Cochrane Library

Cochrane Database of Systematic Reviews

The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults (Protocol)

Linde K, Olm M, Teusen C, Akturk Z, von Schrottenberg V, Hapfelmeier A, Dawson S, Rücker G, Löwe B, Schneider A

Linde K, Olm M, Teusen C, Akturk Z, von Schrottenberg V, Hapfelmeier A, Dawson S, Rücker G, Löwe B, Schneider A.
The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults (Protocol).
Cochrane Database of Systematic Reviews 2022, Issue 9. Art. No.: CD015292.
DOI: 10.1002/14651858.CD015292.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Diagnostic Test Accuracy Protocol]

The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults

Klaus Linde[1], Michaela Olm[1], Clara Teusen[1], Zekeriya Akturk[1], Victoria von Schrottenberg[1], Alexander Hapfelmeier[1,2], Sarah Dawson[3,4], Gerta Rücker[5], Bernd Löwe[6], Antonius Schneider[1]

[1]Institute of General Practice and Health Services Research, School of Medicine, Technical University of Munich, Munich, Germany. [2]Institute for AI and Informatics in Medicine, School of Medicine, Technical University of Munich, Munich, Germany. [3]Cochrane Common Mental Disorders, University of York, York, UK. [4]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [5]Institute of Medical Biometry and Statistics, University Medical Center – University of Freiburg, Freiburg, Germany. [6]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Contact: Klaus Linde, klaus.linde@mri.tum.de.

Editorial group: Cochrane Common Mental Disorders Group.
Publication status and date: New, published in Issue 9, 2022.

Citation: Linde K, Olm M, Teusen C, Akturk Z, von Schrottenberg V, Hapfelmeier A, Dawson S, Rücker G, Löwe B, Schneider A. The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults (Protocol). *Cochrane Database of Systematic Reviews* 2022, Issue 9. Art. No.: CD015292. DOI: 10.1002/14651858.CD015292.

ABSTRACT

Objectives

This is a protocol for a Cochrane Review (diagnostic). The objectives are as follows:

To determine the diagnostic test accuracy (DTA) of the following five widely used self-reporting questionnaires for detecting anxiety disorders against standardized or structured clinical interviews as the reference standard among adults in any setting. This is a generic protocol for four parallel Cochrane Reviews of DTA.

- Review 1. Generalized Anxiety Disorder 7-item (GAD-7) Scale and its short version Generalized Anxiety Disorder 2-item (GAD-2) Scale
- Review 2. Hospital Anxiety and Depression Scale (HADS)
- Review 3. Beck Anxiety Inventory (BAI)
- Review 4. State-Trait Anxiety Inventory (STAI)

In the primary analyses, DTA will be determined separately for the detection of the two target conditions 'any anxiety disorder' and 'generalized anxiety disorder'. If at least three studies present data on other specified anxiety disorders (e.g. panic disorder or social phobia), DTA will be determined also for these conditions.

Secondary objectives

To investigate whether the DTA varies depending on the prevalence of anxiety disorders in the study population, setting, the reference standard and the risk of bias and to understand how diagnostic performance changes with test threshold.

BACKGROUND

This is a generic protocol for four parallel Cochrane Reviews of diagnostic test accuracy (DTA).

Large epidemiological surveys indicate that anxiety disorders are the most prevalent mental disorders worldwide (Greenberg 1999; Kessler 2001; Leon 1995; Weissman 1988). One large high-quality meta-analysis of population-based studies estimated the 12-month prevalence of anxiety disorders as 7%, and the lifetime prevalence as 13% (Steel 2014). Women are more likely to be affected than men. Anxiety disorders are classified in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) or Fifth Edition (DSM-5) (American Psychiatric Association 2000; American Psychiatric Association 2013), as well as the International Statistical Classification of Diseases and Related Health Conditions (ICD-10, WHO 1993; ICD-11, WHO 2019). Anxiety disorders such as generalized anxiety disorder (GAD) and other forms such as panic disorder and phobias are specified in structured chapters. Besides their high prevalence, anxiety disorders are often chronic and can severely limit the ability to carry out activities of daily live (Hoffman 2008; Schonfeld 1997). Beyond that, anxiety disorder is a risk factor for hospitalization and increased mortality in chronic illness (Gudmundsson 2005; Schneider 2008; Vongmany 2016). Measured by years of life lived with disability (YLDs), anxiety disorders were the sixth leading cause of all disability (e.g. after low back pain or major depressive disorder) in high-income as well as low- to middle-income countries (Baxter 2014).

The vast majority of anxiety disorders remain undetected and untreated by healthcare systems, even in economically advanced countries (Craske 2017). One reason might be that symptoms of anxiety overlap with those of other mental diseases (Olariu 2015; Sherbourne 1996; Toft 2005). Moreover, symptoms of anxiety disorders often mimic those of somatic (physical) diseases (e.g. breathlessness is a symptom of both panic disorder and several somatic diseases such as heart failure, asthma or chronic obstructive pulmonary disease). Therefore, anxiety symptoms are often hidden by somatic diagnoses, which might in turn lead to under-recognition (Olariu 2015; Walters 2012; Wittchen 2001), and diagnostic mismatch between patients and doctors (Schneider 2013). People with anxiety disorders are commonly seen by primary care providers (Lieb 2005; Wittchen 2002). They exhibit increased healthcare use, for example high referral rates (Schneider 2011), and long duration of sick certificates (Schneider 2017a). Although increasing attention has been paid to anxiety, it still lags far behind depression in terms of research in case finding and improvement of diagnostic decision-making (Kroenke 2007). General recommendations for screening for anxiety in larger populations based on one systematic review of evidence for effectiveness do not exist apart from the recent recommendation of the US Women's Preventive Services Initiative (WPSI) for adolescent and adult women (Gregory 2020). The WPSI recommends that anxiety disorders be screened with a standardized brief questionnaire in order to make timely diagnoses and initiate early treatment. But even in this specific case relatively little is known on whether systematic screening in larger groups of the general population provides more benefit than harm (Nelson 2020).

Structured self-report questionnaires are sometimes used in non-mental healthcare settings to screen for anxiety or as a diagnostic aid complementing the physician's judgement (Kroenke 2007). In addition, they are used for monitoring the severity of anxiety symptoms during the course of treatment or as measurement tools in epidemiological and clinical research. There is a plethora of anxiety 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 few languages. However, a number of instruments that have undergone intense testing are available in several languages and widely used. These include the Generalized Anxiety Disorder 7-item scale (GAD-7; Löwe 2008; Spitzer 2006); its short version Generalized Anxiety Disorder 2-item scale (GAD-2; Kroenke 2007), the Hospital Anxiety and Depression Scale (HADS; Zigmond 1983); the Beck Anxiety Inventory (BAI; Beck 1988), and the State-Trait Anxiety Inventory (STAI; Spielberger 1970).

Systematic reviews including a meta-analysis of DTA for anxiety disorders across a variety of different populations were published in 2016 for the GAD-7 and the GAD-2 (Plummer 2016), and in 2010 for the HADS (Brennan 2010), but a relevant number of new studies have been published since the searches for these reviews were completed. Systematic reviews across conditions are not available for the BAI and the STAI. Small subsets of the available DTA studies on one or more of the five questionnaires are included in condition-specific or setting-specific reviews (e.g. Benjamin 2011; Creighton 2018; Litster 2016; Mele 2018; Sinesi 2019). However, while systematic reviews of the evidence for screening tools for anxiety in highly restricted populations (e.g. people with epilepsy) have some merit, the small number of studies included strongly limits the feasibility or conclusiveness of any meta-analysis. Therefore, there is a need for systematic reviews comprehensively summarizing the available evidence on the DTA of the most widely used self-report questionnaires. This is a generic protocol for four individual reviews.

Clinical pathway

While our reviews will assess the 'diagnostic test accuracy' of the selected questionnaires, it must be emphasized that these questionnaires alone are not sufficient to make a diagnosis of a specific anxiety disorder or guiding treatment decisions. They must always be complemented by the assessment of a competent clinician. We see four main scenarios for the use of our index tests as diagnostic or screening tools: 1. as a screening tool in epidemiological studies (in the general population or in more selected populations); 2. as a screening tool in unselected primary care patients or other clinical settings; 3. as a screening tool in people with a specific disease (e.g. cancer, epilepsy, etc.) or specific risk factors associated with an increased prevalence of anxiety; and 4. as a diagnostic aid for people with suspected but uncertain anxiety disorder. Scenario 1 is not a clinical scenario, but is relevant to the estimation of the prevalence of anxiety in a population and to what extent the screening questionnaires can provide a reliable estimate of the true prevalence. The clinical scenarios 2 and 3 address typical screening situations: screening is a routine activity directed at people not identified or seeking assistance for anxiety disorder. A positive screening result (suspected anxiety disorder) will then be followed by a detailed clinical assessment to make a diagnosis (anxiety disorder yes/no), while such an assessment is usually not indicated if the screening result is negative. This process is likely to be very similar, regardless of the specific settings or population. In scenario 4, after an initial assessment, the doctor has a specific need for more detailed clarification regarding the presence or absence of an anxiety disorder. The questionnaire is used here *individually* to confirm or weaken the suspicion in a specific patient. Depending on whether the doctor is sufficiently sure after using the questionnaire or not, a more extensive clinical assessment may be carried out. The use of questionnaires for monitoring severity and the course of disease is beyond the interest of our review.

In principle, any anxiety disorders listed under the Chapters F.40 (phobic anxiety disorders, agoraphobia, social phobias, other or unspecified phobic anxiety disorders) and F.41 (panic disorder, GAD, mixed anxiety and depressive disorder, and other mixed, specified or unspecified anxiety disorders) in the ICD revision 10 or 11 (WHO 1993; WHO 2019), or in related chapters of the DSM revisions (IV, IV-Text Revision or 5; American Psychiatric Association 2000; American Psychiatric Association 2013) are of interest in the context of the planned reviews. However, the *primary* target conditions being diagnosed in our four reviews will be 'GAD' and 'any anxiety disorder'. There are several reasons for this decision. 1. For diagnosing specific phobias or panic disorders, specifically targeted questionnaires are likely to have better DTA than general anxiety questionnaires (e.g. Connor 2000; Wuyek 2011). 2. For a general screening in non-psychiatric settings, the overarching category 'any anxiety disorder' is clinically important and, accordingly, it is the most frequently investigated and reported category among the available diagnostic studies (Brennan 2010; Litster 2016; Mele 2018; Plummer 2016; Sinesi 2019). GAD is both clinically relevant and, according to our preliminary screening of potentially eligible articles, the most frequently analysed specific anxiety disorder. 3. According to our screening of the available studies, diagnostic findings for other specific diagnoses are rarely ever reported in a format allowing meta-analysis unless the study addresses the diagnosis exclusively. Nevertheless, we will also summarize the findings on important other specific anxiety disorders (panic disorder, social phobia, mixed anxiety and depressive disorder) for which the use of generic anxiety screening questionnaires seems adequate as *secondary* target conditions if sufficient data are available in at least three studies on a questionnaire.

OBJECTIVES

To determine the diagnostic test accuracy (DTA) of the following five widely used self-reporting questionnaires for detecting anxiety disorders against standardized or structured clinical interviews as the reference standard among adults in any setting. This is a generic protocol for four parallel Cochrane Reviews of DTA.

- Review 1. Generalized Anxiety Disorder 7-item (GAD-7) Scale and its short version Generalized Anxiety Disorder 2-item (GAD-2) Scale
- Review 2. Hospital Anxiety and Depression Scale (HADS)
- Review 3. Beck Anxiety Inventory (BAI)
- Review 4. State-Trait Anxiety Inventory (STAI)

In the primary analyses, DTA will be determined separately for the detection of the two target conditions 'any anxiety disorder' and 'generalized anxiety disorder'. If at least three studies present data on other specified anxiety disorders (e.g. panic disorder or social phobia), DTA will be determined also for these conditions.

Secondary objectives

To investigate whether the DTA varies depending on the prevalence of anxiety disorders in the study population, setting, the reference standard and the risk of bias and to understand how diagnostic performance changes with test threshold.

METHODS

Criteria for considering studies for this review

Types of studies

For all reviews, we will include cross-sectional studies comparing findings from the specific index test with the diagnostic status according to a reference standard that allow the generation of 2 × 2 tables (number of true positive, false positive, false negative and true negative index tests results). We will exclude case-control studies as estimates of DTA derived from case-control-type studies might not apply to the diagnostic process in real situations due to the selection involved.

Participants

To be included, studies must have recruited adults (aged 18 years or older). There will be no restrictions regarding setting or comorbidity; participants of primary studies can have been sampled from the general population, from primary care or other settings; participants may have experienced a specific condition (e.g. epilepsy), from various diseases or may be without any prediagnosed mental or physical health condition. We will exclude studies in children as these have to use specific, age-adapted self-report questionnaires (Lazor 2017).

Index tests

To be included, studies must have administered at least one of the following five widely used self-reporting questionnaires concerning the diagnosis of anxiety disorders: GAD-7, GAD-2, HADS Anxiety (HADS-A), BAI or STAI.

The GAD-7 scale was designed originally as a screening tool for GAD but is also used for screening for anxiety disorders in a broader manner (Kroenke 2007; Löwe 2008; Spitzer 2006). It has seven items with four answer options each (as all questionnaires addressed in this protocol). The item scores are summed up resulting in a score ranging from 0 to 21, with a higher score indicating more anxiety. The cut-off point (anxiety/no anxiety) recommended by the scale authors is 10 or greater (Spitzer 2006).

The GAD-2 scale is the short version of the GAD-7. It has two items, a score range from 0 to 6, with a higher score indicating more anxiety, and the recommend cut-off is 3 or greater (Kroenke 2007).

The HADS is a widely used questionnaire to measure symptoms of anxiety (HADS-A) and depression (HADS Depression or HADS-D) (Zigmond 1983). In our review, we will exclusively focus on the HADS-A. The scale has seven items and a score range from 0 to 21, with a higher score indicating more anxiety. A widely used cut-off point is 8 or greater (Bjelland 2002).

The BAI has 21 items and a score range from 0 to 63, with a higher score indicating more anxiety. A widely used cut-off point also recommended by the scale developers is 16 or greater (Beck 1988; Beck 1993). Compared to other questionnaires, the BAI has a comparably strong focus on physical symptoms of anxiety.

The 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, with a higher score indicating more anxiety. Recommendations for cut-offs are rare and less clear than for the other questionnaires. For the State anxiety subscale (STAI-S) a cut-off of 40 or greater is commonly used (Emons 2019). For the Trait anxiety subscale (STAI-T), a value of 44 or greater has been suggested (Ercan 2015).

Target conditions

Studies must have addressed at least one of the following three diagnostic categories:

- presence or absence of any anxiety disorder;
- presence or absence of GAD;
- presence or absence of social phobia, panic disorder, or mixed anxiety and depressive disorder.

Reference standards

Validated structured and semi-structured interviews are accepted as state-of-the-art procedures in epidemiological and diagnostic studies in mental disorders (Levis 2019; Suppiger 2009). They ensure sufficient quality and a certain reproducibility of diagnosis. Thus, the diagnosis of anxiety of all participants in primary studies must have been made or ruled out using a validated structured or semi-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); DIA-X (diagnostic expert system for mental disorders) (Wittchen 1997); SADS (Schedule for Affective Disorders and Schizophrenia) (Endicott 1978); and DIPS (Margraf 1994), or Mini-DIPS (Margraf 2013) (German: "Diagnostisches Interview bei psychischen Störungen"). We will exclude studies in which the reference standard is informally based on a checklist based on ICD or DSM or a clinical diagnosis without operationalization or only another questionnaire.

There is evidence that different interviews are not fully equivalent in the case of depression (Wu 2021). As the structured and semi-structured interviews used for anxiety disorders are very similar, it is to be expected that this finding might also apply here. Therefore, we will include the type of reference standard as a covariate into our analyses of heterogeneity.

Search methods for identification of studies

Electronic searches

In order to identify potentially eligible studies, we will search the following databases using relevant subject headings (controlled vocabularies), text-words and search syntax, appropriate to each resource (1990 onwards).

- MEDLINE (Ovid);
- Embase (Ovid) (search strategy Appendix 1);
- PubMed-not-MEDLINE subset (NLM);
- PsycINFO (Ovid);
- US National Institutes of Health Ongoing Trials Register ClinicalTrials.gov (www.clinicaltrials.gov);
- World Health Organization International Clinical Trials Registry Platform (trialsearch.who.int).

We will not apply any restrictions on language or publication status to the searches. However, we will apply a date limit (1990 onwards) due to the differences in older diagnostic classification schemes, together with concerns regarding the quality of earlier diagnostic studies (the DSM-IV classification was introduced in 1994, the ICD-10 in 1989).

Search structure

One of the main challenges in identifying the evidence for a diagnostic review in psychiatry 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 study retrieved by a search (DTA study or intervention study). In order to balance the sensitivity (recall) and specificity (precision) of the search, we will structure it around the following key concepts:

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

#2 Target Condition + Index Test(s) + DTA Filter

#3 (#1 OR #2)

Lead by a Cochrane information specialist (SD), we have developed an initial Embase search (Appendix 1), benchmarking the strategy against a set of known studies for each of the five self-report instruments (index tests): GAD-2; GAD-7; HADS; STAI, BAI. We have paid special attention to the search terms used for the reference standard and DTA 'filters'.

The information specialist will perform a separate search for each of the five index tests, with minor adaptations to the search terms for the target condition where appropriate (e.g. in the search for HADS), but applying the same reference standard and DTA 'filters' to all searches. We did consider running one search across all index tests, to minimize the screening burden, but our scoping searches have shown little duplication across the individual searches.

Searching other resources

Grey literature

We will search for conference abstracts and theses using a more focused set of search terms (which we may adapt after screening the results of the main search, for journal articles).

Conference abstracts (1990 onwards):

- Embase (Ovid);
- Web of Science Conference Proceedings Citation Index – Science (CPCI-S) (Clarivate).

Theses (1990 onwards):

- Electronic Theses Online Service (EThOS) – British Library (ethos.bl.uk/Home.do);
- DART – Europe e-theses Portal (dart-europe.eu/basic-search.php);

- Networked Digital Library of Theses and Dissertations (NDLTD) (search.ndltd.org/);
- Open Access Theses and Dissertations (OATD) (oatd.org);
- Proquest Dissertations & Theses Global (search.proquest.com/pqdtglobal/dissertations/).

Other

To help identify further published, unpublished or ongoing research, we will scan the reference lists of included studies and any relevant systematic reviews. We will check for any relevant retraction statements or errata of the included studies. We may contact original authors for clarification and further data if any of the study reports are unclear.

Data collection and analysis

Selection of studies

After searching, we will collate titles from the databases in EPPI-Reviewer Web software (EPPI-Reviewer). Two review authors will independently screen titles and abstracts of search hits to exclude clearly irrelevant papers. If the relevance remains unclear after title and abstract screening, we will obtain and screen full texts. In the second round, the review authors will independently decide to include or exclude a potentially eligible study based on the review of full texts using the criteria listed above. We will document disagreements, and resolve them by discussion. If consensus cannot be achieved between the original review authors, an additional review author will be asked to make a final decision. We will present the study selection process in a PRISMA flow diagram (McInnes 2018). We will document reasons for excluding studies in a characteristics of excluded studies table.

Following the approach used by Quinn and colleagues, we will contact the corresponding author if potentially relevant articles do not contain data that are required for the analyses (Quinn 2020). When the study authors do not submit a response or the relevant data are not available, we will not include the study and rate it as 'data not available for analyses'. Where studies have multiple publications, we will collate the reports of the same study so that each study, rather than each report, is the unit of interest for the review, and such studies have a single identifier with multiple references. In case of studies that are only presented in abstract form, we will contact the main author to enquire whether the full paper has been published.

Data extraction and management

Two review authors will independently extract primary study characteristics and results using a pretested form. In particular, we will extract diagnoses and main inclusion criteria for participants; age; gender; setting; details regarding the reference standard (type, person doing the assessment, timing): language of the index test; predefined cut-off points/methods to select presented cut-off points; country of origin; study design; number and type of study centres; numbers of participants who were included and analysed; the number and reasons for dropouts and withdrawals. We will summarize the details in a characteristic of included studies table. For all cut-off points available, we will extract diagnostic 2 × 2 tables (true positive, false positive, true negative and false negative index test results) from the publications, or, if not available, reconstruct them using information about relevant parameters (prevalence, sensitivity, specificity or predictive values). We will try to obtain missing information from authors.

We assume that in some publications of primary studies 2 × 2 diagnostic tables will only be available for a single cut-off (predefined or selected based on findings), for other studies for a small number of cut-offs and for a few studies for a large number of cut-offs. We will try to obtain additional data from study authors at least for the cut-off recommended by scale creators (with respect to the primary analysis) and for a core range of cut-offs for as many studies as possible.

Review authors who are codevelopers of the scales will not be involved in the extraction and assessment of studies using these instruments.

Assessment of methodological quality

At least two review authors will independently assess the risk of bias and the applicability of the included studies using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (Whiting 2011). Following the recommendations of the Cochrane Diagnostic Test Accuracy Working Group (Davenport 2014), we have tailored QUADAS-2 to our specific subject and we have developed coding guidelines for each item (Appendix 2). After reviewing the flow diagram of a primary study (or after drawing a flow diagram) we will make a risk of bias judgement ('high', 'low' or 'unclear') for the four domains of participant selection, index test, reference standard, and flow and timing based on signalling questions. If the answers to all signalling questions within a domain are judged as 'yes' (indicating low risk of bias for each question), then the domain will be judged at low risk of bias. If any signalling question is judged as 'no', indicating a high risk of bias, the domain will be scored at high risk of bias. 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). This will be followed by a judgement about concerns regarding applicability for the participant selection, index test and reference standard domains. We will consider overall risk of bias 'low' if at least three domains are scored 'low' and none 'high'. We will document disagreements between rating review authors and resolved them by discussion (if needed by a third review author).

To adapt QUADAS-2 to our expected set of primary studies we modified the signalling question 2 in the domain index test. As we expect many studies presenting data for several cut-offs, we will ask one of the following two questions (instead of the single question "if a threshold was used, was it prespecified?"): either "if a single threshold was used, was it prespecified?", or, in studies presenting more than one threshold, "is there a risk that the cut-offs presented/available have been chosen post-hoc due to optimal performance values".

We will present the QUADAS-2 results in graphical form, as well as narrative text.

Review authors who are codevelopers of the scales will not be involved in the extraction and assessment of studies using these instruments.

Statistical analysis and data synthesis

Our primary analyses of interest will be the accuracy of the five reviewed self-report questionnaires for each of the two target conditions 'any anxiety disorder' (present/not present) and 'GAD' (present/not present) at the cut-off closest to and within a 'core range' around the recommended (primary) cut-off (see also section index tests). This will allow us to include studies in the primary analyses for which data are not available for the recommended cut-off but for a cut-off within a clinically meaningful range ('core range'). For each scale, the core-range will be defined based on available reviews (Bjelland 2002; Julian 2011; Plummer 2016), screening of already identified potentially eligible primary studies and scale characteristics:

- GAD-7: primary cut-off 10 or greater; core range 7 to 13;
- GAD-2: primary cut-off 3 or greater; core range 2 to 4;
- HADS: primary cut-off 8 or greater; core range 5 to 11;
- BAI: primary cut-off 16 or greater; core range 12 to 20;
- STAI-S: primary cut-off 40 or greater; core range 35 to 45;
- STAI-T: primary cut-off 44 or greater; core range 38 to 50.

The definition of core ranges can be changed before running the statistical analysis if relevant new information becomes available (e.g. empirical data from other reviews or original studies published after completion of this protocol).

In addition to primary analyses, we will investigate the DTA for each cut-off for which data will be available for at least three studies in secondary analyses for the two primary target conditions as well as for other anxiety disorders.

In all of these analyses (primary and secondary), we will calculate study-specific pairs of sensitivity and specificity with 95% confidence intervals using 2 × 2 diagnostic tables and present them in joint forest plots separately for the two target conditions ('any anxiety disorder' and 'GAD') for each cut-off value.

For the primary analyses, we will use scatter-plots of study-specific estimates of sensitivity and 1 – specificity with 95% confidence intervals in addition to display data in the Receiver Operating Characteristic (ROC) space, that is to produce summary ROC plots, separately for the two target conditions and for each of the five instruments. We will use the bivariate model to obtain summary estimates of sensitivity and specificity with 95% confidence regions for each test and for each cut-off (Chu 2006; Reitsma 2005).

For making better use of all available information on all cut-offs and for investigating optimal cut-offs more efficiently in another secondary analysis, we will use the multiple thresholds model which is implemented in the R package diagmeta (Rücker 2020). This is a multilevel random-effects model which creates a link between the range of thresholds and the respective pairs of sensitivity and specificity and thus allows identifying thresholds at which the test is likely to perform best (Schneider 2017b; Steinhauser 2016). At the meta-analytic level, the model fits data of both groups (with or without the target condition, as determined by the reference test) and all available thresholds over all studies. Based on a log-logistic model, it provides estimates of the two cumulative distribution functions (cdfs) for the two groups across all studies, accounting for the between-study heterogeneity and correlation between groups. Then we will extract the summary sensitivity and specificity values at every threshold, followed by the creation of a multiple thresholds summary ROC (mtsROC) curve (preserving the threshold information).

We will conduct analyses using Review Manager 5 (Review Manager 2020) and 'R' (R Core Team 2021). We will implement the bivariate model and other analyses in R as recommended by the *Cochrane Handbook Systematic Reviews of Diagnostic Test Accuracy* (Macaskill 2021). We will summarize the main findings in a summary of findings table.

Investigations of heterogeneity

As we will include a diverse set of studies, statistical heterogeneity of our findings is, to some extent, expected. Predefined key covariates or subgrouping variables of interest are the prevalence of anxiety disorders in the study population (metric/dichotomized by median), clinical setting (categorical; exact way of categorization to be defined before starting the analysis), reference standards (categorical; different types of interviews) and risk of bias (low versus unclear/high risk of bias). Additional subgrouping variables will be considered after data extraction, but before running any analyses. We will perform systematic analyses of heterogeneity in regard to the 'primary analyses'. In a first step, we will visually inspect forest plots of sensitivity and specificity and summary ROC plots of the primary analysis (e.g. for outliers and using symbols or colours to visualize the relation to the aforementioned covariates). We will use 95% prediction intervals of summary point estimates to quantify heterogeneity. If possible, we will add study-level key covariates as fixed effects to the bivariate model in order to explain potential heterogeneity in a second step. We will conduct the analyses in concordance with general recommendations from the *Cochrane Handbook Systematic Reviews of Diagnostic Test Accuracy* (Macaskill 2021).

Sensitivity analyses

We have no prespecified sensitivity analyses.

Assessment of reporting bias

We will not conduct tests concerning reporting bias as there exists discussions about the most robust approach (Wilson 2015), and uncertainty how to use funnel plots (van Enst 2014), which have the potential to cause misleading results (Deeks 2005; Leeflang 2008).

We will conduct the reviews according to this published protocol, and report any deviations from it in the 'Differences between protocol and review' section.

ACKNOWLEDGEMENTS

We thank: the Cochrane Common Mental Disorders Group, especially Rachel Churchill, Jessica Hendon and Nick Meader for their help in the various steps of the project; the National Institute for Health and Care Research (NIHR) Complex Review Support Unit, especially Alex Sutton, Terry Quinn, Nicola Cooper and Enzo Cerullo for their comments and feedback; our advisory patient and provider team Kerstin Schaeffer, Markus Hendel, Frederike Kneissl, Tom Brandhuber and Jan Gehrmann.

Cochrane Common Mental Disorders (CCMD) and Cochrane Diagnostic Test Accuracy Editorial Team (DTA ET) supported the authors in the development of this protocol.

The following people conducted the editorial process for this article.

- Sign-off Editor, CCMD (final joint editorial decision with DTA ET): Nick Meader, Newcastle University.
- Contact Editor, Cochrane DTA ET (provided comments and recommended an editorial decision): Mia Schmidt-Hansen, Cardiff University.
- Managing Editor, CCMD (selected peer reviewers, collated peer-reviewer comments, provided editorial guidance to authors, edited the article): Jessica Hendon, Centre for Reviews and Dissemination, University of York.
- Managing Editor, Cochrane DTA ET (selected peer reviewers, collated peer-reviewer comments): Sophie Beese, University of Birmingham.
- Peer-reviewers CCMD (provided comments and recommended an editorial decision): Mark Haddad, School of Health Sciences, City, University of London (content review); Karen Morley (consumer review).
- Peer-reviewers Cochrane DTA ET (provided comments and recommended an editorial decision): Julie Glanville (Information Specialist – methods review); Mark Simmonds, University of York (Statistician – methods review).

The authors and the CCMD Editorial Team are grateful to the peer reviewers for their time and comments. They would also like to thank Cochrane Copy Edit Support for the team's help.

Cochrane Group funding acknowledgement. The UK NIHR is the largest single funder of the Cochrane Common Mental Disorders Group.

Disclaimer: the views and opinions expressed herein are those of the review authors and do not necessarily reflect those of the NIHR, National Health Service (NHS), the Department of Health and Social Care or the UK Government.

REFERENCES

Additional references

American Psychiatric Association 2000

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

American Psychiatric Association 2013

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Arlington (VA): American Psychiatric Association, 2013.

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 JW Jr. 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.

Bjelland 2002

Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. *Journal of Psychosomatic Research* 2002;**52**(2):69-77.

Brennan 2010

Brennan C, Worrall-Davies A, McMillan D, Gilbody S, House A. The Hospital Anxiety and Depression Scale: a diagnostic meta-analysis of case-finding ability. *Journal of Psychosomatic Research* 2010;**69**(4):371-8.

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.

Connor 2000

Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. *British Journal of Psychiatry* 2000;**176**:379-86.

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.

Davenport 2014

Davenport CF, Leeflang MM, Takwoingi Y, Deeks JJ. Use of QUADAS-2. Lesson 6.3: Cochrane Collaboration DTA Online Learning Materials. training.cochrane.org/uploads/resources/embedded_resources/DTA_modules/6.3_Use_of_QUADAS-2/story_html5.html (accessed 15 August 2022).

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.

Emons 2019

Emons WH, Habibović M, Pedersen SS. Prevalence of anxiety in patients with an implantable cardioverter defibrillator: measurement equivalence of the HADS-A and the STAI-S. *Quality of Life Research* 2019;**28**(11):3107-16.

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.

EPPI-Reviewer [Computer program]

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

Ercan 2015

Ercan I, Hafizoglu S, Ozkaya G, Kirli S, Yalcintas E, Akaya C. Examining cut-off values for the state-trait anxiety inventory [Examinando los puntajes de corte para el inventario de ansiedad estado-rasgo]. *Revista Argentina de Clínica Psicológica* 2015;**24**:143-8.

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.

Greenberg 1999

Greenberg PE, Sisitsky T, Kessler RC, Finkelstein SN, Berndt ER, Davidson JR, et al. The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry* 1999;**60**(7):427-35.

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.

Hoffman 2008

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

Julian 2011

Julian LJ. Measures of anxiety: State-Trait Anxiety Inventory (STAI), Beck Anxiety Inventory (BAI), and Hospital Anxiety and Depression Scale-Anxiety (HADS-A). *Arthritis Care & Research* 2011;**63**(Suppl 11):S467-72.

Kessler 2001

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

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.

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 2008

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

Leon 1995

Leon AC, Portera L, Weissman MM. The social costs of anxiety disorders. *British Journal of Psychiatry* 1995;**166**(S27):19-22.

Levis 2019

Levis B, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;**365**:1476.

Lian 2019

Lian Q, Hodges JS, Chu H. A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for network meta-analysis of diagnostic tests. *Journal of the American Statistical Association* 2019;**114**(527):949-61.

Lieb 2005

Lieb R, Becker E, Altamura C. The epidemiology of generalized anxiety disorder in Europe. *European Neuropsychopharmacology* 2005;**15**(4):445-52.

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.

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.

Ma 2018

Ma X, Lian Q, Chu H, Ibrahim JG, Chen Y. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. *Biostatistics* 2018;**19**(1):87-102.

Macaskill 2021

Macaskill P, Takwoingi Y, Deeks JJ, Gatsonis C. Chapter 10: Understanding meta-analysis. Draft version (17 June 2021). In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors(s). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2. London (UK): Cochrane, 2021.

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 (Germany): Springer-Verlag, 2013.

McInnes 2018

McInnes MD, Moher D, Thombs BD, McGrath TA, Bossuyt PM, and the PRISMA-DTA Group. 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.

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.

Nyaga 2018

Nyaga VN, Arbyn M, Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. *Statistical Methods in Medical Research* 2018;**27**(8):2554-66.

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 metaanalysis. *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 Core Team 2021 [Computer program]

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

Reitsma 2005

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, 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.

Review Manager 2020 [Computer program]

The Cochrane Collaboration Review Manager (RevMan). Version 5.4. Copenhagen: The Cochrane Collaboration, 2020.

Rücker 2018

Rücker G. Network meta-analysis of diagnostic test accuracy studies. In: Biondi-Zoccai Giuseppe, editors(s). Diagnostic Meta-Analysis: a Useful Tool for Clinical Decision-Making. Berlin (Germany): Springer, 2018:183-97.

Rücker 2020 [Computer program]

R Foundation for Statistical Computing diagmeta: Meta-analysis of diagnostic accuracy studies with several cutpoints. R package version 0.4-0. Rücker G, Steinhauser S, Kolampally S, Schwarzer G. Vienna, Austria: R Foundation for Statistical Computing, 2020.

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 2011

Schneider A, Hörlein E, Wartner E, Schumann I, Henningsen P, Linde K. Unlimited access to health care-impact of psychosomatic co-morbidity on utilisation in German general practices. *BMC Family Practice* 2011;**12**(1):51.

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 2017a

Schneider A, Hilbert S, Hamann J, Skadsem S, Glaser J, Löwe B, et al. The implications of psychological symptoms for length of sick leave: burnout, depression, and anxiety as predictors in a primary care setting. *Deutsches Ärzteblatt International* 2017;**114**(17):291-7.

Schneider 2017b

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.

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.

Spielberger 1970

Spielberger CD, Gorsuch RL, Lushene RE. STAI Manual for the State-Trait Anxiety Inventory. Palo Alto (CA): 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.

Steel 2014

Steel Z, Marnane C, Iranpour C, Chey T, Jackson JW, Patel V, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology* 2014;**43**:476-93.

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.

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.

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.

Walters 2012

Walters K, Rait G, Griffin M, Buszewicz M, Nazareth I. Recent trends in the incidence of anxiety diagnoses and symptoms in primary care. *PLOS One* 2012;**7**(8):e41670.

Weissman 1988

Weissman MM. The epidemiology of anxiety disorders: rates, risks and familial patterns. *Journal of Psychiatric Research* 1988;**22**:99-114.

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. The ICD-10 Classification of Mental and Behavioural Disorders: diagnostic criteria for research. 2nd 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.

Wittchen 2001

Wittchen HU, Hoyer J. Generalized anxiety disorder: nature and course. *Journal of Clinical Psychiatry* 2001;**62**(Suppl 11):15-9; discussion 20-1.

Wittchen 2002

Wittchen HU. Generalized anxiety disorder: prevalence, burden, and cost to society. *Depression and Anxiety* 2002;**16**(4):162-71.

Wu 2021

Wu Y, Levis B, Ioannidis JP, Benedetti A, Thombs B, DEPRESSion Screening Data (DEPRESSD) Collaboration. Probability of major depression classification based on the SCID, CIDI, and MINI diagnostic interviews: a synthesis of three individual participant data meta-analyses. *Psychotherapy and Psychosomatics* 2021;**90**:28-40.

Wuyek 2011

Wuyek LA, Antony MM, McCabe RE. Psychometric properties of the panic disorder severity scale: clinician-administered and self-report versions. *Clinical Psychology & Psychotherapy* 2011;**18**(3):234-43.

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**:1592-9.

Zigmond 1983

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

APPENDICES

Appendix 1. Embase search strategy

Generalized Anxiety Disorder 2-item (GAD-2), Generalized Anxiety Disorder 7-item (GAD-7)

Embase (Ovid) <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)

****

Hospital Anxiety and Depression Scale (HADS)

Embase (Ovid) <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 social* or unspecif*)).tw,kf.

9 GAD.tw,kf.

10 ADNOS*.tw,kf.

11 (agoraphobi* or panic*).tw,kf.

12 *Anxiety/ or Anxiety/di

13 anxiety.ti,kf.

14 anxi*.ab. /freq=3

15 or/1-14

16 "Hospital Anxiety and Depression Scale"/

17 (HADS or HAD-S).tw,kf.

18 (hospital* adj2 anxi* adj2 depress* adj2 (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.

19 or/16-18

20 (15 and 19)

21 Gold Standard/

22 (reference standard? or gold standard?).tw,kf.

23 Psychiatric Diagnosis/

24 Interview/

25 Structured Interview/

26 Structured Clinical Interview for DSM Disorders/

27 (structured clinic* interview* or SCI or SCID).tw,kf.

28 ((standard* adj2 clinic* adj2 interview* adj2 psychiatr*) or SCIP).tw,kf.

29 (schedul* adj2 clinic* assess* adj2 neuropsych*).tw,kf.

30 Mini International Neuropsychiatric Interview/

31 (international neuropsych* interview* or MINI).tw,kf.

32 Mini Mental State Examination/

33 (mental state examination* or MMSE).tw,kf.

34 (composite international diagnos* interview* or CIDI).tw,kf.

35 ((anxi* adj2 interview* adj2 schedule*) or ADIS).tw,kf.

36 Diagnostic Interview Schedule/

37 (diagnos* interview* schedule* or DIS).tw,kf.

38 General Mental Disease Assessment/

39 (geriatric* mental state* or GMSA).tw,kf.

40 ((cambridge examination adj3 mental disorder* adj3 elder*) or CAMDEX).tw,kf.

41 ((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.

42 ((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.

43 clinical diagnosis.mp.

44 ((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.

45 ((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.

46 or/21-45

47 DIAGNOSTIC TEST ACCURACY STUDY/

48 DIAGNOSTIC ACCURACY/

49 VALIDATION STUDY/

50 "SENSITIVITY and SPECIFICITY"/

51 specificity.tw,kf.

52 RECEIVER OPERATING CHARACTERISTIC/

53 RELIABILITY/

54 INTERNAL VALIDITY/

55 INTERNAL CONSISTENCY/

56 (validat* or validity).tw,kf.

57 likelihood ratio*.tw,kf.

58 ((re-test or retest or test-retest) adj reliability).tw,kf.

59 receiver operating characteristic*.tw,kf.

60 ROC.tw,kf.

61 (DTA or (diagnos* adj2 accura*)).tw,kf.

62 (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.

63 ((degree? or rate* or rating) adj3 agreement?).tw,kf.

64 or/47-63

65 (sensitivity or specificity or validity or accuracy or gold standard* or reference standard* or ROC).tw,kf,hw.

66 (46 or 64)

67 limit 66 to yr="1990 -Current"

68 (20 and 67)

69 limit 68 to conference abstract status

70 (68 not 69)

71 (65 and 69)

72 (70 or 71)

****

Beck Anxiety Inventory (BAI)

Embase (Ovid) <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 Beck Anxiety Inventory/

20 BAI.tw,kf.

21 (beck* adj2 anxi* adj2 (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.

22 or/19-21

23 (18 and 22)

24 Gold Standard/

25 (reference standard? or gold standard?).tw,kf.

26 Psychiatric Diagnosis/

27 Interview/

28 Structured Interview/

29 Structured Clinical Interview for DSM Disorders/

30 (structured clinic* interview* or SCI or SCID).tw,kf.

31 ((standard* adj2 clinic* adj2 interview* adj2 psychiatr*) or SCIP).tw,kf.

32 (schedul* adj2 clinic* assess* adj2 neuropsych*).tw,kf.

33 Mini International Neuropsychiatric Interview/

34 (international neuropsych* interview* or MINI).tw,kf.

35 Mini Mental State Examination/

36 (mental state examination* or MMSE).tw,kf.

37 (composite international diagnos* interview* or CIDI).tw,kf.

38 ((anxi* adj2 interview* adj2 schedule*) or ADIS).tw,kf.

39 Diagnostic Interview Schedule/

40 (diagnos* interview* schedule* or DIS).tw,kf.

41 General Mental Disease Assessment/

42 (geriatric* mental state* or GMSA).tw,kf.

43 ((cambridge examination adj3 mental disorder* adj3 elder*) or CAMDEX).tw,kf.

44 ((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.

45 ((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.

46 clinical diagnosis.mp.

47 ((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.

48 ((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.

49 or/24-48

50 DIAGNOSTIC TEST ACCURACY STUDY/

51 DIAGNOSTIC ACCURACY/

52 VALIDATION STUDY/

53 "SENSITIVITY and SPECIFICITY"/

54 specificity.tw,kf.

55 RECEIVER OPERATING CHARACTERISTIC/

56 RELIABILITY/

57 INTERNAL VALIDITY/

58 INTERNAL CONSISTENCY/

59 (validat* or validity).tw,kf.

60 likelihood ratio*.tw,kf.

61 ((re-test or retest or test-retest) adj reliability).tw,kf.

62 receiver operating characteristic*.tw,kf.

63 ROC.tw,kf.

64 (DTA or (diagnos* adj2 accura*)).tw,kf.

65 (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.

66 ((degree? or rate* or rating) adj3 agreement?).tw,kf.

67 or/50-66

68 (sensitivity or specificity or validity or accuracy or gold standard* or reference standard* or ROC).tw,kf,hw.

69 (49 or 67)

70 limit 69 to yr="1990 -Current"

71 (23 and 70)

72 limit 71 to conference abstract status

73 (71 not 72)

74 (68 and 72)

75 (73 or 74)

****

State Trait Anxiety Inventory (STAI)

Embase (Ovid) <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 State Trait Anxiety Inventory/

20 (STAI or STAI1* or STAI2* or STAI6*).tw,kf.

21 (Spielberger and state and trait and anxiety).tw,kf.

22 (state adj3 trait adj3 anxi* adj3 (self-report* or scale? or scor* or checklist* or check-list* or criteria or form? or index or indexes or indices or inventory or inventories or instrument? or measure* or procedure? or questionnaire? or screen* or tool*)).tw,kf.

23 or/19-22

24 (18 and 23)

25 Gold Standard/

26 (reference standard? or gold standard?).tw,kf.

27 Psychiatric Diagnosis/

28 Interview/

29 Structured Interview/

30 Structured Clinical Interview for DSM Disorders/

31 (structured clinic* interview* or SCI or SCID).tw,kf.

32 ((standard* adj2 clinic* adj2 interview* adj2 psychiatr*) or SCIP).tw,kf.

33 (schedul* adj2 clinic* assess* adj2 neuropsych*).tw,kf.

34 Mini International Neuropsychiatric Interview/

35 (international neuropsych* interview* or MINI).tw,kf.

36 Mini Mental State Examination/

37 (mental state examination* or MMSE).tw,kf.

38 (composite international diagnos* interview* or CIDI).tw,kf.

39 ((anxi* adj2 interview* adj2 schedule*) or ADIS).tw,kf.

40 Diagnostic Interview Schedule/

41 (diagnos* interview* schedule* or DIS).tw,kf.

42 General Mental Disease Assessment/

43 (geriatric* mental state* or GMSA).tw,kf.

44 ((cambridge examination adj3 mental disorder* adj3 elder*) or CAMDEX).tw,kf.

45 ((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.

46 ((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.

47 clinical diagnosis.mp.

48 ((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.

49 ((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.

50 or/25-49

51 DIAGNOSTIC TEST ACCURACY STUDY/

52 DIAGNOSTIC ACCURACY/

53 VALIDATION STUDY/

54 "SENSITIVITY and SPECIFICITY"/

55 specificity.tw,kf.

56 RECEIVER OPERATING CHARACTERISTIC/

57 RELIABILITY/

58 INTERNAL VALIDITY/

59 INTERNAL CONSISTENCY/

60 (validat* or validity).tw,kf.

61 likelihood ratio*.tw,kf.

62 ((re-test or retest or test-retest) adj reliability).tw,kf.

63 receiver operating characteristic*.tw,kf.

64 ROC.tw,kf.

65 (DTA or (diagnos* adj2 accura*)).tw,kf.

66 (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.

67 ((degree? or rate* or rating) adj3 agreement?).tw,kf.

68 or/51-67

69 (sensitivity or specificity or validity or accuracy or gold standard* or reference standard* or ROC).tw,kf,hw.

70 (50 or 68)

71 limit 70 to yr="1990 -Current"

72 (24 and 71)

73 limit 72 to conference abstract status

74 (72 not 73)

75 (69 and 73)

76 (74 or 75)

Search narrative

Searching for DTA studies in psychiatry is challenging, because unlike many other medical specialties, 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 suite of reviews (for self-reporting questionnaires to detect anxiety disorders in adults). 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 will include a sensitive list of terms for validated psychiatric interview schedules or generic terms for clinical interviews, or clinician administered checklists/questionnaires, or broad database subject headings for 'interviews'.

Diagnostic test accuracy filter

We will use a study design filter created by the Cochrane Common Mental Disorders' 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 (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). It will be revalidated in a similar manner, when the search is updated prior to publication.

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.

While the reviews will focus 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). While 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). Please note that the search already contains a sensitive list of terms for specific anxiety disorders (agoraphobia, generalized anxiety disorder, panic, social phobia, mixed anxiety and depression).

Appendix 2. QUADAS rating guidelines

Rating guidelines for QUADAS-2

Version 7 April 2022

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 participants (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 participants 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 the specific domain

The assessment of risk of bias focuses 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'; when one question was answered 'unclear' you might rate 'low risk of bias' if you consider this adequate (please give reasons).
- Rate 'high risk of bias' if one or more signalling questions were rated 'no'.
- If one or more signalling questions were 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).

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.

Domain I: patient selection

Could the selection of participants have introduced bias?

[Additional remark for the published protocol not included in the version used by review authors doing the assessment: we do not use the original QUADAS-2 'signalling question 2: was a case-control design avoided?' as we exclude any case-control studies. We merged the original 'signalling question I.1. Was a consecutive or random sample of patients enrolled' and 'signalling question I.3. Did the study avoid unnecessary exclusions' because the issues are closely linked in our specific study set (see also the remark for the *review authors doing the assessment below)]*

Signalling question I.1: did the study avoid unnecessary exclusions?/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 preselected regarding issues not related to anxiety diagnoses (e.g. they were participants in a study on quality of life among participants with epilepsy, or they were frequent attenders participating in a case management programme) or for other reasons (e.g. they responded to advertisements of the study). However, we will ignore such preselection as it usually is mainly relevant to generalizability and less to bias. We will only focus on the question whether among the potential participants (e.g. the participants with epilepsy included in the quality of life study) sampling avoid selection bias.

- Rate 'yes' (low risk of bias) if:
  - the description (e.g. 'consecutive or random sampling', 'all potential participants were invited') makes clear that the sampling procedure avoided systematic selection with the potential to influence diagnostic accuracy AND
  - 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 OR
  - when you are uncertain whether the sampling process might have been biased.
- Rate 'no' (high risk of bias) if:
  - the sampling allowed selection of participants with specific characteristics potentially influencing diagnostic accuracy.

Domain II: index test

Could the conduct or interpretation of the index test have introduced bias?

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 it clear that the index test(s) was (were) filled before the structured interview was performed OR
    - there was an explicit statement that questionnaire rating and structured interview were performed 'independently', that participants and raters were 'blinded' for each other rating, etc.
- Rate 'unclear' if:
    - the process description is unclear, but there is 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 diagnostic test accuracy?

- 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 or greater recommended by the scale developers') is described;
    - (in some exceptional cases there might be two predefined 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);
    - a cut-off different from the cut-off recommended by the scale authors (e.g. GAD-7 score 10 or greater; GAD-2 score 3 or greater) without being explicitly predefined.
- Rate 'unclear' if:
    - the cut-off recommended by the scale authors (GAD-7 score 10 or greater; GAD-2 score 3 or greater) was used but there is no statement whether this choice was made in advance.
- Rate 'not applicable' if:
    - diagnostic accuracy (sensitivity/specificity) is reported for more than one cut-off.

Signalling question II.3: if findings for several threshold 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 and the HADS-A for ≥ 10 cut-offs (covering a range including at least ± 3 around the recommended cut-offs); for the BAI and the STAI ≥ for 15 cut-offs (± 5 around the recommended cut-off).
- Rate 'no' if:
    - the conditions for the answer 'yes' are not met.
- Rate 'not applicable' if:
    - diagnostic accuracy is reported for only one cut-off.

Domain III: reference standard

Could the reference standard, its conduct or its interpretation have introduced bias?

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 whether specific interviews or interviews in general are 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 people).

We will not use an 'unclear' option for this signalling question. According to our selection criteria, we will only include studies using accepted standardized or semi-standardized validated structured and semi-structured interviews as reference standard. So, by definition the use of one of these interviews usually leads to a 'yes' answer – unless there are active hints that the interview was implemented or applied incorrectly. Any such deviation from good practice will result in an answer 'no'.

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 questionnaire rating and structured interview were performed 'independently', that participants and raters were 'blinded' for each other rating, etc.
- Rate 'unclear' if:
  - the process description is unclear, but there is 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

Could the patient flow have introduced bias?

Signalling question 1: was there an appropriate interval between index test and reference standard?

- Rate 'yes' if:
  - there is an explicit statement that the time lag between questionnaire rating and structured interview was two weeks or less.
- Rate 'unclear' if:
  - the process description is unclear, but there is no indication of high risk of bias (see answer 'no').
- Rate 'no' if:
  - the time lag is greater than two weeks (if it exceeds four 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 result of the index test(s).
- Rate 'unclear' if:
  - the process description is unclear, but there is no indication of high risk of bias (see answer 'no').
- Rate 'no' if:
  - less than 90% 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 result 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 index test/reference standard or other factors likely to cause bias.
- Rate 'unclear' if:
  - the process description is unclear, but there is no indication of high risk of bias (see answer 'no').
- Rate 'no' if:
  - less than 90% of participants were included in the analyses OR there are reasons to believe that exclusions were related to the results of index test/reference standard or other likely to induce bias OR both.

Applicability

Applicability question A.1: are there concerns that the included participants do not match the review question?

Given that our review question is very inclusive, this question usually should be rated 'no' (no concerns). Rate 'yes' (or 'unclear') only if the participants are irrelevant for our questions (even though the study meets inclusion criteria).

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 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 people).

CONTRIBUTIONS OF AUTHORS

AS, KL, MO, AH, GR and BL conceptualized the overall review project and obtained funding.

MO and KL drafted the protocol, selection and extraction forms and adapted scoring rules for the risk of bias assessment with QUADAS.

SD developed and implemented the search in collaboration with MO and KL.

All authors contributed to the protocol development and commented on protocol drafts.

DECLARATIONS OF INTEREST

KL: none.

MO: none.

CT: none.

ZA: none.

VS: none.

AH: none.

SD: none.

GR: none.

BL: is a codeveloper of the GAD-7 and GAD-2 scales. Both instruments are in the public domain and there is no financial conflict of interest. BL will not be involved in the extraction and assessment of primary studies using these instruments.

AS: none.

SOURCES OF SUPPORT

Internal sources

- No sources of support provided

External sources

- German Federal Ministry of Education and Research Grant 01KG2105, Other

NOTES

The overall project, in which the Cochrane Reviews described above are embedded, comprises an additional subproject 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 subproject will be based on the data collected in the four Cochrane Reviews. However, the findings of this additional project part will not be addressed within the Cochrane Reviews but published separately. In addition, we will also search studies for two additional, newer questionnaires: the PROMIS (Patient Reported Outcomes Measurement Information System) Anxiety Short Form and the OASIS (Overall Anxiety Severity and Impairment Scale). If the search identifies a sufficient number of eligible studies, we will consider reviewing the DTA of these scales with the methods described in this protocol.

If the studies included in the four Cochrane Reviews or a subset of studies is clinically sufficiently comparable (and the assumption of transitivity plausible), we will conduct a network meta-analysis of multiple diagnostic tests based on a single common threshold. Currently, we plan to use the beta-binomial model by Nyaga and colleagues (Nyaga 2018) or the Bayesian hierarchical model (Lian 2019; Ma 2018). However, network meta-analysis of DTA studies (with multiple cut-offs) is a new field (Rücker 2018), and it is difficult to predict which methods will be available when our project approaches the analysis phase. If new methods become available that allow to use multiple thresholds data, we will use these if feasible. We will use the QUADAS-C instrument to assess the risk of bias in studies using two or more index instruments directly (Yang 2021).

We will prepare an analysis plan in the late phase of data extraction.