

Risk of bias, applicability concerns and data reporting in diagnostic test accuracy studies of questionnaires used for anxiety screening –lessons from six systematic reviews

Klaus Linde, Alexey Fomenko, Zekeriya Aktürk, Stefanie Eck, Alexander Hapfelmeier, Antonius Schneider

Angaben zur Veröffentlichung / Publication details:

Linde, Klaus, Alexey Fomenko, Zekeriya Aktürk, Stefanie Eck, Alexander Hapfelmeier, and Antonius Schneider. 2026. "Risk of bias, applicability concerns and data reporting in diagnostic test accuracy studies of questionnaires used for anxiety screening –lessons from six systematic reviews." *Journal of Psychosomatic Research* 210: 112967.
<https://doi.org/10.1016/j.jpsychores.2026.112967>.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
<https://creativecommons.org/licenses/by/4.0/deed.de>





Review article

Risk of bias, applicability concerns and data reporting in diagnostic test accuracy studies of questionnaires used for anxiety screening – Lessons from six systematic reviews

Klaus Linde^{a,*}, Alexey Fomenko^a, Zekeriya Aktürk^b, Stefanie Eck^a, Alexander Hapfelmeier^{a,c}, Antonius Schneider^a

^a Technical University Munich, TUM School of Medicine and Health, Department Clinical Medicine, Institute for General Practice and Health Services Research, Munich, Germany

^b University of Augsburg, Faculty of Medicine, General Practice, Germany

^c Technical University Munich, TUM School of Medicine and Health, Department Clinical Medicine, Department of AI and Informatics in Medicine, Munich, Germany

ARTICLE INFO

Keywords:

Anxiety disorders
Mass screening
Surveys and questionnaires
Sensitivity and specificity
Bias
Publication Bias
Selection Bias

ABSTRACT

Background: This article aims to provide an overview of the risk of bias, concerns regarding applicability, and the reporting of findings from studies examining the diagnostic test accuracy (DTA) of self-report questionnaires used to screen for anxiety disorders.

Methods: We merged data from 103 studies included in six systematic reviews on the diagnostic accuracy of eight anxiety questionnaires (HADS-A, GAD-7, GAD-2, BAI, STAI-S, STAI-T, PROMIS-A-SF8a, and OASIS) in adults. To be included, primary studies had to report the sensitivity and specificity for at least one cut-off when tested against a (semi-)structured clinical interview used as a reference standard. Risk of bias and applicability were rated with the QUADAS-2 tool (Quality Assessment of Diagnostic Accuracy Studies).

Results: The 103 studies involved a total of 34,833 participants (range 48 to 3215). The most commonly used questionnaires were the HADS-A (54 studies), the GAD-7 (46), and the GAD-2 (31). Overall, 80% of the studies were rated as having unclear or high risk of bias, and for 88%, concerns regarding applicability were rated as unclear or high. Uncertainties or concerns were particularly common regarding participant recruitment and whether the samples included reflected realistic screening populations. Publications often reported sensitivity and specificity in a very selective manner for a single or a few cut-offs only.

Conclusions: Most DTA studies on questionnaires used to screen for anxiety have significant shortcomings in the reporting of methods and results. Our findings provide a basis for recommendations for future studies.

1. Introduction

While anxiety disorders are widespread, they often go undiagnosed and untreated outside mental health care settings [1,2]. For this reason, several organizations, including the US Preventive Services Task Force (USPSTF), recommend screening for anxiety disorders using self-assessment questionnaires [3–5]. Several questionnaires have been validated in multiple languages and used for screening [6,7]. A key prerequisite for efficient screening is that the questionnaires reliably distinguish between individuals with and without an anxiety disorder, and that the cut-off scores used are appropriate for the population under study. Therefore, diagnostic test accuracy (DTA) studies investigating

the sensitivity and specificity of the questionnaires by comparing their findings with those from a validated structured or semi-structured clinical interview as a reference standard are of major importance.

We recently conducted six systematic reviews of diagnostic studies investigating the accuracy of eight widely used anxiety questionnaires against validated structured or semi-structured interviews for screening in adult populations [8–13]. Our reviews indicate that, despite the large number of published studies evaluating screening instruments, it remains unclear whether their diagnostic accuracy is sufficiently high and which cutoff values are most appropriate, for example, depending on whether the aim is to optimize rule-in or rule-out performance. This uncertainty is partly attributable to the numerous factors that can

* Corresponding author at: Institute of General Practice and Health Service Research, TUM Klinikum Rechts der Isar Orleansstr. 47, 81667 München, Germany.
E-mail address: klaus.linde@mri.tum.de (K. Linde).

influence DTA, including characteristics of the target population, disease prevalence, the reference standard employed, and linguistic and cultural differences. However, this uncertainty also reflects important methodological and reporting limitations in the existing body of evidence. Further studies that avoid these limitations are needed.

While our individual reviews included risk-of-bias assessments, their primary focus was on DTA findings, leaving limited scope for a comprehensive evaluation and discussion of methodological and reporting quality. Therefore, we undertook a systematic analysis of quality and reporting across all questionnaires using the aggregated data from the six reviews. The primary objective of our analysis was to identify the most important methodological limitations and reporting deficiencies in existing DTA studies. Building on these findings, we explore in the Discussion how future studies can overcome these limitations to generate more robust evidence and support better-informed evidence-based decisions.

2. Methods

As the general methods used in our reviews have already been described in detail [8–14], we will briefly discuss only those directly relevant to this article.

2.1. Eligibility criteria

To be eligible, studies had to meet the following criteria: 1) allowing the generation of contingency (2×2) tables (number of true positive, false positive, false negative, and true negative index test results); 2) including adult participants (18 years and older); we excluded studies that recruited participants seeking help in mental health settings or participants seeking help for mental health complaints in other settings as they represent a diagnostic (symptomatic) rather than a screening population; 3) applying at least one of the eight selected questionnaires, regardless of the language: the Hospital Anxiety and Depression Scale – Anxiety Subscale (HADS-A [15]); the Generalized Anxiety Disorder 7-item scale (GAD-7) and its two-item version GAD-2 [16]; the Beck Anxiety Inventory (BAI [17]); the two subscales of the State and Trait Anxiety Inventory (STAI-S and STAI-T [18]); the Patient-Reported Outcomes Measurement Information System Anxiety Short Forms (PROMIS-A-SF [19]); or the Overall Anxiety Severity and Impairment Scale (OASIS [20]); 4) providing data on any type of anxiety disorder (AAD) and/or generalized anxiety disorder (GAD) according to DSM-IV or DSM-5 as target condition; 5) using validated semi-structured or structured clinical interviews as reference standard such as the SCID (Structured Clinical Interview for DSM) or the MINI (Mini-International Neuropsychiatric Interview). Furthermore, we excluded case-control studies; studies in which the reference standard was informally based on an ICD- or DSM-based checklist, a clinical diagnostic interview without operationalization, or another (anxiety) questionnaire; and studies with a time lag of over four weeks between the administration of the index test and the reference standard. For this article focusing on published DTA studies, we additionally excluded 23 studies from our original reviews. In these studies, both an eligible index test and an adequate reference standard were applied, but DTA was neither a primary nor a secondary objective, and no DTA findings had been published. In response to our request, the authors of these studies had provided us with unpublished DTA results for the reviews. These studies were excluded from the present meta-research analysis because the objective here was to evaluate the reporting and methodological transparency of published DTA studies, and studies without published DTA analyses could not contribute to that question.

2.2. Information sources

For each of the underlying systematic reviews, we separately searched the following databases using relevant subject headings

(controlled vocabularies), text-words, and search syntax: Embase (Ovid), MEDLINE (Ovid), PubMed-NOT-MEDLINE (NLM) PsycINFO (Ovid). The last update ranged from 18 January 2024 to 15 July 2025, depending on the individual review.

2.3. Study selection, data extraction, assessment of risk of bias, and applicability

Two reviewers independently selected, extracted, and assessed studies. We extracted information (list not exhaustive) on study design, objective, mean age, sex distribution, setting, details regarding the reference standard and the index test, types of anxiety disorders identified in the study sample, country of origin, numbers of patients who were included and analyzed, and available DTA data (most often prevalence of the condition, cut-off used, sensitivity, specificity). We tried to contact the corresponding authors of eligible studies and asked them to provide us with unpublished information about cut-offs and their DTA, as well as details of study methods.

We applied the Quality Assessment of Diagnostic Accuracy Studies tool version 2 (QUADAS-2, see Box 1 for details) to assess the risk of bias and concerns regarding the applicability [21]. The risk of bias assessment addresses four domains: participant selection, index test, reference standard, and flow and timing. Each domain contains several signaling questions tailored for this review to judge the ‘risk of bias’. In addition to the stricter original QUADAS approach, we also used a slightly more liberal approach for assessing the overall risk of bias (see Box 1) to ensure greater discriminatory power for the meta-regression analyses used to investigate whether risk of bias was associated with the study results. Applicability concerns were considered low when the study’s patient population, index test, and reference standard closely aligned with the systematic review’s research question focusing on DTA in a screening pathway. Where the authors of the primary studies provided additional methodological information, this was taken into account in the assessment. For further details on how judgments of each domain were determined and the QUADAS-2 assessment guidance tailored for this review, see Appendix A.

2.4. Analyses

For this article, the databases of the individual reviews were merged into a single file. Descriptive statistical analyses were performed using IBM SPSS, Version 30 (IBM Corp., Armonk, NY).

3. Results

3.1. Characteristics of included studies

Overall, 103 unique studies were included in our analyses (see Appendix B for references). The studies were published between 1998 and 2024 and included a total of 34,833 participants (median sample size 170; Table 1). Sixty-three studies (61%) investigated a single index test, 31 (30%) two index tests, and 9 (9%) three to five index tests. The most frequently investigated index tests were the HADS-A (54 studies; 52%), the GAD-7 (46 studies; 45%), and the GAD-2 (31 studies; 30%). The index tests were administered in 30 different languages, with English questionnaire versions being evaluated most often (in 34 studies using English-language questionnaires only, and an additional five studies using questionnaires in two to five languages). The MINI was the most frequently used reference standard (50 studies; 48%), followed by the SCID (39 studies; 38%). Seventy-nine studies (77%) were performed in participants having specific medical conditions (e.g., epilepsy, cancer), 14 (14%) in clinical settings regardless of conditions (e.g., unselected primary care patients), and 10 (10%) in non-healthcare study populations (e.g., the general population, elderly, students) and. We received additional information on methods and/or findings from authors for 63 studies (61%).

Box 1

Assessment of risk of bias and applicability concerns with the QUADAS-2 tool as used in the six reviews.

RISK OF BIAS DOMAIN I: Patient Selection

1. Was a consecutive or random sample of patients enrolled?
2. Did the study avoid inappropriate exclusions?

RISK OF BIAS DOMAIN II: Index Test

1. Were the index test results interpreted without knowledge of the results of the reference standard?
2. Was there a predefined threshold for a primary analysis of diagnostic accuracy? OR
3. If findings for several thresholds are presented, was the range of thresholds prespecified OR was a broad range of cut-offs reported?

RISK OF BIAS DOMAIN III: Reference Standard

1. Is the reference standards likely to correctly classify the target condition?
2. Were the reference standard results interpreted without knowledge of the results of the index tests?

RISK OF BIAS DOMAIN IV: Flow and Timing

1. Was there an appropriate interval between index test and reference standard?
2. Were all patients included in the analysis?
3. Did all patients receive a reference standard?

DOMAIN RATING: 'low risk' if all answers to questions are 'yes', 'high risk' if one or more questions are answered 'no'

OVERALL RATING OF RISK OF BIAS (STRICT – recommended by developers): 'low risk' only if all domains are 'low risk', 'high risk' if one or more domains are 'high risk'

OVERALL RATING OF RISK OF BIAS (LIBERAL - used in our reviews): 'low risk' if no domain is 'high risk' and no more than one 'unclear', 'high risk' if one or more domains are 'high risk'

RISK OF BIAS APPLICABILITY

1. Are there concerns that the included patients and setting do not match the review question?
2. Are there concerns that the index test, its conduct, or interpretation differ from the review question?
3. Are there concerns that the target condition as defined by the reference standard does not match the question?

OVERALL RATING OF APPLICABILITY: 'low/no concern' if all three questions answered 'no', 'high' if one of more questions are answered 'yes'

Accuracy findings for the target conditions AAD and GAD were available for 72 studies (70%) and 64 studies (62%), respectively. Findings for both AAD and GAD were available from 33 studies (32%), for AAD only from 39 (38%), and for GAD only from 31 (30%) studies. The median prevalence was 18% for AAD and 10% for GAD.

3.2. Risk of bias and applicability

When using the original QUADAS assessment rule, overall risk of bias was considered low in 15 studies (15%), unclear in 15 (15%), and high in 73 (71%), while these figures were 21 (20%), 9 (9%), and 73 (71%), respectively, when using our more liberal assessment (Fig. 1). At the domain level, risk of bias was rated as unclear or high in 59 studies (57%) for the patient selection, in 57 (56) for the index test, in 37 (36%) for the reference standard, and in 44 (43%) for the flow and timing. Methodological and reporting problems were most frequently observed for the 'sampling' question ("Was a consecutive or random sample of participants/patients enrolled?"). In 33 studies (32%), the description was insufficient to make a reliable assessment, and in an additional 22 (21%), convenience sampling was used. A flow diagram depicting the number of individuals approached for participation, included in the study, and contributing data to the analysis was presented in only 9 studies (9%); a further 15 (15%) reported all relevant flow details in the

text. Key methodological issues, such as blinding, were not mentioned in more than a third of publications; in some cases, blinding of the index test and the reference standard could be confirmed by contacting the study authors.

Due to the use of clearly defined questionnaires and validated diagnostic interviews as reference standards, there were almost no applicability concerns regarding the index tests or the confirmation of the target condition. However, in only 14 studies (14%), we were confident that participants were likely to represent a screening population in a real-world screening approach. The main issue here was whether patients with a known current anxiety disorder or who were undergoing treatment for such a disorder were excluded (which would be the case in a classical screening approach). In 68 studies (66%), the description of the recruitment process was insufficient, and in 21 (20%), we had applicability concerns. Taking all three domains into account, applicability concerns were considered low in only 12 (12%) studies.

3.3. Reporting of study findings

After obtaining additional information from authors for 38 studies (37%), we were able to include 1398 diagnostic 2×2 tables for the target condition AAD and 1246 for the target condition GAD in our meta-analyses (see Table 2). While 33 studies (32%) were included into

Table 1
Characteristics of included studies.

Total number of studies included in the six reviews	103
Publication years	
– 1998–2008	16 (15%)
– 2009– 2016	42 (41%)
– 2017–2024	45 (44%)
Diagnostic accuracy	
– Primary study objective	85 (83%)
– Only secondary objective	18 (17%)
Number of participants included in analyses	
– Sum	34,833
– Median (quartile 1/3)	170 (100/385)
– Minimum/Maximum	48/3215
Mean age in years: median (Quartile 1/3)	47 (37/60)
Percentage female: median (Quartile 1/3)	56 (40/72)
Number of eligible index tests	
– One	63 (61%)
– Two	31 (30%)
– Three	4 (4%)
– Four	2 (2%)
– Five	3 (3%)
Index tests investigated	
– HADS-A	54 (52%)
– GAD-7	46 (45%)
– GAD-2	31 (30%)
– BAI	9 (9%)
– STAI-S	7 (7%)
– STAI-T	5 (5%)
– PROMIS-A SF8	5 (5%)
– OASIS	3 (3%)
Index test languages	30
– English	39 (38%)*
– German	9 (9%)
– Chinese	9 (9%)
– Spanish	8 (6%)**
Reference standard	
– MINI	50 (48%)
– SCID	39 (38%)
– Other structured interviews	14 (14%)
Diagnostic classification (5 studies unclear or missing information)	
– DSM-IV	87 (84%)
– DSM-5	11 (11%)
Study population	
– Patients with specific conditions	79 (77%)
o Most frequent: neurology (epilepsy alone)	21 (11)
o 2nd most frequent: oncology	15
o 3rd most frequent: cardiology	9
o 4th most frequent: pregnancy & childbirth	5
– Patients with various medical conditions	14 (14%)
o Most frequent: unselected primary care patients	6
o Non-health care	10 (10%)
o Most frequent: general population	2
Studies on target conditions AAD/GAD	72 (70%) / 64
Prevalence of target conditions	(62%)
– AAD: median (Quartile 1/3)	18% (13%/29%)
– GAD: median (Quartile 1/3)	10% (5%/23%)

*including five bi- or multilingual studies, including English; **including two bi- or multilingual studies, including Spanish.

the meta-analyses for both target conditions, 39 (38%) could only be included for AAD and 31 (30%) only for GAD. For most studies, 2×2 tables were re-constructed from pairs of sensitivity and specificity, the prevalence of the target condition, and the number of participants in the analysis. Based on the published reports alone, the number of diagnostic 2×2 tables was 560 (40% of those analyzed after author requests) for AAD and 404 (32%) for GAD. The median number of diagnostic 2×2 tables obtained from publications ranged from 0 (for the GAD-2, mainly because results were only presented for the GAD-7, although any study investigating the GAD-7 could also investigate the GAD-2), and 8 across questionnaires. Eleven (11%) studies had published pairs of sensitivity and specificity for the full or a large range of cut-off points.

4. Discussion

4.1. Summary of main findings

Our analysis of 103 DTA studies, which examined eight widely used anxiety assessment questionnaires, shows that reporting on methods and results is often incomplete and contains several significant methodological shortcomings. In particular, many studies provided insufficient information on participant recruitment and on whether patients with a known current anxiety disorder or who were undergoing treatment for such a disorder were excluded. Reporting was also often unclear regarding the independence of the index test and reference standard assessment. Furthermore, diagnostic accuracy results were frequently presented selectively for a single or a few cut-offs only.

4.2. Assessing risk of bias and applicability - comparison with other systematic reviews

Our assessment of the risk of bias differs significantly from that presented in a recent systematic review of 45 DTA studies on GAD-7 and GAD-2 by Park and Park [22]. These authors also applied QUADAS-2 but considered the overall risk of bias to be low for 21 studies (47%). In the index test and flow and timing domains, all studies were judged to have low risk of bias in the reference standard domain 73%, and in the patient selection domain 65%. Such a positive assessment is unusual in systematic reviews of DTA studies (see, for example, [23–25]). Park and Park neither report ratings for the individual studies nor provide their rating guidelines. While their study sample does not exactly match our GAD-7/ GAD-2 sub-sample (due to slightly different eligibility criteria and a probably less sensitive search), it seems unlikely that these differences alone explain the differences in the assessments. Although QUADAS-2 provides a well-established assessment framework [21], rating guidelines need to be tailored to the specific field and research question. In developing our assessment guidelines, we largely followed the approach of the DEPRESSD Project (Depression Screening Data; [26]), a pioneering project aimed at synthesizing global data on depression screening. Park and Park's judgments appear substantially more permissive than those typically observed in DTA reviews and those obtained with our tailored criteria. For example, they judged the risk of bias low for all studies in the index test domain "because the GAD scale was self-assessed" [22]. Selective reporting of favorable outcomes – a central issue in the index test domain and evident in a substantial proportion of studies – does not appear to have been addressed.

Similar to Park and Park [22], we had little applicability concerns regarding the index test and the target condition (as defined by the reference standard). The questionnaires examined in the studies would be (and are) used in the same or a similar way in true screening situations, and validated (semi-)structured diagnostic interviews are considered adequate for making anxiety diagnoses. However, we had significantly more concerns than Park and Park regarding patient selection. This is probably because (following a debate within the team) we took the following problem more seriously: Screening is usually conducted to identify previously unrecognized cases. When studies aiming to assess the DTA of an anxiety questionnaire for screening purposes include individuals who currently have a known anxiety disorder or are being treated for such a disorder, this affects the prevalence, positive and negative predictive values, and might also influence sensitivity and specificity. This issue has been intensively discussed for depression screening studies [27,28]. As with depression, many DTA studies (and reviews) on anxiety claim to examine screening but actually focus on the detection of anxiety disorders in a mixed sample of individuals with both unknown and known disease status.

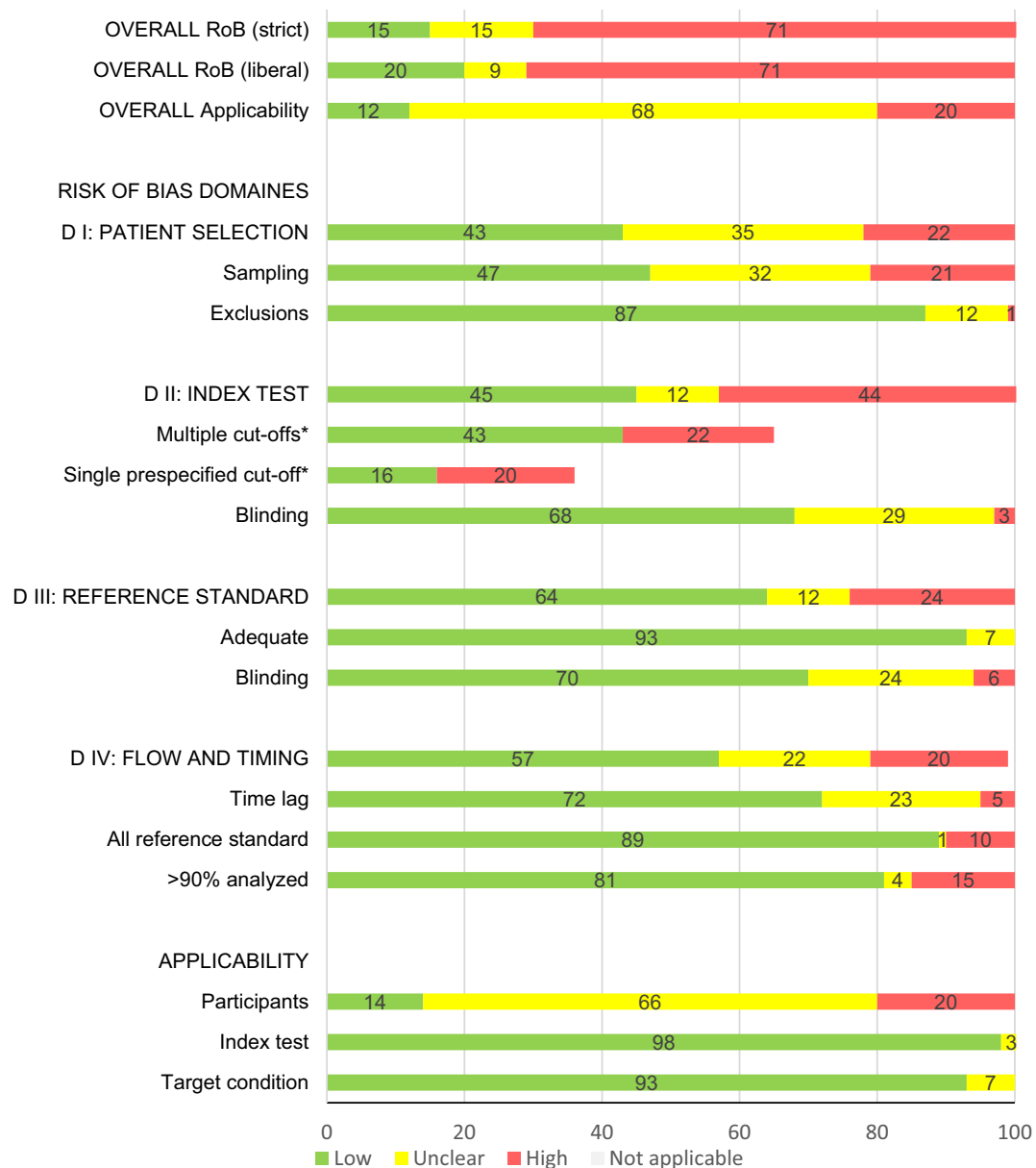


Fig. 1. Risk of bias and applicability assessment (percentage of studies; $n = 103$) rated as low, unclear or high: overall assessments, in domains (capital letters) and for the respective individual signaling questions. *In the domain 'index test' EITHER the question on 'single pre-specified cut-off' OR on the presentation of 'multiple cut-offs' could be rated.

Table 2

Data reporting: median/sum number of cut-offs (CO) for which diagnostic data were reported in original publications (CO published) and which could be used in meta-analyses after obtaining additional information from authors (CO analyzed).

Questionnaire (Scale range)	Target condition AAD			Target condition GAD		
	Studies	CO published median/sum	CO analyzed median/sum	Studies	CO published median/sum	CO analyzed median/sum
HADS (0–21)	44	1/174	7/413	27	1/90	11/305
GAD-7 (0–21)	25	2/92	18/366	37	1/140	13/470
GAD-2 (0–6)	22	0/35	6/113	26	1/42	6/130
BAI (0–63)	7	8/107	27/192	5	3/62	29/132
STAI-S (20–80)	7	1/44	1/45	1	–/31	–/31
STAI-T (20–80)	5	6/72	6/98	1	–/31	–/31
PROMIS-ASF8a (8–40)	5	2/32	25/123	4	1/5	25/100
OASIS (0–20)	3	2/4	16/48	3	1/3	16/47
Total	72	n.a./560	n.a./1398	64	n.a./404	n.a./1246

n.a. = not applicable (due to the highly variable scale ranges).

4.3. Reporting of methods and methodological issues – recommendations for future studies

One key area where improvement in reporting of methods in anxiety screening studies is needed concerns participant recruitment and the transparency of the patient flow. Almost half of the studies we analyzed reported using ‘consecutive’ or ‘random’ sample selection. However, as patient flow charts or detailed descriptions were missing in the majority of studies, it usually remained unclear how many of the people approached did not participate and why. Although we rated this sampling approach appropriate under our coding rules, doubts remained given the sometimes high non-participation rates in studies with adequate reporting. Despite a similarly generous assessment in the flow and timing domain, we were often unsure whether the figures relating to individuals included in the study but not in the analysis were truly reliable. Reporting guidelines clearly recommend that all publications of DTA studies include a flow diagram detailing the number of people approached, the frequency and reasons for non-participation, the number of participants included, the number of completed index tests and interviews, and the number of participants included in the final analysis [29]. Future studies should adhere to this recommendation. This is important not only for a more reliable assessment of the risk of bias in future systematic reviews, but also for a better understanding of whether the study populations are representative of, and transferable to, routine clinical screening practice.

If studies are genuinely focused on evaluating diagnostic test accuracy (DTA) in a screening setting, individuals with a known current anxiety disorder or who are undergoing treatment for such a disorder should be excluded. In many studies, however, this does not appear to have occurred, presumably to facilitate recruitment. If such individuals are included, consideration should be given to conducting an additional analysis excluding them.

Since it is of such fundamental importance that the completion of the questionnaire and the diagnostic interview be conducted strictly independently of one another, the measures taken to ensure ‘blinding’ should be described in detail. If the accuracy of several index tests (questionnaires) is to be compared, it should be considered whether their order should be randomized, as it is unclear whether and how completing the individual tests influences one another. Furthermore, the time lag between the index test and the reference standard should be kept as short as possible and reported accurately.

4.4. Reporting of study findings – recommendations for future studies

The second key area where we see a particularly great need for improvement is the reporting of study findings. A first problem is that many studies have reported prevalence and DTA findings only for a single target condition (typically AAD or GAD), even though the clinical interviews often captured a broader range of specific anxiety disorders. While comprehensive reporting of findings for specific disorders in the main text may not be practical due to space constraints and low case numbers, this would be easily possible in appendices. At the very least, the prevalence of the specific anxiety diagnoses assessed should always be reported to characterize the sample.

A second important issue is that in most studies, the DTA is reported for only one or a few threshold values. Only 11 studies (11%) had published DTA findings across a broad range or reported all possible cut-offs. About a quarter of studies focused on a more or less clearly pre-defined, previously recommended ‘main’ cut-off (and possibly a few surrounding cut-offs). But in the majority, the best-performing cut-off(s) were selected post-hoc using the Youden-Index or similar approaches. These ‘optimal’ thresholds could vary significantly from one study to another. The selective publication of DTA results for variable ‘optimal’ cut-offs is not only questionable from the perspective of clinical interpretation but also poses a major problem for meta-analyses conducted using the currently standard bivariate model. With the bivariate model

[30], a meta-analysis can only be performed for one cut-off point per study. Reviewers are therefore faced with an unsatisfactory choice: either they include only those studies that have published results for a specific cut-off point and, as far possible, conduct separate meta-analyses for other specific cut-offs - this is the usual approach (see, for example, the review by Plummer et al. [23] on GAD-7 and GAD-2); or they combine study results for different cut-off points (see, for example, in the review by Park and Park [22], where GAD-7 cut-offs ranged from ≥ 4 to ≥ 12). In both cases, however, there is a significant risk that the resulting summary estimates are biased [31].

In recent years, new statistical methods have been developed that allow for meta-analysis of all available data, taking into account the level of the cut-off points (e.g., [32]). It has been shown empirically that such models are less susceptible to reporting bias [33]. Above all, however, these models can estimate the DTA at all possible cut-off points of a scale, and the more comprehensively the results of the individual studies are reported, the more reliable these estimates are. This is particularly important for the planning of screening programs in routine care. Traditional methods for identifying ‘optimal’ cut-offs for screening, both in the area of depression [34] and anxiety, focus on balancing sensitivity and specificity. However, as Brehaut et al. [34] wrote, there is no rationale “why a balance between these test characteristics would best select people most likely to engage and benefit from mental health services and reduce, as much as possible, the resources needed to find them.” Optimal cut-offs are likely to differ between screening settings (e.g., primary care versus specialized epilepsy outpatient clinics), reflecting differences in the relative importance placed on ruling in versus ruling out a condition, the acceptability of false-negative results that may lead to missed cases, and the resources required to investigate false-positive findings. Therefore, reporting findings for a wide range of cut-offs should become standard practice for DTA studies.

Finally, we would like to mention an important overall impression that emerged as we compiled our reviews: The currently available DTA studies consistently lacked a genuine connection to the use of questionnaires as screening tools in daily clinical practice. Neither the question of what rates of false-positive or false-negative results would be acceptable, nor the consequences of the screening results in routine practices, were addressed or discussed in depth. At the same time, however, the studies did not address whether anxiety questionnaires could be useful as part of the standard diagnostic process in daily practice. None of the included studies examined whether the use of these questionnaires makes physicians' diagnoses more reliable. Future research on anxiety questionnaires should evaluate DTA not only in terms of formal validation but also in relation to its practical implications for clinical decision-making, taking into account the specific factors and trade-offs discussed above.

4.5. Strengths and limitations

The main strength of our analysis lies in the large number of primary studies included. Although not all existing questionnaires used for screening purposes were investigated, it seems unlikely that the issues identified differ fundamentally from those in DTA studies of other questionnaires.

Nevertheless, several limitations should be taken into account when interpreting our findings. First, we applied QUADAS-2, the most widely used tool for assessing the risk of bias and the applicability of DTA studies. Yet, the comparison of our assessment with that of Park and Park [22] highlights that QUADAS-2 still permits considerable subjective judgment. Second, our recommendations focus on issues addressed in our assessments and related to reporting of results. They do not cover all issues relevant to DTA studies on anxiety screening and they are based primarily on the experiences made in our reviews. Third, because our reviews included only screening studies, our recommendations also apply exclusively to such studies. We excluded “case-finding” studies

that examined the questionnaires as a diagnostic aid for identifying cases among individuals were already suspected of having an anxiety disorder or seeking help at a mental health facility. Evidence from studies across different conditions indicates that higher disease prevalence tends to be associated with increased sensitivity and decreased specificity [35]. Forth, our manuscript focused on the methodological quality and reporting of the currently available DTA studies. Adherence to these recommendations should enable future research to generate findings with greater clinical relevance and methodological rigor. Clinicians should remain mindful that the existing evidence base is constrained by significant methodological limitations. As a detailed examination of the diagnostic test accuracy (DTA) of individual screening questionnaires and their clinical implications falls outside the scope of the present article, interested readers are directed to the corresponding systematic reviews [8–13].

5. Conclusion

In conclusion, most published anxiety-screening DTA studies are limited by incomplete reporting, risk of bias concerns, and highly selective presentation of cut-off-specific results. As a result, they are less informative for clinical and health policy decision-making than they could be. Most of the reporting and methodological shortcomings seem easily avoidable. Future studies should align with the STARD reporting recommendations [29]. Beyond the recommendations provided by STARD, our article offers additional guidance specifically targeted at improving the design, conduct, and reporting of DTA studies of anxiety screening. Authors of DTA studies should bear in mind that their studies are likely to be included in systematic reviews and meta-analyses, which in turn are of particular importance for the development of guidelines and screening recommendations.

Declaration of generative AI use

We used DeepL (<https://www.deepl.com/de/translator>) and M365 Copilot for checking and correcting the language in our manuscript.

CRedit authorship contribution statement

Klaus Linde: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Alexey Fomenko:** Investigation, Methodology, Writing – review & editing. **Zekeriya Aktürk:** Conceptualization, Investigation, Methodology, Writing – review & editing. **Stefanie Eck:** Investigation, Writing – review & editing. **Alexander Hapfelmeier:** Formal analysis, Methodology, Writing – review & editing. **Antonius Schneider:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Ethics approval and consent to participate

Not applicable

Funding

The review project was funded by a grant (01KG2105) from the German Federal Ministry of Research, Technology and Space, www.bmfr.bund.de

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Antonius Schneider reports financial support was provided by German Federal Ministry of Research, Technology and Space. If there are other

authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We would like to thank all the authors who provided us with additional information about their studies. We would also like to thank Tithi Patel for her assistance with correcting the language in our manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jpsychores.2026.112967>.

Data availability

The full dataset is available from the corresponding author on reasonable request.

References

- [1] K.L. Szuhany, N.M. Simon, Anxiety disorders: a review, *JAMA* 328 (24) (2022) 2431–2445, <https://doi.org/10.1001/jama.2022.22744>.
- [2] GBD 2019 Mental Disorders Collaborators, Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, *Lancet Psychiatry* (2) (2022) 137–150, [https://doi.org/10.1016/S2215-0366\(21\)00395-3](https://doi.org/10.1016/S2215-0366(21)00395-3).
- [3] US Preventive Services Task Force, M.J. Barry, W.K. Nicholson, M. Silverstein, et al., Screening for anxiety disorders in adults: US Preventive Services Task Force recommendation statement, *JAMA* 329 (24) (2023) 2163–2170, <https://doi.org/10.1001/jama.2023.9301>.
- [4] American College of Obstetricians and Gynecologists, Screening and diagnosis of mental health conditions during pregnancy and postpartum (Clinical Practice Guideline No. 4), *Obstet. Gynecol.* 141 (6) (2023) 1232–1261.
- [5] M.P. Austin, P. Middleton, N.M. Reilly, N.J. Highet, Detection and management of mood disorders in the maternity setting: the Australian Clinical Practice Guidelines, *Women Birth* 26 (1) (2013) 2–9, <https://doi.org/10.1016/j.wombi.2011.12.001>.
- [6] A.Y. Mughal, J. Devadas, E. Ardman, B. Levis, V.F. Go, B.N. Gaynes, A systematic review of validated screening tools for anxiety disorders and PTSD in low to middle income countries, *BMC Psychiatry* 20 (1) (2020) 338, <https://doi.org/10.1186/s12888-020-02753-3>.
- [7] M. Corso, K. Martin, L.R. Salmi, et al., Screening tools for ruling out mood and anxiety disorders in adults in primary care: a rapid systematic review, *BMC Prim. Care* 27 (1) (2025) 36, <https://doi.org/10.1186/s12875-025-03132-1>.
- [8] Z. Aktürk, A. Hapfelmeier, A. Fomenko, et al., Generalized Anxiety Disorder 7-item (GAD-7) and 2-item (GAD-2) scales for detecting anxiety disorders in adults, *Cochrane Database Syst. Rev.* 3 (3) (2025) CD015455, <https://doi.org/10.1002/14651858.CD015455>.
- [9] A. Fomenko, D. Dümmler, Z. Aktürk, et al., Hospital Anxiety and Depression Scale Anxiety subscale (HADS-A) for detecting anxiety disorders in adults, *Cochrane Database Syst. Rev.* 7 (7) (2025) CD015456, <https://doi.org/10.1002/14651858.CD015456>.
- [10] S. Eck, D. Dümmler, Z. Aktürk, et al., Beck Anxiety Inventory (BAI) for detecting anxiety disorders in adults, *Cochrane Database Syst. Rev.* 12 (12) (2025) CD015457, <https://doi.org/10.1002/14651858.CD015457>.
- [11] D. Dümmler, S. Eck, A. Hapfelmeier, et al., State-Trait Anxiety Inventory (STAI) for detecting anxiety disorders in adults, *Cochrane Database Syst. Rev.* 12 (12) (2025) CD015458, <https://doi.org/10.1002/14651858.CD015458>.
- [12] A. Fomenko, F.F. Fischer, R.A. Marrie, et al., The screening accuracy of the PROMIS® anxiety measures in adults – a systematic review and multiple-thresholds meta-analysis, *J. Psychosom. Res.* 205 (2026) 112611, <https://doi.org/10.1016/j.jpsychores.2026.112611>.
- [13] A. Fomenko, R.A. Marrie, D. Dümmler, et al., Test accuracy of the Overall Anxiety severity and impairment Scale (OASIS) for screening anxiety in adults – systematic review and multiple thresholds meta-analysis, *J. Affect. Disord.* 400 (2026) 121126.
- [14] K. Linde, M. Olm, C. Teusen, et al., The diagnostic accuracy of widely used self-report questionnaires for detecting anxiety disorders in adults (protocol), *Cochrane Database Syst. Rev.* 9 (2022) CD015292, <https://doi.org/10.1002/14651858.CD015292>.
- [15] A.S. Zigmond, R.P. Snaith, The hospital anxiety and depression scale, *Acta Psychiatr. Scand.* 67 (6) (1983) 361–370.
- [16] R.L. Spitzer, K. Kroenke, J.B. Williams, B. Löwe, A brief measure for assessing generalized anxiety disorder: the GAD-7, *Arch. Intern. Med.* 166 (10) (2006) 1092–1097, <https://doi.org/10.1001/archinte.166.10.1092>.

- [17] A.T. Beck, N. Epstein, G. Brown, R.A. Steer, An inventory for measuring clinical anxiety: psychometric properties, *J. Consult. Clin. Psychol.* 56 (6) (1988) 893–897, <https://doi.org/10.1037//0022-006x.56.6.893>.
- [18] C.D. Spielberger, R.L. Gorsuch, R.E. Lushene, *Manual for the State-Trait Anxiety Inventory*, Consulting Psychologists' Press, Inc, Palo Alto, California, 1970.
- [19] M. Bevans, A. Ross, D. Cella, Patient-Reported Outcomes Measurement Information System (PROMIS): efficient, standardized tools to measure self-reported health and quality of life, *Nurs. Outlook* 62 (5) (2014) 339–345, <https://doi.org/10.1016/j.outlook.2014.05.009>.
- [20] S.B. Norman, S.H. Cissell, A.J. Means-Christensen, M.B. Stein, Development and validation of an Overall Anxiety Severity And Impairment Scale (OASIS), *Depress. Anxiety* 23 (4) (2006) 245–249, <https://doi.org/10.1002/da.20182>.
- [21] P.F. Whiting, A.W. Rutjes, M.E. Westwood, et al., QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies, *Ann. Intern. Med.* 155 (8) (2011) 529–536, <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>.
- [22] S.H. Park, S.K. Park, An updated systematic review and meta-analysis of the predictive validity of the general anxiety disorder (GAD)-7 and GAD-2 in screening for anxiety disorders, *J. Affect. Disord.* 391 (2025) 119913, <https://doi.org/10.1016/j.jad.2025.119913>.
- [23] F. Plummer, L. Manea, D. Trepel, D. McMillan, Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis, *Gen. Hosp. Psychiatry* 39 (2016) 24–31, <https://doi.org/10.1016/j.genhosppsych.2015.11.005>.
- [24] A.W. Rutjes, J.B. Reitsma, M. Di Nisio, N. Smidt, J.C. van Rijn, P.M. Bossuyt, Evidence of bias and variation in diagnostic accuracy studies, *CMAJ* 174 (4) (2006) 469–476, <https://doi.org/10.1503/cmaj.050090>.
- [25] L. Faes, P.M. Bossuyt, L.M. Bachmann, M. Nagler, Design-related bias in studies investigating diagnostic tests for venous thromboembolic diseases: a systematic review and meta-analysis, *Front. Cardiovasc. Med.* 11 (2024) 1420000, <https://doi.org/10.3389/fcvm.2024.1420000>.
- [26] B. Levis, A. Benedetti, B.D. Thombs, 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* 365 (2019) 11476, <https://doi.org/10.1136/bmj.11476>.
- [27] B.D. Thombs, E. Arthurs, G. El-Baalbaki, A. Meijer, R.C. Ziegelstein, R.J. Steele, 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* 343 (2011) d4825, <https://doi.org/10.1136/bmj.d4825>.
- [28] D.B. Rice, B.D. Thombs, Risk of bias from inclusion of currently diagnosed or treated patients in studies of depression screening tool accuracy: a cross-sectional analysis of recently published primary studies and meta-analyses, *PLoS One* 11 (2) (2016) e0150067, <https://doi.org/10.1371/journal.pone.0150067>.
- [29] P.M. Bossuyt, J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L. Irwig, et al., STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies, *BMJ* 351 (2015) h5527, <https://doi.org/10.1136/bmj.h5527>.
- [30] J.B. Reitsma, A.S. Glas, A.W. Rutjes, R.J. Scholten, P.M. Bossuyt, A.H. Zwinderman, Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews, *J. Clin. Epidemiol.* 58 (10) (2005) 982–990, <https://doi.org/10.1016/j.jclinepi.2005.02.022>.
- [31] D. Neupane, B. Levis, P.M. Bhandari, B.D. Thombs, A. Benedetti, 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, *Int. J. Methods Psychiatr. Res.* 30 (3) (2021) e1873, <https://doi.org/10.1002/mpr.1873>.
- [32] S. Steinhauser, M. Schumacher, G. Rücker, Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies, *BMC Med. Res. Methodol.* 16 (1) (2021) 97, <https://doi.org/10.1186/s12874-016-0196-1>.
- [33] A. Benedetti, B. Levis, G. Rücker, et al., An empirical comparison of three methods for multiple cutoff diagnostic test meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool using published data vs individual level data, *Res. Synth. Methods* 11 (2020) 833–848, <https://doi.org/10.1002/jrsm.1443>.
- [34] E. Brehaut, D. Neupane, B. Levis, Y. Wu, Y. Sun, J.P.A. Ioannidis, S. Markham, P. Cuijpers, S.B. Patten, A. Benedetti, B.D. Thombs, 'Optimal' cutoff selection in studies of depression screening tool accuracy using the PHQ-9, EPDS, or HADS-D: a meta-research study, *Int. J. Methods Psychiatr. Res.* 32 (3) (2023) e1956, <https://doi.org/10.1002/mpr.1956>.
- [35] M.H. Murad, L. Lin, H. Chu, B. Hasan, R.A. Alsibai, A.S. Abbas, R.A. Mustafa, Z. Wang, The association of sensitivity and specificity with disease prevalence: analysis of 6909 studies of diagnostic test accuracy, *CMAJ* 195 (27) (2023) E925–E931, <https://doi.org/10.1503/cmaj.221802>.