Short Screening Tools for Cancer-Related Distress: A Review and Diagnostic Validity Meta-Analysis

Alex J. Mitchell, MSc, MRCPsych, Leicester, United Kingdom

Key Words
Distress, depression, cancer, diagnostic accuracy, screening

Abstract
Clinicians are increasingly seeking efficient methods to identify distress in cancer settings, using short screening tools with fewer than 14 items that take less than 5 minutes to complete. This article examines the value of these tools for identifying cancer-related distress, defined by semi-structured interview. An updated search, appraisal, and meta-analysis, with adjustments made for heterogeneity and underlying prevalence variations, identified 45 potentially useful short and ultra-short tools, although most were intended to help diagnose depression, with few targeted at distress (or anxiety). Very few studies attempted robust validation in cancer settings. When studies were limited to those tested against distress defined by semi-structured interview, only 6 methods had been validated, namely the Hospital Anxiety and Depression Scale (HADS; 13 studies, 14 items), the Distress Thermometer (DT; 4 studies, 1 item), a single verbal question (4 studies, 1 item), the Psychological Distress Inventory (PDI; 1 study, 13 items), combined DT and an impact thermometer (1 study, 2 items), and combined 2 verbal questions (1 study, 2 items). Comparing these 6 approaches side-by-side suggests that for screening, all tools have approximately the same accuracy. Therefore, choice of a short screening tool for distress can be based on acceptability or cost-effectiveness. Here, best evidence supports use of the DT or single verbal question. Remarkably, the overall accuracy of these single-item approaches seems comparable to that of the 14-item HADS (total score), whereas their efficiency is superior. For case-finding, data are sparse but no method seems to be entirely satisfactory. Current evidence suggests that the optimal short methods for identifying distress are 2 verbal questions or PDI. Of these approaches, the 2 verbal questions has superior efficiency. All short methods may be augmented by repeated application, an assessment of unmet needs (problem list), and clarification regarding the need for professional help. No screening tool should be seen as an alternative to careful clinical assessment and management. Despite much interest in the development of short and ultra-short tools, data on validation and implementation are currently incomplete. Nevertheless, short methods seem to be at least as successful as the HADS, although substantially more efficient and hence more acceptable, and therefore may be a suitable initial method of assessment in busy clinical settings. (JNCCN 2010;8:487–494)

Perhaps the most important development in psychooncology in the past 10 years has been the development and testing of short, user-friendly screening tools for distress. Attempts to validate these tools have helped crystallize the concept of distress, which had previously received little attention compared with depression. Distress is a very common complication of cancer at any stage and often occurs when multiple needs are unmet. The presence of distress is also linked with reduced health-related quality of life, poor satisfaction with medical care, and possibly reduced survival. Early psychometric research focused on diagnostic accuracy (performance against another scale) and diagnostic validity (performance against a true criterion standard) of longer tools involving 10 items or more that typically took at least 5 minutes to administer. It is now known that fewer than 15% of cancer professionals are prepared to use these questionnaires in clinical practice, with most relying on their own clinical judgement. It is also known that self-assessed judgement is inaccurate, just as accuracy in vignette studies overestimates clini-
clinical performance. Direct observation of clinicians’ habits suggests modest clinical accuracy with a high proportion of both false-negative and positive errors. These errors are seen in several professional groups and in all settings, leading to the interpretation that distress can be difficult to detect in busy clinical settings.

Patient-reported information such as distress and quality of life can be evaluated through interview (verbally), pencil and paper, or computer. Most questionnaires use a multiple-choice format with a Likert scale (e.g., mild, moderate, severe), but some use a visual analog format. This review and brief meta-analysis examines the merits of short tools with 5 to 14 items taking 2 to 5 minutes to complete, and ultra-short screening tools of 1 to 4 items taking less than 2 minutes to complete.

The first application of a visual analog scale (VAS) was in 1921, although it was not used to measure mood until 1969. In 1969, Zealley and Aitken compared patient and nurse twice-daily VAS ratings. A high correlation was found between the patients’ self-rated VAS score and their Hamilton score, although interestingly nurses’ VAS ratings tended to cluster centrally and lag behind patient-reported change.

In the 1970s, several independent groups experimented with VAS ratings for mood, suicidal thoughts, pain, and quality of life. The development of simple but structured verbal methods has been a more recent development, given that verbal enquiry may be more acceptable than even a single self-report item (e.g., “are you depressed?” or “how distressed have you been in the previous week?”). In cancer settings, use of the VAS for emotional difficulties has followed from its use in pain measurement (e.g., in the Memorial Pain Assessment Card). In 1997, Chochinov et al. examined a VAS from “worst possible mood” to “best possible mood.” However, the best known example and most adequately studied is the Distress Thermometer (DT), which was developed by a panel of 23 health professionals along with a patient representative working in collaboration with the NCCN. The DT is a simple pencil and paper measure consisting of a line with a 0-to-10 scale, indicating “No Distress” at 0, and at 10, indicating “Extreme Distress.” The words no distress, moderate distress, and severe distress are present as anchors along the scale in some versions. Patients are asked to answer the question “How distressed have you been during the past week on a scale of 0 to 10?” A cutoff of 4 or above is recommended to be significant, with 4 or 5 indicating mild distress, 6 or 7 denoting moderate distress, and 8 or higher indicating severe distress. An important addition is a problem checklist that highlights potential areas of difficulty for a patient that may be linked with perceived distress.

Recently, several other methods of measuring distress have been proposed. The Psychological Distress Inventory (PDI) is a 13-item scale first proposed to measure distress in patients with breast cancer. It was tested against a structured clinical interview as the criterion, with a cut-off of 28 or 29 considered clinically significant. Several variants on the thermometer format have also been developed. Lees and Lloyd-Williams tested a VAS anchored with a sad face and happy face. They reported a high correlation with the Hospital Anxiety and Depression Scale total score (HADS-T) but did not report sensitivity or specificity. García et al. tested a VAS for anxiety and depression but did not use a robust outcome standard.

In a multicenter study in Europe, Gil et al. assessed the value of both the DT and a Mood Thermometer (MT), although the comparator was the HADS. Interestingly, the DT was more highly associated with HADS anxiety scores than depression scores, whereas the MT was related to both HADS anxiety and depression scores. Recently, Mitchell et al. developed a 5-item Emotion Thermometer designed to measure multidomain emotional complications of cancer. It had good validity against Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)-defined depression and HADS total scores in early cancer, but studies are awaited regarding interview-defined distress.

In addition to these custom scales, abbreviated versions of every major mood scale have been published using factor analysis or Rasch analysis. A good example is the adaptation of 10-item Edinburgh Postnatal Depression Scale (EPDS) into an 8-item version. An important caveat is that often the abbreviated version is untested in an independent sample, making interpretation difficult.

The purpose of the following data synthesis is to examine which short scales have been adequately validated against defined distress in cancer or pallia-
tive settings and to compare the accuracy of single and multiple applications of the scale.

**Methods**

A systematic search and appraisal was conducted through Medline/PubMed, Embase, SCOPUS, and Web of Knowledge. The search focused on diagnostic validity studies of tools to identify distress in the cancer and palliative settings. Related methodologic details were reported previously. The author’s 2007 search and analysis was updated to August 2009 and limited to studies measured against interview-defined distress. An analysis was conducted regarding the merits of single versus multiple application of short screening using a typical distress rate of 30%. To account for sample size variations, a meta-analytic weighted rate for sensitivity and specificity was calculated. When heterogeneity was high, a random-effects meta-analysis was performed using StatsDirect 2.6.2 (StatsDirect Ltd, Cheshire, United Kingdom). To account for variations in prevalence, a Bayesian plot of conditional probabilities was calculated. This shows conditional post-test probabilities from all pre-test probabilities regardless of prevalence. The area under the Bayesian positive curve (AUC+) allows statistical comparison of rule-in success, and 1– AUC (or area above the negative curve [AUC–]) allows statistical comparison of rule-out success. This article defines case-identification as the application of a tool to identify (rule-in) individuals with the index disorder, minimizing false-positives. Screening was defined as the systematic application of a tool to rule-out individuals with the index disorder, minimizing false-negatives.

**Results**

Table 1 shows 45 potentially useful short and ultra-short tools that were identified. Of those, only 3 methods had been tested against robustly defined distress (i.e., distress defined by semi-structured interview) in multiple samples. Validation of these 6 tools was attempted in 24 studies. Only the DT and single verbal question (1Q) were specifically validated against interview-defined distress in more than 1 independent sample. Table 1 uses meta-analytic weighting to show the pooled results for the DT versus the single verbal question.

**Single Application of a Short Screen**

At the optimal cutoff suggested by the original authors, the DT seems nonsignificantly better at screening than the single verbal question (AUC$_{DT}$ 0.682 vs. AUC$_{1Q}$ 0.648). However, the 1Q is slightly better at case-finding than the DT (AUC$_{1Q}$ 0.687 vs. AUC$_{DT}$ 0.643). However both methods seem to be improved by adding a second domain. Essentially 2 questions (or 2 thermometers) appear to be more accurate than 1. This is most clearly illustrated by plotting post-test probabilities (positive predictive value [PPV] and negative predictive value [NPV]) according to pre-test probability (prevalence) in a Bayesian conditional probability graph (Figure 1). The case-finding ability is given by an AUC$_{2Q}$ of 0.831 and the screening ability AUC$_{2Q}$ of 0.673 (Table 2).

**Multiple Applications of a Short Screen**

Although historically the science behind algorithmic testing has been unclear, new analysis shows gain from repeated application for any given prevalence. In an algorithmic application, those who test negative after initial screen undergo no further testing because NPV is high, but those who test positive may undergo a second test. Assuming sensitivity and specificity of the test hold true (Table 2), and at a prevalence of 30%, the overall accuracy (fraction correctly identified) is 70% using the DT and 78% using the HADS from a single application. If the DT is applied twice to the same patient group, sensitivity would decline, but specificity would improve and overall accuracy would become 81%. Notably 3 applications often yield no further incremental benefit in accuracy over 2 applications. If the DT is used first, and then the HADS applied as a second step, the overall accuracy would be 82%. Thus, in low prevalence settings, an algorithmic combination of tests usually improves on the yield compared with the single application of any one scale.

**Discussion**

Distress is a common but treatable complication of cancer, present in at least one third of people undergoing active cancer treatment. Distress may comprise predominant depression, anxiety, or anger, or present in mixed broadly defined forms. Although evidence shows that willingness to screen is increasing among cancer professionals, clinicians still prefer brief tests that are easy to administer and score.
Table 1  Short and Ultra-Short Tools for Emotional Complications of Cancer

<table>
<thead>
<tr>
<th>Items</th>
<th>Scale</th>
<th>Abbreviation</th>
<th>Primary Focus</th>
<th>Diagnostic Accuracy or Validity in Cancer Settings?</th>
</tr>
</thead>
<tbody>
<tr>
<td>14 Items</td>
<td>Hospital Anxiety and Depression Scale</td>
<td>HADS</td>
<td>Distress</td>
<td>Yes</td>
</tr>
<tr>
<td>13 Items</td>
<td>Beck Depression Inventory-Short Form*</td>
<td>BDI-SF</td>
<td>Depression</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Psychological Distress Inventory</td>
<td>PDI</td>
<td>Distress</td>
<td>Yes</td>
</tr>
<tr>
<td>12 Items</td>
<td>General Health Questionnaire - 12*</td>
<td>GHQ-12</td>
<td>Distress</td>
<td>Yes</td>
</tr>
<tr>
<td>11 Items</td>
<td>Bech-Rafaelsen Melancholia Scale</td>
<td>MES</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td>10 Items</td>
<td>Edinburgh Postnatal Depression Scale</td>
<td>EPDS (original)</td>
<td>Depression</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Montgomery Asberg Depression Rating Scale (MADRAS)</td>
<td>MADRS (original)</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Zung Self-Rating Depression Scale, Short Form*</td>
<td>SDS-10</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Center for Epidemiologic Studies Depression Scale - 10*</td>
<td>CES-D-10</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Depression Scale</td>
<td>DEPS-10</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td>9 Items</td>
<td>Patient Health Questionnaire - 9</td>
<td>PHQ9</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Hornheide Short Form*</td>
<td>HSF</td>
<td>Depression</td>
<td>Yes</td>
</tr>
<tr>
<td>8 Items</td>
<td>Medical Outcomes Scale - 8</td>
<td>MOS-8</td>
<td>Depression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Even Briefer Assessment Scale for Depression*</td>
<td>EBAS-Dep</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Edinburgh Postnatal Depression Scale - 8*</td>
<td>EPDS-8</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Patient Health Questionnaire - 8*</td>
<td>PHQ-8</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td>7 Items</td>
<td>Hospital Anxiety and Depression scale - depression subscale</td>
<td>HADS-D</td>
<td>Depression</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Hospital Anxiety and Depression scale - anxiety subscale</td>
<td>HADS-A</td>
<td>Anxiety</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Hamilton Depression Scale - 7*</td>
<td>HAM-D-7</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Beck Depression Inventory - 7*</td>
<td>BDI-7</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Duke Anxiety-Depression Scale</td>
<td>DADS-7</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Edinburgh Postnatal Depression Scale - depression items</td>
<td>EPDS-7</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Hornheide Screening Instrument*</td>
<td>HSI</td>
<td>Depression</td>
<td>Yes</td>
</tr>
<tr>
<td>6 Items</td>
<td>Brief Edinburgh Postnatal Depression Scale*</td>
<td>BEDS</td>
<td>Depression</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Hamilton Depression Scale - 6*</td>
<td>HAM-D-6</td>
<td>Depression</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Center for Epidemiologic Studies Depression Scale - 6*</td>
<td>CES-D-6</td>
<td>Depression</td>
<td>No</td>
</tr>
</tbody>
</table>
search shows that at least 45 potentially valuable short and ultra-short tools are available (Table 1), but most are intended to help diagnose depression, and few studies have attempted validation in cancer settings. When limited to those testing against robustly defined distress according to semi-structured interview, only 6 methods have been validated across 24 studies (Table 1). Comparing these 6 tools side-by-side suggests that in a screening capacity all tools have approximately comparable accuracy with no significant differences (side-by-side implies an indirect comparison across different studies, whereas head-to-head is a direct comparison in the same sample). Thus, choice of tool should be based mostly on acceptability or cost-effectiveness. Although best evidence supports the use of the DT or a single verbal question, their rule-in accuracy is modest because they generate a significant number of false-negatives. However, their overall accuracy is almost comparable with the HADS-T, which is clearly a more inefficient (14-item) test. Although data are sparse and await replication, current information suggests that 2 verbal questions, the DT/Impact Thermometer combined, or the PDI are the optimal short tools for determining distress.
Clinicians wanting to know the best simple tool for determining distress may be disappointed, because data are currently too incomplete to make a strong recommendation. Issues of acceptability are likely to guide implementation as much as or more than accuracy alone. Lessons from studies with the Edmonton Symptom Assessment Scale (ESAS) show that completion rates vary by increasing age, higher opioid dose, and the presence of confusion.\textsuperscript{36-38} VAS scales are usually highly acceptable but may have lower completion rates than verbal or categorical scales.\textsuperscript{39}

In attempting to determine the optimal tool, a limited number of head-to-head studies have been conducted, but these are restricted to clinical depression. For example, Mitchell et al.\textsuperscript{40} recently examined several screening tools in a sample of 215 chemotherapy attendees who had early cancer. Comparing the accuracy of the Patient Health Questionnaire 2 (PHQ2), PHQ9, HADS-T, and HADS-depression subscale showed that the PHQ2 was the optimal strategy for detecting DSM-IV–defined major or minor depressions. Akizuki et al.\textsuperscript{41} found the HADS superior to either the DT or single verbal question. Grassi et al.\textsuperscript{42} also showed superiority of the HADS over ultra-short methods with an overall accuracy for ICD-10 cases, but this did not reach statistical significance.\textsuperscript{43} In the most extensive comparison to date, Singer et al.\textsuperscript{43} compared 6 scales in 250 individuals diagnosed according to DSM-IV. Against the HADS, the subscale “Emotional Functioning” of the European Organisation for Research and Treatment of Cancer Quality of Life Core Questionnaire (EORTC QLQ-C30) and a single-item VAS were all highly accurate.

Despite many worthwhile developments, the accuracy of short and ultra-short tools is modest and no tool should be relied on in isolation. Short tools can potentially be supplemented in 5 ways: 1) they can be used as part of a package of assessment that includes unmet needs (e.g., in a problem list);\textsuperscript{44,45} 2) they can be incorporated into a validated global assessment, such as the ESAS, in which the sum scores can generate an overall distress score;\textsuperscript{45} 3) they can also be combined with each other, or with longer methods...
to further improve accuracy, albeit at a cost of increased length;46 4) their accuracy can be potentially increased by repeated application, as with any tool; and 5) they can be usefully combined with questions about whether an individual needs professional help for their distress.47

**Conclusions**

This article summarizes the diagnostic validity of short tools for distress. The next step is examination of proof of benefit in implementation studies.48 No group has yet conducted a randomized trial of screening involving a short method, although these studies are currently underway and results awaited eagerly (http://prsinfoclinicaltrial.gov/ct2/show/NCT00960466).

**References**


