Generalized anxiety disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults

Zekeriya Aktürk, Alexander Hapfelmeier, Alexey Fomenko, Daniel Dümmler, Stefanie Eck, Michaela Olm, Jan Gehrmann, Victoria von Schrottenberg, Rahel Rehder, Sarah Dawson, Bernd Löwe, Gerta Rücker, Antonius Schneider, Klaus Linde

Angaben zur Veröffentlichung / Publication details:

Aktürk, Zekeriya, Alexander Hapfelmeier, Alexey Fomenko, Daniel Dümmler, Stefanie Eck, Michaela Olm, Jan Gehrmann, et al. 2025. "Generalized anxiety disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults." *Cochrane Database of Systematic Reviews* 2025 (3): CD015455. https://doi.org/10.1002/14651858.cd015455.

Cochrane Library

Cochrane Database of Systematic Reviews

Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults (Review)

Aktürk Z, Hapfelmeier A, Fomenko A, Dümmler D, Eck S, Olm M, Gehrmann J, von Schrottenberg V, Rehder R, Dawson S, Löwe B, Rücker G, Schneider A, Linde K

Aktürk Z, Hapfelmeier A, Fomenko A, Dümmler D, Eck S, Olm M, Gehrmann J, von Schrottenberg V, Rehder R, Dawson S, Löwe B, Rücker G, Schneider A, Linde K.
Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults.
Cochrane Database of Systematic Reviews 2025, Issue 3. Art. No.: CD015455.
DOI: 10.1002/14651858.CD015455.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Diagnostic Test Accuracy Review]

Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults

Zekeriya Aktürk[1,2], Alexander Hapfelmeier[1,3], Alexey Fomenko[1], Daniel Dümmler[1], Stefanie Eck[1], Michaela Olm[1], Jan Gehrmann[1,4], Victoria von Schrottenberg[1], Rahel Rehder[1], Sarah Dawson[5], Bernd Löwe[6], Gerta Rücker[7], Antonius Schneider[1], Klaus Linde[1]

[1]Institute of General Practice and Health Services Research, TUM School of Medicine and Health, Department of Clinical Medicine, Technical University of Munich, Munich, Germany. [2]General Practice, Faculty of Medicine, University of Augsburg, Augsburg, Germany. [3]Institute of AI and Informatics in Medicine, TUM School of Medicine and Health, Department of Clinical Medicine, Technical University of Munich, Munich, Germany. [4]Chair of Social Determinants of Health, TUM School of Medicine and Health, Department of Health and Sport Sciences, Technical University of Munich, Munich, Germany. [5]Bristol Medical School, University of Bristol, Bristol, UK. [6]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [7]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

Contact: Zekeriya Aktürk, zekeriya.aktuerk@med.uni-augsburg.de.

Editorial group: Cochrane Central Editorial Service.
Publication status and date: New, published in Issue 3, 2025.

Citation: Aktürk Z, Hapfelmeier A, Fomenko A, Dümmler D, Eck S, Olm M, Gehrmann J, von Schrottenberg V, Rehder R, Dawson S, Löwe B, Rücker G, Schneider A, Linde K. Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults. *Cochrane Database of Systematic Reviews* 2025, Issue 3. Art. No.: CD015455. DOI: 10.1002/14651858.CD015455.

ABSTRACT

Background

Anxiety disorders often remain undetected and can cause substantial burden. Amongst the many anxiety screening tools, the 7-item Generalized Anxiety Disorder (GAD-7) scale and its short version, the 2-item Generalized Anxiety Disorder (GAD-2) scale, are the most frequently used instruments.

Objectives

Primary: to determine the diagnostic accuracy of GAD-7 and GAD-2 to detect generalised anxiety disorder (GAD) and any anxiety disorder (AAD) in adults.

Secondary: to investigate whether their diagnostic accuracy varies by setting, anxiety disorder prevalence, reference standard, and risk of bias; to compare the diagnostic accuracy of GAD-7 and GAD-2; to investigate how diagnostic performance changes with the test threshold.

Search methods

We searched MEDLINE, Embase, PubMed-not-MEDLINE subset, and PsycINFO from 1990 to 18 January 2024. We checked reference lists of included studies and review articles.

Selection criteria

We included cross-sectional studies conducted in adults, containing diagnostic accuracy information on GAD-7 and/or GAD-2 questionnaires for the target conditions generalised anxiety disorder and/or any anxiety disorder, and allowing the generation of 2x2 tables. The target conditions must have been diagnosed using a structured or semi-structured clinical interview. We excluded case-control studies and studies in which the time elapsed between the index tests and reference standards exceeded four weeks. We excluded studies involving people (1) seeking help in mental health settings or (2) recruited specifically due to mental health symptoms in other settings.

Data collection and analysis

At least two review authors independently decided on study eligibility, extracted data, and assessed the risk of bias and applicability of included studies. For each questionnaire and each target condition, we present sensitivity and specificity with 95% confidence intervals (95% CI) in forest plots. We used the bivariate model to obtain summary estimates based on cut-offs closest to the recommended values (i.e. within a core range). In secondary analyses, we used the bivariate model and the multiple thresholds model to obtain summary estimates for all available cut-off points. Using the multiple thresholds model, we also calculated the area under the receiver operating characteristic curve to obtain a general indicator of the diagnostic accuracy of GAD-7 and GAD-2.

Main results

We included 48 studies with 19,228 participants from 27 different countries, evaluating the GAD-7 and the GAD-2 in 24 different languages. Seven studies were performed in non-clinical settings, nine in clinical settings recruiting participants across conditions, and 32 in clinical settings with participants having specific conditions. Even after categorisation into three settings, the study populations were substantially different. The most frequently studied populations were people: with epilepsy (nine studies); with cancer (five studies); with cardiovascular disease (five studies); and in primary care regardless of their condition (five studies). We considered the risk of bias low in eight studies, and we had low concerns about the applicability of findings in three studies.

Thirty-five studies contributed to the primary analyses of GAD-7 for detecting generalised anxiety disorder (median prevalence 12%); 22 studies to analyses of GAD-7 for any anxiety disorder (median prevalence 19%); 24 studies to analyses of GAD-2 for generalised anxiety disorder (median prevalence 9%); and 19 studies to analyses of GAD-2 for any anxiety disorder (median prevalence 19%).

At the recommended cut-off of 10 or higher (or the closest available cut-off), the GAD-7 questionnaire yielded a summary sensitivity of 0.64 (95% CI 0.56 to 0.72) and a summary specificity of 0.91 (95% CI 0.87 to 0.93) in detecting generalised anxiety disorder. For detecting any anxiety disorder, summary sensitivity was 0.48 (95% CI 0.40 to 0.57) and summary specificity 0.91 (95% CI 0.89 to 0.93).

At the recommended cut-off of 3 or higher (or the closest available cut-off), the GAD-2 yielded a summary sensitivity of 0.68 (95% CI 0.59 to 0.75) and a summary specificity of 0.86 (95% CI 0.82 to 0.89) for detecting generalised anxiety disorder. For detecting any anxiety disorder, the summary sensitivity was 0.53 (95% CI 0.44 to 0.62) and the summary specificity was 0.89 (95% CI 0.86 to 0.91).

The 95% prediction region of GAD-7 for detecting generalised anxiety disorder was larger (indicating pronounced statistical heterogeneity) than for the three other analyses. Specificity varied by setting in the analysis of GAD-7 and GAD-2 for detecting any anxiety disorder, and by reference standard in the analysis of GAD-2 for detecting generalised anxiety disorder. Sensitivity varied with prevalence in the analysis of GAD-7 for generalised anxiety disorder. Other investigations of potential sources of heterogeneity did not show statistically significant associations with test accuracy. In all analyses, sensitivity tended to be higher and specificity lower in participants with specific conditions compared to the other two settings. Overall, the heterogeneity in the subgroup analyses remained high.

The area under the receiver operating characteristic curve in the multiple thresholds model was 0.86 (95% CI 0.84 to 0.88) for the GAD-7 scale in detecting generalised anxiety disorder, and 0.80 (95% CI 0.78 to 0.82) in detecting any anxiety disorders. For the GAD-2 scale, the value was 0.82 (95% CI 0.81 to 0.86) for detecting generalised anxiety disorder, and 0.77 (95% CI 0.76 to 0.82) for detecting any anxiety disorders. Comparative bivariate analyses revealed no statistically significant differences between the diagnostic test accuracy of GAD-7 and GAD-2.

Authors' conclusions

The GAD-7 and the GAD-2 scales have been tested in numerous languages and different populations. Overall, the GAD-7 and the GAD-2 seem to have acceptable or good diagnostic accuracy for both generalised anxiety disorder and any anxiety disorder. The GAD-2 scale seems to have similar diagnostic accuracy as the GAD-7 scale. However, due to the diversity of the included studies and the heterogeneity of our findings, our summary estimates of sensitivity and specificity should be interpreted as rough averages. The performance of GAD-7 and GAD-2 may deviate substantially from these values in specific situations.

PLAIN LANGUAGE SUMMARY

How accurate are GAD-7 and GAD-2 questionnaires for detecting anxiety disorders?

Key messages

• The GAD-7 and GAD-2 questionnaires *alone* cannot be used to diagnose or rule out an anxiety disorder.

• However, they provide an indication of whether an anxiety disorder may be present.

• The interpretation of a "negative" or "positive" questionnaire finding of an individual depends on the context.

What are anxiety disorders?

'Anxiety disorder' is an umbrella term which refers to mental health conditions including (but not limited to):

• generalised anxiety disorder: when someone experiences excessive anxiousness for most days over six months, difficulty controlling worry, plus at least three of six symptoms: restlessness, fatigue, difficulty concentrating, irritability, muscle tension, and sleep disturbance;

• social phobia: a strong fear of being judged, embarrassed, or humiliated in social or performance situations;

• panic disorder: repeatedly experiencing unexpected and intense periods of fear or discomfort, which may involve a variety of physical and emotional symptoms.

Why is an early diagnosis of anxiety disorders important?

Anxiety disorders are common and often remain undetected, even in people for whom treatment can be regarded as necessary. Not recognising an anxiety disorder when it is present (false negative) means that you miss the opportunity of timely treatment with medicine, psychotherapy, or both. For this reason, some experts and patient initiatives advocate screening; that is, the systematic examination of a social group or population for the presence of anxiety disorders, even in people who do not seek help for corresponding symptoms. Other experts argue against screening as there is no evidence that such screening has more benefit than harm (e.g. through misdiagnosis and side effects of unnecessarily prescribed medication).

What are the GAD-7 and GAD-2 questionnaires?

The GAD-7 (Generalized Anxiety Disorder 7-item scale) and the GAD-2 (the first two items of the GAD-7 scale) are user-friendly, self-completed questionnaires designed for laypeople. They help provide information on whether an individual may be suffering from an anxiety disorder. People completing the questionnaire indicate on a scale from 0 (never present) to 3 (present daily) how often they experience important anxiety symptoms. Responses are added up to obtain a total score ranging from 0 to 21. The GAD-7 is deemed 'test positive' when the total score reaches 10 or above, suggesting a potential anxiety disorder. A score of 3 or higher on the GAD-2 indicates the same. If the total score is below 10 for GAD-7 or below 3 for GAD-2, the result is considered 'test negative'.

What did we want to find out?

We aimed to find out how accurately the GAD-7 and GAD-2 scales can tell whether a person has an anxiety disorder or not.

What did we do?

We searched for studies that investigated the accuracy of the GAD-7 or GAD-2 scales (or both) by comparing them with a full diagnostic interview conducted by a healthcare professional, and we combined the results from these studies.

What did we find?

Our review includes results from 48 studies with 19,228 people from 27 different countries, evaluating the GAD-7 and the GAD-2 questionnaire in 24 different languages. Seven studies were undertaken in non-clinical settings (e.g. general population), nine in clinical settings recruiting participants with different conditions (e.g. people who sought primary care), and 32 in clinical settings with participants having specific conditions (e.g. people with epilepsy). Many of these studies investigated the accuracy of the questionnaires for one specific type of anxiety disorder (generalised anxiety disorder), others for any type of anxiety disorder, or for both.

Overall, the GAD-7 and GAD-2 questionnaires were acceptable or good at determining whether a person was suffering from an anxiety disorder.

To illustrate, if the GAD-7 questionnaire was used by a group of 1000 people, 120 (12%) of whom actually have a generalised anxiety disorder, one would expect the following results: an estimated 156 people (16%) would be classified 'test positive' by the GAD-7. However, of these, only about half (77 people or 8%) actually would have generalised anxiety disorder, while the other half (79 people or 8%) would be "false positives." Of the 844 (84%) people who tested negative, the vast majority (801 people or 80%) would be correctly classified as not suffering from generalised anxiety disorder, but 43 people (4%) with generalised anxiety disorder would be "false negatives".

The shorter GAD-2 questionnaire seems as accurate as the longer GAD-7 scale in detecting anxiety disorders. Both questionnaires seem slightly better at detecting generalised anxiety disorder than detecting any anxiety disorder.

What are the limitations of the evidence?

The studies varied in several ways, generally demonstrating poor methodological quality. The variations included differences in the population to whom the questionnaires were administered, the language of the questionnaire, methodological quality, and the method used for making the final diagnosis. These variations might explain why the accuracy of the GAD-7 and GAD-2 significantly differs across individual studies. This means that the diagnostic accuracy of these questionnaires could be better or worse in specific situations than in the summary illustration we provided above.

How current is the evidence?

The evidence is current to January 2024.

SUMMARY OF FINDINGS

Summary of findings 1. GAD-7 for generalised anxiety disorder

What is the diagnostic accuracy of GAD-7 for generalised anxiety disorder (GAD)?

Number of studies: 36 (35 included in the primary analysis)*
Settings and study populations: adults with heterogeneous backgrounds

- Non-clinical settings (n = 5): general population, migrant workers, students, elderly, mixed healthy persons/patients
- Clinical settings – across conditions (n = 6): unselected, primary care (n = 5), high utilisers, mixed inpatients
- Clinical settings – specific conditions (n = 24): epilepsy (n = 9), pain conditions (n = 4), cancer (n = 3), other (n = 8)

Index test: Generalized Anxiety Disorder seven-item questionnaire (GAD-7)

Reference standard: validated structured or semi-structured diagnostic interviews

Study design: cross-sectional

Certainty of evidence:

- Risk of bias: overall risk of bias was considered unclear or high in 27 studies.
- Applicability: none of the studies entailed concerns regarding applicability in the index test or reference standard domains. However, there were concerns about whether the included participants matched the review question in 33 studies.
- Between-study variability: point estimates from individual studies for sensitivity ranged widely from 0.09 to 0.91, those for specificity from 0.60 to 0.99.

Summary sensitivity, specificity, PPV, and NPV (and 95% CIs) across all studies and setting subgroups based on the core range cut-offs of 8 to 10 according to the bivariate model:

Setting	Studies/ participants	Summary sensitivity	Summary specificity	PPV	NPV	Meaning of the findings**
Any setting (Median prevalence: 12%)	35/15,274*	0.64 (0.56 to 0.72)	0.91 (0.87 to 0.93)	0.48 (0.42 to 0.55)	0.95 (0.94 to 0.96)	TP: 77 FP: 79 FN: 43 TN: 801
Non-clinical (Median prevalence: 8%)	5/3458	0.50 (0.30 to 0.70)	0.93 (0.84 to 0.97)	0.37 (0.23 to 0.54)	0.96 (0.93 to 0.97)	TP: 40 FP: 64 FN: 40 TN: 856

Clinical – across conditions (Median prevalence: 8%)	7/5486	0.57 (0.39 to 0.73)	0.93 (0.86 to 0.97)	0.42 (0.28 to 0.56)	0.96 (0.94 to 0.97)	TP: 46 FP: 64 FN: 34 TN: 856
Clinical – specific conditions (Median prevalence: 16%)	23/6330	0.70 (0.60 to 0.78)	0.89 (0.85 to 0.93)	0.55 (0.47 to 0.64)	0.94 (0.92 to 0.95)	TP: 112 FP: 92 FN: 48 TN: 748
Area under the curve from the multiple thresholds analysis across all settings: 0.86 (95% CI 0.84 to 0.88)						

CI: confidence interval; **NPV:** negative predictive value; **PPV:** positive predictive value
***Mughal 2021** not included in the primary analyses since the cut-off value reported was outside the core range.
**True positive (TP: appropriate further diagnostics), false positive (FP: unnecessary further diagnostic inquiry), false negative (FN: no further diagnostics - wrong reassurance), and true negative (TN: no further diagnostics - appropriate reassurance); values based on a hypothetical cohort of 1000 people

Summary of findings 2. GAD-7 for any anxiety disorder

What is the diagnostic accuracy of GAD-7 for any anxiety disorder (AAD)?

Number of studies: 22

Settings and study populations: adults with heterogeneous backgrounds

- Non-clinical settings (n = 3): general population, poor elderly, mixed health/patients
- Clinical settings – across conditions (n = 4): unselected primary care, sick-listed workers, under-/uninsured adults
- Clinical settings – specific conditions (n = 15): epilepsy (n = 3), cancer (n = 4), heart diseases (n = 2), other (n = 6)

Index test: Generalized Anxiety Disorder seven-item questionnaire (GAD-7)

Reference standard: validated structured or semi-structured diagnostic interviews

Study design: cross-sectional

Certainty of evidence:

- Risk of bias: overall risk of bias was unclear or high in 16 studies.
- Applicability: none of the studies entailed concerns regarding applicability in the index test or reference standard domains. However, there were concerns about whether the included participants matched the review question in 21 studies.

- Between-study variability: point estimates from individual studies for sensitivity ranged widely from 0.25 to 0.86, those for specificity from 0.80 to 1.00.

Summary sensitivity, specificity, PPV, and NPV (and 95% CIs) across all studies and for setting subgroups based on the core range cut-offs 8 to 10 according to the bivariate model:

Setting	Studies/ participants	Summary sensitivity	Summary specificity	PPV	NPV	Meaning of the findings*
Any setting (Median prevalence: 19%)	22/10245	0.48 (0.40 to 0.57)	0.91 (0.89 to 0.93)	0.57 (0.51 to 0.63)	0.88 (0.87 to 0.90)	TP: 91 FP: 73 FN: 99 TN: 737
Non-clinical (Median prevalence: 19%)	3/1810	0.43 (0.23 to 0.64)	0.96 (0.93 to 0.98)	0.72 (0.54 to 0.85)	0.88 (0.85 to 0.9)	TP: 82 FP: 32 FN: 108 TN: 778
Clinical – across conditions (Median prevalence: 18%)	4/2652	0.48 (0.29 to 0.67)	0.89 (0.83 to 0.93)	0.49 (0.34 to 0.64)	0.89 (0.86 to 0.91)	TP: 86 FP: 90 FN: 94 TN: 730
Clinical – specific conditions (Median prevalence: 19%)	15/5046	0.50 (0.40 to 0.60)	0.90 (0.87 to 0.93)	0.54 (0.46 to 0.62)	0.88 (0.87 to 0.9)	TP: 95 FP: 81 FN: 95 TN: 824
Area under the curve from the multiple thresholds analysis across all settings: 0.80 (95% CI 0.78 to 0.82)						

CI: confidence interval; **NPV:** negative predictive value; **PPV:** positive predictive value
*True positive (TP: appropriate further diagnostics), false positive (FP: unnecessary further diagnostic inquiry), false negative (FN: no further diagnostics - wrong reassurance), and true negative (TN: no further diagnostics - appropriate reassurance); values based on a hypothetical cohort of 1000 people.

Summary of findings 3. GAD-2 for generalised anxiety disorder

What is the diagnostic accuracy of GAD-2 for generalised anxiety disorder (GAD)?

Number of studies included in the primary analysis: 24

Settings and study populations: adults with heterogeneous backgrounds

- Non-clinical settings (n = 5): general population (n = 2), students, poor elderly, mixed healthy persons/patients
- Clinical settings – across conditions (n = 5): unselected primary care (n = 4), high utilisers
- Clinical settings – specific conditions (n = 14): pain conditions (n = 4), epilepsy (n = 3), other (n = 7)

Index test: Generalized Anxiety Disorder two-item questionnaire (GAD-2)

Reference standard: validated structured or semi-structured diagnostic interviews

Study design: cross-sectional

Certainty of evidence:

- Risk of bias: overall risk of bias was considered unclear or high in 17 studies
- Applicability: none of the studies entailed concerns regarding applicability in the index test or reference standard domains. However, there were concerns about whether the included participants matched the review question in 22 studies.
- Between-study variability: point estimates from individual studies for sensitivity ranged widely from 0.12 to 0.89, those for specificity from 0.60 to 0.96.

Summary sensitivity, specificity, PPV, and NPV (and 95% CIs) across all studies and for setting subgroups based on the core range cut-offs 2 and 3 according to the bivariate model:

Setting	Studies/ participants	Summary sensitivity	Summary specificity	PPV	NPV	Meaning of the findings**
Any setting (Median prevalence: 9%)	24/11,428	0.68 (0.59 to 0.75)	0.86 (0.82 to 0.89)	0.32 (0.28 to 0.36)	0.96 (0.96 to 0.97)	TP: 61 FP: 127 FN: 29 TN: 783
Non-clinical (Median prevalence: 8%)	5/3942	0.66 (0.49 to 0.80)	0.88 (0.81 to 0.93)	0.32 (0.22 to 0.45)	0.97 (0.96 to 0.98)	TP: 53 FP: 110 FN: 27 TN: 810

Clinical – across conditions (Median prevalence: 8%)	5/3348	0.63 (0.45 to 0.78)	0.86 (0.78 to 0.92)	0.29 (0.19 to 0.41)	0.96 (0.95 to 0.97)	TP: 50 FP: 129 FN: 30 TN: 791
Clinical – specific conditions (Median prevalence: 11%)	14/4138	0.70 (0.59 to 0.79)	0.84 (0.79 to 0.89)	0.35 (0.28 to 0.44)	0.96 (0.95 to 0.96)	TP: 77 FP: 142 FN: 33 TN: 748
Area under the curve from the multiple thresholds analysis across all settings: 0.82 (95% CI 0.81 to 0.86)						

CI: confidence interval; **NPV:** negative predictive value; **PPV:** positive predictive value
*True positive (TP: appropriate further diagnostics), false positive (FP: unnecessary further diagnostic inquiry), false negative (FN: no further diagnostics - wrong reassurance), and true negative (TN: no further diagnostics - appropriate reassurance); values based on a hypothetical cohort of 1000 people.

Summary of findings 4. GAD-2 for any anxiety disorder

What is the diagnostic accuracy of GAD-2 for any anxiety disorder (AAD)?

Number of studies included in the primary analysis: 19

Settings and study populations: adults with heterogeneous backgrounds

- Non-clinical settings (n = 4): general population (n = 2), poor elderly, mixed healthy persons/patients
- Clinical settings – across conditions (n = 2): unselected primary care
- Clinical settings – specific conditions (n = 13): epilepsy (n = 2), cancer (n = 2), heart diseases (n = 2), pregnancy (n = 2), other (n = 5)

Population and setting: adults with heterogeneous backgrounds. Most studies (n = 16) were conducted in people with specific diseases. Others included people with various conditions (n = 2), and non-clinical general populations (n = 4). Amongst the specific diseases were cancer (n = 2), heart diseases (n = 3), and epilepsy (n = 2).

Index test: Generalized Anxiety Disorder two-item questionnaire (GAD-2)

Reference standard: validated structured or semi-structured diagnostic interviews

Study design: cross-sectional

Certainty of evidence:

- Risk of bias: overall risk of bias was unclear or high in 13 studies.

- Applicability: none of the studies entailed concerns regarding applicability in the index test or reference standard domains. However, there were concerns about whether the included participants matched the review question in all 19 studies.
- Between-study variability: point estimates from individual studies for sensitivity ranged widely from 0.22 to 0.96, those for specificity from 0.74 to 0.96.

Summary sensitivity, specificity, PPV, and NPV (and 95% CIs) across all studies and setting/pathway subgroups based on the core range cut-offs of 2 and 3 according to the bivariate model:

Setting	Studies/ participants	Summary sensitivity	Summary specificity	PPV	NPV	Meaning of the findings*
Any setting (Median prevalence: 19%)	19/9973	0.53 (0.44 to 0.62)	0.89 (0.86 to 0.91)	0.53 (0.47 to 0.58)	0.89 (0.87 to 0.91)	TP: 101 FP: 89 FN: 89 TN: 721
Non-clinical (Median prevalence: 20%)	4/3134	0.50 (0.32 to 0.68)	0.93 (0.90 to 0.95)	0.64 (0.48 to 0.78)	0.88 (0.86 to 0.90)	TP: 100 FP: 56 FN: 100 TN: 744
Clinical – across conditions (Median prevalence: 18%)	2/2432	0.53 (0.28 to 0.76)	0.87 (0.81 to 0.92)	0.47 (0.28 to 0.67)	0.89 (0.87 to 0.91)	TP: 94 FP: 107 FN: 86 TN: 713
Clinical – specific conditions (Median prevalence: 19%)	13/4407	0.54 (0.43 to 0.65)	0.87 (0.84 to 0.89)	0.49 (0.4 to 0.58)	0.89 (0.88 to 0.90)	TP: 103 FP: 105 FN: 87 TN: 705
Area under the curve from the multiple thresholds analysis across all settings: 0.77 (95% CI 0.76 to 0.82)						

CI: confidence interval; **NPV:** negative predictive value; **PPV:** positive predictive value
*True positive (TP: appropriate further diagnostics), false positive (FP: unnecessary further diagnostic inquiry), false negative (FN: no further diagnostics - wrong reassurance), and true negative (TN: no further diagnostics - appropriate reassurance); values based on a hypothetical cohort of 1000 people.

BACKGROUND

This is the first of a series of four Cochrane reviews following a joint protocol (Linde 2022), addressing the diagnostic accuracy of the Generalized Anxiety Disorder 7-item (GAD-7) Scale and the Generalized Anxiety Disorder 2-item (GAD-2) Scale, the Hospital Anxiety and Depression Scale (HADS), the Beck Anxiety Inventory (BAI), and the State-Trait Anxiety Inventory (STAI).

Target condition being diagnosed

Anxiety disorders are the most prevalent mental disorders worldwide (Bandelow 2015; Kessler 2001; Remes 2016). Estimates of the prevalence of anxiety disorders within a 12-month period range from 8% to 21% (Alonso 2007; Kessler 2012a; Kessler 2012b). Besides their high prevalence, which further increased during the COVID-19 pandemic (Santamauro 2021), anxiety disorders are often chronic and associated with physical, social, emotional, and functional impairments (Hoffman 2008; Schonfeld 1997). Anxiety disorders are also risk factors for hospitalisation and increased mortality in chronic illness (Gudmundsson 2005; Schneider 2008; Vongmany 2016). Measured by years of life lived with disability, anxiety disorders were the sixth leading cause of all disability (Baxter 2014).

Anxiety disorders are classified in the Diagnostic and Statistical Manual of Mental Disorders (DSM IV or DSM 5) (APA 2000; APA 2013), and the International Statistical Classification of Diseases and Related Health Problems (ICD-10 or ICD-11) (WHO 1993; WHO 2019). Important types of anxiety disorder include generalised anxiety disorder (GAD), agoraphobia, specific phobia, social anxiety disorder (social phobia), post-traumatic stress disorder, obsessive-compulsive disorder, and separation anxiety disorder. Obsessive-compulsive disorder and post-traumatic stress disorder were included in former classifications but were removed from anxiety disorders with the introduction of DSM-V (Bandelow 2015). People with generalised anxiety disorder demonstrate excessive anxiety and worry for most days over six months, with difficulty in controlling worry, along with at least three out of six associated symptoms: restlessness, fatigue, difficulty in concentrating, irritability, muscle tension, and sleep disturbance (APA 2013).

Most anxiety disorders go undetected and untreated (Craske 2017), partly due to the difficulty of diagnosis. This is because anxiety symptoms can overlap with those of other mental diseases (Olariu 2015; Sherbourne 1996; Toft 2005), and with symptoms caused by organic diseases. Symptoms such as breathlessness often mimic other conditions such as somatic symptom disorder, heart failure, asthma, or chronic obstructive pulmonary disease (Schneider 2013), which makes correct labelling of people with anxiety disorders even more important.

Although anxiety disorders often go unrecognised, there are effective treatment options. Therefore, some expert panels have recommended screening with self-report questionnaires. In such questionnaires, individuals are asked to rate the presence and severity of common anxiety symptoms. In 2023, for example, the US Preventive Services Task Force recommended screening for anxiety disorders in adults (USPSTF 2023). Similar recommendations have been published for more restricted populations, such as adolescent and adult women (Gregory 2020), adults with cancer (Howell 2015), and stroke patients (Crow 2023). Some experts consider such recommendations to be premature, as there is a lack of evidence from randomised trials to show that anxiety screening brings more benefit than harm. They argue that routine screening will lead to many false positives, increase the likelihood of over-treatment, and harm some people due to the adverse effects of medications (Thombs 2023).

An important prerequisite for effective screening is that self-report questionnaires can reliably discriminate between individuals with and without anxiety disorders. Our series of reviews summarises studies which investigate this discriminatory ability of questionnaires, referred to as 'diagnostic accuracy'. To investigate diagnostic accuracy, studies compare the results of a self-report questionnaire against a 'reference standard'. The reference standard is a validated structured or semi-structured interview based on DSM or ICD criteria administered by a clinician or a well-trained person. For our reviews, we decided to focus on generalised anxiety disorder (GAD) and any anxiety disorder (AAD) as target conditions for primary analyses.

For general screening in non-psychiatric settings, GAD is both clinically relevant and was the most frequently analysed specific anxiety disorder in an important previous systematic review (Plummer 2016). The overarching category AAD was the second most frequently investigated and reported category in this review. Unfortunately, the list of specific anxiety disorders included in this category can vary between studies, depending on the classification used. Literature searches in the planning phase of our reviews revealed that diagnostic results for other specific diagnoses were rarely reported.

Index test(s)

The 7-item Generalized Anxiety Disorder (GAD-7) scale (Löwe 2008; Spitzer 2006), and its short version, the 2-item Generalized Anxiety Disorder (GAD-2) scale (Kroenke 2007), are amongst the most frequently used instruments to screen for anxiety disorders.

The GAD-7 scale was originally designed as a screening tool for generalised anxiety disorder, but is also used for screening for anxiety disorders in a broader manner (Kroenke 2007; Löwe 2008; Spitzer 2006). It has seven items, each with four answer options (0 – Not at all; 1 – Several days; 2 – More than half the days; 3 – Nearly every day). The item scores are added up, resulting in a score ranging from 0 to 21. The cut-off point (above which a finding is considered as "test positive" or "screen positive") recommended by the scale authors is 10 or higher (Spitzer 2006). The GAD-2 scale is the short version of the GAD-7 scale. It has two items, which produce a score ranging from 0 to 6, and a recommended cut-off of three or higher (Kroenke 2007). In addition to their use as screening instruments, the GAD-7 and GAD-2 scales are often used for monitoring the severity of anxiety symptoms during the course of treatment or as measurement tools in epidemiological and clinical research. However, our review does not address these latter uses. A recent review by the US Preventive Services Task Force reported that the GAD-7 and GAD-2 demonstrated acceptable sensitivity (0.79, 95% confidence interval (CI) 0.65 to 0.94, and 0.76, 95% CI 0.68 to 0.85, respectively) and specificity (0.89, 95% 0.83 to 0.94, and 0.88, 95% CI 0.87 to 0.88, respectively) for detecting generalised anxiety disorder (O'Connor 2023a; O'Connor 2023b). For detecting any anxiety disorders, values were slightly lower. However, these estimates were based on only two and three studies, respectively.

Clinical pathway

In this review, we focus on studies relevant for a screening pathway. The screening pathway comprises the following steps.

- The GAD-7 or GAD-2 test is administered to all individuals in a particular setting who are not known to have an anxiety disorder or who seek psychiatric evaluation or treatment.
- Individuals scoring below a predefined cut-off would be considered screen-negative and not further evaluated, whereas those at or above the cut-off would be considered screen-positive and in need of evaluation. A mental health evaluation should be conducted by a competent health professional, who decides whether an anxiety disorder is present.
- Individuals deemed in need of management would be referred, and discussions regarding treatment options and subsequent treatment would be offered.
- Ongoing follow-up care would be provided for those undergoing management.

We explored the following three distinct basic settings for implementing a screening pathway.

- Non-clinical settings, such as screening in the general adult population or specific groups not seeking healthcare (e.g. students, elderly).
- Clinical settings across conditions, including screening unselected primary care patients or hospital patients, who are not seeking care specifically for anxiety.
- Clinical settings for specific conditions, such as screening patients seeking care for cancer or epilepsy.

Prior test(s)

There are no prior tests specifically for anxiety disorder in the populations included in the review.

Role of index test(s)

In the pathway addressed in our review, the GAD-7 or GAD-2 can be considered as triage tools.

Alternative test(s)

There is a plethora of anxiety self-report questionnaires (at least 25), which have been used in a variety of settings and populations (Benjamin 2011; Creighton 2018; Lazor 2017; Litster 2016; Mele 2018; Sinesi 2019). Most of these instruments have been investigated in a limited number of studies and seem to be used infrequently. In addition, some are only available in one or a few languages. However, some instruments, which have undergone intense testing, are available in several languages and are commonly used. For example, the Hospital Anxiety and Depression Scale (HADS) is a widely used questionnaire with seven items each for measuring anxiety (HADS Anxiety or HADS-A) and depression (HADS Depression or HADS-D) (Zigmond 1983), with a score range from 0 to 21. The Beck Anxiety Inventory (BAI) has 21 items and a score range from 0 to 63 (Beck 1988; Beck 1993). The State-Trait Anxiety Inventory (STAI) aims to measure two types of anxiety: state anxiety (anxiety about an event) and trait anxiety (anxiety level as a personal characteristic) (Spielberger 1970). Both subscales have 20 items and a score range from 20 to 80. The diagnostic accuracy of the HADS, BAI, and STAI will be the focus of upcoming reviews, as illustrated in our protocol (Linde 2022).

Rationale

For this review, we had several rationales for using the GAD-7 scale and its shortened version, GAD-2. First, despite their more recent development, these instruments are amongst the most commonly used anxiety screening tools, have been translated to over 70 languages (https://www.phqscreeners.com/), and validated in many specific populations. Second, they are self-administered, easy-to-interpret scales requiring one to two minutes to complete (Löwe 2008; Spitzer 2006). Third, the GAD-7 scale has high reliability (Cronbach alpha = 0.92) and procedural validity (intraclass correlation = 0.83). Further psychometric properties of the two scales were tested in several studies, and good test-retest reliability, criterion validity, construct validity, concurrent validity, and convergent validity were demonstrated (Kertz 2013; Löwe 2008; Spitzer 2006). Fourth, a systematic review and meta-analysis of diagnostic test accuracy for anxiety disorders was published in 2016 for the GAD-7 and the GAD-2 scales (Plummer 2016), but new studies have since been published, and advances in methodology of diagnostic accuracy meta-analysis have made it possible to include all available cut-off values simultaneously (Steinhauser 2016). The above-mentioned recent evidence summary of the US Preventive Services Task Force (USPSTF) included only a small fraction of the available diagnostic accuracy studies relevant to the screening pathway (O'Connor 2023a; O'Connor 2023b). Therefore, there was a need for a systematic review comprehensively summarising the available evidence on the diagnostic accuracy of the GAD-7 and the GAD-2.

OBJECTIVES

To determine the diagnostic accuracy of GAD-7 and GAD-2 to detect generalised anxiety disorder (GAD) and any anxiety disorder (AAD) in adults.

Secondary objectives

- To investigate whether their diagnostic accuracy varies by setting, anxiety disorder prevalence, reference standard, and risk of bias;
- To compare the diagnostic accuracy of GAD-7 and GAD-2;
- To investigate how diagnostic performance changes with the test threshold.

METHODS

Criteria for considering studies for this review

Types of studies

We included studies comparing findings from the GAD-7, GAD-2, or both, with the diagnostic status according to a reference standard that allowed the generation of 2×2 tables (number of true-positive, false-positive, false-negative and true-negative index test results). We excluded case-control studies in which cases suffered from anxiety disorders, as estimates of diagnostic test accuracy derived from such studies might not apply to the diagnostic process in real situations due to the selection involved.

Participants

We included studies involving adults (18 years or older). Participants could have been recruited from: (1) non-clinical settings, such as from the general adult population or any

subpopulation groups not seeking healthcare (e.g. students, elderly); (2) clinical settings across conditions, such as unselected primary care patients or hospital patients not seeking care specifically for anxiety; (3) clinical settings for specific conditions known to have an increased risk of anxiety disorders, such as outpatient units for cancer, cardiovascular disease, or pregnancy.

We excluded studies involving people seeking help in mental health settings or recruited specifically due to mental health symptoms in other settings. Furthermore, we excluded studies involving children, as this population requires specific, age-adapted self-report questionnaires (Lazor 2017).

Index tests

We included studies evaluating the GAD-7 or GAD-2 scales, or both, as screening tools for anxiety disorders. These index tests are described in detail in the Background section.

Target conditions

Studies had to address at least one of the following diagnostic categories: (1) presence or absence of generalised anxiety disorder; (2) presence or absence of any anxiety disorder.

Reference standards

Validated semi-structured and structured interviews are accepted as state-of-the-art procedures in epidemiological and diagnostic studies on mental disorders (Suppiger 2009). They ensure sufficient quality and a certain reproducibility of diagnosis. Thus, the diagnosis of anxiety in primary studies must have been made or ruled out using a validated standardised or structured clinical interview, such as the SCID (Structured Clinical Interview for DSM) (First 1996; First 2015), the CIDI (Composite International Diagnostic Interview) (WHO 1990), the MINI (Mini-International Neuropsychiatric Interview) (Sheehan 1998), CIS-R (Clinical Interview Schedule-Revised) (Lewis 1992), DIA-X (diagnostic expert system for mental disorders) (Wittchen 1997), SADS (Schedule for Affective Disorders and Schizophrenia) (Endicott 1978), DIPS (German: 'Diagnostisches Interview bei psychischen Störungen') (Margraf 1994), or Mini-DIPS (Margraf 2013).

We excluded studies in which the reference standard was informally based on a checklist based on ICD, DSM, or a clinical diagnosis without operationalisation or only another questionnaire. Also, we excluded studies using the diagnostic criteria of DSM-III or earlier versions. We excluded studies with a time lag of over four weeks between the index test(s) and the reference standard to limit the risk of misclassification.

Search methods for identification of studies

Electronic searches

In order to identify potentially eligible studies, we searched the following databases using relevant subject headings (controlled vocabularies), text-words, and search syntax, appropriate to each resource. All databases were searched from 1990 to 30 September 2022 (search strategies given in Appendix 1):

- MEDLINE (Ovid);
- Embase (Ovid);
- PubMed-not-MEDLINE subset (NLM);
- PsycINFO (Ovid).

We did not apply any restrictions on language or publication status to the searches.

Search structure

One of the main challenges in identifying the evidence for a diagnostic review for mental health conditions is that the index tests are also symptom inventories used to measure treatment outcomes in intervention studies. Therefore, it is difficult to differentiate between the types of studies retrieved by a search (diagnostic accuracy (diagnostic test accuracy) study or intervention study). In order to balance the sensitivity (recall) and specificity (precision) of the search, we structured it around the following key concepts:

#1 Target Condition + Index Test(s) + Reference Standard

#2 Target Condition + Index Test(s) + diagnostic accuracy Filter

#3 (#1 OR #2)

Lead by a Cochrane information specialist (co-author SD), we developed an initial Embase search benchmarking the strategy against a set of known studies before translating the search across to other databases. We paid special attention to the search terms used for the reference standard and diagnostic accuracy 'filters'. In response to peer-review, we made some amendments to our initial search, but this did not affect the number of relevant studies retrieved. The final search strategies are presented in Appendix 1.

Searching other resources

To identify further published, unpublished, or ongoing research, we scanned the reference lists of included studies and any relevant systematic reviews. We checked for any relevant retraction statements or errata of the included studies.

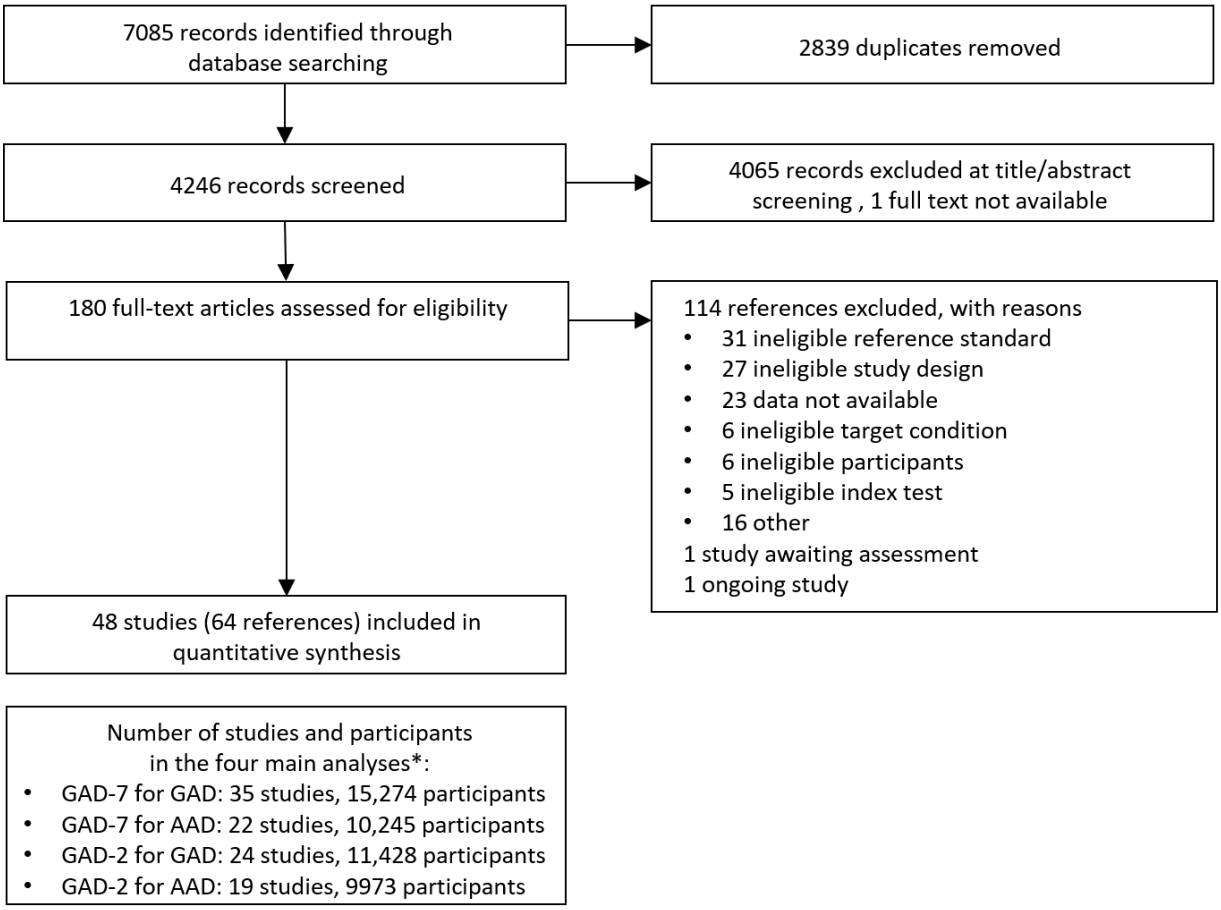
Data collection and analysis

After searching, titles from the databases were collated into the EPPI-Reviewer Web software (Thomas 2020).

Selection of studies

Pairs of review authors independently screened titles and abstracts of search hits to identify potentially relevant studies and to exclude clearly irrelevant papers. If the relevance remained unclear after title and abstract screening, we obtained and screened full texts. In the second round, two authors independently decided to include or exclude a potentially eligible study based on the review of full texts using the criteria listed above. We resolved any disagreements through discussion, involving a third review author when necessary. The study selection process is presented in a PRISMA flow diagram (McInnes 2018) (Figure 1).

Figure 1. Study flow diagram *One study not included in primary analyses (Mughal 2021). AAD: any anxiety disorder; GAD: generalised anxiety disorder; GAD-7/2: Generalised Anxiety Disorder 7-item/2-item questionnaire



Following the approach used by Quinn 2020, we contacted the corresponding study authors/investigators if eligibility remained unclear and if potentially relevant articles did not contain data required for the analyses. This approach was of particular relevance if a study collected GAD-7 or GAD-2 scores and also administered a diagnostic interview but did not report diagnostic accuracy data. When the study authors/investigators did not respond or relevant data were not provided for extraction, we excluded the study with the reason 'data not available for analysis'. If the same dataset was reported in multiple reports, we included the main (primary) report but referred to the other secondary reports if they contained relevant additional information. For studies presented only in abstract form, we contacted the main presenting study author/investigator to enquire whether the full paper had been published.

Data extraction and management

Working independently, at least two review authors extracted primary study characteristics and results using a pre-tested form. In particular, we extracted diagnoses and main inclusion criteria for participants, age, sex, setting, details regarding the reference standard (type, person doing the assessment, timing), language of the index test, predefined cut-off points ("predefined" as per study reports) or methods to select presented cut-off-points, country of origin, and numbers of participants included and analysed. For all cut-off points available, we extracted diagnostic 2x2 tables (true-positive, false-positive, true-negative, and false-negative index test results) from the publications, or if not available, reconstructed them using information about relevant parameters (prevalence, sensitivity, specificity, or predictive values). We did not ask for study protocols to verify whether cut-offs were predefined.

We extracted data from two Chinese studies after translating them with Google Translate (Qu 2016; Zeng 2013). Three articles did not report prevalence but reported positive and negative predictive values in addition to sensitivity and specificity (Ahmadi 2019; Nath 2018; Van Heyningen 2018). For these, we calculated prevalence using the formula suggested by Taylor 2021 (see https://www.cebm.ox.ac.uk/files/data-extraction-tips/blog-2.pdf). In two studies (Esser 2018; Nath 2018), we weighted the data to correct potential partial verification bias and to mimic the original sample size (i.e. the number of participants who underwent both the index test and reference standard). We corrected the data we extracted from Sidik 2012 using Kohn's formula (Kohn 2022), and weighted the data to mimic the original sample size.

We reached out to the original authors of the articles for clarification and additional data when information was missing or any of the study reports were unclear, doing so at least twice with a minimum interval of two weeks. Given that the GAD-7

questionnaire encompasses the GAD-2 questions, we contacted the authors of studies that reported results solely on the GAD-7 scale to enquire about unpublished data on the GAD-2 questionnaire. Furthermore, if the publication lacked diagnostic accuracy information for all cut-offs, we contacted study authors to request sensitivity and specificity data for additional cut-offs. For this review, we attempted to contact all authors of the included studies for extra information, resulting in additional data from 21 (44%) of them. Fourteen authors (29%) provided further cut-off data that were not disclosed in the original study reports. Finally, we sent the extracted information to the study authors for quality check and approval.

If necessary, we also asked study authors to re-analyse their data and submit summary tables or provide raw data. Ten authors provided raw data (Clover 2020; Conway 2016; Esser 2018; Kroenke 2007; Maric 2022; Marrie 2018; Michaelis 2022; Pranckeviciene 2022; Scott 2019; Simning 2012). In studies that presented receiver operating characteristic (ROC) curves that included all possible cut-offs but no sensitivity/specificity information was available (Baker 2018; Mughal 2021; Shih 2022; Veisy 2021), we used WebPlotDigitizer to extract data (Rohatgi 2014). In addition, we asked study authors to provide missing information to populate the Characteristics of included studies section and to assess the methodological quality of the included studies.

Assessment of methodological quality

At least two review authors independently assessed the risk of bias and the external validity of the included studies using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (Whiting 2011). We made a risk of bias judgement ('high', 'low', or 'unclear') for the four domains (participant selection, index test, reference standard, and flow and timing) based on signalling questions (see Appendix 2). To adapt QUADAS-2 to our aims and the expected set of primary studies, we modified signalling question 2 in the index test domain at the protocol stage. As we expected many studies would present data for several cut-offs, we asked one of the following two questions (instead of the single question "if a threshold was used, was it prespecified?"): (1) "Was there a predefined threshold for a primary analysis of DTA?"; or in studies presenting more than one threshold, (2) "If findings for several thresholds are presented, was the range of thresholds prespecified?" We implemented this adjustment because we view studies reporting sensitivity and specificity across a wide range of cut-offs as not being at a higher risk of bias compared to studies that have prespecified a specific threshold.

In addition, we used the QUADAS-C tool for studies providing diagnostic accuracy data for both GAD-7 and GAD-2 (Yang 2021). The QUADAS-C tool was developed as an extension of QUADAS-2 to assess the risk of bias specifically for the comparison of two or more diagnostic tests. The instrument uses the same domains as QUADAS-2.

If the answers to all signalling questions within a domain were judged as 'yes' (indicating low risk of bias for each question), then the domain was labelled as 'low risk of bias'. If any signalling question was judged as 'no', indicating a high risk of bias, the domain was scored as at 'high risk of bias'. Otherwise, the domain was rated 'unclear'. This was followed by a judgement about concerns regarding applicability for the participant selection, index test, and reference standard domains.

Following the recommendations of the Cochrane Diagnostic Test Accuracy Working Group (Davenport 2014), we developed coding guidelines for each item, both for QUADAS-2 and QUADAS-C (Appendix 2). The overall risk of bias was considered 'low' if at least three domains were scored 'low' and none 'high'. Disagreements amongst review authors regarding ratings were recorded and resolved through discussion, involving a third review author if necessary. We used protocol documents or secondary publications, when available, in the rating process. If publications omitted details pertinent to the assessment of bias risk, we reached out to study investigators/authors for additional information. We did not ask for unpublished protocols.

After completing risk of bias assessment, data extraction, and addition of unpublished information provided by authors to the extraction database, we invited authors to check the correctness of the collected data. For this purpose, we sent the extracted information (including QUADAS-2/QUADAS-C assessment and diagnostic cut-off information) to the individual authors for quality check. We made amendments to the extracted information according to the received feedback. We received confirmation of data extraction accuracy and minor amendments for 17 studies (35%).

Statistical analysis and data synthesis

Our predefined primary analyses of interest concern the sensitivity and specificity of the two self-report questionnaires for each of the two target conditions: 'any anxiety disorder' (yes/no) and 'generalised anxiety disorder' (yes/no) at the cut-off closest to and within a 'core range' around the recommended (primary) cut-off. This approach allowed us to include in the primary analyses studies which had no data available for the recommended cut-off but did have data for a cut-off within a clinically meaningful core range. We specified the following core ranges based on findings from a previous, important systematic review (Plummer 2016), screening of already identified potentially eligible primary studies, and scale characteristics:

- GAD-7: primary cut-off of 10 or higher; core range from 7 to 13;
- GAD-2: primary cut-off of 3 or higher; core range from 2 to 4.

In addition to primary analyses, we investigated the sensitivity and specificity of each cut-off for which data were available for at least three studies in secondary analyses for the two target conditions. Estimates of summary sensitivity and specificity with 95% confidence intervals are considered to be the main outcomes of this meta-analysis. We also provide descriptive P values as an additional measure of the strength of the evidence (Wasserstein 2019).

In all primary and secondary analyses, we calculated study-specific pairs of sensitivity and specificity with 95% confidence intervals using separate 2x2 diagnostic tables for the two target conditions and the two questionnaires. For these four primary analyses, we present forest plots summarising study findings according to settings and study populations. We used the bivariate model to obtain and display summary estimates of sensitivity and specificity with 95% confidence intervals, 95% confidence regions, and 95% prediction regions (Chu 2006; Reitsma 2005). We used univariate models in the case of fitting issues due to boundary fits of the between-study correlation. The models were implemented by generalised linear mixed models and all corresponding statistics

were derived from them as recommended (Deeks 2023). We present summary estimates of sensitivity and specificity both across all settings and for the three settings separately. We used scatter-plots to display these estimates in the receiver operating characteristic (ROC) space; that is, to produce summary ROC (SROC) plots. We calculated positive and negative predictive values with 95% confidence intervals based on the median prevalence of the target conditions.

To simultaneously exploit all available information on all cut-offs and to investigate cut-offs which maximise the Youden index (= sensitivity + specificity - 1) more efficiently in another secondary analysis, we used the multiple thresholds model (Rücker 2020). This is a multilevel random-effects model which uses the data of all studies to model observed sensitivities and specificities in dependence of the given cut-off values, taking between- and within-study covariance into account (Schneider 2017; Steinhauser 2016). Summary estimates of sensitivity and specificity can be derived from the model for every cut-off, followed by the creation of a multiple thresholds summary ROC (mtsROC) curve. The area under the mtsROC curve (AUC) is presented as a summary measure of diagnostic accuracy, and 95% confidence intervals of the AUC are determined by the percentile bootstrap. For this purpose, 500 samples of the data were drawn with replacement at study level to preserve the within-study correlation structure. The cut-off 0 was not included in the modelling as it relates to the deterministic values of 1 and 0 for sensitivity and specificity, which greatly deteriorated the model's goodness-of-fit. The model was implemented with different random intercepts and different random slopes ('DIDS' model) for logit sensitivity and logit specificity. In the case of questionable model fit (i.e. due to boundary fits), a model with common random intercept and different random slopes ('CIDS' model) was used as indicated. While labelling AUC is to some extent arbitrary, we followed the frequently used approach to interpret an AUC above 0.90 as an indicator of very good diagnostic test accuracy, values between 0.80 and 0.89 as good diagnostic test accuracy, between 0.70 and 0.79 as acceptable diagnostic test accuracy, and between 0.60 and 0.69 as low diagnostic test accuracy (De Hond 2022).

In another secondary analysis, we compared the logit sensitivity and logit specificity of the recommended (primary) cut-offs, again using a core range, between questionnaires within the same target condition and included only studies that used a paired design. A bivariate model allowing unequal between-study variance and unstructured between-questionnaire/condition covariance was fitted as recommended (Deeks 2023). Statistics were calculated as described above.

Analyses were conducted using RevMan 2024 and R. The bivariate model and other analyses were implemented in R as recommended by the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks 2023; Freeman 2019). The multiple thresholds model was implemented using the R package diagmeta (Rücker 2020). Exploratory hypothesis tests were performed at a two-sided 5% significance level. Cut-offs maximising the Youden index (sensitivity + specificity - 1) were determined for each analysis.

Investigations of heterogeneity

As we included a diverse set of studies, we expected our findings to be statistically heterogeneous. Therefore, we predefined four key variables or subgroupings in our protocol, as follows.

- Setting: as defined in the pathways section, we categorised settings into (1) non-clinical settings, (2) clinical settings – across conditions, and (3) clinical settings – specific conditions.
- Type of reference standard: we categorised reference standards into (1) semi-structured (SCID), (2) fully structured (CIDI, CIS-R, and Mini-DIPS), and (3) MINI, as applied by Levis 2019.
- Risk of bias: (1) studies with an overall risk of bias categorised as low versus (2) studies with unclear/high risk of bias overall.
- Prevalence of generalised anxiety disorder or any anxiety disorder (continuous variable).

As the potential differences between the settings are of substantial interest for the implementation of the screening, we prioritised this subgrouping variable.

In a first step, we visually inspected forest plots of sensitivity and specificity and summary ROC plots of the primary analysis (e.g. for outliers and using symbols or colours to illustrate the relation to the aforementioned variables). We used meta-regression to compare subgroups or to obtain covariate dependent estimates of logit sensitivity and logit specificity by extending the bivariate model with respective interaction effects, assuming equal between-study variance. The analyses were conducted in concordance with general recommendations from the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks 2023). We assessed between-study heterogeneity by examining the area of the 95% prediction regions, which we compared to the respective statistics of the primary analysis to identify possible sources of heterogeneity. We interpret the area of the 95% prediction region in reference to the total area of the ROC region. Therefore, a value of 0.5 indicates very high heterogeneity. We assessed the statistical significance of the relation of covariates to logit sensitivity and logit specificity using likelihood-ratio tests. Additionally, we assessed between-study heterogeneity by calculating the between-study variance (Tau^2).

Sensitivity analyses

As planned in our protocol (Linde 2022), we did not perform sensitivity analyses.

Assessment of reporting bias

We did not conduct tests to assess possible reporting bias, because consensus is lacking about the most robust approach (Wilson 2015), and there is uncertainty about how to use funnel plots (Van Enst 2014), which have the potential to cause misleading results (Deeks 2005; Leeflang 2008a).

Patient involvement

This review benefitted from the input of a Patient and Public Involvement (PPI) committee. We organised four meetings at approximately equal intervals, with the participation of two practising family physicians, two representatives from patient initiatives, and most of the review authors, from the planning phase of the review project through to the finalisation of the research. In these meetings, we discussed the status of the review, and gathered feedback suggesting possible future directions. Additionally, we explored the findings during a recent session designed to collect opinions on the report's acceptability for practising doctors and the public, as well as to reflect the group's viewpoints. Overall, the report was well received by the PPI committee, including in terms

of its interpretation, understanding, and the potential application of the results.

RESULTS

Results of the search

Our search identified 7085 records (Figure 1). After the removal of duplicates, we screened 4246 records by title and abstract, and excluded 4065 clearly irrelevant records. We retrieved and assessed 180 full-text reports against our eligibility criteria; we were unable to obtain the full-text report for one study. We classified one eligible study identified in our last update search as 'awaiting classification' (Burger 2023), because we were unable to obtain additional information from authors before conducting the final review analyses. For another potentially eligible study, only a protocol published in 2023 was available (Aslan 2023). We classified this study as 'ongoing', although the author did not respond to our emails enquiring about the status of the study.

We excluded 114 articles for a variety of reasons. The most frequent reasons were: no appropriate reference standard, ineligible study design, and no diagnostic accuracy data to extract (either the authors could not be contacted, or they were contacted but could not provide data). Reasons for exclusion are summarised in the Characteristics of excluded studies table and in Figure 1.

Description of the included studies

Overall, we included 48 studies described in 64 full-text reports (total sample size = 19,228 participants) in the qualitative and quantitative synthesis. The first diagnostic accuracy study on GAD-7 and GAD-2 was published in 2007 (Kroenke 2007), followed by a gap until 2012. Eighteen studies were published between 2012 and 2016, and a further 29 between 2017 and 2022. The median analysed sample size of the included studies was 217 participants (range 48 to 2142). Most of the included study reports were original articles published in scientific journals. However, one study was available as a conference abstract only (Seo 2017), and two studies were dissertation theses (Homans 2012; Makulowich 2018). Studies came from 27 different countries, most commonly from Europe (n = 16; 33%), followed by Asia (n = 15; 31%), and North America (n = 7; 15%). Although English was the most frequent language (n = 13; 27%), most of the index tests were in other languages (n = 35; 73%), with a total of 24 different languages.

Summary characteristics of the overall study sample are presented in Appendix 3, Table 1. An overview of the included studies is available in Appendix 3, Table 2. Extensive details of the included studies are provided in the Characteristics of included studies table.

Populations and settings

The median of the reported mean ages of the participants was 41 years (range 23 to 67). Only four studies had a mean participant age of 65 and above (five studies did not report mean age of participants). Twenty studies (42%) did not collect or did not report ethnicity information. Of those studies reporting participants' ethnicity, 12 (43%) studies were conducted amongst people with white ethnicities, and nine studies (32%) amongst Asians.

Seven studies (15%) were performed in non-clinical settings, nine (19%) in clinical settings recruiting participants across conditions, and 32 (67%) studies in participants with specific conditions.

Of the seven studies in non-clinical settings,

- two were conducted in the general population (Christodoulaki 2022; Maric 2022);
- four in other diverse populations: amongst migrant domestic workers (Garabiles 2020), students (Pranckeviciene 2022), socio-economically disadvantaged people (Simning 2012), and elderly people living at home (Wild 2014);
- one study that we categorised as non-clinical actually included a mixed sample of healthy people and hospitalised people (Ahn 2019).

Of the nine studies in clinical settings across conditions,

- five included largely unselected, consecutive participants from primary care practices (Belus 2021; Ivanovs 2018; Kroenke 2007; Chibanda 2016 – a study performed in Zimbabwe in a primary care clinic with very high HIV prevalence; Sidik 2012 – a study amongst women);
- four in diverse other clinical settings: amongst sick-listed workers in an occupational health service organisation (Homans 2012); high utilisers in community clinics (Kujanpää 2014); under- or uninsured people seen in an integrated community care centre (Makulowich 2018); and hospital patients of an internal medicine department of a hospital for traditional Chinese medicine (Zeng 2013).

Of the 32 studies in clinical settings involving participants with specific conditions,

- 14 were performed in neurology departments or outpatient clinics: nine of these studies were in people with epilepsy (Budikayanti 2019; Michaelis 2022; Micoulaud-Franchi 2016; Micoulaud-Franchi 2022; Scott 2019; Seo 2014; Shih 2022; Tong 2016; Zinchuk 2021), four in people with pain conditions (Bisby 2022; Seo 2015; Seo 2017; Veisy 2021), and one in people with multiple sclerosis (Marrie 2018);
- five in people with various types of cancer (Clover 2020; Esser 2018; Lickova 2021; Osorio 2015; Qu 2016);
- five in people with heart diseases: three in people with coronary heart disease (Ahmadi 2019; Grech 2019; Li 2014), one amongst long-term survivors of heart transplantation (Conway 2016), and one in people with heart failure (Fischer 2014);
- three in pregnant women (Nath 2018; Van Heyningen 2018; Zhong 2015);
- two in individuals with substance abuse (Delgadillo 2012; Mughal 2021);
- three in people with other specific diseases: chronic obstructive pulmonary disease (COPD) (Baker 2018), inflammatory bowel disease (Bernstein 2018), or rheumatoid arthritis (Hitchon 2020).

Study designs

As we excluded case-control studies, all included primary studies were cross-sectional in regard to analysis of diagnostic accuracy. However, in six studies (13%), the analysis of cross-sectional diagnostic data was based on baseline data from randomised controlled trials (Bisby 2022; Grech 2019; Homans 2012) or cohort studies (Bernstein 2018; Hitchon 2020; Marrie 2018). Amongst the 42 original cross-sectional studies, 39 (94%) were planned as diagnostic accuracy studies. The investigation of the diagnostic accuracy of GAD-7/GAD-2 was not originally an objective in three

(6%) studies, but unpublished diagnostic accuracy data were provided by authors (Maric 2022; Simning 2012; Zinchuk 2022).

Reference standards

The most commonly used reference standard was MINI (n = 28; 58%), followed by SCID (n = 13; 27%), and other structured instruments (CIDI: n = 4; 8%; CIS-R: n = 2; 4%; and Mini-DIPS: n = 1; 2%). Most of the reference standard interviews were administered by professionals (n = 41; 85%). However, four studies (8%) only stated that they used trained interviewers, and three (6%) did not report the interviewing person. While 38 studies (79%) investigated the target condition generalised anxiety disorder, 26 studied (54%) any anxiety disorder. Sixteen studies (33%) investigated both any anxiety disorder and generalised anxiety disorder. Diagnostic criteria were mostly based on DSM-IV (n = 17; 35%) or DSM-IV +ICD-10 (n = 9; 19%) (Appendix 3, Table 1). Six studies (13%) used DSM-V as diagnostic criteria.

Post hoc cut-off derivation

Of the 37 studies (77%) which used post hoc cut-off derivation, 13 (35%) applied the Youden index, five (14%) used likelihood ratio, two (5%) the minimum Euclidean distance, one (3%) the product of sensitivity and specificity (the Liu method) (Liu 2012), and 16 (43%) did not clearly mention the criterion.

Methodological quality of included studies

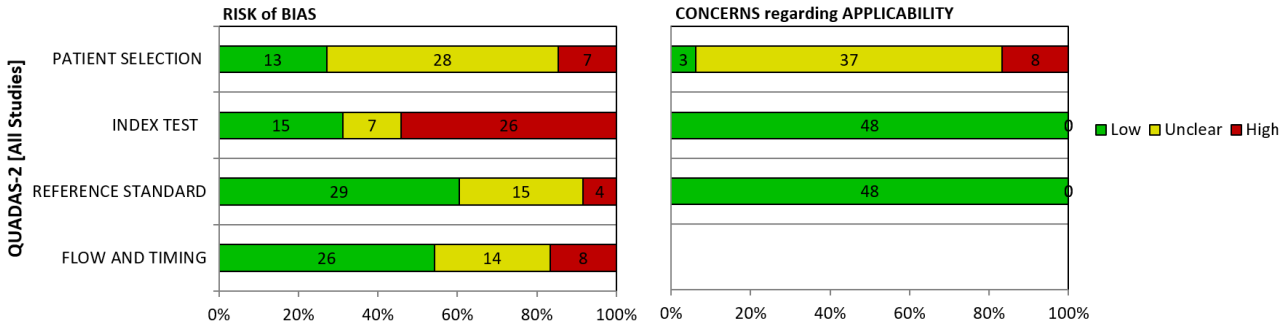
In the following, we first describe the results of our assessment of methodological quality using the QUADAS-2 tool for all 48 included studies, and then separately for the four combinations of two index tests and two target conditions. We then describe the QUADAS-C assessments of the 25 studies in which a direct comparison of GAD-7 and GAD-2 was available (see Table 3 in Appendix 3 for the ratings of individual studies). Details on how we reached these judgements are given in the Characteristics of included studies.

QUADAS-2

Methodological quality of all studies

Of the 48 included studies, we assessed eight (17%) as having low overall risk of bias (out of the four domains, at least three were low risk of bias and none were high risk of bias) (Belus 2021; Bernstein 2018; Budikayanti 2019; Hitchon 2020; Ivanovs 2018; Kroenke 2007; Marrie 2018; Scott 2019). The highest risk of bias concerns were in the index test domain (Figure 2; Table 1 in Appendix 3). Only four studies had a high risk of bias in the reference standard domain. In the patient selection domain, we judged 13 studies (27%) to have a low risk of bias. We assessed 15 studies (31%) as low risk in the index test domain. Most of our high risk of bias assessments in the index test domain stemmed from post hoc selection of cut-offs or concerns related to blinding. In the reference standard domain, we considered 29 studies (60%) to have a low risk of bias. Reasons for unclear or high risk of bias in the reference standard domain were related to blinding concerns. In the flow and timing domain, we considered 26 studies (54%) to have a low risk of bias. In 14 studies (29%), we could not establish whether the interval between the index test(s) and the reference standard was less than two weeks (assessed as unclear risk of bias). We rated the remaining eight studies (17%) as high risk in this domain because they either included less than 90% of the recruited patients in the analysis or did not disclose this information.

Figure 2. Risk of bias and applicability concerns for all studies (n = 48): review authors' judgements for QUADAS-2 domains



None of the studies entailed concerns regarding applicability in the index test or reference standard domains. However, we had concerns about whether the included patients matched the review question. In the patient selection domain, we assessed 37 (77%) studies as having unclear concerns, as we could not determine from the study methods whether: (a) participants constituted a representative sample with regard to the setting/screening pathway, and/or (b) anxiety diagnoses in the eligible participants were pre-existing/pre-diagnosed before study procedure. In eight (17%) studies, we had high concerns for applicability as participants with pre-existing anxiety diagnoses were included, although the proportion of such participants was very low. We categorised three (6%) studies as low concern (Figure 2, Table 3 in Appendix 3).

Methodological quality of the studies pertinent to the four review questions

Risk of bias (RoB) assessments of the different studies answering the four review questions were similar. Summary graphs showing our QUADAS-2 risk of bias assessments for the four review questions are presented in Figure 12, Figure 13, Figure 14, and Figure 15 in Appendix 4.

GAD-7 for generalised anxiety disorder (n = 35): the numbers and proportions of studies with low risk of bias in the patient selection, index test, reference standard, and flow and timing domains were 11 (31%), 14 (40%), 23 (66%), and 18 (51%), respectively. We considered the overall risk of bias as low in eight (22%) studies.

GAD-7 for any anxiety disorder (n = 22): the numbers and proportions of studies with low risk of bias in the patient selection, index test, reference standard, and flow and timing domains were five (23%), nine (41%), 14 (64%), and 15 (68%), respectively. We considered the overall risk of bias as low in six (27%) studies.

GAD-2 for generalised anxiety disorder (n = 24): the numbers and proportions of studies with low risk of bias in the patient selection, index test, reference standard, and flow and timing domains were eight (33%), 10 (42%), 17 (71%), and 13 (54%), respectively. We considered the overall risk of bias as low in seven (29%) studies.

GAD-2 for any anxiety disorder (n = 19): the numbers and proportions of studies with low risk of bias in the patient selection, index test, reference standard, and flow and timing domains were five (26%), nine (47%), 14 (74%), and 13 (68%), respectively. We considered the overall risk of bias as low in six (32%) studies.

QUADAS-C

Diagnostic data from 25 studies were available for the comparison of the diagnostic accuracy of GAD-7 and GAD-2 in detecting either any anxiety disorder or generalised anxiety disorder. We judged three studies (12%) as having a low risk of bias in all four domains. We assessed 16 studies (64% of the studies with a direct comparison) as low risk of bias in the patient selection domain, 13 studies (52%) as low risk in the index test domain, nine studies (36%) as low risk in the reference standard domain, and nine studies (36%) as low risk of bias in the flow and timing domain. Three studies were at high risk of bias in all four domains (Bisby 2022; Kujanpää 2014; Seo 2017) (Figure 16 in Appendix 4).

When the two target conditions were evaluated separately, in the generalised anxiety disorder group (n = 23), the number of studies with low risk of bias in the patient selection, index test, reference standard, and flow and timing domains were 15 (65%), 12 (52%), seven (30%), and nine (39%), respectively. In the 'any anxiety disorder' group (n = 15), the number of studies with low risk of bias in the patient selection, index test, reference standard, and flow and timing domains were 10 (67%), six (40%), four (27%), and four (27%), respectively (Figure 17 and Figure 18 in Appendix 4).

Findings

Below we present the findings addressing our four primary questions and our secondary objectives.

Accuracy of the GAD-7 questionnaire for identifying generalised anxiety disorder

See Summary of findings 1.

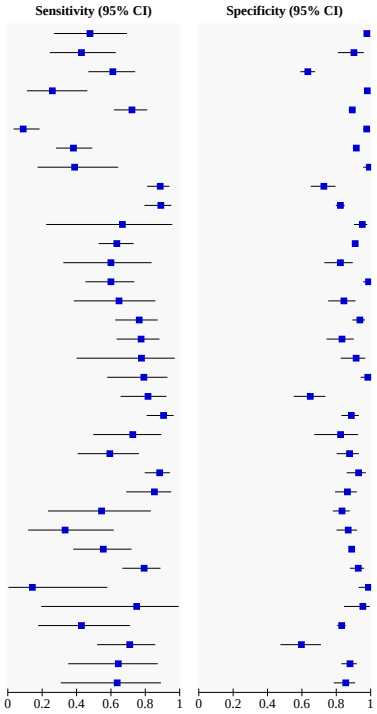
Primary analysis using the recommended cut-offs/core ranges

Thirty-six studies investigated GAD-7 in the assessment of generalised anxiety disorder. However, we did not include one study, Mughal 2021, in the primary analysis because it reported cut-offs outside the core range.

Of the 35 studies included in the primary analysis using the bivariate model, diagnostic accuracy data were available for the recommended cut-off of 10 or higher from 28 studies, a cut-off of 8 or higher from four studies (Sidik 2012; Micoulaud-Franchi 2022; Seo 2015; Seo 2017), and a cut-off of 9 or higher from three studies (Tong 2016; Seo 2014; Lickova 2021).The overall median prevalence was 12% (range 1 to 45) for generalised anxiety disorder. The median prevalence for generalised anxiety disorder in the subgroups was as follows: 8% (range 2 to 28) in non-clinical settings, 8% (range 4 to 40) in clinical settings across conditions, and 16% (range 1 to 45) in clinical settings in people with specific conditions. Figure 3 displays the sensitivity and specificity estimates and their respective 95% CIs for the individual studies. Point estimates from individual studies for sensitivity ranged widely from 0.09 to 0.91, those for specificity from 0.60 to 0.99. Also, estimates of sensitivity were less precise than those for specificity, as reflected by the larger 95% confidence intervals. This can be attributed to the statistical property of a binomially distributed random variable having larger confidence intervals around the centre of the value range, but also to the smaller numbers of participants with generalised anxiety disorder (true positive (TP) + false negative (FN)) across the studies (median: 31, range: 4 to 107) compared to those without generalised anxiety disorder (false positive (FP) + true negative (TN)) (median: 144, range: 40 to 1900).

Figure 3. Forest plot of GAD-7 for generalised anxiety disorder (GAD) TP = true positive; FP = false positive; FN = false negative; TN = true negative; Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions; CI = confidence interval

Study	TP	FP	FN	TN	Setting	Population	Sensitivity (95% CI)	Specificity (95% CI)
Maric 2022	11	26	12	1154	I	A general population	0.48 [0.27, 0.69]	0.98 [0.97, 0.99]
Garabiles 2020	12	7	16	65	I	B migrant domestic workers	0.43 [0.24, 0.63]	0.90 [0.81, 0.96]
Pranckeviciene 2022	33	185	21	321	I	C students	0.61 [0.47, 0.74]	0.63 [0.59, 0.68]
Wild 2014	7	8	20	403	I	E elderly living at home	0.26 [0.11, 0.46]	0.98 [0.96, 0.99]
Ahn 2019	65	114	25	953	I	F mixed healthy/patients	0.72 [0.62, 0.81]	0.89 [0.87, 0.91]
Belus 2021	6	10	61	425	II	G unselected primary care	0.09 [0.03, 0.18]	0.98 [0.96, 0.99]
Ivanovs 2018	34	115	55	1263	II	G unselected primary care	0.38 [0.28, 0.49]	0.92 [0.90, 0.93]
Sidik 2012	7	1	11	126	II	G unselected primary care	0.39 [0.17, 0.64]	0.99 [0.96, 1.00]
Chibanda 2016	94	43	12	115	II	G unselected primary care	0.89 [0.81, 0.94]	0.73 [0.65, 0.80]
Kroenke 2007	65	157	8	735	II	G unselected primary care	0.89 [0.80, 0.95]	0.82 [0.80, 0.85]
Kujanpää 2014	4	7	2	137	II	I high utilisers primary care	0.67 [0.22, 0.96]	0.95 [0.90, 0.98]
Zeng 2013	59	171	34	1729	II	K mixed inpatients	0.63 [0.53, 0.73]	0.91 [0.90, 0.92]
Scott 2019	9	16	6	75	III	L neurology: epilepsy	0.60 [0.32, 0.84]	0.82 [0.73, 0.90]
Tong 2016	30	2	20	161	III	L neurology: epilepsy	0.60 [0.45, 0.74]	0.99 [0.96, 1.00]
Shih 2022	11	14	6	76	III	L neurology: epilepsy	0.65 [0.38, 0.86]	0.84 [0.75, 0.91]
Seo 2014	39	12	12	180	III	L neurology: epilepsy	0.76 [0.63, 0.87]	0.94 [0.89, 0.97]
Micoulaud-Franchi 2016	38	16	11	80	III	L neurology: epilepsy	0.78 [0.63, 0.88]	0.83 [0.74, 0.90]
Michaelis 2022	7	6	2	65	III	L neurology: epilepsy	0.78 [0.40, 0.97]	0.92 [0.83, 0.97]
Budikayanti 2019	19	2	5	120	III	L neurology: epilepsy	0.79 [0.58, 0.93]	0.98 [0.94, 1.00]
Micoulaud-Franchi 2022	31	39	7	72	III	L neurology: epilepsy	0.82 [0.66, 0.92]	0.65 [0.55, 0.74]
Zinchuk 2021	58	19	6	150	III	L neurology: epilepsy	0.91 [0.81, 0.96]	0.89 [0.83, 0.93]
Bisby 2022	16	7	6	33	III	M chronic pain	0.73 [0.50, 0.89]	0.82 [0.67, 0.93]
Seo 2015	19	14	13	100	III	N neurology: mgraine	0.59 [0.41, 0.76]	0.88 [0.80, 0.93]
Veisy 2021	76	7	10	93	III	N neurology: mgraine	0.88 [0.80, 0.94]	0.93 [0.86, 0.97]
Seo 2017	29	17	5	109	III	O tension-type headache	0.85 [0.69, 0.95]	0.87 [0.79, 0.92]
Marrie 2018	6	40	5	200	III	P multiple sclerosis	0.55 [0.23, 0.83]	0.83 [0.78, 0.88]
Clover 2020	5	18	10	120	III	Q various cancers	0.33 [0.12, 0.62]	0.87 [0.80, 0.92]
Esser 2018	20	232	16	1872	III	Q various cancers	0.56 [0.38, 0.72]	0.89 [0.88, 0.90]
Lickova 2021	46	13	12	167	III	Q various cancers	0.79 [0.67, 0.89]	0.93 [0.88, 0.96]
Grech 2019	1	1	6	77	III	S coronary heart disease	0.14 [0.00, 0.58]	0.99 [0.93, 1.00]
Conway 2016	3	2	1	42	III	T heart transplantation survivors	0.75 [0.19, 0.99]	0.95 [0.85, 0.99]
Zhong 2015	6	157	8	775	III	V pregnancy	0.43 [0.18, 0.71]	0.83 [0.81, 0.86]
Delgadillo 2012	22	29	9	43	III	W substance abuse	0.71 [0.52, 0.86]	0.60 [0.47, 0.71]
Bernstein 2018	9	27	5	198	III	Y inflammatory bowel disease	0.64 [0.35, 0.87]	0.88 [0.83, 0.92]
Hitchon 2020	7	20	4	118	III	Z rheumatoid arthritis	0.64 [0.31, 0.89]	0.86 [0.79, 0.91]



Across all studies, we estimated a summary sensitivity of 0.64 (95% CI 0.56 to 0.72) and a summary specificity of 0.91 (95% CI 0.87 to 0.93). The area of the 95% prediction region was 0.30, indicating pronounced heterogeneity. The summary point, along with the 95% prediction region, is presented in Appendix 5.

Investigations of sources of heterogeneity

There was no statistical evidence for differences in test accuracy according to the three settings (P > 0.05). For five studies performed in non-clinical settings, the summary estimates were 0.50 (95% CI 0.30 to 0.70) for sensitivity and 0.93 (95% CI 0.84 to 0.97) for specificity. For seven studies performed in clinical settings across conditions, the summary estimates for sensitivity and specificity were 0.57 (95% CI 0.39 to 0.73) and 0.93 (95% CI 0.86 to 0.97), respectively. Sensitivity was somewhat higher (0.70, 95% CI 0.60 to 0.78) and specificity slightly lower (0.89, 95% CI 0.85 to 0.93) for the 23 studies in clinical settings in people with specific conditions. Heterogeneity remained high within the three settings (Appendix 5).

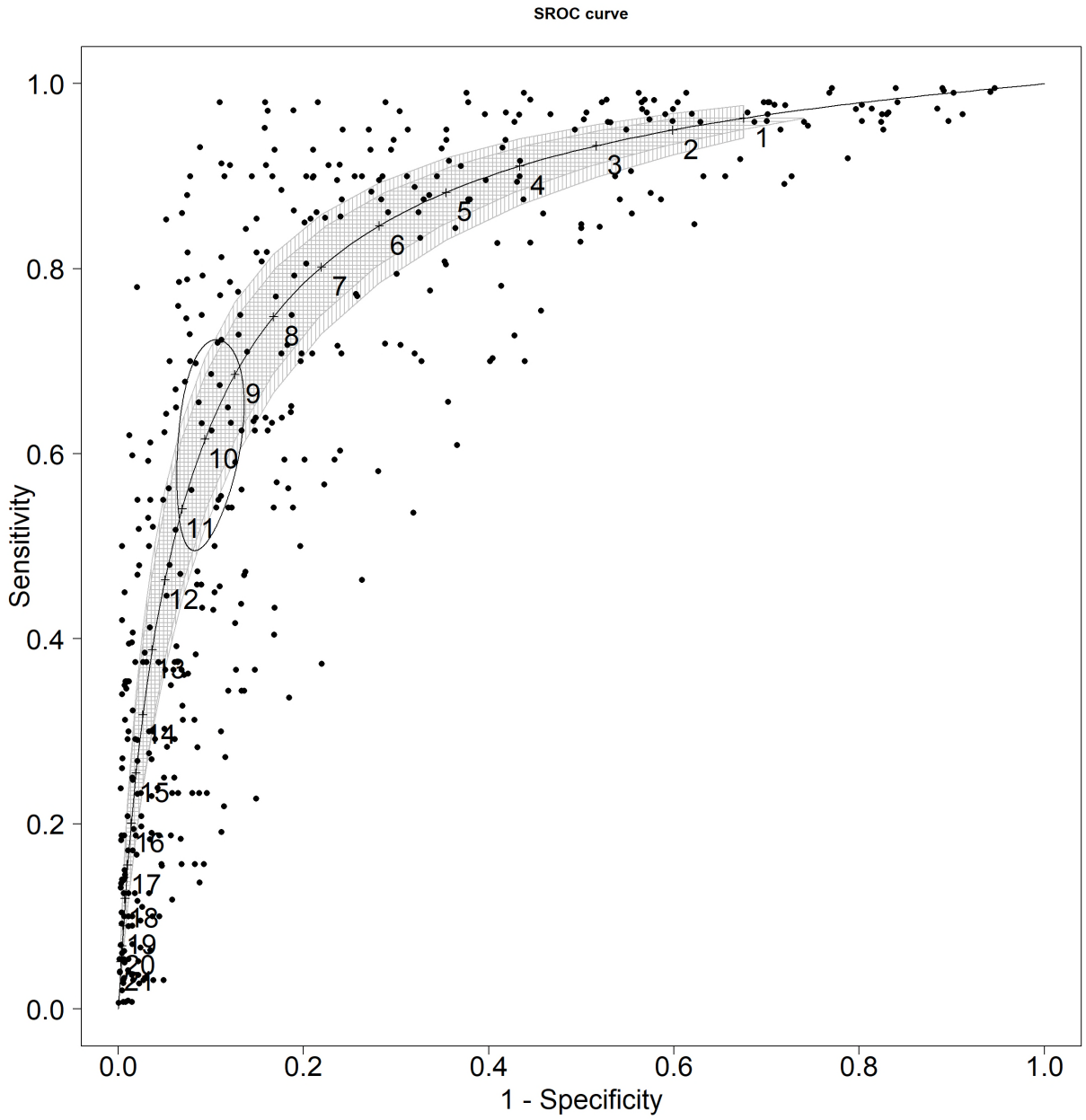
There were no statistically significant subgroup differences for risk of bias and type of reference standard (P > 0.05). However, there was a significant association between prevalence and diagnostic accuracy (overall P = 0.01). Increasing prevalence of generalised anxiety disorder was significantly associated with higher sensitivity (P = 0.003), but there was no statistically significant association with specificity (P = 0.27; for details see Appendix 5).

Secondary analyses

Summary accuracy estimates based on the bivariate model for each of the 21 GAD-7 cut-offs assessing generalised anxiety disorder can be found in Appendix 5. At the cut-off recommended by the scale authors (≥ 10), summary sensitivity was 0.62 (95% CI 0.52 to 0.71) and summary specificity 0.90 (95% CI 0.86 to 0.93). Using bivariate models, the highest Youden index (i.e. 0.62) was found at a cut-off of 7 or higher, with a sensitivity of 0.84 (95% CI 0.77 to 0.90) and a specificity of 0.78 (95% CI 0.72 to 0.83).

Diagnostic accuracy information from 36 studies comprising a total of 447 2x2-tables was available for analyses using the multiple thresholds model. The area under the curve (AUC) was determined as 0.86 (95% CI 0.84 to 0.88; Figure 4). Based on the multiple thresholds model, summary sensitivity was 0.62 (95% CI 0.57 to 0.70) and summary specificity 0.91 (95% CI 0.86 to 0.93) for the cut-off recommended by the scale authors (≥ 10). The point estimate for specificity is identical to that found in the primary analysis using the core range, while it is slightly lower for sensitivity. The Youden index was maximal (0.58) at the cut-off of 7 or higher (Appendix 5), with a summary sensitivity of 0.80 (95% CI 0.73 to 0.86) and summary specificity of 0.78 (95% CI 0.72 to 0.83).

Figure 4. Multiple thresholds summary receiver operating characteristic (mtsROC) curve based on all available data for all cut-offs for GAD-7 for detecting generalised anxiety disorder (GAD) The area under the curve (AUC) is 0.86 (95% CI 0.84 to 0.88). The vertical hatching along the curve corresponds to pointwise 95% confidence intervals for sensitivity, given specificity; horizontal hatching corresponds to pointwise confidence intervals for specificity, given sensitivity. The numbers indicate the cut-off points. The 95% confidence region for the recommended cut-off ≥ 10 is represented by the ellipse.



Accuracy of the GAD-7 questionnaire for identifying any anxiety disorder

See Summary of findings 2.

Primary analyses using the recommended cut-offs/core ranges

Twenty-two studies investigated GAD-7 in the evaluation of any anxiety disorder. All studies could be included in the primary analysis.

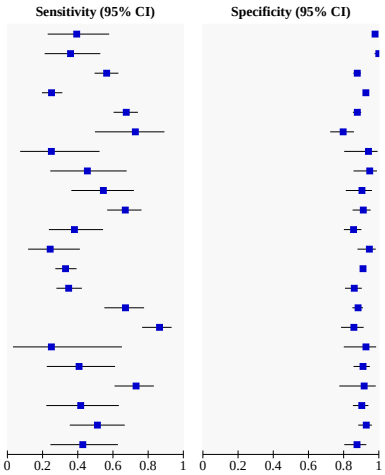
The median prevalence was 19% (range 3 to 65) for any anxiety disorder. The median prevalence for any anxiety disorder in the subgroups was as follows: 19% (range 3 to 21) in non-clinical settings, 18% (range 13 to 32) in clinical settings across conditions, and 19% (range 11 to 65) in clinical settings in people with specific conditions.

Diagnostic accuracy data were available for the recommended cut-off (of 10 or higher) from 18 studies, a cut-off of 8 or higher from

one study (Ahn 2019), and a cut-off of 9 or higher from three studies (Makulowich 2018; Osorio 2015; Clover 2020). Figure 5 displays the sensitivity and specificity estimates and their respective 95% CIs for the individual studies. Point estimates from individual studies for sensitivity ranged widely from 0.25 to 0.86, those for specificity from 0.80 to 1.00. Also, estimates of sensitivity were less precise than those for specificity, as reflected by the larger 95% confidence intervals. The numbers of participants with any anxiety disorder (TP +FN) across the studies (median: 41, range: 8 to 249) were smaller compared to those without any anxiety disorder (FP+TN) (median: 159, range: 34 to 1893).

Figure 5. Forest plot of GAD-7 for detecting any anxiety disorder (AAD)
TP = true positive; FP = false positive; FN = false negative; TN = true negative; Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions; CI = confidence interval

Study	TP	FP	FN	TN	Setting	Population	Sensitivity (95% CI)	Specificity (95% CI)
Maric 2022	13	25	20	1145	I	A general population	0.39 [0.23, 0.58]	0.98 [0.97, 0.99]
Simning 2012	14	0	25	151	I	D poor elderly	0.36 [0.21, 0.53]	1.00 [0.98, 1.00]
Ahn 2019	123	115	95	824	I	F mixed healthy/patients	0.56 [0.50, 0.63]	0.88 [0.85, 0.90]
Ivanovs 2018	58	91	173	1145	II	G unselected primary care	0.25 [0.20, 0.31]	0.93 [0.91, 0.94]
Kroenke 2007	127	95	61	682	II	G unselected primary care	0.68 [0.60, 0.74]	0.88 [0.85, 0.90]
Homans 2012	16	30	6	118	II	H sick-listed workers	0.73 [0.50, 0.89]	0.80 [0.72, 0.86]
Makulowich 2018	4	2	12	32	II	J under-/uninsured adults	0.25 [0.07, 0.52]	0.94 [0.80, 0.99]
Michaelis 2022	10	3	12	55	III	L neurology: epilepsy	0.45 [0.24, 0.68]	0.95 [0.86, 0.99]
Scott 2019	18	7	15	66	III	L neurology: epilepsy	0.55 [0.36, 0.72]	0.90 [0.81, 0.96]
Zinchuk 2021	65	12	32	124	III	L neurology: epilepsy	0.67 [0.57, 0.76]	0.91 [0.85, 0.95]
Marrie 2018	16	30	26	179	III	P multiple sclerosis	0.38 [0.24, 0.54]	0.86 [0.80, 0.90]
Clover 2020	9	5	28	88	III	Q various cancers	0.24 [0.12, 0.41]	0.95 [0.88, 0.98]
Esser 2018	82	170	166	1722	III	Q various cancers	0.33 [0.27, 0.39]	0.91 [0.90, 0.92]
Osorio 2015	62	31	116	191	III	Q various cancers	0.35 [0.28, 0.42]	0.86 [0.81, 0.90]
Qu 2016	49	56	24	419	III	R leukaemia	0.67 [0.55, 0.78]	0.88 [0.85, 0.91]
Li 2014	64	18	10	109	III	S coronary heart disease	0.86 [0.77, 0.93]	0.86 [0.79, 0.91]
Conway 2016	2	3	6	38	III	T heart transplantation survivors	0.25 [0.03, 0.65]	0.93 [0.80, 0.98]
Fischer 2014	11	15	16	152	III	U heart failure	0.41 [0.22, 0.61]	0.91 [0.86, 0.95]
Delgadillo 2012	49	3	18	33	III	W substance abuse	0.73 [0.61, 0.83]	0.92 [0.78, 0.98]
Baker 2018	10	19	14	177	III	X COPD	0.42 [0.22, 0.63]	0.90 [0.85, 0.94]
Bernstein 2018	22	14	21	182	III	Y inflammatory bowel disease	0.51 [0.35, 0.67]	0.93 [0.88, 0.96]
Hitchon 2020	12	15	16	106	III	Z rheumatoid arthritis	0.43 [0.24, 0.63]	0.88 [0.80, 0.93]



Across all studies, we estimated a summary sensitivity of 0.48 (95% CI 0.40 to 0.57) and a summary specificity of 0.91 (95% CI 0.89 to 0.93). The area of the 95% prediction region was 0.16, indicating less heterogeneity compared to the GAD-7 studies with generalised anxiety disorder as the target condition. The summary point, along with the 95% prediction region, is presented in Appendix 6.

Investigations of sources of heterogeneity

There was no statistical evidence for differences in sensitivity with regard to settings (P = 0.86), but specificity differed significantly between the subgroups (P = 0.019). For three studies performed in non-clinical settings, the summary estimates were 0.43 (95% CI 0.24 to 0.64) for sensitivity and 0.96 (95% CI 0.93 to 0.98) for specificity. For four studies performed in clinical settings across conditions, the summary estimates for sensitivity and specificity were 0.48 (95% CI 0.29 to 0.67) and 0.89 (95% CI 0.83 to 0.93), respectively. For the 15 studies in clinical settings in people with specific conditions, sensitivity was 0.50 (95% CI 0.40 to 0.60) and specificity was 0.90 (95% CI 0.87 to 0.93). Heterogeneity remained high within the three settings (Appendix 6).

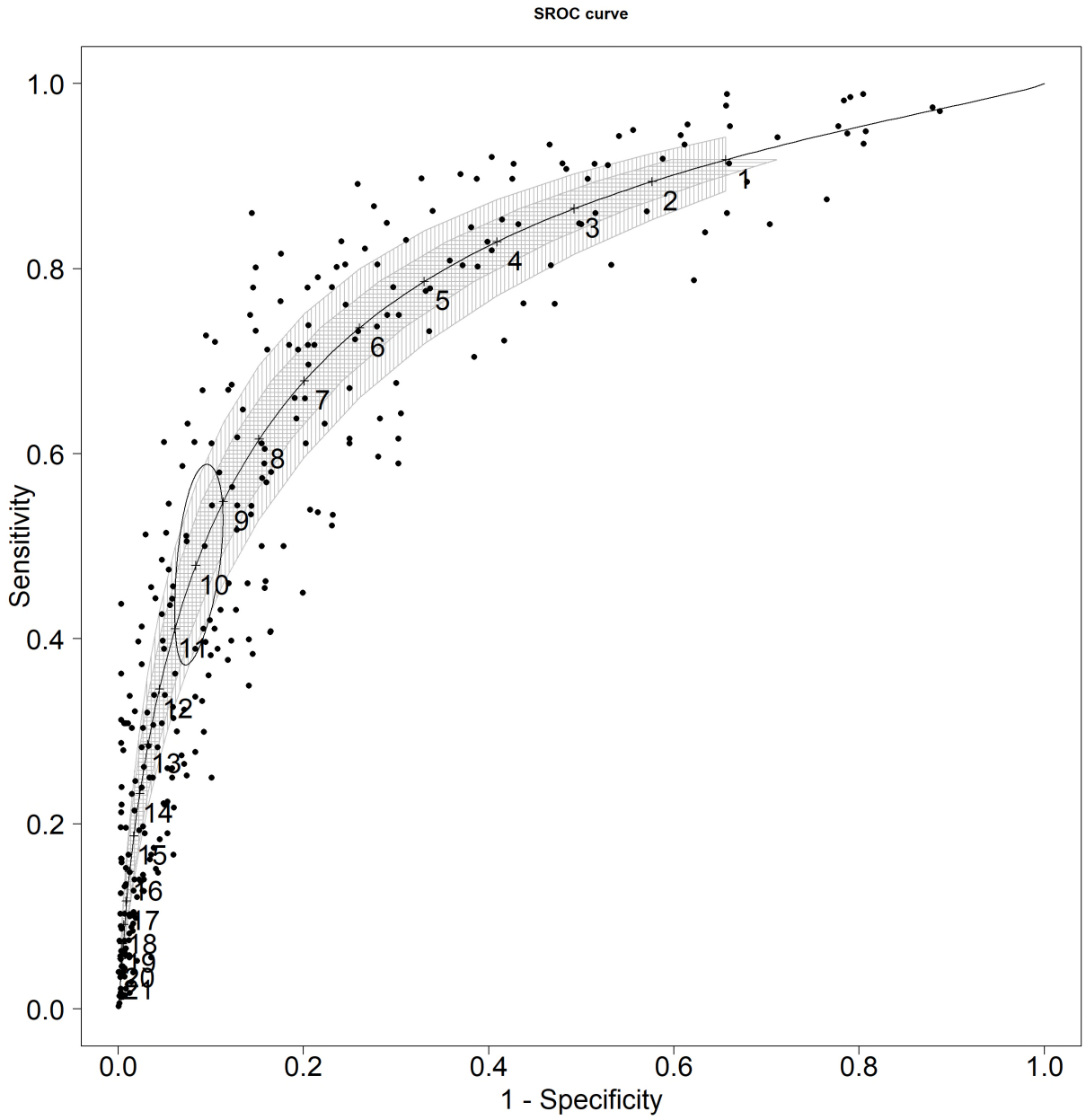
There were no statistically significant subgroup differences for risk of bias, type of reference standard, or prevalence (P > 0.05) (Appendix 6).

Secondary analyses

Summary accuracy estimates based on the bivariate model for each of the 21 GAD-7 cut-offs assessing any anxiety disorder can be found in Appendix 6. At the cut-off recommended by the scale authors (≥ 10; 18 studies), summary sensitivity was 0.52 (95% CI 0.42 to 0.61) and summary specificity 0.92 (95% CI 0.89 to 0.94). Using bivariate models, the highest Youden index (i.e. 0.49) was found at a cut-off of 5 or higher with a sensitivity of 0.77 (95% CI 0.71 to 0.82) and a specificity of 0.72 (95% CI 0.67 to 0.76).

Diagnostic accuracy information from 22 studies comprising a total of 308 2x2-tables was available for analyses using the multiple thresholds model. The area under the curve (AUC) was determined as 0.80 (95% CI 0.78 to 0.82; Figure 6). Based on the multiple thresholds model, summary sensitivity was 0.48 (95% CI 0.39 to 0.57) and summary specificity 0.92 (95% CI 0.89 to 0.93) for the cut-off recommended by the scale authors (≥ 10). The point estimate for specificity is identical to that found in the primary analysis using the core range, while it is slightly lower for sensitivity. The Youden index was maximal (0.48) at the cut-off of 7 or higher (Appendix 6), with a summary sensitivity of 0.68 (95% CI 0.60 to 0.75) and summary specificity of 0.80 (95% CI 0.76 to 0.83).

Figure 6. Multiple thresholds summary receiver operating characteristic (mtsROC) curve based on all available data on all available data for all cut-offs for GAD-7 for detecting any anxiety disorder (AAD) The area under the curve (AUC) is 0.80 (95% CI 0.78 to 0.82). The vertical hatching along the curve corresponds to pointwise 95% confidence intervals for sensitivity, given specificity; horizontal hatching corresponds to pointwise confidence intervals for specificity, given sensitivity. The numbers indicate the cut-off points. The 95% confidence region for the recommended cut-off ≥ 10 is represented by the ellipse.



Accuracy of the GAD-2 questionnaire for identifying generalised anxiety disorder

See Summary of findings 3.

Primary analyses using the recommended cut-offs/core ranges

Twenty-four studies investigated GAD-2 in the evaluation of generalised anxiety disorder. All studies could be included in the primary analysis.

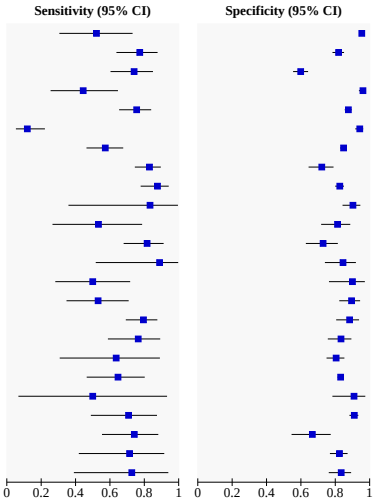
The median prevalence was 9% (range 2 to 45) for generalised anxiety disorder. The median prevalence for generalised anxiety disorder in the subgroups was as follows: 8% (range 2 to 10) in non-clinical settings, 8% (range 4 to 40) in clinical settings across conditions, and 11% (range 2 to 45) in clinical settings in people with specific conditions.

Diagnostic test accuracy data for the recommended cut-off of 3 or higher were available for all 24 studies. Figure 7 displays

the sensitivity and specificity estimates and their respective 95% CIs for the individual studies. Point estimates from individual studies for sensitivity ranged widely from 0.12 to 0.89, those for specificity from 0.60 to 0.96. Also, estimates of sensitivity were less precise than those for specificity, as reflected by the larger 95% confidence intervals. The numbers of participants with generalised anxiety disorder (TP+FN) across the studies (median: 32, range: 4 to 106) were smaller compared to those without generalised anxiety disorder (FP+TN) (median: 193, range: 40 to 1896).

Figure 7. Forest plot of GAD-2 for detecting generalised anxiety disorder (GAD)
TP = true positive; FP = false positive; FN = false negative; TN = true negative; Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions; CI = confidence interval

Study	TP	FP	FN	TN	Setting	Population	Sensitivity (95% CI)	Specificity (95% CI)
Maric 2022	12	54	11	1126	I	A general population	0.52 [0.31, 0.73]	0.95 [0.94, 0.97]
Christodoulaki 2022	41	96	12	435	I	A general population	0.77 [0.64, 0.88]	0.82 [0.78, 0.85]
Pranckeviciene 2022	40	203	14	303	I	C students	0.74 [0.60, 0.85]	0.60 [0.55, 0.64]
Wild 2014	12	16	15	395	I	E elderly living at home	0.44 [0.25, 0.65]	0.96 [0.94, 0.98]
Ahn 2019	68	132	22	935	I	F mixed healthy/patients	0.76 [0.65, 0.84]	0.88 [0.86, 0.90]
Belus 2021	8	25	59	410	II	G unselected primary care	0.12 [0.05, 0.22]	0.94 [0.92, 0.96]
Ivanovs 2018	51	209	38	1169	II	G unselected primary care	0.57 [0.46, 0.68]	0.85 [0.83, 0.87]
Chibanda 2016	88	44	18	114	II	G unselected primary care	0.83 [0.74, 0.90]	0.72 [0.64, 0.79]
Kroenke 2007	64	155	9	737	II	G unselected primary care	0.88 [0.78, 0.94]	0.83 [0.80, 0.85]
Kujanpää 2014	5	14	1	130	II	I high utilisers primary care	0.83 [0.36, 1.00]	0.90 [0.84, 0.95]
Scott 2019	8	17	7	74	III	L neurology: epilepsy	0.53 [0.27, 0.79]	0.81 [0.72, 0.89]
Micoulaud-Franchi 2016	40	26	9	70	III	L neurology: epilepsy	0.82 [0.68, 0.91]	0.73 [0.63, 0.81]
Michaelis 2022	8	11	1	60	III	L neurology: epilepsy	0.89 [0.52, 1.00]	0.85 [0.74, 0.92]
Bisby 2022	11	4	11	36	III	M chronic pain	0.50 [0.28, 0.72]	0.90 [0.76, 0.97]
Seo 2015	17	12	15	102	III	N neurology: mgraine	0.53 [0.35, 0.71]	0.89 [0.82, 0.94]
Veisy 2021	66	12	17	91	III	N neurology: mgraine	0.80 [0.69, 0.88]	0.88 [0.81, 0.94]
Seo 2017	26	21	8	105	III	O tension-type headache	0.76 [0.59, 0.89]	0.83 [0.76, 0.89]
Marrie 2018	7	47	4	194	III	P multiple sclerosis	0.64 [0.31, 0.89]	0.80 [0.75, 0.85]
Esser 2018	22	320	12	1576	III	Q various cancers	0.65 [0.46, 0.80]	0.83 [0.81, 0.85]
Conway 2016	2	4	2	40	III	T heart transplantation survivors	0.50 [0.07, 0.93]	0.91 [0.78, 0.97]
Nath 2018	17	45	7	459	III	V pregnancy	0.71 [0.49, 0.87]	0.91 [0.88, 0.93]
Delgadillo 2012	23	24	8	48	III	W substance abuse	0.74 [0.55, 0.88]	0.67 [0.55, 0.77]
Bernstein 2018	10	40	4	188	III	Y inflammatory bowel disease	0.71 [0.42, 0.92]	0.82 [0.77, 0.87]
Hitchon 2020	8	23	3	116	III	Z rheumatoid arthritis	0.73 [0.39, 0.94]	0.83 [0.76, 0.89]



Across all studies, we estimated a summary sensitivity of 0.68 (95% CI 0.59 to 0.75) and a summary specificity of 0.86 (95% CI 0.82 to 0.89). The area of the 95% prediction region was 0.19, indicating less heterogeneity compared to the GAD-7 studies with generalised anxiety disorder as the target condition. The summary point, along with the 95% prediction region, is presented in Appendix 7.

Investigations of sources of heterogeneity

There was no statistical difference in test accuracy according to settings (P > 0.05). For five studies performed in non-clinical settings, the summary estimates were 0.66 (95% CI 0.49 to 0.80) for sensitivity and 0.88 (95% CI 0.81 to 0.93) for specificity. For five studies performed in clinical settings across conditions, the summary estimates for sensitivity and specificity were 0.63 (95% CI 0.45 to 0.78) and 0.86 (95% CI 0.78 to 0.92), respectively. For the 14 studies in clinical settings in people with specific conditions, sensitivity was 0.70 (95% CI 0.59 to 0.79) and specificity was 0.84 (95% CI 0.79 to 0.89) (Appendix 7).

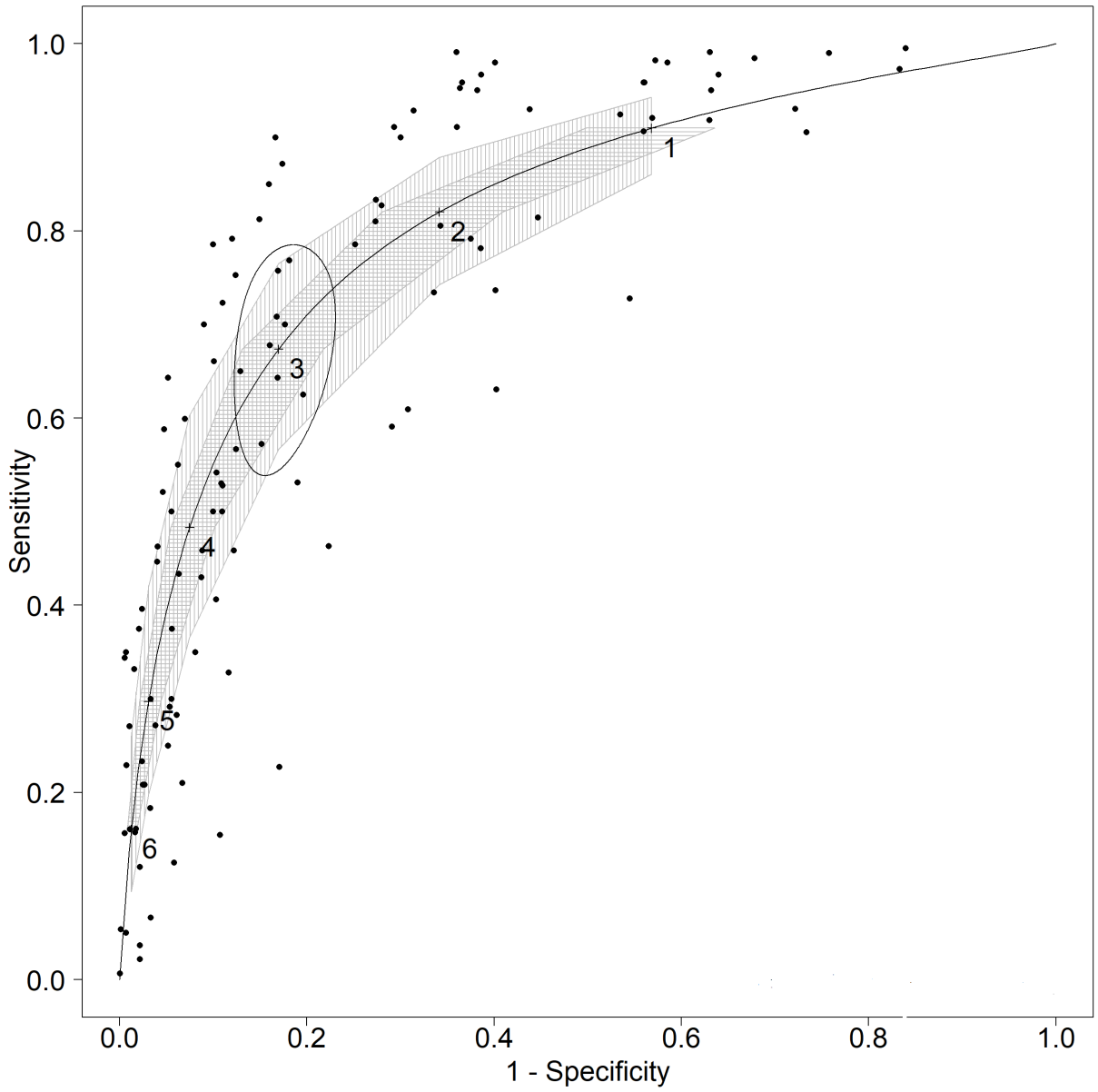
There were no statistically significant subgroup differences for risk of bias or prevalence (P > 0.05). However, type of reference standard was significantly associated with specificity, with studies using SCID and MINI interviews reporting higher values than studies using other structured interviews (overall P = 0.03, for sensitivity P = 0.14, for specificity P = 0.04) (Appendix 7).

Secondary analyses

Summary accuracy estimates based on the bivariate model for each of the six GAD-2 cut-offs can be found in Appendix 7. At the cut-off recommended by the scale authors (≥ 3), summary sensitivity was 0.67 (95% CI 0.59 to 0.75) and summary specificity 0.86 (95% CI 0.82 to 0.89). Using bivariate models, the highest Youden index (i.e. 0.56) was found at a cut-off of 2 or higher, with a sensitivity of 0.89 (95% CI 0.83 to 0.93) and a specificity of 0.67 (95% CI 0.60 to 0.72).

Diagnostic accuracy information from a total of 115 2x2-tables was available for analyses using the multiple thresholds model. The area under the curve (AUC) was determined as 0.82 (95% CI 0.81 to 0.86; Figure 8). Based on the multiple thresholds model, summary sensitivity was 0.67 (95% CI 0.57 to 0.77) and summary specificity 0.83 (95% CI 0.78 to 0.87) for the cut-off recommended by the scale authors (≥ 3); the point estimates were similar to that found in the primary analysis using the core range. The Youden index was maximal (0.50) at the cut-off of 3 or higher (Appendix 7).

Figure 8. Multiple thresholds summary receiver operating characteristic (mtsROC) curve based on all available data or GAD-2 for detecting generalised anxiety disorder (GAD) The area under the curve (AUC) is 0.82 (95% CI 0.81 to 0.86). The vertical hatching along the curve corresponds to pointwise 95% confidence intervals for sensitivity, given specificity; horizontal hatching corresponds to pointwise confidence intervals for specificity, given sensitivity. The numbers indicate the cut-off points. The 95% confidence region for the recommended cut-off ≥ 10 is represented by the ellipse.



Accuracy of the GAD-2 questionnaire for identifying any anxiety disorder

See Summary of findings 4.

Primary analyses using the recommended cut-offs/core ranges

Nineteen studies investigated GAD-2 in the assessment of any anxiety disorder. All studies could be included in the primary analysis using the bivariate model.

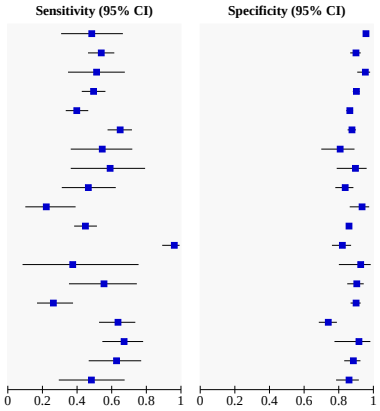
The median prevalence was 19% (range 3 to 65) for any anxiety disorder. The median prevalence for any anxiety disorder in the subgroups was as follows: 20% (range 3 to 30) in non-clinical settings, 18% (range 16 to 19) in clinical settings across conditions, and 19% (range 12 to 65) in clinical settings in people with specific conditions.

Diagnostic test accuracy data were available for the recommended cut-off (≥ 3) from 18 studies, while one study used a cut-off of 2 or

higher (Van Heyningen 2018). Figure 9 displays the sensitivity and specificity estimates and their respective 95% CIs for the individual studies. Point estimates from individual studies for sensitivity ranged widely from 0.22 to 0.96, those for specificity from 0.74 to 0.96. Also, estimates of sensitivity were less precise than those for specificity, as reflected by the larger 95% confidence intervals. The numbers of participants with any anxiety disorder (TP+FN) across the studies (median: 43, range: 8 to 232) was smaller compared to those without any anxiety disorder (FP+TN) (median: 199, range: 36 to 1699).

Figure 9. Forest plot of GAD-2 for detecting any anxiety disorder (AAD)
TP = true positive; FP = false positive; FN = false negative; TN = true negative; Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions; CI = confidence interval

Study	TP	FP	FN	TN	Setting	Population	Sensitivity (95% CI)	Specificity (95% CI)
Maric 2022	16	50	17	1120	I	A general population	0.48 [0.31, 0.66]	0.96 [0.94, 0.97]
Christodoulaki 2022	95	41	81	367	I	A general population	0.54 [0.46, 0.62]	0.90 [0.87, 0.93]
Simning 2012	20	7	19	144	I	D poor elderly	0.51 [0.35, 0.68]	0.95 [0.91, 0.98]
Ahn 2019	108	92	110	847	I	F mixed healthy/patients	0.50 [0.43, 0.56]	0.90 [0.88, 0.92]
Ivanovs 2018	92	168	139	1068	II	G unselected primary care	0.40 [0.33, 0.46]	0.86 [0.84, 0.88]
Kroenke 2007	122	96	66	681	II	G unselected primary care	0.65 [0.58, 0.72]	0.88 [0.85, 0.90]
Scott 2019	18	14	15	59	III	L neurology: epilepsy	0.55 [0.36, 0.72]	0.81 [0.70, 0.89]
Michaelis 2022	13	6	9	52	III	L neurology: epilepsy	0.59 [0.36, 0.79]	0.90 [0.79, 0.96]
Marrie 2018	20	34	23	175	III	P multiple sclerosis	0.47 [0.31, 0.62]	0.84 [0.78, 0.88]
Clover 2020	8	6	28	86	III	Q various cancers	0.22 [0.10, 0.39]	0.93 [0.86, 0.98]
Esser 2018	104	239	128	1460	III	Q various cancers	0.45 [0.38, 0.51]	0.86 [0.84, 0.88]
Ahmadi 2019	75	36	3	165	III	S coronary heart disease	0.96 [0.89, 0.99]	0.82 [0.76, 0.87]
Conway 2016	3	3	5	38	III	T heart transplantation survivors	0.38 [0.09, 0.76]	0.93 [0.80, 0.98]
Fischer 2014	15	16	12	151	III	U heart failure	0.56 [0.35, 0.75]	0.90 [0.85, 0.94]
Nath 2018	20	44	56	397	III	V pregnancy	0.26 [0.17, 0.38]	0.90 [0.87, 0.93]
Van Heyningen 2018	56	75	32	213	III	V pregnancy	0.64 [0.53, 0.74]	0.74 [0.68, 0.79]
Delgadillo 2012	45	3	22	33	III	W substance abuse	0.67 [0.55, 0.78]	0.92 [0.78, 0.98]
Bernstein 2018	27	23	16	176	III	Y inflammatory bowel disease	0.63 [0.47, 0.77]	0.88 [0.83, 0.93]
Hitchon 2020	14	17	15	104	III	Z rheumatoid arthritis	0.48 [0.29, 0.67]	0.86 [0.78, 0.92]



Across all studies, we estimated a summary sensitivity of 0.53 (95% CI 0.44 to 0.62) and a summary specificity of 0.89 (95% CI 0.86 to 0.91). The area of the 95% prediction region was 0.16, indicating less heterogeneity compared to the GAD-7 studies with generalised anxiety disorder as the target condition. The summary point, along with the 95% prediction region, is presented in Appendix 8.

Investigations of sources of heterogeneity

There was no statistical evidence of differences in sensitivity with regard to settings (P = 0.93), but specificity differed significantly between the subgroups (P = 0.019). Studies in non-clinical settings reported higher specificities. For four studies performed in non-clinical settings, the summary estimates were 0.50 (95% CI 0.32 to 0.68) for sensitivity and 0.93 (95% CI 0.90 to 0.95) for specificity. For two studies performed in clinical settings across conditions, the summary estimates were 0.53 (95% CI 0.28 to 0.76) and 0.87 (95% CI 0.81 to 0.92), respectively. For the 13 studies in clinical settings in people with specific conditions, sensitivity was 0.54 (95% CI 0.43 to 0.65) and specificity was 0.87 (95% CI 0.84 to 0.89). Heterogeneity remained high within the three settings (Appendix 8).

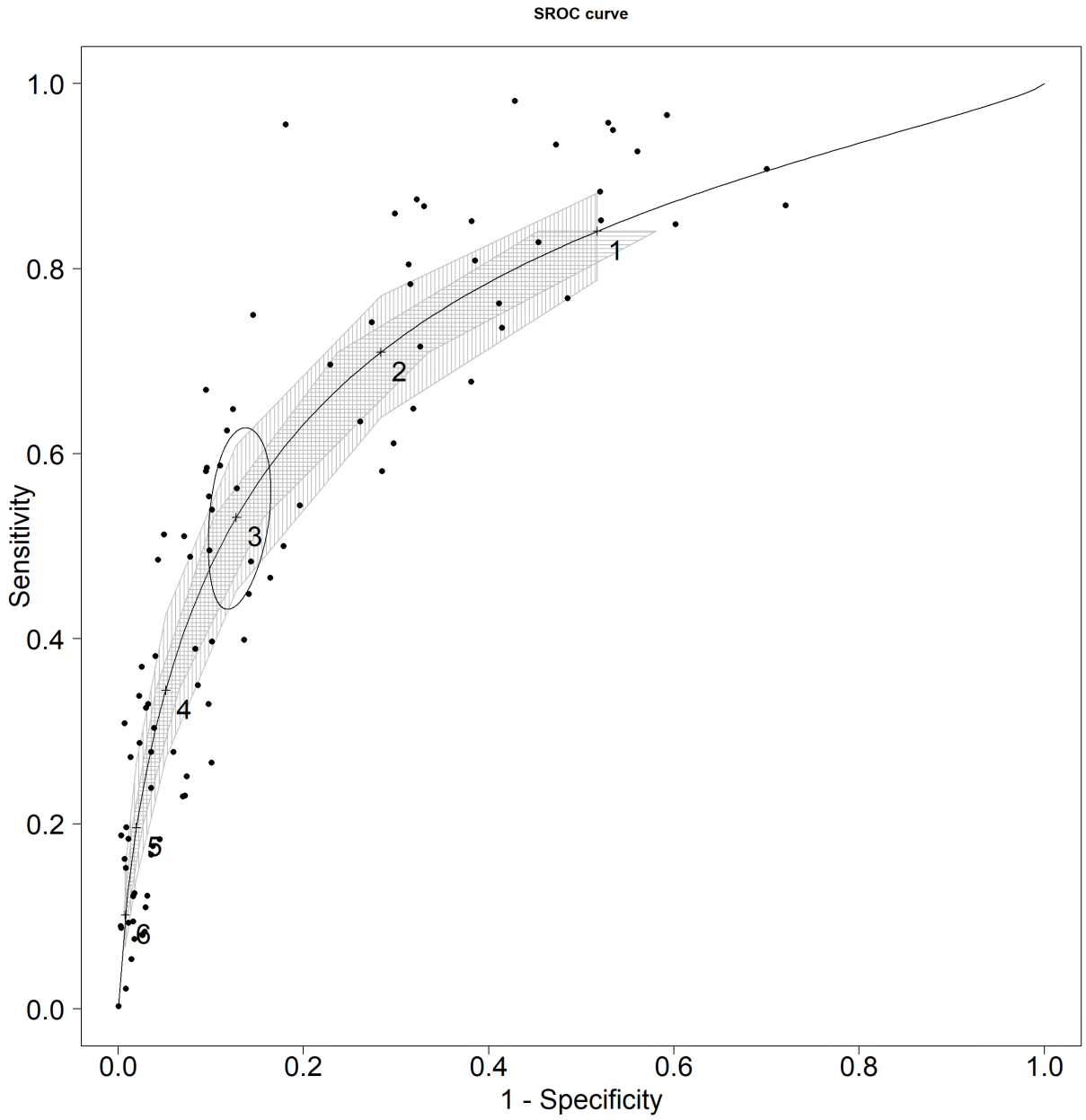
There were no statistically significant subgroup differences for risk of bias, type of reference standard, or any anxiety disorder prevalence (P > 0.05) (Appendix 8).

Secondary analyses

Summary accuracy estimates based on the bivariate model for each of the six GAD-2 cut-offs can be found in Appendix 8. At the cut-off recommended by the scale authors (≥ 3), summary sensitivity was 0.53 (95% CI 0.43 to 0.62) and summary specificity was 0.89 (95% CI 0.87 to 0.91). Using bivariate models, the highest Youden index (i.e. 0.49) was found at a cut-off of 2 or higher, with a sensitivity of 0.77 (95% CI 0.69 to 0.83) and a specificity of 0.73 (95% CI 0.67 to 0.78).

Diagnostic accuracy information from a total of 94 2x2-tables was available for analyses using the multiple thresholds model. The area under the curve (AUC) was determined as 0.77 (95% CI 0.76 to 0.82; Figure 10). Based on the multiple thresholds model, summary sensitivity was 0.53 (95% CI 0.45 to 0.61) and summary specificity was 0.87 (95% CI 0.84 to 0.90) for the cut-off recommended by the scale authors (≥ 3); the point estimates were similar to that found in the primary analysis using the core range. The Youden index was maximal (0.43) at the cut-off of 2 or higher, which revealed a sensitivity of 0.71 (95% CI 0.64 to 0.77) and a specificity of 0.72 (95% CI 0.67 to 0.76) (Appendix 8).

Figure 10. Multiple thresholds summary receiver operating characteristic (mtsROC) curve based on all available data or GAD-2 for detecting any anxiety disorder (AAD) The area under the curve (AUC) is 0.77 (95% CI 0.76 to 0.82). The vertical hatching along the curve corresponds to pointwise 95% confidence intervals for sensitivity, given specificity; horizontal hatching corresponds to pointwise confidence intervals for specificity, given sensitivity. The numbers indicate the cut-off points. The 95% confidence region for the recommended cut-off ≥ 10 is represented by the ellipse.



Comparison of the GAD-7 and GAD-2 questionnaires

The diagnostic accuracy of GAD-7 and GAD-2 did not differ significantly in the detection of GAD or AAD.

Twenty-two studies contained paired information about GAD-7 and GAD-2 for generalised anxiety disorder (Ahn 2019; Belus 2021; Bernstein 2018; Bisby 2022; Chibanda 2016; Conway 2016; Delgadillo 2012; Esser 2018; Hitchon 2020; Ivanovs 2018; Kroenke 2007; Kujanpää 2014; Maric 2022; Marrie 2018; Michaelis 2022; Micoulaud-Franchi 2016; Pranckeviciene 2022; Scott 2019; Seo 2015; Seo 2017; Veisy 2021; Wild 2014). The sensitivity of diagnosing generalised anxiety disorder was 0.65 (95% CI 0.54 to 0.74) for GAD-7 and 0.67 (95% CI 0.58 to 0.75) for GAD-2 (P = 0.78). The specificity of diagnosing generalised anxiety disorder was 0.89 (95% CI 0.85 to 0.92) for GAD-7 and 0.86 (95% CI 0.81 to 0.89) for GAD-2 (P = 0.21). The summary ROC figure showing the comparison of GAD-7 and

GAD-2 for generalised anxiety disorder is given in Appendix 9 (Figure 35).

Fifteen studies contained paired information about GAD-7 and GAD-2 for any anxiety disorder (Ahn 2019; Bernstein 2018; Clover 2020; Conway 2016; Delgadillo 2012; Esser 2018; Fischer 2014; Hitchon 2020; Ivanovs 2018; Kroenke 2007; Maric 2022; Marrie 2018; Michaelis 2022; Scott 2019; Simning 2012). The sensitivity of diagnosing any anxiety disorder was 0.44 (95% CI 0.36 to 0.52) for GAD-7 and 0.51 (95% CI 0.45 to 0.57) for GAD-2 (P = 0.20). The specificity of diagnosing any anxiety disorder was 0.93 (95% CI 0.90 to 0.95) for GAD-7 and 0.90 (95% CI 0.87 to 0.92) for GAD-2 (P = 0.09). The summary ROC figure showing the comparison of GAD-7 and GAD-2 for any anxiety disorder is given in Appendix 9 (Figure 36).

DISCUSSION

Summary of main results

In this review, we summarised the findings of a total of 48 studies (with 19,228 participants) investigating the diagnostic test accuracy of the GAD-7 scale, the GAD-2 scale, or both, for detecting generalised anxiety disorder, any anxiety disorder, or both, against validated structured or semi-structured clinical interviews as reference standard amongst adults. As we investigated both versions of the questionnaire for both generalised anxiety disorder and any anxiety disorder, this review has four primary analyses including between 19 and 35 studies (See Summary of findings 1; Summary of findings 2; Summary of findings 3; Summary of findings 4). The included studies are clinically and methodologically diverse (regarding setting, language of the questionnaire, participants, diagnostic interview, and risk of bias), and their diagnostic accuracy estimates are statistically heterogeneous, making the clinical interpretation of our findings difficult. We considered the overall risk of bias as low in only eight studies, and we had low concerns about the applicability of the findings in only three studies.

At the recommended cut-off of 10 or higher (or the closest available cut-off), the GAD-7 questionnaire yielded a summary sensitivity of 0.64 (95% CI 0.56 to 0.72) and a summary specificity of 0.91 (95% CI 0.87 to 0.93) in detecting generalised anxiety disorder. For diagnosing any anxiety disorder, the summary sensitivity was 0.48 (95% CI 0.40 to 0.57), and summary specificity was 0.91 (95% CI 0.89 to 0.93). At the recommended cut-off of 3 or higher (or the closest available cut-off), the GAD-2 yielded a summary sensitivity of 0.68 (95% CI 0.59 to 0.75) and a summary specificity of 0.86 (95% CI 0.82 to 0.89) in detecting generalised anxiety disorder. For detecting any anxiety disorder, the summary sensitivity was 0.53 (95% CI 0.44 to 0.62), and the summary specificity was 0.89 (95% CI 0.86 to 0.91). In all analyses, sensitivity tended to be somewhat higher and specificity lower in people with specific conditions than in the other two settings. All analyses showed a considerable degree of heterogeneity. The analysis of GAD-7 for detecting generalised anxiety disorder showed the most heterogeneous results compared to the other three analyses.

The area under the ROC curve in the multiple thresholds model was 0.86 (95% CI 0.84 to 0.88) for the GAD-7 scale in detecting generalised anxiety disorders, and 0.80 (95% CI 0.78 to 0.82) in detecting any anxiety disorders. For the GAD-2 scale, these values were 0.82 (95% CI 0.81 to 0.86) and 0.77 (95% CI 0.76 to 0.82), respectively. Comparative bivariate analyses revealed no statistically significant differences between the diagnostic test accuracy of GAD-7 and GAD-2.

Strengths and weaknesses of the review

Compared to the most comprehensive systematic review of the GAD-7 and the GAD-2 scales available so far (Plummer 2016), our review includes four times as many studies. Most studies were published after the completion of the literature search by Plummer and colleagues in 2014, and many investigated translations of the GAD-7 and the GAD-2 into further languages. The evidence synthesis backing the recommendation of the US Preventive Services Task Force (USPSTF) on anxiety screening included seven studies investigating GAD-7, GAD-2, or both (O'Connor 2023a; O'Connor 2023b). The summary estimates calculated were based on only three and two studies, respectively. Five of the studies were also included in our review, two were excluded by us, as we considered the reference standards not appropriate. The USPTF evidence synthesis did not include studies in clinical settings involving people with specific conditions, but did include studies in pregnant women.

In addition to using standard Cochrane methods for preparing systematic reviews of test accuracy (Deeks 2023), we made every effort to obtain additional data on diagnostic accuracy for unreported cut-off points from authors of primary studies. As the authors of 14 studies generously shared such data with us (see Acknowledgements and Characteristics of included studies), we could use new statistical methods to make the best use of all available data simultaneously, in addition to the standard bivariate analyses for single cut-offs. We sent the data we extracted from primary studies to authors/investigators for quality checking and approval; 17 authors/investigators responded. However, we were unable to obtain additional data, methodological information, or quality checking from the authors for most studies.

Our review addressed broad questions and used liberal eligibility criteria. Therefore, it provides a comprehensive overview of the available evidence. However, this also implies pronounced clinical, methodological, and statistical heterogeneity. Our subgroup analyses can only explain a modest part of this heterogeneity. The subgroup analysis according to setting was one of our four pre-defined subgroup analyses. However, we did not explicitly state in our protocol that we would prioritise them in a way that would make them almost part of the primary analyses.

A general problem for primary studies and systematic reviews is the diversity of anxiety disorders. Our decision to focus on generalised anxiety disorder (GAD) and any anxiety disorder (AAD) as target conditions for analysis was a pragmatic compromise. What was summarised under the category 'any anxiety disorder' differed to some extent between studies and with time (Bandelow 2015). We also considered analysing the other anxiety disorders less frequently reported, but abandoned this plan given the already high complexity of the review and the small number of studies reporting such information.

During the review process, we modified our coding guidelines for assessing the risk of bias and applicability in the patient selection domain. Initially, our approach for rating this domain was quite liberal, as we incorporated studies from 'any setting' into our review. Upon deciding to give precedence to subgroup analyses by setting, we adopted a more critical approach in evaluating whether

the participants were likely to 'represent' the specific setting in question. Furthermore, similar to most diagnostic accuracy reviews (Thombs 2011), our analysis does not consider the potential bias introduced by including patients who have already been diagnosed and might be undergoing treatment, which could alter their signs and symptoms. Nonetheless, such information was seldom reported in the studies we included.

For our primary analyses, we used diagnostic accuracy data for the cut-offs recommended by the scale authors whenever available. However, in the case of studies that did not provide data for these cut-offs, we used the data for the closest available cut-offs if they were within a pre-defined core range. We described this approach in our peer-reviewed published protocol. However, it may have increased heterogeneity and introduced bias due to selective reporting of cut-offs termed 'best' or 'optimal' (Hartzes 2019; Neupane 2021; Trikalinos 2012). Although it is known that a data-driven selection of cut-off values leads to a bias in the estimation, it has been shown that this problem is of minor importance with increasing sample size (Leeflang 2008b). By using all available data, the multiple thresholds model (MTM) offers a more comprehensive and unbiased analysis of the diagnostic performance of the GAD-7 and GAD-2 scales. Furthermore, we also performed secondary analyses, including only studies reporting diagnostic accuracy at the recommended cut-offs. Both analyses yielded estimates of sensitivity and specificity that were identical to or differed only minimally from those of our primary analyses.

We cannot rule out that our findings might be affected to some extent by publication bias or selective cut-off reporting. When assessing the eligibility of identified studies, we tried to contact the authors of all studies with other primary objectives, not reporting any diagnostic accuracy findings but having collected the relevant data. Nineteen authors did not respond to our requests or could not provide the information needed. For four studies, authors either performed diagnostic accuracy analyses in response to our request or provided the raw data necessary for such analysis. In general, the actual diagnostic accuracy results did not seem to have an influence on (non-)publication. Amongst the published studies presenting diagnostic accuracy findings, lack of a clearly pre-defined primary cut-off and selective reporting of cut-offs with a favourable balance of sensitivity and specificity was frequent. For 14 studies, the authors provided unpublished diagnostic accuracy results for additional cut-offs, and by using multiple threshold analysis, we attempted to reduce the risk of bias due to selective reporting.

In addressing potential methodological concerns, one issue that merits discussion is the risk of incorporation bias. Incorporation bias occurs when the index test influences the reference standard, potentially skewing accuracy measures (Schmidt 2013). The GAD-7 and GAD-2 scales, while aligned with DSM and ICD diagnostic criteria, do not replicate the exact wording or criteria used in the structured interviews that serve as our reference standards. Importantly, the index tests are completed independently by patients, and the reference standards are applied under blinded conditions by professionals, ensuring that our setup avoids classical incorporation bias. This separation minimises the risk that the diagnostic outcomes are unduly influenced by the index tests themselves. We have thoroughly assessed this aspect using the QUADAS-2 tool, confirming that the independence of the index test and reference standard is maintained, thereby reducing the potential for bias in our findings.

Applicability of findings to the review question

Why our diagnostic accuracy estimates might be difficult to apply

Overall, our review provides evidence that the GAD-7 and its shorter version GAD-2 have acceptable to good diagnostic accuracy for detecting generalised anxiety disorder and any anxiety disorder.

However, the available evidence seems insufficient to reliably predict the diagnostic performance of the questionnaires in specific contexts. Sensitivity and specificity estimates varied widely between the individual studies, and this heterogeneity remained high in the three settings and even in most groups of studies investigating similar populations. Our subgroup analyses suggest that differences between settings, prevalence, and type of reference standard can contribute to this heterogeneity. However, statistically significant subgroup differences were not observed consistently for the four study questions. Most of the observed heterogeneity remains unexplained, and we can only speculate on other reasons and their complex interaction. Language and translation issues, socioeconomic and cultural influences on the manifestation of anxiety disorders, the inherent vagueness of the concept of anxiety disorders, national differences in diagnosis, quality of the implementation of the reference standard, methodological quality in general, and differences in participant selection, in particular, seem plausible candidates. To be included in our review, the primary studies had to be compatible with the screening pathway described in the Background section. However, we very rarely (if ever) had the impression that the studies really tested such systematic screening approaches. Instead, most studies seemed to investigate the diagnostic accuracy of GAD-7, GAD-2, or both in pragmatically selected participants giving consent in the settings accessible to the respective study authors. This might explain the large diversity of study populations and why so few studies excluded (or mentioned whether they excluded) individuals with known anxiety disorders. In a number of studies, the investigation of diagnostic accuracy was part of a validation process of a new translation of the questionnaires. The prevalence of generalised anxiety disorder and any anxiety disorder within the three settings varied substantially even between studies in which selection criteria seemed similar. Sampling was often described as ‘random’ or ‘consecutive’, but the recruitment and eligibility processes were rarely reported in detail. And when reported, it was clear that the selection process was often hampered by situations such as lacking consent.

Applying our findings to different scenarios

Due to the problems described above, our results can only give a rough indication of what a specific test result may mean in practice. In the following, we present four scenarios for illustration. These simple scenarios give a first impression of the complexity of the interpretation in real life situations. They do not take the uncertainty of the point estimates and variable pre-test probabilities into account. Furthermore, the scenarios described are based on the analysis of GAD-7 for generalised anxiety disorders, which have been the focus of investigation most frequently.

Scenario 1 – applying the overall summary estimates across all settings (Figure 11, A)

Figure 11. Graphical presentation of the screening scenarios applying to GAD-7 to detect generalised anxiety disorder (GAD) A: across all studies, B: clinical - across conditions, C: clinical - specific conditions, D: across all studies using the cut-off ≥7 maximising the Youden index multiple thresholds model across all studies at cut-off 7 (maximal Youden). Scenarios A, B, and C are based on analyses using the bivariate model for the cut-off ≥10 or the

closest cut-off. Scenario D is based on an analysis using the multiple thresholds model. (Figures produced at https://testbaum.de/)

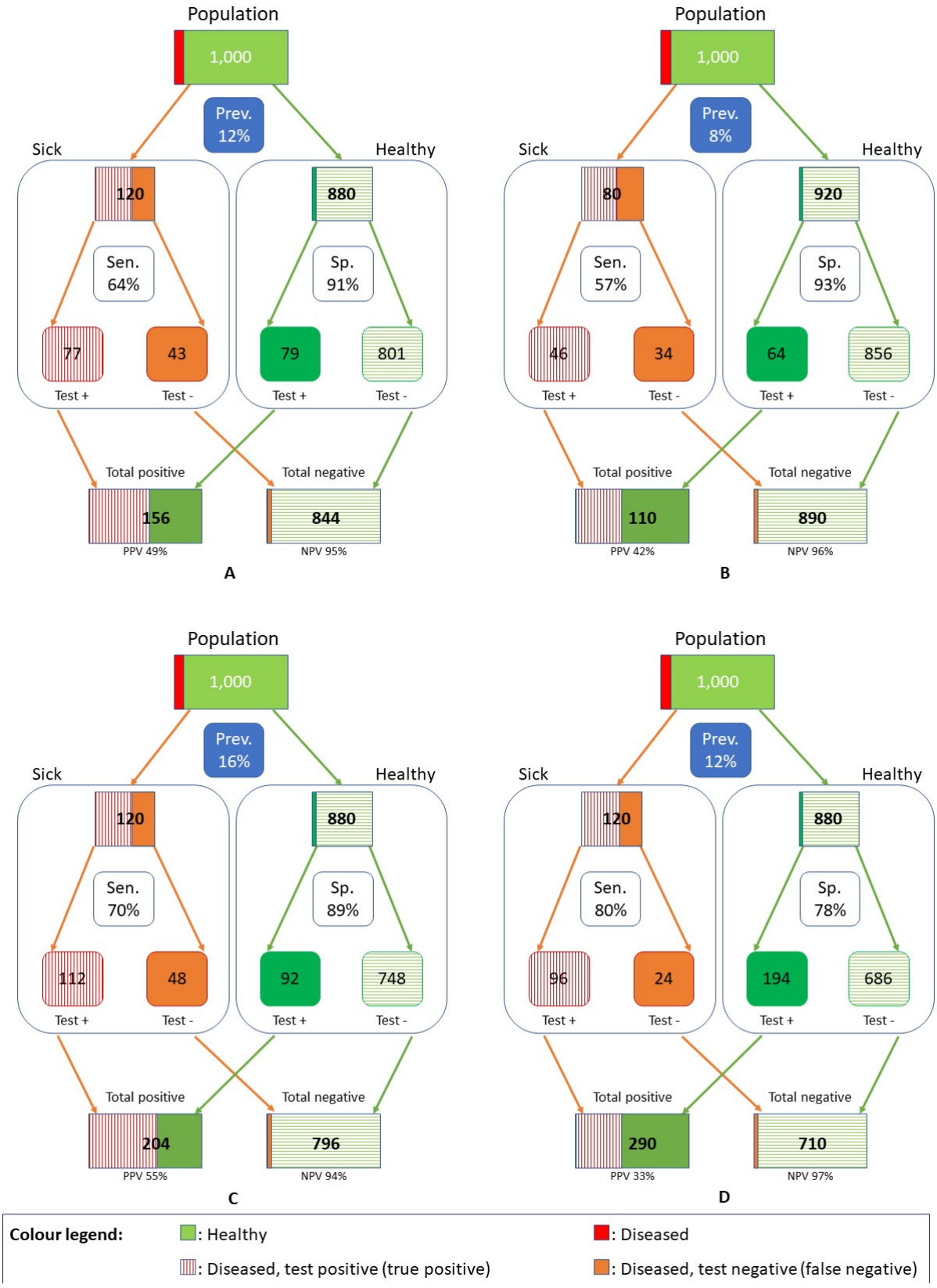
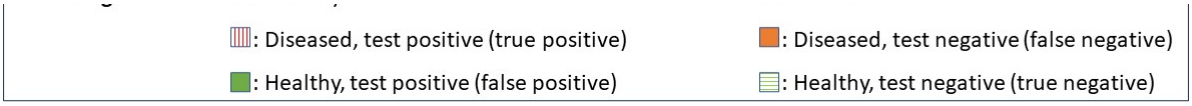


Figure 11. (Continued)



As our analyses provide only a weak indication of relevant subgroup differences, our summary estimates across all studies can be considered as an acceptable average to summarise the overall results. When applying the summary estimates of a sensitivity of 0.64 and a specificity of 0.91 for GAD-7 at a cut-off of 10 or higher (or the closest lower cut-off lying in our pre-defined core range) to a screening scenario of 1000 consecutive patients in whom the prevalence of generalised anxiety disorder is 12% (the median prevalence in the studies included in the analysis), we would expect a positive test result in 156 patients (16%). In the case of such a result, the probability that a patient suffers from generalised anxiety disorder would increase from 12% before screening to 49% after screening. The screening finding would be true positive in 77 patients (8%) and false positive in 79 (8%). A negative test result would be expected in 844 patients (84%), of whom 43 (4%) would be false negative and actually have generalised anxiety disorder, while 801 patients (80%) could be correctly classified as negative. In the case of a negative test result, the probability of not having generalised anxiety disorder would increase from 88% before filling in the questionnaire to 95%.

Scenario 2 - using the questionnaire in consecutive primary care patients (Figure 11, B)

While the summary estimates from the analysis across all studies can be considered as an overall indicator of the diagnostic accuracy of GAD-7, it seems preferable to use the estimates in the respective setting subgroup when considering the use of GAD-7 in practice. In the studies in clinical settings recruiting patients regardless of their conditions, the median prevalence of generalised anxiety disorder was lower (8% in both settings) than in the total group (12%). Sensitivity was slightly lower and specificity was slightly higher.

A setting for which anxiety screening is discussed is general practice and family medicine (Thombs 2023; USPSTF 2023). When applying our point estimates of a sensitivity of 0.57 and a specificity of 0.93 for GAD-7 from the subgroup 'clinical settings across conditions' to a scenario of 1000 consecutive primary care patients in whom the prevalence of generalised anxiety disorder is 8%, the findings would be somewhat different from scenario 1. We would expect a positive test result in 110 (11% of patients). In the case of such a result, the probability that a patient suffers from generalised anxiety disorder would increase from 8% before filling out the questionnaire to 42% afterwards (when completely ignoring all other information about the patient). The questionnaire finding would be true positive in 46 patients (5%) and false positive in 64 (6%). A negative test result would be expected in 890 (89%) patients, of whom 34 (3%) would be false negative and actually have generalised anxiety disorder. In the case of a negative test result, the probability of not having generalised anxiety disorder would increase from 92% before filling in the questionnaire to 96%.

Scenario 3 - screening in consecutive patients in an epilepsy outpatient clinic (Figure 11, C)

Amongst the studies recruiting people with specific conditions in clinical settings, the median prevalence was markedly higher (16%) than in studies in clinical settings recruiting people regardless of their conditions (8%). Sensitivity was higher and specificity somewhat lower.

When applying our point estimates of a sensitivity of 0.70 and a specificity of 0.89 from the subgroup 'clinical settings for specific conditions' to a screening scenario of 1000 consecutive patients with epilepsy seeking care in a specialised outpatient clinic in whom the prevalence of generalised anxiety disorder is 16%, we would expect a positive test result in 204 patients (20%). In this case, the probability that a patient suffers from generalised anxiety disorder would increase from 16% before screening to 55% after screening (when completely ignoring all other information about the patient). The screening finding would be true positive in 112 patients (11%) and false positive in 92 (9%). A negative test result would be expected in 796 (80%) patients, of whom 48 (5%) would be false negative and actually have generalised anxiety disorder. In the case of a negative test result, the probability of not having generalised anxiety disorder would increase from 80% before filling in the questionnaire to 94%.

Scenario 4 - using a lower cut-off with maximised Youden index (Figure 11, D)

A logical question is whether the performance of the GAD-7 would be better at another cut-off. Of the studies included in our review, many used the Youden index for identifying the 'best' cut-off. The Youden index combines sensitivity and specificity into a single measure (sensitivity + specificity - 1). In our multiple thresholds analysis of all available data, the highest Youden index was found for the cut-off of 7 or higher. When applying the respective point estimates of a sensitivity of 0.80 and a specificity of 0.78 to a scenario of 1000 individuals in whom the prevalence of generalised anxiety disorder is 12%, the findings would be the following: we would expect a positive test result in 290 (29% of patients). In the case of such a result, the probability that a patient suffers from generalised anxiety disorder would increase from 12% before filling the questionnaire to 33% afterwards. The questionnaire finding would be true positive in 96 patients (10%) but false positive in 194 (19%). A negative test result would be expected in 710 (71%) patients, of whom 24 (2%) would be false negative and actually have generalised anxiety disorder. In the case of a negative test result, the probability of not having generalised anxiety disorder would increase from 88% before filling in the questionnaire to 97%.

Above all, our scenarios demonstrate that GAD-7 and GAD-2 are not suitable for making the 'diagnosis' of anxiety disorder. Negative predictive values were around 95% in the case of the less prevalent target condition 'generalised anxiety disorder'. For the more frequent target condition 'any anxiety disorder', however, the negative predictive values were slightly below 90%. Our scenarios illustrate the interplay of positive and predictive values, sensitivity and specificity, and prevalence, which complicates

prediction and interpretation of test results. In the past, sensitivity and specificity were often considered stable characteristics of a diagnostic test, irrespective of the population being tested. However, there are many examples showing that sensitivity and specificity can vary in different populations (Knottnerus 1992; Mulherin 2002; Ransohoff 1978). Increasing prevalence is often associated with increasing sensitivity and decreasing specificity (Leeflang 2009), a finding observed to some extent in our data. Yet, the differences between positive and negative predictive values in the first three scenarios are relatively small. In our opinion, this indicates that the summarised estimates from the analyses across all studies provide a good indication of the average performance of the questionnaires.

Comparing scenarios 1 and 4 nicely illustrates the problems of using the Youden index to identify the 'best' cut-off. This approach makes sense only if sensitivity and specificity are diagnostically equally important. While the lower cut-off in scenario 4 led to 19 more (96 instead of 77 in scenario 1) individuals with generalised anxiety disorder being detected, the number of false positives increased by 115 (194 instead of 79). The positive predictive value of a positive test result decreased to 33% (compared to 49%), while the negative predictive value of a negative test result increased from 95% to 97%.

There are two sides to every coin: on one side, unnecessary emotional, physical, or financial harm in the form of referrals, labelling, unease, and stigma related to the diagnostic process, as well as potentially unnecessary treatments resulting from screening that have side effects, must be considered (Lin 2007); on the other side, there is the risk of delay in diagnosing an anxiety disorder, which also must be regarded as a burden for the person affected, as well as society. Additionally, many conditions, including chronic diseases, somatoform disorders, and ageing, may exhibit anxiety symptoms, making it more complicated to decide to implement anxiety screening tools in health practice. Therefore, the possible benefits and harms of identifying or missing patients as well as adverse effects and costs of further investigations must be considered in selecting cut-offs. If further testing brings small burden to the patient, provider, or the health system, but the benefits of identifying cases are high, lower cut-offs may be selected, leading to higher sensitivity but lower specificity (Smits 2007).

Another aspect to consider in the context of anxiety disorders is the potentially low motivation amongst patients with mild and transient symptoms to seek medical advice. Diagnosing anxiety relies on time-intensive psychiatric evaluations by healthcare professionals, as no laboratory-based diagnostic test exists. Consequently, false positive results from screening could demand significant human resources, burdening clinical staff who are already under pressure to assess or refer all cases identified by positive screening. Additionally, there is a real risk that patients could suffer harm from incorrect labelling, side effects of medications, or co-payments for services. From this viewpoint, the screening tools GAD-7 and GAD-2 might offer greater value in scenarios where the pre-test probability of anxiety is higher. Thus, screening appears most justified amongst individuals with an elevated likelihood of having anxiety disorders, such as those exhibiting symptoms or with comorbid health conditions that increase the risk of such disorders.

AUTHORS' CONCLUSIONS

Implications for practice

The Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales have been tested in many languages and different populations. The findings of our review indicate that, on a technical level, both GAD-7 and GAD-2 have good or acceptable diagnostic accuracy for detecting our target conditions, generalised anxiety disorder and any anxiety disorder. When interpreting the scores of GAD-7 and GAD-2 in practice, clinicians might refer primarily to the estimates of sensitivity and specificity of the setting / pathway subgroup fitting best with the specific clinical situation. However, due to the diversity of the studies and the heterogeneity of the findings, our summary estimates of sensitivity and specificity should be interpreted as rough averages. The performance of GAD-7 and GAD-2 may deviate substantially from these values in specific situations. Interestingly, the GAD-2 scale seems to have similar diagnostic accuracy as the GAD-7 scale.

In our opinion, our results and scenarios cast doubt on whether using the GAD-7 and GAD-2 scales (and probably other questionnaires) with fixed cut-offs for screening in non-healthcare settings and in clinical settings amongst unselected patients seeking medical care is an efficient method for identifying patients with anxiety disorders. With the observed level of diagnostic accuracy, the number of misclassifications is high. This problem is unlikely to be resolved by using the lower ‘optimal’ cut-off values found in our secondary analyses. Large numbers of people with positive screening findings would need further clarification in diagnostic interviews. In people seeking specialised healthcare services for specific diseases shown to be associated with an increased prevalence of anxiety disorders, the questionnaires might be more efficient for screening than in the ‘non-healthcare’ and the 'clinical setting across conditions' settings because the positive predictive values increase by increasing prevalence (pre-test probability) and because implementation of screening is probably easier. In general, if the GAD-7 and the GAD-2 scales are used in populations with low prevalence, they are likely to be useful for ruling out an anxiety disorder due to high specificity. Based on the available evidence, a ruling-in based on the currently recommended cut-off does not seem possible in screening populations.

Other approaches for using the GAD-7 and GAD-2 scales might be clinically more useful in practice, but they do not follow a classical screening approach and have not yet been investigated. For example, it seems plausible to use the summary score of GAD-7 or GAD-2 as an indicator of the presence and severity of anxiety symptoms together with other clinical information. A very low score without any further hints about the possible presence of an anxiety disorder would result in no further action. At a moderate score level (around the currently suggested cut-offs) and in the absence of further diagnostic hints, one might consider re-applying the questionnaire after an adequate period of time, but apply a closer clinical enquiry in case of further clues. High scores might immediately lead to further clinical investigation. Which score levels are to be considered low or high might also depend on the specific context.

It is critical to acknowledge that the GAD-7 and GAD-2 scales, while effective for screening, are not standalone diagnostic tools. Comprehensive clinical evaluations are crucial to confirm

these preliminary findings and guide treatment decisions. This clarification is vital to ensure that these tools are used appropriately within the spectrum of clinical assessment and are not over-interpreted beyond their intended utility.

In 2023, the US Preventive Services Task Force on anxiety screening recommended screening for anxiety in adults aged 64 years or younger, including pregnant and postpartum persons (USPSTF 2023). To the best of our knowledge, similar recommendations do not exist in other countries. Although there are advocates suggesting improved outcomes and economic benefits of anxiety screening (Katon 2007; Lang 2002), the evidence for screening anxiety disorders is insufficient (Nelson 2020). Our review focused on the diagnostic accuracy of the GAD-7 and GAD-2 questionnaires. We did not explore whether screening for anxiety with these tools leads to improved clinical outcomes. Nonetheless, if the diagnostic accuracy of these instruments is not sufficiently high, it seems improbable that screening would enhance clinical outcomes, and there may be an elevated risk of harm.

As the diagnostic accuracy of GAD-7 and GAD-2 is acceptable to good, but far from optimal, and there is limited evidence indicating that anxiety screening is beneficial, we think that the decision to use these questionnaires should be left to the individual healthcare professional, institution, or initiative.

Implications for research

In principle, diagnostic accuracy studies of self-report questionnaires against diagnostic interviews as reference standard are straightforward, but we were surprised to see how many studies had flaws that could easily be avoided. As simple examples, a sufficient number of individuals should be enrolled consecutively or randomly; interviewers have to be blinded regarding questionnaire findings; and the time interval between the index test and the reference standard should be as short as methodologically feasible to minimise bias. Many studies did not report on whether they excluded individuals with a current diagnosis or receiving treatment for an anxiety disorder. If studies aim to investigate a systematic screening approach, they should exclude such people. If they are not excluded in the studies, this should be carefully documented and reported.

When planning and reporting studies, we strongly recommend researchers carefully follow the well-established guidelines for reporting diagnostic accuracy studies and the accompanying explanation and elaboration documents (Bossuyt 2015; Cohen 2016). In particular, we emphasise that the recruitment and selection process have to be documented in a detailed flow chart. It is of great importance to allow the reader to understand which population the study sample 'represents'. Furthermore, studies should routinely report findings on all cut-offs and all anxiety disorders classified in the diagnostic interviews (in an appendix) to avoid selective reporting and allow reliable and powerful future meta-analyses (Brehaut 2022). If GAD-7 is used, analyses should always be performed on GAD-2 as well.

Many of the existing studies on the diagnostic accuracy of the GAD-7 and GAD-2 scales do not constitute true screening research in populations where systematic screening could feasibly be implemented. Nor do they examine the effectiveness of more adaptable strategies that utilise the questionnaire scores as part of a diagnostic process or indicated screening approach. Future research should aim to concentrate more consistently on one of these two avenues.

Much research is done on the development and validation of anxiety screening tools, adaptation of the available tools in different languages and cultures, or testing the instruments in different clinical and non-clinical populations. This research is valuable and important. However, research is lacking on how and how frequently questionnaires such as GAD-7 and GAD-2 are used by individuals and healthcare professionals for screening, as a diagnostic aid, and other diagnostic purposes. Available research suggests that general practitioners rarely use anxiety questionnaires (see, for example, Turner 2020). Qualitative studies might help to understand obstacles to a wider use. Furthermore, more evidence is needed on whether screening for anxiety with these tools leads to improved clinical outcomes.

However, apart from the question of the diagnostic accuracy of GAD-7, GAD-2, and other screening questionnaires at the currently recommended or other cut-off values, the following should be noted: an evidence-based decision on whether to introduce anxiety screening requires high-quality randomised trials examining the benefits and harms of screening.

ACKNOWLEDGEMENTS

We thank Terry Quinn for his advice on handling parallel reviews based on a joint protocol; Bada Yang for his advice concerning implementation of the QUADAS-C; our advisory patient and provider team Alexandra Dührssen, Kerstin Schaeffer, Markus Hendel, Frederike Kneissl, and Tom Brandhuber. We thank our colleagues Clara Teusen and Siranush Karapetyan for their help in screening, selecting studies, and data extraction.

We thank the study authors for supporting us with additional information for the following studies: Beard 2014; Belus 2021; Bernstein 2018; Bisby 2022; Chibanda 2016; Christodoulaki 2022; Conway 2016; Donker 2011; Esser 2018; Fischer 2014; Grech 2019; Hitchon 2020; Kroenke 2007; Lickova 2021; Maric 2022; Marrie 2018; Michaelis 2022; Micoulaud-Franchi 2016; Nath 2018; Pranckeviciene 2022; Scott 2019; Simning 2012; Van Heyningen 2018; Zhong 2015; Zinchuk 2021; Zinchuk 2022. We also thank the study authors for providing data quality-check feedback for the following studies: Beard 2014; Belus 2021; Bernstein 2018; Bisby 2022; Budikayanti 2019; Donker 2011; Esser 2018; Grech 2019; Hitchon 2020; Ivanovs 2018; Kertz 2013; Marrie 2018; Michaelis 2022; Munoz-Navarro 2017b; Pranckeviciene 2022; Scott 2019; Seo 2014; Seo 2015; Seo 2017; Simning 2012; Zinchuk 2021; Zinchuk 2022.

Editorial and peer-review contributions

Cochrane Common Mental Disorders supported the authors in the development of this review up to the Group's closure in March 2023. We acknowledge Rachel Churchill, Jessica Hendon, and Nick Meader for their help in the various steps of the review development. The following people conducted the editorial process for this review.

- Sign-off Editors (final editorial decision)
  - Mia Schmidt-Hansen, National Guideline Alliance; Royal College of Obstetricians and Gynaecologists, London

- Gerald Gartlehner, Cochrane Austria, Department for Evidence-based Medicine and Evaluation, Danube University Krems

- Managing Editors (selected content peer reviewers, collated comments, provided editorial comments/guidance to authors, edited the article): Gail Quinn and Joey Kwong, Cochrane Central Editorial Service
- Editorial Assistant (conducted editorial policy checks, supported editorial team): Lisa Wydrzynski, Cochrane Central Editorial Service
- Copy Editor (copy-editing and production): Faith Armitage, Cochrane Central Production Service
- Peer-reviewers (provided comments and recommended an editorial decision): Cochrane DTA Editorial Team (methods review); Susan van Wyk (general methods review); Zosia Beckles (search review); Francesca Chappell (statistical review); Zahra Goodarzi, Division of Geriatrics, Department of Medicine, University of Calgary (clinical/content review); Brett D. Thombs, Lady Davis Institute of the Jewish General Hospital and McGill University, Montreal, Quebec, Canada (clinical/content review); Leslie A. Perdue, Kaiser Permanente Evidence-based Practice Center, Center for Health Research, Portland, OR, USA (clinical/content review); Bradley N. Gaynes, Department of Psychiatry, University of North Carolina at Chapel Hill (clinical/content review)

REFERENCES

References to studies included in this review

Ahmadi 2019 *{published data only}*

Ahmadi SM, Arani AM, Bakhtiari M, Emamy MH. Psychometric properties of Persian version of patient health questionnaires-4 (PHQ-4) in coronary heart disease patients. *Iranian Journal of Psychiatry and Behavioral Sciences* 2019;**13**(4):e85820.

Ahn 2019 *{published data only}*

* Ahn JK, Kim Y, Choi KH. The psychometric properties and clinical utility of the Korean version of GAD-7 and GAD-2. *Frontiers in Psychiatry* 2019;**10**:127.

Kim SH, Park K, Yoon S, Choi Y, Lee SH, Choi KH. A brief online and offline (paper-and-pencil) screening tool for generalized anxiety disorder: the final phase in the development and validation of the mental health screening tool for anxiety disorders (MHS: A). *Frontiers in Psychology* 2021;**12**:639366.

Kim Y, Park Y, Cho G, Park K, Kim SH, Baik SY, et al. Screening tool for anxiety disorders: development and validation of the Korean anxiety screening assessment. *Psychiatry Investigation* 2018;**15**(11):1053.

Baker 2018 *{published data only}*

Baker AM, Holbrook JT, Yohannes AM, Eakin MN, Sugar EA, Henderson RJ, et al. Test performance characteristics of the air, GAD-7, and HADS-Anxiety screening questionnaires for anxiety in chronic obstructive pulmonary disease. *Annals of the American Thoracic Society* 2018;**15**(8):926-34.

Belus 2021 *{published and unpublished data}*

Belus JM, Muanido A, Cumbe VF, Manaca MN, Wagenaar BH. Psychometric validation of a combined assessment for anxiety and depression in primary care in Mozambique (CAD-MZ). *Assessment* 2022;**29**(8):1890-900. [DOI: 10.1177/10731911211032285]

Bernstein 2018 *{published and unpublished data}*

Bernstein CN, Zhang L, Lix LM, Graff LA, Walker JR, Fisk JD, et al. The validity and reliability of screening measures for depression and anxiety disorders in inflammatory bowel disease. *Inflammatory Bowel Diseases* 2018;**24**(9):1867-75.

Bisby 2022 *{published data only}*

Bisby MA, Karin E, Scott AJ, Dudeney J, Fisher A, Gandy M, et al. Examining the psychometric properties of brief screening measures of depression and anxiety in chronic pain: The Patient Health Questionnaire 2-item and Generalized Anxiety Disorder 2-item. *Pain Practice* 2022;**22**(4):478-86.

Budikayanti 2019 *{published data only}*

Budikayanti A, Larasari A, Malik K, Syeban Z, Indrawati LA, Octaviana F. Screening of generalized anxiety disorder in patients with epilepsy: using a valid and reliable Indonesian version of Generalized Anxiety Disorder-7 (GAD-7). *Neurology Research International* 2019;**2019**(2):5902610. [DOI: 10.1155/2019/5902610]

Chibanda 2016 *{published and unpublished data}*

Chibanda D, Verhey R, Gibson LJ, Munetsi E, Machando D, Rusakaniko S, et al. Validation of screening tools for depression and anxiety disorders in a primary care population with high HIV prevalence in Zimbabwe. *Journal of Affective Disorders* 2016;**198**:50-5.

Christodoulaki 2022 *{published data only}*

Christodoulaki A, Baralou V, Konstantakopoulos G, Touloumi G. Validation of the Patient Health Questionnaire-4 (PHQ-4) to screen for depression and anxiety in the Greek general population. *Journal of Psychosomatic Research* 2022;**160**:110970. [DOI: 10.1016/j.jpsychores.2022.110970]

Clover 2020 *{published data only}*

* Clover K, Lambert SD, Oldmeadow C, Britton B, King MT, Mitchell AJ, et al. Apples to apples? Comparison of the measurement properties of Hospital Anxiety and Depression-Anxiety subscale (HADS-A), Depression, Anxiety and Stress Scale-Anxiety subscale (DASS-A), and Generalised Anxiety Disorder (GAD-7) scale in an oncology setting using Rasch analysis and diagnostic accuracy statistics. *Current Psychology* 2020;**41**(15):4592-601. [DOI: 10.1007/s12144-020-00906-x]

Clover K, Lambert SD, Oldmeadow C, Britton B, Mitchell AJ, Carter G, et al. Convergent and criterion validity of PROMIS anxiety measures relative to six legacy measures and a structured diagnostic interview for anxiety in cancer patients. *Current Psychology* 2022;**41**(15):4592-601.

Conway 2016 *{published and unpublished data}*

Conway A, Sheridan J, Maddicks-Law J, Fulbrook P, Ski CF, Thompson DR, et al. Accuracy of anxiety and depression screening tools in heart transplant recipients. *Applied Nursing Research* 2016;**32**:177-81.

Delgadillo 2012 *{published data only}*

Delgadillo J, Payne S, Gilbody S, Godfrey C, Gore S, Jessop D, et al. Brief case finding tools for anxiety disorders: validation of GAD-7 and GAD-2 in addictions treatment. *Drug and Alcohol Dependence* 2012;**125**(1-2):37-42.

Esser 2018 *{published and unpublished data}*

Esser P, Hartung TJ, Friedrich M, Hermann F, Koch U, Braehler E, et al. Diagnostic accuracy of the HADS-A and GAD-7 as a screening tool for generalized anxiety disorder among cancer patients. *Psycho-Oncology* 2017;**26**:40. [DOI: 10.1002/pon.4476]

* Esser P, Hartung TJ, Friedrich M, Johansen C, Wittchen HU, Faller H, et al. The Generalized Anxiety Disorder Screener (GAD-7) and the anxiety module of the Hospital and Depression Scale (HADS-A) as screening tools for generalized anxiety disorder among cancer patients. *Psycho-Oncology* 2018;**27**(6):1509-16.

Faller H, Weis J, Koch U, Brähler E, Härter M, Keller M, et al. Utilization of professional psychological care in a large German sample of cancer patients. *Psycho-Oncology* 2017;**26**(4):537-43. [PMID: https://pubmed.ncbi.nlm.nih.gov/27327213/]

Geue K, Brähler E, Faller H, Härter M, Schulz H, Weis J, et al. Prevalence of mental disorders and psychosocial distress in German adolescent and young adult cancer patients (AYA). *Psycho-Oncology* 2018;**27**(7):1802-9. [PMID: https://pubmed.ncbi.nlm.nih.gov/29644783/]

Fischer 2014 *{published and unpublished data}*

Fischer HF, Klug C, Roeper K, Blozik E, Edelmann F, Eisele M, et al. Screening for mental disorders in heart failure patients using computer-adaptive tests. *Quality of Life Research* 2014;**23**(5):1609-18.

Garabiles 2020 *{published data only}*

Garabiles MR, Lao CK, Yip P, Chan EW, Mordeno I, Hall BJ. Psychometric validation of PHQ-9 and GAD-7 in Filipino migrant domestic workers in Macao (SAR), China. *Journal of Personality Assessment* 2020;**102**(6):833-44.

Grech 2019 *{published and unpublished data}*

Grech M, Turnbull DA, Wittert GA, Tully PJ, CHAMPS Investigators. Identifying the internalizing disorder clusters among recently hospitalized cardiovascular disease patients: a receiver operating characteristics study. *Frontiers in Psychology* 2019;**10**:2829.

Hitchon 2020 *{published and unpublished data}*

Hitchon C, Zhang L, Bolton J, Graff L, Lix L, Peschken C, et al. Validity of screening tools for depression and anxiety in rheumatoid arthritis. *Arthritis Care & Research* 2018;**45**:1007. [PMID: https://pubmed.ncbi.nlm.nih.gov/31199570/]

Hitchon CA, Zhang L, Peschken CA, Lix LM, Graff LA, Fisk JD, et al. Validity and reliability of screening measures for depression and anxiety disorders in rheumatoid arthritis. *Arthritis Care & Research* 2020;**72**(8):1130-9.

Homans 2012 *{published data only}*

Homans WA. The Validity of the PHQ-9 and the GAD-7 for Screening Depressive and Anxiety Disorders in Sick-listed Workers [Masters thesis]. Utrecht, the Netherlands: Clinical and Health Psychology, Utrecht University Trimbos Institute, 2012. [STUDENT THESES: https://studenttheses.uu.nl/bitstream/handle/20.500.12932/19666/Homans%200472956.pdf?sequence=1]

Ivanovs 2018 *{published data only}*

* Ivanovs R, Kivite A, Ziedonis D, Mintale I, Vrublevska J, Rancans E. Association of depression and anxiety with cardiovascular co-morbidity in a primary care population in Latvia: a cross-sectional study. *BMC Public Health* 2018;**18**(1):328. [PMID: https://pubmed.ncbi.nlm.nih.gov/29510681/]

Ivanovs R, Kivite A, Ziedonis D, Mintale I, Vrublevska J, Rancans E. Association of depression and anxiety with the 10-year risk of cardiovascular mortality in a primary care population of Latvia using the SCORE system. *Frontiers in Psychiatry* 2018;**9**:276.

Vrublevska J, Renemane L, Kivite-Urtane A, Rancans E. Validation of the generalized anxiety disorder scales (GAD-7 and GAD-2) in primary care settings in Latvia. *Frontiers in Psychiatry / Frontiers Research Foundation* 2022;**13**(972628):1-10. [PMID: https://pubmed.ncbi.nlm.nih.gov/36276320/]

Kroenke 2007 *{published and unpublished data}*

* Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine* 2007;**146**(5):317-25.

Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* 2006;**166**(10):1092-7.

Kujanpää 2014 *{published data only}*

Kujanpää T, Ylisaukko-Oja T, Jokelainen J, Hirsikangas S, Kanste O, Kyngäs H, et al. Prevalence of anxiety disorders among Finnish primary care high utilizers and validation of Finnish translation of GAD-7 and GAD-2 screening tools. *Scandinavian Journal of Primary Health Care* 2014;**32**(2):78-83.

Li 2014 *{published data only}*

Li WL, Rongjing D, Dayi H, Sheng L. The value of Chinese version GAD-7 and PHQ-9 to screen anxiety and depression in cardiovascular outpatients. *Journal of the American College of Cardiology* 2014;**64**(16 Suppl):C222.

Lickova 2021 *{published data only}*

Licková K, Čoček A, Ambruš M, Soumarová R, Vránová J, Klézl P, et al. Rapid screening of depression and anxiety in cancer patients: Interview validation of emotion thermometers. *Asian Journal of Psychiatry* 2021;**65**:102827.

Makulowich 2018 *{published data only}*

Makulowich AA. Identification of patients with anxiety symptoms in an integrated community care clinic among a predominantly Latino patient population [Doctoral dissertation]. Chicago (IL): Department of Psychology. Rosalind Franklin University of Medicine and Science, 2018. [https://www.proquest.com/openview/8c59d6c5209850724113b56a6aa36085/1?cbl=18750&pq-origsite=gscholar&parentSessionId=lxqweNxT%2FOYhE%2F2UIfEr2SBQ85TrZbZM%2BXDURt4VajY%3D]

Maric 2022 *{unpublished data only}*

Maric NP, Lazarevic LJ, Priebe S, Mihic LJ, Pejovic-Milovancevic M, Terzic-Supic Z, et al. Covid-19-related stressors, mental disorders, depressive and anxiety symptoms: a cross-sectional, nationally-representative, face-to-face survey in Serbia. *Epidemiology and Psychiatric Sciences* 2022;**31**:e36. [PMID: https://pubmed.ncbi.nlm.nih.gov/35607805/]

Marrie 2018 *{published and unpublished data}*

Marrie RA, Zhang L, Lix LM, Graff LA, Walker JR, Fisk JD, et al. The validity and reliability of screening measures for depression and anxiety disorders in multiple sclerosis. *Multiple Sclerosis and Related Disorders* 2018;**20**:9-15.

Michaelis 2022 *{published and unpublished data}*

Michaelis R, Schlömer S, Popkirov S, Krämer G, Lindemann A, Cosentino M, et al. German translation and validation of the

brief Epilepsy Anxiety SurveyInstrument (brEASI). *Epilepsy & Behavior* 2022;**134**:108857.

Micoulaud-Franchi 2016 *{published and unpublished data}*

Micoulaud-Franchi JA, Bartolomei F, Duncan R, McGonigal A. Evaluating quality of life in epilepsy: the role of screening for adverse drug effects, depression, and anxiety. *Epilepsy & Behavior* 2017;**75**:18-24.

Micoulaud-Franchi JA, Bartolomei F, McGonigal A. Ultra-short screening instruments for major depressive episode and generalized anxiety disorder in epilepsy: the NDDIE-2 and the GAD-SI. *Journal of Affective Disorders* 2017;**210**:237-40.

* Micoulaud-Franchi JA, Lagarde S, Barkate G, Dufournet B, Besancon C, Trébuchon-Da Fonseca A, et al. Rapid detection of generalized anxiety disorder and major depression in epilepsy: validation of the GAD-7 as a complementary tool to the NDDI-E in a French sample. *Epilepsy & Behavior* 2016;**57**:211-6.

Micoulaud-Franchi 2022 *{published data only}*

Micoulaud-Franchi JA, Kotwas I, Arthuis M, Soncin L, Rontani P, Daquin G, et al. Screening for epilepsy-specific anxiety symptoms: French validation of the EASI. *Epilepsy & Behavior* 2022;**128**:108585.

Mughal 2021 *{published data only}*

Mughal AY, Stockton MA, Bui Q, Go V, Ha TV, Pence BW, et al. Validation of screening tools for common mental health disorders in the methadone maintenance population in Hanoi, Vietnam. *BMC Psychiatry* 2021;**21**(1):1-9.

Nath 2018 *{published data only}*

Nath S, Ryan EG, Trevillion K, Bick D, Demilew J, Milgrom J, et al. Prevalence and identification of anxiety disorders in pregnancy: the diagnostic accuracy of the two-item Generalised Anxiety Disorder scale (GAD-2). *BMJ Open* 2018;**8**(9):e023766.

Osorio 2015 *{published data only}*

Osorio FL, Lima MP, Chagas MH. Screening tools for psychiatry disorders in cancer setting: caution when using. *Psychiatry Research* 2015;**229**(3):739-42.

Pranckeviciene 2022 *{published data only}*

Pranckeviciene A, Saudargiene A, Gecaite-Stonciene J, Liaugaudaite V, Griskova-Bulanova I, Simkute D, et al. Validation of the patient health questionnaire-9 and the generalized anxiety disorder-7 in Lithuanian student sample. *PloS One* 2022;**17**(1):e0263027.

Qu 2016 *{published data only}*

Qu S, Cheng YH, Xu LP, Zhang XH, Wang Y, Sheng L, et al. A cross-sectional investigation of anxiety disorder in acute leukemia patients. *Chinese Mental Health Journal* 2016;**30**(12):891-5.

Scott 2019 *{published and unpublished data}*

Scott AJ, Sharpe L, Thayer Z Miller LA, Hunt C, MacCann C, et al. Design and validation of two measures to detect anxiety disorders in epilepsy: The Epilepsy Anxiety Survey Instrument and its brief counterpart. *Epilepsia* 2019;**60**(10):2068-77. [PMID: https://pubmed.ncbi.nlm.nih.gov/31560136/]

Seo 2014 *{published data only}*

Park SP, Seo JG. Validation of the Generalized Anxiety Disorder-7 in people with epilepsy: a MEPSY study. In: 11th European Congress on Epileptology. Stockholm, Sweden. Vol. 2. 2014:226. [DOI: 10.1111/epi.12675]

* Seo JG, Cho YW, Lee SJ, Lee JJ, Kim JE, Moon HJ, et al. Validation of the generalized anxiety disorder-7 in people with epilepsy: a MEPSY study. *Epilepsy & Behavior* 2014;**35**:59-63. [PMID: https://pubmed.ncbi.nlm.nih.gov/24798411/]

Seo 2015 *{published data only}*

Seo JG, Park SP. Validation of the Generalized Anxiety Disorder-7 (GAD-7) and GAD-2 in patients with migraine. *Journal of Headache and Pain* 2015;**16**(1):1-7.

Seo 2017 *{published data only}*

Seo JG, Kim SY, Moon HJ, Do JK, Park SP. Validation of the Patients Health Questionnaire-9 (PHQ-9), PHQ-2, Generalized Anxiety Disorder-7 (GAD-7), and GAD-2 in patients with tension-type headache. *Cephalalgia* 2017;**37 (1 Supplement 1)**:230. [DOI: 10.1177/0333102417719590]

Shih 2022 *{published data only}*

Shih YC, Chou CC, Lu YJ, Yu HY. Reliability and validity of the traditional Chinese version of the GAD-7 in Taiwanese patients with epilepsy. *Journal of the Formosan Medical Association / Taiwan Yi Zhi* 2022;**121**(11):2324-30. [PMID: https://pubmed.ncbi.nlm.nih.gov/35584970/]

Sidik 2012 *{published data only}*

Sidik SM, Arroll B, Goodyear-Smith F. Validation of the GAD-7 (Malay version) among women attending a primary care clinic in Malaysia. *Journal of Primary Health Care* 2012;**4**(1):5-11.

Simning 2012 *{unpublished data only}*

Simning A, van Wijngaarden E, Fisher SG, Richardson TM, Conwell Y. Mental healthcare need and service utilization in older adults living in public housing. *American Journal of Geriatric Psychiatry* 2012;**20**(5):441-51.

Tong 2016 *{published data only}*

Tong X, An D, McGonigal A, Park SP, Zhou D. Validation of the Generalized Anxiety Disorder-7 (GAD-7) among Chinese people with epilepsy. *Epilepsy Research* 2016;**120**:31-6.

Van Heyningen 2018 *{published data only}*

Van Heyningen T, Honikman S, Tomlinson M, Field S, Myer L. Comparison of mental health screening tools for detecting antenatal depression and anxiety disorders in South African women. *PloS One* 2018;**13**(4):e0193697.

Veisy 2021 *{published data only}*

Veisy F, Farahani H, Togha M, Gharaee B, Janani L, Aghebati A. Rapid screening for generalized anxiety disorder in patients with migraine. *Current Journal of Neurology* 2021;**20**(2):102-10. [PMID: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8743176/]

Wild 2014 *{published data only}*

Wild B, Eckl A, Herzog W, Niehoff D, Lechner S, Maatouk I, et al. Assessing generalized anxiety disorder in elderly people using the GAD-7 and GAD-2 scales: results of a validation study. *American Journal of Geriatric Psychiatry* 2014;**22**(10):1029-38.

Zeng 2013 *{published data only}*

Liu H, He YL, Miao JM, Wang JY, Zeng QZ. Validity and reliability of the 12-item Short Form Health Survey in outpatients from traditional Chinese internal department. *Chinese Mental Health Journal* 2016;**30**(1):6-12. [http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc13&NEWS=N&AN=2016-06986-001]

* Zeng QZ, He YL, Liu H, Miao JM, Chen JX, Xu HN, et al. Reliability and validity of Chinese version of the Generalized Anxiety Disorder 7-item (GAD-7) scale in screening anxiety disorders in outpatients from traditional Chinese internal department. *Chinese Mental Health Journal* 2013;**27**(3):163-8.

Zhong 2015 *{published data only}*

Zhong QY, Gelaye B, Zaslavsky AM, Fann JR, Rondon MB, Sánchez Sixto E, et al. Diagnostic validity of the Generalized Anxiety Disorder-7 (GAD-7) among pregnant women. *PloS One* 2015;**10**(4):e0125096.

Zinchuk 2021 *{published data only}*

* Zinchuk M, Kustov G, Pashnin E, Gersamia A, Rider F, Yakovlev A, et al. Validation of the Generalized Anxiety Disorder-7 (GAD-7) in Russian people with epilepsy. *Epilepsy & Behavior* 2021;**123**:108269.

Zinchuk M, Kustov G, Rider F, Pashnin E, Voinova N, Popova S, et al. Psychometric properties of the Russian language version of the GAD-7 in patients with epilepsy. *Journal of the Neurological Sciences* 2021;**429**:119186. [DOI: 10.1016/j.jns.2021.119186]

References to studies excluded from this review

Afshari 2020 *{published data only}*

Afshari B. Personality and anxiety disorders: examination of revised reinforcement sensitivity theory in clinical generalized anxiety disorder, social anxiety disorder, and panic disorder. *Current Issues in Personality Psychology* 2020;**8**(1):52-60.

Ahmad 2017 *{published data only}*

Ahmad S, Hussain S, Shah FS, Akhtar F. Urdu translation and validation of GAD-7: a screening and rating tool for anxiety symptoms in primary health care. *Journal of the Pakistan Medical Association* 2017;**67**(10):1536-40.

AlKhathami 2022 *{published data only}*

AlKhathami AD. An innovative 5-Step Patient Interview approach for integrating mental healthcare into primary care centre services: a validation study. *General Psychiatry* 2022;**35**(4):e100693. [PMID: https://pubmed.ncbi.nlm.nih.gov/36189183/]

Andersen 2023 *{published data only}*

Andersen K, Nielsen B, Hesse M, Mejldal A, Bilberg R, Nielsen AS. Screening of coexisting depression or anxiety disorders among Danish patients with alcohol use disorder in outpatient treatment [Andersen (2023)]. *Nordic Journal of Psychiatry* 2023;**77(6)**:608-16. [DOI: 10.1080/08039488.2023.2200764]

Austin 2022 *{published data only}*

Austin MP, Mule V, Hadzi-Pavlovic D, Reilly N. Screening for anxiety disorders in third trimester pregnancy: a comparison of four brief measures. *Archives of Women's Mental Health* 2022;**25**(2):389-97.

Basaraba 2023 *{published data only}*

Basaraba CN, Stockton MA, Sweetland A, Medina-Marino A, Lovero KL, Oquendo MA, et al. Does it matter what screener we use? A comparison of Ultra-brief PHQ-4 and E-mwTool-3 screeners for anxiety and depression among people with and without HIV. *AIDS and Behavior* 2023;**27(4)**:1154-61. [DOI: 10.1007/s10461-022-03852-w]

Baykan 2019 *{published data only}*

Baykan H, Baykan O, Durmaz O, Kara H, Hisamiogullari AA, Karlidere T. Plasma neuropeptide-S levels in populations diagnosed with generalized anxiety disorder: A controlled study. *Archives of Neuropsychiatry* 2019;**56**(1):52.

Beard 2014 *{published data only}*

Beard C, Björgvinsson T. Beyond generalized anxiety disorder: psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. *Journal of Anxiety Disorders* 2014;**28**(6):547-52.

Beesdo-Baum 2012 *{published data only}*

Beesdo-Baum K, Klotsche J, Knappe S, Craske MG, LeBeau RT, Hoyer J, et al. Psychometric properties of the dimensional anxiety scales for DSM-V in an unselected sample of German treatment seeking patients. *Depression and Anxiety* 2012;**29**(12):1014-24.

Belk 2016 *{published data only}*

Belk RA, Pilling M, Rogers KD, Lovell K, Young A. The theoretical and practical determination of clinical cut-offs for the British Sign Language versions of PHQ-9 and GAD-7. *BMC Psychiatry* 2016;**16**(1):1-12.

Brünahl 2017 *{published data only}*

Brünahl C, Dybowski C, Albrecht R, Riegel B, Höink J, Fisch M, et al. Mental disorders in patients with chronic pelvic pain syndrome (CPPS). *Journal of Psychosomatic Research* 2017;**98**:19-26.

Bunevicius 2013 *{published data only}*

Bunevicius A, Staniute M, Brozaitiene J, Pop VJ, Neverauskas J, Bunevicius R. Screening for anxiety disorders in patients with coronary artery disease. *Health and Quality of Life Outcomes* 2013;**11**(1):1-9.

Bushey 2021 *{published data only}*

Bushey MA, Kroenke K, Baye F, Lourens S. Composite measures of pain, anxiety, and depressive (PAD) symptoms: Construct and predictive validity. *General Hospital Psychiatry* 2021;**72**:1-6.

Byrd-Bredbenner 2021 *{published data only}*

Byrd-Bredbenner C, Eck K, Quick V. GAD-7, GAD-2, and GAD-mini: Psychometric properties and norms of university students in the United States. *General Hospital Psychiatry* 2021;**69**:61-6.

Camarata 2023 *{published data only}*

Camarata MA, Ala A, Coskun AK, Deng Y, Embel VK, Gonzalez-Peralta R, et al. Major depressive disorder in an international multisite Wilson Disease Registry. *Journal of the Academy of Consultation-Liaison Psychiatry* 2023;**64(2)**:106-17. [DOI: 10.1016/j.jaclp.2022.12.001]

Cano-Vindel 2018 *{published data only}*

Cano-Vindel A, Muñoz-Navarro R, Medrano LA, Ruiz-Rodríguez P, González-Blanch C, Gómez-Castillo MD, et al, PsicAP Research Group. A computerized version of the Patient Health Questionnaire-4 as an ultra-brief screening tool to detect emotional disorders in primary care. *Journal of Affective Disorders* 2018;**234**:247-55.

Castro 2023 *{published data only}*

Castro A, Gili M, Visser M, Penninx BW, Brouwer IA, Montano JJ, et al. Soft drinks and symptoms of depression and anxiety in overweight subjects: a longitudinal analysis of an European cohort. *Nutrients* 2023;**15(18)**:3865. [DOI: 10.3390/nu15183865]

Celano 2013 *{published data only}*

Celano CM, Suarez L, Mastromauro C, Januzzi JL, Huffman JC. Feasibility and utility of screening for depression and anxiety disorders in patients with cardiovascular disease. *Circulation: Cardiovascular Quality and Outcomes* 2013;**6**(4):498-504.

Christensen 2011 *{published data only}*

Christensen H, Batterham PJ, Grant JB, Griffiths KM, Mackinnon AJ. A population study comparing screening performance of prototypes for depression and anxiety with standard scales. *BMC Medical Research Methodology* 2011;**11**(1):1-9.

Christensen 2014 *{published data only}*

Christensen H, Batterham P, Mackinnon A, Griffiths KM, Hehir KK, Kenardy J, et al. Prevention of generalized anxiety disorder using a web intervention, iChill: randomized controlled trial. *Journal of Medical Internet Research* 2014;**16**(9):e3507.

Chun 2017 *{published data only}***http://dx.doi.org/10.1177/2396987317705242**

Chun HY, Dennis M, Mead G, Whiteley W, Carson A. The accuracy of 7-item Generalized Anxiety Disorder questionnaire and the Fear Questionnaire in detecting clinically diagnosed generalized anxiety disorder and phobic disorder in stroke. *European Stroke Journal* 2017;**2 (1 Suppl 1)**:366. [DOI: 10.1177/2396987317705242]

Clary 2021 *{published data only}*

Clary HM, Wan M, Conner K, Brenes GA, Kimball J, Kim E, et al. Examining brief and ultra-brief anxiety and depression screening methods in a real-world epilepsy clinic sample. *Epilepsy & Behavior* 2021;**118**:107943.

Dadi 2016 *{published data only}*

Dadi AF, Dachew BA, Kisi T, Yigzaw N, Azale T. Anxiety and associated factors among prisoners in North West of Amhara Regional State, Ethiopia. *BMC Psychiatry* 2016;**16**(1):1-6.

Dear 2011 *{published data only}*

Dear BF, Titov N, Sunderland M, McMillan D, Anderson T, Lorian C, et al. Psychometric comparison of the Generalized Anxiety Disorder scale-7 and the Penn State Worry Questionnaire for measuring response during treatment of generalised anxiety disorder. *Cognitive Behaviour Therapy* 2011;**40**(3):216-27.

Donker 2011 *{published data only}*

Donker T, van Straten A, Marks I, Cuijpers P. Quick and easy self-rating of Generalized Anxiety Disorder: validity of the Dutch web-based GAD-7, GAD-2 and GAD-SI. *Psychiatry Research* 2011;**188**(1):58-64.

Dyukova 2021 *{published data only}*

Dyukova GM, Kryukov AI, Makarov SA, Guseva AL, Olimpieva SP. A method for prediction functional dizziness after benign paroxysmal positional vertigo. *Zhurnal Nevrologii i Psikhiatrii Imeni SS Korsakova* 2021;**121**(5):120-5.

Fairbrother 2019 *{published data only}*

Fairbrother N, Corbyn B, Thordarson DS, Ma A, Surm D. Screening for perinatal anxiety disorders: room to grow. *Journal of Affective Disorders* 2019;**250**:363-70.

Fatima 2021 *{published data only}*

Fatima N, Chinnakali P, Rajaa S, Menon V, Mondal N, Chandrasekaran V. Prevalence of depression and anxiety among mothers of children with neuro-developmental disorders at a tertiary care centre, Puducherry. *Clinical Epidemiology and Global Health* 2021;**11**:100792.

García-Campayo 2009 *{published data only}*

García-Campayo J, Zamorano E, Ruíz MA, Pardo A, Freire O, Pérez-Páramo M, et al. Cultural adaptation into Spanish of the Generalized Anxiety Disorder scale-7 (GAD-7) scale. *European Psychiatry* 2009;**24**(S1):1-1.

García-Campayo 2010 *{published data only}*

García-Campayo J, Zamorano E, Ruíz MA, Pardo A, Pérez-Péramo M, López-Gómez V, et al. Cultural adaptation into Spanish of the Generalized Anxiety Disorder-7 (GAD-7) scale as a screening tool. *Health and Quality of Life Outcomes* 2010;**8**(1):1-11.

García-Campayo 2012a *{published data only}*

García-Campayo J, Zamorano E, Ruíz MA, Pérez-Páramo M, López-Gómez V, Rejas J. The assessment of generalized anxiety disorder: psychometric validation of the Spanish version of the self-administered GAD-2 scale in daily medical practice. *Health and Quality of Life Outcomes* 2012;**10**(1):1-10.

García-Campayo 2012b *{published data only}*

García-Campayo J, Caballero F, Pérez M, López V. Pain related factors in newly diagnosed Generalized Anxiety Disorder

patients. *Actas Españolas de Psiquiatría.* 2012;**40**(4):177-86. [PMID: https://pubmed.ncbi.nlm.nih.gov/22851478/]

García-Campayo 2012c *{published data only}*

García-Campayo J, Zamorano E, Ruíz MA, Pardo A, Pérez-Páramo M, López-Gómez V, et al. P-149 - Psychometric validation of the Spanish version of the GAD-2 scale for screening generalized anxiety disorder. *European Psychiatry* 2012;**27**:1.

García-Campayo 2012d *{published data only}*

García-Campayo J, Caballero F, Pérez M, Lopez V. Prevalence and clinical features of newly diagnosed generalized anxiety disorder patients in Spanish primary care settings: the GADAP study. *Actas Españolas de Psiquiatría* 2012;**40**(3):105-13.

Giuliani 2021 *{published data only}*

Giuliani M, Gorini A, Barbieri S, Veglia F, Tremoli E. Examination of the best cut-off points of PHQ-2 and GAD-2 for detecting depression and anxiety in Italian cardiovascular inpatients. *Psychology & Health* 2021;**36**(9):1088-101.

Goldstein 2023 *{published data only}*

Goldstein LH, Vitoratou S, Stone J, Chalder T, Baldellou Lopez, M, et al. Performance of the GAD-7 in adults with dissociative seizures. *Seizure* 2023;**104**:15-21. [DOI: 10.1016/j.seizure.2022.11.011]

Gong 2021 *{published data only}*

Gong Y, Zhou H, Zhang Y, Zhu X, Wang X, Shen B, et al. Validation of the 7-item Generalized Anxiety Disorder scale (GAD-7) as a screening tool for anxiety among pregnant Chinese women. *Journal of Affective Disorders* 2021;**282**:98-103.

Graham 2019 *{published data only}*

Graham AK, Minc A, Staab E, Beiser DG, Gibbons RD, Laiteerapong N. Validation of the computerized adaptive test for mental health in primary care. *Annals of Family Medicine* 2019;**17**(1):23-30.

Greenhalgh 2016 *{published data only}*

Greenhalgh T, Mitchell AJ, Symonds P. Which self-report anxiety scale is optimal for measuring cancer related anxiety? A comparison of GAD7, GAD-DSMIV, and ET Anxiety Thermometer vs MINI semi-structured interview. *Psycho-Oncology* 2016;**25**:150-1. [DOI: 10.1002/pon.4272]

Gul 2015 *{published data only}*

Gül AI, Simsek G, Karaaslan Ö, Inanir S. Comparison of automatical thoughts among generalized anxiety disorder, major depressive disorder and generalized social phobia patients. *European Review for Medical and Pharmacological Sciences* 2015;**19**(15):2916-21.

Gulpers 2022 *{published data only}*

Gulpers BJ, Verhey FR, Eussen SJ, Schram MT, de Galan BE, van Boxtel MP, et al. Anxiety and cognitive functioning in the Maastricht study: A cross-sectional population study. *Journal of Affective Disorders* 2022;**319**:570-9. [PMID: https://pubmed.ncbi.nlm.nih.gov/36162695/]

Gündüz 2021 *{published data only}*

Gündüz, Gündoğmuş I, Sertçelik S, Engin BH, İşler A, Yaşar AB, et al. Validity and reliability of the Turkish version of Generalized Anxiety Disorder scale-Revised. *Psychiatry Investigation* 2021;**18**(10):949.

Hallgren 2007 *{published data only}*

Hallgren JD, Morton JR. What's the best way to screen for anxiety and panic disorders? *Journal of Familiy Practice* 2007;**56**(7):579-80. [PMID: https://pubmed.ncbi.nlm.nih.gov/17605953/]

Henderson 2014 *{published data only}*

Henderson LC, Antony MM, Koerner N. Psychometric properties of the Generalized Anxiety Disorder inventory in a Canadian sample. *Journal of Psychopharmacology (Oxford, England)* 2014;**28**(5):440-8.

Hinz 2017 *{published data only}*

Hinz A, Klein AM, Brähler E, Glaesmer H, Luck T, Riedel-Heller SG, et al. Psychometric evaluation of the Generalized Anxiety Disorder Screener GAD-7, based on a large German general population sample. *Journal of Affective Disorders* 2017;**210**:338-44.

Holper 2021 *{published data only}*

Holper S, Foster E, Lloyd M, Rayner G, Rychkova M, Ali R, et al. Clinical predictors of discordance between screening tests and psychiatric assessment for depressive and anxiety disorders among patients being evaluated for seizure disorders. *Epilepsia* 2021;**62**(5):1170-83.

Houston 2011 *{published data only}*

Houston JP, Kroenke K, Faries DE, Doebbeling CC, Adler LA, Ahl J, et al. A provisional screening instrument for four common mental disorders in adult primary care patients. *Psychosomatics* 2011;**52**(1):48-55.

Hughes 2018 *{published data only}*

Hughes J, Dunn KM, Chaffee T, Bhattarai JJ, Beier M. Diagnostic and clinical utility of the GAD-2 for screening anxiety symptoms in individuals with multiple sclerosis. *Archives of Physical Medicine and Rehabilitation* 2018;**99**(10):2045-9.

Jansen 2019 *{published data only}*

Jansen C, Francomme L, Vignal JP, Jacquot C, Schwan R, Tyvaert L, et al. Interictal psychiatric comorbidities of drug-resistant focal epilepsy: prevalence and influence of the localization of the epilepsy. *Epilepsy & Behavior* 2019;**94**:288-96.

Jin 2020 *{published data only}*

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

Jyothi 2020 *{published data only}*

Jyothi KS, Kannan GK, Viswanathan S, Ranganathan S, Menon J, Ramanathan S. Antenatal depression and generalized anxiety disorder in a tertiary hospital in South India. *Indian Journal of Psychological Medicine* 2020;**42**(6):513-8.

Kemp 2023 *{published data only}*

Kemp J, Chenji S, MacMaster F, Bray S, Kopala-Sibley DC. Associations between parental depression and anxiety symptom severity and their offspring's cortical thickness and subcortical volume. *Journal of Psychiatric Research* 2023;**166**:139-46. [DOI: 10.1016/j.jpsychires.2023.09.014]

Kertz 2013 *{published data only}*

Kertz S, Bigda-Peyton J, Bjorgvinsson T. Validity of the Generalized Anxiety Disorder-7 Scale in an acute psychiatric sample. *Clinical Psychology & Psychotherapy* 2013;**20**(5):456-64.

Khani 2017 *{published data only}*

Khani M, Dadashi M. Comparison of brain/behavioral system activity, intolerance of uncertainty and cognitive avoidance among individuals with generalized anxiety disorder and normal subjects. *Journal of Advances in Medical and Biomedical Research* 2017;**25**(111):67-81.

Kim 2023 *{published data only}*

Kim HJ, Kim S, Son Y, Youn I, Lee K. Reliability and validity of the Korean version of the Ask Suicide-Screening Questions (ASQ). *Journal of Korean Medical Science* 2023;**38(6)**:e41. [DOI: 10.3346/jkms.2023.38.e41]

Konkan 2013 *{published data only}*

Konkan R, Senormanci O, Guclu O, Aydin E, Sungur MZ. Validity and reliability study for the Turkish adaptation of the Generalized Anxiety Disorder-7 (GAD-7) scale. *Archives of Neuropsychiatry* 2013;**50**:53-8. [DOI: 10.4274/npa.y6308]

Konstantakopoulos 2013 *{published data only}*

Konstantakopoulos G, Sofianopoulou E, Touloumi G, Ploumpidis D. Ultra-short questionnaires for the detection of depression and anxiety. *Psychiatrike* 2013;**24**(4):288-97.

Kounali 2020 *{published data only}*

Kounali D, Button KS, Lewis G, Gilbody S, Kessler D, Araya R, et al. How much change is enough? Evidence from a longitudinal study on depression in UK primary care. *Psychological Medicine* 2020;**52**(10):1875-82. [PMID: https://pubmed.ncbi.nlm.nih.gov/33138872/]

Kroenke 2010 *{published data only}*

Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General Hospital Psychiatry* 2010;**32**(4):345-59.

Kumar 2018 *{published data only}*

Kumar V, Avasthi A, Grover S. Somatosensory amplification, health anxiety, and alexithymia in generalized anxiety disorder. *Industrial Psychiatry Journal* 2018;**27**(1):47.

Ladouceur 1999 *{published data only}*

Ladouceur R, Dugas MJ, Freeston MH, Rhéaume J, Blais F, Boisvert JM, et al. Specificity of generalized anxiety disorder symptoms and processes. *Behavior Therapy* 1999;**30**(2):191-207.

Lan 2021 *{published data only}*

Lan H, Jamil F, Al Kaabi N, Aser R, Gyatso K, Habbal H, et al. Two-step screening for anxiety symptoms in solid organ transplant recipients. In: American Transplant Congress Meeting Abstracts; 2021 June 4-9; virtual meeting. Vol. 21. 2021:592. [ATC: https://atcmeetingabstracts.com/abstract/two-step-screening-for-anxiety-symptoms-in-solid-organ-transplant-recipients/]

Lee 2020 *{published data only}*

Lee SH, Shin C, Kim H, Jeon SW, Yoon HK, Ko YH, et al. Validation of the Korean version of the Generalized Anxiety Disorder 7 self-rating scale. *Asian Pacific Psychiatry* 2020;**14**(1):e12421. [DOI: 10.1111/appy.12421]

Levitt 2021 *{published data only}*

Levitt EE, Syan SK, Sousa S, Costello MJ, Rush B, Samokhvalov AV, et al. Optimizing screening for depression, anxiety disorders, and post-traumatic stress disorder in inpatient addiction treatment: A preliminary investigation. *Addictive Behaviors* 2021;**112**:106649.

Li 2022 *{published data only}*

Li Y, Jin S, Li Y, Guo F, Luo T, Pan B, et al. The effects of research activities on biomedical students' mental health: A national cross-sectional study. *Frontiers in Psychiatry* 2022;**13**:796697. [DOI: 10.3389/fpsyt.2022.796697]

Lin 2021 *{published data only}*

Lin Q, Bonkano O, Wu K, Liu Q, Ibrahim TA, Liu L. The value of Chinese version GAD-7 and PHQ-9 to screen anxiety and depression in Chinese outpatients with atypical chest pain. *Therapeutics and Clinical Risk Management* 2021;**17**:423.

Liu 2023 *{published data only}*

Liu M, Wang D, Fang J, Chang Y, Hu Y, Huang K. Validation of the Generalized Anxiety Disorder-7 in patients with COPD: a cross-sectional study. *BMC Psychiatry* 2023;**23**(1):593. [DOI: 10.1186/s12888-023-05072-5]

Lowe 2008 *{published data only}*

Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care* 2008;**46**(3):266-74. [PMID: https://pubmed.ncbi.nlm.nih.gov/18388841/]

Ma 2021 *{published data only}*

Ma M, Shi Z, Wu H, Ma X. Clinical implications of panic attack in Chinese patients with somatoform disorders. *Journal of Psychosomatic Research* 2021;**146**:110509.

Mansbach 2015 *{published data only}*

Mansbach WE, Mace RA, Clark KM. The Brief Anxiety and Depression Scale (BADS): a new instrument for detecting anxiety and depression in long-term care residents. *International Psychogeriatrics* 2015;**27**(4):673-81.

Means-Christensen 2006 *{published data only}*

Means-Christensen AJ, Sherbourne CD, Roy-Byrne PP, Craske MG, Stein MB. Using five questions to screen for five

common mental disorders in primary care: diagnostic accuracy of the Anxiety and Depression Detector. *General Hospital Psychiatry* 2006;**28**(2):108-18.

Minc 2017 *{published data only}*

Minc A, Kass AE, Laiteerapong N, Gibbons R. Validation of the computerized adaptive test for mental health in primary care. In: Journal of General Internal Medicine. Vol. 32. Washington (DC): Springer, 2017 April 19-22:S373-4. [https://link.springer.com/article/10.1007/s11606-017-4028-8]

Monteiro 2020 *{published data only}*

Monteiro RP, Nascimento BS, Monteiro TM, Da Silva PD, Ferreira AJ. Psychometric evidence of the 7-item Generalized Anxiety Disorder questionnaire in Brazil. *International Journal of Mental Health and Addiction* 2022;**20**:1023-34. [DOI: 10.1007/s11469-020-00423-9]

Monterrosa-Blanco 2021 *{published data only}*

Monterrosa-Blanco A, Cassiani-Miranda CA, Scoppetta O, Monterrosa-Castro A. Generalized anxiety disorder scale (GAD-7) has adequate psychometric properties in Colombian general practitioners during COVID-19 pandemic. *General Hospital Psychiatry* 2021;**70**:147.

Munoz-Navarro 2017a *{published data only}*

Munoz-Navarro R, Cano-Vindel A, Ruiz-Rodriguez P, Adrian ML, Gonzalez-Blanch C, Antonio MJ, et al. Diagnostic and referral hierarchical model of common mental disorders in primary care centers. The PsicAP clinical trial approach. *Ansiedad Y Estres-Anxiety and Stress* 2017;**23**(2-3):124-9.

Munoz-Navarro 2017b *{published data only}*

Cano-Vindel A, Munoz-Navarro R, Medrano LA, Ruiz-Rodriguez P, Gonzalez-Blanch C, Gomez-Castillo MD, et al. A computerized version of the Patient Health Questionnaire-4 as an ultra-brief screening tool to detect emotional disorders in primary care. *Journal of Affectective Disorders* 2018;**234**:247-55. [PMID: https://pubmed.ncbi.nlm.nih.gov/29549826/]

* Munoz-Navarro R, Cano-Vindel A, Moriana JA, Medrano LA, Ruiz-Rodríguez P, Agüero-Gento L, et al. Screening for generalized anxiety disorder in Spanish primary care centers with the GAD-7. *Psychiatry Research* 2017;**256**:312-7.

Muntingh 2013 *{published data only}*

Muntingh AD, De Heer EW, Van Marwijk HW, Adèr HJ, Van Balkom AJ, Spinhoven P, et al. Screening high-risk patients and assisting in diagnosing anxiety in primary care: the Patient Health Questionnaire evaluated. *BMC Psychiatry* 2013;**13**(1):1-9.

Mussell 2008 *{published data only}*

Mussell M, Kroenke K, Spitzer RL, Williams JB, Herzog W, Löwe B. Gastrointestinal symptoms in primary care: prevalence and association with depression and anxiety. *Journal of Psychosomatic Research* 2008;**64**(6):605-12.

Nath 2019 *{published data only}***http://dx.doi.org/10.1007/s00737-019-00996-y**

Nath S, Ryan EG, Trevillion K, Bick D, Demilew J, Milgrom J, et al. The effectiveness of the two-item Generalized Anxiety Disorder Scale (GAD-2) in identifying anxiety disorders in early pregnancy. *Archives of Women's Mental Health* 2019;**22(5)**:659.

Nath 2020 *{published data only}*

Nath S, Busuulwa P, Ryan EG, Challacombe FL, Howard LM. The characteristics and prevalence of phobias in pregnancy. *Midwifery* 2020;**82**:102590.

Priyanka 2010 *{published data only}*

Priyanka P, Boyle LL, Tu XM, Conwell Y. Inter-rater reliability and validity of the PHQ-9 and GAD-7 to identify depression and anxiety in older adults receiving aging services care management. *American Journal of Geriatric Psychiatry* 2010;**18**:S113-4.

Purves 2019 *{published data only}*

Purves K, Coleman J, Gaspar HA, Breen G, Eley T. Genome-wide association study of generalized anxiety disorder symptoms and self-reported anxiety disorder: an investigation of unique and shared genomic contributions. *European Neuropsychopharmacology* 2019;**29**:S775-6.

Purvis 2019 *{published data only}*

Purvis TE, Neuman BJ, Riley LH, Skolasky RL. Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression among patients presenting for spine surgery. *Journal of Neurosurgery: Spine* 2019;**30**(4):524-31.

Qu 2015 *{published data only}*

Qu S, Sheng L. Diagnostic test of screening generalized anxiety disorders in general hospital psychological department with GAD-7. *Chinese Mental Health Journal* 2015;**29**(12):939-44.

Rady 2021 *{published data only}*

Rady A, Alamrawy R, Ramadan I, Abd El Raouf M. Does alexithymia, independent of depressive and anxiety disorders, correlate with the severity of somatic manifestations among patients with medically unexplained physical symptoms? *Journal of Experimental Psychopathology* 2021;**12**(4):20438087211043729.

Ringoir 2014 *{published data only}*

Ringoir L, Pedersen SS, Widdershoven JW, Pop VJ. Prevalence of psychological distress in elderly hypertension patients in primary care. *Netherlands Heart Journal* 2014;**22**(2):71-6.

Rogers 2021 *{published data only}*

Rogers MA, Appaneal RN, Hughes D, Vlahovich Ne, Waddington G, Burke LM, et al. Prevalence of impaired physiological function consistent with Relative Energy Deficiency in Sport (RED-S): an Australian elite and pre-elite cohort. *British Journal of Sports Medicine* 2021;**55**(1):38-45.

Ruiz 2009 *{published data only}*

Ruiz MA, Zamorano E, García-Campayo J, Pardo A, Freire O, Rejas J. PMH59 Further validity evidence of the Generalized Anxiety Disorder-7 scale (GAD-7). *Value in Health* 2009;**7**(12):A362.

Sawaya 2016 *{published data only}*

Sawaya H, Atoui M, Hamadeh A, Zeinoun P, Nahas Z. Adaptation and initial validation of the Patient Health Questionnaire-9 (PHQ-9) and the Generalized Anxiety Disorder-7 Questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample. *Psychiatry Research* 2016;**239**:245-52.

Short 2019 *{published data only}*

Short H, Al Sayah F, Ohinmaa A, Johnson JA. Screening for anxiety and depressive symptoms using the EQ-5D-3L post-hospital discharge. *This is Public Health Week* 2019;**28**:S181. [UNIVERSITY OF ALBERTA: https://www.ualberta.ca/public-health/about/academic-events/this-is-public-health-week/program/poster-presentations/short-h.html]

Shrestha 2020 *{published data only}*

Shrestha S, Ramos K, Fletcher TL, Kraus-Schuman C, Stanley MA, Ramsey D, et al. Psychometric properties of worry and anxiety measures in a sample of African American and Caucasian older adults. *Aging & Mental Health* 2020;**24**(2):315-21.

Silva 2018 *{published data only}*

Silva MT, Roa MC, Martins SS, Da Silva AT, Galvao TF. Generalized anxiety disorder and associated factors in adults in the Amazon, Brazil: A population-based study. *Journal of Affective Disorders* 2018;**236**:180-6.

Simpson 2014 *{published data only}*

Simpson W, Glazer M, Michalski N, Steiner M, Frey BN. Comparative efficacy of the Generalized Anxiety Disorder 7-item scale and the Edinburgh Postnatal Depression Scale as screening tools for generalized anxiety disorder in pregnancy and the postpartum period. *Canadian Journal of Psychiatry* 2014;**59**(8):434-40.

Sinesi 2019 *{published data only}*

Sinesi A, Cheyne H, Maxwell M, O'Carroll R. The development and initial psychometric validation of a screening scale for antenatal anxiety. *Journal of Reproductive and Infant Psychology* 2019;**37**:e1-e48. [DOI: 10.1080/02646838.2019.1673570]

Skapinakis 2007 *{published data only}*

Skapinakis P. The 2-item Generalized Anxiety Disorder scale had high sensitivity and specificity for detecting GAD in primary care. *Evidence-Based Medicine* 2007;**12**(5):149. [PMID: https://pubmed.ncbi.nlm.nih.gov/17909240/]

Snijkers 2021 *{published data only}*

Snijkers JT, Van Den Oever W, Weerts ZZ, Vork L, Mujagic Z, Leue C, et al. Examining the optimal cutoff values of HADS, PHQ-9 and GAD-7 as screening instruments for depression and anxiety in irritable bowel syndrome. *Neurogastroenterology and Motility* 2021;**33**(12):e14161.

Soni 2019 *{published data only}*

Soni A, Fahey N, Prabhakaran A, Varma J, Raithatha N, Byatt N, et al. How and when to screen for symptoms of postpartum depression: results from a prospective cohort of pregnant rural Indian women. *Archives of Women's Mental Health* 2019;**22**:678. [https://link.springer.com/content/pdf/10.1007/s00737-019-00996-y.pdf]

Sousa 2015 *{published data only}*

Sousa TV, Viveiros V, Chai MV, Vicente FL, Jesus G, Carnot MJ, et al. Reliability and validity of the Portuguese version of the Generalized Anxiety Disorder (GAD-7) scale. *Health and Quality of Life Outcomes* 2015;**13**(1):1-8.

Staples 2019 *{published data only}*

Staples LG, Dear BF, Gandy M, Fogliati V, Fogliati R, Karin E, et al. Psychometric properties and clinical utility of brief measures of depression, anxiety, and general distress: The PHQ-2, GAD-2, and K-6. *General Hospital Psychiatry* 2019;**56**:13-8.

Swinson 2006 *{published data only}*

Swinson RP. The GAD-7 scale was accurate for diagnosing generalised anxiety disorder. *Evidence-based Medicine* 2006;**11**(6):184.

Tang 2019 *{published data only}*

Tang E, Dano S, Ayub A, Li V, Escudero I, Ullah N, et al. Clinical utility of the Edmonton Symptom Assessment System to screen for depression and anxiety in kidney transplant recipients. In: American Transplant Congress Meeting Abstracts; 2019 June 1-5; Boston (MA). Vol. 19. 2019:927-8. [ATC: https://atcmeetingabstracts.com/abstract/clinical-utility-of-the-edmonton-symptom-assessment-system-to-screen-for-depression-and-anxiety-in-kidney-transplant-recipients/]

Terluin 2012 *{published data only}*

Terluin B. Malaysian GAD-7 less sensitive than reported. *Journal of Primary Health Care* 2012;**4**(3):262.

Tonev 2019 *{published data only}*

Tonev I, Conev NV, Valev SS, Panayotova T, Karanikolova T, Donev I. Fast time estimation to predict levels of anxiety in cancer patients. Journal of Clinical Oncology 2019;**37**(15 suppl):e23162-e23162. [DOI: 10.1200/JCO.2019.37.15_suppl.e23162]

Tran 2020 *{published data only}*

Tran AG. Using the GAD-7 and GAD-2 generalized anxiety disorder screeners with student-athletes: empirical and clinical perspectives. *Sport Psychologist* 2020;**34**(4):300-9. [DOI: 10.1123/tsp.2020-0028]

Van Dooren 2016 *{published data only}*

Van Dooren FE, Denollet J, Verhey FR, Stehouwer CD, Sep SJ, Henry R, et al. Psychological and personality factors in type 2 diabetes mellitus, presenting the rationale and exploratory results from The Maastricht Study, a population-based cohort study. *BMC Psychiatry* 2016;**16**(1):1-11.

Vasiliadis 2015 *{published data only}*

Vasiliadis HM, Chudzinski V, Gontijo-Guerra S, Préville M. Screening instruments for a population of older adults: The 10-item Kessler Psychological Distress Scale (K10) and the 7-item Generalized Anxiety Disorder Scale (GAD-7). *Psychiatry Research* 2015;**228**(1):89-94.

Villarreal-Zegarra 2023 *{published data only}*

Villarreal-Zegarra D, Barrera-Begazo J, Otazu-Alfaro S, Mayo-Puchoc N, Bazo-Alvarez JC, Huarcaya-Victoria J. Sensitivity and specificity of the Patient Health Questionnaire (PHQ-9, PHQ-8, PHQ-2) and General Anxiety Disorder scale (GAD-7, GAD-2) for depression and anxiety diagnosis: A cross-sectional study in a Peruvian hospital population. *BMJ Open* 2023;**13(9)**:e076193. [DOI: 10.1136/bmjopen-2023-076193]

Wang 2021 *{published data only}*

Wang YY, Yan JC, Li CY, Zhong L, Sun Y, Fu LL. Development and preliminary validation of a self-rating anxiety inventory for maintenance haemodialysis patients. *Psychology, Health & Medicine* 2021;**27**(7):1482-94. [DOI: 10.1080/13548506.2021.1890159]

Webb 2008 *{published data only}*

Webb SA, Diefenbach G, Wagener P, Novy DM, Kunik M, Rhoades HM, et al. Comparison of self-report measures for identifying late-life generalized anxiety in primary care. *Journal of Geriatric Psychiatry and Neurology* 2008;**21**(4):223-31.

Wiltink 2017 *{published data only}*

Wiltink J, Kliem S, Michal M, Subic-Wrana C, Reiner I, Beutel ME, et al. Mini-Social Phobia INventory (mini-SPIN): psychometric properties and population based norms of the German version. *BMC Psychiatry* 2017;**17**(1):1-10.

Zachar-Tirado 2021 *{published data only}*

Zachar-Tirado CN, Donders J. Clinical utility of the GAD-7 in identifying anxiety disorders after traumatic brain injury. *Brain Injury* 2021;**35**(6):655-60.

Zhang 2021 *{published data only}*

Zhang C, Wang T, Zeng P, Zhao M, Zhang G, Zhai S, et al. Reliability, validity, and measurement invariance of the General Anxiety Disorder scale among Chinese medical university students. *Frontiers in Psychiatry* 2021;**12**:648755.

Zinchuk 2022 *{unpublished data only}*

Zinchuk M, Kustov G, Pashnin E, Gersamia A, Rider F, Voinova N, et al. Not always that EASI: validating the Russian version of the Epilepsy Anxiety Survey Instrument and its brief counterpart. *Epilepsy & Behavior* 2022;**133**:108801. [DOI: 10.1016/j.yebeh.2022.108801]

Zirke 2013 *{published data only}*

Zirke N, Seydel C, Arsoy D, Klapp BF, Haupt H, Szczepek AJ, et al. Analysis of mental disorders in tinnitus patients performed with Composite International Diagnostic Interview. *Quality of Life Research* 2013;**22**(8):2095-104.

References to studies awaiting assessment

Burger 2023 *{published data only}*

Burger JW, Andersen LS, Joska JA. Baseline concussion assessments can identify mental disorders: SCAT-5 and other screening tools in South African club rugby. *Physician and Sportsmedicine* 2023;**51**(5):472-81.

References to ongoing studies

Aslan 2023 *{published data only}*

Aslan S, Nyundo A. Incidence and predictors of neuropsychiatric manifestations following a traumatic brain injury in referral hospitals in Dodoma, Tanzania: a protocol of a prospective longitudinal observational study. *medRxiv* 2023;**21**:(no pagination). [DOI: 10.1101/2023.08.20.23294327]

Additional references

Alonso 2007

Alonso J, Lépine JP. Overview of key data from the European Study of the Epidemiology of Mental Disorders (ESEMeD). *Journal of Clinical Psychiatry* 2007;**68**(Suppl 2):3-9.

APA 2000

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR). Washington, DC: American Psychiatric Association, 2000.

APA 2013

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Arlington, VA: American Psychiatric Association, 2013. [DOI: 10.1176/appi.books.9780890425596]

Bandelow 2015

Bandelow B, Michaelis S. Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neuroscience* 2015;**17**(3):327-35. [PMID: 26487813]

Baxter 2014

Baxter AJ, Vos T, Scott KM, Ferrari AJ, Whiteford HA. The global burden of anxiety disorders in 2010. *Psychological Medicine* 2014;**44**(11):2363-74.

Beck 1988

Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. *Journal of Consulting and Clinical Psychology* 1988;**56**(6):893-7.

Beck 1993

Beck AT, Steer RA. Beck Anxiety Inventory Manual. San Antonio, TX: Psychological Corporation, 1993.

Benjamin 2011

Benjamin S, Herr NR, McDuffie J, Nagi A, Williams Jr JW. Performance Characteristics of Self-Report Instruments for Diagnosing Generalized Anxiety and Panic Disorders in Primary Care: A Systematic Review. Washington (DC): Department of Veterans Affairs (US), 2011.

Bossuyt 2015

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ (Clinical Research Ed.)* 2015;**351**:h5527. [PMID: 26511519/]

Brehaut 2022

Brehaut E, Neupane D, Levis B, Wu Y, Sun Y, Ioannidis JP, et al. 'Optimal' cutoff selection in studies of depression screening tool accuracy using the PHQ-9, EPDS, or HADS-D: a meta-research study. *International Journal of Methods in Psychiatric Research* 2022;**31**:e1956. [DOI: 10.1002/mpr.1956]

Chu 2006

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *Journal of Clinical Epidemiology* 2006;**59**(12):1331.

Cohen 2016

Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;**6**(11):e012799.

Craske 2017

Craske MG, Stein MB, Eley TC, Milad MR, Holmes A, Rapee RM, et al. Anxiety disorders. *Nature Reviews Disease Primers* 2017;**3**(1):17024.

Creighton 2018

Creighton AS, Davison TE, Kissane DW. The assessment of anxiety in aged care residents: a systematic review of the psychometric properties of commonly used measures. *International Psychogeriatrics* 2018;**30**(7):967-79.

Crow 2023

Crow J, Smith A. National Clinical Guideline for Stroke for the United Kingdom and Ireland: Part I - An overview of the updated recommendations. *British Journal of Occupational Therapy* 2023;**86**(10):661-4.

Davenport 2014

Davenport CF, Leeflang MM, Takwoingi Y, Deeks JJ. Use of QUADAS-2. Lesson 6.3: Cochrane Collaboration DTA Online Learning Materials. The Cochrane Collaboration September 2012 (available at http://training.cochrane.org).

De Hond 2022

De Hond AA, Steyerberg EW, Van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digital Health* 2022;**4**:e853-e855.

Deeks 2005

Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *Journal of Clinical Epidemiology* 2005;**58**(9):882-93.

Deeks 2023

Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.

Endicott 1978

Endicott J, Spitzer RL. A diagnostic interview: the schedule for affective disorders and schizophrenia. *Archives of General Psychiatry* 1978;**35**(7):837-44.

First 1996

First MB, Spitzer RL, Gibbon M, Williams JB. Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I), Clinician's Version. Arlington, VA: American Psychiatric Association, 1996.

First 2015

First MB, Williams JB, Karg RS, Spitzer RL. Structured Clinical Interview for DSM-5 - Research Version (SCID-5 for DSM-5, Research Version; SCID-5-RV, Version 1.0.0). Arlington, VA: American Psychiatric Association, 2015.

Freeman 2019

Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. *BMC Medical Research Methodology* 2019;**19**(81):1-11. [DOI: 10.1186/s12874-019-0724-x]

Gregory 2020

Gregory KD, Chelmow D, Nelson HD, Van Niel MS, Conry JA, Garcia F, et al. Screening for anxiety in adolescent and adult women: a recommendation from the Women's Preventive Services Initiative. *Annals of Internal Medicine* 2020;**173**(1):48-56.

Gudmundsson 2005

Gudmundsson G, Gislason T, Janson C, Lindberg E, Hallin R, Ulrik CS, et al. Risk factors for rehospitalisation in COPD: role of health status, anxiety and depression. *European Respiratory Journal* 2005;**26**(3):414-9.

Hartzes 2019

Hartzes AM, Morgan CJ. Meta-analysis for diagnostic tests. *Journal of Nuclear Cardiology* 2019;**26**(1):68-71.

Hoffman 2008

Hoffman DL, Dukes EM, Wittchen HU. Human and economic burden of generalized anxiety disorder. *Depression and Anxiety* 2008;**25**(1):72-90.

Howell 2015

Howell D, Keshavarz H, Esplen MJ, Hack T, Hamel M, Howes J, et al. Pan-Canadian Practice Guideline: Screening, Assessment and Management of Psychosocial Distress, Depression and Anxiety in Adults with Cancer; July 2015. Canadian Association of Psychosocial Oncology; Canadian Partnership Against Cancer 2015. [URL: https://www.capo.ca/guidelines]

Katon 2007

Katon W, Roy-Byrne P. Anxiety disorders: efficient screening is the first step in improving outcomes. *Annals of Internal Medicine* 2007;**146**(5):390-2.

Kessler 2001

Kessler RC, Keller MB, Wittchen HU. The epidemiology of generalized anxiety disorder. *Psychiatric Clinics of North America* 2001;**24**(1):19-39.

Kessler 2012a

Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen HU. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research* 2012;**21**(3):169-84.

Kessler 2012b

Kessler RC, Avenevoli S, Costello EJ, Georgiades K, Green JG, Gruber MJ, et al. Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. *Archives of General Psychiatry* 2012;**69**(4):372-80.

Knottnerus 1992

Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *Journal of Clinical Epidemiology* 1992;**45**:1143-54.

Kohn 2022

Kohn MA. Studies of diagnostic test accuracy: partial verification bias and test result-based sampling. *Journal of Clinical Epidemiology* 2022;**145**:179-82. [PMID: https://pubmed.ncbi.nlm.nih.gov/35124189/]

Kroenke 2007

Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Annals of Internal Medicine* 2007;**146**(5):317-25.

Lang 2002

Lang AJ, Stein MB. Screening for anxiety in primary care: why bother? *General Hospital Psychiatry* 2002;**24**(6):365-6.

Lazor 2017

Lazor T, Tigelaar L, Pole JD, De Souza C, Tomlinson D, Sung L. Instruments to measure anxiety in children, adolescents, and young adults with cancer: a systematic review. *Supportive Care in Cancer* 2017;**25**(9):2921-31.

Leeflang 2008a

Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine* 2008;**149**(12):889-97.

Leeflang 2008b

Leeflang MM, Moons KG, Reisma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical Chemistry* 2008;**54**(4):729-37. [DOI: 10.1373/clinchem.2007.096032]

Leeflang 2009

Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology* 2009;**62**(1):5-12.

Levis 2019

Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ (Clinical Research Ed.)* 2019;**365**:1476. [PMID: https://pubmed.ncbi.nlm.nih.gov/30967483/]

Lewis 1992

Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychological Medicine* 1992;**22**(2):465-86.

Lin 2007

Lin KW, Duane MR. Are some screening tests doing more harm than good? American Family Physician 2007;**76**(3):351-2. [PMID: 17708134]

Litster 2016

Litster B, Fiest KM, Patten SB, Fisk JD, Walker JR, Graff LA, et al. Screening tools for anxiety in people with multiple sclerosis: a systematic review. *International Journal of MS Care* 2016;**18**(6):273-81.

Liu 2012

Liu X. Classification accuracy and cut point selection. *Statistics in Medicine* 2012;**31**(23):2676-86.

Löwe 2008

Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Medical Care* 2008;**46**(3):266-74.

Margraf 1994

Margraf J, Schneider S, Ehlers A. DIPS - Diagnostisches Interview bei Psychischen Störungen. Berlin, Germany: Springer, 1994.

Margraf 2013

Margraf J. Mini-DIPS: Diagnostisches Kurz-Interview bei Psychischen Störungen. Berlin; Heidelberg (Germany): Springer-Verlag, 2013.

McInnes 2018

McInnes MD, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018;**319**(4):388-96.

Mele 2018

Mele B, Holroyd-Leduc J, Smith EE, Pringsheim T, Ismail Z, Goodarzi Z. Detecting anxiety in individuals with Parkinson disease: a systematic review. *Neurology* 2018;**90**(1):e39-47.

Mulherin 2002

Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine* 2002;**137**(7):598-602.

Nelson 2020

Nelson HD, Cantor A, Pappas M, Weeks C. Screening for anxiety in adolescent and adult women: a systematic review for the Women's Preventive Services Initiative. *Annals of Internal Medicine* 2020;**173**(1):29-41.

Neupane 2021

Neupane D, Levis B, Bhandari PM, Thombs BD, Benedetti A, DEPRESsion Screening Data (DEPRESSD) Collaboration. Selective cutoff reporting in studies of the accuracy of the Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale: comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis. *International Journal of Methods in Psychiatric Research* 2021;**30**(3):e1873.

O'Connor 2023a

O'Connor E, Henninger M, Perdue LA, Coppola EL, Thomas R, Gaynes BN. Screening for Depression, Anxiety, and Suicide Risk in Adults: A Systematic Evidence Review for the US Preventive Services Task Force [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (USA) June 2023;**Report No.: 22-05295-EF-1**. [PMID: 37406149]

O'Connor 2023b

O'Connor EA, Henninger ML, Perdue LA, Coppola EL, Thomas RG, Gaynes BN. Anxiety screening: evidence report and systematic review for the US Preventive Services Task Force. *JAMA* 2023;**329**(24):2171-84. [DOI: 10.1001/jama.2023.6369]

Olariu 2015

Olariu E, Forero CG, Castro-Rodriguez JI, Rodrigo-Calvo MT, Álvarez P, Martín-López LM, et al. Detection of anxiety disorders in primary care: a meta-analysis of assisted and unassisted diagnoses. *Depression and Anxiety* 2015;**32**(7):471-84.

Plummer 2016

Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic meta-analysis. *General Hospital Psychiatry* 2016;**39**:24-31.

Quinn 2020

Quinn TJ, McCleery J, Hietamies TM, Abakar IF. Diagnostic test accuracy of self-administered cognitive assessment questionnaires for dementia. *Cochrane Database of Systematic Reviews* 2020, Issue 9. Art. No: CD013725. [DOI: 10.1002/14651858.CD013725]

R [Computer program]

R: A language and environment for statistical computing. Version 4.2.2. Vienna, Austria: R Foundation for Statistical Computing, 2021. Available at www.R-project.org.

Ransohoff 1978

Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* 1978;**299**(17):926-30.

Reitsma 2005

Reitsma JB, Glas AS, Rutjes AW, Scholten R, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005;**58**(10):982-90.

Remes 2016

Remes O, Brayne C, Van Der Linde R, Lafortune L. A systematic review of reviews on the prevalence of anxiety disorders in adult populations. *Brain and Behavior* 2016;**6**(7):e00497.

RevMan 2024 [Computer program]

Review Manager (RevMan). Version 7.2.0. The Cochrane Collaboration, 2024. Available at revman.cochrane.org.

Rohatgi 2014

Rohatgi A. Webplotdigitizer: Version 3.9 Of Webplotdigitizer. Automeris Software 2014. [AUTOMERIS: https://automeris.io/WebPlotDigitizer/]

Rücker 2020 [Computer program]

diagmeta: Meta-analysis of diagnostic accuracy studies with several cutpoints. R package version 0.5-1. Rücker G, Steinhauser S, Kolampally S, Schwarzer G. R Foundation for Statistical Computing, 2022.

Santamauro 2021

Santomauro DF, Herrera AM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet* 2021;**398**(10312):1700-12. [PMID: 34634250]

Schmidt 2013

Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. *Archives of Pathology & Laboratory Medicine* 2013;**137**(4):558-65. [DOI: 10.5858/arpa.2012-0198-RA]

Schneider 2008

Schneider A, Löwe B, Meyer FJ, Biessecker K, Joos S, Szecsenyi J. Depression and panic disorder as predictors of health outcomes for patients with asthma in primary care. *Respiratory Medicine* 2008;**102**(3):359-66.

Schneider 2013

Schneider A, Wartner E, Schumann I, Hörlein E, Henningsen P, Linde K. The impact of psychosomatic co-morbidity on discordance with respect to reasons for encounter in general practice. *Journal of Psychosomatic Research* 2013;**74**(1):82-5.

Schneider 2017

Schneider A, Linde K, Reitsma JB, Steinhauser S, Rücker G. A novel statistical model for analyzing data of a systematic review generates optimal cutoff values for fractional exhaled nitric oxide for asthma diagnosis. *Journal of Clinical Epidemiology* 2017;**92**:69-78.

Schonfeld 1997

Schonfeld WH, Verboncoeur CJ, Fifer SK, Lipschutz RC, Lubeck DP, Buesching DP. The functioning and well-being of patients with unrecognized anxiety disorders and major depressive disorder. *Journal of Affective Disorders* 1997;**43**(2):105-19.

Sheehan 1998

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry* 1998;**59 Suppl 20**:22-33; quiz 34-57.

Sherbourne 1996

Sherbourne CD, Jackson CA, Meredith LS, Camp P, Wells KB. Prevalence of comorbid anxiety disorders in primary care outpatients. *Archives of Family Medicine* 1996;**5**(1):27-34.

Sinesi 2019

Sinesi A, Maxwell M, O'Carroll R, Cheyne H. Anxiety scales used in pregnancy: systematic review. *BJPsych Open* 2019;**5**(1):e5.

Smits 2007

Smits N, Smit F, Cuijpers P, De Graaf R. Using decision theory to derive optimal cut-off scores of screening instruments: an illustration explicating costs and benefits of mental health screening. *International Journal of Methods in Psychiatric Research* 2007;**16**(4):219-29.

Spielberger 1970

Spielberger CD, Gorsuch RL, Lushene RE. STAI Manual for the State-Trait Anxiety Inventory. Palo Alto, California: Consulting Psychologists' Press. Inc, 1970.

Spitzer 2006

Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine* 2006;**166**(10):1092-7.

Steinhauser 2016

Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Medical Research Methodology* 2016;**16**(1):97.

Suppiger 2009

Suppiger A, In-Albon T, Hendriksen S, Hermann E, Margraf J, Schneider S. Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behavior Therapy* 2009;**40**(3):272-9.

Taylor 2021

Taylor KS, Mahtani KR, Aronson JK. Extracting data from diagnostic test accuracy studies for meta-analysis. *BMJ Evidence Based Medicine* 2021;**26**(1):19-21. [DOI: 10.1136/bmjebm-2020-111650]

Thomas 2020 [Computer program]

EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M. EPPI-Centre, UCL Social Research Institute, University College London, 2020.

Thombs 2011

Thombs BD, Arthurs E, El-Baalbaki G, Meijer A, Ziegelstein RC, Steele RJ. Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. *BMJ (Clinical Research Ed.)* 2011;**343**:d4825. [PMID: https://pubmed.ncbi.nlm.nih.gov/21852353/]

Thombs 2023

Thombs BD, Markham S, Rice DB, Ziegelstein RC. Screening for depression and anxiety in general practice. BMJ 2023;**382**:1615. [PMID: 37460129]

Toft 2005

Toft T, Fink P, Oernboel E, Christensen K, Frostholm L, Olesen F. Mental disorders in primary care: prevalence and co-morbidity among disorders. Results from the functional illness in primary care (FIP) study. *Psychological Medicine* 2005;**35**(8):1175-84.

Trikalinos 2012

Trikalinos TA, Balion CM, Coleman CI, Griffith L, Santaguida PL, Vandermeer B, et al. Meta-analysis of test performance when there is a “Gold Standard”. In: Chang SM, Matchar DB, Smetana GW, Umscheid CA, editors(s). Methods Guide for Medical Test Reviews. 1 edition. Vol. **27**. Rockville (MD): Agency for Healthcare Research and Quality (US), 2012:56-66. [PMID: 22834018]

Turner 2020

Turner GM, Litchfield I, Finnikin S, Aiyegbusi OL, Calvert M. General practitioners’ views on use of patient reported outcome measures in primary care: a cross-sectional survey and qualitative study. *BMC Family Practice* 2020;**21**(1):14. [PMID: https://pubmed.ncbi.nlm.nih.gov/31980021/]

USPSTF 2023

US Preventive Services Task Force. Screening for anxiety disorders in adults: US Preventive Services Task Force recommendation statement. *JAMA* 2023;**329**(24):2163-70. [DOI: 10.1001/jama.2023.9301]

Van Enst 2014

Van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC Medical Research Methodology* 2014;**14**:70.

Vongmany 2016

Vongmany J, Hickman LD, Lewis J, Newton PJ, Phillips JL. Anxiety in chronic heart failure and the risk of increased hospitalisations and mortality: A systematic review. *European Journal of Cardiovascular Nursing* 2016;**15**(7):478-85.

Wasserstein 2019

Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond *"p < 0.05"*. *American Statistician* 2019;**73**(Suppl 1):1-19. [DOI: 10.1080/00031305.2019.1583913]

Whiting 2011

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011;**155**(8):529-36.

WHO 1990

World Health Organization. Composite International Diagnostic Interview. Geneva, Switzerland: World Health Organization, 1990.

WHO 1993

World Health Organization. ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research. 2 edition. Geneva, Switzerland: World Health Organization, 1993.

WHO 2019

World Health Organization. ICD-11 for Mortality and Morbidity Statistics: Mental, Behavioural Or Neurodevelopmental Disorders. Geneva, Switzerland: World Health Organization, 2019.

Wilson 2015

Wilson C, Kerr D, Noel-Storr A, Quinn TJ. Associations with publication and assessing publication bias in dementia diagnostic test accuracy studies. *International Journal of Geriatric Psychiatry* 2015;**30**(12):1250-6.

Wittchen 1997

Wittchen HU, Pfister H. DIA-X-Interviews: Manual Für Screening-Verfahren und Interview; Interviewheft. Frankfurt, Germany: Swets & Zeitlinger, 1997.

Yang 2021

Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Annals of Internal Medicine* 2021;**174**(11):1592-9. [PMID: 34698503]

Zigmond 1983

Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica* 1983;**67**(6):361-70.

References to other published versions of this review

Linde 2022

Linde K, Olm M, Teusen C, Akturk Z, von Schrottenberg V, Hapfelmeier A, et al. The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults. *Cochrane Database of Systematic Reviews* 2022, Issue 9. Art. No: CD015292. [DOI: 10.1002/14651858.CD015292]

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Ahmadi 2019

Study characteristics	
Patient Sampling	**Sampling method**: No detailed information but reasonable exclusions. **Sampling place**: Hospital in Karmanash, Iran **Exclusion criteria**: Heart disease patients not able to participate in the study due to severe heart problems, those with cognitive problems (such as Alzheimer's), severe psychological problems (such as psychotic disorders), and physical problems (such as thyroid) were excluded.
Patient characteristics and setting	**Country**: Iran **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 279 **n analysed**: 279 **% female participants**: 46 **Mean age**: 60 years

Ahmadi 2019 *(Continued)*

	Patient characteristics: Patients with coronary heart disease
Index tests	**Index tests used**: GAD-7 and GAD-2 **Index test administration**: Mixed **Index test language**: Persian **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-2: 2 **Post hoc 'best' cut-off based on**: Youden Index
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained professionals **Target condition(s)**: AAD **Prevalence AAD (%)**: 0.28 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The primary aim of the study was DTA assessment. No extra information could be obtained from the authors. Prevalence was calculated from positive predictive value, sensitivity, and specificity using the formula suggested by Taylor and colleagues (Taylor 2021).

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		

Ahmadi 2019 *(Continued)*

Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Unclear		
Could the patient flow have introduced bias?		Unclear risk	

Ahn 2019

Study characteristics	
Patient Sampling	**Sampling method**: Potential research participants were included by hospital staff **Sampling place**: Internet and one research lab and two general hospitals **Exclusion criteria**: Exclusion criteria were not specified.
Patient characteristics and setting	**Country**: South Korea **Setting**: Screening in patients from non-clinical setting **Ethnicity**: Asian **n included**: 1228 **n analysed**: 1157

Ahn 2019 *(Continued)*

	% female participants: 66.7 **Mean age**: 37.3 years **Patient characteristics**: Undefined adults
Index tests	**Index tests used**: GAD-7, GAD-2, and BAI **Index test administration**: Mixed self-report and interview **Index test language**: Korean **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-2: 2 **Post hoc 'best' cut-off based on**: Youden Index
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: trained professionals **Target condition(s):** AAD, GAD **Prevalence AAD (%)**: 0.19 **Prevalence GAD (%)**: 0.08 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. No extra information could be obtained from the authors.

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		

Ahn 2019 *(Continued)*

If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Baker 2018

Study characteristics	
Patient Sampling	**Sampling method**: No information **Sampling place**: Participants were recruited at 16 centres of the American Lung Association Airways Clinical Research Centers Network. **Exclusion criteria**: Individuals with unstable coronary heart disease or a major psychiatric disorder were excluded.
Patient characteristics and setting	**Country**: United States of America **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information

Baker 2018 *(Continued)*

	n included: 223 **n analysed**: 220 **% female participants**: 46 **Median age**: 65 years **Patient characteristics**: 40 years of age or older, stable COPD
Index tests	**Index tests used**: GAD-7 and HADS-A **Index test administration**: Interview **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 5 **Post hoc 'best' cut-off based on**: Likelihood ratio +
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: local study coordinators (trained). **Target condition(s):** AAD **Prevalence AAD (%)**: 0.11 **Diagnostic criteria**: DSM-V **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: Same day
Comparative	
Notes	The objective of the study was DTA assessment. No extra information could be obtained from the authors. Diagnostic information for the missing cut-offs was extracted by the review authors from the appendix, Figure 3, using WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/).

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			High

Baker 2018 *(Continued)*

DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Belus 2021

Study characteristics	
Patient Sampling	**Sampling method**: Research assistants approached unselected adults in waiting rooms. Interested participants were recruited **Sampling place**: Three primary care health facilities in the Sofala province, Mozambique **Exclusion criteria**: People were excluded if they had an acute health condition or disability impeding their ability to complete the survey.

Belus 2021 *(Continued)*

Patient characteristics and setting	**Country**: Mozambique **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: No information **n included**: 534 **n analysed**: 502 **% female participants**: 74 **Mean age**: 28 years **Patient characteristics**: Unselected adults
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Interview **Index test language**: Mozambican Portuguese **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: NA **Post hoc 'best' cut-off based on**: NA
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: trained mental health professionals **Target condition(s):** GAD **Prevalence GAD**: 0.13 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Same day
Comparative	
Notes	The objective of the study was DTA assessment. Authors provided extra information. All cut-offs could be extracted. Additional information obtained: there was no time lag between the index tests and the reference standard. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		

Belus 2021 *(Continued)*

Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Bernstein 2018

Study characteristics			

Bernstein 2018 *(Continued)*

Patient Sampling	**Sampling method**: Posters, social media, self-help groups; contacting inflammatory bowel disease (IBD) patients in the clinic and in a regional IBD registry. **Sampling place**: University hospital department for internal medicine, Manitoba, Canada. **Exclusion criteria**: Not reported.
Patient characteristics and setting	**Country**: Canada **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: White **n included**: 247 **n analysed**: 242 **% female participants**: 63 **Mean age**: 48 years **Patient characteristics**: Adults with inflammatory bowel disease
Index tests	**Index tests used**: GAD-7, HADS-A **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 8, HADS-A: 9 **Post hoc 'best' cut-off based on**: Youden's index
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained personnel (graduate psychology students, nurses, research coordinators) supervised by a clinical psychologist **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.18 **Prevalence GAD**: 0.06 **Diagnostic criteria**: DSM-IV-TR **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Within two weeks (114 participants the same day)
Comparative	
Notes	The objective of the study was DTA assessment. The authors responded and provided extra information. All cut-offs could be extracted. The

Bernstein 2018 *(Continued)*

	study shares a protocol with Hitchon 2020 and Marrie 2018. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		

Bernstein 2018 *(Continued)*

Were all patients included in the analysis?	Yes
Could the patient flow have introduced bias?	Low risk

Bisby 2022

Study characteristics	
Patient Sampling	**Sampling method**: Recruitment via advertisements directing interested individuals to the trial website (this publication presents a re-analysis of a randomised controlled study (RCT)) **Sampling place**: Internet **Exclusion criteria**: Participants who reported very severe depressive symptoms on the PHQ-9 (i.e. total score ≥ 22, item 9 score ≥ 2) were excluded.
Patient characteristics and setting	**Country**: Australia **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 63 **n analysed**: 62 **% female participants**: 85 **Mean age**: 49 years **Patient characteristics**: Patients with chronic pain recruited for an RCT investigating the effectiveness of internet-delivered CBT programme.
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Computer-based self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 11, GAD-2: 3 **Post hoc 'best' cut-off based on**: LR+
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: One clinical psychologist with postgraduate qualifications and several years' clinical experience with chronic pain **Target condition(s):** GAD **Prevalence GAD**: 0.35 **Diagnostic criteria**: DSM-IV

Bisby 2022 *(Continued)*

	Were the researchers administering the RS blinded for the index test results? No
Flow and timing	**Time lag between index test and RS**: 0–14 days
Comparative	
Notes	Data originating from a randomised controlled study. Extra information was provided by the authors. Additional information obtained: Diagnostic criteria were based on the DSM-IV (MINI version 5). The time lag between the index tests and the reference standard was within 0–14 days. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	No		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	

Bisby 2022 *(Continued)*

Are there concerns that the target condition as defined by the reference standard does not match the question?		Low concern
DOMAIN 4: Flow and timing		
Was there an appropriate interval between index test and reference standard?	Yes	
Did all patients receive the same reference standard?	No	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		High risk

Budikayanti 2019

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive **Sampling place**: Epilepsy outpatient clinic in Cipto Mangunkusumo General Hospital, Indonesia. **Exclusion criteria**: Having hearing problems, severe visual impairment, mental retardation, history of neurological, psychiatric, or severe medical problems, and history of alcohol use or drug abuse.
Patient characteristics and setting	**Country**: Indonesia **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 146 **n analysed**: 146 **% female participants**: 46 **Mean age**: 36 years **Patient characteristics**: Adults with epilepsy
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Bahasa Indonesia **Index test validation**: Yes **Post hoc cut-off based on**: Cut-off point of the sensitivity and specificity **Were the researchers administering the RS blinded for the index test results?** Yes
Target condition and reference standard(s)	**Reference standard**: MINI

Budikayanti 2019 *(Continued)*

	Reference standard applied by: Neurology resident trained by the psychiatrist in the team **Target condition(s):** GAD **Prevalence GAD**: 0.16 **Diagnostic criteria**: ICD-10 **Time lag between index test and RS**: Same day
Flow and timing	**Post hoc best cut-off (≥)**: GAD-7 > 6
Comparative	
Notes	The objective of the study was DTA assessment. DTA information could be calculated for all cut-offs. Authors provided extra information: the reference standard was performed by a neurology resident who was trained by the psychiatrist in the team. The point with the highest sensitivity and specificity was selected as the best cut-off. There was no gap between administrating MINI and GAD-7. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			

Budikayanti 2019 *(Continued)*

Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Chibanda 2016

Study characteristics	
Patient Sampling	**Sampling method**: Research assistants invited clinic attendees randomly selected from the daily list for participation. **Sampling place**: Primary care clinic in a suburb of Mbare, Zimbabwe. **Exclusion criteria**: Pregnant women in their last trimester and women within the 3-month post-natal period were excluded as were those who were unable to understand the purpose of the study.
Patient characteristics and setting	**Country**: Zimbabwe **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: Black **n included**: 264 **n analysed**: 264 **% female participants**: 79 **Mean age**: 37.6 years **Patient Characteristics**: Unselected primary clinic attendees (165 (63% of 264) were HIV positive)
Index tests	**Index tests used**: GAD-7, GAD-2

Chibanda 2016 *(Continued)*

	Index test administration: Interview **Index test language**: Shona **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: No information **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Four trained psychiatrists **Target condition(s):** GAD **Prevalence GAD**: 0.4 **Diagnostic criteria**: DSM-IV **Time lag between index test and RS**: Same day
Flow and timing	**Were the researchers administering the RS blinded for the index test results?** Yes
Comparative	
Notes	The objective of the study was DTA assessment. Authors provided extra information on additional cut-offs. DTA information could be calculated for all cut-offs. Additional information obtained: all participants were black Africans.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Unclear		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Unclear		

Chibanda 2016 *(Continued)*

Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Christodoulaki 2022

Study characteristics	
Patient Sampling	**Sampling method**: A cross-sectional health examination survey conducted from May 2013 until June 2016, aiming to describe morbidity along with risk factors and preventive measures in a large, randomly selected sample, representative of the general population in Greece. The sample was selected using the multistage stratified random sampling method, and 6006 adults were recruited in total. **Sampling place**: Part of the National Survey of Morbidity and Risk Factors (EMENO), a cross-sectional health examination survey amongst individuals living in Central Greece. **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: Greece **Setting**: Screening in non-clinical setting **Ethnicity**: White **n included**: 591 **n analysed**: 584

Christodoulaki 2022 *(Continued)*

	% female participants: 56.1 **Mean age**: 53 years **Patient characteristics**: Random individuals living in Central Greece from different households. A cross-sectional sample from the general population.
Index tests	**Index tests used**: GAD-2 **Index test administration**: Self-report **Index test language**: Greek **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: 2 **Post-hoc 'best' cut-off based on**: Maximisation of sensitivity and specificity.
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Psychiatrist **Target condition(s):** GAD **Prevalence GAD**: 0.9 **Prevalence AAD:** 0.30 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: No longer than one week.
Comparative	
Notes	The objective of the study was DTA assessment. The study investigated target conditions AAD and GAD. DTA information could be calculated for all cut-offs. Diagnostic information on GAD were obtained from the article supplement provided by the author. Also, the authors confirmed that the time elapsed between the index test and reference standard was less than a week, and that the researchers administering the reference standard were blinded for the index test results.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	

Christodoulaki 2022 *(Continued)*

Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Clover 2020

Study characteristics			

Clover 2020 *(Continued)*

Patient Sampling	**Sampling method**: Convenience sample. A research assistant approached participants at an outpatient appointment. Consenting patients were included. **Sampling place**: Cancer centre in New South Wales, Australia. Outpatients from a cancer centre. **Exclusion criteria**: Patients attending their first clinic visit, insufficient English language skills, too unwell to consider participation.
Patient characteristics and setting	**Country**: Australia **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 322 **n analysed**: 153 (GAD), 132 (AAD) **% female participants**: 66 **Mean age**: 59 years **Patient Characteristics**: Adult outpatients with a variety of cancers, unselected for cancer type or stage from a cancer centre.
Index tests	**Index tests used**: GAD-7, GAD-2, HADS-A **Index test administration**: Computer-based self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: for AAD: all cut-offs provided, for GAD: 4 ('best' balance) **Post hoc 'best' cut-off based on**: for GAD: 'best' balance between sensitivity and specificity
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Two trained registered psychologists **Target condition(s):** GAD, AAD **Prevalence AAD**: 0.28 **Prevalence GAD**: 0.09 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Same day Of the 322 people consenting, 168 completed all instruments and 153 were included in the GAD analysis and 132 in the AAD analysis. While many participants giving consent were not included in the analysis, this seems to be more of an organisational issue than selection.

Clover 2020 *(Continued)*

Comparative	
Notes	The objective of the study was DTA assessment. The principal investigator could be contacted for additional information. GAD data were reported in Clover 2020, while AAD in Clover 2022 (additional record for Clover 2020). Raw data were provided by the principal investigator for AAD, and DTA information could be calculated for all cut-offs.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	

Clover 2020 *(Continued)*

Are there concerns that the target condition as defined by the reference standard does not match the question?		Low concern
DOMAIN 4: Flow and timing		
Was there an appropriate interval between index test and reference standard?	Yes	
Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	No	
Could the patient flow have introduced bias?		High risk

Conway 2016

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive patients, January to September 2018. **Sampling place**: Outpatient clinic at a major metropolitan hospital in Australia **Exclusion criteria**: Patients who were less than three months post-transplant as well as those who were cognitively impaired (as confirmed by a treating clinician), unable to understand and speak English, had a diagnosed major psychiatric comorbidity (schizophrenia, bipolar disorder, dementia) or terminal illness were excluded.
Patient characteristics and setting	**Country**: Australia **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 80 **n analysed**: 48 **% female participants**: 24 **Mean age**: 63 years **Patient characteristics**: Heart transplant recipients (long-term survivors) over 18 years.
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 6

Conway 2016 *(Continued)*

	Post hoc 'best' cut-off based on: Liu method
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Provisional psychologist undertaking a Doctor of Clinical Psychology **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.16 **Prevalence GAD**: 0.08 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Mean 21.8 days
Comparative	
Notes	The objective of the study was DTA assessment. Raw data were provided by the principal investigator. DTA information could be calculated for all cut-offs. Extra information provided by the authors: ethnicity of the participants was not queried. There was a time lag between the GAD-7 and MINI. The Liu method was used to identify the optimal cut-off point that maximised the sensitivity and specificity of the screening tools in the detection of psychopathology.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		

Conway 2016 *(Continued)*

Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		High risk	

Delgadillo 2012

Study characteristics	
Patient Sampling	**Sampling method:** Participants were recruited via sequential contacts during a full calendar year **Sampling place:** A community drugs treatment service in Leeds, UK **Exclusion criteria:** Patients with severe mental illness such as psychotic disorders identified in clinical records.
Patient characteristics and setting	**Country:** United Kingdom **Setting:** Screening in patients from clinical setting with a specific condition **Ethnicity:** White **n included:** 103 **n analysed:** 103 **% female participants:** 23

Delgadillo 2012 *(Continued)*

	Mean age: 35 years **Patient characteristics:** Outpatients applying to a community drug treatment service.
Index tests	**Index tests used:** GAD-7, GAD-2 **Index test administration:** Self-report **Index test language:** English **Index test validation:** Yes **Post hoc 'best' cut-off (≥):** AAD: GAD-7:9, GAD-2:2; GAD: GAD-7:9, GAD-2:3 **Post hoc 'best' cut-off based on:** Youden's index
Target condition and reference standard(s)	**Reference standard**: CIS-R **Reference standard applied by**: Trained interviewers **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.65 **Prevalence GAD**: 0.3 **Diagnostic criteria**: ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS:** Same day
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not be contacted for additional information.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		

Delgadillo 2012 *(Continued)*

If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Esser 2018

Study characteristics	
Patient Sampling	**Sampling method**: Eligible participants were consecutively approached by study research assistants. **Sampling place**: Five centres (Hamburg, Freiburg, Heidelberg, Leipzig, Würzburg) **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: Germany **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 4020

Esser 2018 *(Continued)*

	n analysed: 2142 (GAD-7), 1931 (GAD-2)
	% female participants: 52
	Mean age: 58 years
	Patient characteristics: Adults with cancer
Index tests	**Index tests used**: GAD-7, GAD-2, HADS-A
	Index test administration: Self-report
	Index test language: German
	Index test validation: Yes
	Post hoc 'best' cut-off (≥): GAD-7: 7
	Post hoc 'best' cut-off based on: Youden's index
Target condition and reference standard(s)	**Reference standard**: CIDI
	Reference standard applied by: Trained and supervised interviewers
	Target condition(s): AAD, GAD
	Prevalence AAD: 0.12
	Prevalence GAD: 0.02
	Diagnostic criteria: DSM-IV, ICD-10
	Were the researchers administering the RS blinded for the index test results? Yes
Flow and timing	**Time lag between index test and RS**: Mean 8.7 days (SD 15)
Comparative	
Notes	The objective of the study was DTA assessment. Raw data were received from the authors. DTA information could be calculated for all cut-offs. Weighting was done to correct potential bias by test-result-based sampling. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear

Esser 2018 *(Continued)*

DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		High risk	

Fischer 2014

Study characteristics	
Patient Sampling	**Sampling method**: Recruitment (via email) was conducted amongst a cohort of outpatients with confirmed heart failure under treatment at the University Clinic Göttingen. Convenience sample. **Sampling place**: University Clinic Göttingen **Exclusion criteria**: Patients were excluded if they suffered from acute decompensation, acute myocardial infarction or heart surgery in the past four weeks or were unable to participate due to cognitive impairment, lan-

Fischer 2014 *(Continued)*

	guage problems, psychosis or addiction, or refused to give informed consent.
Patient characteristics and setting	**Country**: Germany **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: White **n included**: 194 **n analysed**: 194 **% female participants**: 21.1 **Mean age**: 65.9 years **Patient characteristics**: Confirmed heart failure patients with ejection fractions less than 45%.
Index tests	**Index tests used**: GAD-7, GAD-2, HADS-A **Index test administration**: Self-report **Index test language**: German **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 6 **Post hoc 'best' cut-off based on**: Youden Index
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained psychologists **Target condition(s):** AAD **Prevalence AAD**: 0.14 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: Same day
Comparative	
Notes	The objective of the study was DTA assessment. Extra information was received from the author. DTA information could be calculated for all cut-offs. Additional information obtained: the participants were sampled with convenience sample. The GAD-7 was administered as self-report. There was no time gap between the index test and the reference standard. Data were collected on the same day. No clear information about blinding. Anxiety in the text refers to 'any anxiety disorder'.
Methodological quality	

Fischer 2014 *(Continued)*

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		

Fischer 2014 *(Continued)*

Could the patient flow have introduced bias?	Low risk

Garabiles 2020

Study characteristics	
Patient Sampling	**Sampling method**: Respondent-driven sampling **Sampling place**: Local partner non-governmental organisation **Exclusion criteria**: None reported
Patient characteristics and setting	**Country**: Philippines **Setting**: Screening in non-clinical setting **Ethnicity**: Asian **n included**: 100 **n analysed**: 100 **% female participants**: 100 **Mean age**: 41.2 years **Patient characteristics**: Female Filipino domestic workers aged 18 or above, who worked in Macao
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Filipino **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 7 **Post hoc 'best' cut-off based on**: Balancing between sensitivity and specificity
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Licenced female Filipino clinical psychologist **Target condition(s):** GAD **Prevalence GAD**: 0.28 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Same day
Comparative	

Garabiles 2020 *(Continued)*

Notes	The objective of the study was DTA assessment. No extra information could be obtained from the authors.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		

Garabiles 2020 *(Continued)*

Were all patients included in the analysis?	Yes
Could the patient flow have introduced bias?	Low risk

Grech 2019

Study characteristics	
Patient Sampling	**Sampling method**: Recently hospitalised people **Sampling place**: Tertiary hospital in Adelaide, South Australia **Exclusion criteria**: "Psychosis or bipolar disorder diagnosis determined by medical history or randomisation-naïve assessors, high suicide risk at psychiatric interview. Observed cognitive impairment or dementia impeding delivery of psychotherapy or ability to provide informed consent, neurodegenerative condition (for example, Parkinson's disease or multiple sclerosis), in receipt of GP, psychologist or psychiatrist counselling elsewhere, a diagnosis of drug and alcohol dependence or abuse determined by randomization-naïve assessors, medical condition likely to be fatal within one year."
Patient characteristics and setting	**Country**: Australia **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 85 **n analysed**: 85 **% female participants**: 31 **Mean age**: 63 years **Patient characteristics**: People recently hospitalised for CVD participating in an RCT on trans-diagnostic treatment of emotional disorders.
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 4 **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: The assessor (graduate from a 4th year psychology degree) was trained, and supervised, in all aspects of clinical assessments. **Target condition(s):** GAD **Prevalence GAD**: 0.08

Grech 2019 *(Continued)*

	Diagnostic criteria: DSM-IV (reviewer assumption based on the use of the MINI version 5.0.0) **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: Unclear (GAD-7 used at baseline and at two weeks, MINI at about two weeks)
Comparative	
Notes	Data from a randomised controlled study. Authors provided extra information. Diagnostic information on all possible cut-offs could be extracted. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		

Grech 2019 *(Continued)*

Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Hitchon 2020

Study characteristics	
Patient Sampling	**Sampling method**: Posters, social media, self-help groups; contacts individually on a local registry. **Sampling place**: University hospital, arthritis centre in Manitoba, Canada. **Exclusion criteria**: Not reported.
Patient characteristics and setting	**Country**: Canada **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Mixed **n included**: 154 **n analysed**: 150 **% female participants**: 85 **Mean age**: 60 years **Patient characteristics**: Adults with rheumatoid arthritis
Index tests	**Index tests used**: GAD-7, GAD-2, HADS-A **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 9, HADS-A: 9 **Post hoc 'best' cut-off based on**: Youden's index

Hitchon 2020 *(Continued)*

Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained personnel (graduate psychology students, nurses, research coordinators) supervised by a clinical psychologist. **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.19 **Prevalence GAD**: 0.07 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Within two weeks (114 participants on same day).
Comparative	
Notes	The objective of the study was DTA assessment. The authors responded and provided extra information. All cut-offs could be extracted. The study shares a protocol with Bernstein 2018 and Marrie 2018. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	

Hitchon 2020 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Homans 2012

Study characteristics	
Patient Sampling	**Sampling method**: In the recruitment phase from an RCT (sick-listed workers). **Sampling place**: In collaboration with the occupational health service, Trimbis Institute leaflet. **Exclusion criteria**: Excluded from participation were workers who were not sick-listed between four and 26 weeks, those who did not have sufficient command of the Dutch language and those who were pregnant.
Patient characteristics and setting	**Country**: Netherlands **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 170 **n analysed**: 170 **% female participants**: 50 **Mean age**: 45.4 years

Homans 2012 *(Continued)*

	Patient characteristics: Sick-listed workers due to any cause. Long-term sick-listed workers (4–26 weeks sick).
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Dutch **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: 9 **Post hoc 'best' cut-off based on**: Sum of sensitivity and specificity
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Trained interviewers **Target condition(s):** AAD **Prevalence AAD**: 0.13 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Less than 4 weeks; mean 13.6 days, SD: 7.1 days
Comparative	
Notes	The objective of the study was not DTA assessment. Master's thesis in Clinical Health and Psychology. Author could not be contacted for extra information.

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		

Homans 2012 *(Continued)*

Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Ivanovs 2018

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive adult primary care patients. **Sampling place**: 24 primary care facilities in Latvia. **Exclusion criteria**: People who refused to participate in the study; younger than 18 years of age, and people with acute medical conditions requiring urgent hospitalisation.
Patient characteristics and setting	**Country**: Latvia **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: White **n included**: 1602 **n analysed**: 1485

Ivanovs 2018 *(Continued)*

	% female participants: 31 **Mean age**: 56% ≥ 55 years **Patient characteristics**: Primary care patients all over the country
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: Latvian/Russian **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: No information **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: trained psychiatrists (telephone) **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.16 **Prevalence GAD**: 0.06 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Within two weeks
Comparative	
Notes	The objective of the study was not DTA assessment. Unpublished diagnostic information on most cut-offs (GAD-2: 1-6, GAD-7: 3-15) were provided by the authors. The paper was also published split by language (Latvian and Russian), but we used the unsplit data from Vrublevska 2022 (see secondary reference) for our analyses. Authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			

Ivanovs 2018 *(Continued)*

Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Kroenke 2007

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive patients were approached **Sampling place**: 15 American primary care clinics **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: USA **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: White

Kroenke 2007 *(Continued)*

	n included: 965 **n analysed**: 965 **% female participants**: 69 **Mean age**: 47.1 **Patient characteristics**: Post-traumatic stress disorder: 83 participants (8.6%), generalised anxiety disorder: 73 participants (7.6%), panic disorder: 66 participants (6.8%), social anxiety disorder: 60 participants (6.2%).
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 8, GAD-2: 3 **Post hoc 'best' cut-off based on**: LR+
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Clinical psychologist (with a PhD) or a senior psychiatric social worker. **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.19 **Prevalence GAD**: 0.08 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Within one week
Comparative	
Notes	The objective of the study was DTA assessment. The authors provided raw data, enabling the calculation of diagnostic information for all possible cut-offs.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Unclear risk	

Kroenke 2007 *(Continued)*

Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Kujanpää 2014

Study characteristics	
Patient Sampling	**Sampling method**: High utilisers taking part in a case management intervention giving consent to the additional diagnostic study. **Sampling place**: Four municipal health centres in Finland **Exclusion criteria**: Visits due to pregnancy or delivery, serial treatment for the same illness, terminal hospice, cancer palliative care, psychotic illness,

Kujanpää 2014 *(Continued)*

	dementia, mental retardation, inability to give informed consent, and other study intervention at the same time or just prior to this study.
Patient characteristics and setting	**Country**: Finland **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 150 **n analysed**: 150 **% female participants**: 69 **Mean age**: 63 years **Patient characteristics**: At least 18 years of age and having eight visits per year to the general practitioner in the local health centre or four visits per year to the university hospital.
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: Finnish **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 7 **Post hoc 'best' cut-off based on**: Youden's index
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: A trained general practitioner **Target condition(s):** GAD **Prevalence GAD**: 0.04 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: Not described, but very likely after inclusion/GAD-7
Comparative	
Notes	The objective of the study was DTA assessment. No additional information could be obtained from the authors. Although the sample is described as "high utilisers", people seeking serial treatment for the same illness were excluded, leaving mostly participants with mental disorders.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			

Kujanpää 2014 *(Continued)*

Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Li 2014

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive outpatients **Sampling place**: Hospital outpatient units for cardiovascular diseases **Exclusion criteria**: None reported
Patient characteristics and setting	**Country**: China **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Asian **n included**: No information **n analysed**: 201 **% female participants**: No information **Mean age**: No information **Patient characteristics**: Cardiovascular outpatients
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Chinese **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: 10 **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: CIDI **Reference standard applied by**: psychiatrist **Target condition(s):** AAD **Prevalence AAD**: 0.37 **Diagnostic criteria**: No information **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. Primarily a DTA study. No additional information could be obtained from the authors.
Methodological quality	

Li 2014 *(Continued)*

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Unclear		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Unclear		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Unclear		
Could the patient flow have introduced bias?		Unclear risk	

Lickova 2021

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive **Sampling place**: Radiotherapy and Oncology Clinic of the Third Faculty of Medicine of Charles University and the Faculty Hospital in Prague. **Exclusion criteria**: Severe neurodegenerative diseases, people with brain tumours, people in the final stage of the disease, and people who were unable or unwilling to cooperate and complete the required information.
Patient characteristics and setting	**Country**: Czech Republic **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: White **n included**: 238 **n analysed**: 238 **% female participants**: 40.3 **Mean age**: 66.6 years **Patient characteristics**: Hospitalised patients and outpatients, regardless of diagnosis, stage or time since diagnosis
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Czech **Index test validation**: No **Post-hoc best cut-off (≥)**: 9 **Post-hoc best cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: The researcher **Target condition(s):** GAD **Prevalence GAD**: 0.24 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No
Flow and timing	**Time lag between index test and RS**: No time lag
Comparative	
Notes	The objective of the study was DTA assessment. No diagnostic information could be obtained for additional cut-offs. The authors provided additional information as follows: "The index test has not been

Lickova 2021 *(Continued)*

validated for Czech. The post-hoc best cut-off was decided using the Youden index. The researchers administering the reference standard were not blinded for the results of the index test."

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		

Lickova 2021 *(Continued)*

Were all patients included in the analysis?	Yes
Could the patient flow have introduced bias?	Low risk

Makulowich 2018

Study characteristics	
Patient Sampling	**Sampling method**: No information **Sampling place**: Integrated community clinic, Rosalind Franklin University of Medicine and Science Interprofessional Community Clinic. **Exclusion criteria**: Women who were pregnant or nursing were excluded from the study. One individual was excluded due to not speaking English or Spanish.
Patient characteristics and setting	**Country**: USA **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: Mixed **n included**: 50 **n analysed**: 50 **% female participants**: 70 **Mean age**: 49.6 years **Patient characteristics**: Patients in the Interprofessional Community Clinic
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: English+Spanish **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: no information **Post hoc 'best' cut-off based on**: Identified taking into account anxiety prevalence in this population (36%) and cost of false and true positive and negative decisions.
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Doctoral students in clinical psychology or Master's students in clinical counselling. **Target condition(s):** AAD **Prevalence AAD**: 0.32 **Diagnostic criteria**: DSM-IV

Makulowich 2018 *(Continued)*

	Were the researchers administering the RS blinded for the index test results? No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. No additional information could be obtained from the authors.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern

Makulowich 2018 *(Continued)*

DOMAIN 4: Flow and timing		
Was there an appropriate interval between index test and reference standard?	Unclear	
Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		Unclear risk

Maric 2022

Study characteristics	
Patient Sampling	**Sampling method**: Cross-sectional observational epidemiological survey with multistage probabilistic household sampling. **Sampling place**: Representative household sampling of the adult population in 60 municipalities. **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: Serbia **Setting**: Screening in non-clinical setting **Ethnicity**: White **n included**: 1203 **n analysed**: 1203 **% female participants**: 51.3 **Mean age**: 43.7 years **Patient characteristics**: Adults from general population
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: Serbian **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: NA **Post hoc 'best' cut-off based on**: NA
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Psychologists, medical doctors or senior medical students **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.03

Maric 2022 *(Continued)*

	Prevalence GAD: 0.03 **Diagnostic criteria**: DSM-V **Were the researchers administering the RS blinded for the index test results?** No
Flow and timing	**Time lag between index test and RS**: There was no time lag between the index test and reference standard administrations.
Comparative	
Notes	The objective of the study was not DTA assessment. The authors responded and provided unpublished information and raw data. All cut-offs could be extracted. Additional information obtained: 97.2% of the participants were white. GAD-7 was validated for Serbian. The index test was self-administered. There was no time lag between the index test and reference standard administrations. The assessors were not blinded.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		

Maric 2022 *(Continued)*

Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Marrie 2018

Study characteristics	
Patient Sampling	**Sampling method**: Posters, social media, self-help groups; contacting patients in the clinic and in a regional registry. **Sampling place**: Provincial multiple sclerosis (MS) clinic. The MS Clinic in Manitoba. **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: Canada **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: White **n included**: 255 **n analysed**: 253 **% female participants**: 81.4 **Mean age**: 51 years **Patient characteristics**: MS patients admitted to an MS clinic
Index tests	**Index tests used**: GAD-7, GAD-2, HADS-A **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 5

Marrie 2018 *(Continued)*

	Post hoc 'best' cut-off based on: Youden's Index
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained interviewers **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.17 **Prevalence GAD**: 0.04 **Diagnostic criteria**: DSM-IV-TR **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Usually on the same day, otherwise within two to four weeks.
Comparative	
Notes	The objective of the study was DTA assessment. The authors responded and provided extra information and raw data. All cut-offs could be extracted. The study shares a protocol with Bernstein 2018 and Hitchon 2020. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	

Marrie 2018 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Michaelis 2022

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive sampling **Sampling place**: Ruhr-Epileptology, an academic reference centre nested within the Neurology Department of the University Hospital Knappschaftskrankenhaus in Bochum, Germany **Exclusion criteria**: Clinically relevant conditions that could potentially compromise reliable completion of the questionnaires, such as language barriers, severe cognitive impairment, and severe psychiatric disorders.
Patient characteristics and setting	**Country**: Germany **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 80 **n analysed**: 80 **% female participants**: 60 **Mean age**: 39 years

Michaelis 2022 *(Continued)*

	Patient characteristics: Adults with confirmed epilepsy
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: German **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: NA **Post hoc 'best' cut-off based on**: NA
Target condition and reference standard(s)	**Reference standard**: Mini-DIPS **Reference standard applied by**: Trained psychotherapists with experience in epileptology. **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.28 **Prevalence GAD**: 0.11 **Diagnostic criteria**: ICD-10 **Were the researchers administering the RS blinded for the index test results?** No
Flow and timing	**Time lag between index test and RS**: Maximum three days.
Comparative	
Notes	The objective of the study was DTA assessment. The authors provided raw data and diagnostic information on additional cut-offs. DTA information could be calculated for all cut-offs. Additional information obtained: the interviews were conducted within three workdays after the collection of the questionnaires. Researchers administering the interviews were not blinded for the index test results. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			

Michaelis 2022 *(Continued)*

Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	No		
Could the reference standard, its conduct, or its interpretation have introduced bias?		High risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Micoulaud-Franchi 2016

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive sampling **Sampling place**: Clinical Neurophysiology Department of the Marseille University Hospital and the Hôpital Henri Gastaut, Marseille. **Exclusion criteria**: Insufficient capacity to consent and to understand and answer the self-report questionnaires and the presence of other severe chronic medical, neurological, and psychiatric conditions.
Patient characteristics and setting	**Country**: France

Micoulaud-Franchi 2016 *(Continued)*

	Setting: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 145 **n analysed**: 145 **% female participants**: 63.4 **Mean age**: 39.3 years **Patient characteristics**: Native French-speaking adults with any type of active epilepsy according to the ILAE (International League Against Epilepsy) criteria
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: French **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 8 **Post hoc 'best' cut-off based on**: by selecting the point on the ROC curve that maximised both sensitivity and specificity.
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: physicians with expertise in the field of epilepsy and psychiatry for whom an explanation and a training session with the MINI module was performed by an experienced psychiatrist **Target condition(s):** GAD **Prevalence GAD**: 0.34 **Diagnostic criteria**: DSM-IV-TR **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Same session
Comparative	
Notes	The objective of the study was DTA assessment. The authors responded and provided extra information. All cut-offs could be extracted.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		

Micoulaud-Franchi 2016 *(Continued)*

Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Micoulaud-Franchi 2022

Study characteristics	
Patient Sampling	**Sampling method**: People were invited to participate in the study during their routine neurological evaluation.

Micoulaud-Franchi 2022 *(Continued)*

	Sampling place: Clinical Neurophysiology Department of the Marseille University Hospital, Marseille. **Exclusion criteria**: Insufficient capacity to consent and to understand and answer the self-report questionnaires.
Patient characteristics and setting	**Country**: France **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 149 **n analysed**: 149 **% female participants**: 57 **Mean age**: 38.5 years **Patient characteristics**: people with epilepsy
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: French **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 8 **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: A neurologist with expertise in the field of epilepsy and its psychiatric aspects **Prevalence GAD**: 0.26 **Target condition(s):** GAD **Diagnostic criteria**: DSM-IV-TR **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. No additional information could be obtained from the authors.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			

Micoulaud-Franchi 2022 *(Continued)*

Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Mughal 2021

Study characteristics

Mughal 2021 *(Continued)*

Patient Sampling	**Sampling method:** Consecutive sampling. **Sampling place:** Urban methadone maintenance therapy clinic. **Exclusion criteria:** None reported.
Patient characteristics and setting	**Country:** Vietnam **Setting:** Screening in patients from clinical setting with a specific condition **Ethnicity:** No information **n included:** 420 **n analysed:** 400 **% female participants:** 80 **Mean age:** 41.3 years **Patient characteristics:** Adults enroled in an urban methadone maintenance therapy clinic
Index tests	**Index tests used:** GAD-7 **Index test administration:** Interview **Index test language:** Vietnamese **Index test validation:** No information **Post hoc 'best' cut-off (≥):** 3 **Post hoc 'best' cut-off based on:** Youden's index
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: A research assistant and a physician who were trained by a US-based psychiatrist and epidemiologist in administration of all tools. **Target condition(s):** GAD **Prevalence GAD**: 0.04 **Diagnostic criteria**: DSM-V **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS:** 0 to 1 week
Comparative	
Notes	The objective of the study was DTA assessment. No additional information could be obtained from the authors. Diagnostic information for some cut-offs was extracted using the WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/). The reported cut-offs 1, 2, and 3 were outside the core range defined for this review. Thus, this study was not included in the primary analyses.
Methodological quality	

Mughal 2021 *(Continued)*

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Unclear		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Nath 2018

Study characteristics	
Patient Sampling	**Sampling method**: Using a sampling design stratified according to answering positive or negative (saying yes or no, respectively) on either of the two Whooley questions. **Sampling place**: Inner-city maternity service. **Exclusion criteria**: Women aged under 16 years, women who declined to answer Whooley questions, those who had a termination or miscarriage prior to baseline interview, or had already attended for their maternity booking appointment elsewhere in the UK.
Patient characteristics and setting	**Country**: United Kingdom **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Mixed **n included**: 545 **n analysed**: 528 **% female participants**: 100 **Mean age**: 32.8 years **Patient characteristics**: Pregnant women from an inner-city maternity service; "ethnically and socioeconomically diverse population."
Index tests	**Index tests used**: GAD-2 **Index test administration**: Web-based self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: NA **Post hoc 'best' cut-off based on**: NA
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Researcher **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.15 **Prevalence GAD**: 0.05 **Diagnostic criteria**: DSM-V **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: No time lag
Comparative	

Nath 2018 *(Continued)*

Notes	The objective of the study was DTA assessment. Prevalence was calculated from positive predictive value, sensitivity and specificity using the formula suggested by Taylor and colleagues (Taylor 2021). Additional information obtained: researchers administering the reference standard were blinded for the index test results.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		

Nath 2018 *(Continued)*

Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		Low risk

Osorio 2015

Study characteristics	
Patient Sampling	**Sampling method**: From outpatient applicants using random numbers table. Approximately one-third of the participants were randomly selected. **Sampling place**: Specialised outpatient cancer hospital. **Exclusion criteria**: Severe cognitive impairment as qualitatively evaluated by the applicator, and the absence of clinical conditions that would affect responses to the instruments.
Patient characteristics and setting	**Country**: Brazil **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 434 **n analysed**: 400 **% female participants**: 61.5 **Mean age**: No information **Patient characteristics**: Outpatient cancer population
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Brazilian Portuguese **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: 6 **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: By professionals who were trained to administer the instrument **Target condition(s):** AAD **Prevalence AAD**: 0.45 **Diagnostic criteria**: DSM-IV

Osorio 2015 *(Continued)*

	Were the researchers administering the RS blinded for the index test results? Yes
Flow and timing	**Time lag between index test and RS**: At least 7 and at most 14 days after the first stage of data collection.
Comparative	
Notes	The objective of the study was DTA assessment. No additional information could be obtained from the authors.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern

Osorio 2015 *(Continued)*

DOMAIN 4: Flow and timing	
Was there an appropriate interval between index test and reference standard?	Yes
Did all patients receive the same reference standard?	Yes
Were all patients included in the analysis?	Yes
Could the patient flow have introduced bias?	Low risk

Pranckeviciene 2022

Study characteristics	
Patient Sampling	**Sampling method**: The study invitation with a link to an online survey was disseminated through social media, university websites and university emails. **Sampling place**: Various universities in Lithuania **Exclusion criteria**: None reported
Patient characteristics and setting	**Country**: Lithuania **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: White **n included**: 560 **n analysed**: 560 **% female participants**: 82 **Mean age**: 22.7 years **Patient characteristics**: Lithuanian students
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Computer-based self-report **Index test language**: Lithuanian **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 7, GAD-2: 3 **Post hoc 'best' cut-off based on**: Youden's Index
Target condition and reference standard(s)	**Reference standard**: CIS-R **Reference standard applied by**: professional psychologists trained to perform CIS-R assessment **Target condition(s):** GAD **Prevalence GAD**: 0.1

Pranckeviciene 2022 *(Continued)*

	Diagnostic criteria: ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Four weeks
Comparative	
Notes	The objective of the study was DTA assessment. Raw data were provided by the principal investigator. DTA information could be calculated for all cut-offs. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	

Pranckeviciene 2022 *(Continued)*

Are there concerns that the target condition as defined by the reference standard does not match the question?		Low concern
DOMAIN 4: Flow and timing		
Was there an appropriate interval between index test and reference standard?	No	
Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		High risk

Qu 2016

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive **Sampling place**: General hospital haematology department **Exclusion criteria**: Previous history of severe mental illness including schizophrenia, schizoaffective disorder, persistent delusions, sexual disorders, bipolar disorder, psychiatric disorders due to epilepsy, psychiatric disorders, developmental retardation with psychiatric disorders, currently taking medications that can cause anxiety, depressive symptoms, speech impairment and hearing impairment; those who have received less than 6 years of education.
Patient characteristics and setting	**Country**: China **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Asian **n included**: 585 **n analysed**: 548 **% female participants**: 41.2 **Mean age**: No information **Patient characteristics**: Leukaemia patients from a general hospital haematology department
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Chinese **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: No information **Post hoc 'best' cut-off based on**: No information

Qu 2016 *(Continued)*

Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Three psychiatrists **Target condition(s):** AAD **Prevalence AAD**: 0.13 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. No additional information could be obtained from the authors.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		

Qu 2016 *(Continued)*

Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Scott 2019

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive sampling. **Sampling place**: Royal Prince Alfred Hospital Comprehensive Epilepsy Service in Sydney, Australia. **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: Australia **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 106 **n analysed**: 106 **% female participants**: 57.5 **Mean age**: 40.1 years **Patient characteristics**: Adults with confirmed epilepsy
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: No information **Post hoc 'best' cut-off based on**: No information

Scott 2019 *(Continued)*

Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Psychologist **Target condition(s)**: AAD, GAD **Prevalence AAD**: 0.31 **Prevalence GAD**: 0.14 **Diagnostic criteria**: DSM-V **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Same day
Comparative	
Notes	The objective of the study was DTA assessment. Raw data were provided by the principal investigator. DTA information could be calculated for all cut-offs. A typo in the dataset was corrected. Additional information obtained: there was no time lag between the index test (GAD-7) and the reference standard (MINI); they were conducted at the same time. Researchers administering the MINI were blinded for the GAD-7 results. The questionnaires were completed in pen-and-paper format in a waiting room, and then the participants were administered the MINI once they had finished. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	

Scott 2019 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Seo 2014

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive sampling. **Sampling place**: Outpatient epilepsy clinics of one secondary and four tertiary care hospitals at Daegu city. **Exclusion criteria**: People who had insufficient information in their medical records, who had mental retardation or serious medical, neurological, or psychiatric disorders that prevented them from understanding the questionnaire and cooperating with the study, and who declined to participate in answering the questionnaires were excluded.
Patient characteristics and setting	**Country**: South Korea **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Asian **n included**: 243 **n analysed**: 243 **% female participants**: 39.5 **Mean age**: 39.8 years

Seo 2014 *(Continued)*

	Patient characteristics: Current diagnosis of epilepsy; were taking one or more antiepileptic drugs for at least one year.
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Korean **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 6 **Post hoc 'best' cut-off based on**: Minimising the Euclidean distance from point (sensitivity and specificity) to point in the x–y plane.
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: neuropsychologist **Target condition(s):** GAD **Prevalence GAD**: 0.21 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No time lag
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not provide additional information. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		

Seo 2014 *(Continued)*

Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Seo 2015

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive new patients with migraine who had consecutively visited an outpatients' headache clinic. **Sampling place**: Department of Neurology at Kyungpook National University Hospital outpatients headache clinic. **Exclusion criteria**: People were excluded if they were unable to cooperate in the psychiatric interview or had difficulty understanding the questionnaire because of illiteracy, mental retardation, serious medical, neurological, or psychiatric disorders, and alcohol or drug abuse. People with a probable migraine and those declining the interview were also excluded.
Patient characteristics and setting	**Country**: South Korea **Setting**: Screening in patients from clinical setting with a specific condition

Seo 2015 *(Continued)*

	Ethnicity: Asian **n included**: 146 **n analysed**: 146 **% female participants**: 88.6 **Mean age**: 40.7 years **Patient characteristics**: People with migraine
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: Korean **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: GAD-7: 6, GAD-2: 2 **Post hoc 'best' cut-off based on**: Minimise the Euclidean distance from point (sensitivity and specificity) to point in the x-y plane.
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: No information **Target condition(s):** GAD **Prevalence GAD**: 0.22 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not provide additional information. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Low concern

Seo 2015 *(Continued)*

DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Seo 2017

Study characteristics	
Patient Sampling	**Sampling method**: No information. **Sampling place**: Four tertiary care hospitals. **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: South Korea **Setting**: Screening in patients from clinical setting with a specific condition

Seo 2017 *(Continued)*

	Ethnicity: Asian **n included**: No information **n analysed**: 160 **% female participants**: No information **Mean age**: No information **Patient characteristics**: People with tension-type headache
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: Korean **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: GAD-7: 8, GAD-2: 2 **Post hoc 'best' cut-off based on**: Sum of sensitivity and specificity
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: No information **Target condition(s):** GAD **Prevalence GAD**: 0.21 **Diagnostic criteria**: DSM-IV (reviewer assumption based on the use of the MINI version 5.0.0) **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not provide additional information. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear

Seo 2017 *(Continued)*

DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Unclear		
Could the patient flow have introduced bias?		Unclear risk	

Shih 2022

Study characteristics	
Patient Sampling	**Sampling method**: All patients were invited to participate in the study during a routine neurological evaluation. **Sampling place**: Neurology ward and clinic of Taipei Veterans General Hospital. **Exclusion criteria**: Severe chronic medical conditions other than epilepsy and insufficient capacity to consent, to understand, or to answer the self-report questionnaires.

Shih 2022 *(Continued)*

Patient characteristics and setting	**Country**: Taiwan **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Asian **n included**: 107 **n analysed**: 107 **% female participants**: 56.1 **Mean age**: 33.2 years **Patient characteristics**: Native Taiwanese patients with epilepsy diagnosed based on ILAE (International League Against Epilepsy) criteria.
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Mandarin **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 8 **Post hoc 'best' cut-off based on**: Maximisation of sensitivity and specificity
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: The structured questionnaire was administered either by a neurologist or by a psychiatrist with expertise in the field of epilepsy. **Target condition(s):** GAD **Prevalence GAD**: 0.16 **Diagnostic criteria**: DSM-IV-TR **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: No information.
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not be contacted. Cut-offs presented in Table 4 are given with a '>' sign, which gives the best post hoc cut-off of 8 (>7). However, in the text, the cut-off point that maximised both sensitivity and specificity is given as 7. Diagnostic information for some cut-offs was extracted using Webplotdigitizer (Rohatgi 2014). DTA information could be calculated for all cut-offs.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns

Shih 2022 *(Continued)*

DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Sidik 2012

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive sampling with random assignment to CIDI with 1:10 ratio for GAD-7 < 5 / PHQ < 10 and 1:2 for GAD-7 ≥ 5/ PHQ ≥ 10. **Sampling place**: Government primary care clinic headed by a family physician in an urban district in Malaysia. **Exclusion criteria**: People who were acutely ill and needed immediate medical attention, and those with communication problems.
Patient characteristics and setting	**Country**: Malaysia **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: No information **n included**: 895 **n analysed**: 146 **% female participants**: 100 **Mean age**: 30.9 years **Patient characteristics**: Women primary care patients
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Malay **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: NA **Post hoc 'best' cut-off based on**: NA
Target condition and reference standard(s)	**Reference standard**: CIDI **Reference standard applied by**: Principal investigator of the study who was a family physician **Target condition(s):** GAD **Prevalence GAD**: 0.12 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Less than four weeks
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not provide additional information. To prevent potential bias due to

Sidik 2012 *(Continued)*

partial verification, sensitivities and specificities were adjusted according to the method suggested by Kohn (Kohn 2022).

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	No		

Sidik 2012 *(Continued)*

Were all patients included in the analysis?	Yes
Could the patient flow have introduced bias?	High risk

Simning 2012

Study characteristics	
Patient Sampling	**Sampling method**: Random (a series of mailings sent (in English and Spanish) to all residents) **Sampling place**: Four public housing high-rises reserved for older adults in Rochester, New York **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: USA **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Black **n included**: 190 **n analysed**: 190 **% female participants**: 58 **Mean age**: 66 years **Patient characteristics**: Socioeconomically disadvantaged older adults
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: English **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: NA **Post hoc 'best' cut-off based on**: NA
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: The first author who completed his psychiatry residency and geriatric psychiatry fellowship **Target condition(s):** AAD **Prevalence AAD**: 0.21 **Diagnostic criteria**: DSM-IV-TR **Were the researchers administering the RS blinded for the index test results?** No

Simning 2012 *(Continued)*

Flow and timing	**Time lag between index test and RS**: No time lag
Comparative	
Notes	The objective of the study was not DTA assessment. Authors provided unpublished raw data enabling calculation of all possible cut-offs. The authors responded to data quality check request and approved the extraction.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			High
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	No		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			

Simning 2012 *(Continued)*

Was there an appropriate interval between index test and reference standard?	Yes
Did all patients receive the same reference standard?	Yes
Were all patients included in the analysis?	Yes
Could the patient flow have introduced bias?	Low risk

Tong 2016

Study characteristics	
Patient Sampling	**Sampling method**: Consecutively recruited from epilepsy outpatients. **Sampling place**: Epilepsy outpatient clinic, West China Hospital. **Exclusion criteria**: Patients with psychogenic non-epileptic seizures or other significant neurological/psychiatric disorders, such as cognitive deficits, aphasia or schizophrenia, which might hamper appropriate understanding and completion of the questionnaire, were excluded.
Patient characteristics and setting	**Country**: China **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Asian **n included**: No information **n analysed**: 213 **% female participants**: 48.8 **Mean age**: 29.8 years **Patient characteristics**: Chinese citizens currently diagnosed with epilepsy
Index tests	**Index tests used**: GAD-7 **Index test administration**: Self-report **Index test language**: Chinese **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: > 6 **Post hoc 'best' cut-off based on**: Youden's index
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Qualified psychiatrist **Target condition(s):** GAD

Tong 2016 *(Continued)*

	Prevalence GAD: 0.23 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not provide additional information.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Unclear risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	

Tong 2016 *(Continued)*

Are there concerns that the target condition as defined by the reference standard does not match the question?		Low concern
DOMAIN 4: Flow and timing		
Was there an appropriate interval between index test and reference standard?	Unclear	
Did all patients receive the same reference standard?	Yes	
Were all patients included in the analysis?	Yes	
Could the patient flow have introduced bias?		Unclear risk

Van Heyningen 2018

Study characteristics	
Patient Sampling	**Sampling method**: Systematic sampling. Every third woman attending their first antenatal visit was invited. **Sampling place**: The study was located at a primary care antenatal clinic within a busy urban community health centre in the Cape Town metropolis. **Exclusion criteria**: Those who were aged below 18 years or unable to converse in any of the three local languages (English, Afrikaans or isiXhosa) were excluded.
Patient characteristics and setting	**Country**: South Africa **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Mixed **n included**: 376 **n analysed**: 376 **% female participants**: 100 **Mean age**: 26.8 years **Patient characteristics**: Pregnant women attending a primary care antenatal clinic
Index tests	**Index tests used**: GAD-2 **Index test administration**: Interview **Index test language**: English (n = 312), IsiXhosa (n = 37), Afrikaans (n = 27) **Index test validation**: Yes for English **Post hoc 'best' cut-off (≥)**: 2 **Post hoc 'best' cut-off based on**: LR+

Van Heyningen 2018 *(Continued)*

Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: registered counsellor, with a Bachelor of Psychology (Honours) degree in counselling **Target condition(s):** AAD **Prevalence AAD**: 0.23 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Consecutively
Comparative	
Notes	The objective of the study was DTA assessment. Authors provided additional information. Additional information obtained: target condition: "any anxiety disorder" covers all DSM-IV diagnostic categories; namely, generalised anxiety disorder, social phobia, agoraphobia, specific phobia, post-traumatic stress disorder, and obsessive-compulsive disorder. Researchers administering the reference standard were blinded for the results of the index test. Prevalence was calculated from positive predictive value, sensitivity, and specificity using the formula suggested by Taylor 2021.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	

Van Heyningen 2018 *(Continued)*

Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

Veisy 2021

Study characteristics	
Patient Sampling	**Sampling method**: Participants were recruited from the headache clinics of university hospitals and headache speciality centres by convenience sampling. **Sampling place**: Headache clinics of university hospitals and headache speciality centres in Tehran, Iran **Exclusion criteria**: Inability to participate in the interview, inability to perceive self-completion questionnaire due to medical condition, mental defectiveness, or receiving medical treatment that impairs comprehension of the questionnaire.
Patient characteristics and setting	**Country**: Iran **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: No information **n included**: 188 **n analysed**: 186 **% female participants**: 70

Veisy 2021 *(Continued)*

	Mean age: 37 years **Patient characteristics**: Adult migraine patients from the headache clinics of university hospitals and headache speciality centres, not on psychotropic medication.
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Self-report **Index test language**: Persian **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: GAD-7: 10, GAD-2: 3 **Post hoc 'best' cut-off based on**: Optimal balance between sensitivity and specificity
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained clinician (PhD candidate in clinical psychology) **Target condition(s):** GAD **Prevalence GAD**: 0.45 **Diagnostic criteria**: DSM-V **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not be contacted. Diagnostic information for some cut-offs was extracted using Webplotdigitizer (Rohatgi 2014).

Methodological quality

Item	**Authors' judgement**	**Risk of bias**	**Applicability concerns**
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	No		
Could the selection of patients have introduced bias?		High risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		

Veisy 2021 *(Continued)*

If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Wild 2014

Study characteristics	
Patient Sampling	**Sampling method**: Population-based cohort (ESTHER). **Sampling place**: General elderly population living at home. **Exclusion criteria**: Not reported.
Patient characteristics and setting	**Country**: Germany **Setting**: Screening in non-clinical setting **Ethnicity**: White **n included**: 509 **n analysed**: 438

Wild 2014 *(Continued)*

	% female participants: 55.2 **Mean age**: 60 years **Patient characteristics**: Elderly people in the community
Index tests	**Index tests used**: GAD-7, GAD-2 **Index test administration**: Interview **Index test language**: German **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: No information **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: SCID **Reference standard applied by**: Trained study doctors **Target condition(s):** GAD **Prevalence GAD**: 0.06 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: 0 to 4 weeks
Comparative	
Notes	The objective of the study was DTA assessment. Authors provided additional diagnostic information for three cut-offs. Additional information obtained: mean age of the participants was 60 years.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		

Wild 2014 *(Continued)*

Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	No		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		High risk	

Zeng 2013

Study characteristics	
Patient Sampling	**Sampling method**: No information. **Sampling place**: Internal department of traditional Chinese medicine. **Exclusion criteria**: None reported.
Patient characteristics and setting	**Country**: China **Setting**: Screening in patients from clinical setting (across conditions) **Ethnicity**: Asian **n included**: 2011 **n analysed**: 1993

Zeng 2013 *(Continued)*

	% female participants: 69.3 **Mean age**: 37 years **Patient characteristics**: Outpatients aged 18 to 65 from internal department of traditional Chinese medicine
Index tests	**Index tests used**: GAD-7 **Index test administration**: Interview **Index test language**: Chinese **Index test validation**: No information **Post hoc 'best' cut-off (≥)**: 6 **Post hoc 'best' cut-off based on**: No information
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: No information **Target condition(s):** GAD **Prevalence GAD**: 0.05 **Diagnostic criteria**: DSM-IV **Were the researchers administering the RS blinded for the index test results?** No information
Flow and timing	**Time lag between index test and RS**: No information
Comparative	
Notes	The objective of the study was DTA assessment. Authors could not provide additional information.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Unclear		
If a threshold was used, was it pre-specified?	No		

Zeng 2013 *(Continued)*

Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Unclear		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Unclear risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Unclear		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Unclear risk	

Zhong 2015

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive. **Sampling place**: Prenatal care clinics. **Exclusion criteria**: This cross-sectional study was a part of the Pregnancy Outcomes, Maternal and Infant Study (PrOMIS). Change of address or inaccurate contact information was an exclusion criterion in PrOMIS.
Patient characteristics and setting	**Country**: Peru **Setting**: Screening in patients from clinical setting with a specific condition **Ethnicity**: Hispanic or Latino **n included**: 2978

Zhong 2015 *(Continued)*

	n analysed: 946 **% female participants**: 100 **Mean age**: 28 years **Patient characteristics**: Pregnant women enroled in prenatal care clinics
Index tests	**Index tests used**: GAD-7 **Index test administration**: Interview **Index test language**: Spanish **Index test validation**: Yes **Post hoc 'best' cut-off (≥)**: 7 **Post hoc 'best' cut-off based on**: Youden's index
Target condition and reference standard(s)	**Reference standard**: CIDI **Reference standard applied by**: Licenced research psychologists were recruited and received structured training on administration of the CIDI **Target condition(s):** GAD **Prevalence GAD**: 0.01 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: 15 days
Comparative	
Notes	The objective of the study was DTA assessment. Authors provided additional information. DTA information could be calculated for all cut-offs. Additional information obtained: researchers administering the reference standard were blinded for the results of the index test.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			
Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Yes		
Could the selection of patients have introduced bias?		Low risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear

Zhong 2015 *(Continued)*

DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	Yes		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	Yes		
Could the conduct or interpretation of the index test have introduced bias?		Low risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	No		
Were all patients included in the analysis?	No		
Could the patient flow have introduced bias?		High risk	

Zinchuk 2021

Study characteristics	
Patient Sampling	**Sampling method**: Consecutive. **Sampling place**: Clinical centre for neuropsychiatry inpatients and outpatients. Moscow Research and Clinical Centre for Neuropsychiatry. **Exclusion criteria**: Cognitive impairment which prevented a patient from understanding the meaning of the questionnaires and the interview, significant somatic and neurological comorbidities, history of interictal or postictal psychotic disorders, alcohol and substance use disorders in the

Zinchuk 2021 *(Continued)*

	past 12 months, an epileptic seizure 48 hours before or after completing the questionnaires.
Patient characteristics and setting	**Country**: Russia **Setting**: Screening in patients with a specific disease/risk **Ethnicity**: White **n included**: 233 **n analysed**: 233 **% female participants**: 65.2 **Mean age**: 41.1 years **Patient characteristics**: People with epilepsy
Index tests	**Index tests used**: GAD-7, HADS-A **Index test administration**: Self-report **Index test language**: Russian **Index test validation**: No **Post hoc 'best' cut-off (≥)**: AAD: 9, GAD: 10 **Posthoc 'best' cut-off based on**: Youden's index
Target condition and reference standard(s)	**Reference standard**: MINI **Reference standard applied by**: Experienced psychiatrists **Target condition(s):** AAD, GAD **Prevalence AAD**: 0.42 **Prevalence GAD**: 0.27 **Diagnostic criteria**: DSM-IV+ICD-10 **Were the researchers administering the RS blinded for the index test results?** Yes
Flow and timing	**Time lag between index test and RS**: Same day
Comparative	
Notes	The objective of the study was DTA assessment. Authors provided additional information. Diagnostic information on additional GAD-7 cut-offs was obtained. Additional information obtained: this study validated the Generalised Anxiety Disorder-7 (GAD-7) in Russian people with epilepsy. There was no time lag between the GAD-7 and the MINI interview. The researchers administering the MINI were blinded for the results of the GAD-7.

Methodological quality

Item	Authors' judgement	Risk of bias	Applicability concerns
DOMAIN 1: Patient selection			

Zinchuk 2021 *(Continued)*

Did the study avoid unnecessary exclusions?	Yes		
Was a consecutive or random sample of patients enrolled?	Unclear		
Could the selection of patients have introduced bias?		Unclear risk	
Are there concerns that the included patients and setting do not match the review question?			Unclear
DOMAIN 2: Index test (All tests)			
Were the index test results interpreted without knowledge of the results of the reference standard?	Yes		
If a threshold was used, was it pre-specified?	No		
Is the risk avoided that the cut-off(s) presented/available have been chosen post-hoc due to optimal performance values? (NA if there is a predefined main cut-off)	No		
Could the conduct or interpretation of the index test have introduced bias?		High risk	
Are there concerns that the index test, its conduct, or interpretation differ from the review question?			Low concern
DOMAIN 3: Reference standard			
Is the reference standard likely to correctly classify the target condition?	Yes		
Were the reference standard results interpreted without knowledge of the results of the index tests?	Yes		
Could the reference standard, its conduct, or its interpretation have introduced bias?		Low risk	
Are there concerns that the target condition as defined by the reference standard does not match the question?			Low concern
DOMAIN 4: Flow and timing			
Was there an appropriate interval between index test and reference standard?	Yes		
Did all patients receive the same reference standard?	Yes		
Were all patients included in the analysis?	Yes		
Could the patient flow have introduced bias?		Low risk	

AAD: any anxiety disorder; **BAI:** Beck Anxiety Inventory; **CBT:** cognitive behavioural therapy; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **COPD:** chronic obstructive pulmonary disease; **CVD:** cardiovascular disease; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **DSM(-IV-TR/-5):** Diagnostic and Statistical Manual of Mental

Disorders (4th edition Text Revision/5th edition); **DTA:** diagnostic test accuracy; **GAD:** generalised anxiety disorder; **GAD-7/2:** 7-item/2-item Generalized Anxiety Disorder scales; **HADS(-A):** Hospital Anxiety and Depression Scale (-Anxiety subscale); **ICD:** International Statistical Classification of Diseases and Related Health Problems; **LR+:** positive likelihood ratio; **MINI:** Mini-International Neuropsychiatric Interview; **PHQ:** Patient Health Questionnaire; **RCT:** randomised controlled trial; **ROC curve:** receiver operating characteristic curve; **RS:** reference standard; **SCID:** Structured Clinical Interview for DSM; **SD:** standard deviation

Characteristics of excluded studies *[ordered by study ID]*

Study	Reason for exclusion
Afshari 2020	Not appropriate type of study: examination of revised reinforcement sensitivity theory
Ahmad 2017	No appropriate reference standard: no reference standard used
AlKhathami 2022	No diagnostic data available to extract. The corresponding author did not respond to our information request.
Andersen 2023	Ineligible participants: pre-selected participants had to report "severe depression or anxiety during the last month"
Austin 2022	No appropriate reference standards. Reference standard used: SAGE (Screening Assessment for Guiding Evaluation)
Basaraba 2023	DTA information not available in the article (reports findings for subgroups only, but prevalence for all participants). No response from the corresponding author.
Baykan 2019	Ineligible study design: case-control study
Beard 2014	Other: did not fit the screening pathway. Mental health pathway
Beesdo-Baum 2012	DTA information not available in the article. No response from the corresponding author.
Belk 2016	No appropriate reference standard. Did not use a structured clinical interview as reference standard.
Brünahl 2017	No data to extract. No response from the corresponding author.
Bunevicius 2013	No appropriate index test. Index tests used: HADS-A, STAI and SSAI.
Bushey 2021	No appropriate target conditions. Target condition: global improvement.
Byrd-Bredbenner 2021	No appropriate reference standard. No reference standard used. Only exploratory principal components analysis.
Camarata 2023	No appropriate target conditions. Reference standard seems to have addressed depression only.
Cano-Vindel 2018	Other. Secondary publication.
Castro 2023	No appropriate target conditions. Reference standard addressed depression only.
Celano 2013	Ineligible study type. No appropriate reference standard. Aimed to identify depression and anxiety disorders without a reference standard.
Christensen 2011	Ineligible design: study design has elements of case-control. Participants proven to be healthy at the study onset. Incidence of GAD was assessed at six months.
Christensen 2014	No data to extract. No response from the corresponding author.

Study	Reason for exclusion
Chun 2017	No data to extract. Authors could not be contacted.
Clary 2021	Other. Secondary publication.
Dadi 2016	No data to extract. Authors could not be contacted.
Dear 2011	Ineligible study design: RCT comparing two CBT methods in people with anxiety. People meeting the Diagnostic and Statistical Manual of Mental Disorders criteria for a diagnosis of GAD were recruited.
Donker 2011	Did not fit the screening pathways. Mental health pathway.
Dyukova 2021	No data to extract. No response from the corresponding author.
Fairbrother 2019	Other. Time between index test and reference standard was six weeks.
Fatima 2021	Ineligible study type: not a diagnostic study (only prevalence).
García-Campayo 2009	No DTA information to extract. Authors could not be contacted.
García-Campayo 2010	No appropriate reference standard: "diagnosis performed under standard medical practice conditions according to the DSM-IV-TR classification"; comparison of GAD-7 and HADS-A with "clinical diagnosis of reference". Ineligible study design: case-control study.
García-Campayo 2012a	Ineligible study design: case-control study.
García-Campayo 2012b	No DTA information to extract. Authors could not be contacted.
García-Campayo 2012c	Ineligible study design: case-control study
García-Campayo 2012d	No DTA information to extract. Author responded that the study was conducted a long time ago, and that the database was not maintained.
Giuliani 2021	No appropriate reference standard. PHQ-2 and GAD-2 psychometric properties were compared with PHQ-9 and GAD-7.
Goldstein 2023	Ineligible participants: participants suffered from mental health conditions (dissociative seizures).
Gong 2021	Other. The study used matching based on GAD-7 and HADS scores, which introduced a spectrum effect. Additionally, the population size was not known. Hence, no correction of the number of people in the different diagnostic categories could be undertaken.
Graham 2019	No DTA information to extract. The corresponding author could not provide data.
Greenhalgh 2016	Other. Poster presentation only. Limited information received from co-author Alex Mitchell. Excluded because no reliable 2x2 table could be produced with the available data and communications with the author.
Gul 2015	Ineligible study design: case-control study. The control group did not receive GAD-7.
Gulpers 2022	No appropriate target conditions. Target conditions: panic disorder with and without agoraphobia, agoraphobia, lifetime panic disorder.
Gündüz 2021	Ineligible study design: case-control study
Hallgren 2007	Ineligible study type: short review of anxiety screening instruments - not an original study.

Study	Reason for exclusion
Henderson 2014	Ineligible participants: participants were between the ages of 17 and 59 years (mean = 23.57 years; 44 participants did not provide their age). No separate data available for adults.
Hinz 2017	No appropriate reference standard. Confirmatory factor analyses were done. No reference standard was used.
Holper 2021	No appropriate reference standard. Diagnoses made by psychiatrists - structured interview unlikely.
Houston 2011	No DTA information to extract. No response from the corresponding author.
Hughes 2018	No appropriate reference standard. GAD-7 was used as reference standard.
Jansen 2019	Study does not address DTA, but has, in principle, collected the data. Authors did not reply to data requests.
Jin 2020	No appropriate reference standard. In-person interviewer assessment without a reference standard.
Jyothi 2020	Ineligible study design: MINI only applied to participants who were screened positive.
Kemp 2023	Data not available. Published analyses did not address diagnostic accuracy. Author did not respond to information requests.
Kertz 2013	Other. Did not fit the screening pathways. Mental health pathway.
Khani 2017	No DTA information to extract. No response from the authors.
Kim 2023	Ineligible participants: people admitted to a psychiatric department. Ineligible target condition (suicidality).
Konkan 2013	Ineligible study design: case-control study.
Konstantakopoulos 2013	Review; not a primary study.
Kounali 2020	No appropriate target conditions. The reference standard CIS-R was self-administered. Our inclusion criteria required that the reference standard be administered by interviewing the participants.
Kroenke 2010	Systematic review; not a primary study.
Kumar 2018	Not a diagnostic accuracy study. Ineligible participants. Study included only people already diagnosed with generalised anxiety disorder, followed by administering the GAD-7 scale.
Ladouceur 1999	No appropriate index test. Awkward case-control study not using the GAD-7/-2 but the 10 item GAD-Q. Furthermore, it is not fully clear whether the reference standard would be adequate.
Lan 2021	No appropriate reference standard (Patient Reported Outcomes Measurement Information System Anxiety Computer Adaptive Test (PROMIS-A-CAT)).
Lee 2020	Did not fit the screening pathways. Mental health pathway.
Levitt 2021	No appropriate reference standard. Reference standard DART (Diagnostic Assessment Research Tool).
Li 2022	No appropriate reference standard, and no appropriate target conditions. Only anxiety symptoms were assessed; no reference standard to diagnose the target conditions.

Study	Reason for exclusion
Lin 2021	No appropriate reference standard. No diagnosis was made. GAD-7 and Self-rating Anxiety Scale (SAS) were compared.
Liu 2023	No appropriate reference standard (no specific validated interview).
Lowe 2008	Ineligible study type: validation study, not a diagnostic test accuracy study (no reference standard used).
Ma 2021	Ineligible study type: it exclusively included people with somatoform disorders and then constructed subgroups (e.g. panic disorder - yes/no). Diagnostic accuracy data not presented.
Mansbach 2015	No DTA information to extract. No response from the authors.
Means-Christensen 2006	No appropriate index test. Five-item Anxiety and Depression Detector (ADD).
Minc 2017	Other. Secondary publication.
Monteiro 2020	No appropriate reference standard. No reference standard; only factor analysis was performed.
Monterrosa-Blanco 2021	Not an eligible primary study, and no appropriate reference standard. Letter to the editor describing psychometric properties of the GAD-7 scale.
Munoz-Navarro 2017a	Other. Secondary publication without relevant additional information. The article is primarily a conceptual paper referring to the results of the Cano-Vindel 2018 study.
Munoz-Navarro 2017b	Did not fit the screening pathways. Mental health pathway.
Muntingh 2013	No appropriate index test (Patient Health Questionnaire).
Mussell 2008	Other. Secondary publication to Spitzer 2006 and Kroenke 2007, without relevant additional information.
Nath 2019	Other. Secondary publication without relevant additional information. This is an abstract of the same dataset published by Nath 2020.
Nath 2020	No DTA information to extract. No response from authors.
Priyanka 2010	Data not available. Conference presentation of a small sample (n = 33), which could generate diagnostic accuracy data, but had no data to extract. No response from the authors.
Purves 2019	Not a diagnostic study. The researchers investigated GAD-specific genome-wide associations in a biobank.
Purvis 2019	No appropriate reference standard (expert clinical consensus).
Qu 2015	Did not fit the screening pathways. Mental health pathway.
Rady 2021	Ineligible study type: included only people who scored at least 5 on PHQ-9/GAD-7.
Ringoir 2014	No appropriate target conditions. No anxiety diagnosis was made. Symptoms of anxiety were measured with the Generalized Anxiety Disorder 7-item scale.
Rogers 2021	No appropriate participants. Participants aged 15 to 32 years. No separate data available for adults.
Ruiz 2009	No DTA information to extract. Authors could not be contacted.

Study	Reason for exclusion
Sawaya 2016	No appropriate reference standard. No reference standard used: one of four psychiatrists or two clinical psychologists assessed participants.
Short 2019	No appropriate reference standard. Conference proceeding of a study likely not to have an adequate reference standard (possibly using GAD-2 as reference standard). Authors could not be contacted.
Shrestha 2020	No appropriate participants. For inclusion in the study, participants had to screen high for anxiety (PSWQ-A: Penn State Worry Questionnaire - Abbreviated).
Silva 2018	No appropriate reference standard. GAD-7 cut-off was used to assess the prevalence of GAD.
Simpson 2014	No appropriate reference standard. No reference standard used. Diagnoses were made by experienced psychiatrists.
Sinesi 2019	Other. Pre-selection based on GAD score. Mental health pathway.
Skapinakis 2007	Not a primary study; reports the study of Kroenke 2007
Snijkers 2021	No appropriate reference standard: comparison of GAD-7 and HADS-A; no diagnostic interview.
Soni 2019	Conference proceeding. No GAD-7 data to extract. Author contacted but could not provide data.
Sousa 2015	No appropriate reference standard. Criterion validity was assessed by comparing GAD-7 scores with those obtained by HADS and EQ-5D (a quality of life questionnaire). STAI was mainly used as a screening indicator for patient inclusion.
Staples 2019	Ineligible study design: the study included two data samples. The first sample came from four related RCTs, which used GAD-7 cut-off as an exclusion criterion.
Swinson 2006	Not a primary study: commentary on the study by Spitzer 2006.
Tang 2019	No appropriate reference standard. Abstract on an analysis evaluation of the Edmonton Symptom Assessment System-Revised against the GAD-7. A standardised interview as reference standard is not mentioned.
Terluin 2012	Not a primary study: letter to the editor, commenting on Sidik 2012.
Tonev 2019	No appropriate reference standard. Conference proceeding probably using the GAD-7 as reference standard.
Tran 2020	No appropriate reference standard. The authors could not be contacted.
Van Dooren 2016	No DTA information to extract. The authors could not be contacted.
Vasiliadis 2015	No appropriate reference standard. Used ESA (Enquête sur la Santé des Aînés survey) as reference standard.
Villarreal-Zegarra 2023	No appropriate reference standard. No specific validated interview.
Wang 2021	No appropriate reference standard. No reference standard (people were diagnosed with anxiety disorders by specialists).
Webb 2008	No appropriate index test. Index test used: Generalized Anxiety Disorder Questionnaire-IV.

Study	Reason for exclusion
Wiltink 2017	No appropriate reference standard. Two studies focusing on the evaluation of the mini-SPIN (Mini-Social Phobia INventory): substudy 1 (sample of psychiatry in- and outpatients) with clinical assessment as reference standard (but no data presented for the GAD-7); substudy 2 (community sample) without any reference standard.
Zachar-Tirado 2021	No appropriate reference standard. Anxiety and depression diagnoses were based on "Neuropsychological assessment", not diagnostic interviews.
Zhang 2021	Ineligible study type: not a diagnostic study; only psychometric validation.
Zinchuk 2022	Did not fit the screening pathways. Mental health pathway.
Zirke 2013	No data to extract. Authors contacted, but data no longer available.

BAI: Beck Anxiety Inventory; **CBT:** cognitive behavioural therapy; **CIS-R:** Clinical Interview Schedule-Revised; **DTA:** diagnostic test accuracy; **GAD:** generalised anxiety disorder; **HADS-A:** Hospital Anxiety and Depression Scale - Anxiety subscale; **MINI:** Mini-International Neuropsychiatric Interview; **PHQ:** Patient Health Questionnaire; **RCT:** randomised controlled trial; **SSAI:** Spielberger State Anxiety Inventory; **STAI:** State-Trait Anxiety Inventory

Characteristics of studies awaiting classification *[ordered by study ID]*

Burger 2023

Patient Sampling	Convenience sampling
Patient characteristics and setting	71 South African adult male rugby union players
Index tests	Generalized Anxiety Disorder 7-item (GAD-7) questionnaire
Target condition and reference standard(s)	Any anxiety disorder / MINI (Mini-International Neuropsychiatric Interview)
Flow and timing	
Comparative	
Notes	Study identified in last update search (January 2024) – likely to be included in future updates. Publication only presents a Receiver Operating Characteristic (ROC) curve – author is willing to provide data on sensitivity and specificity at multiple cut-off points. Communication ongoing.

Characteristics of ongoing studies *[ordered by study ID]*

Aslan 2023

Study name	Incidence and predictors of neuropsychiatric manifestations following a traumatic brain injury in referral hospitals in Dodoma, Tanzania: a protocol of a prospective longitudinal observational study
Target condition and reference standard(s)	Generalised anxiety disorder, MINI (Mini-International Neuropsychiatric Interview)
Index and comparator tests	Generalized Anxiety Disorder 7-item (GAD-7) questionnaire

Aslan 2023 *(Continued)*

Starting date	June 2023
Contact information	azannaj@gmail.com
Notes	

DATA

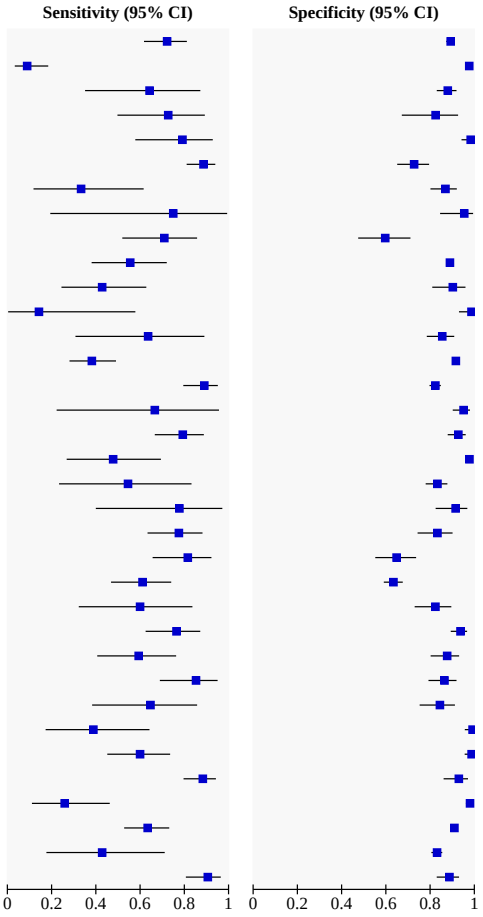
Presented below are all the data for all of the tests entered into the review.

Table Tests. Data tables by test

Test	No. of studies	No. of participants
1 GAD-7 - GAD	35	15272
2 GAD-7 - AAD	22	10245
3 GAD-2 - GAD	24	11428
4 GAD-2 - AAD	19	9973

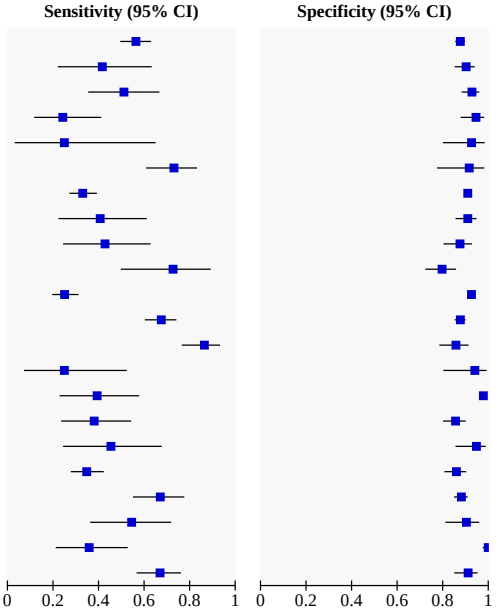
Test 1. GAD-7 - GAD

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ahn 2019	65	114	25	953	0.72 [0.62, 0.81]	0.89 [0.87, 0.91]
Belus 2021	6	10	61	425	0.09 [0.03, 0.18]	0.98 [0.96, 0.99]
Bernstein 2018	9	27	5	198	0.64 [0.35, 0.87]	0.88 [0.83, 0.92]
Bisby 2022	16	7	6	33	0.73 [0.50, 0.89]	0.82 [0.67, 0.93]
Budikayanti 2019	19	2	5	120	0.79 [0.58, 0.93]	0.98 [0.94, 1.00]
Chibanda 2016	94	43	12	115	0.89 [0.81, 0.94]	0.73 [0.65, 0.80]
Clover 2020	5	18	10	120	0.33 [0.12, 0.62]	0.87 [0.80, 0.92]
Conway 2016	3	2	1	42	0.75 [0.19, 0.99]	0.95 [0.85, 0.99]
Delgadillo 2012	22	29	9	43	0.71 [0.52, 0.86]	0.60 [0.47, 0.71]
Esser 2018	20	232	16	1872	0.56 [0.38, 0.72]	0.89 [0.88, 0.90]
Garabiles 2020	12	7	16	65	0.43 [0.24, 0.63]	0.90 [0.81, 0.96]
Grech 2019	1	1	6	77	0.14 [0.00, 0.58]	0.99 [0.93, 1.00]
Hitchon 2020	7	20	4	118	0.64 [0.31, 0.89]	0.86 [0.79, 0.91]
Ivanovs 2018	34	115	55	1263	0.38 [0.28, 0.49]	0.92 [0.90, 0.93]
Kroenke 2007	65	157	8	735	0.89 [0.80, 0.95]	0.82 [0.80, 0.85]
Kujanpää 2014	4	7	2	137	0.67 [0.22, 0.96]	0.95 [0.90, 0.98]
Lickova 2021	46	13	12	167	0.79 [0.67, 0.89]	0.93 [0.88, 0.96]
Maric 2022	11	26	12	1154	0.48 [0.27, 0.69]	0.98 [0.97, 0.99]
Marrie 2018	6	40	5	200	0.55 [0.23, 0.83]	0.83 [0.78, 0.88]
Michaelis 2022	7	6	2	65	0.78 [0.40, 0.97]	0.92 [0.83, 0.97]
Micoulaud-Franchi 2016	38	16	11	80	0.78 [0.63, 0.88]	0.83 [0.74, 0.90]
Micoulaud-Franchi 2022	31	39	7	72	0.82 [0.66, 0.92]	0.65 [0.55, 0.74]
Pranckeviciene 2022	33	185	21	321	0.61 [0.47, 0.74]	0.63 [0.59, 0.68]
Scott 2019	9	16	6	75	0.60 [0.32, 0.84]	0.82 [0.73, 0.90]
Seo 2014	39	12	12	180	0.76 [0.63, 0.87]	0.94 [0.89, 0.97]
Seo 2015	19	14	13	100	0.59 [0.41, 0.76]	0.88 [0.80, 0.93]
Seo 2017	29	17	5	109	0.85 [0.69, 0.95]	0.87 [0.79, 0.92]
Shih 2022	11	14	6	76	0.65 [0.38, 0.86]	0.84 [0.75, 0.91]
Sidik 2012	7	1	11	126	0.39 [0.17, 0.64]	0.99 [0.96, 1.00]
Tong 2016	30	2	20	161	0.60 [0.45, 0.74]	0.99 [0.96, 1.00]
Veisy 2021	76	7	10	93	0.88 [0.80, 0.94]	0.93 [0.86, 0.97]
Wild 2014	7	8	20	403	0.26 [0.11, 0.46]	0.98 [0.96, 0.99]
Zeng 2013	59	171	34	1729	0.63 [0.53, 0.73]	0.91 [0.90, 0.92]
Zhong 2015	6	157	8	775	0.43 [0.18, 0.71]	0.83 [0.81, 0.86]
Zinchuk 2021	58	19	6	150	0.91 [0.81, 0.96]	0.89 [0.83, 0.93]



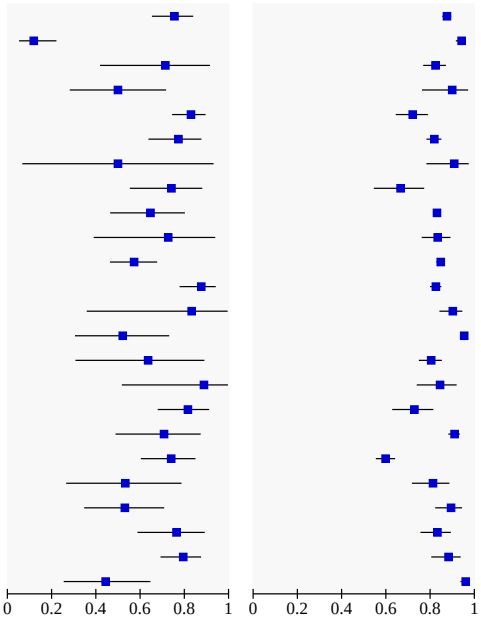
Test 2. GAD-7 - AAD

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ahn 2019	123	115	95	824	0.56 [0.50, 0.63]	0.88 [0.85, 0.90]
Baker 2018	10	19	14	177	0.42 [0.22, 0.63]	0.90 [0.85, 0.94]
Bernstein 2018	22	14	21	182	0.51 [0.35, 0.67]	0.93 [0.88, 0.96]
Clover 2020	9	5	28	88	0.24 [0.12, 0.41]	0.95 [0.88, 0.98]
Conway 2016	2	3	6	38	0.25 [0.03, 0.65]	0.93 [0.80, 0.98]
Delgadillo 2012	49	3	18	33	0.73 [0.61, 0.83]	0.92 [0.78, 0.98]
Esser 2018	82	170	166	1722	0.33 [0.27, 0.39]	0.91 [0.90, 0.92]
Fischer 2014	11	15	16	152	0.41 [0.22, 0.61]	0.91 [0.86, 0.95]
Hitchon 2020	12	15	16	106	0.43 [0.24, 0.63]	0.88 [0.80, 0.93]
Homans 2012	16	30	6	118	0.73 [0.50, 0.89]	0.80 [0.72, 0.86]
Ivanovs 2018	58	91	173	1145	0.25 [0.20, 0.31]	0.93 [0.91, 0.94]
Kroenke 2007	127	95	61	682	0.68 [0.60, 0.74]	0.88 [0.85, 0.90]
Li 2014	64	18	10	109	0.86 [0.77, 0.93]	0.86 [0.79, 0.91]
Makulowich 2018	4	2	12	32	0.25 [0.07, 0.52]	0.94 [0.80, 0.99]
Maric 2022	13	25	20	1145	0.39 [0.23, 0.58]	0.98 [0.97, 0.99]
Marrie 2018	16	30	26	179	0.38 [0.24, 0.54]	0.86 [0.80, 0.90]
Michaelis 2022	10	3	12	55	0.45 [0.24, 0.68]	0.95 [0.86, 0.99]
Osorio 2015	62	31	116	191	0.35 [0.28, 0.42]	0.86 [0.81, 0.90]
Qu 2016	49	56	24	419	0.67 [0.55, 0.78]	0.88 [0.85, 0.91]
Scott 2019	18	7	15	66	0.55 [0.36, 0.72]	0.90 [0.81, 0.96]
Simning 2012	14	0	25	151	0.36 [0.21, 0.53]	1.00 [0.98, 1.00]
Zinchuk 2021	65	12	32	124	0.67 [0.57, 0.76]	0.91 [0.85, 0.95]



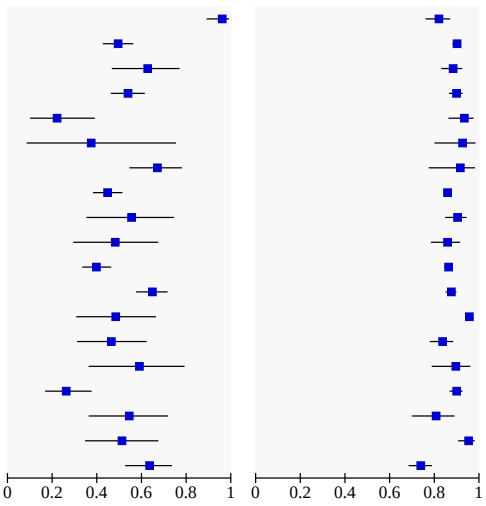
Test 3. GAD-2 - GAD

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ahn 2019	68	132	22	935	0.76 [0.65, 0.84]	0.88 [0.86, 0.90]
Belus 2021	8	25	59	410	0.12 [0.05, 0.22]	0.94 [0.92, 0.96]
Bernstein 2018	10	40	4	188	0.71 [0.42, 0.92]	0.82 [0.77, 0.87]
Bisby 2022	11	4	11	36	0.50 [0.28, 0.72]	0.90 [0.76, 0.97]
Chibanda 2016	88	44	18	114	0.83 [0.74, 0.90]	0.72 [0.64, 0.79]
Christodoulaki 2022	41	96	12	435	0.77 [0.64, 0.88]	0.82 [0.78, 0.85]
Conway 2016	2	4	2	40	0.50 [0.07, 0.93]	0.91 [0.78, 0.97]
Delgadillo 2012	23	24	8	48	0.74 [0.55, 0.88]	0.67 [0.55, 0.77]
Esser 2018	22	320	12	1576	0.65 [0.46, 0.80]	0.83 [0.81, 0.85]
Hitchon 2020	8	23	3	116	0.73 [0.39, 0.94]	0.83 [0.76, 0.89]
Ivanovs 2018	51	209	38	1169	0.57 [0.46, 0.68]	0.85 [0.83, 0.87]
Kroenke 2007	64	155	9	737	0.88 [0.78, 0.94]	0.83 [0.80, 0.85]
Kujanpää 2014	5	14	1	130	0.83 [0.36, 1.00]	0.90 [0.84, 0.95]
Maric 2022	12	54	11	1126	0.52 [0.31, 0.73]	0.95 [0.94, 0.97]
Marrie 2018	7	47	4	194	0.64 [0.31, 0.89]	0.80 [0.75, 0.85]
Michaelis 2022	8	11	1	60	0.89 [0.52, 1.00]	0.85 [0.74, 0.92]
Micoulaud-Franchi 2016	40	26	9	70	0.82 [0.68, 0.91]	0.73 [0.63, 0.81]
Nath 2018	17	45	7	459	0.71 [0.49, 0.87]	0.91 [0.88, 0.93]
Pranckeviciene 2022	40	203	14	303	0.74 [0.60, 0.85]	0.60 [0.55, 0.64]
Scott 2019	8	17	7	74	0.53 [0.27, 0.79]	0.81 [0.72, 0.89]
Seo 2015	17	12	15	102	0.53 [0.35, 0.71]	0.89 [0.82, 0.94]
Seo 2017	26	21	8	105	0.76 [0.59, 0.89]	0.83 [0.76, 0.89]
Veisy 2021	66	12	17	91	0.80 [0.69, 0.88]	0.88 [0.81, 0.94]
Wild 2014	12	16	15	395	0.44 [0.25, 0.65]	0.96 [0.94, 0.98]



Test 4. GAD-2 - AAD

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Ahmadi 2019	75	36	3	165	0.96 [0.89, 0.99]	0.82 [0.76, 0.87]
Ahn 2019	108	92	110	847	0.50 [0.43, 0.56]	0.90 [0.88, 0.92]
Bernstein 2018	27	23	16	176	0.63 [0.47, 0.77]	0.88 [0.83, 0.93]
Christodoulaki 2022	95	41	81	367	0.54 [0.46, 0.62]	0.90 [0.87, 0.93]
Clover 2020	8	6	28	86	0.22 [0.10, 0.39]	0.93 [0.86, 0.98]
Conway 2016	3	3	5	38	0.38 [0.09, 0.76]	0.93 [0.80, 0.98]
Delgadillo 2012	45	3	22	33	0.67 [0.55, 0.78]	0.92 [0.78, 0.98]
Esser 2018	104	239	128	1460	0.45 [0.38, 0.51]	0.86 [0.84, 0.88]
Fischer 2014	15	16	12	151	0.56 [0.35, 0.75]	0.90 [0.85, 0.94]
Hitchon 2020	14	17	15	104	0.48 [0.29, 0.67]	0.86 [0.78, 0.92]
Ivanovs 2018	92	168	139	1068	0.40 [0.33, 0.46]	0.86 [0.84, 0.88]
Kroenke 2007	122	96	66	681	0.65 [0.58, 0.72]	0.88 [0.85, 0.90]
Maric 2022	16	50	17	1120	0.48 [0.31, 0.66]	0.96 [0.94, 0.97]
Marrie 2018	20	34	23	175	0.47 [0.31, 0.62]	0.84 [0.78, 0.88]
Michaelis 2022	13	6	9	52	0.59 [0.36, 0.79]	0.90 [0.79, 0.96]
Nath 2018	20	44	56	397	0.26 [0.17, 0.38]	0.90 [0.87, 0.93]
Scott 2019	18	14	15	59	0.55 [0.36, 0.72]	0.81 [0.70, 0.89]
Simning 2012	20	7	19	144	0.51 [0.35, 0.68]	0.95 [0.91, 0.98]
Van Heyningen 2018	56	75	32	213	0.64 [0.53, 0.74]	0.74 [0.68, 0.79]



ADDITIONAL TABLES

Table 1. Assessment of risk of bias and applicability – QUADAS-2 and QUADAS-C assessments of included studies

Study ID	GAD-7		GAD-2		Risk of bias (QUADAS-2)					Applicability concerns (QUADAS-2)			Risk of bias (QUADAS-C)			
	AAD	GAD	AAD	GAD	P	I	R	FT	ORoB	P	I	R	P	I	R	FT
Ahmadi 2019			✓		?	#	?	?	H	?	✓	✓				
Ahn 2019	✓	✓	✓	✓	?	#	✓	?	H	?	✓	✓	#	#	✓	#
Baker 2018	✓		✓		✓	#	#	✓	H	#	✓	✓				
Belus 2021		✓		✓	✓	✓	✓	✓	L	?	✓	✓	✓	✓	✓	✓
Bernstein 2018	✓	✓	✓	✓	?	✓	✓	✓	L	?	✓	✓	#	✓	✓	✓
Bisby 2022		✓		✓	#	#	#	#	H	#	✓	✓	#	#	#	#
Budikayanti 2019		✓			✓	✓	✓	✓	L	?	✓	✓				
Chibanda 2016		✓		✓	?	?	✓	✓	H	✓	✓	✓	#	#	✓	✓
Christodoulaki 2022			✓	✓	?	#	✓	?	H	?	✓	✓				
Clover 2020	✓	✓	✓		#	✓	✓	#	H	?	✓	✓	#	✓	✓	#
Conway 2016	✓	✓	✓	✓	?	✓	✓	#	H	?	✓	✓	#	✓	✓	#
Delgadillo 2012	✓	✓	✓	✓	#	#	✓	✓	H	?	✓	✓	#	#	✓	✓
Esser 2018	✓	✓	✓	✓	#	#	✓	#	H	?	✓	✓	#	#	✓	#
Fischer 2014	✓		✓		?	✓	?	✓	H	?	✓	✓	#	✓	#	✓
Garabiles 2020		✓			?	#	✓	✓	H	?	✓	✓				
Grech 2019		✓			#	?	?	✓	H	#	✓	✓				
Hitchon 2020	✓	✓	✓	✓	?	✓	✓	✓	L	?	✓	✓	#	✓	✓	✓

Table 1. Assessment of risk of bias and applicability – QUADAS-2 and QUADAS-C assessments of included studies *(Continued)*

Homans 2012	✓				?	#	✓	✓	H	?	✓	✓				
Ivanovs 2018	✓	✓	✓	✓	✓	✓	✓	✓	L	#	✓	✓	✓	✓	✓	✓
Kroenke 2007	✓	✓	✓	✓	?	✓	✓	✓	L	?	✓	✓	#	✓	✓	✓
Kujanpää 2014		✓		✓	?	#	?	?	H	#	✓	✓	#	#	#	#
Li 2014	✓				?	?	?	?	H	?	✓	✓				
Lickova 2021		✓			?	#	?	✓	H	?	✓	✓				
Makulowich 2018	✓				?	?	?	?	H	?	✓	✓				
Maric 2022	✓	✓	✓	✓	✓	#	#	✓	H	#	✓	✓	✓	#	#	✓
Marrie 2018	✓	✓	✓	✓	?	✓	✓	✓	L	?	✓	✓	#	✓	✓	✓
Michaelis 2022	✓	✓	✓	✓	✓	#	#	✓	H	#	✓	✓	✓	#	#	✓
Micoulaud-Franchi 2016		✓		✓	?	#	✓	✓	H	?	✓	✓	#	#	✓	✓
Micoulaud-Franchi 2022		✓			?	#	✓	?	H	?	✓	✓				
Mughal 2021*		✓			?	#	✓	✓	H	?	✓	✓				
Nath 2018			✓	✓	✓	#	✓	✓	H	?	✓	✓				
Osorio 2015	✓				?	#	✓	✓	H	?	✓	✓				
Pranckeviciene 2022		✓		✓	?	✓	✓	#	H	?	✓	✓	✓	✓	✓	#
Qu 2016	✓				?	?	?	?	H	✓	✓	✓				
Scott 2019	✓	✓	✓	✓	✓	✓	✓	✓	L	?	✓	✓	✓	✓	✓	✓
Seo 2014		✓			✓	#	?	✓	H	?	✓	✓				
Seo 2015		✓		✓	✓	#	?	?	H	✓	✓	✓	✓	#	#	✓
Seo 2017		✓		✓	?	#	?	?	H	?	✓	✓	#	#	#	#

Table 1. Assessment of risk of bias and applicability – QUADAS-2 and QUADAS-C assessments of included studies *(Continued)*

Shih 2022		✓			?	?	✓	?	H	?	✓	✓				
Sidik 2012		✓			✓	✓	✓	#	H	?	✓	✓				
Simning 2012	✓		✓		#	#	?	✓	H	#	✓	✓	✓	#	#	✓
Tong 2016		✓			?	?	?	?	H	?	✓	✓				
Van Heyningen 2018			✓		?	#	✓	✓	H	?	✓	✓				
Veisy 2021		✓		✓	#	#	?	?	H	?	✓	✓	#	#	#	✓
Wild 2014		✓		✓	✓	✓	✓	#	H	?	✓	✓	✓	✓	✓	#
Zeng 2013		✓			?	#	?	?	H	?	✓	✓				
Zhong 2015		✓			✓	✓	✓	#	H	?	✓	✓				
Zinchuk 2021	✓	✓			?	#	✓	✓	H	?	✓	✓				

AAD: any anxiety disorder, **GAD:** generalised anxiety disorder
✓: Yes/low risk, **#:** No/ high risk, **?:** Unclear
*extracted but not included in the primary analysis because cut-offs reported were beyond the defined core range.
P: patient selection domain
I: index test domain (since GAD-7 and GAD-2 had identical assessments, only one column is shown)
R: reference standard domain
FT: flow and timing domain
ORoB: Overall risk of bias ('low' if at least three domains were scored 'low' and none were 'high') (**L** = low, **H** = high)

Table 2. Summary characteristics of the overall study sample

Study characteristics	No. of studies	%
Target conditions, index tests (48 studies)		
Target condition GAD, index test GAD-7 (n = 15,672)	36	75
Target condition AAD, index test GAD-7 (n = 10,245)	22	46
Target condition GAD, index test GAD-2 (n = 11,428)	24	50
Target condition AAD, index test GAD-2 (n = 9973)	19	40
Studies implementing GAD-7 and GAD-2		
Studies which used GAD-7	44	92
Studies which used GAD-2	29	60
GAD-7 and GAD-2	25	52
Target condition		
Any anxiety disorder	26	54
Generalised anxiety disorder	38	79
GAD and any anxiety disorder	16	33
Continent		
Europe	16	33
Asia	15	31
North America	7	15
Australia	5	10
Africa	3	6
South America	2	4
Ethnicity		
White	12	25
Asian	9	19
Other	7	15
No information	20	42
Clinical pathway/Setting		
Clinical – across conditions	9	19

Table 2. Summary characteristics of the overall study sample *(Continued)*

Clinical – specific conditions	32	67
Non-clinical	7	15
Reference standard		
MINI	28	58
Semi-structured (SCID)	13	27
Structured (CIDI, CIS-R, DIPS)	7	15
Sampling method		
Random	35	73
Non-random	8	17
No information	5	10
Language of the index test and reference standard		
English	13	27
Other	35	73
Reference standard administered by		
Professionals	41	85
Trained interviewers	4	8
No information	3	6
Diagnostic criteria		
DSM-IV	17	35
DSM-IV+ICD-10	9	19
DSM-IV-TR	7	15
DSM-V	6	13
ICD-10	4	8
No information	5	10
Numerical variables	**Median**	**Min-Max**
Prevalence, AAD (%)	19.0	3-65
Prevalence, GAD (%)	10.5	1-45
No. analysed	217	48-2142

Table 2. Summary characteristics of the overall study sample *(Continued)*

Age (years)	41	23-67
Female proportion (%)	60.8	0.8-100

AAD: any anxiety disorder; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **DSM:** Diagnostic and Statistical Manual of Mental Disorders; **GAD:** generalised anxiety disorder; **ICD:** International Statistical Classification of Diseases and Related Health Conditions; **MINI:** Mini-International Neuropsychiatric Interview; **SCID:** Structured Clinical Interview for DSM

Table 3. Characteristics of included studies

Study ID year (country of conduct)	Study population (mean age / % female)	Reference standard (diagnostic criteria)	Target condition (% prevalence)	Sample size		Cut-offs (≥) included in primary analysis	Cut-offs (≥) included in secondary analysis	Low risk of bias?
				n included	n analysed			
Setting: Non-clinical								
Ahn 2019 (South Korea)	Mixed healthy persons and hospital patients (37 / 67)	GAD-7 & GAD-2 / MINI (DSM-IV)	GAD: 8%; AAD: 19%	1228	1157	GAD-7: 10 (for GAD) & 8 (for AAD); GAD-2: 3	GAD-7 for GAD: 7; GAD-7 for AAD: 5; GAD-2 for GAD: 3; GAD-2 for AAD:3	No
Christodoulaki 2022 (Greece)	Random sample of the general population (53 (57)	GAD-2 / MINI (DSM-IV)	GAD: 9%; AAD 30%	591	584	GAD-2: 3	GAD-2 for GAD: 6; GAD-2 for AAD: 6	No
Garabiles 2020 (Philippines)	Female migrant domestic workers (41 / 100)	GAD-7 / MINI (DSM-IV)	GAD: 28%	100	100	GAD-7: 10	GAD-7: 8	No
Maric 2022 (Serbia)	Random sample of the general population (44 / 51)	GAD-7 & GAD-2 / MINI (DSM-V)	GAD: 2%; AAD: 3%	1203	1203	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 17; GAD-7 for AAD: 18; GAD-2 for GAD: 5; GAD-2 for AAD: 4	No
Pranckeviciene 2022 (Lithuania)	Students (23 / 82)	GAD-7 & GAD-2 / CIS-R (ICD-10)	GAD: 10%	560	560	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 21; GAD-2 for GAD: 6	No
Simning 2012 (USA)	Socioeconomically disadvantaged older adults (66 / 58)	GAD-7 & GAD-2 / SCID (DSM-IV-TR)	AAD: 2%	190	190	GAD-7: 10	GAD-7 for AAD: 17; GAD-2 for AAD: 6	No
Wild 2014 (Germany)	Elderly living at home (60 / 55)	GAD-7 & GAD-2 / SCID (DSM-IV)	GAD: 6%	509	438	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 17; GAD-2 for GAD: 6	No
Setting: Clinical – across conditions								
Belus 2021 (Mozambique)	Unselected primary care patients (28 / 74)	GAD-7 & GAD-2 / MINI (DSM-IV)	GAD: 13%	534	502	GAD-7: 10; GAD-2: 3	GAD-7: 17; GAD-2 for GAD: 6	Yes

Table 3. Characteristics of included studies *(Continued)*

Chibanda 2016 (Zimbabwe)	Unselected primary patients - high HIV prevalence (38 / 79)	GAD-7 & GAD-2 / SCID (DSM-IV)	GAD: 40%	264	264	GAD-7: 10; GAD-2: 3	GAD-7: 21; GAD-2 for GAD: 6	No
Homans 2012 (Netherlands)	Long term sick-listed workers (45 / 50)	GAD-7 / MINI (DSM-IV)	AAD: 13%	170	170	GAD-7: 10	GAD-7 for AAD: 4	No
Ivanovs 2018 (Latvia)	Unselected primary care patients (56% ≥55 / 31)	GAD-7 & GAD-2 / MINI (DSM-IV)	GAD: 6%; AAD: 16%	1602	1485	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 13; GAD-7 for AAD: 13; GAD-2 for GAD: 6; GAD-2 for AAD: 6	Yes
Kroenke 2007 (USA)	Unselected primary care patients (47 / 69)	GAD-7 & GAD-2 / SCID (DSM-IV)	GAD: 8%; AAD: 19%	965	965	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 21; GAD-7 for AAD: 21; GAD-2 for GAD: 6; GAD-2 for AAD: 6	Yes
Kujanpää 2014 (Finland)	High utilisers in four community clinics (63 / 69)	GAD-7 & GAD-2 / MINI (DSM-IV)	GAD: 4%	150	150	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 6; GAD-2 for GAD: 3	No
Makulowich 2018 (USA)	Under-insured or uninsured persons (50 / 70)	GAD-7 / MINI (DSM-IV)	AAD: 32%	50	50	GAD-7: 9	GAD-7 for AAD: 10	No
Sidik 2012 (Malaysia)	Female primary care patients (31 / 100)	GAD-7 / CIDI (DSM-IV and ICD-10)	GAD: 12%	895	146	GAD-7: 8	GAD-7 for GAD: 1	No
Zeng 2013 (China)	Outpatients from an internal medicine department of traditional Chinese medicine (37 / 69)	GAD-7 / MINI (no information)	GAD: 5%	2011	1993	GAD-7: 10	GAD-7 for GAD: 6	No
Setting: Clinical – specific conditions								
Ahmadi 2019 (Iran)	People with coronary heart disease (60 / 46)	GAD-2 / SCID (DSM-IV)	AAD: 28%	279	279	GAD-2: 3	GAD-2 for AAD: 2	No
Baker 2018 (USA)	People with stable COPD (65 (median age) / 46)	GAD-7 / MINI (DSM-IV)	AAD: 11%	223	220	GAD-7: 10	GAD-7: 19	No
Bernstein 2018 (Canada)	People with inflammatory bowel disease (48 / 63)	GAD-7 & GAD-2 / SCID (DSM-IV-TR)	GAD: 6%; AAD: 18%	247	242	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 21; GAD-7 for AAD: 21;	Yes

Table 3. Characteristics of included studies *(Continued)*

							GAD-2 for GAD: 6; GAD-2 for AAD: 6	
Bisby 2022 (Australia)	People with chronic pain (49 / 85)	GAD-7 & GAD-2 / MINI (DSM-IV)	GAD: 35%	63	62	GAD-7: 10; GAD-2: 3	GAD-7: 9; GAD-2 for GAD: 3	No
Budikayanti 2019 (Indonesia)	People with epilepsy (36 / 46)	GAD-7 / MINI (ICD-10)	GAD: 16%	146	146	GAD-7: 10	GAD-7: 18	Yes
Clover 2020 (Australia)	Adult outpatients with cancer (59 / 66)	GAD-7 & GAD-2 / SCID (DSM-IV-TR)	GAD: 9%; AAD: 28%	322	153	GAD-7: 10	GAD-7 for GAD: 8; GAD-7 for AAD: 16; GAD-2 for AAD: 6	No
Conway 2016 (Australia)	Heart transplant recipients - long-term survivors (63 / 24)	GAD-7 & GAD-2 / MINI (DSM-IV and ICD-10))	GAD: 8%; AAD: 16%	80	48	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 13; GAD-7 for AAD: 13; GAD-2 for GAD: 6; GAD-2 for AAD: 6	No
Delgadillo 2012 (UK)	Outpatients applying to a community drugs treatment service (35 / 23)	GAD-7 & GAD-2 / CIS-R (ICD-10)	GAD: 30%; AAD: 65%	103	103	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 4; GAD-7 for AAD: 4; GAD-2 for GAD: 4; GAD-2 for AAD: 4	No
Esser 2018 (Germany)	People with cancer (58 / 52)	GAD-7 & GAD-2 / CIDI (DSM-IV and ICD 10)	GAD: 2%; AAD: 12%	4020	2142 (GAD-7), 1931 (GAD-2)	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 21; GAD-7 for AAD: 21; GAD-2 for GAD: 6; GAD-2 for AAD: 6	No
Fischer 2014 (Germany)	People with heart failure (66 / 21)	GAD-7 & GAD-2 / SCID (DSM-IV)	AAD: 14%	194	194	GAD-7: 10; GAD-2: 3	GAD-7 for AAD: 21; GAD-2 for AAD:6	No
Grech 2019 (Australia)	People recently hospitalised for coronary heart disease (63 / 31)	GAD-7 / MINI (no information)	GAD: 8%	85	85	GAD-7: 10	GAD-7: 18	No
Hitchon 2020 (Canada)	People with rheumatoid arthritis (60 / 85)	GAD-7 & GAD-2 / SCID (DSM-IV)	GAD: 7%; AAD: 19%	154	150	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 19; GAD-7 for AAD: 19; GAD-2 for GAD: 6; GAD-2 for AAD: 6	Yes
Li 2014 (China)	Cardiovascular outpatients (- / -)	GAD-7 / CIDI (no information)	AAD: 37%	No information	201	GAD-7: 10	GAD-7 for AAD: 1	No

Table 3. Characteristics of included studies *(Continued)*

Lickova 2021 (Czech Republic)	People with cancer (67 / 40)	GAD-7 / MINI (DSM-IV)	GAD: 24%	238	238	GAD-7: 9	GAD-7 for GAD: 1	No
Marrie 2018 (Canada)	People with multiple sclerosis (51 / 81)	GAD-7 & GAD-2 / SCID (DSM-IV-TR)	GAD: 4%; AAD: 17%	255	253	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 21; GAD-7 for AAD: 21; GAD-2 for GAD: 6; GAD-2 for AAD: 6	Yes
Michaelis 2022 (Germany)	People with epilepsy (39 / 60)	GAD-7 & GAD-2 / Mini-DIPS (ICD-10)	GAD: 11; AAD: 28%	80	80	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 18; GAD-7 for AAD: 18; GAD-2 for GAD: 6; GAD-2 for AAD: 6	No
Micoulaud-Franchi 2016 (France)	People with epilepsy (39 / 63)	GAD-7 & GAD-2 / MINI (DSM-IV-TR)	GAD: 34%	145	145	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 21; GAD-2 for GAD: 6	No
Micoulaud-Franchi 2022 (France)	People with epilepsy (39 / 57)	GAD-7 / MINI (DSM-IV-TR)	GAD: 26%	149	149	GAD-7: 8	GAD-7 for GAD: 1	No
Mughal 2021 (Vietnam)	People enroled in an urban methadone maintenance therapy clinic (41 / 80)	GAD-7 / MINI (DSM-V)	GAD: 4%	420	400	Cut-offs outside core range	GAD-7 for GAD: 3	No
Nath 2018 (UK)	Pregnant women (33 / 100)	GAD-2 / SCID (DSM-V)	GAD: 5%; AAD: 15%	545	528	GAD-2: 3	GAD-2 for GAD: 2; GAD-2 for AAD: 2	No
Osorio 2015 (Brazil)	Outpatients with cancer (- / 62)	GAD-7 / SCID (DSM-IV)	AAD: 45%	434	400	GAD-7: 9	GAD-7 for AAD: 5	No
Qu 2016 (China)	People with leukaemia (- / 41)	GAD-7 / MINI (DSM-IV and ICD-10))	AAD: 13%	585	548	GAD-7: 10	GAD-7 for AAD: 1	No
Scott 2019 (Australia)	People with epilepsy (40 / 58)	GAD-7 & GAD-2 / MINI (DSM-V)	GAD: 14%; AAD: 31%	106	106	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 19; GAD-7 for AAD: 19; GAD-2 for GAD: 6; GAD-2 for AAD: 6	Yes
Seo 2014 (South Korea)	People with epilepsy (40 / 40)	GAD-7 / MINI (DSM-IV and ICD-10)	GAD: 21%	243	243	GAD-7: 9	GAD-7 for GAD: 5	No

Table 3. Characteristics of included studies *(Continued)*

Seo 2015 (South Korea)	People with migraine (41 /89)	GAD-7 & GAD-2 / MINI (no information)	GAD: 22%	146	146	GAD-7: 8; GAD-2:	GAD-7 for GAD: 5; GAD-2 for GAD: 3	No
Seo 2017 (South Korea)	People with tension-type headache (- / -)	GAD-7 & GAD-2 / MINI (no information)	GAD: 21%	No information	160	GAD-7: 8; GAD-2: 2	GAD-7 for GAD: 1; GAD-2 for GAD: 1	No
Shih 2022 (Taiwan)	People with epilepsy (33 / 56)	GAD-7 / MINI (DSM-IV-TR)	GAD: 16%	107	107	GAD-7: 10	GAD-7 for GAD: 19	No
Tong 2016 (China)	People with epilepsy (30 / 49)	GAD-7 / MINI (DSM-IV and ICD-10)	GAD: 23%	No information	213	GAD-7: 9	GAD-7 for GAD: 4	No
Van Heyningen 2018 (South Africa)	Pregnant women (27 / 100)	GAD-2 / MINI (DSM-IV and ICD-10)	AAD: 23%	376	376	GAD-2: 2	GAD-2 for AAD: 1	No
Veisy 2021 (Iran)	People with migraine (73 / 70)	GAD-7 & GAD-2 / SCID (DSM-V)	GAD: 45%	188	186	GAD-7: 10; GAD-2: 3	GAD-7 for GAD: 1; GAD-2 for GAD: 3	No
Zhong 2015 (Peru)	Pregnant women (28 / 100)	GAD-7 / CIDI (DSM-IV and ICD-10)	GAD: 1%	2978	946	GAD-7: 10	GAD-7 for GAD: 21	No
Zinchuk 2021 (Russia)	People with epilepsy (41 / 65)	GAD-7/ MINI (DSM-IV and ICD-10)	GAD: 27%; AAD: 42%	233	233	GAD-7: 10	GAD-7 for GAD: 21; GAD-7 for AAD: 21	No

AAD: any anxiety disorder; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **DSM:** Diagnostic and Statistical Manual of Mental Disorders; **GAD:** generalised anxiety disorder; **GAD-7/2:** 7-item/2-item Generalized Anxiety Disorder scales; **ICD:** International Statistical Classification of Diseases and Related Health Conditions; **MINI:** Mini-International Neuropsychiatric Interview; **SCID:** Structured Clinical Interview for DSM

Table 4. GAD-7 for generalised anxiety disorder – heterogeneity and comparisons between covariates

Study question: GAD-7–GAD (Core range cut-offs 8-10, n = 35)	**Heterogeneity measures:**
Summary sensitivity (95% CI): 0.64 (0.56 to 0.72)	Variance of logit sensitivity (Tau^2): 0.902
Summary specificity (95% CI): 0.91 (0.87 to 0.93)	Variance of logit specificity (Tau^2): 0.942
	Area of the 95% prediction region: 0.295

Covariate	Covariate categories	Sensitivity (95% CI)	Specificity (95% CI)
Setting	Clinical – across conditions (n = 7)	0.57 (0.39 to 0.73)	0.93 (0.86 to 0.97)
	Clinical – specific conditions (n = 23)	0.70 (0.60 to 0.78)	0.89 (0.85 to 0.93)
	Non-clinical (n = 5)	0.50 (0.30 to 0.70)	0.93 (0.84 to 0.97)
		P = 0.16	P = 0.47
		Tau^2 = 0.785	Tau^2 = 0.897
Reference standard	MINI (n = 21)	0.64 (0.53 to 0.74)	0.92 (0.89 to 0.95)
	Semi-structured (SCID) (n = 8)	0.70 (0.53 to 0.83)	0.88 (0.80 to 0.94)
	Structured (CIDI, CIS-R, DIPS) (n = 6)	0.57 (0.37 to 0.75)	0.85 (0.73 to 0.93)
		P = 0.62	P = 0.20
		Tau^2 = 0.869	Tau^2 = 0.820
Risk of bias	Unclear/High (n = 27)	0.66 (0.57 to 0.74)	0.91 (0.87 to 0.94)
	Low (n = 8)	0.57 (0.40 to 0.75)	0.91 (0.83 to 0.95)
		P = 0.37	P = 0.98
		Tau^2 = 0.867	Tau^2 = 0.944
GAD prevalence*		**P = 0.003**	P = 0.27

*Prevalence was assessed as a continuous variable.
AAD: any anxiety disorder; **CI:** confidence interval; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **GAD:** generalised anxiety disorder; **GAD-7/2:** 7-item/2-item Generalized Anxiety Disorder scale; **MINI:** Mini-International Neuropsychiatric Interview; **SCID:** Structured Clinical Interview for DSM

Table 5. GAD-7 for generalised anxiety disorder – summary estimates for each cut-off based on bivariate and multiple threshold models

			Bivariate model			Multiple thresholds model		
Cut-off (≥)	**n studies**	**n participants**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**
1	19	7806	0.99 (0.96 to 1.00)	0.21 (0.13 to 0.3)	0.20	0.96 (0.94 to 0.98)	0.32 (0.26 to 0.4)	0.29
2	20	9009	0.98 (0.94 to 0.99)	0.36 (0.26 to 0.47)	0.34	0.95 (0.92 to 0.97)	0.40 (0.33 to 0.48)	0.35
3	22	10,629	0.96 (0.92 to 0.98)	0.49 (0.39 to 0.59)	0.45	0.93 (0.9 to 0.96)	0.48 (0.41 to 0.56)	0.42
4	23	10,475	0.95 (0.89 to 0.98)	0.56 (0.48 to 0.64)	0.51	0.91 (0.87 to 0.94)	0.57 (0.49 to 0.64)	0.48
5	28	14,080	0.93 (0.87 to 0.96)	0.65 (0.58 to 0.72)	0.58	0.88 (0.83 to 0.92)	0.65 (0.57 to 0.71)	0.53
6	28	14,140	0.89 (0.82 to 0.94)	0.73 (0.66 to 0.78)	0.62	0.85 (0.78 to 0.89)	0.72 (0.65 to 0.78)	0.56
7	28	14,195	**0.84 (0.77 to 0.90)**	**0.78 (0.72 to 0.83)**	**0.62**	**0.80 (0.73 to 0.86)**	**0.78 (0.72 to 0.83)**	**0.58**
8	32	14,698	0.75 (0.67 to 0.82)	0.85 (0.8 to 0.89)	0.60	0.75 (0.67 to 0.82)	0.83 (0.78 to 0.87)	0.58
9	29	14,335	0.70 (0.61 to 0.77)	0.88 (0.84 to 0.91)	0.57	0.69 (0.60 to 0.76)	0.87 (0.83 to 0.91)	0.56
10	28	13,980	0.62 (0.52 to 0.71)	0.90 (0.86 to 0.93)	0.52	0.62 (0.52 to 0.70)	0.91 (0.87 to 0.93)	0.52
11	24	11,548	0.52 (0.41 to 0.62)	0.92 (0.89 to 0.94)	0.44	0.54 (0.44 to 0.64)	0.93 (0.91 to 0.95)	0.47
12	19	9588	0.48 (0.37 to 0.58)	0.93 (0.90 to 0.96)	0.41	0.46 (0.37 to 0.56)	0.95 (0.93 to 0.96)	0.41
13	21	10,243	0.35 (0.25 to 0.46)	0.95 (0.92 to 0.97)	0.29	0.39 (0.30 to 0.49)	0.96 (0.95 to 0.97)	0.35
14	20	9727	0.33 (0.25 to 0.42)	0.96 (0.94 to 0.97)	0.29	0.32 (0.24 to 0.41)	0.97 (0.96 to 0.98)	0.29
15	19	10,075	0.25 (0.17 to 0.35)	0.97 (0.96 to 0.98)	0.22	0.25 (0.18 to 0.34)	0.98 (0.97 to 0.99)	0.24
16	17	8481	0.21 (0.14 to 0.31)	0.98 (0.96 to 0.99)	0.19	0.2 (0.14 to 0.28)	0.99 (0.98 to 0.99)	0.19
17	15	7807	0.23 (0.17 to 0.31)	0.99 (0.97 to 0.99)	0.22	0.16 (0.11 to 0.22)	0.99 (0.98 to 0.99)	0.15
18	15	7894	0.13 (0.07 to 0.21)	0.99 (0.98 to 0.99)	0.11	0.12 (0.08 to 0.17)	0.99 (0.99 to 1.00)	0.11
19	15	6737	0.09 (0.05 to 0.16)	0.99 (0.98 to 0.99)	0.08	0.09 (0.06 to 0.14)	0.99 (0.99 to 1.00)	0.09

Table 5. GAD-7 for generalised anxiety disorder – summary estimates for each cut-off based on bivariate and multiple threshold models *(Continued)*

20	11	5997	0.08 (0.04 to 0.13)	0.99 (0.99 to 1.00)	0.07	0.07 (0.04 to 0.1)	1.00 (0.99 to 1.00)	0.06
21	14	6630	0.01 (0.00 to 0.06)	1.00 (0.99 to 1.00)	0.01	0.05 (0.03 to 0.08)	1.00 (1.00 to 1.00)	0.05

CI: confidence interval. Bold numbers indicate the cut-off with the highest Youden index.

Table 6. GAD-7 for any anxiety disorder – heterogeneity and comparisons between covariates

Study question: GAD-7–AAD (Core range cut-offs 8-10, n = 22)	**Heterogeneity measures:**
Summary sensitivity (95% CI): 0.48 (0.40 to 0.57)	Variance of logit sensitivity (Tau^2): 0.54
Summary specificity (95% CI): 0.91 (0.89 to 0.93)	Variance of logit specificity (Tau^2): 0.34
	Area of the 95% prediction region: 0.156

Covariate	Covariate categories	Sensitivity (95% CI)	Specificity (95% CI)
Setting	Clinical – across conditions (n = 4)	0.48 (0.29 to 0.67)	0.89 (0.83 to 0.93)
	Clinical – specific conditions (n = 15)	0.50 (0.40 to 0.60)	0.90 (0.87 to 0.93)
	Non-clinical (n = 3)	0.43 (0.24 to 0.64)	0.96 (0.93 to 0.98)
		P = 0.86	**P = 0.019**
		Tau^2 = 0.53	Tau^2 = 0.23
Reference standard	MINI (n = 10)	0.49 (0.37 to 0.6)	0.92 (0.88 to 0.94)
	Semi-structured (SCID) (n = 8)	0.42 (0.3 to 0.55)	0.92 (0.87 to 0.95)
	Structured (CIDI, CIS-R, DIPS) (n = 4)	0.61 (0.43 to 0.77)	0.91 (0.83 to 0.95)
		P = 0.25	P = 0.94
		Tau^2 = 0.48	Tau^2 = 0.33
AAD prevalence*		P = 0.11	P = 0.43
Risk of bias	Unclear/High (n = 16)	0.49 (0.39 to 0.59)	0.92 (0.89 to 0.94)
	Low (n = 6)	0.46 (0.31 to 0.62)	0.90 (0.84 to 0.94)
		P = 0.75	P = 0.38
		Tau^2 = 0.54	Tau^2 = 0.33

*Prevalence was assessed as a continuous variable.
AAD: any anxiety disorder; **CI:** confidence interval; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **GAD:** generalised anxiety disorder; **GAD-7/2:** 7-item/2-item Generalized Anxiety Disorder scale; **MINI:** Mini-International Neuropsychiatric Interview; **SCID:** Structured Clinical Interview for DSM

Table 7. GAD-7 for any anxiety disorder – summary estimates for each cut-off based on bivariate and multiple threshold models

			Bivariate model			Multiple thresholds model		
Cut-off (≥)	**n studies**	**n participants**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**
1	14	4998	0.96 (0.92 to 0.97)	0.24 (0.19 to 0.30)	0.20	0.92 (0.88 to 0.94)	0.34 (0.29 to 0.40)	0.26
2	15	6201	0.92 (0.89 to 0.95)	0.41 (0.34 to 0.48)	0.33	0.89 (0.85 to 0.92)	0.42 (0.37 to 0.48)	0.32
3	16	7668	0.87 (0.82 to 0.90)	0.54 (0.47 to 0.61)	0.41	0.86 (0.82 to 0.90)	0.51 (0.45 to 0.57)	0.37
4	17	8825	0.82 (0.77 to 0.86)	0.65 (0.59 to 0.71)	0.47	0.83 (0.77 to 0.87)	0.59 (0.53 to 0.65)	0.42
5	17	9175	**0.77 (0.71 to 0.82)**	**0.72 (0.67 to 0.76)**	**0.49**	0.79 (0.72 to 0.84)	0.67 (0.62 to 0.72)	0.46
6	18	9225	0.67 (0.60 to 0.73)	0.79 (0.75 to 0.82)	0.46	0.74 (0.66 to 0.80)	0.74 (0.69 to 0.78)	0.48
7	18	9279	0.62 (0.55 to 0.69)	0.83 (0.79 to 0.86)	0.45	**0.68 (0.60 to 0.75)**	**0.80 (0.76 to 0.83)**	**0.48**
8	19	9448	0.57 (0.48 to 0.65)	0.87 (0.84 to 0.89)	0.43	0.62 (0.53 to 0.70)	0.85 (0.81 to 0.88)	0.46
9	19	8341	0.5 (0.41 to 0.59)	0.90 (0.87 to 0.92)	0.40	0.55 (0.46 to 0.63)	0.89 (0.86 to 0.91)	0.43
10	18	8510	0.52 (0.42 to 0.61)	0.92 (0.89 to 0.94)	0.43	0.48 (0.39 to 0.57)	0.92 (0.89 to 0.93)	0.40
11	16	7788	0.38 (0.30 to 0.46)	0.95 (0.92 to 0.96)	0.32	0.41 (0.33 to 0.50)	0.94 (0.92 to 0.95)	0.35
12	14	7569	0.32 (0.26 to 0.39)	0.96 (0.94 to 0.97)	0.28	0.35 (0.27 to 0.43)	0.96 (0.94 to 0.97)	0.30
13	15	7619	0.26 (0.21 to 0.33)	0.97 (0.96 to 0.98)	0.23	0.29 (0.22 to 0.36)	0.97 (0.96 to 0.98)	0.25
14	15	7538	0.22 (0.17 to 0.28)	0.98 (0.97 to 0.98)	0.19	0.23 (0.17 to 0.30)	0.98 (0.97 to 0.98)	0.21
15	13	7463	0.17 (0.12 to 0.23)	0.98 (0.97 to 0.99)	0.15	0.19 (0.14 to 0.25)	0.98 (0.98 to 0.99)	0.17
16	11	5892	0.15 (0.01 to 0.22)	0.99 (0.98 to 0.99)	0.14	0.15 (0.11 to 0.20)	0.99 (0.98 to 0.99)	0.14
17	10	5676	0.13 (0.09 to 0.19)*	0.99 (0.99 to 0.99)*	0.12	0.12 (0.08 to 0.16)	0.99 (0.99 to 0.99)	0.11
18	11	5873	0.08 (0.05 to 0.13)*	0.99 (0.99 to 1.00)*	0.07	0.09 (0.06 to 0.13)	0.99 (0.99 to 1.00)	0.08
19	11	4628	0.06 (0.03 to 0.10)*	0.99 (0.99 to 1.00)*	0.05	0.07 (0.05 to 0.10)	1.00 (0.99 to 1.00)	0.07

Table 7. GAD-7 for any anxiety disorder – summary estimates for each cut-off based on bivariate and multiple threshold models *(Continued)*

20	9	5463	0.03 (0.02 to 0.06)*	1.00 (1.00 to 1.00)*	0.03	0.05 (0.04 to 0.08)	1.00 (1.00 to 1.00)	0.05
21	12	4588	0.01 (0.00 to 0.02)*	1 (1.00 to 1.00)*	0	0.04 (0.03 to 0.06)	1.00 (1.00 to 1.00)	0.04

CI: confidence interval. Bold numbers indicate the cut-off with the highest Youden index
*using the bivariate model for this cut-off resulted in a boundary fit

Table 8. GAD-2 for generalised anxiety disorder – heterogeneity and comparisons between covariates

Study question: GAD-2–GAD (Core range cut-offs 2-3, n = 24) **Summary sensitivity (95% CI):** 0.68 (0.59 to 0.75) **Summary specificity (95% CI):** 0.86 (0.82 to 0.89)		**Heterogeneity measures:** Variance of logit sensitivity (Tau^2): 0.59 Variance of logit specificity (Tau^2): 0.43 Area of the 95% prediction region: 0.187	
Covariate	**Covariate categories**	**Sensitivity (95% CI)**	**Specificity (95% CI)**
Setting	Clinical – across conditions (n = 5)	0.63 (0.45 to 0.78)	0.86 (0.78 to 0.92)
	Clinical – specific conditions (n = 14)	0.70 (0.59 to 0.79)	0.84 (0.79 to 0.89)
	Non-clinical (n = 5)	0.66 (0.49 to 0.80)	0.88 (0.81 to 0.93)
		P = 0.79	P = 0.65
		Tau^2 = 0.57	Tau^2 = 0.41
Reference standard	MINI (n = 12)	0.6 (0.48 to 0.70)	0.88 (0.84 to 0.91)
	Semi-structured (SCID) (n = 8)	0.75 (0.63 to 0.84)	0.86 (0.80 to 0.90)
	Structured (CIDI, CIS-R, DIPS) (n = 4)	0.73 (0.54 to 0.86)	0.74 (0.62 to 0.84)
		P = 0.14	**P = 0.037**
		Tau^2 = 0.47	Tau^2 = 0.32
GAD prevalence*		P = 0.29	P = 0.14
Risk of bias	Unclear/High (n = 17)	0.70 (0.61 to 0.78)	0.86 (0.81 to 0.90)
	Low (n = 7)	0.60 (0.44 to 0.74)	0.85 (0.78 to 0.91)
		P = 0.22	P = 0.86
		Tau^2 = 0.53	Tau^2 = 0.43

*Prevalence was assessed as a continuous variable.
AAD: any anxiety disorder; **CI:** confidence interval; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **GAD:** generalised anxiety disorder; **GAD-7/2:** 7-item/2-item Generalized Anxiety Disorder scale; **MINI:** Mini-International Neuropsychiatric Interview; **SCID:** Structured Clinical Interview for DSM

Table 9. GAD-2 for generalised anxiety disorder – summary estimates for each cut-off based on bivariate and multiple threshold models

			Bivariate model			Multiple thresholds model		
Cut-off (≥)	**n studies**	**n participants**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**
1	18	8594	0.98 (0.95 to 0.99)	0.39 (0.32 to 0.46)	0.37	0.91 (0.86 to 0.94)	0.43 (0.36 to 0.50)	0.34
2	23	10,901	**0.89 (0.83 to 0.93)**	**0.67 (0.60 to 0.72)**	**0.56**	0.82 (0.74 to 0.88)	0.66 (0.59 to 0.72)	0.48
3	23	11,269	0.67 (0.59 to 0.75)	0.86 (0.82 to 0.89)	0.53	**0.67 (0.57 to 0.77)**	**0.83 (0.78 to 0.87)**	**0.50**
4	20	10,409	0.48 (0.39 to 0.57)	0.92 (0.88 to 0.94)	0.40	0.48 (0.37 to 0.6)	0.93 (0.90 to 0.95)	0.41
5	17	9040	0.28 (0.21 to 0.36)*	0.96 (0.94 to 0.98)*	0.24	0.30 (0.20 to 0.42)	0.97 (0.96 to 0.98)	0.27
6	16	8937	0.11 (0.11 to 0.11)*	0.99 (0.99 to 0.99)*	0.10	0.16 (0.09 to 0.26)	0.99 (0.98 to 0.99)	0.15

CI: confidence interval. Bold numbers indicate the cut-off with the highest Youden index.
*using the bivariate model for this cut-off resulted in a boundary fit

Table 10. GAD-2 for any anxiety disorder – heterogeneity and comparisons between covariates

Study question: GAD-2–AAD (Core range cut-offs 2-3, n = 19) **Summary sensitivity (95% CI):** 0.53 (0.44 to 0.62) **Summary specificity (95% CI):** 0.89 (0.86 to 0.91)	**Heterogeneity measures:** Variance of logit sensitivity (Tau^2): 0.54 Variance of logit specificity (Tau^2): 0.21 Area of the 95% prediction region: 0.158

Covariate	Covariate categories	Sensitivity (95% CI)	Specificity (95% CI)
Setting	Clinical – across conditions (n = 2)	0.53 (0.28 to 0.76)	0.87 (0.81 to 0.92)
	Clinical – specific conditions (n = 13)	0.54 (0.43 to 0.65)	0.87 (0.84 to 0.89)
	Non-clinical (n = 4)	0.50 (0.32 to 0.68)	0.93 (0.90 to 0.95)
		P = 0.93	**P = 0.019**
		Tau^2 = 0.54	Tau^2 = 0.11
Reference standard	MINI (n = 7)	0.50 (0.36 to 0.64)	0.89 (0.84 to 0.92)
	Semi-structured (SCID) (n = 9)	0.55 (0.42 to 0.67)	0.89 (0.85 to 0.92)
	Structured (CIDI, CIS-R, DIPS) (n = 3)	0.57 (0.35 to 0.76)	0.88 (0.79 to 0.93)
		P = 0.84	P = 0.96
		Tau^2 = 0.53	Tau^2 = 0.21
AAD prevalence*		P = 0.15	P = 0.45
Risk of bias	Unclear/High (n = 13)	0.53 (0.43 to 0.64)	0.90 (0.87 to 0.92)
	Low (n = 6)	0.53 (0.38 to 0.68)	0.86 (0.81 to 0.90)
		P = 0.97	P = 0.13
		Tau^2 = 0.55	Tau^2 = 0.19

*Prevalence was assessed as a continuous variable.
AAD: any anxiety disorder; **CI:** confidence interval; **CIDI:** Composite International Diagnostic Interview; **CIS-R:** Clinical Interview Schedule-Revised; **DIPS:** [German] 'Diagnostisches Interview bei psychischen Störungen'; **GAD:** generalised anxiety disorder; **GAD-7/2:** 7-item/2-item Generalized Anxiety Disorder scale; **MINI:** Mini-International Neuropsychiatric Interview; **SCID:** Structured Clinical Interview for DSM

Table 11. GAD-2 for any anxiety disorder – summary estimates for each cut-off based on bivariate and multiple threshold models

			Bivariate model			Multiple thresholds model		
Cut-off (≥)	**n studies**	**n participants**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**Youden index**
1	16	9214	0.88 (0.83 to 0.91)	0.50 (0.43 to 0.56)	0.37	0.84 (0.79 to 0.88)	0.48 (0.42 to 0.55)	0.32
2	18	9456	**0.77 (0.69 to 0.83)**	**0.73 (0.67 to 0.78)**	**0.49**	**0.71 (0.64 to 0.77)**	**0.72 (0.67 to 0.76)**	**0.43**
3	18	9596	0.53 (0.43 to 0.62)	0.89 (0.87 to 0.91)	0.42	0.53 (0.45 to 0.61)	0.87 (0.84 to 0.90)	0.40
4	15	7644	0.34 (0.28 to 0.41)	0.95 (0.93 to 0.96)	0.29	0.34 (0.27 to 0.43)	0.95 (0.93 to 0.96)	0.29
5	14	6441	0.20 (0.15 to 0.24)*	0.98 (0.97 to 0.99)*	0.18	0.20 (0.14 to 0.27)	0.98 (0.97 to 0.99)	0.18
6	13	6338	0.06 (0.03 to 0.09)*	0.99 (0.98 to 0.99)*	0.05	0.10 (0.07 to 0.15)	0.99 (0.99 to 0.99)	0.09

CI: confidence interval. Bold numbers indicate the cut-off with the highest Youden index.
*using the bivariate model for this cut-off resulted in a boundary fit

APPENDICES

Appendix 1. Search strategies

Ovid Embase <1974 to 2022 September 30>

[Date limited: 1990 onwards]

1 anxiety disorder/

2 anxiety neurosis/

3 generalized anxiety disorder/

4 "mixed anxiety and depression"/

5 Agoraphobia/

6 Panic/

7 social* phobi*.tw,kf,hw.

8 (anxiety adj3 (disorder* or neuros* or general* or depress* or social* or unspecif*)).tw,kf.

9 GAD.tw,kf.

10 ADNOS*.tw,kf.

11 (agoraphobi* or panic*).tw,kf.

12 or/1-11

13 anxiety.ti,kf.

14 anxi*.ab. /freq=3

15 Anxiety/ and diagnosis.fs.

16 *Anxiety/

17 or/13-16

18 12 or 17

19 (GAD-7* or GAD7* or (GAD adj3 seven) or (GAD adj3 "7")).tw,kf.

20 (generali#ed anxiety disorder adj3 (seven or "7")).tw,kf.

21 (GAD-2* or GAD2* or (GAD adj3 two) or (GAD adj3 "2")).tw,kf.

22 (generali#ed anxiety disorder adj3 (two or "2")).tw,kf.

23 ((GAD or generali#ed anxiety disorder) adj3 (self-report* or scale? or scor* or checklist* or check-list* or criteria or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire? or screen* or tool*)).tw,kf.

24 or/19-23

25 18 and 24

26 Gold Standard/

27 (reference standard? or gold standard?).tw,kf.

28 psychiatric diagnosis/

29 Interview/

30 Structured Interview/

31 Structured Clinical Interview for DSM Disorders/

32 (structured clinic* interview* or SCI or SCID).tw,kf.

33 ((standard* adj2 clinic* adj2 interview* adj2 psychiatr*) or SCIP).tw,kf.

34 (schedul* adj2 clinic* assess* adj2 neuropsych*).tw,kf.

35 Mini International Neuropsychiatric Interview/

36 (international neuropsych* interview* or MINI).tw,kf.

37 Mini Mental State Examination/

38 (mental state examination* or MMSE).tw,kf.

39 (composite international diagnos* interview* or CIDI).tw,kf.

40 ((anxi* adj2 interview* adj2 schedule*) or ADIS).tw,kf.

41 Diagnostic Interview Schedule/

42 (diagnos* interview* schedule* or DIS).tw,kf.

43 General Mental Disease Assessment/

44 (geriatric* mental state* or GMSA).tw,kf.

45 ((cambridge examination adj3 mental disorder* adj3 elder*) or CAMDEX).tw,kf.

46 ((clinical* or clinician* or diagnos* or neuropsych* or neuro-psych* or psychiatri* or schedul* or structured or semi-structured or symptom scale*) adj3 interview*).tw,kf.

47 ((clinical* or clinician* or diagnos* or neuropsych* or neuro-psych* or psychiatri* or schedul* or structured or semi-structured or symptom scale*) and interview*).kw.

48 clinical diagnosis.mp.

49 ((clinical* or clinician*) adj3 (administered or checklist* or check list* or index or indexes or indices or inventory or inventories or instrument? or questionnaire* or scale? or tool*)).tw,kf.

50 ((ICD10 or ICD-10 or ICD11 or ICD-11 or (international classification adj2 disease?) or DSM* or (diagnostic adj2 statistical manual adj2 mental disorder*)) adj diagnos*).tw,kf.

51 or/26-50

52 DIAGNOSTIC TEST ACCURACY STUDY/

53 DIAGNOSTIC ACCURACY/

54 VALIDATION STUDY/

55 "SENSITIVITY and SPECIFICITY"/

56 specificity.tw,kf.

57 RECEIVER OPERATING CHARACTERISTIC/

58 RELIABILITY/

59 INTERNAL VALIDITY/

60 INTERNAL CONSISTENCY/

61 (validat* or validity).tw,kf.

62 likelihood ratio*.tw,kf.

63 ((re-test or retest or test-retest) adj reliability).tw,kf.

64 receiver operating characteristic*.tw,kf.

65 ROC.tw,kf.

66 (DTA or (diagnos* adj2 accura*)).tw,kf.

67 (performance adj5 (self-report* or scale? or scor* or checklist* or check-list* or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire? or screen* or tool*)).tw,kf.

68 ((degree? or rate* or rating) adj3 agreement?).tw,kf.

69 or/52-68

70 (sensitivity or specificity or validity or accuracy or gold standard* or reference standard* or ROC).tw,kf,hw.

71 51 or 69

72 limit 71 to yr="1990 -Current"

73 25 and 72

74 limit 73 to conference abstract status

75 73 not 74

76 70 and 74

77 75 or 76

****

Ovid MEDLINE(R) ALL <1946 to September 30, 2022>

[Date limited: 1990 onwards]

1 anxiety disorders/ or agoraphobia/ or panic disorder/ or phobia, social/

2 generali#ed anxiety disorder.mp.

3 (anxiety adj3 (disorder* or neuros* or general* or depress* or social* or unspecif*)).tw,kf.

4 social* phobi*.tw,kf.

5 GAD.tw,kf.

6 ADNOS*.tw,kf.

7 (agoraphobi* or panic*).tw,kf.

8 or/1-7

9 anxiety.ti,kw.

10 anxi*.ab. /freq=3

11 Anxiety/di [Diagnosis]

12 *Anxiety/

13 or/9-12

14 (8 or 13)

15 (GAD-7* or GAD7* or (GAD adj3 seven) or (GAD adj3 "7")).tw,kf.

16 (generali#ed anxiety disorder adj3 (seven or "7")).tw,kf.

17 (GAD-2* or GAD2* or (GAD adj3 two) or (GAD adj3 "2")).tw,kf.

18 (generali#ed anxiety disorder adj3 (two or "2")).tw,kw.

19 (GAD adj3 (self-report* or scale? or scor* or checklist* or check-list* or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire? or screen* or tool*)).tw,kf.

20 or/15-19

21 (14 and 20)

22 (reference standard? or gold standard?).tw,kf.

23 psychiatric diagnosis.mp.

24 Interview, Psychological/ or Interview/

25 ((structured adj2 interview*) or SCI or SCID).tw,kf.

26 ((standard* adj2 clinic* adj2 interview* adj2 psychiatr*) or SCIP).tw,kf.

27 (schedul* adj2 clinic* assess* adj2 neuropsych*).tw,kf.

28 (international neuropsych* interview* or MINI).mp.

29 (mental state examination* or MMSE).mp.

30 mental status schedule/ or psychiatric status rating scales/ 85212

31 composite international diagnos* interview*.mp. or CIDI.tw,kf. 3495

32 ((anxi* adj2 interview* adj2 schedule*) or ADIS).tw,kf. 609

33 (diagnos* interview* schedule* or DIS).tw,kf.

34 (geriatric* mental state* or GMSA).mp.

35 ((cambridge examination adj3 mental disorder* adj3 elder*) or CAMDEX).tw,kf.

36 ((clinical* or clinician* or diagnos* or neuropsych* or neuro-psych* or psychiatri* or schedul* or structured or semi-structured or symptom scale*) adj3 interview*).tw,kf.

37 ((clinical* or clinician* or diagnos* or neuropsych* or neuro-psych* or psychiatri* or schedul* or structured or semi-structured or symptom scale*) and interview*).kw.

38 clinical diagnosis.mp.

39 ((clinical* or clinician*) adj3 (administered or checklist* or check-list* or index or indexes or indices or inventory or inventories or instrument? or questionnaire* or scale* or tool*)).tw,kf.

40 ((ICD10 or ICD-10 or ICD11 or ICD-11 or (international classification adj2 disease?) or DSM* or (diagnostic adj2 statistical manual adj2 mental disorder*)) adj diagnos*).tw,kf.

41 or/22-40

42 "sensitivity and specificity"/ or "predictive value of tests"/ or roc curve/ or "limit of detection"/

43 specificity.tw,kf.

44 Validation Study/

45 "reproducibility of results"/

46 (validat* or validity or cross-validat* or internal consistency).tw,kw.

47 likelihood ratio*.tw,kf.

48 ((re-test or retest or test-retest) adj reliability).tw,kf.

49 (ROC or receiver operating characteristic*).tw,kf.

50 (DTA or (diagnos* adj2 accura*)).tw,kf.

51 (performance adj5 (self-report* or scale? or scor* or checklist* or check-list* or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire* or screen* or tool*)).tw,kf.

52 ((degree? or rate* or rating) adj3 agreement?).tw,kf.

53 or/42-52

54 (41 or 53)

55 (21 and 54)

56 anxiety disorders/di or agoraphobia/di or panic disorder/di or phobia, social/di

57 (20 and 56)

58 (55 or 57)

59 limit 58 to yr="1990 -Current"

Note: After peer review, Line-19 was amended to read: ((GAD or generali#ed anxiety disorder) adj3 (self-report* or scale? or scor* or checklist* or check-list* or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire? or screen* or tool*)).tw,kf.

The MEDLINE search was amended after we received this comment in the first round of peer review. All update searches have been correct. The search was also backdated accordingly, to account for this omission.

****

Ovid APA PsycINFO <1806 to September Week 4 2022>

[Date limited: 1990 onwards]

1 (GAD-7* or GAD7* or (GAD adj3 seven) or (GAD adj3 "7")).tw,id,tm.

2 (generali#ed anxiety disorder adj3 (seven or "7")).tw,id,tm.

3 (GAD-2* or GAD2* or (GAD adj3 two) or (GAD adj3 "2")).tw,id,tm.

4 (generali#ed anxiety disorder adj3 (two or "2")).tw,id,tm.

5 ((GAD or generali#ed anxiety disorder) adj3 (self-report* or scale? or scor* or checklist* or check-list* or criteria or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire? or screen* or tool*)).tw,id,tm.

6 or/1-5

7 anxiety disorders/ or generalized anxiety disorder/ or panic attack/ or panic disorder/ or social phobia/

8 (anxiety adj3 (disorder* or neuros* or general* or depress* or social* or unspecif*)).tw,id.

9 social* phobi*.tw,id.

10 GAD.tw,id.

11 ADNOS*.tw,id.

12 (agoraphobi* or panic*).tw,id.

13 anxiety.ti,id.

14 anxi*.ab. /freq=3

15 Anxiety/

16 or/7-15

17 6 and 16

18 psychodiagnosis/ or diagnosis/ or diagnostic criteria/ or ((clinical or psychiatric) adj (diagnosi* or diagnosti*)).tw,id,tm.

19 psychological assessment/ or neuropsychological assessment/

20 (reference standard? or gold standard?).tw,id.

21 *psychometrics/ or rating scales/ or screening tests/

22 interviews/

23 interview schedules/ or diagnostic interview schedule/ or psychodiagnostic interview/

24 structured clinical interview/

25 ((structured adj2 interview*) or SCI or SCID).tw,id,tm.

26 ((standard* adj2 clinic* adj2 interview* adj2 psychiatr*) or SCIP).tw,id,tm.

27 (schedul* adj2 clinic* assess* adj2 neuropsych*).tw,id,tm.

28 (international neuropsych* interview* or MINI).tw,id,tm.

29 mini mental state examination/

30 (mental state examination* or MMSE).tw,id,tm.

31 (composite international diagnos* interview* or CIDI).tw,id,tm.

32 ((anxi* adj2 interview* adj2 schedule*) or ADIS).tw,id,tm.

33 (diagnos* interview* schedule* or DIS).tw,id,tm.

34 (geriatric* mental state* or GMSA).tw,id,tm.

35 ((cambridge examination adj3 mental disorder* adj3 elder*) or CAMDEX).tw,id,tm.

36 "diagnostic and statistical manual"/ or psychodiagnostic typologies/ or "international classification of diseases"/ or research diagnostic criteria/

37 ((clinical* or clinician* or diagnos* or neuropsych* or neuro-psych* or psychiatri* or schedul* or structured or semi-structured or symptom scale*) adj3 interview*).tw,id,tm.

38 ((ICD10 or ICD-10 or ICD11 or ICD-11 or (international classification adj2 disease?) or DSM* or (diagnostic adj2 statistical manual adj2 mental disorder*)) adj diagnos*).tw,id,tm.

39 or/18-38

40 (specificity or (limit adj2 detection) or predictive value).tw,id.

41 test sensitivity/ or exp test reliability/ or exp test validity/

42 predictive validity/

43 (validat* or validity or internal consistency).tw,id.

44 validation study.tw,id.

45 reproducibility.tw,id.

46 likelihood ratio*.tw,id.

47 ((re-test or retest) adj reliability).tw,id.

48 (ROC or receiver operating characteristic*).tw,id.

49 (DTA or (diagnos* adj2 accura*)).tw,id.

50 (performance adj5 (self-report or scale? or scor* or checklist* or check-list* or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire* or screen* or tool*)).tw,id.

51 ((degree? or rate* or rating) adj3 agreement?).tw,id.

52 or/40-51

53 39 or 52

54 17 and 53

55 limit 54 to yr="1990 -Current"

****

NLM PubMed (30 Sep 2022)

(anxiety OR agoraphobia OR agoraphobic OR panic OR ADNOS) AND ("GAD-7" OR GAD7 OR GAD7* OR "GAD-2" OR GAD2 OR GAD2*) AND (DTA OR diagnosis or reference standard OR gold standard OR interview OR SCI OR SCID OR SCIP OR MINI OR schedule OR rating scale OR rating scales OR examination OR MMSE OR CIDI OR ADIS OR DIS OR GMSA OR CAMDEX OR specificity OR predictive OR validation OR validity OR reproducibility OR internal consistency OR likelihood OR reliability OR performance OR agreement OR ROC or receiver operating characteristic) AND (publisher[sb] OR pubmednotmedline[sb])

****

Search narrative

Searching for DTA studies in psychiatry is challenging, because unlike many other medical specialities, psychiatry relies almost exclusively on patient interviews and symptom inventories (questionnaires), used for both diagnostic purposes and to measure treatment outcomes in intervention studies. So, the recommended DTA search structure, using terms for index tests and target condition (only), without a third search concept, would be far too sensitive in the context of this review.
To balance the sensitivity (recall) and specificity (precision) we decided to structure the search around the following concepts: ((Target Condition AND Index Test) AND (Reference Standard OR DTA Filter)).

Reference standard

For the reference standard, we've included a sensitive list of terms for validated psychiatric interview schedules together with generic terms for clinical interviews, or clinician-administered checklists/questionnaires, or broad database subject headings for 'interviews'.

DTA filter

We used a study design filter created by CCMD's experienced information specialist, one which has evolved over many years of searching. It contains search terms included in diagnostic filters listed on websites, such as InterTASC-ISSG (https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home/diagnostic-test-accuracy), together with terms included in known study reports and other DTA reviews. Search terms were also supplied by experienced authors and psychiatrists, working in diagnostics research. Although the filter has not been validated, it went through a number of iterations, checking relative recall, against sets of known study reports (marker papers).

Target condition

Another challenge is that the index test and some of the reference standards, include the word 'anxiety' in the title, which is also a term for the target condition.

Whilst the review focusses on 'anxiety disorders', we appreciate that some reports may use the word 'anxiety' unqualified, to describe the study population. As anxiety is a natural human emotion (often studied experimentally in healthy volunteers), and subclinical anxiety symptoms are common in other branches of medicine, not just psychiatry, we need to address this issue. To balance the sensitivity of the search, we will narrow the choice of search fields for 'anxiety' (unqualified), to the title field, author assigned keywords or subject headings. In the abstract we will use the frequency operator, where the word stem 'anxi*' has to appear three or more times, for the record to be recalled (anxi*.ab. /freq=3). Whilst there is the risk that this approach may drop studies, it retrieved all marker papers when designing the search. The word anxiety was mentioned multiple times in the abstract (i.e. background section, methods, results and conclusion). The search already contains a sensitive list of terms for specific anxiety disorders (agoraphobia, GAD, panic, social phobia, mixed anxiety and depression).

Appendix 2. Rating guidelines for QUADAS-2 and QUADAS-C

PART I: QUADAS-2

Before doing the assessment

Flow diagram

Review the published flow diagram for the primary study or draw one if none is reported or the published diagram is not adequate. The flow diagram will facilitate judgements of risk of bias, and should provide information about the method of recruitment of patients (e.g. based

on a consecutive series of patients with specific symptoms suspected of having the target condition, or of cases and controls), the order of test execution, and the number of patients undergoing the index test and the reference standard. A hand-drawn diagram is sufficient as this step does not need to be reported as part of the QUADAS-2 assessment.

General recommendations/rules

Assessing risk of bias in a specific domain

The assessment of risk of bias focusses on whether specific problems or issues in a study may have led to a systematic over- or underestimation of sensitivity/specificity.

- Rate "low risk of bias" if all signalling questions were rated "yes".

- Rate "high risk of bias" if one or more signalling questions were rated "no".

- Otherwise, rate as "unclear".

Assessing applicability

The assessment of applicability does not focus on bias, but on whether a study fits our review question. Our review has a broad question (diagnostic accuracy of the questionnaires in highly variable populations), and at the same time, the principle of the studies meeting our selection criteria is very straightforward (e.g., all participants giving consent and meeting relatively broad inclusion criteria within the specific setting fill in a questionnaire and get a structured diagnostic review). Therefore, we expect that concerns regarding applicability will be rare.

Give a reason for your judgement if it is not self-evident!

Domain I: Patient selection

Could the selection of participants have introduced bias?

Rationale: "A study should ideally enrol all consecutive, or a random sample of, eligible patients with suspected disease – otherwise there is potential for bias. Studies that make inappropriate exclusions, e.g. excluding "difficult to diagnose" patients, may result in overoptimistic estimates of diagnostic accuracy. Similarly, studies enrolling patients with known diseases and a control group without the condition may exaggerate diagnostic accuracy. Exclusion of patients with "red flags" for the target condition, who may be easier to diagnose, may lead to underestimation of diagnostic accuracy."

Signalling question I.1: Was a consecutive or random sample of patients enrolled?

Remark: In our study set, we expect a number of studies in which potential participants were pre-selected regarding issues not related to anxiety diagnoses (e.g., they were participants in a study on quality of life amongst patients with epilepsy, or they were frequent attenders participating in a case management program) or for other reasons (e.g., they responded to advertisements of the study).

Rate "yes" (low risk of bias) if:

- the study used adequate random or consecutive sampling and the percentage of consented and eligible patients who participated is ≥75%.

Rate "unclear" if:

- the description is insufficient to make a judgement OR

- if the study used random or consecutive sampling, but the percentage of eligible patients participating is unclear or below 75%

Rate "no" (high risk of bias) if:

- if convenience sampling was used

- if the study claimed to use random or consecutive sampling, but the method seems inadequate (give reasons)

- if the sample came from a study with a primary objective other than DTA assessment of an anxiety questionnaire, and the sampling in this study cannot be considered random or consecutive in respect to DTA anxiety

Signalling question I.2: Did the study avoid unnecessary exclusions?

Rate "yes" (low risk of bias) if:

- there were no exclusion criteria at study enrolment (beyond informed consent or age) OR exclusions were unavoidable (e.g. emergencies, language problems) OR were necessary to avoid bias (e.g., known major psychiatric disease)

Rate “unclear” if:

- the description is insufficient to make a judgement

Rate “no” (high risk of bias) if:

- the sampling allowed exclusion of participants with specific characteristics potentially adversely influencing diagnostic accuracy.

Domain II: Index test

Signalling question II.1: Were the index test results interpreted without knowledge of the results of the reference standard?

Rate “yes” if

- The process description makes clear that the index test(s) was (were) filled before the structured interview was performed OR

- There was an explicit statement that the questionnaire rating and structured interview were performed “independently”, that participants and raters were “blinded” to each other's rating...

Rate “unclear”

- if the process description is unclear, but no indication of high risk of bias (see answer “no”)

Rate “no” if:

- participants were aware or likely to be aware of the result of the reference standard

Signalling question II.2: Was there a predefined threshold for a primary analysis of DTA?

Rate “yes”

- if the cut-off value was explicitly prespecified or a clear a-priori reason for the choice (e.g. “we used the cut-off of ≥10 recommended by the scale developers”) is described
[in some exceptional cases there might be two pre-defined cut-offs based on literature recommendations – this is acceptable, too]

Rate “no”

- if the cut-off was chosen post-hoc (e.g. as it maximized sensitivity and specificity)

- if a cut-off different from the cut-off recommended by the scale authors (GAD-7 ≥10; GAD-2 ≥3) without being explicitly pre-defined

Rate-“unclear”

- if the cut-off recommended by the scale authors (GAD-7 ≥10; GAD-2 ≥3) was used, but there is no statement about whether this choice was made in advance

Rate “not applicable”

- if there was no predefined main cut-off

Signalling question II.3: If findings for several thresholds are presented, was the range of thresholds prespecified?

Rate “yes”

- if the range of cut-off values presented was explicitly prespecified OR if for the GAD-2 findings for all cut-offs were presented; for the GAD-7 for at least 10 cut-offs (covering a range including at least +/- 3 around the recommended cut-offs)

Rate “no”

- if the conditions for the answer “yes” are not met

Rate “not applicable”

- if there was a pre-defined main cut-off

Domain III: Reference Standard

Signalling question III.1: Is the reference standard likely to correctly classify the target condition?

Remark: for the assessment of risk of bias, we assume that the structured clinical interviews accepted for inclusion in the review are adequate reference standards, unless they have been applied inadequately. The problem of whether specific interviews or interviews in general are the true gold standard will be addressed in the discussion of the reviews.

Rate "yes"

- if the study uses a structured interview meeting our inclusion criteria and there are no reasons to assume that it had been performed inadequately

Rate "no"

- if the reference standard has been applied in an inadequate manner (e.g., untrained students or lay persons)

Signalling question III.2: Were the reference standard results interpreted without knowledge of the results of the index test?

Rate "yes" if

- there is an explicit statement that the questionnaire rating and structured interview were performed "independently", that participants and raters were "blinded" to each other's rating...

Rate "unclear"

- if the process description is unclear, but no indication of high risk of bias (see answer "no")

Rate "no" if:

- interviewers were aware or likely to be aware of the questionnaire rating

Domain IV: Study Flow and Timing

Signalling question 1: Was there an appropriate interval between the index test and the reference standard?

Rate "yes" if

- there is an explicit statement that the time lag between the questionnaire rating and structured interview was not more than two weeks

Rate "unclear"

- if the process description is unclear, but no indication of high risk of bias (see answer "no")

Rate "no" if:

- the time lag is greater than 2 weeks (if it exceeds 4 weeks, exclude the study)

Signalling question 2: Did all patients receive a reference standard?

Rate "yes" if:

- at least 90% of participants completing the index text(s) also underwent the reference standard AND there are no hints that missing reference standards were due to reasons related to the outcome of the index test(s)

Rate "unclear"

- the process description is unclear, but no indication of high risk of bias (see answer "no")

Rate "no" if:

- less than 95% of participants completing the index text(s) also underwent the reference standard OR

- there are hints that missing reference standards were due to reasons related to the outcome of the index test(s)

Signalling question 3: Were all patients included in the analysis?

Rate "yes" if

- at least 90% of participants (individuals meeting inclusion criteria and giving consent) were also included in the analyses, AND there is no indication that exclusions were related to the results of the index test/reference standard or other factors likely to cause bias

Rate “unclear”

- if the process description is unclear, but no indication of high risk of bias (see answer “no”)

Rate “no” if:

- less than 90% of participants were included in the analyses, AND/OR there are reasons to believe that exclusions were related to the results of the index test/reference standard or other likely to induce bias

Applicability

Applicability question A.1: Are there concerns that the included participants do not match the review question?

Rate “no” if the study is likely to include a representative sample of patients in respect to our four pathways/settings and to the presence of a screening scenario. Rate “yes” if the study setting and/or the sample of participants are likely to be not representative in respect to our three pathways/settings. Rate “unclear” in the remaining studies. When checking whether a screening scenario was present, consider whether participants' anxiety diagnoses were pre-existing; if participants with pre-existing anxiety were included, rate as "yes."

Applicability question A.2: Are there concerns that the index test, its conduct, or interpretation differ from the review question?

Rate “no” if the questionnaire is filled in by participants themselves (on paper or electronically) or by an interviewer who is reading the questions and answer options for the participants literally, and then documents the answers. Rate “yes” (or “unclear”) only if the questionnaire has been used in an inadequate or highly unusual manner (e.g.: modified questionnaire items).

Applicability question A.3: Are there concerns that the target condition as defined by the reference standard does not match the review question?

Rate “no” unless the structured diagnostic interview (meeting our selection criteria) is performed in an inadequate manner (e.g.: untrained lay persons).

PART II: Rating Guidelines QUADAS-C

When do I have to do the QUADAS-C assessment?

Only if the study includes a comparison of the diagnostic accuracy of two or more of the questionnaires reviewed by us (GAD-7 and GAD-2).

The general rules for assessing the risk of bias for a specific domain based on the signalling questions are the same as for QUADAS-2 (see Page 1).

Domain C-I: Patient selection

Signalling question C--I.1: Was the risk of bias in each index test judged ‘low’ for this domain?

If the accuracy estimates of one or more index tests are considered to be at high risk of bias, their comparison will also be at high risk of bias. Therefore, each domain in QUADAS-C starts with the question of whether the risk of bias for that domain for each index test (as judged using QUADAS-2) is ‘low’. If one or more index tests in the comparison is judged ‘unclear’ or ‘high’ in a QUADAS-2 domain, this question should be answered ‘no’.

Rate “yes” if:in

- you judged the risk of bias for this domain ‘low’ in QUADAS-2

Rate “no” if:

- you judged the risk of bias in this domain ‘unclear’ or ‘high’ in QUADAS-2

Signalling question C-I.2: Was a fully paired or randomized design used?

Rate “yes” if:

- all participants (were intended to) get both all index tests and the reference standard (fully paired) OR if

- all participants (were intended to) get both all index tests + all participants screening positive in any of the index tests and a random subgroup of those screening negative (were intended to) undergo the reference standard

Rate “unclear” if:

- other forms of randomization were used

Rate “no” if:

- a fully paired or a randomized design were not used

Domain C-II: Index test

Signalling question C-II.1: Was the risk of bias in each index test judged 'low' for this domain?

See C-I.1

Signalling question C-II.2: Were the index test results interpreted without knowledge of the results of the other index test(s)?

You should copy your rate from QUADAS-2 to here

Signalling question C-II.3: Is undergoing one index test unlikely to affect the performance of the other index test(s)?

This is a tricky question in our case, because in our study set, usually one index questionnaire is filled in immediately after the other. We will discuss this problem on a general level, but rate the item pragmatically: Rate "yes" unless there are reasons that in the specific study the risk of bias is high or unclear.

Signalling question C-II.4: Were the index tests conducted and interpreted without advantaging one of the tests?

In our specific study set (using self-report questionnaires with a simple answer count for interpretation), this should rarely ever be a problem. Rate "yes" unless there are reasons that in the specific study the risk of bias is high or unclear.

Domain C-III: Reference standard

Signalling question C-III.1: Was the risk of bias in each index test judged 'low' for this domain?

See C-I.1

Signalling question C-III.2: Did the reference standard avoid incorporating any of the index tests?

Again, this is a tricky question in our case, because obviously, questionnaires and interviews overlap considerably. We will discuss this problem on a general level, but rate the item pragmatically: Rate "yes" unless there are reasons that in the specific study the risk of bias is high or unclear (different from other studies).

Domain C-IV: Study flow and timing

Signalling question C-IV.1: Was the risk of bias in each index test judged 'low' for this domain?

See C-I.1

Signalling question C-IV.2: Was there an appropriate interval between the index tests?

Rate "yes" if:

- the study report explicitly states or clearly suggests (description such as: "after giving consent participants filled in the questionnaires A and B") that questionnaires were filled immediately one after the other

Rate "unclear" if:

- there was an interval up to one week

Rate "no" if:

- there was a longer interval

Signalling question C-IV.3: Was the same reference standard used for all index tests?

Rate "yes" if:

- this was the case (we expect that this applies to all included studies)

Rate "no" if:

- this was not the case

Signalling question C-IV.4: Are the proportions and reasons for missing data similar across index tests?

Rate "yes" if:

- the difference(s) in the number of participants with missing data is very minor (<3% in studies with up to 100 participants, <2% up to 200, <1% if more than 200 participants) and the reasons are similar

Rate "no" if:

- if this is not the case, or reasons for missing data differ in an important way

Rate “unclear” if:

- there is no information on the number of participants analysed, and the question cannot be answered

Appendix 3. Summary tables

Summary characteristics of the overall study sample in Table 2; characteristics of the individual studies in Table 3; assessment of risk of bias and applicability (QUADAS-2 and QUADAS-C) for individual studies in Table 1.

Appendix 4. QUADAS-2 and QUADAS-C risk of bias assessment summary graphs

Summary graphs showing the risk of bias assessments for the four study questions are given in Figure 12, Figure 13, Figure 14, and Figure 15. QUADAS-C risk of bias assessment summary graphs comparing GAD-7 and GAD-2 in all studies and the two target conditions are shown in Figure 16, Figure 17, and Figure 18.

Figure 12. Risk of bias and applicability concerns for studies (n = 35) evaluating the GAD-7 for detecting generalised anxiety disorder (GAD): review authors' judgements for QUADAS-2 domains

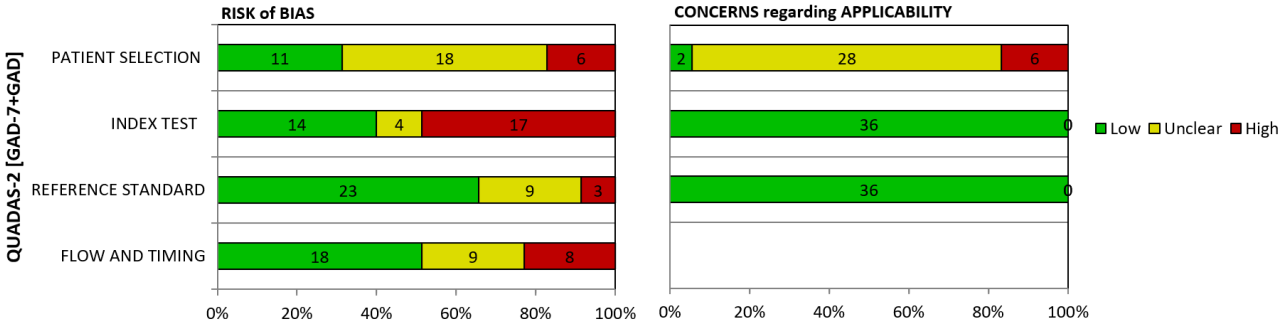


Figure 13. Risk of bias and applicability concerns for studies (n=22) evaluating the GAD-7 for detecting any anxiety disorder (AAD): review authors' judgements for QUADAS-2 domains

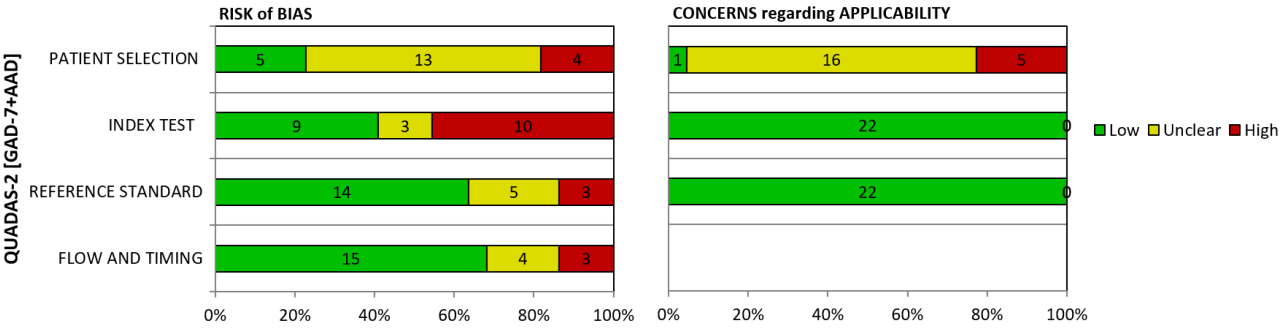


Figure 14. Risk of bias and applicability concerns for studies (n=24) evaluating the GAD-2 for detecting generalised anxiety disorder (GAD): review authors' judgements for QUADAS-2 domains

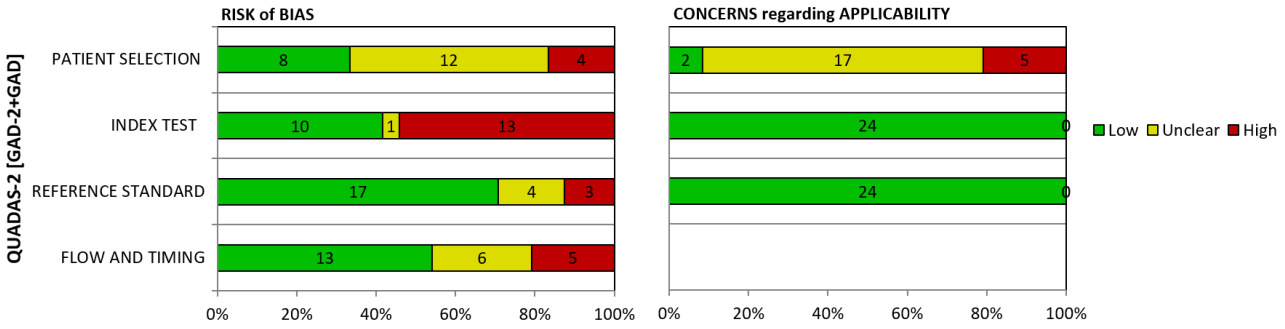


Figure 15. Risk of bias and applicability concerns for studies (n=19) evaluating the GAD-2 for detecting any anxiety disorder (AAD): review authors' judgements for QUADAS-2 domains

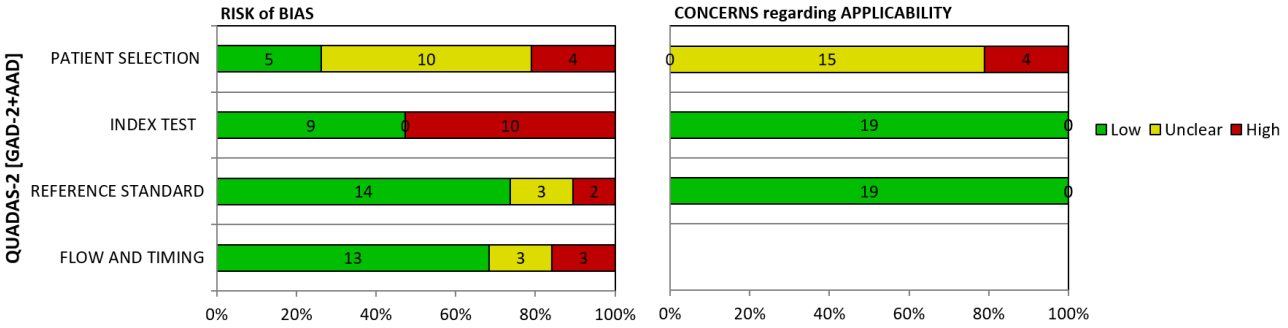


Figure 16. Risk of bias in all studies (n=25) allowing a comparison of the diagnostic accuracy GAD-7 and GAD-2 for detecting generalised anxiety disorder and/or any anxiety disorder (AAD): review authors' judgements for QUADAS-C domains

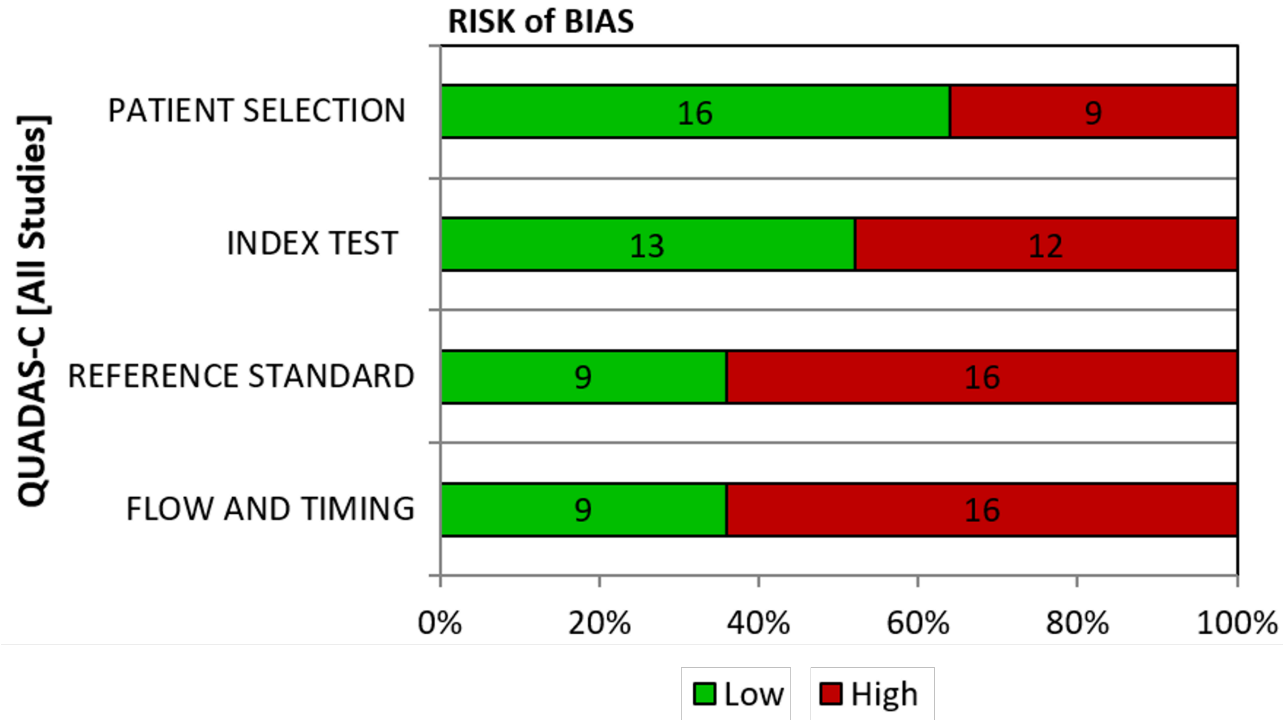


Figure 17. Risk of bias in studies (n=22) allowing a comparison of the diagnostic accuracy GAD-7 and GAD-2 for detecting generalised anxiety disorder (GAD): review authors' judgements for QUADAS-C domains

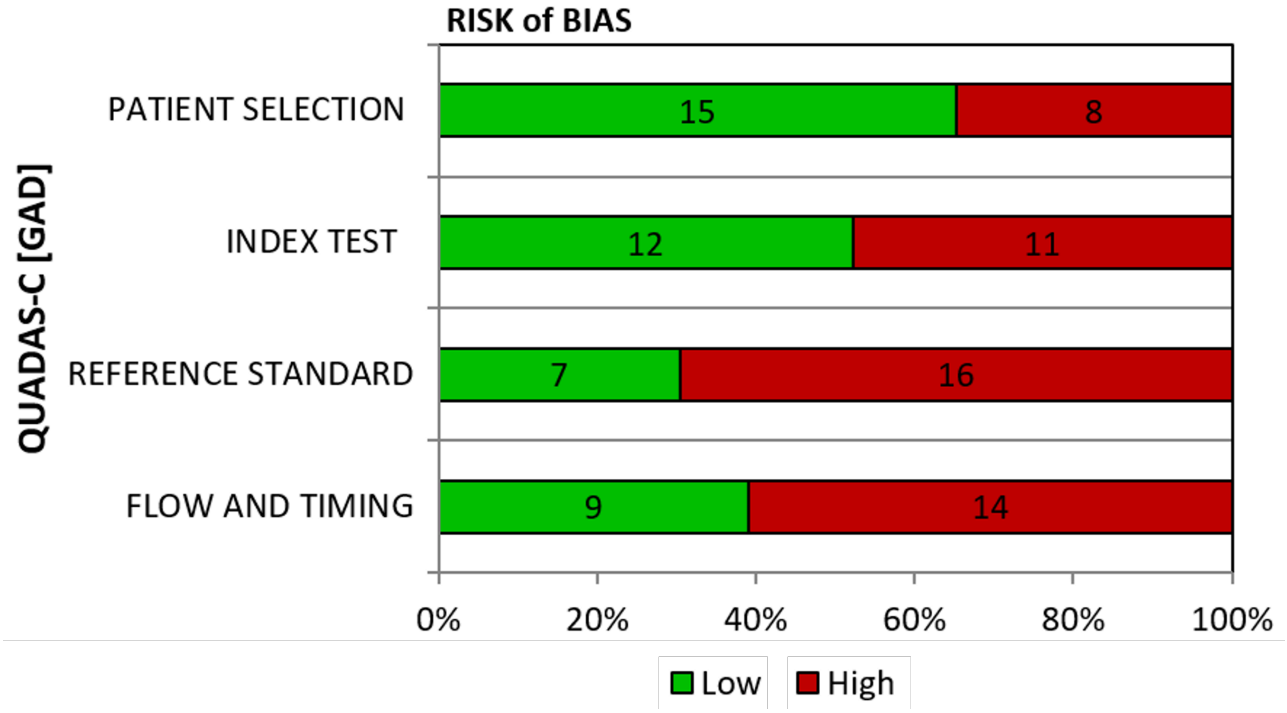
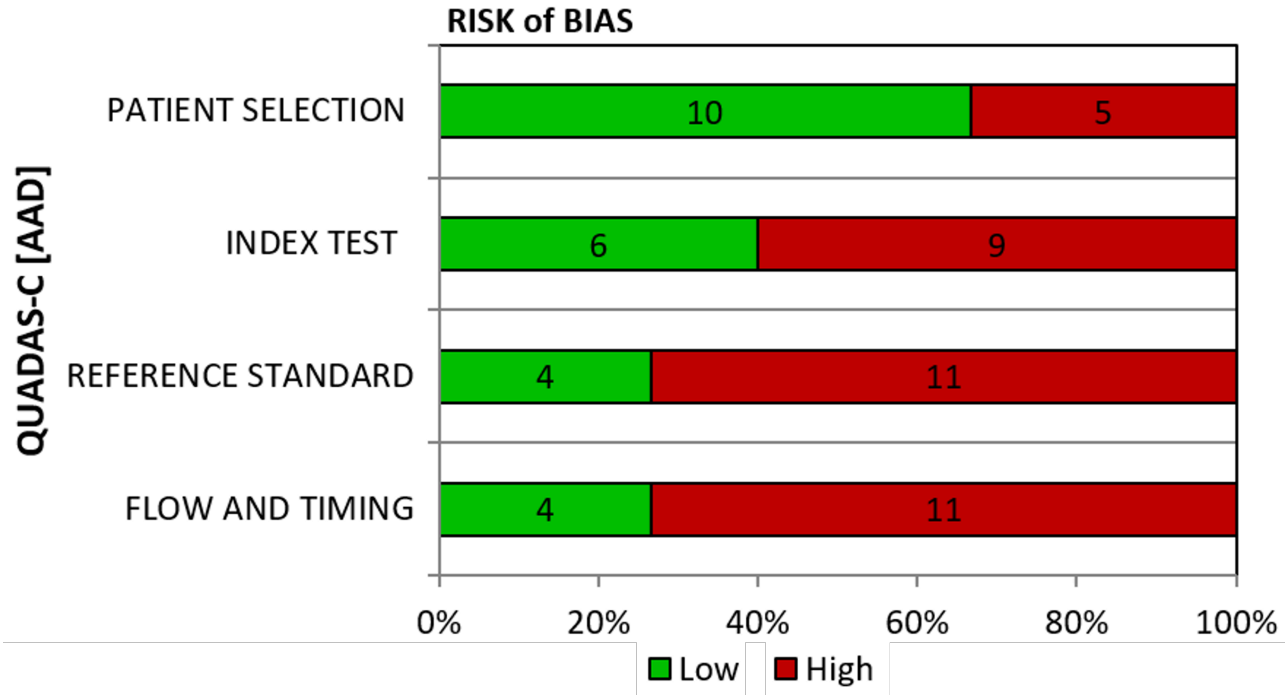


Figure 18. Risk of bias in studies (n=15) allowing a comparison of the diagnostic accuracy GAD-7 and GAD-2 for detecting any anxiety disorder (AAD): review authors' judgements for QUADAS-C domains



Appendix 5. GAD-7 for detecting generalised anxiety disorder – additional figures and tables

The SROC plot across studies is given in Figure 19; SROC plots for subgroups in Figure 19, Figure 20, Figure 21, and Figure 22; heterogeneity and comparisons between covariates in Table 4; secondary analyses (summary estimates for each cut-off based on bivariate and multiple threshold models) in Table 5.

Figure 19. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting generalised anxiety disorder (GAD) Open circles indicate the individual sensitivity and specificity for each study. The filled circle indicates the summary estimate. Circle sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

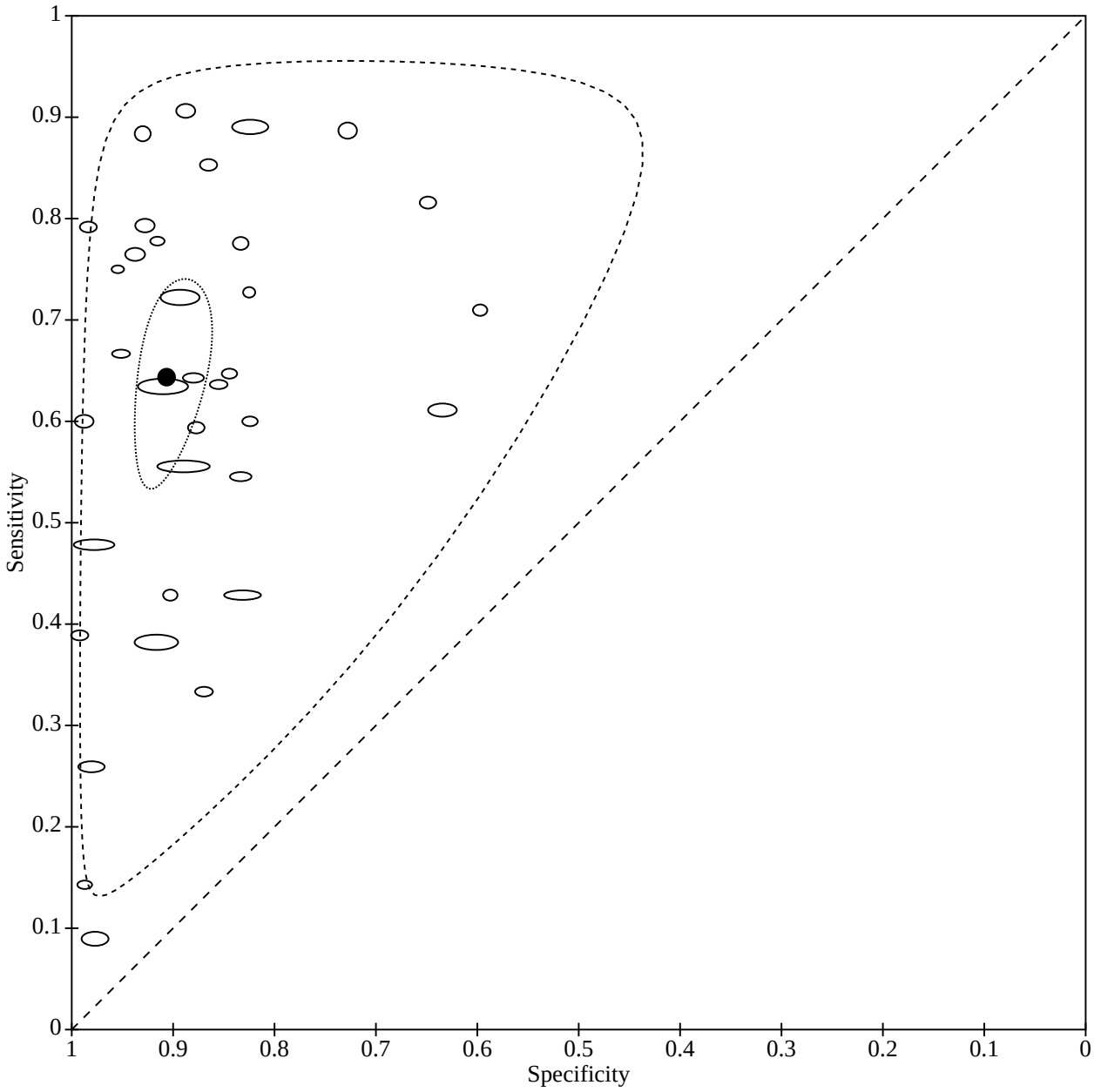


Figure 20. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting generalised anxiety disorder (GAD) for the setting subgroups Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles, rectangles and diamonds indicate the summary estimate. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

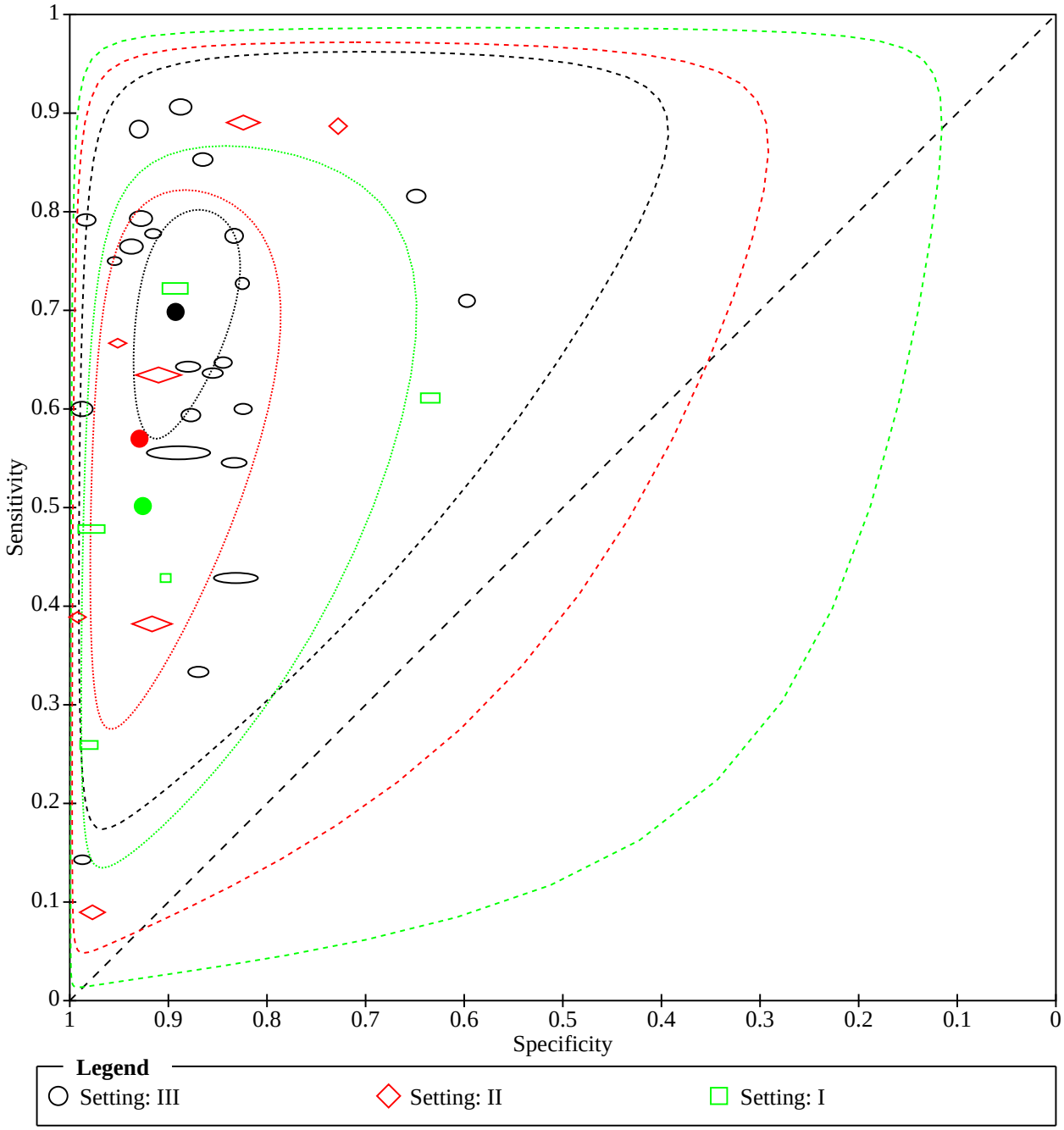


Figure 21. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting generalised anxiety disorder (GAD) for the reference standard subgroups Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles, rectangles and diamonds indicate the summary estimate. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

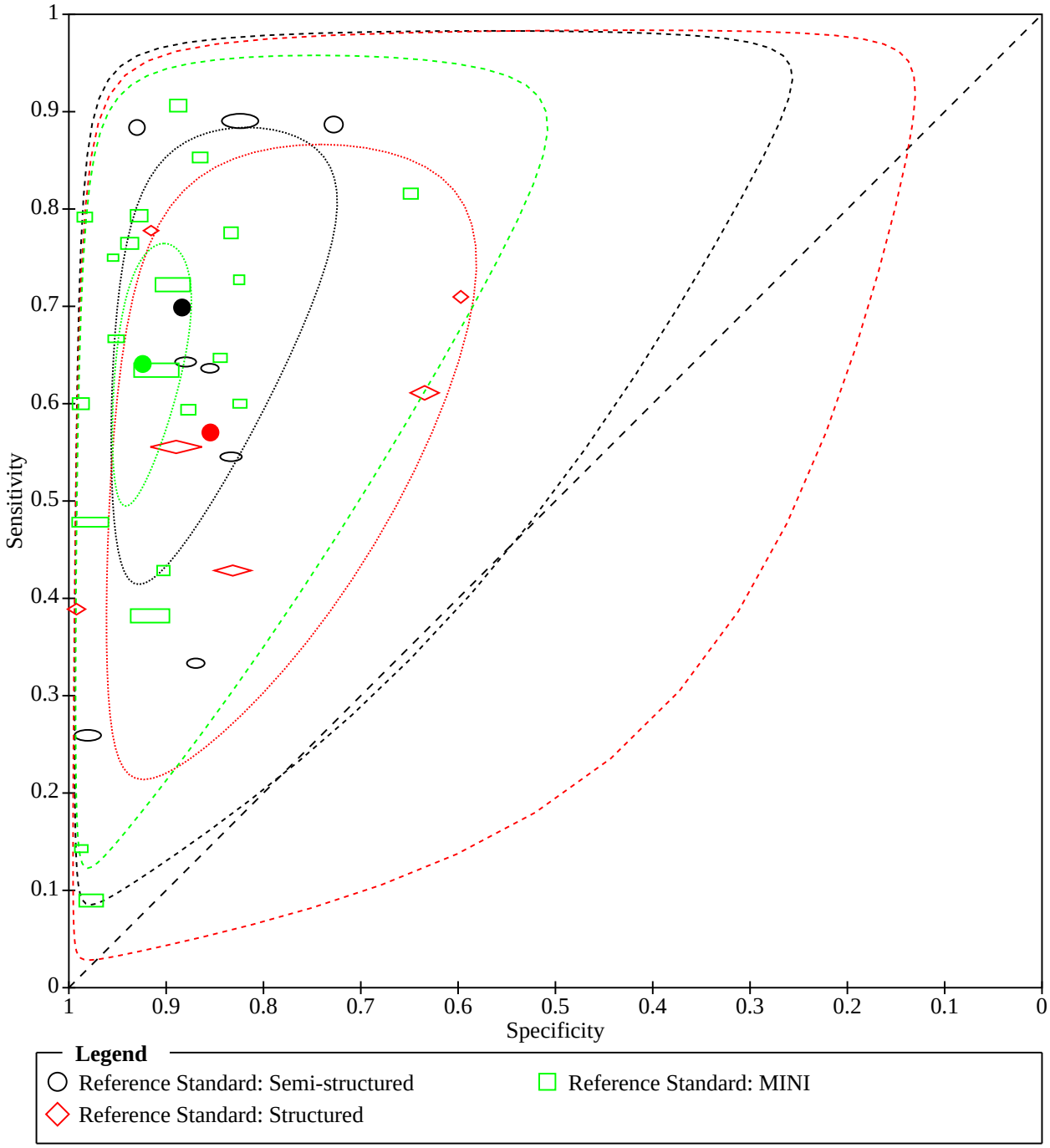
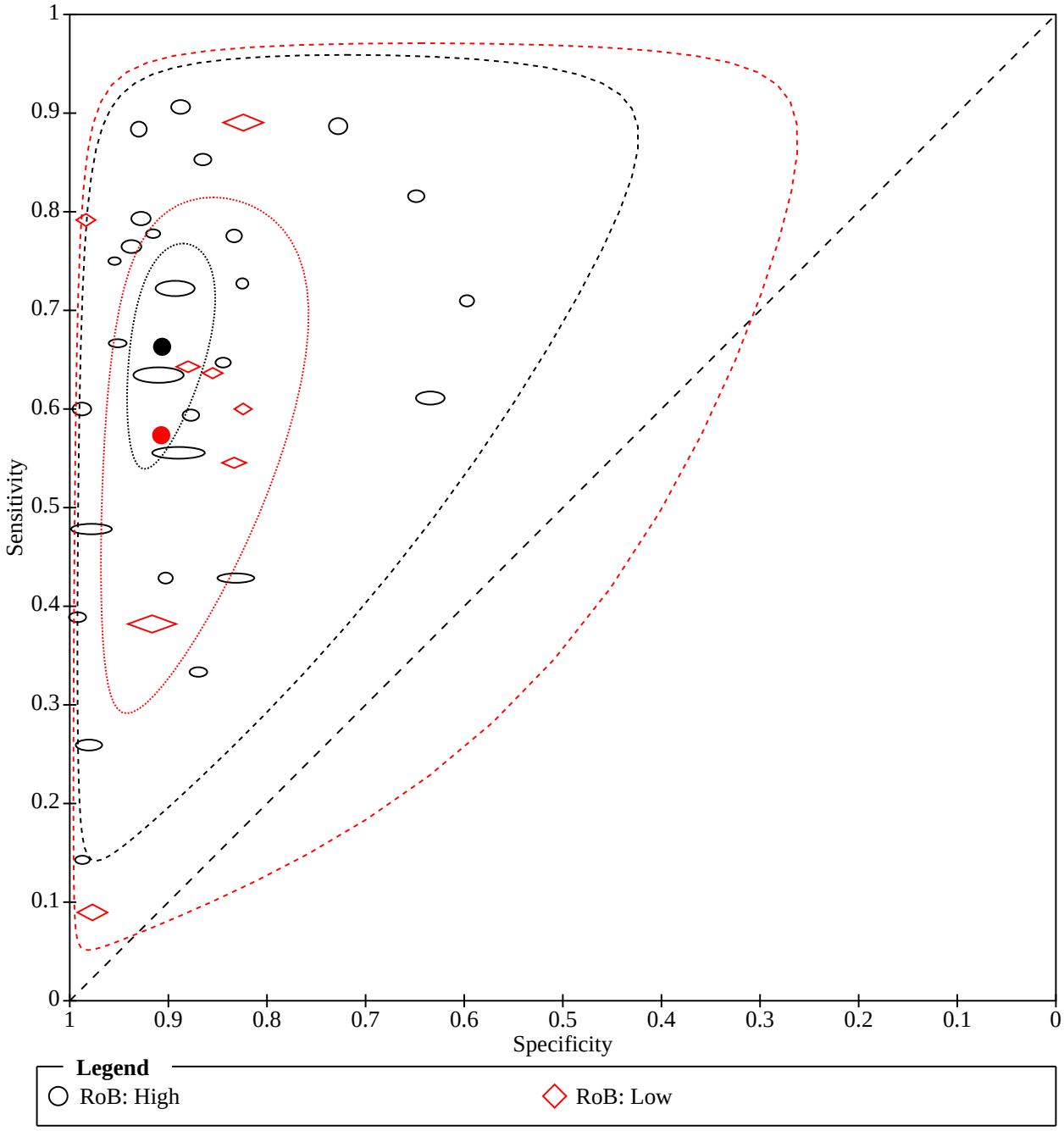


Figure 22. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting generalised anxiety disorder (GAD) for high and low risk of bias (RoB) subgroups Open circles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimate. Circle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.



Appendix 6. GAD-7 for detecting any anxiety disorder – additional figures and tables

The SROC plot across studies in Figure 23; SROC plots for subgroups in Figure 19; Figure 24; Figure 25; Figure 26; heterogeneity and comparisons between covariates in Table 6; secondary analyses (summary estimates for each cut-off based on bivariate and multiple threshold models) in Table 7.

Figure 23. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting any anxiety disorder (AAD) Open circles indicate the individual sensitivity and specificity for each study. The filled circle indicates the summary estimate. Circle sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

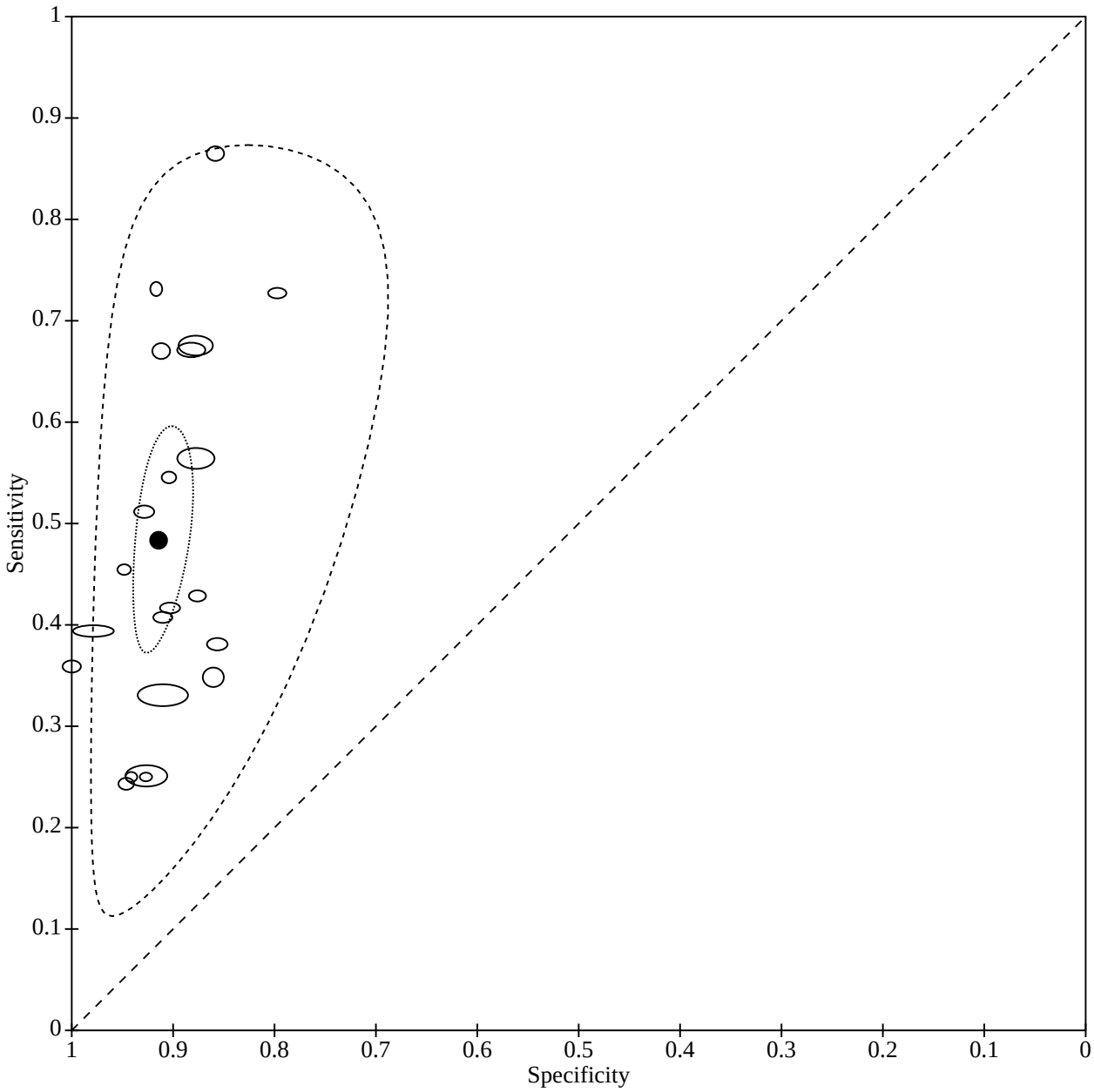


Figure 24. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting any anxiety disorder (AAD) for the setting subgroups Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimate. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

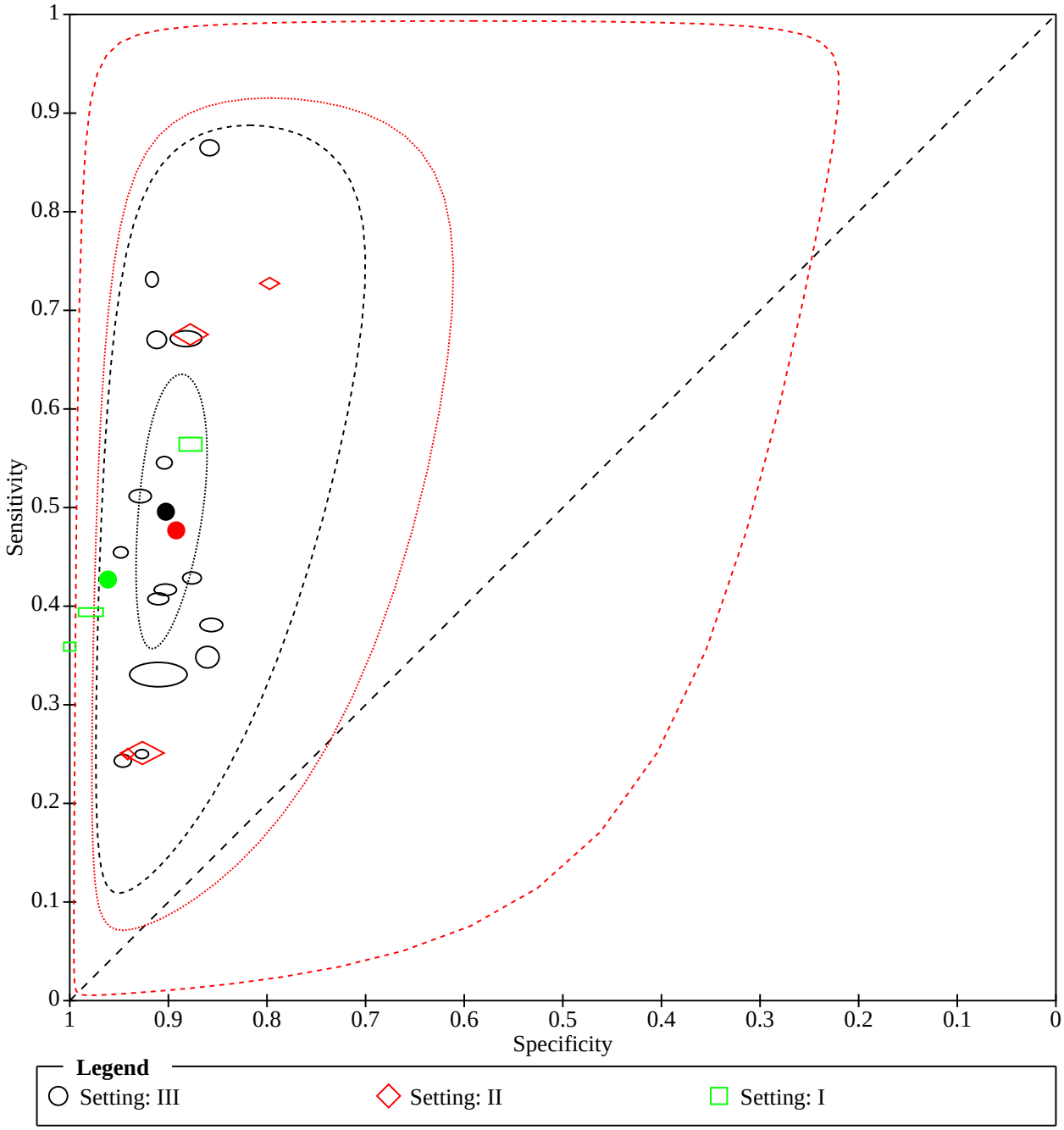


Figure 25. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting any anxiety disorder (AAD) for the reference standard subgroups Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

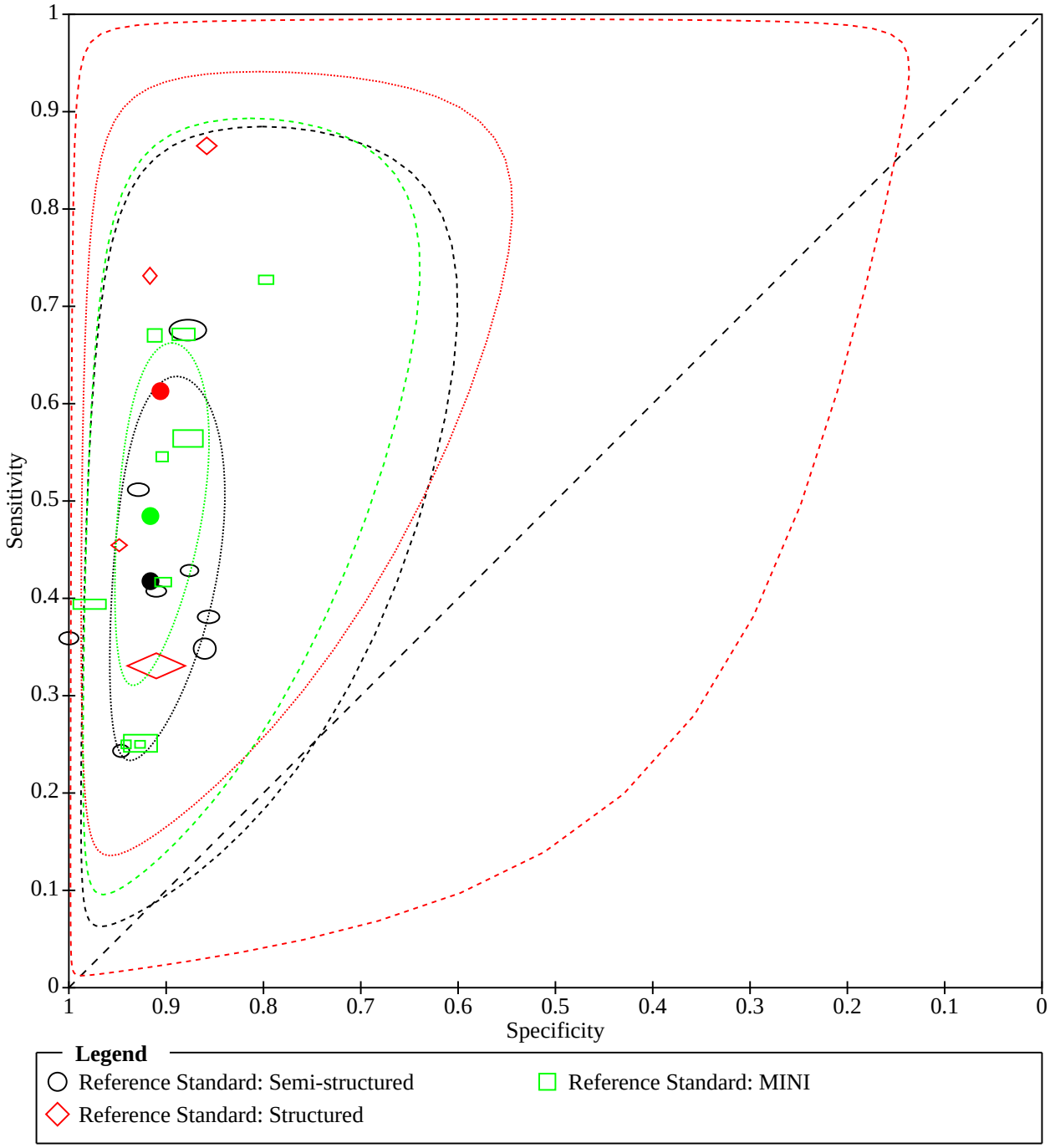
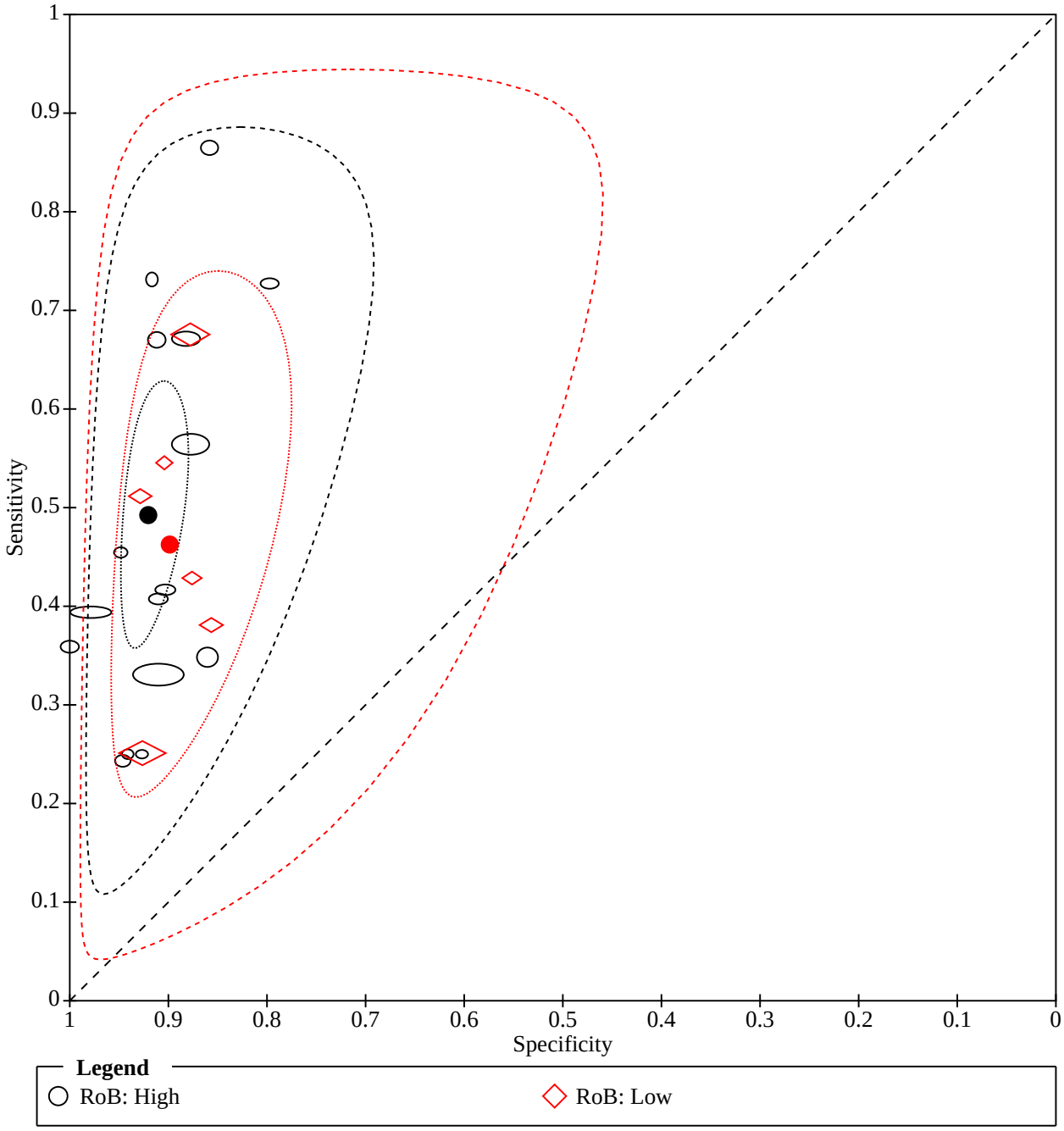


Figure 26. Summary receiver operating characteristic (SROC) plot of GAD-7 for detecting any anxiety disorder (AAD) for high and low risk of bias (RoB) subgroups Open circles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.



Appendix 7. GAD-2 for detecting generalised anxiety disorder – additional figures and tables

SROC plot across studies in Figure 27; SROC plots for subgroups in Figure 28, Figure 29, and Figure 30; heterogeneity and comparisons between covariates in Table 8; secondary analyses (summary estimates for each cut-off based on bivariate and multiple threshold models) in Table 9.

Figure 27. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting generalised anxiety disorder (GAD) Open circles indicate the individual sensitivity and specificity for each study. The filled circle indicates the summary estimate. Circle sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

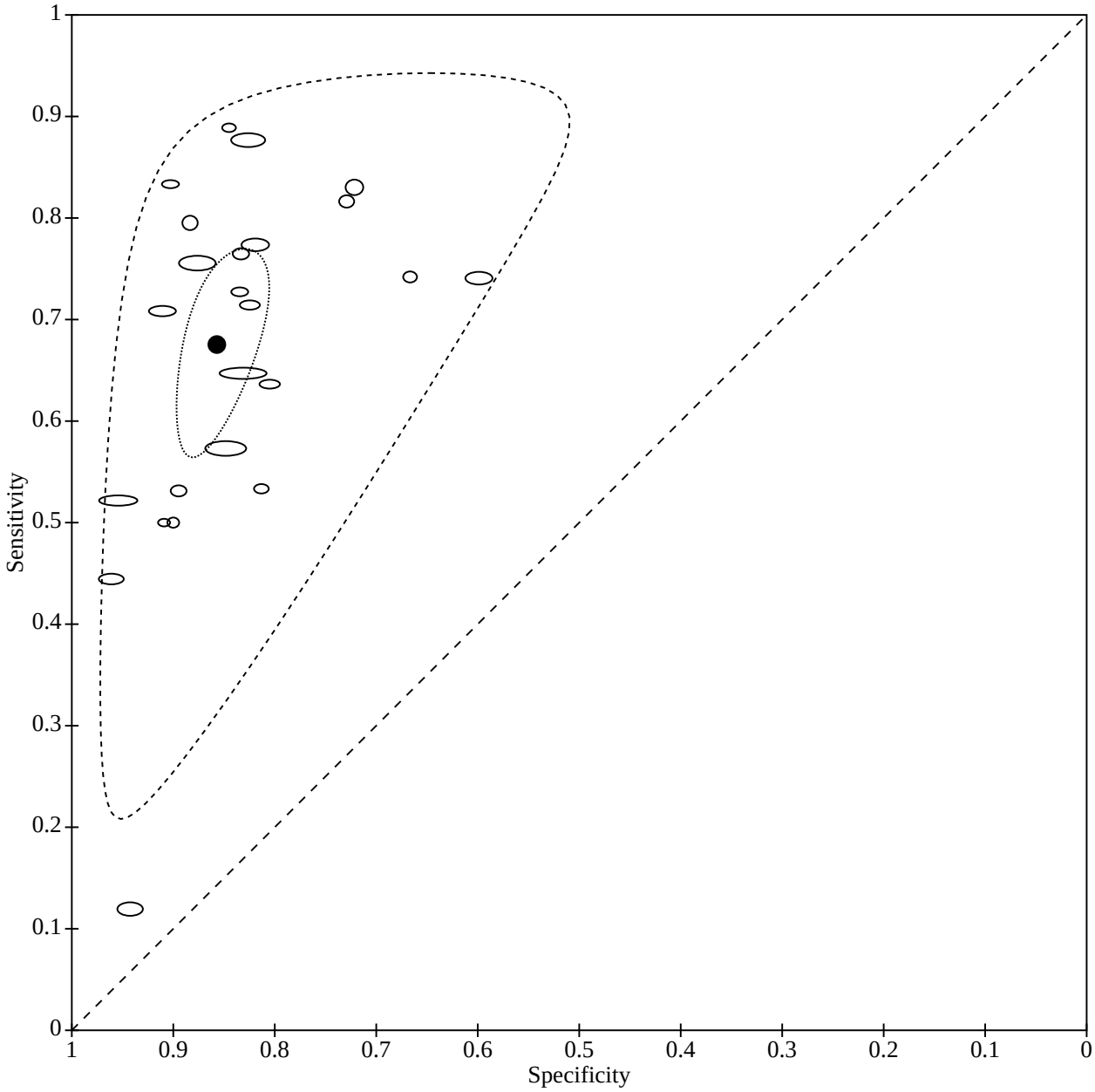


Figure 28. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting generalised anxiety disorder (GAD) for the individual setting subgroups Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

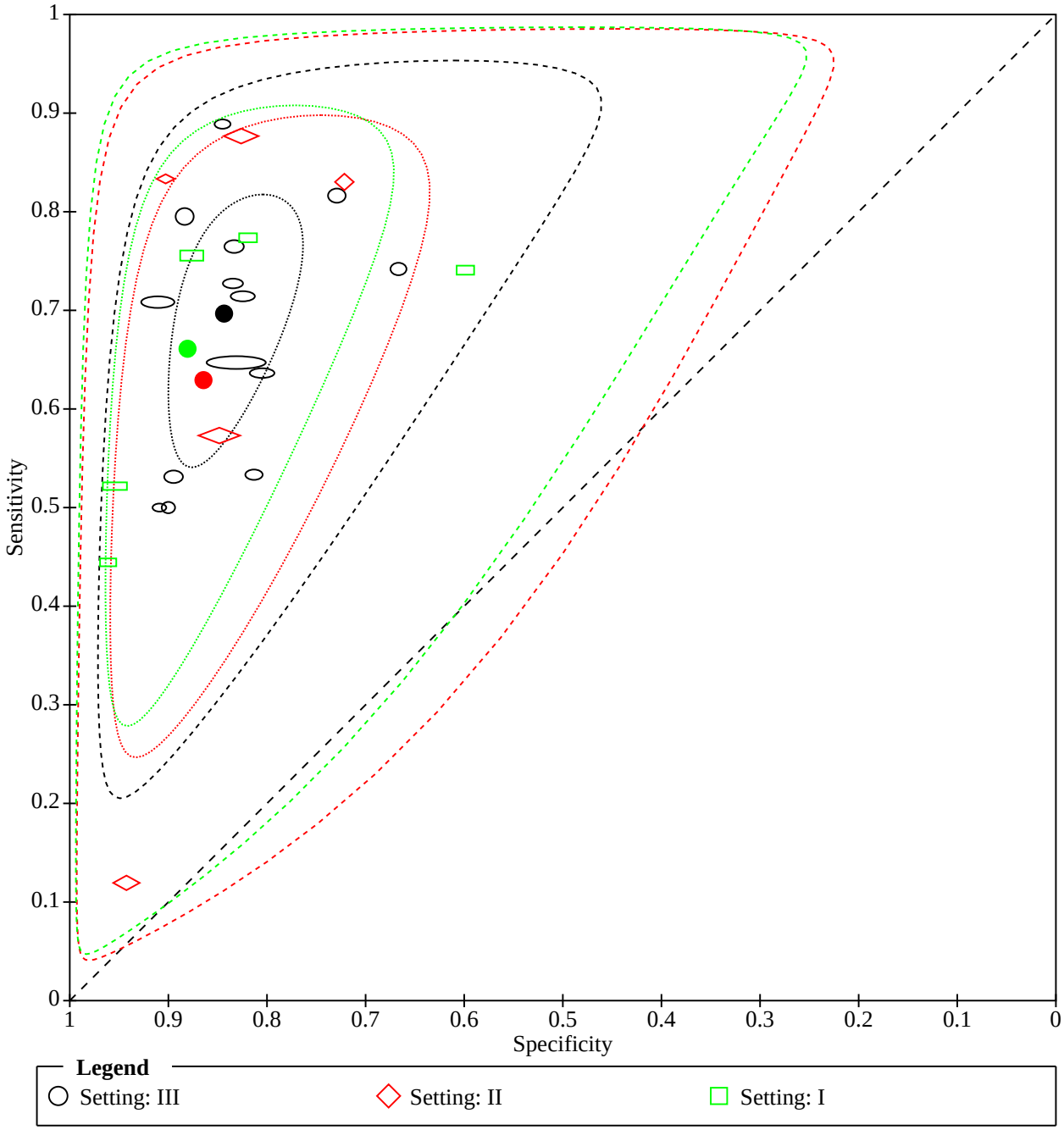


Figure 29. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting generalised anxiety disorder (GAD) for the reference standard subgroups Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

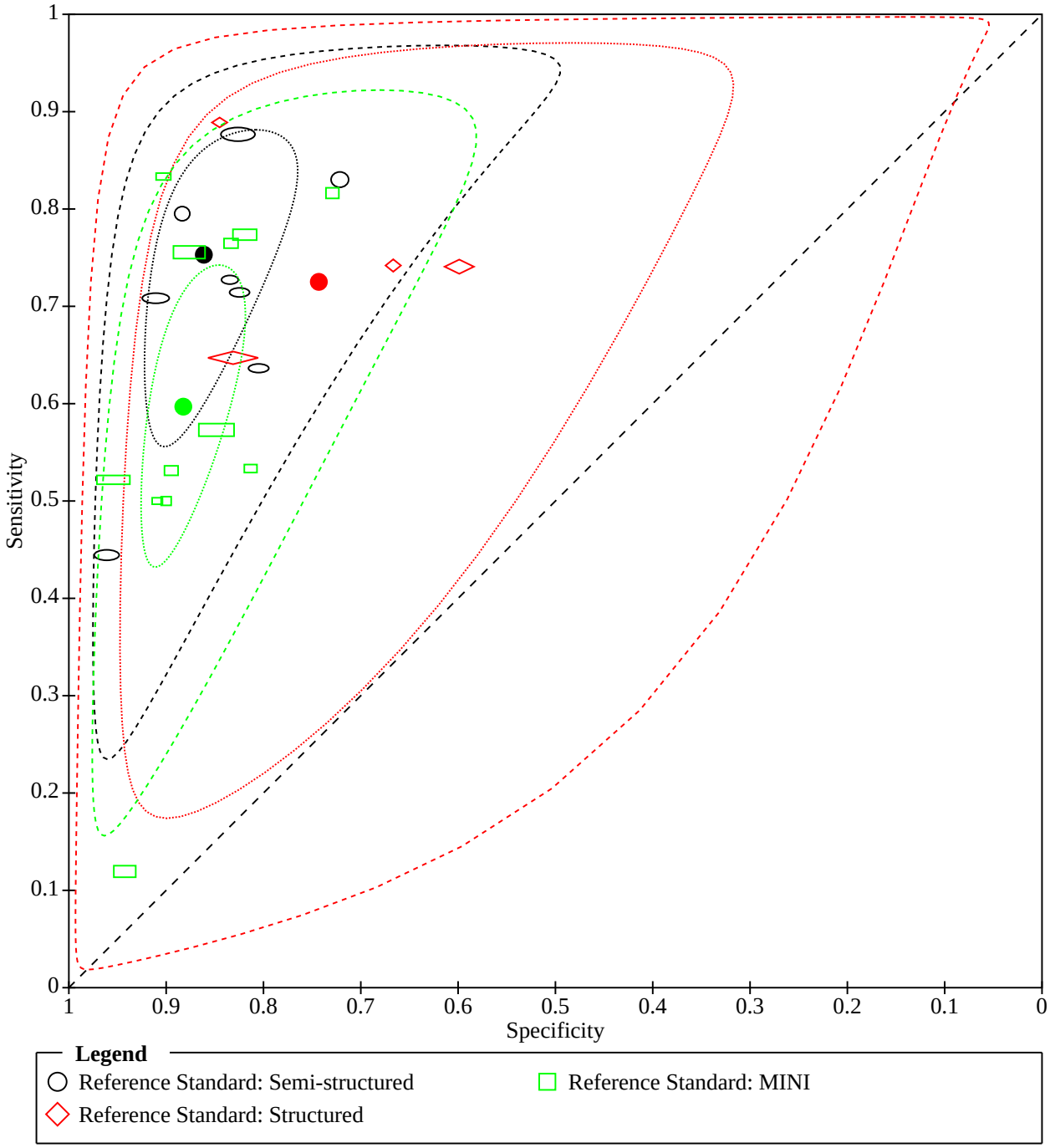
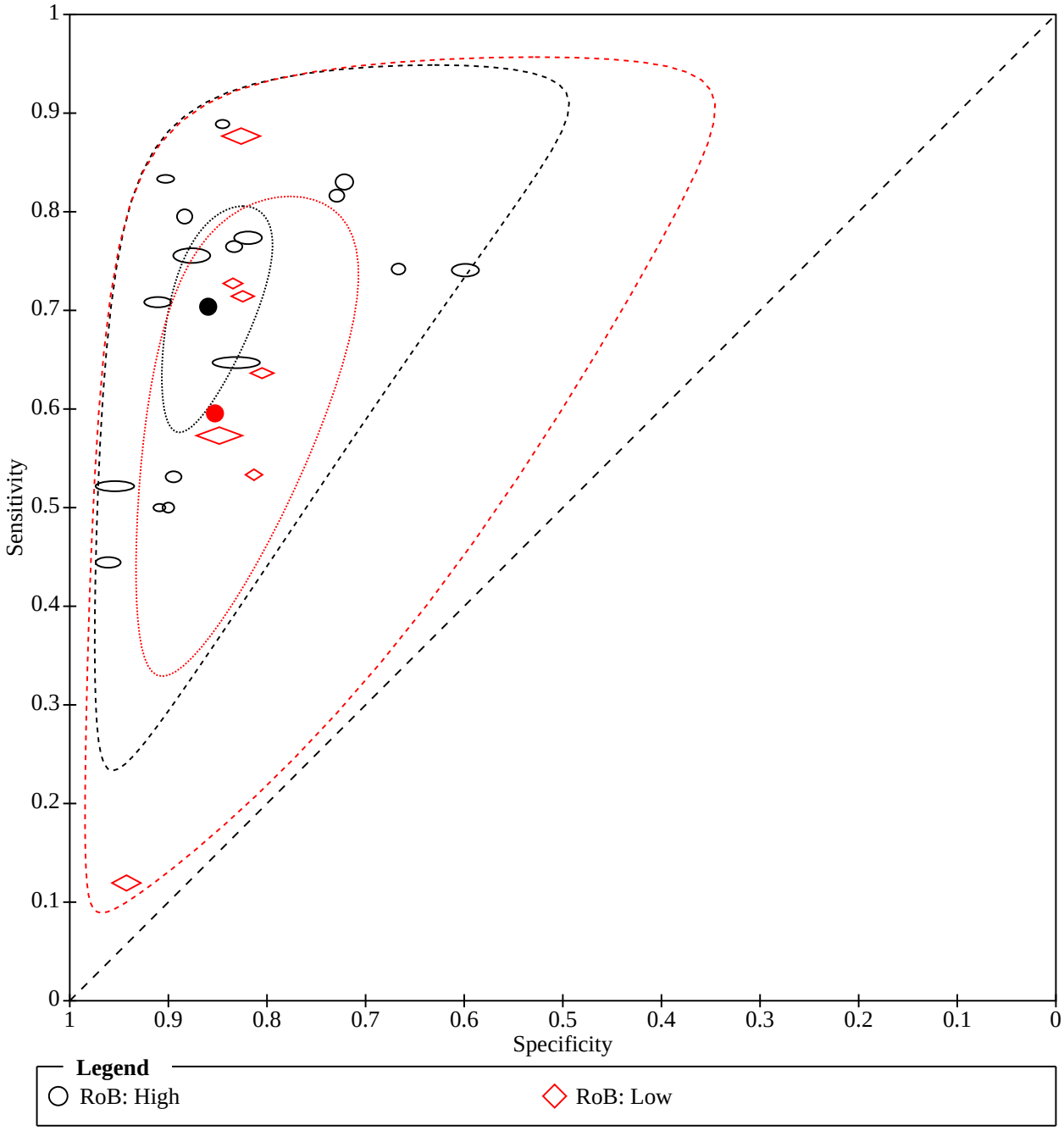


Figure 30. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting generalised anxiety disorder (GAD) for high and low risk of bias (RoB) subgroups Open circlesand diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.



Appendix 8. GAD-2 for detecting any anxiety disorder – additional figures and tables

SROC plot across studies in Figure 31; SROC plots for subgroups in Figure 32, Figure 33, and Figure 34; heterogeneity and comparisons between covariates in Table 10; secondary analyses (summary estimates for each cut-off based on bivariate and multiple threshold models) in Table 11.

Figure 31. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting any anxiety disorder (AAD) Open circles indicate the individual sensitivity and specificity for each study. The filled circle indicates the summary estimate. Circle sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

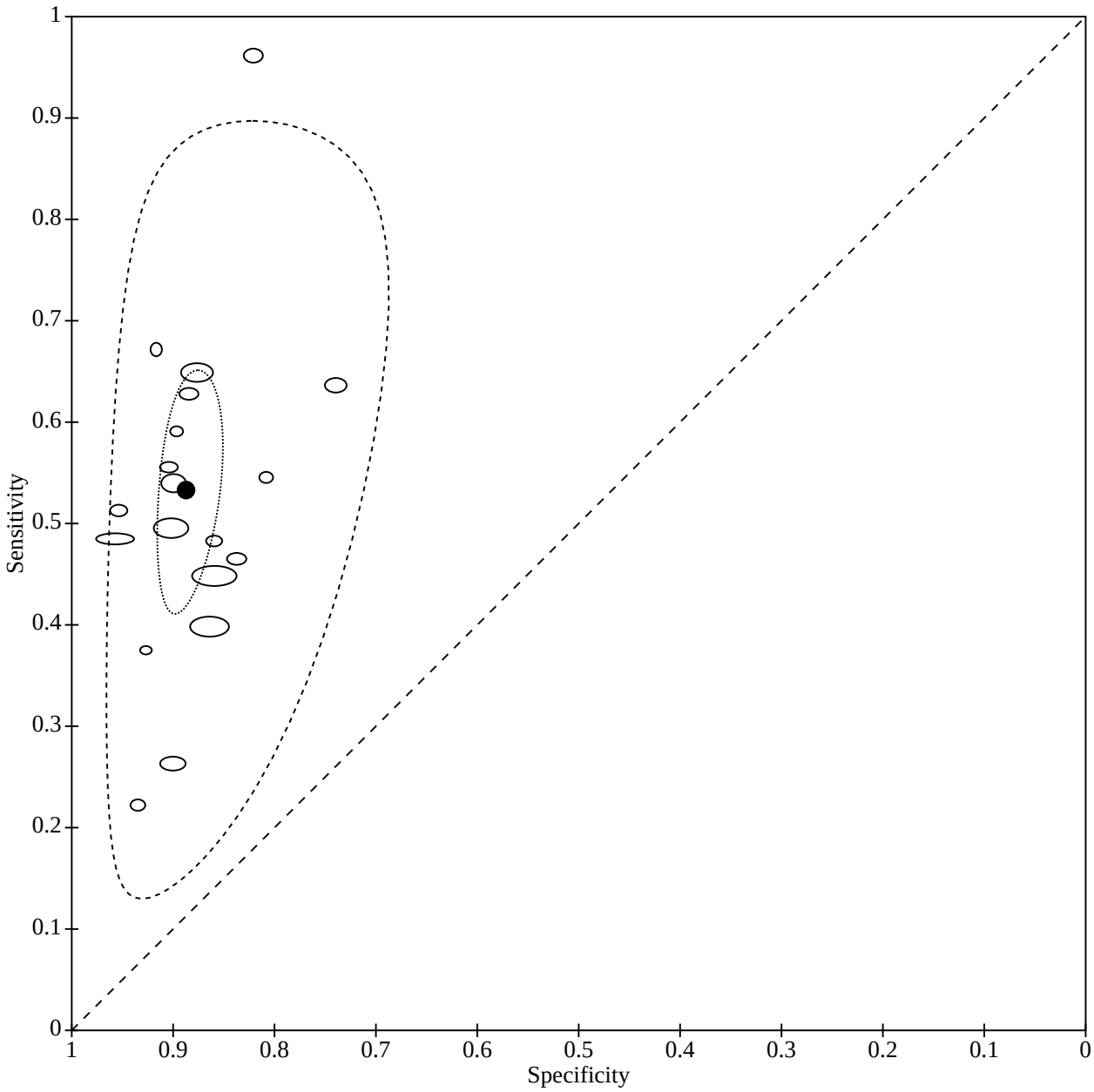


Figure 32. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting any anxiety disorder (AAD) for the individual setting subgroups Setting: I = non-clinical, II = clinical – across conditions, III = clinical – specific conditions Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

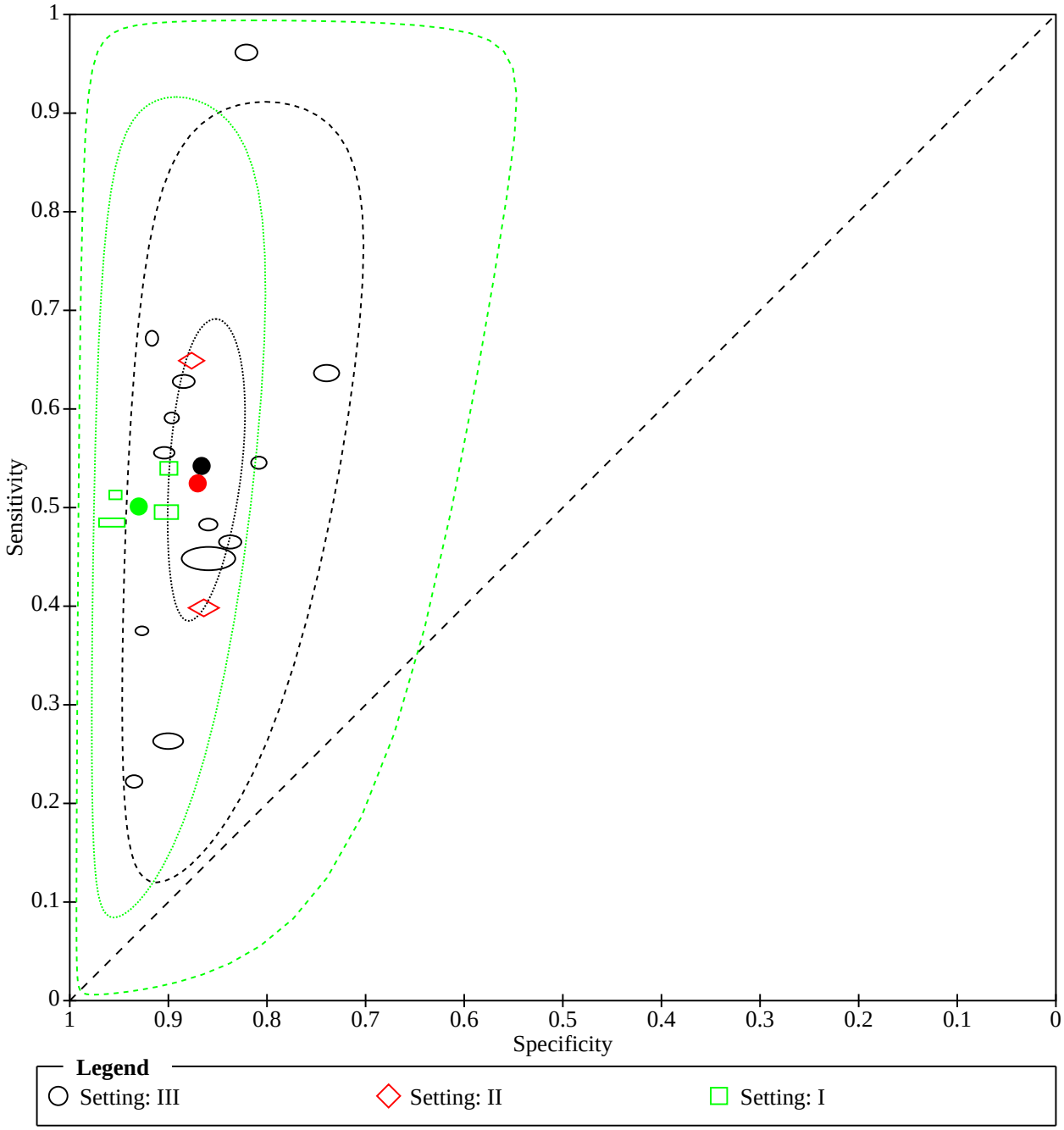


Figure 33. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting any anxiety disorder (AAD) for the reference standard subgroups Open circles, rectangles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle, rectangle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

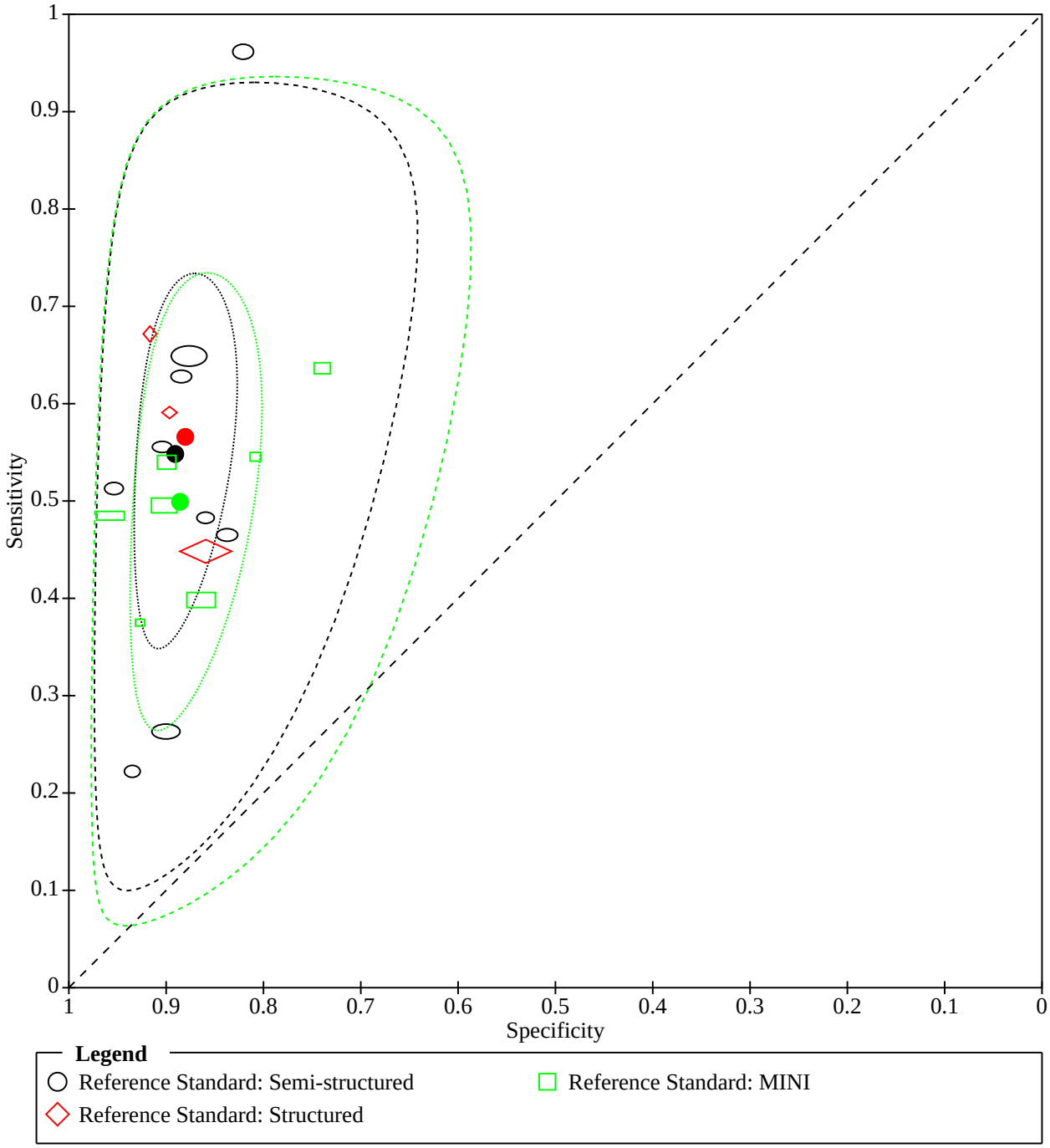
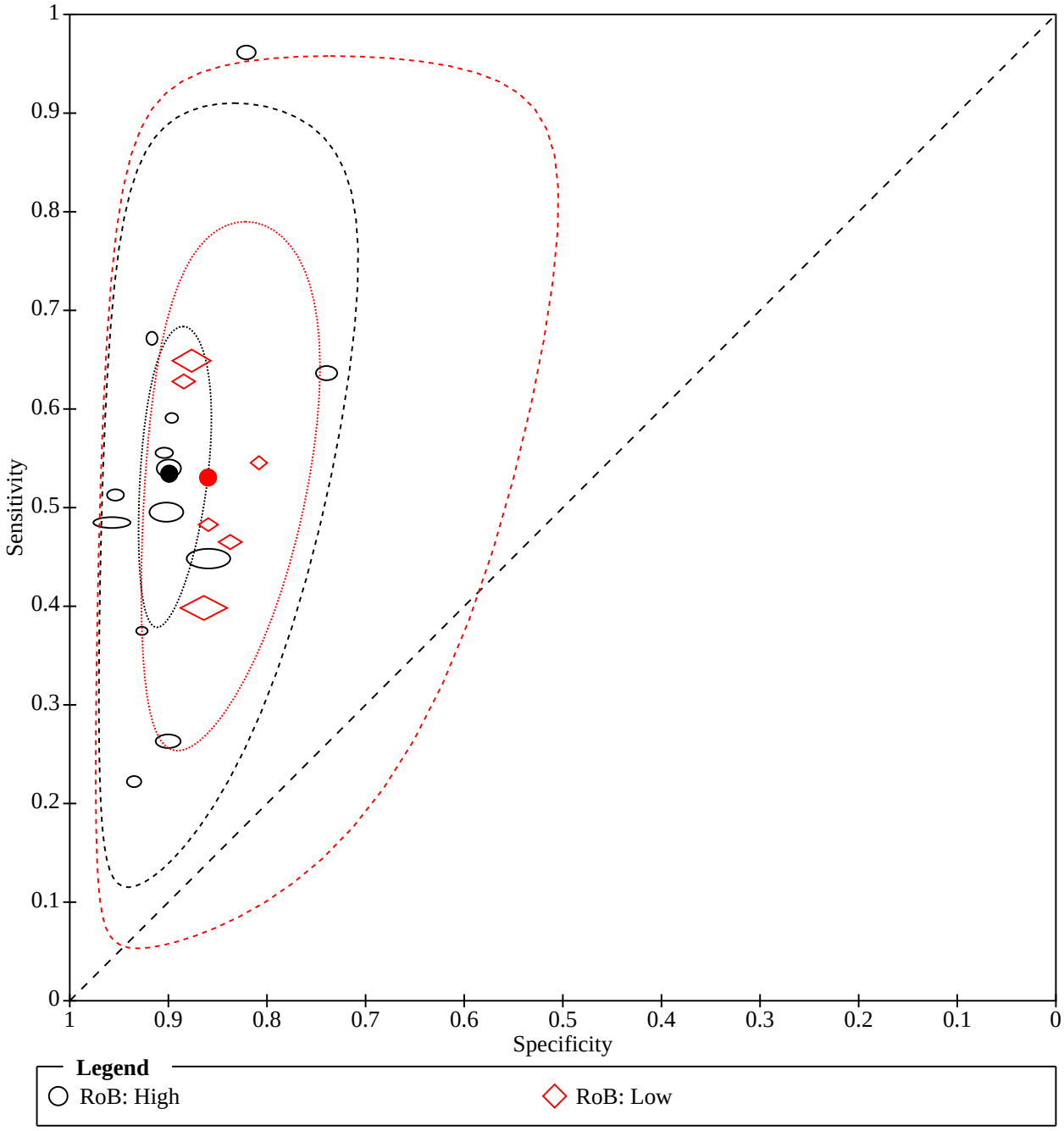


Figure 34. Summary receiver operating characteristic (SROC) plot of GAD-2 for detecting any anxiety disorder (AAD) for high and low risk of bias (RoB) subgroups Open circles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.



Appendix 9. SROC plots showing comparisons between the index tests and target conditions

SROC plots comparing the GAD-7 and GAD-2 questionnaires for generalised anxiety disorder and any anxiety disorder are shown in Figure 35 and Figure 36.

Figure 35. Summary receiver operating characteristic (SROC) plot of studies comparing the GAD-7 and the GAD-2 for detecting generalised anxiety disorder (GAD) Dashed straight lines connect the estimates for GAD-7 and GAD-2 for individual studies. Open circles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.

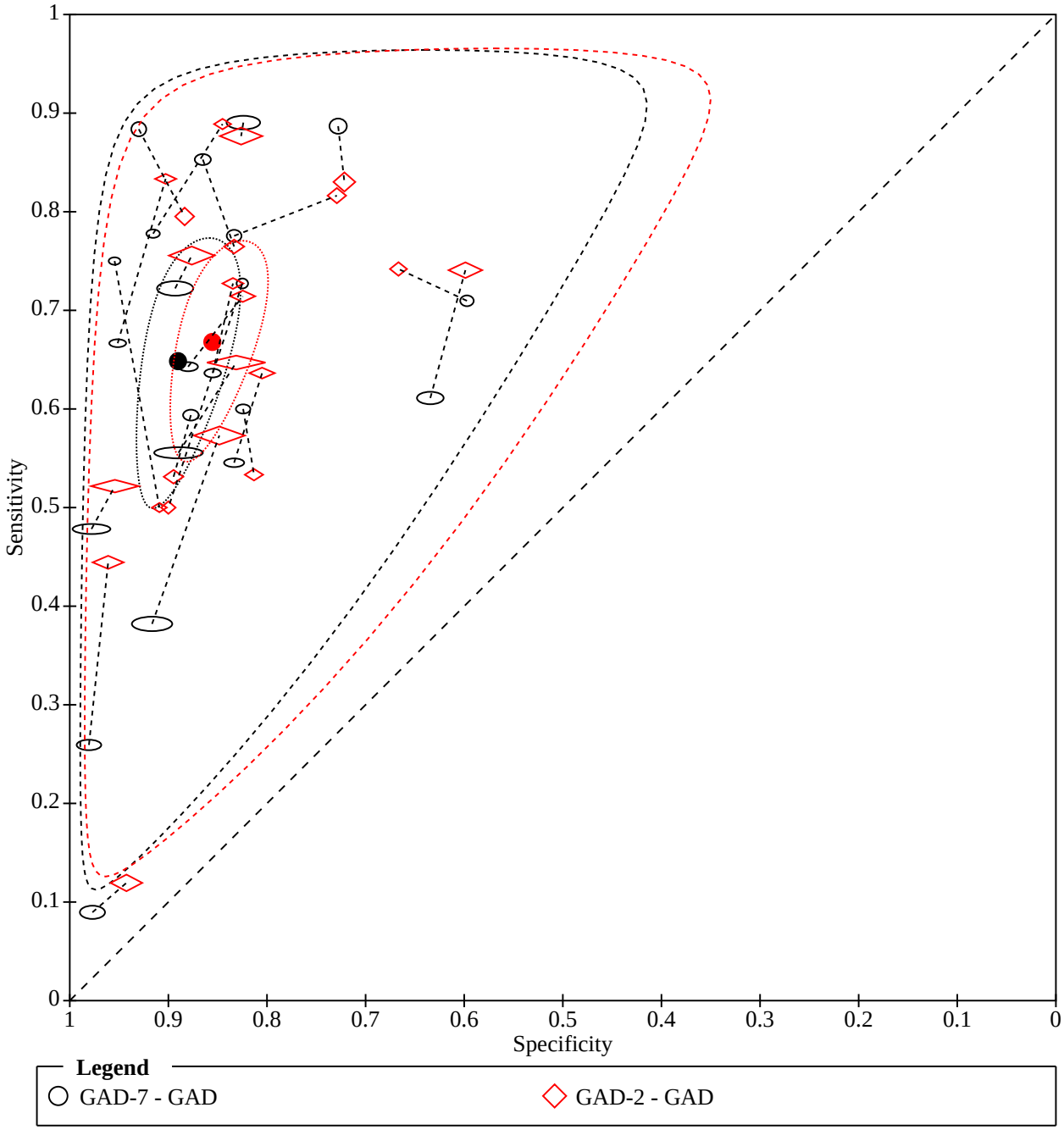
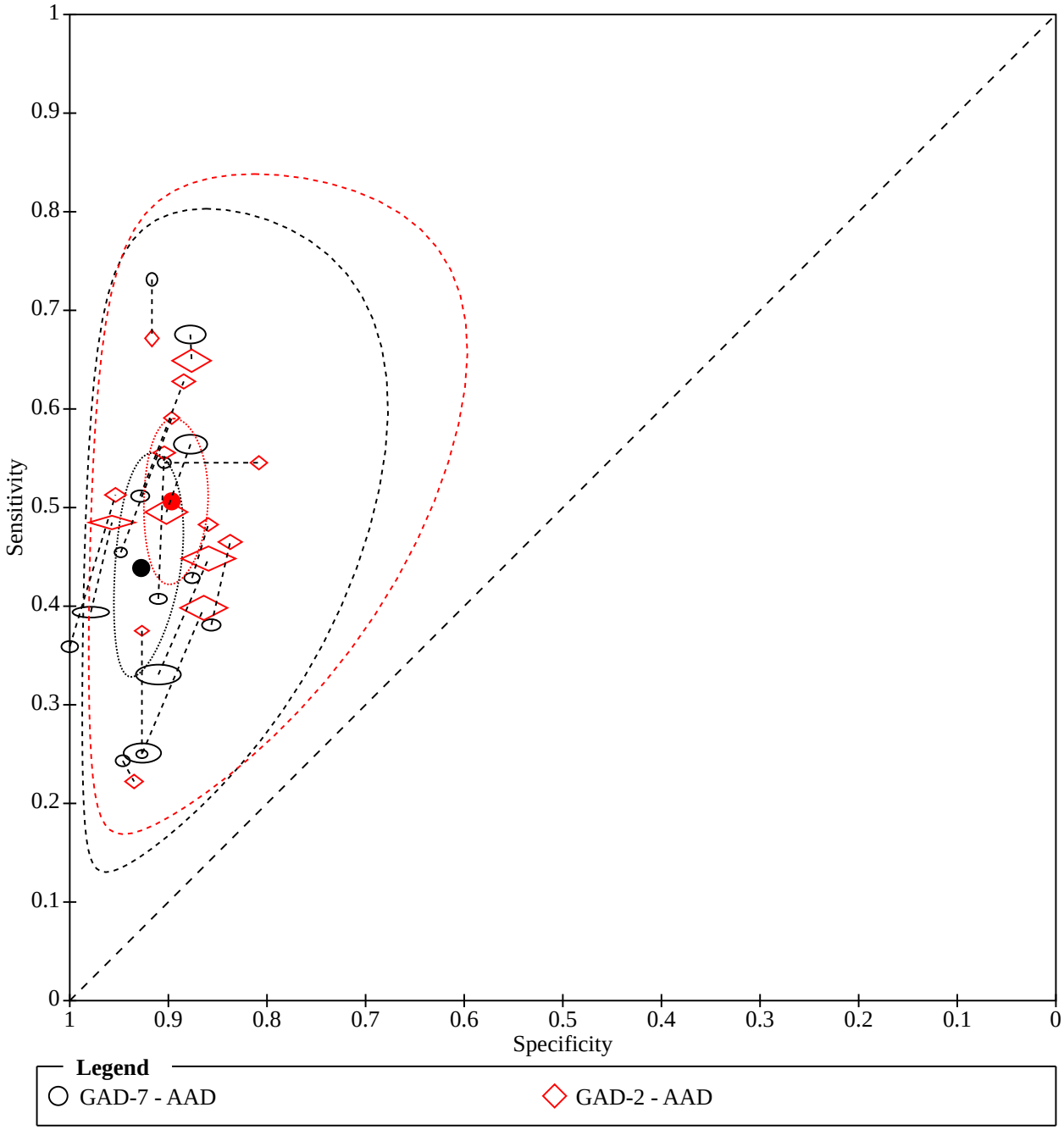


Figure 36. Summary receiver operating characteristic (SROC) plot of studies comparing the GAD-7 and the GAD-2 for detecting any anxiety disorder (AAD) Dashed straight lines connect the estimates for GAD-7 and GAD-2 for individual studies. Open circles and diamonds indicate the individual sensitivity and specificity for each study. The filled circles indicate the summary estimates. Circle and diamond sizes are scaled for sample size. Confidence (dotted) and prediction (dashed) regions are for 95%.



CONTRIBUTIONS OF AUTHORS

Antonius Schneider, Klaus Linde, Michaela Olm, Alexander Hapfelmeier, Gerta Rücker, and Bernd Löwe conceptualised the overall review project and obtained funding.

Zekeriya Aktürk, Michaela Olm, and Klaus Linde drafted selection and extraction forms and adapted scoring rules for the risk of bias assessment with QUADAS.

Sarah Dawson designed the search strategy and performed the searches.

Zekeriya Aktürk, Alexey Fomenko, Daniel Dümmler, Stefanie Eck, Michaela Olm, Jan Gehrmann, Victoria von Schrottenberg, Rahel Rehder, and Klaus Linde screened search hits, screened and selected studies, extracted data, and assessed risk of bias and applicability.

Zekeriya Aktürk performed the analyses using RevMan.

Alexander Hapfelmeier conducted the statistical analyses with R. Klaus Linde, Zekeriya Aktürk, and Alexey Fomenko drafted the manuscript.

All authors contributed to the data interpretation, finalisation of the manuscript, and commented on drafts.

DECLARATIONS OF INTEREST

Zekeriya Aktürk: none known.

Alexander Hapfelmeier: none known.

Alexey Fomenko: none known.

Daniel Dümmler: none known.

Stefanie Eck: none known.

Michaela Olm: none known.

Jan Gehrmann: none known.

Victoria von Schrottenberg: none known.

Rahel Rehder: none known.

Sarah Dawson: none known.

Bernd Löwe: collaborative development and validation of GAD-7 and GAD-2 (intellectual property). Both instruments are in the public domain and there is no financial conflict of interest. The GAD-7 scale and the GAD-2 scale were collaboratively developed and validated by Robert L. Spitzer, Kurt Kroenke, Janet B. Williams, and Bernd Löwe in the USA in 2003 and 2004. The development of the GAD-7 scale was underwritten by an unrestricted educational grant from Pfizer Inc. (New York, NY) to the first three authors. Bernd Löwe did not receive any grant support, and he did not receive an honorarium. BL was not involved in the extraction and assessment of studies using these instruments. See publications:

1. Spitzer RL, Kroenke K, Williams JB, Löwe B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med 166, 1092-1097. The development of the GAD-7 scale was underwritten by an unrestricted educational grant from Pfizer Inc. (New York, NY) to the first three authors. Bernd Löwe did not receive any grant support, and he did not receive an honorarium.
2. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann. Intern. Med 146, 317-325. The development of the GAD-7 scale was underwritten by an unrestricted educational grant from Pfizer Inc. (New York, NY) to the first three authors. Bernd Löwe did not receive any grant support, and he did not receive an honorarium.
3. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, Herzberg PY. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med. Care 46, 266-274. No external funding, no honorarium.
4. Kroenke K, Spitzer RL, Williams JB, Löwe B. (2010). The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen. Hosp. Psychiatry 32, 345-359. The development of the GAD-7 scale was underwritten by an unrestricted educational grant from Pfizer Inc. (New York, NY) to the first three authors. Bernd Löwe did not receive any grant support, and he did not receive an honorarium.

Gerta Rücker: none known.

Antonius Schneider: none known.

Klaus Linde: none known.

SOURCES OF SUPPORT

Internal sources

- None, Other

  None

External sources

- German Federal Ministry of Education and Research Grant 01KG2105, Germany

  The study project with the DIAQANDI acronym was approved on 2 February 2021. Funding period: 1 July 2021 to 31 December 2023

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

Objectives

- In the protocol, we planned to investigate the diagnostic test accuracy (DTA) of the Generalized Anxiety Disorder 7-item (GAD-7) scale and the Generalized Anxiety Disorder 2-item (GAD-2) scale for specified anxiety disorders (e.g. panic disorder or social phobia), other than generalised anxiety disorder (GAD) and any anxiety disorder (AAD), if at least three studies presented data. We decided not to undertake these investigations in the extraction phase of the review because (1) such data were rarely reported; and (2) the complexity of the review was already high.
- We decided to investigate whether the DTA of GAD-7 and GAD-2 differs (because we considered this relevant for the feasibility of screening).

Clinical pathway and selection criteria

- We had originally planned to also investigate the use of GAD-7 and GAD-2 “as a diagnostic aid for people with suspected but uncertain anxiety disorder” (pathway section of the protocol). This led to the inclusion of studies that were clinically and statistically different from those that fit a screening pathway, adding further complexity to the already complex review. This approach was criticised during the peer review process. Therefore, we decided to focus on the screening pathway and to exclude studies investigating the use of the scales as diagnostic aids.
- Therefore, we excluded studies involving people seeking help in mental health settings or patients who were recruited specifically due to mental health symptoms in other settings, contrary to what was planned in the protocol.

Search methods for identification of studies

- In the protocol, we planned to search trial registries and grey literature sources. As preliminary searches in trial registries yielded a very large number of clearly irrelevant hits (mainly due to the frequent use of GAD-7 – and less frequently of GAD-2 – as outcome measures in clinical trials), we refrained from doing systematic searches in trial registers. We did not search any grey literature sources directly. However, some grey literature items have been captured via Embase (conference abstracts) and PsycINFO (theses).

Assessment of methodological quality

- In the protocol, we stated that, "If one or more signalling questions are rated as 'unclear', this will usually lead to an overall assessment of 'unclear' risk of bias; however, if review authors give explicit reasons, they can rate risk of bias as low (e.g. if an 'unclear' for a single signalling question is considered unlikely to imply a relevant risk of bias) or 'high' (more than one 'unclear' rating for an overall dubious study).” During the extraction process, we revised our protocol decision. Instead, we rated a domain as ‘unclear risk of bias’ even if only one signalling question was answered ‘unclear’.
- Rating guidelines QUADAS-2, domain I: in the protocol, we had planned to merge signalling questions 1 and 2 due to conceptual considerations. In the course of the assessment, however, this approach proved unsatisfactory, so we decided to evaluate the questions separately as usual in QUADAS-2. In this process, we also added details for the rating of signalling question 1.
- Rating guidelines QUADAS-2, applicability question 1: in the protocol, we had planned to answer this question for a study with "yes" only "if the participants are irrelevant for our questions". Due to the changes described above in "Clinical pathway and selection criteria" and following the rating guidelines of a large individual patient data meta-analysis on screening tools for depression (Levis 2019), we used the more strict approach described for the final assessments.
- Use of the QUADAS-C: after the completion of the protocol, we became aware of the newly developed QUADAS-C tool. As we also aimed to compare the DTA of GAD-7 and GAD-2, we decided to also use this tool for studies allowing such a comparison.

Statistical analysis

- As the potential differences between the settings are of great interest for the implementation of the screening, we decided to prioritise this subgrouping variable before starting the analyses.

NOTES

The overall project, in which the Cochrane reviews described above are embedded, comprises an additional sub-project with the aim to perform a network meta-analysis to compare the DTA of the five questionnaires, making efficient use of all available data. This sub-project will be based on the data collected in the four Cochrane reviews. However, the findings of this additional project will not be addressed within the Cochrane reviews but published separately.

INDEX TERMS

Medical Subject Headings (MeSH)

*Anxiety Disorders [diagnosis] [epidemiology]; Bias; Cross-Sectional Studies; Generalized Anxiety Disorder; *Psychiatric Status Rating Scales [standards]; Sensitivity and Specificity; Surveys and Questionnaires [standards]

MeSH check words

Adult; Humans