later if there was more than one. Table 1 lists the outcome and other main variables extracted. All eligible outcomes (measures of variation for PHQ-9, PHQ-2, BDI, HADS) were included.

If patient characteristic data was reported for a whole study population, but variability data was only reported for a sub-population and the details for the subpopulation were not reported, the patient characteristics data of the whole population patient was used as an approximation of the subgroup data.

Table 1
Variables extracted from papers.

Variable	Definition and rules
Author name and year of publication	Full author list extracted. Short title given as first author and year of publication
Depression measurement instrument	Depression measurement instrument reported in paper: PHQ-9, PHQ-2, BDI, HADS (or multiple)
Study design	Cohort, RCT etc
Number of subjects	Number of subjects used to calculate variability measure
Age	Average age of subjects used to calculate variability measure. If the subjects used to calculate variability measure is a subset of the study population and only the age is given for the whole population, the age of the whole population is used.
Sex	Percentage of subjects used to calculate variability measure who were male. If the subjects used to calculate variability measure is a subset of the study population and only the sex is given for the whole population, the sex of the whole population is used.
Ethnicity	Ethnicity recorded in study.
Setting	Primary care/community or secondary/tertiary/laboratory setting
Health status	Healthy/depressed/other mental health condition/physical health condition/mixture
Number of measurements	Number of repeated applications of depression measurement instrument to the same subject
Time interval between measurements	Length of time between measurements in days
CVI	Coefficient of variation of repeated measures within a subject
SD	Standard deviation of repeated measures within a subject
ICC	Intraclass correlation of repeated measures within a subject

A risk of bias tool adapted from the COSMIN risk of bias tool for test reliability (Mokkink et al., 2020) was used. (Appendix 4). Risk of bias was assessed using seven questions regarding patient stability, time between measurements, differences in measurement conditions, administering the measurement and assigning scores without knowledge of previous scores, any other design flaws, and whether the variability measure was adequately described. These were rated on a four-point scale from very good to inadequate. Risk of bias was scored by AG and the scores were reviewed by EF. Disagreements were resolved by discussion.

The primary outcome measure was variability measure of repeated measurements within the same subject. Studies were pooled to provide combined results where possible (i.e. the same depression measurement instrument and the same measure of variability).

Where multiple measures of variability were given in a single study, the primary population only was analysed for the main meta-analysis. The primary population was the full study population (as opposed to subgroups), and if the full study population was not given, the primary outcome was identified using the following hierarchy: 1. Healthy population; 2. Most stable population (i.e. subjectively judged to be in the most steady state such as disease course or treatment); 3. First outcome listed in the paper. If multiple depression measurement instruments were included, the data was extracted separately for each instrument.

A forest plot was generated in Stata using the metan command to graphically display the ICC results. See Appendix 5 for code. Stata SE 17 was used to perform statistical calculations. ICCs were transformed using Fisher's Z transformation (Field, 2005), : $Z = 0.5 * \ln((1 + \text{ICC}) / (1 - \text{ICC}))$. Standard error was calculated using the formula SE = sqrt (1/(N - 3) Z scores were then back-transformed. The metan command was used to generate a forest plot. 95 % confidence intervals of the ICC were calculated. Other variability measures were described without performing metanalysis as they were too few to pool.

Subgroup analysis was performed by health status of the study population, setting (primary care or community versus secondary or tertiary), length of time between measurements and number of subjects (>100). A sensitivity analysis was performed excluding all studies with a risk of bias of 3 or more (not including questions on blinding which were all scored high). Subgroup analyses were only performed on the PHQ-9 depression measurement instrument since it had the largest number of included studies.

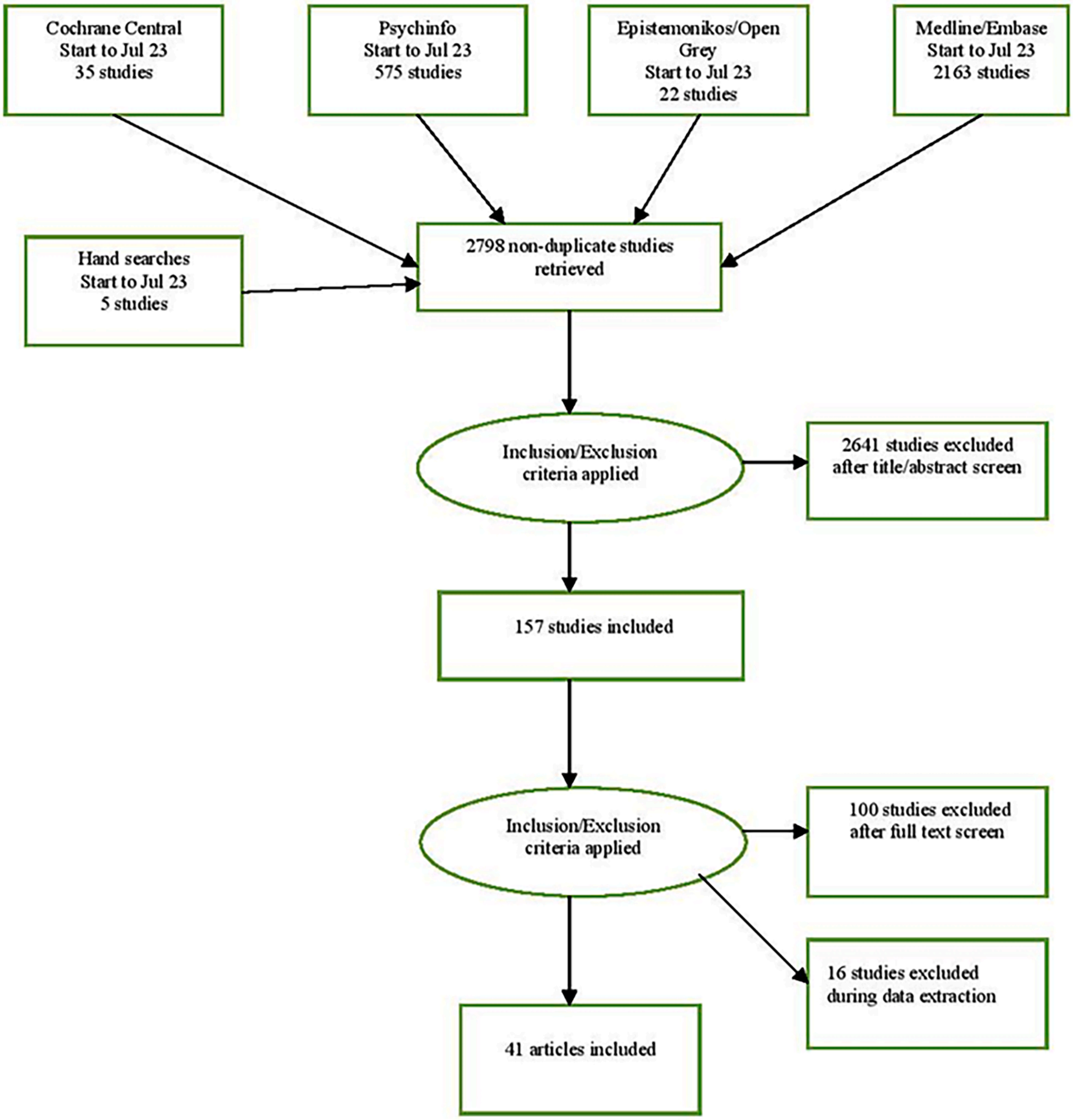
Cochrane Central
Start to Jul 23
35 studies
Psychinfo
Start to Jul 23
575 studies
Epistemonikos/Open Grey
Start to Jul 23
22 studies
Medline/Embase
Start to Jul 23
2163 studies
Hand searches
Start to Jul 23
5 studies
2798 non-duplicate studies retrieved
Inclusion/Exclusion criteria applied
2641 studies excluded after title/abstract screen
157 studies included
Inclusion/Exclusion criteria applied
100 studies excluded after full text screen
16 studies excluded during data extraction
41 articles included

Fig. 1. PRISMA flow diagram describing selection of studies.

Where studies reported demographic data for the whole population but not for the subgroup for which variability data was reported, the demographic data for the whole study population was used and was assumed to be close to the variability subgroup.

2.1. Deviations from protocol

A bespoke risk of bias tool was developed and included in the original protocol recorded in Prospero, but it was ultimately decided preferable to use a validated risk of bias tool, and the COSMIN tool was selected.

3. Results

2798 non-duplicate citations were retrieved after database searches of Embase, Medline, Psychinfo, Cochrane, Epistemonokis and hand searching of reference lists of included studies. 2641 articles were excluded after title and abstract screening leaving 157 articles included see Fig. 1. After full text screening, 100 articles were excluded, and a further 16 were excluded during data extraction, leaving 41 included articles for analysis. Most full text exclusions were due to lack of reporting of a recognised measure of variability.

Forty-one studies met the inclusion criteria, including between 15 and 458 subjects. Study populations were diverse in terms of age, gender and health status. Only two studies reported more than two repeat measures of the tests within the same subjects. The majority had time intervals between tests of one to two weeks. Ten papers included healthy populations or subpopulations. No studies looked at patients with depression as the main population or subpopulation.

See Appendix 6 for details.

Risk of bias due to missing results was deemed to be low, since in most cases the measure of variability was not the primary outcome of the individual studies, and therefore systematically withholding publication based on measurement variability is unlikely. Some papers with repeated measures do not report variability data, but this is likely to be because it is not of relevance to the study's primary outcome, and therefore is unlikely to introduce a systematic bias.

No study scored more than doubtful for question B4 *knowledge of previous scores when administering the test* and B5, *when assigning the scores* (questions B4 and B5). Appendix 6 presents the maximum risk of bias score for each study with questions B4 and B5 excluded, since every study would be rated at a high risk of bias with these scores included. eight studies scored the lowest risk of bias, very good. 15 were ranked adequate and 18 were ranked doubtful. No studies were ranked inadequate, since these should have been excluded at the screening stage, for example due to an unstable population or lack of calculation of a recognised measure of variability. Appendix 7 lists the full risk of bias scoring for each study.

3.1. Results of individual studies

The following summary statistics refer to the primary population as defined in Synthesis Methods above, but include the papers by Karmisholt and Andersen (2019) and Shelton et al. (2015) which were excluded from the pooled results. All 41 papers reported number of subjects (range 15 to 458), age (range 15.9 to 71.4), number of measurements (range 2 to 13); 38 reported gender (range 0 % to 100 % male); and 39 reported time between measurements (range 1 to 84 days). 2 out of 41 papers were available only as abstract, the rest had full texts available. 15 of the 41 papers explicitly recorded ethnicity. 40 studies reported an ICC, with results ranging from 0.232 to 0.980. One paper (Karmisholt and Andersen, 2019) reported a CVI of 0.416 in patients with subclinical hypothyroidism. Table 2 shows the number of specific depression measurement instruments which were studied (53 depression measurement instruments in 41 papers).. For the five papers where an ICC was reported and the time interval between measurements was 28 days or more, the ICC of the primary measurement ranged from 0.66 to 0.98. The number of measurements ranged from 2 to 13 but only three papers had more than two measurements.

Table 2
Depression measurement instruments analysed for variability in the included studies (total 53 groups or subgroups in 41 papers).

Depression measurement instrument	Number of studies
BDI-I	2
BDI-II	5
HADS	9
HADS - depression subscale only	5
PHQ-2	7
PHQ-9	25

All studies were cohort design except for two which used the patients from one arm of a randomised controlled trial (Weobong et al., 2009 and Löwe, 2004). 11/41 studies were conducted in primary care (26.8 %), 28/41 (68.3.1 %) studies were conducted in hospitals, laboratories and other secondary or tertiary care settings and in 2/41 studies the setting was unclear (4.9 %). In 30/40 (75 %) studies that reported ICCs, the method of calculating the ICCs was not reported.

Table 3 shows the health status of the patients included in the 41 studies. Note that some of the papers included more than one health condition, so the denominator for number of health conditions in the included studies is 44. None of the studies identified expressly concerned individuals with depression.

Two papers were excluded from the pooled estimates, Karmisholt and Andersen (2019) because it reported a CV not an ICC and Shelton et al. (2015) because it reported a negative confidence interval which is mathematically impossible for an ICC.

The tau-squared result for the meta-analysis of the ICCs ranged from to 0.006, suggesting a low between study heterogeneity. There were few outliers, and the 95 % confidence intervals of the weighted average results for the ICC are small. Appendix 8 shows the tau-squared results and predictive intervals for subgroup analyses.

On analysis grouped by depression measurement instrument, pooled estimates (95 % CI) of ICC were similar for BDI, HADS, PHQ-9: 0.89 (0.84–0.93), 0.89 (0.85–0.93), 0.84 (0.81–0.88) respectively. For PHQ-2, ICC was 0.75 (0.56–0.93), see Fig. 2.

There was little difference between the pooled estimates for primary care/community versus secondary care or other care with ICCs and 95 % CIs of 0.86 (0.80–0.92) and 0.83 (0.79–0.87) respectively (Fig. 3).

The pooled estimate of ICC in the healthy subgroup was lower, 0.79 (95 % CI 0.73–0.85) than in the physical illness subgroup, ICC 0.88 (95 % CI 0.84–0.91) although the confidence intervals overlapped (Fig. 4). When studies of 100 or less subjects were excluded (Fig. 5), the pooled ICC estimate for PHQ-9 was 0.83 (95 % CI 0.80–0.86). Studies in which measurements were taken more than 7 days apart (Fig. 6) had a pooled estimate for PHQ-9 of 0.83 (95 %CI 0.80–0.86). When studies with a risk of bias of 3 or more were excluded (Fig. 7), the pooled estimate for ICC was 0.82 (95 %CI 0.78–0.86).When these results are compared to the full dataset where the pooled ICC estimate for PHQ-9 was 0.84 (0.81–0.88) it can be seen that there is little effect of sample size, time between measurements and risk of bias on the robustness of the result.

Table 3
Health status of patients included in studies.

Health status	Number of results (/44)	%
Healthy	12	27.3
Mental health condition	1	2.3
Physical health condition	28	63.6
Mix of mental and physical conditions	2	4.5

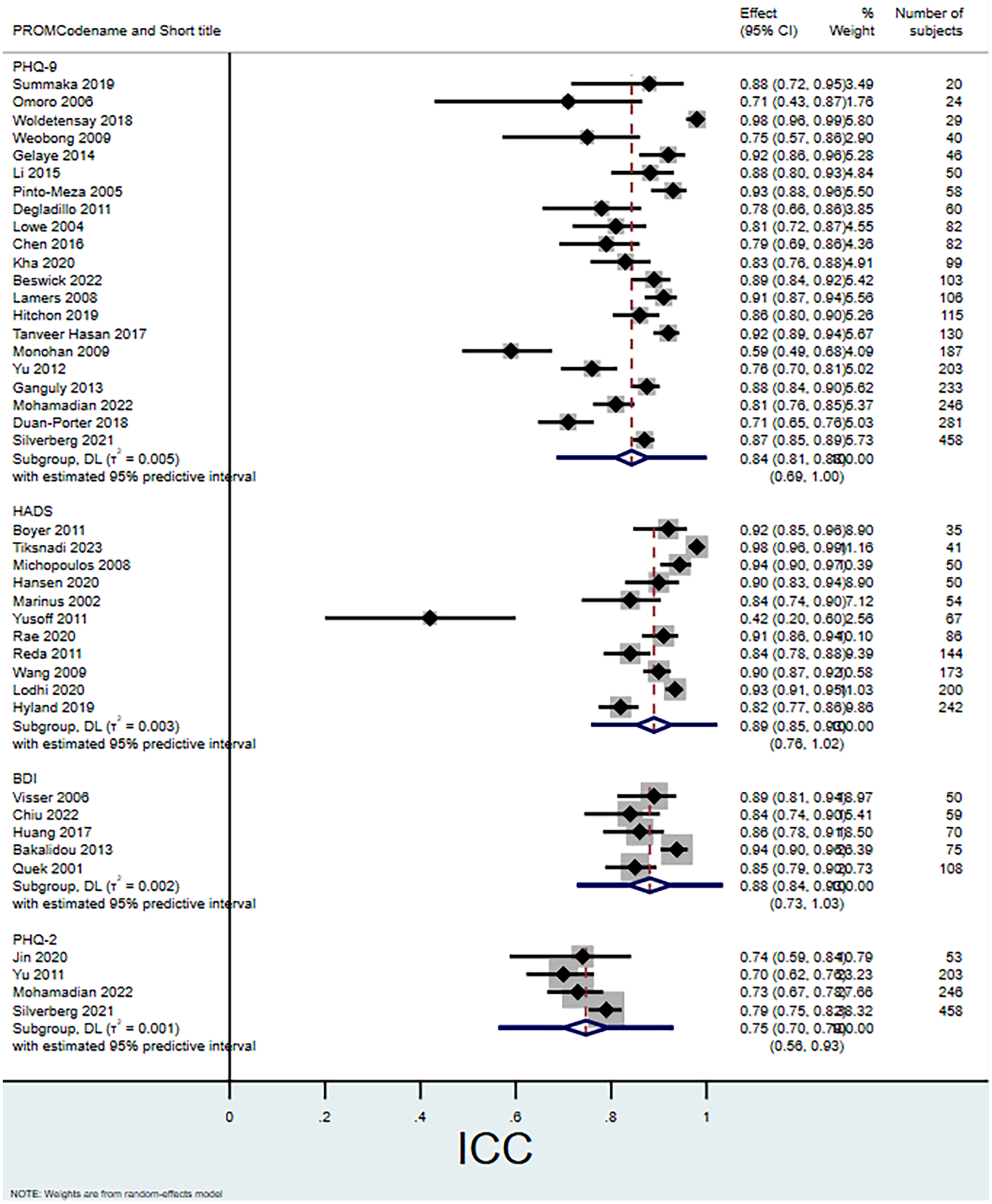

Fig. 2. Forest plot of ICCs with 95 % CI, grouped by depression measurement instrument.

4. Discussion

This review describes the current literature on variability of scores derived from depression measurement instruments within an individual. The intraclass correlation (ICC) has been used in all but one of the included papers here to measure test-test reliability or variability. The pooled estimate of ICCs ranged from 0.84 to 0.89 for three of the depression measurement instruments (PHQ-9, BDI and HADS), suggesting generally these depression measurement instruments have excellent reliability, and the short-term variability is low. The healthy subgroup had higher variability than those with a physical illness, and the primary or community setting subgroup had lower variability than subgroups in other settings, though it is not clear why this is the case. Most papers used ICCs as the measure of reliability/variability, but one paper (Karmisholt and Andersen, 2019) used coefficient of variation (CV), and reported a markedly high variability (CV=0.416) which contrasts with the high reliability/repeatability shown by the other variability estimates.

The main aims of the included studies were to validate the selected tests. They were therefore not designed to measure variability. Sample sizes were small by general medical study standards, with the largest study population numbering only 458. In only three studies identified were more than two repeated measures performed. Two of these studies showed higher variability than most of the studies with only two measurements. Duan-Porter et al. (2018) had 12 measurements one month apart, which resulted in an ICC for PHQ-9 of 0.71 which is lower than the pooled average. Karmisholt and Andersen (2019) had 13 measurements of HADS one month apart, which resulted in a high estimate of

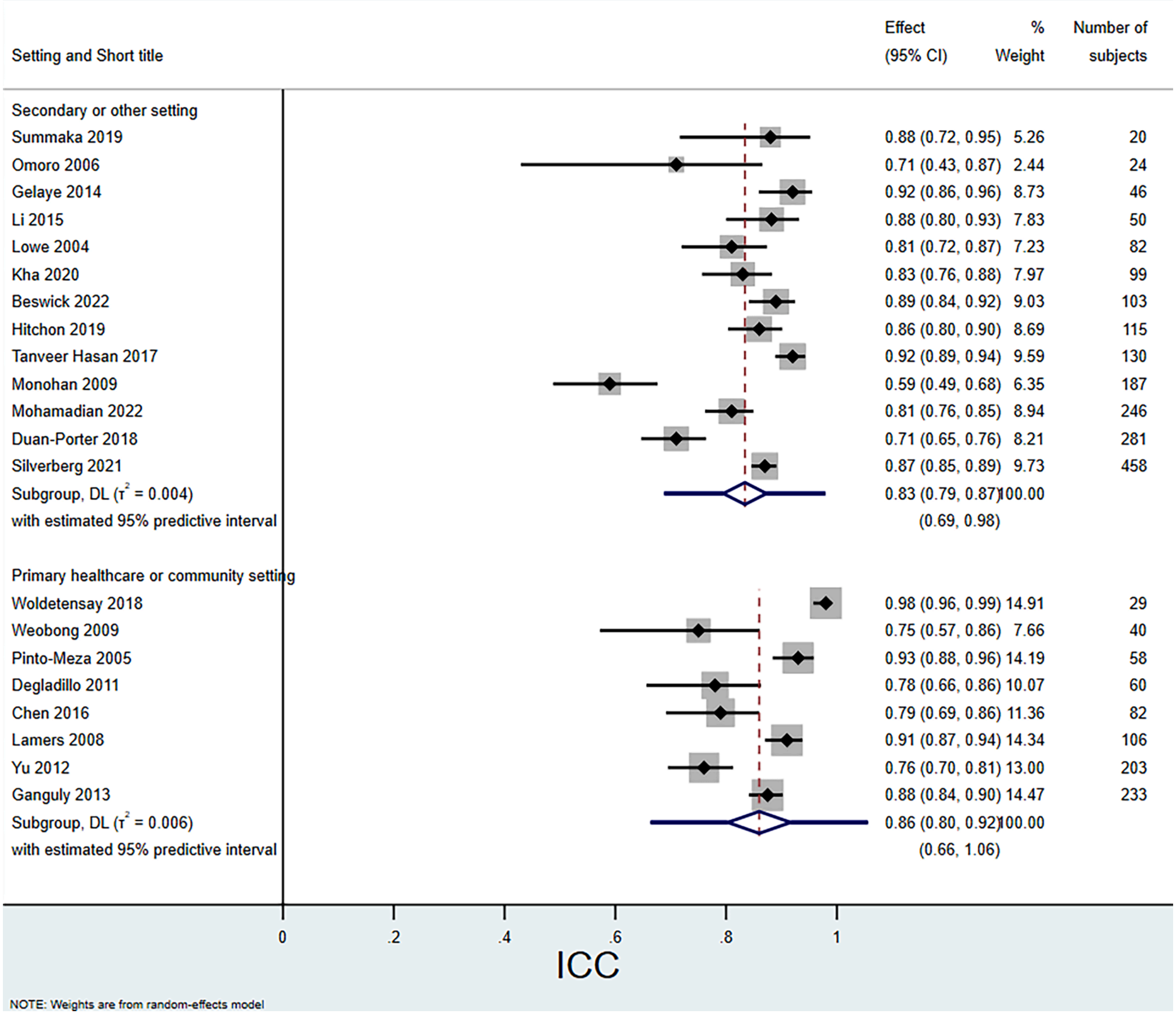

Fig. 3. Forest plot of ICCs with 95 % CI grouped by setting, PHQ-9 only.

variability. By contrast, Beswick et al. (2022) had a lower variability, but only had three measurements performed over four weeks. The studies with more measurements over a longer time period had higher variability than most of the studies with only two measurements.

In only seven studies identified were ICCs calculated for measures repeated 28 days or more apart (Delgadillo et al., 2011; Hyland et al., 2019; Quek et al., 2001; Shelton et al., 2015; Yusoff et al., 2011; Silverberg et al., 2020 and Tiksnadi et al., 2023). Six of these studies had lower ICCs than the overall pooled estimates, suggesting that variability of depression measurement instruments is higher than studies looking only at short term test-retest reliability might suggest. The majority had time intervals between tests of one to two weeks, making them poor measures of long-term variability. For more accurate estimates of how the results of these tests vary over time, more repeat measures, over longer time scales, in more individuals is needed. It is also of note that the variability of the measurement instruments appears to be greater in patients with mental health diagnoses than with physical health diagnoses, although the number of included studies with mental health diagnoses is too small to draw firm conclusions.

Even for the main aims of these individual studies, there are some weaknesses. There are at least ten ways of calculating an ICC, all of which can give different results (Koo and Li, 2016). Only ten of the studies included in this review provide the method by which they calculate ICC, which limits the ability to interpret their results. None of the studies were blinded, (it was not ensured that the professionals administered the tests and assigned the scores without knowledge of previous patient scores). This may not have had a marked effect on the test results since the questionnaires are standardised and there is no possibility of interpretation by the investigator regarding the numerical scores. However, it is possible that the subjects remembered their answers from their prior tests which could influence their scores (Terwee et al., 2007), especially where the interval between tests was short. Information on ethnicity was often not recorded.

The patient populations of the included studies were diverse in terms of age, ethnicity, sex and health status. Many of the studies provided analysis of the variability of the outcome measures in subgroups. To choose which group to include in the meta-analysis was problematic, since it had to involve arbitrary choices. However, the rules to decide which group to include were applied systematically, using the following hierarchy: 1. Healthiest population; 2. Most stable population; 3. First listed population. There were missing values in a number of studies for age, sex and ethnicity, with ethnicity being particularly poorly recorded.

The limited number of measures and the short time periods between measurements mean that the current literature provides very limited information on how measures of depression symptoms vary over time within a subject. This lack of knowledge of how depression measures vary could make clinical decision making difficult. If a test of depression symptoms with high within-subject variability is performed once by a clinician, it will be unclear whether the result obtained will represent a good estimate of the true mean state of that subject, or whether due to

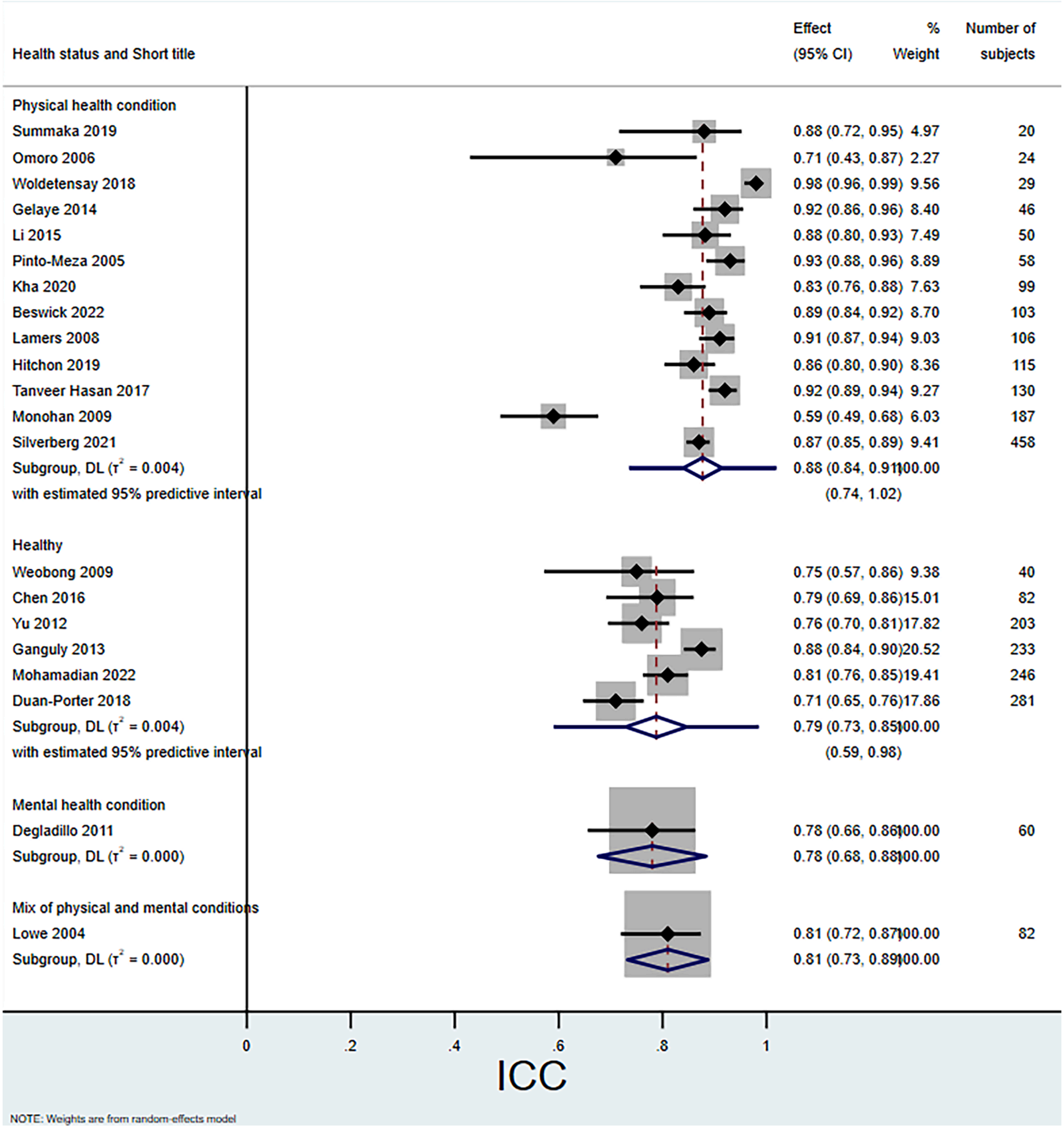

Fig. 4. Forest plot of ICCs with 95 % CI by health status, PHQ-9 only.

random changes in situational stressors or other temporary factors the measurement obtained is in reality markedly lower or higher than the true mean. This could lead to inappropriate influence on clinical management decisions, even when other clinical factors are taken into account.

It is surprising that none of the studies identified specifically related to patients with depression.

Studies are therefore needed that more accurately demonstrate the within-subject variation of tests of depression over time. Studies on depression measurement instruments such as those in this review with multiple measures over longer periods of time would demonstrate the total within-subject variation of these tests, encompassing the measurement error and the biological variation. This total variation is the most useful measure for clinicians since it provides guidance on the “real world” variability of the selected tests, and hence the best idea of whether a single measurement can be relied upon, and if not, now many measurements are required to make meaningful clinical decisions. Studies assessing the variability of measurements of depression in patients with depression are currently lacking and are needed.

4.1. Limitations

There are a few limitations with this review. One is the possibility of relevant studies being missed by the search strategy. Since few of the papers were interested in measuring variability as a primary objective, but were focussed mainly on reliability of the tests, there may have been papers that report variability as a secondary outcome that have not been included. However, the search strategy was as broad as possible, and a large number of abstracts and papers were screened. On the other hand, this broad search strategy could have led to the inclusion of poor-quality studies. There is no risk of bias tool currently in existence that is ideal for this review, but it was decided to use a validated risk of bias tool instead

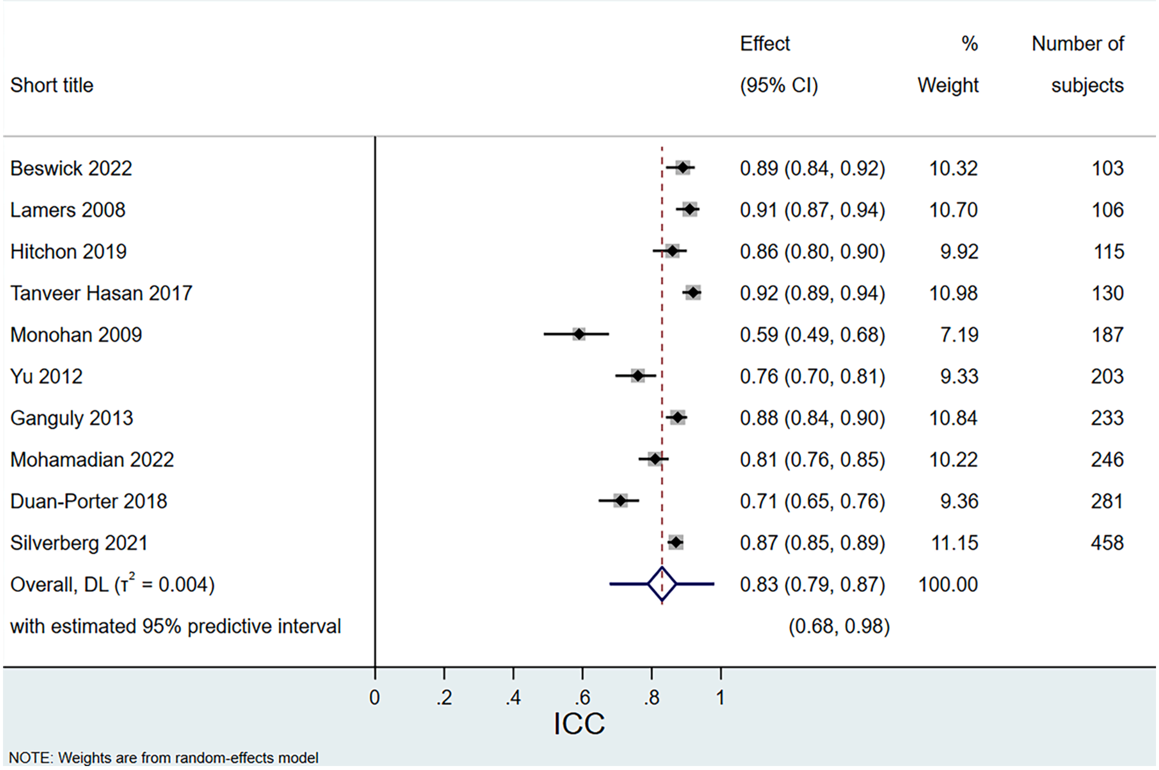

Fig. 5. Forest plot of ICCs with 95 % CI for those studies with more than 100 subjects, PHQ-9 only.

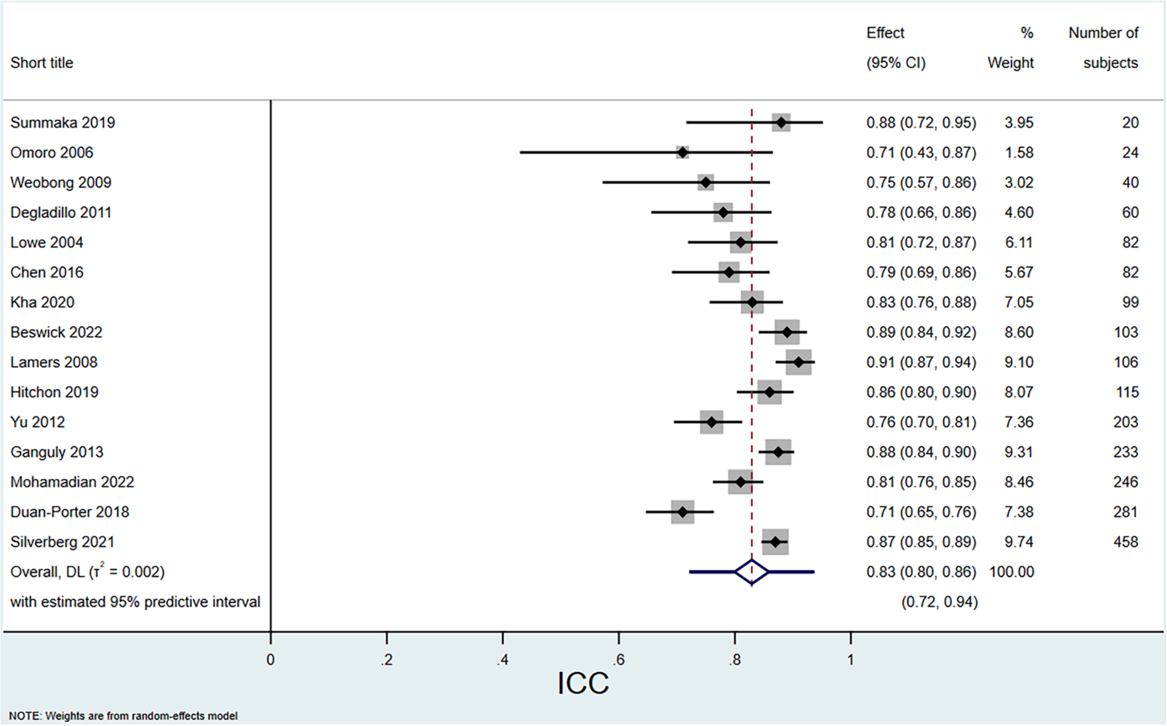

Fig. 6. Forest plot of ICCs with 95 % CI for those studies where measurements were performed more than 7 days apart, PHQ-9 only.

of a bespoke tool, and the COSMIN tool was selected. This enabled some estimate of risk of bias of each paper, although the questions on blinding were problematic, given that none of the papers included stated that the investigators performed the measurements or assigned scores without knowledge of previous scores by the individual. Nevertheless, since the questionnaires were largely self-administered by the subjects, and provided numerical scores, this was unlikely to bias the findings significantly. The risk of bias scores are therefore presented in this study with the two questions on blinding excluded.

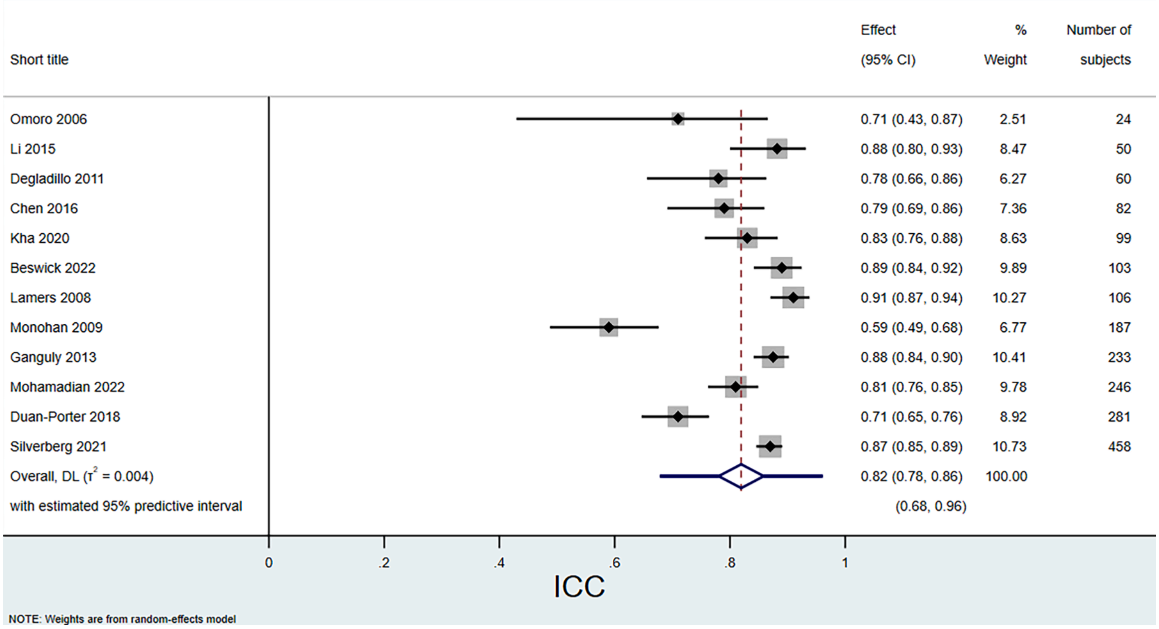

Fig. 7. Forest plot of ICCs with 95 % CI for those studies with a risk of bias of 2 or less, PHQ-9 only.

Other information

Registration

Prospero CRD42020213398 22/10/2020

The review protocol is available at https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020213398

The protocol describes a risk of bias tool which was adapted from the QUADAS tool. The study used the COSMIN tool for risk of bias assessment as being more appropriate and validated.

Funding

No external funding was provided for this review. It was performed as part of a PhD thesis at the University of Birmingham which is self-funded by AG.

Author statement

All authors read and approved the manuscript.

Availability of code, data and other materials

Risk of bias data and data extraction tables are available from the author on reasonable request.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgements

Thanks to Konstantinos Tryposkiadis for assistance with the statistical methods and Sophie Beese for assistance with developing the search strategies.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jadr.2023.100675.

References

American Psychiatric Association aQAPA. Diagnostic and Statistical Manual of Mental Disorders : DSM-5/American Psychiatric Association, 5th ed., 2013. American Psychiatric Publishing, Washington, District of Columbia London, England.

Anonymous. NICE clinical guideline CG90: depression in adults: recognition and management 2009 [Available from: https://www.nice.org.uk/guidance/cg90].

Anonymous, 2019. Depression. the Treatment and Management of Depression in Adults (Updated edition). The British Psychological Society and The Royal College of Psychiatrists, Leicester.

Beswick, E., Quigley, S., Macdonald, P., Patrick, S., Colville, S., Chandran, S., et al., 2022. The Patient Health Questionnaire (PHQ-9) as a tool to screen for depression in people with multiple sclerosis: a cross-sectional validation study. BMC Psychol. 10 (1), 281.

Brodey, B., Purcell, S.E., Rhea, K., Maier, P., First, M., Zweede, L., et al., 2018. Rapid and accurate behavioral health diagnostic screening: initial validation study of a web-based, self-report tool (the SAGE-SR). J. Med. Internet Res. 20 (3), e108.

British Medical Association Ne. Revisions to the GMS Contract, 2006/7: Delivering Investment in General Practice, 2006. BMA, London.

Byron C. Wallace K.S., C.E. Brodley, J. Lau and T.A. Trikalinos, editor. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. ACM International Health Informatics Symposium (IHI); 2012.

Delgadillo, J., Payne, S., Gilbody, S., Godfrey, C., Gore, S., Jessop, D., et al., 2011. How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. J. Affect. Disord. 134 (1–3), 266–271.

Duan-Porter, W., Hatch, D., Pendergast, J.F., Freude, G., Rose, U., Burr, H., et al., 2018. 12-month trajectories of depressive symptoms among nurses-contribution of personality, job characteristics, coping, and burnout. J. Affect. Disord. 234, 67–73.

Field, A.P., 2005. Is the meta-analysis of correlation coefficients accurate when population correlations vary? Psychol. Methods 10 (4), 444–467.

Fraser Callum, G, 2001. Biological Variation: From Principles to Practice. AACC Press, USA.

Hyland, K.A., Hoogland, A.I., Gonzalez, B.D., Nelson, A.M., Lechner, S., Tyson, D.M., et al., 2019. Evaluation of the psychometric and structural properties of the Spanish version of the hospital anxiety and depression scale in Latina cancer patients. J. Pain Symptom Manage. 58 (2), 289–296.e2.

Karmisholt, J., Andersen, S., 2019. Detecting true change in the hospital anxiety and depression scale, SF-36, and hypothyroid score when monitoring patients with subclinical hypothyroidism. Eur. Thyroid J. 8 (3), 144–151.

Kendrick, T., Dowrick, C., McBride, A., Howe, A., Clarke, P., Maisey, S., et al., 2009. Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. BMJ 338, b750.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163.

Lakkis, N.A., Mahmassani, D.M., 2015. Screening instruments for depression in primary care: a concise review for clinicians. Postgrad. Med. 127 (1), 99–106.

Löwe, B., 2004. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med. Care 42 (12), 1194–1202.

Matheson, G.J., Schain, M., Almeida, R., Lundberg, J., Cselényi, Z., Borg, J., et al., 2015. Diurnal and seasonal variation of the brain serotonin system in healthy male subjects. Neuroimage 112, 225–231.

Mokkink, L.B., Boers, M., van der Vleuten, C.P.M., Bouter, L.M., Alonso, J., Patrick, D.L., et al., 2020. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med. Res. Methodol. 20 (1), 293.

Osório, F.L., Loureiro, S.R., Hallak, J.E.C., Machado-de-Sousa, J.P., Ushirohira, J.M., Baes, C.V.W., et al., 2019. Clinical validity and intrarater and test–retest reliability of the Structured Clinical Interview for DSM-5 – Clinician Version (SCID-5-CV). Psychiatry Clin. Neurosci. 73 (12), 754–760.

Quek, K.F., Low, W.Y., Razack, A.H., Loh, C.S., 2001. Beck Depression Inventory (BDI): a reliability and validity test in the Malaysian urological population. Med. J. Malaysia 56 (3), 285–292.

Shelton, A.T., Houghton, N.Y., Morris, D.O., Latchford, G.L., Bekker, H.L., Munyombwe, T., 2015. The development and validation of a psychological questionnaire for patients undergoing orthognathic treatment. Orthod. Craniofac. Res. 18 (1), 51–64.

Silverberg, J., Lee, B., Lei, D., Yousaf, M., Janmohamed, S., Mann, C., et al., 2020. PMH39 Measurement properties of Patient Health Questionnaire (PHQ)-9 and PHQ-2 in adult patients with atopic dermatitis. Value Health 23 (Supplement 2), S590–S591.

Sitch, A.J., 2019. Modelling Optimal Use of Tests For Monitoring [Disease Progression And Recurrence]. University of Birmingham.

Terwee, C.B., Bot, S.D.M., de Boer, M.R., van der Windt, D.A.W.M., Knol, D.L., Dekker, J., et al., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J. Clin. Epidemiol. 60 (1), 34–42.

Tiksnadi, B.B., Triani, N., Fihaya, F.Y., Turu' Allo, I.J., Iskandar, S., Putri, D.A.E., 2023. Validation of hospital anxiety and depression scale in an Indonesian population: a scale adaptation study. Fam. Med. Commun. Health 11 (2).

Weobong, B., Akpalu, B., Doku, V., Owusu-Agyei, S., Hurt, L., Kirkwood, B., et al., 2009. The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana. J. Affect. Disord. 113 (1-2), 109–117.

Yusoff, N., Low, W.Y., Yip, C.H., 2011. Psychometric properties of the Malay Version of the hospital anxiety and depression scale: a study of husbands of breast cancer patients in Kuala Lumpur, Malaysia. Asian Pac. J. Cancer Prev. 12 (4), 915–917.

Further reading

Bakalidou, D., Skordilis, E.K., Giannopoulos, S., Stamboulis, E., Voumvourakis, K., 2013. Validity and reliability of the FSS in Greek MS patients. Springerplus 2 (1), 304.

Bocéréan, C., Dupret, E., 2014. A validation study of the Hospital Anxiety and Depression Scale (HADS) in a large sample of French employees. BMC Psychiatry 14 (1), 354.

Boyer, F.C., Rapin, A., Calmus, A., Percebois-Macadre, L., Tambosco, L., Bertaud, S., et al., 2011. HADS scale in adults suffering from Steinert myotonia: reproducibility and internal consistency. Ann. Phys. Rehabil. Med. 54, e238.

Button, K.S., Kounali, D., Thomas, L., Wiles, N.J., Peters, T.J., Welton, N.J., 2015. Minimal clinically important difference on the Beck Depression Inventory - II according to the patient's perspective. Psychol. Med. 3269–3279.

Chen, I.P., Liu, S.I., Huang, H.C., Sun, F.J., Huang, C.R., Sung, M.R., 2016. Validation of the Patient Health Questionnaire for depression screening among the elderly patients in Taiwan. Int. J. Gerontol. 193–197.

Chiu, E.C., Chen, Y.J., Wu, W.C., Chou, C.X., Yu, M.Y., 2022. Psychometric comparisons of three depression measures for patients with stroke. Am. J. Occup. Ther. 76 (4).

Dozois, D.J.A., Dobson, K.S., Ahnberg, J.L., 1998. A psychometric evaluation of the beck depression inventory-II. Psychol. Assess. 10 (2), 83–89.

Ganguly, S., Samanta, M., Roy, P., Chatterjee, S., Kaplan, D.W., Basu, B., 2013. Patient health questionnaire-9 as an effective tool for screening of depression among Indian adolescents. J. Adolesc. Health 52 (5), 546–551.

Gelaye, B., 2014. Diagnostic validity of patient health questionnaire-9 and composite international diagnostic interview assessment scales for depression in East Africa. Dissertation Abstr. Int. Sect. B Sci. Eng. 74 (9–B(E)) No Pagination Specified.

Hitchon, C.A., Zhang, L., Peschken, C.A., Lix, L.M., Graff, L.A., Fisk, J.D., et al., 2019. The validity and reliability of screening measures for depression and anxiety disorders in rheumatoid arthritis. Arthritis Care Res 14. Hoboken.

Huang, S.L., Hsieh, C.L., Wu, R.M., Lu, W.S., 2017. Test-retest reliability and minimal detectable change of the Beck Depression Inventory and the Taiwan Geriatric Depression Scale in patients with Parkinson's disease. PLoS ONE 12 (9).

Jin, H., Wu, S., 2020. Text messaging as a screening tool for depression and related conditions in underserved, predominantly minority safety net primary care patients: validity study. J. Med. Internet Res. 22 (3), e17282.

Kha, T.V., Stenager, E., Hoang, H., Bruun-Plesner, K., Fuglsang, K.S., la Cour, B.S., et al., 2020. Preliminary validity and test-retest reliability of two depression questionnaires compared with a diagnostic interview in 99 patients with chronic pain seeking specialist pain treatment. Scand. J. Pain 20 (4), 717–726.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9), 606–613.

Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2003. The patient health questionnaire-2: validity of a two-item depression screener. Med. Care 41 (11), 1284–1292.

Lamers, F., Jonkers, C.C.M., Bosma, H., Penninx, B.W.J.H., Knottnerus, J.A., van Eijk, J. T.M., 2008. Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. J. Clin. Epidemiol. 61 (7), 679–687.

Li, W., Kai, L., Jianchao, L., Li, S., Rongjing, D., Dayi, H., 2015. Value of patient health questionnaires (PHQ)-9 and PHQ-2 for screening depression disorders in cardiovascular outpatients. [Chinese]. Zhonghua Xin Xue Guan Bing Za Zhi 43 (5), 428–431.

Lodhi, F.S., Elsous, A.M., Irum, S., Khan, A.A., Rabbani, U., 2020. Psychometric properties of the Urdu version of the Hospital Anxiety and Depression Scale (HADS) among pregnant women in Abbottabad. Pakistan. Gen. Psychiatr. 33 (5), e100276.

Löwe, B., Unützer, J., Callahan, C.M., Perkins, A.J., Kroenke, K., 2004. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med. Care 42 (12), 1194–1201.

Manea, L., Gilbody, S., McMillan, D., 2011. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. CMAJ 184 (3), E191–E196.

Marinus, J., Leentjens, A.F.G., Visser, M., Stiggelbout, A.M., Van Hilten, J.J., 2002. Evaluation of the hospital anxiety and depression scale in patients with Parkinson's disease. Clin. Neuropharmacol. 25 (6), 318–324.

Michopoulos, I., Douzenis, A., Kalkavoura, C., Christodoulou, C., Michalopoulou, P., Kalemi, G., et al., 2008. Hospital anxiety and depression scale (HADS): validation in a Greek general hospital sample. Ann. Gen. Psychiatry 7 no pagination.

Mohamadian, R., Khazaie, H., Ahmadi, S.M., Fatmizade, M., Ghahremani, S., Sadeghi, H., et al., 2022. The psychometric properties of the persian versions of the patient health questionnaires 9 and 2 as screening tools for detecting depression among university students. Int. J. Prev. Med. 13 (1), 116.

Monahan, P.O., Shacham, E., Reece, M., Kroenke, K., Ong'Or, W.O., Omollo, O., et al., 2009. Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in Western Kenya. J. Gen. Intern. Med. 24 (2), 189–197.

Moore, M., Ali, S., Stuart, B., Leydon, G.M., Ovens, J., Goodall, C., et al., 2012. Depression management in primary care: an observational study of management changes related to PHQ-9 score for depression monitoring. Br. J. Gen. Pract. 62 (599), e451–e457.

Omoro, S., Fann, J., Weymuller, E., Macharia, I., Yueh, B., 2006. Swahili translation and validation of the Patient Health Questionnaire-9 depression scale in the Kenyan head and neck cancer patient population. Int. J. Psychiatry Med. 36 (3), 367–381.

Pinto-Meza, A., Serrano-Blanco, A., Penarrubia, M.T., Blanco, E., Haro, J.M., 2005. Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone? J. Gen. Intern. Med. 20 (8), 738–742.

Rae, C.S., Tsangaris, E., Klassen, A.F., Breakey, V., D'Agostino, N., 2020. Comparison of patient-reported outcome measures for use as performance metrics in adolescent and young adult psychosocial cancer care. J. Adolesc. Young Adult Oncol. 9 (2), 262–270.

Reda, A.A., 2011. Reliability and validity of the Ethiopian version of the hospital anxiety and depression scale (HADS) in HIV infected patients. PLoS ONE 6 (1).

Silverberg, J.I., Lee, B., Lei, D., Yousaf, M., Janmohamed, S.R., Vakharia, P.P., et al., 2021. Measurement properties of patient health questionnaire 9 and patient health questionnaire 2 in adult patients with atopic dermatitis. Dermatitis 32 (4), 225–231.

Smarr, K.L., Keefer, A.L., 2011. Measures of depression and depressive symptoms: beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). Arthritis Care Res. (Hoboken) 63 (S11), S454–S466.

Summaka, M., Zein, H., Abbas, L.A., Elias, C., Elias, E., Fares, Y., et al., 2019. Validity and reliability of the arabic patient health questionnaire-9 in patients with spinal cord injury in Lebanon. World Neurosurg 125, e1016–e1e22.

Tanveer Hasan, A.T.M., Haq, S.A., Ahmed, S., Abdal, S.J., Majumder, M.S.M., Nahiduzzamane Shazzad, M., et al., 2017. Translation & cross-cultural adaptation of the English patient health questionnaire-9 (PHQ-9) to Bangla and its validation in adult SLE patients. Int. J. Rheum. Dis. 20 (Supplement 1), 57.

Visser, M., Leentjens, A.F.G., Marinus, J., Stiggelbout, A.M., van Hilten, J.J., 2006. Reliability and validity of the Beck Depression Inventory in patients with Parkinson's disease. Movement Disord 21 (5), 668–672.

von Glischinski, M., 2019. How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). Qual. Life Res. 28 (5), 1111–1119.

Wang, W., SY, Chair, Thompson, D.R., Twinn, S.F., 2009. A psychometric evaluation of the Chinese version of the Hospital Anxiety and Depression Scale in patients with coronary heart disease. J. Clin. Nurs. 18 (13), 1908–1915.

Wilkinson, M.J., Barczak, P., 1988. Psychiatric screening in general practice: comparison of the general health questionnaire and the hospital anxiety depression scale. J. R. Coll. Gen. Pract. 38 (312), 311–313.

Woldetensay, Y.K., Belachew, T., Tesfaye, M., Spielman, K., Biesalski, H.K., Kantelhardt, E.J., et al., 2018. Validation of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression in pregnant women: Afaan Oromo version. PLoS ONE 13 (2).

Yeung, A.S., Jing, Y., Brenneman, S.K., Chang, T.E., Baer, L., Hebden, T., et al., 2012. Clinical outcomes in measurement-based treatment (Comet): a trial of depression monitoring and feedback to primary care physicians. Depress Anxiety 29 (10), 865–873.

Yu, X., Stewart, S.M., Wong, P.T., Lam, T.H., 2011. Screening for depression with the Patient Health Questionnaire-2 (PHQ-2) among the general population in Hong Kong. J. Affect. Disord. 134 (1–3), 444–4447.

Yu, X., Tam, W.W., Wong, P.T., Lam, T.H., Stewart, S.M., 2012. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr. Psychiatry 53 (1), 95–102.

Zigmond, A.S., Snaith, R.P., 1983. The hospital anxiety and depression scale. Acta Psychiatr. Scand. 67 (6), 361–370.