Appendix 1: SIGN Checklist for Diagnostic Accuracy

SIGN

Methodology Checklist 5: Studies of Diagnostic Accuracy

This checklist is based on the work of the QUADAS2 team at Bristol University (http://www.bris.ac.uk/quadas/).

Study identification (*Include author, title, reference, year of publication*)

Guideline topic: Screening/Assessment of Cancer-related Fatigue (CRF)	Key Question No:

Before completing this checklist, consider:

1. Is the paper really a study of diagnostic accuracy? It should be comparing a specific diagnostic test against another, and **not** a general paper or comment on diagnosis.
2. Is the paper relevant to key question? Analyse using PICO (Patient or Population Intervention Comparison Outcome). IF NO REJECT (give reason below). IF YES complete the checklist..

Reason for rejection: Reason for rejection: 1. Paper not relevant to key question □ 2. Other reason □ (please specify):

Checklist completed by:

All the questions in the following sections have associated footnotes providing short explanations behind each of the questions. Users who want more detailed explanations should consult the QUADAS-2: Background Document.

DOMAIN 1 – PATIENT SELECTION

Risk of bias

In a well conducted diagnostic study…		***Is that true in this study?***	
1.1	A consecutive sequence or random selection of patients is enrolled.[1]	Yes ☐ No ☐	Can't say ☐
1.2	Case – control methods are not used.[2]	Yes ☐ No ☐	Can't say ☐
1.3	Inappropriate exclusions are avoided.[3]	Yes ☐ No ☐	Can't say ☐

Applicability

1.4	The included patients and settings match the key question.[4]	Yes ☐ No ☐	Can't say ☐

DOMAIN 2 – INDEX TEST

Risk of bias

In a well conducted diagnostic study…		***Is that true in this study?***	
2.1	The index test results interpreted without knowledge of the results of the reference standard.[5]	Yes ☐ No ☐	Can't say ☐
2.2	If a threshold is used, it is pre-specified.[6]	Yes ☐ No ☐	Can't say ☐

Appendix 1: SIGN Checklist for Diagnostic Accuracy

Applicability			
2.3	The index test, its conduct, and its interpretation is similar to that used in practice with the target population of the guideline.[7]	Yes ☐ No ☐	Can't say ☐
DOMAIN 3 – REFERENCE STANDARD			
Risk of bias			
In a well conducted diagnostic study…		***Is that true in this study?***	
3.1	The reference standard is likely to correctly identify the target condition.[8]	Yes ☐ No ☐	Can't say ☐
3.2	Reference standard results are interpreted without knowledge of the results of the index test.[9]	Yes ☐ No ☐	Can't say ☐
Applicability			
3.3	The target condition as defined by the reference standard matches that found in the target population of the guideline.[10]	Yes ☐ No ☐	Can't say ☐
DOMAIN 4 – FLOW AND TIMING			
Risk of bias			
In a well conducted diagnostic study…		***Is that true in this study?***	
4.1	There is an appropriate interval between the index test and reference standard.[11]	Yes ☐ No ☐	Can't say ☐
4.2	All patients receive the same reference standard.[12]	Yes ☐ No ☐	Can't say ☐
4.3	All patients recruited into the study are included in the analysis.[13]	Yes ☐ No ☐	Can't say ☐
SECTION 5: OVERALL ASSESSMENT OF THE STUDY			
5.1	How well was the study done to minimise bias? *Code as follows*[14]	High quality (++)☐ Acceptable (+)☐ Low quality (-)☐ Unacceptable – reject 0 ☐	
5.2	What is your assessment of the **applicability** of this study to our target population?	Directly applicable ☐ Some indirectness ☐ (Please explain in the following section for **Notes**)	
5.2	**Notes.** Summarise the authors conclusions. Add any comments on your own assessment of the study, and the extent to which it answers your question.		

Appendix 1: SIGN Checklist for Diagnostic Accuracy

1. Studies should enrol either all eligible patients suspected of having the target condition during a specified period, or a random sample of those patients. The essential point is that investigators should have no freedom of choice as to which individual patients are or are not included.
2. There is evidence that studies comparing patients with known disease with a control group without the condition tend to exaggerate diagnostic accuracy.
3. Inappropriate exclusions may result in either overestimates (eg by excluding 'difficult to diagnose' patients) or underestimates (eg by excluding patients with 'red flags' suggesting presence of disease) of the degree of diagnostic accuracy.
4. Patients included in the study should match the target population of the guideline in terms of severity of the target condition, demographic features, presence of differential diagnosis or co-morbidity, setting of the study and previous testing protocols.
5. This is similar to the question of 'blinding' in intervention studies. The index test should always been done first, or by a separate investigator with no knowledge of the outcome of the reference test.
6. Bias can be introduced if a threshold level is set after data has been collected. Any minimum threshold should be specified at the start of the trial.
7. Variations in test technology, execution, or interpretation (eg use of a higher ultrasound transducer frequency) may affect estimates of diagnostic accuracy.
8. Estimates of test accuracy are based on the assumption that the reference standard is 100% sensitive (=accurately diagnoses the target condition).
9. This is the similar to question 2.1, but in this case relates to making sure the reference standard is applied without any prior knowledge of the outcome of previous tests.
10. The definition of the target condition used when testing the reference standard may differ from that used by the NHS in Scotland. eg threshold levels used in laboratory cultures may differ.
11. The index test and reference standard should be performed as close together in time as possible, otherwise changes in the patients condition is likely to invalidate the results.
12. In some cases the choice of reference standard may be influenced by the outcome of the index test or the urgency of the need for diagnosis. Use of different reference standards is likely to lead to overestimates of both sensitivity and specificity.
13. Not including all patients in the analysis may lead to bias as there may be some systematic difference between those lost to follow-up and those analysed.
14. Rate the overall methodological quality of the study, using the following as a guide:
    - High quality (++): Majority of criteria met. Little or no risk of bias. Results unlikely to be changed by further research.
    - Acceptable (+): Most criteria met. Some flaws in the study with an associated risk of bias, Conclusions may change in the light of further studies.
    - Low quality (-): Either most criteria not met, or significant flaws relating to key aspects of study design. Conclusions likely to change in the light of further studies.

Appendix 2: COSMIN Boxes 3 and 4 – Structural Validity and Internal Consistency

Box 3. Structural validity

Does the scale consist of effect indicators, i.e. is it based on a reflective model? [1] yes / no

Does the study concern unidimensionality or structural validity? [2] unidimensionality / structural validity

Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factor analysis performed	Not applicable
2 For IRT/Rasch: does the chosen model fit to the research question?	Chosen model fits well to the research question	Assumable that the chosen model fits well to the research question	Doubtful if the chosen model fits well to the research question	Chosen model does not fit to the research question	Not applicable
3 Was the sample size included in the analysis adequate?	FA: 7 times the number of items and ≥100	FA: at least 5 times the number of items and ≥100; OR at least 6 times number of items but <100	FA: 5 times the number of items but <100	FA: < 5 times the number of items	
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models: 100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: < 50 subjects	
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokken scale analysis: 500-999 subjects	2PL parametric IRT models OR Mokken scale analysis: 250-499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects	

Other						
4	Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws (e.g. rotation method not described)	Other important methodological flaws (e.g. inappropriate rotation method)	

[1] If the scale is not based on a reflective model, unidimensionality or structural validity is not relevant.
[2] In a systematic review, it is helpful to make a distinction between studies where factor analysis is performed on each (sub)scale separately to evaluate whether the (sub)scales are unidimensional (unidimensionality studies) and studies where factor analysis is performed on all items of an instrument to evaluate the (expected) number of subscales in the instrument and the clustering of items within subscales (structural validity studies).

Appendix 2: COSMIN Boxes 3 and 4 – Structural Validity and Internal Consistency

Box 4. Internal consistency Does the scale consist of effect indicators, i.e. is it based on a reflective model? [1] yes / no					
Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 Was an internal consistency statistic calculated for each unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale		Unclear whether scale or sub scale is unidimensional	Internal consistency statistic NOT calculated for each unidimensional scale or sub scale	
Statistical methods					
2 For continuous scores: Was Cronbach's alpha or omega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not applicable
3 For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not applicable
4 For IRT-based scores: Was standard error of the theta (SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient calculated	Not NOT applicable
Other					
5 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

[1] If the scale is not based on a reflective model, internal consistency is not relevant

Appendix 3: Levels of Evidence

Level of Evidence	Intervention/ prevention	Pathoanatomic/ risk/ clinical course/ prognosis/ differential diagnosis	Diagnosis/ diagnostic accuracy	Prevalence of condition/ disorder	Exam/ outcomes
I. Evidence obtained from high-quality systematic reviews, diagnostic studies, prospective studies, or randomized controlled trials (RCTs)	Systematic review of high-quality RCTs (a) High-quality RCT (a)	Systematic review of prospective cohort studies High-quality prospective cohort study (b)	Systematic review of high-quality diagnostic studies High-quality diagnostic study with validation (c)	Systematic review of high-quality cross-sectional studies High-quality cross-sectional study (d)	Systematic review of prospective cohort studies High-quality prospective cohort study
I. Evidence obtained from lesser-quality diagnostic studies, prospective studies, or RCTs (eg, weaker diagnostic criteria and reference standards, improper randomization, no blinding, less than 80% follow-up)	Systematic review of high-quality cohort studies High-quality cohort study (b) High-quality outcomes research High-quality quasi-experimental study (g) High-quality Single subject design (h) Lower-quality RCT (e)	Systematic review of retrospective cohort studies Lower-quality prospective cohort study High-quality retrospective cohort study Consecutive cohort study Outcomes study or ecological study (f)	Systematic review of exploratory diagnostic studies or consecutive cohort studies High-quality exploratory diagnostic study Consecutive retrospective cohort study	Systematic review of studies that allows relevant estimate Lower-quality cross-sectional study	Systematic review of lower-quality prospective cohort studies Lower-quality prospective cohort study

Appendix 3: Levels of Evidence

III. Case-con-trolled studies or retrospective studies	Systematic review of case-controlled studies High-quality case-controlled study Outcomes study or ecological study (f) Lower-quality cohort study	Lower-quality retrospective cohort study High-quality cross sectional study Case- controlled study	Lower-quality exploratory diagnostic study Nonconsecutive retrospective cohort study	Local nonrandom study	High-quality cross-sectional study
IV. Case series	Case series	Case series	Case- controlled study		Lower-quality cross-sectional study
V. Expert opinion	Expert opinion	Expert opinion	Expert opinion	Expert opinion	Expert opinion

Appendix 4 – Grade Assignments for Level of Evidence Recommendations("APTA Clinical Practice Guideline Process Manual," 2018)

Letter Grade	Level of Obligation	Definition
A	Strong	A high level of certainty of *moderate to substantial* benefit, harm or cost, or a *moderate* level of certainty for *substantial* benefit, harm or cost (based on a preponderance of Level 1 or 2 evidence with at least 1 level 1 study)
B	Moderate	A high level of certainty of *slight to moderate* benefit, harm or cost, or a *moderate* level of certainty for a *moderate* level of benefit, harm or cost (based on a preponderance of level 2 evidence, or a single high quality RCT)
C	Weak	A moderate level of certainty of *slight* benefit, harm or cost, or a weak level of certainty for moderate to substantial benefit, harm, or cost (based on Level 2 thru 5 evidence)
D	Theoretical / foundational	A preponderance of evidence from animal or cadaver studies, from conceptual/theoretical models/principles, or from basic science/bench research, or published expert opinion in peer-reviewed journals that supports the recommendation
P	Best practice	Recommended practice based on current clinical practice norms, exceptional situations in which validating studies have not or cannot be performed yet there is a clear benefit, harm or cost, expert opinion
R	Research	An absence of research on the topic or disagreement among conclusions from higher-quality studies on the topic

APTA Clinical Practice Guideline Process Manual. (2018). In. Alexandria, VA: American Physical Therapy Association.