Evaluation of Screening Tests for PTSD

10

Scott D. McDonald, Emily L. Gentes, and Patrick S. Calhoun

Contents

Abstract

Screening programs are an essential component of preventative medicine for conditions of public health importance. For a post-traumatic stress disorder

S.D. McDonald (✉)
Psychology Section (116-B), McGuire VA Medical Center, Mental Health Service, Richmond VA, USA
e-mail: scott.mcdonald@va.gov

E.L. Gentes
Durham VA Medical Center, VA Mid-Atlantic Mental Illness Research, Education, and Clinical Center, Durham, NC, USA
e-mail: emily.gentes@va.gov

P.S. Calhoun
Durham VA Medical Center, Durham, NC, USA
e-mail: patrick.calhoun2@va.gov

© Springer International Publishing Switzerland (outside the USA) 2016

C.R. Martin et al. (eds.), *Comprehensive Guide to Post-Traumatic Stress Disorders*,
DOI 10.1007/978-3-319-08359-9_50

(PTSD) screening program to be effective, valid screening tests are essential. Thus, it is essential for clinicians to be aware of how to evaluate studies of screening test utility for applicability and validity. This chapter (1) reviews the principal components of a diagnostic accuracy study for a screening test, (2) describes the method for finding appropriate diagnostic accuracy studies for a screening test of interest, and (3) outlines how to evaluate the quality and applicability of diagnostic accuracy studies for PTSD screening tests.

List of Abbreviations

CAPS	Clinician Administered PTSD Scale
CI	Confidence interval
CPG	Clinical practice guideline
DSM-5	Diagnostic and Statistical Manual of Mental Disorders
LR-	Negative likelihood ratio
LR+	Positive likelihood ratio
PCL	PTSD Checklist
PC-PTSD	Primary care PTSD screen
PTSD	Post-traumatic stress disorder

Introduction

Screening programs are an essential component of preventative medicine for conditions of public health importance. Behavioral health problems such as depression, bipolar disorder, and substance abuse are leading causes of disability (World Health Organization 2008), and early identification can reduce personal and societal economic costs (U.S. Preventive Services Task Force 2009). Screening for post-traumatic stress disorder (PTSD) has been recommended in many settings in which risk of trauma exposure is high such as cancer treatment centers and among at-risk patients, such as combat veterans (Andrykowski et al. 1998; Department of Veterans Affairs and Veterans Health Administration 2004; Freedy et al. 2010; National Collaborating Centre for Mental Health 2005; New York State Department of Health 2006).

For a PTSD screening program to be effective, valid screening tests are essential (UK National Screening Committee 2013; Wilson and Jungner 1968). There are many PTSD screening tests available (Table 1), the majority symptom-based and presented as brief questionnaires or clinical interviews. However, there are limitations in the PTSD screening literature, most notably few replication studies, modest evidence for generalizability, and biases due to deficient study design (Brewin 2005; McDonald et al. in press; McDonald and Calhoun 2010; Spoont et al. 2013). Unfortunately, these limitations may ultimately lead to ineffective screening programs, misguided allocation of clinical resources, and faulty interpretations of epidemiologic and research findings. Thus, it is essential for clinicians to be aware of how to evaluate studies of screening test utility for applicability and validity.

Table 1 Common PTSD measures used as screening tests. This table provides a list of common PTSD measures that have diagnostic accuracy information available

Test	Number of items	Time to administer	Fee for use
Beck Anxiety Inventory – Primary Care (BAI-PC)	7	Not reported	Yes
Davidson Trauma Scale (DTS)	17	10’	Yes
M-3 Checklist	23	<5’	No
Primary Care PTSD Screen (PC-PTSD)	4	<5’	No
Screen for Post-Traumatic Stress Symptoms (SPTSS)	17	Not reported	
Short forms of the PTSD Checklist	2 and 6	<5’	No
Short Screening Scale for PTSD	7	5’	No
Startle, Physiological arousal, Anger, and Numbness (SPAN)	4	<5’	Yes
Short Post-Traumatic Stress Disorder (PTSD) Rating Interview (SPRINT)	8	5–10’	No
Trauma Screening Questionnaire (TSQ)	10	Not reported	No
PTSD Checklist (PCL)	17	5–10’	No

A few details about the scope and approach to this chapter should be mentioned. First, available measures have been reviewed elsewhere (Brewin 2005; McDonald and Calhoun 2010; National Center for PTSD 2014; Spoont et al. 2013) and will only be mentioned here as case examples. Also, this chapter will focus on brief screening tools that can be used efficiently in a busy clinic and as such will not review the growing number studies of possible genetic, biomarker, psychophysiological, and metabolic testing procedures that could provide a tool for PTSD screening. Furthermore, this chapter will focus on screening tests that detect current PTSD rather than those that predict the development of PTSD from prodromal symptoms during the acute stress phase. In summary, this chapter will define “screening tests,” describe the method for testing the diagnostic accuracy of screening tests, and outline an approach for evaluating diagnostic accuracy of screening tests for PTSD.

What Is a Screening Test?

A *screening test* is a brief and efficient tool, administered to asymptomatic or at-risk patients, to provide a probable diagnosis that is confirmed by a subsequent diagnostic procedure (Table 2). The terms “diagnostic test” and “screening test” are sometimes used interchangeably or are differentiated by the comparative rigor and precision of a diagnostic procedure. However, it is useful to define them by the intended use: a screening test is administered to asymptomatic or at-risk individuals to identify cases that may have a condition, whereas diagnostic tests are for confirmation of a condition’s presence (Streiner 2003). Screening tests may be part of a *universal screening* program in which all individuals in a particular setting or category are tested or may be administered only to individuals with risk factors, called *case finding*. Mental health screening tests tend to involve patient self-report

Table 2 Key facts about PTSD screening tests. This table lists the key facts about PTSD screening tests, including purpose, features and format, and scoring

Administered to asymptomatic or at-risk individuals to identify patients who have a probable PTSD diagnosis
Brief, inexpensive, and easy to administer and score
Safe, acceptable to the patient, and easy to understand
Self-report questionnaire or interview format
Generally scored using a threshold to determine probable PTSD diagnosis

via clinical interview or questionnaires, in contrast to the procedures typically found in other areas of medicine (e.g., vital signs, blood tests). They are often face-valid and directly measure diagnostic criteria but may also tap behaviors, counterfactual information, or other factors (e.g., asking about alcohol consumption to infer potential abuse). Although screening tests may offer information on the likelihood of a diagnosis across a range of scores (e.g., scores on a multi-item questionnaire), it is often useful to determine a threshold for a positive screening and a procedure for further clinical assessment, feedback to the patient, and options for treatment.

A good screening test is inexpensive, safe, easy to administer and score, acceptable to the patient, and uses language that is easy to understand. The test should have good reliability and must be internally valid, meaning that the results should be attributable only to the construct of interest rather than to other sources of influence. In contrast, it is not essential that a screening test demonstrate face validity or appear on its face to assess the construct of interest. The overarching measure of the utility of a screening test is how well the screening test accurately and efficiently detects those with the target condition and rules out those who do not have the condition.

Diagnostic Accuracy of a Screening Test

Diagnostic accuracy refers to the degree of agreement between the screening test and a *gold* or *reference standard* that represents the best available indicator of the presence or absence of the condition of interest (Bossuyt et al. 2003a). In the USA, the definitions presented in the Diagnostic and Statistical Manual of Mental Disorders (*DSM*) define mental disorders, leaving it up to the clinician to determine through interview and collateral information whether specific criteria for a disorder are met. In psychiatric research, structured diagnostic interviews are generally used as reference standards, as their standardized approach provides a common set of proscribed questions that leads to optimal diagnostic reliability across raters.

Performance Characteristics and Accuracy of Screening Tests

Various indices are used to characterize the precision of a screening test. A 2 × 2 table, such as the one shown in Table 3, is a simple way to illustrate the

Table 3 Classification of results from a validity study of a screening test. This table illustrates the relationship between the results of a screening test and the diagnosis per a reference standard, such as a structured interview for PTSD. Counts of cases in each cell are used to calculate diagnostic accuracy indices such as sensitivity and specificity

	Reference standard	
Result of screening test	Present	Absent
Present	True positive	False positive
Absent	False negative	True negative

Table 4 Common diagnostic accuracy indices. Commonly used indices of diagnostic accuracy and their calculation

Index	Description	Calculation
Hit rate	Overall classification rate	(True positive + true negative)/total
Sensitivity	Proportion of those *with* the disorder who are correctly identified by the test	True positive/(true positive + false negative)
Specificity	Proportion of those *without* the disorder who are correctly identified by the test	True negative/(true negative + false positive)
Positive predictive power (PPP)	Proportion of those screening *positive* who *have* the condition	True positive/(true positive + false positive)
Negative predictive power (NPP)	Proportion of those screening *negative* who *do not have* the condition	True negative/(true negative + false negative)

relationship between screening test performance and true diagnosis per reference standard. Some of the most commonly used are overall hit rate, sensitivity, specificity, and positive/negative predictive power. These are described below, with calculations shown in Table 4. The overall *hit rate* or *efficiency* of a test is defined as the ratio of the sum of true positives and true negatives by the total number of cases in the sample. A problem with the hit rate is that it does not account for agreement that may occur simply by chance, inflating the apparent accuracy of the test. Cohen's *kappa* is one common example of an adjusted index of overall accuracy that corrects for chance agreement.

Sensitivity is the probability that a patient with a condition will have a positive screening test result. *Specificity* is the probability that a patient who does not have the condition will have a negative test result. Sensitivity and specificity are fixed properties of the test, such that they do not change as long as the test is used within the same population.

Positive and negative predictive power is the converse of sensitivity and specificity. *Positive predictive power* is the proportion of those screening positive that actually have the condition. *Negative predictive power* is the proportion of those screening negative that actually do not have the condition. Unlike sensitivity and specificity, positive predictive power and negative predictive power are dependent on the prevalence or base rate of the condition in the population being tested.

The most prevalent strategy for interpreting results from a screening test with continuous scores is the use of a cutoff score, above which the test is considered positive. A cutoff score is typically identified as the score that optimally distinguishes individuals who have the disorder from those who do not. For example, the PTSD Checklist (PCL), a PTSD symptom severity questionnaire, includes 17 items rated on a scale from 1 ("not at all") to 5 ("extremely"). In the initial validation study, a cutoff score of 50 was identified as optimal for indicating a probable diagnosis of combat-related PTSD (Weathers et al. 1993). When using a cutoff score, there is a tradeoff between sensitivity and specificity that is dependent on the cutoff score that is employed. That is, a lower cutoff score will result in a higher number of individuals screening positive (i.e., sensitivity will improve but specificity will decline). The "optimal" cutoff score therefore will depend in large part on the intended purpose of the test. For example, a lower cutoff score may be most useful if it is desirable to ensure that no true positive cases are missed.

Variability in Performance Characteristics Across Studies

Screening tests do not have inherent and universal performance characteristics. Evidence for this is easy to find, as considerable variation in diagnostic accuracy across studies for each PTSD screening test has been reported. For example, the performance characteristics of the PCL, Civilian Version, varied considerably in one review of 10 studies: at the oft-recommended cut score of 50, sensitivity ranged from.20 to.86 and specificity ranged from.67 to.93 (McDonald and Calhoun 2010). Such diversity in classification accuracy across studies could result in misguided interpretation if used indiscriminately and without an understanding of *why* studies might have such diverse results. Which begs the question, what drives this diversity of outcomes across studies of the same measure? Three sources of variability are highlighted here: differences between two populations, differences between a study sample and its target population or setting, and imprecision (Fig. 1).

Differences between two populations. Two studies may have different results because they drew their samples from different populations. *Variation* between populations such as demographic composition, clinical acuity, context of evaluation (e.g., treatment or disability), and clinical setting may impact a test's performance characteristics. These factors also influence the *generalizability* of a diagnostic accuracy study's results, that is, the degree to which they apply to other populations.

Differences between a study sample and its target population or setting. Variation in sample, context, or setting between two diagnostic accuracy studies targeting the same population is likely to generate disparate performance characteristics. Nonrandom sampling and other problems with study design may result in *bias*, that is, systematic differences between the study sample's performance characteristics and the population's "true" performance characteristics (Ransohoff and Feinstein 1978; Whiting et al. 2013). For example, a study that selects only those patients who are asymptomatic or have severe symptomatology will have inflated diagnostic accuracy, due to the relative ease of correctly classifying extreme cases. In

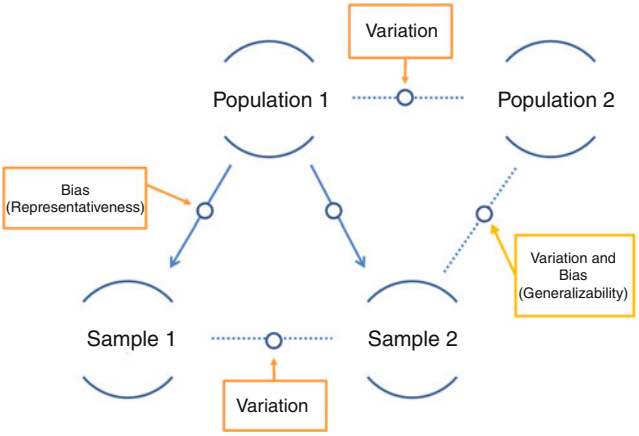

Fig. 1 Relationship between sources of variation and bias. This figure illustrates the relationship between sources of variation and bias

the case of diagnostic accuracy studies, the performance characteristics (e.g., accuracy, sensitivity, and specificity) are biased to the degree that they systematically differ from those of the target population. The degree of bias is inversely related to the *representativeness* of the study's results to the target population.

Imprecision. Even under the best controlled conditions, some random error is unavoidable and contributes to variation in performance characteristics of a screening test across studies. Random error contributes to *precision* of measures, most often expressed as confidence intervals (CIs) for sensitivity, specificity, and other performance characteristics. Precision can be optimized by improving design features such as the reliability of measures and by increasing sample size.

In summary, population-level differences, bias from patient selection, study design, and precision are sources of variability among diagnostic accuracy studies. It is essential for clinicians and investigators to consider the presence and source of these factors when evaluating diagnostic accuracy studies that may inform their work. The next section provides additional detail and examples on how these sources of variability affect diagnostic accuracy study outcomes.

Practices and Procedures

Finding an Applicable Diagnostic Accuracy Study of a PTSD Screening Test

A variation of the PICO format (Population, Index test, Comparator or reference standard, and Outcomes) can assist clinicians to develop a focused question that will guide them toward appropriate diagnostic accuracy studies (Schmidt and Factor 2013). This process involves determining relevant characteristics of the population of interest, desired screening test qualities, acceptable reference standard, and

desired outcome (e.g., is it accurate enough and applicable?). PICO can be used for both clinical and research applications (e.g., systematic reviews). The goal at this stage is to conduct a search of the literature through scanning title and abstract to suggest potential applicability to your patient population and setting (Jaeschke et al. 1994a).

Clinical practice guidelines (CPGs), systematic reviews, and nonsystematic reviews offer useful information about the evidence base for screening tests. Systematic reviews, such as those listed in the Cochrane Library (srdta.cochrane.org), generally provide the best overview of the evidence base and often inform CPGs and policy. Several reviews of PTSD screening tests are available (Brewin 2005; McDonald et al. in press; McDonald and Calhoun 2010; Spoont et al. 2013). Another source of diagnostic accuracy studies can be found through abstracting and indexing databases such as subscription-based PsycINFO and the freely available PubMed (www.ncbi.nlm.nih.gov). Google Scholar (scholar.google.com) can also be helpful, although search results include articles from a variety of sources that include online journals and self-published articles that do not utilize peer review as a quality measure.

Evaluating a Diagnostic Accuracy Study for a PTSD Screening Test

Once a relevant diagnostic accuracy study is identified, a critical appraisal informs applicability and validity. Despite the availability of guidance for the *conduct* of diagnostic accuracy studies (Reid et al. 1995; Reitsma et al. 2009), the *quality* of studies has been criticized as frequently deficient (Bossuyt 2008; Reid et al. 1995; Smidt et al. 2006). Reviews of diagnostic accuracy studies of PTSD screening tests have also found commonplace deficits, most notably the use of nonrepresentative sampling, insufficient description of the sample's clinical characteristics, and lack of detailed, transparent reporting of key aspects of study procedures (McDonald et al. in press; Spoont et al. 2013).

One of the difficulties in critically evaluating diagnostic accuracy studies is that *reporting quality* is often inadequate. In response to this problem, the STARD initiative outlined a framework for the examination of reporting quality of diagnostic accuracy studies (Bossuyt et al. 2003a, b). The STARD initiative also developed a useful and freely available 25-item checklist to assist the evaluation of reporting quality (www.stard-statement.org). Although reporting quality has improved since the STARD initiative, and the STARD statement has been adopted by over 200 journals, several elements are still routinely omitted or not adequately described, including mention of inclusion criteria, method for identifying eligible patients, details of screening test and reference standard administration, and the handling of uninterpretable results (Bossuyt 2008). There are no existing studies of reporting quality for PTSD diagnostic accuracy studies.

Guidance for the *evaluation* of diagnostic accuracy studies is available for many medical specialties (Jaeschke et al. 1994b; Reitsma et al. 2009; Warner 2004) and readily applies to studies of screening tests. As might be expected, key, superordinate

Table 5 Key considerations for evaluating a diagnostic accuracy study for a PTSD screening test. This table lists key considerations for evaluating a diagnostic accuracy study for a PTSD screening test

Was the study's patient sample representative of the patients who will receive the test?
What are the differences in context between the research study real-world clinical practice?
Are biases avoided in the study design, procedures, and analyses?
Was an appropriate reference standard used?
Do results have enough precision to be useful?
Does the study inform your practice?

domains for evaluation include applicability, bias in study design, and precision. The QUADAS-2 tool (www.bris.ac.uk/quadas/quadas-2) assists a reader to evaluate applicability and risk of bias in a diagnostic accuracy study (Whiting et al. 2011). Although designed specifically for selecting quality studies for inclusion in systematic reviews, QUADAS-2 can be an asset for clinicians and researchers who are evaluating a diagnostic accuracy study of a PTSD screening test.

Below are a series of specific questions that will assist the reader in assessing the quality of a diagnostic accuracy study for a PTSD screening test (Table 5). For each question, a rationale is provided within the context of PTSD assessment, as well as an example, when available. It is important to note that there is no consensus regarding the relative importance of these quality indicators, as different factors will be more relevant in some circumstances than others. That said, the Cochrane Collaboration has indicated that key domains include representativeness, an unbiased verification procedure, blinding, and handling of missing data (Reitsma et al. 2009).

Was the Study's Patient Sample Representative of the Patients Who Will Receive the Test?

A common deficit in diagnostic accuracy studies occurs when the population of interest is not well defined and/or the sample is not representative of the target population (Mann et al. 2009; McDonald et al. in press; Wilczynski 2008). This is a critical issue, as variation in patient *spectrum* such as demographics (age, gender, etc.), comorbidities, socioeconomic factors, and condition prevalence and severity can have a substantial impact on a screening test's performance characteristics (Lijmer et al. 1999; Whiting et al. 2013; Willis 2008). Studies that use convenience samples, have restrictive inclusion/exclusion criteria, or recruit beyond the target population are at risk of *selection bias*, that is, practices that result in systematic differences between the study sample and the target population.

Few studies have examined the impact of selection bias on PTSD screening tests, although there is some evidence that certain subgroup differences, comorbidity, and PTSD prevalence affect performance characteristics. For example, Freedy and colleagues (2010) found that optimal cut scores and performance characteristics of several PTSD screening tests differed by sex (Freedy et al. 2010). Although sex did not affect the optional cut score for the primary care PTSD screen (PC-PTSD) (Prins

et al. 2003), the operating characteristics did differ (e.g., sensitivity for men = 1.00, women = 0.83). In another study, the PC-PTSD offered a higher hit rate for military veterans over 35 years old (.91) than younger veterans (.79) (Calhoun et al. 2010). Regarding the impact of mental health morbidity, the Davidson Trauma Scale (Davidson et al. 1997) was considerably more effective at discriminating between US military veterans with PTSD and those with no mental health disorder than it was for discriminating between those with PTSD vs. other mental disorders. The reduced specificity of the test in the mental disorder sample may reflect higher scores due to subthreshold PTSD or elevations on relatively nonspecific symptoms such as difficulty concentration and anhedonia.

The impact of PTSD prevalence (i.e., the base rate) on a screening test's performance characteristics has been well described. A screening test's accuracy is reduced to the degree that the true prevalence is greater or less than 50 % (Meehl and Rosen 1955). Terhakopian and colleagues illustrated this property, by demonstrating that false-positive screens are more likely on the PTSD Checklist when PTSD prevalence was low and false-negative screens are more likely when PTSD prevalence was high (Terhakopian et al. 2008). Using a cut score of 44 on the PTSD Checklist and a weighted average sensitivity and specificity across 14 studies, the authors found that PTSD prevalence was overestimated when the true prevalence was 15 % or less and underestimated when true prevalence was 35 % or higher.

In summary, a screening test's validity and applicability is dependent on the degree of similarity between the sample and the target population. Ideally, a diagnostic accuracy study employs random sampling (e.g., randomly selected or every *n*-th patient appointment in a primary care clinic) to ensure that the sample is representative of the target population. A comparison of demographics, PTSD prevalence, and other factors between sample and population is a first step in evaluating representativeness. Unfortunately, the precise population-level description of patient spectrum and prevalence of PTSD is rarely available, making it difficult for readers to evaluate how well a sample represents a target clinical population (McDonald and Calhoun 2010). It is sometimes preferable to score screening test results by subgroup (e.g., sex) to provide the most accurate outcomes.

What are the Differences in Context Between the Research Study and Real-Word Clinical Practice?

The impact of the setting, context of testing, and interpersonal factors such as the test administrator's interpersonal and communication skills are likely more salient for diagnostic accuracy research in psychiatry than for many other areas of medicine. Whereas a patient has limited ability to impact the outcomes of a radiologic procedure or tissue biopsy, most psychiatric screening and diagnostic tests involve self-report and as such are dependent on a patient's willingness to provide thoughtful and accurate responses. For example, patient's responses to a traumatic stress screen could be very well influenced by the context such as the reason for evaluation (e.g., self-referred for treatment, litigation, mandatory military post-deployment screening), confidentiality of results, rapport with the clinician, and other such factors (McLay et al. 2008; Rona et al. 2005; Rubenzer 2009). Even just the knowledge of

being evaluated within a research context can change behavior (McCambridge et al. 2012, 2014). Further, clinical staff administering tests may be more engaged and likely to adhere to test instructions during the study but revert to less standardized practices after the study is completed. A final consideration is that a patient's willingness to participate in a research study, including trust in the research team, desirability of compensation, transportation difficulties, and available time, may impact representativeness without influencing demographic composition. In sum, a *context bias* resulting from differences between the clinical and research settings must also be considered as a potential impediment to representativeness.

Are Biases Avoided in the Study Design, Procedures, and Analyses?

In addition to selection and context biases, other study design factors may lead to biases, most notably:

- **Was the same reference standard administered to all those receiving the screening test?** Structured interviews for PTSD do not have perfect reliability or concordance. As a result, the performance characteristics of screening tests will likely differ depending on what diagnostic test is used. When multiple diagnostic tests are used interchangeably as reference standards within one study, *differential verification bias* may occur. Diagnostic accuracy studies should use the same reference standard for all participants.
- **Was the reference standard administered to all patients, or a random selection of patients, who received the screening test?** All patients, regardless of screening test performance, should receive the reference standard. *Partial verification bias* may occur when, on the basis of screening test performance, only a subset of patients receives the reference standard. For example, only administering the reference standard to those scoring above a particular threshold on the screening test (assuming low scorers do not have PTSD) does not allow assessment of false negatives below the threshold, potentially artificially inflating concordance. In samples that have low rates of PTSD, it may be useful to administer the reference standard to only every *n*-th patient below a prespecified threshold and then weight the sample accordingly for analyses. This strategy will avoid partial verification bias while improving study efficiency.
- **Was the time between the index test and the reference standard sufficiently short enough to assume no change in the condition between tests?** The performance of a screening test will degrade as the time between the test and the reference standard increases. PTSD symptoms can worsen (i.e., *progression bias*), improve with treatment, or randomly fluctuate over time. Ideally, both the screening test and the reference standard would be conducted on the same day to determine the test's best possible performance. When that is not possible, a maximum of about 2–4 weeks difference has been suggested (McDonald et al. in press), which would put the administration of the screening test within range of the window of symptom endorsement on most structured PTSD interviews. It should be acknowledged that in real-world clinical practice, the diagnostic interview does not usually occur on the same day and potential changes in

PTSD presentation between tests should be considered by the clinician. Regardless, the duration between tests and the order they were given should be reported for a diagnostic accuracy study.

- **Were the reference standard and the screening test interpreted without knowledge of each other and other clinical data?** Masking, or "blinding," protects against artificially inflated agreement between tests due to *review bias*, which occurs when knowledge of the result of one test influences the outcome of the other. For example, the administrator of a reference standard might employ additional probes during a structured diagnostic interview if a positive result on the screening test was known. Masking also minimizes subtle or overt influence of investigator impartiality toward a screening test that can impact concordance. Finally, masking to additional clinical information such as comorbidity, psychosocial functioning, and prior treatment minimizes inflated concordance due to changes in administration or interpretation of tests secondary to this knowledge. Of note, the masking of screening test results and clinical information optimizes internal validity (i.e., the true concordance of the tests) but reduces ecological validity (i.e., the degree to which the study reflects real-world practices), perhaps leading to overestimation of diagnostic accuracy when applied to clinical settings (McDonald et al. 2014).
- **Was the reference standard independent of the screening test?** To avoid *incorporation bias*, the screening test should not constitute a part of the reference standard. Incorporating the result of the screening test or part of the screening test (e.g., establishing trauma history) into the diagnostic formulation would be an example of incorporation bias. This is an uncommon issue in PTSD diagnostic accuracy research, as the reference standard is commonly a stand-alone structured diagnostic interview.
- **Were missing and uninterpretable results reported and explained?** Early withdrawals, incomplete data, and uninterpretable results may introduce biases. For example, if more symptomatic patients tend to not attend the session in which the reference standard was administered, the results would not reflect the population. Uninterpretable results of the diagnostic test may occur because of invalid interviews due to psychosis, suspicion of malingering, or other factors. In any case, these numbers should be reported in addition to how the cases were handled in analyses.

Were the author's conclusions about "optimal" cut scores supported? The reporting of operating characteristics for a wide range of cut scores for tests with a continuous scale of measurement allows the reader to evaluate the investigator's methods and conclusions. Indices of diagnostic accuracy do not always agree regarding the optimal cut score, so a wide range of cut scores provides a fuller picture of results. For example, in a diagnostic accuracy study of the Davidson Trauma Scale, a cut score in the 89–92 range produced the highest hit rate but cut scores in the 68–72 range produced the highest *kappa*. Additionally, having a wide range of cut scores and their operating characteristics available allows the reader to select cut scores that fit their need. For example, a clinical screening program will

want a threshold with high sensitivity, whereas an investigator may want high specificity to ensure participants are likely to have PTSD.

Was an Appropriate Reference Standard Used?

The choice of reference standard is critical for a diagnostic accuracy study, as it defines the condition of interest. Generally, a structured diagnostic interview, most commonly the Clinician Administered PTSD Scale (CAPS), is cited as the "gold" standard in the diagnosis of PTSD (National Center for Posttraumatic Stress Disorder 2014; Weathers et al. 1999). There are several advantages of using structured interview in PTSD research, including reliability across raters and standardization of terms and well-defined diagnostic criteria. However, a drawback is that structured diagnostic interviews are rarely used in actual clinical practice (Aboraya 2009), begging the uncomfortable question of whether the research based on diagnoses from structured interviews applies to real-world clinical practice. This is in contrast to most other areas of medicine, in which the best available diagnostic test is routinely employed in clinical practice as well as research. Other reference standard options include the (less than preferable) clinician diagnosis, combining multiple test results, and relying on consensus diagnosis (Bertens et al. 2013). Readers should consider the extent of the reference test's validation, its characteristics in the population of interest, the training and qualifications of the administrators, and whether the structured interview was administered using procedures intended by the measure's developers.

Do Results Have Enough Precision to Be Useful?

Confidence intervals (CIs) should be reported for a screening test's performance characteristics to inform the reader about precision of measurements (Bossuyt et al. 2003a). Random error contributes to a statistic's precision, which improves with a larger sample size. Reporting of CIs for diagnostic accuracy studies across medical journals have been limited (Smidt et al. 2006) and nearly nonexistent in the PTSD screening test literature (McDonald and Calhoun 2010). This is unfortunate, as they are relatively easy to calculate (Carley et al. 2005) and improve the comparability of results across studies. There are no agreed-upon thresholds for what is considered an acceptable CI, but readers should consider whether the precision of a test as indicated by a study is reasonable for the intended purpose.

Does the Study Inform Your Practice?

Once it is determined that a study is applicable to your population, is of good quality, and offers adequate precision, the operating characteristics will inform utility. There is no consensus regarding what constitutes an acceptable overall classification hit rate, although arbitrary performance ranges for *kappa* have been suggested: 0.01–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect (Landis and Koch 1977). Beyond overall classification, it is important to consider the user's requirement for detecting

PTSD: is it enough to capture 75 % of those with PTSD (i.e., sensitivity = .75) or is 95 % required? Since specificity decreases with increased sensitivity, the increase in false positives with increased sensitivity needs to be considered as well. Thus, the screening test user must consider whether clinical resources are sufficient to provide second-level clinical assessment for all patients with a positive screening. With knowledge of the PTSD base rate, sensitivity, and specificity, the user can estimate the proportion of patients who will screen positive and require follow-up care: [(sensitivity × base rate) + (1 − specificity) × (1 − base rate)]. When planning one's approach, the clinician should also consider that although a positive screening does not necessarily lead to a diagnosis of PTSD, it may reflect subthreshold trauma symptom or other clinically meaningful emotional distress. Finally, it is important to consider how screening test feedback will be presented to patients, particularly when false-positive rates are high.

Concluding Remarks

Evidence-based practice in trauma is adjusting to the new DSM-5 criteria for PTSD and related disorders. Accordingly, this is an opportune time to recalibrate or create new PTSD screening tests and improve screening programs. Limitations of prior work, such as nonrepresentative sampling and study design deficits, can be rectified moving forward. In addition, further work is needed to clarify how specific aspects of patient spectrum, testing context, and study methods impact a PTSD screening test's diagnostic accuracy and thus its utility in supporting effective and efficient patient care.

Summary Points

- Screening programs are an essential component of preventative medicine for conditions of public health importance.
- A screening test is a brief and efficient tool, administered to asymptomatic or at-risk patients, to provide a probable diagnosis that is confirmed by a subsequent diagnostic procedure.
- For a PTSD screening program to be effective, valid screening tests are essential.
- Diagnostic accuracy refers to the degree of agreement between the screening test and a reference standard or "gold standard" that represents the best available indicator of the presence or absence of the condition of interest.
- Issues such as nonrepresentative sampling and deficits in study design introduce bias, limiting the validity and thus utility of a diagnostic accuracy study.
- Factors that are particularly important to diagnostic accuracy study in psychiatry include the study context, the masking of test results, and selection of a reference standard.
- It is essential for clinicians to be aware of how to evaluate studies of screening test utility for applicability and validity.

References

Aboraya A. Use of structured interviews by psychiatrists in real clinical settings: results of an open-question survey. Psychiatry (Edgmont (Pa: Township)). 2009;6:24–8.

Andrykowski MA, Cordova MJ, Studts JL, Miller TW. Posttraumatic stress disorder after treatment for breast cancer: prevalence of diagnosis and use of the PTSD Checklist-Civilian Version (PCL-C) as a screening instrument. J Consult Clin Psychol. 1998;66:586–90.

Bertens LC, Broekhuizen BD, Naaktgeboren CA, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10:e1001531.

Bossuyt PMM. STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiology. 2008;248:713–4.

Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Radiol. 2003a;58:575–80.

Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003b;49:7–18.

Brewin CR. Systematic review of screening instruments for adults at risk of PTSD. J Trauma Stress. 2005;18:53–62.

Calhoun PS, McDonald SD, Guerra VS, et al. Clinical utility of the primary care-PTSD screen among U.S. veterans who served since September 11, 2001. Psychiatry Res. 2010;178:330–5.

Carley S, Dosman S, Jones SR, Harrison M. Simple nomograms to calculate sample size in diagnostic studies. Emerg Med J. 2005;22:180–1.

Davidson JRT, Book SW, Colket JT, et al. Assessment of a new self-rating scale for post-traumatic stress disorder. Psychol Med. 1997;27:153–60.

Department of Veterans Affairs, Veterans Health Administration. Implementation of a new national clinical reminder, the "Afghan and Iraq Post-deployment Screen". Washington, DC: Author; 2004.

Freedy JR, Steenkamp MM, Magruder KM, et al. Post-traumatic stress disorder screening test performance in civilian primary care. Fam Pract. 2010;27:615–24.

Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1994a;271:389–91.

Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA. 1994b;271:703–7.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.

Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061–6.

Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. Soc Psychiatry Psychiatr Epidemiol. 2009;44:300–7.

McCambridge J, de Bruin M, Witton J. The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. PLoS One. 2012;7:e39116.

McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. J Clin Epidemiol. 2014;67:267–77.

McDonald SD, Calhoun PS. The diagnostic accuracy of the PTSD checklist: a critical review. Clin Psychol Rev. 2010;30:976–87.

McDonald SD, Thompson ML, Stratton KJ, Calhoun PS, The VA Mid-Atlantic Mental Illness Research, Education, Clinical Center (MIRECC) Workgroup. Diagnostic accuracy of three scoring methods for the Davidson Trauma Scale among U.S. military veterans. J Anxiety Disord. 2014;28:160–8.

McDonald SD, Brown WL, Benesek JP, Calhoun PS. A systematic review of the PTSD checklist's diagnostic accuracy studies using QUADAS. Psychol Trauma Theory Res Pract Policy. In press.

McLay RN, Deal WE, Murphy JA, Center KB, Kolkow TT, Grieger TA. On-the-record screenings versus anonymous surveys in reporting PTSD. Am J Psychiatry. 2008;165:775–6.
Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychol Bull. 1955;52:194–216.
National Center for Posttraumatic Stress Disorder. Clinician-Administered PTSD Scale for DSM-5 (CAPS-5). 2014;2014(July 30). http://www.ptsd.va.gov/professional/assessment/adult-int/caps.asp. Accessed 30 July 2014.
National Center for PTSD. PTSD screening instruments. 2014. http://www.ptsd.va.gov/professional/assessment/screens/index.asp. Accessed 1 May 2014.
National Collaborating Centre for Mental Health. Post-traumatic stress disorder: the management of PTSD in adults and children in primary and secondary care. London: National Institute for Clinical Excellence (NICE); 2005.
New York State Department of Health. Prevention of secondary disease: mental health care. New York: New York State Department of Health; 2006.
Prins A, Ouimette PC, Kimerling R, et al. The primary care PTSD screen (PC-PTSD): development and operating characteristics. Prim Care Psychiatry. 2003;9:9–14.
Ransohoff D, Feinstein A. Problems of spectrum bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926–30.
Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA. 1995;274:645–51.
Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy version 1.0.0. The cochrane collaboration, 2009. Available from: http://srdta.cochrane.org/.
Rona RJ, Hyams KC, Wessely S. Screening for psychological illness in military personnel. JAMA. 2005;293:1257–60.
Rubenzer S. Posttraumatic stress disorder: assessing response style and malingering. Psychol Inj Law. 2009;2:114–42.
Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Arch Pathol Lab Med. 2013;137:558–65.
Smidt N, Rutjes AWS, van der Windt D, et al. The quality of diagnostic accuracy studies since the STARD statement – Has it improved? Neurology. 2006;67:792–7.
Spoont M, Arbisi P, Fu S, et al. Screening for post-traumatic stress disorder (PTSD) in primary care: a systematic review (VA-ESP Project #09-009). Washington, DC: Department of Veterans Affairs, Health Services Research & Development Services; 2013.
Streiner DL. Diagnosing tests: using and misusing diagnostic and screening tests. J Pers Assess. 2003;81:209–19.
Terhakopian A, Sinaii N, Engel CC, Schnurr PP, Hoge CW. Estimating population prevalence of posttraumatic stress disorder: an example using the PTSD checklist. J Trauma Stress. 2008;21:290–300.
UK National Screening Committee. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme. 2013. http://www.screening.nhs.uk/criteria. Accessed 20 May 2014.
U.S. Preventive Services Task Force. Screening for depression in adults: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2009;151:784–92.
Warner J. Clinicians' guide to evaluating diagnostic and screening tests in psychiatry. Adv Psychiatr Treat. 2004;10:446–54.
Weathers F, Litz B, Herman D, Huska J, Keane T. The PTSD Checklist (PCL): reliability, validity, and diagnostic utility. Poster session presented at the Annual Convention of the International Society for Traumatic Stress Studies, San Antonio, Oct 1993.
Weathers FW, Ruscio AM, Keane TM. Psychometric properties of nine scoring rules for the Clinician-Administered Posttraumatic Stress Disorder Scale. Psychol Assess. 1999;11:124–33.

Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529–536.
Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013;66:1093–104.
Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication – before-and-after study. Radiology. 2008;248:817–23.
Willis BH. Spectrum bias-why clinicians need to be cautious when applying diagnostic test studies. Fam Pract. 2008;25:390–6.
Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: WHO; 1968.
World Health Organization. The global burden of disease: 2004 update. Geneva: World Health Organization; 2008.