Depression occurs in one third of patients with stroke and is associated with increased mortality, poorer functional outcome, and increased caregiver distress. Stroke management guidelines recommend routine screening for depression. Because guidelines also recommend screening for other forms of psychological distress such as anxiety, a general distress measure may be useful provided depressive disorders are detected. Many depression and distress screening tools commonly used in medical settings have not been well validated in stroke populations.

The Patient Health Questionnaire-9 (PHQ-9), Hospital Anxiety and Depression Scale (HADS), Beck Depression Inventory II (BDI-II), Distress Thermometer (DT) and Kessler-10 (K-10) are depression or distress measures commonly used in primary and/or specialist care settings. The PHQ-9 has been recommended for use in stroke due to its brevity and strong psychometric properties. Along with the 2-item PHQ-2, it has been evaluated in patients with non-aphasic stroke with promising results. Similarly, the HADS depression (HADS-D) and anxiety (HADS-A) subscales and total score (HADS-Total) have been validated for use with patients with stroke without communication problems. Several other depression measures commonly used in stroke practice (eg, Geriatric Depression Scale) also have evidence supporting their use. Evidence for diagnostic accuracy of the BDI-II, DT, and K-10 is lacking in stroke populations. Diagnostic accuracy of the original BDI has been evaluated, but not the
BDI-II, although a factor analytic study in a neurorehabilitation population indicated the measure provides a meaningful severity score of depression in patients with stroke. The DT, a single-item distress measure commonly used in oncology, is being used in a UK-based stroke service enhancement project. The K-10, a distress screening scale used in several population-based surveys, is being used by Australian general practitioners within a national government rebated mental healthcare plan.

Validation studies of the PHQ-9, PHQ-2, and HADS have several methodological limitations. Recruitment has occurred between 2 weeks and 6 months poststroke, whereas stroke guidelines recommend depression screening beyond the 6-month period. Furthermore, in the PHQ-9 validation study, a clinician interview was only administered to patients endorsing symptoms of depression on the PHQ-9, and the study population was limited to patients 1 to 2 months after ischemic stroke with minimal aphasia or cognitive impairment. These findings therefore need to be replicated, preferably with a more heterogeneous group of patients with stroke. Stroke clinicians might manage patients for many years postevent and screening measures should be validated across all time points in the recovery pathway.

The aim of the present study was to examine performance of these commonly used depression-specific or distress measures with regard to internal consistency, concurrent validity, and diagnostic accuracy in detecting a major depressive episode (MDE) compared with a clinician-administered structured diagnostic interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition as the gold standard. A heterogeneous sample of patients with stroke representing those seen in postacute inpatient and outpatient stroke services was examined.

Methods

Eligible participants had a confirmed stroke >3 weeks prior, were aged ≥18 years, and able to attend a local health site for assessment in the Hunter region of New South Wales, Australia. Exclusion criteria included inability to read or understand English or severe cognitive or physical impairment precluding participation. Patients with expressive dysphasia or dysarthria with adequate receptive and expressive communication strategies were included. Ethics approval was gained from the Hunter New England Human Research Ethics Committee (No. 08/08/2005.02).

Outpatients were recruited through the Community Stroke Team, Hunter Medical Research Institute Research Register, stroke clubs, and hospital discharge lists. Consecutive rehabilitation unit inpatients were also approached. Consenting participants were interviewed using the Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition MDE module by a trained clinical psychologist at the hospital or health site. Participants self-completed the HADS and BDI-II using pencil and paper and PHQ-9, DT, and K-10 by computer touchscreen. Outpatients completed their assessments during 1 appointment; inpatients had the option of having a break if fatigued.

Measures

The Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition MDE module was used as the gold standard (criterion standard). A depression case was defined as a positive score on the Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition diagnosis of MDE. Selected questionnaires were administered for the depression-specific (“Depression”: PHQ-2, PHQ-9, HADS-D, BDI-II) and multifactorial/general distress (“Distress”: DT, K-10, and HADS-Total) measures of symptom burden.

The 9-item PHQ-9 is based on the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition diagnostic symptoms of MDE. Responses are rated 0 to 3 in reference to the past fortnight (total score range 0–27) with higher scores indicating greater severity. The PHQ-2 is the sum of the first 2 items (range, 0–6). Previously recommended cutoffs for the PHQ-9 and PHQ-2 are ≥10 and ≥2, respectively. The PHQ-9 also provides a preliminary diagnosis of “major depressive syndrome” using a scoring algorithm (≥5 items, including Items 1 and/or 2, are rated ≥2 unless suicidal item endorsed, which can be rated ≥1).

The 14-item HADS consists of 2 7-item subscales assessing depression and anxiety. The HADS was developed for medical settings and omit somatic symptoms of depression. Responses are rated 0 to 3 in reference to the past week (subscale scores range 0–21) with a higher score indicating greater severity. The HADS total score (HADS-Total) is scored by adding the HADS-A and HADS-D subscales (range, 0–42). Clinical cutoffs have been provided for HADS-D (>7), whereas past studies suggest an ideal cutoff for the HADS-Total of 11 in patients with stroke.

The 21-item BDI-II was developed to detect changes in depressive symptoms in mental healthcare settings. Responses are on a scale of 0 to 3 in reference to the past fortnight (total score range 0–63) with higher scores indicating greater severity. A clinical cutoff of ≥13 has been suggested.

The single-item DT is a visual analog scale recommended for use in oncology services. It was found to be comparable to the HADS in detecting distress in patients with cancer. Option ranges range from 0 (“no distress”) to 10 (“extreme distress”) in reference to the past week. The recommended cutoff for clinically significant emotional distress in oncology patients is ≥4.

The 10-item K-10 is a measure of nonspecific psychological distress designed as a brief screen for “serious mental illness.” It discriminates well between people with a diagnosed mental health disorder (anxiety, mood or nonaffective psychosis) and all other community respondents. Responses are on a scale of 1 to 5 in reference to the past month (total score range 10–50) with higher scores indicating greater levels of distress. The recommended cutoff for clinically significant psychological distress is ≥19.

Statistical Analysis

Group differences were examined using independent-samples t tests or χ2 tests. Missing data were replaced by the rounded mean item value if ≤20% of items were missing. Internal consistency was assessed with Cronbach’s α. Concurrent validity was examined as Pearson product-moment correlations between the continuous scale scores. For each scale, receiver operating characteristic curves were constructed and area under the curve (AUC) and CIs (binomial exact) calculated. Each receiver operating characteristic curve AUC was compared with chance (AUC=0.5) and with the receiver operating characteristic curves of the other scales with the method of DeLong et al. Used to calculate SE and results expressed as a z-statistic. The sensitivity, specificity, likelihood ratio positive and negative, and posttest probability positive and negative were calculated for established and ideal cut points. PASW Statistics Version 18.0.0 (SPSS Inc, an IBM Company, Chicago, IL), MedCalc Version 11.3.8 (MedCalc Software, Mariakerke, Belgium), and Microsoft Excel 2007 were used.

Results

Recruitment strategies identified 124 potential participants. Nonparticipation reasons included refusal (n=28), distance (n=5), illness or cognitive impairment (n=5), uncontactable/missed (n=10), and unable to read (n=3). One patient’s data were lost due to technical issues. The final sample (n=72) were aged 25 to 91 years (mean, 66.7 years; SD, 13.1), 38

Downloaded from http://stroke.ahajournals.org/ by guest on June 2, 2017
The algorithm had high specificity but poor sensitivity. The sensitivity and specificity. The PHQ-9 categorical diagnosis of "major depressive syndrome" derived from the diagnostic criteria was indicated acceptable internal consistency. Strong correlations indicated acceptable internal consistency. Strong correlations occurred between measures (Table 1).

AUC was >0.80 for all measures apart from the DT (AUC=0.73; Figures 1 and 2). Z-statistics indicated all AUCs were significantly higher for cases than noncases (P<0.005; Table 1). Cronbach α exceeded 0.70 for all scales indicating acceptable internal consistency. Strong correlations (P<0.001) occurred between measures (Table 1).

The 2 ideal cutoff options displayed in Table 2 are those meeting criteria for a good screening tool (sensitivity ≥0.80 and specificity ≥0.60) and those with highest sum of sensitivity and specificity. The PHQ-9 categorical diagnosis of "major depressive syndrome" derived from the diagnostic algorithm had high specificity but poor sensitivity. The recommended stroke cutoff of >10 for HADS-Total had high sensitivity and adequate specificity; however, a higher cutoff of >14 increased specificity and had the highest sum of sensitivity and specificity. For the HADS-D and BDI-II, the ideal cutoff on both criteria were >5 and >11, respectively, lower than previously recommended. Similarly, the PHQ-9 ideal cutoffs were lower than the recommended cutoff with scores >6 meeting criteria for a good screening tool and scores >8 providing the highest sum of sensitivity and specificity. At no point was sensitivity ≥0.80 and specificity ≥0.60 reached concurrently for the PHQ-2, DT, and K-10. PHQ-2 scores >1 and K-10 scores >17 almost met these criteria. Highest sum of sensitivity and specificity was found at higher than previously recommended cutoffs for the PHQ-2 (>3) and K-10 (>25). For the DT, the highest sum of sensitivity and specificity was found at >1; however, this had poor sensitivity. The BDI-II (>11) had the highest overall sum of sensitivity and specificity.

**Discussion**

We examined the performance of several depression and distress screening tools in a stroke population. Overall, this study of a heterogeneous group of patients with stroke able to complete a self-report measure found that, apart from the DT, selected measures performed comparably. Performance differences between longer and shorter tools were not of sufficient size to justify length of time lost in administration (opportunity cost).

Study findings support previous literature indicating the PHQ-2, PHQ-9, and HADS are appropriate for use in patients with stroke without aphasia. Unlike previous studies of...
these measures, we have included participants who varied in time since stroke and number of strokes, representing patients seen across a range of inpatient and outpatient services. Findings are therefore relevant for clinicians who see patients with stroke at both early and late stages of the recovery pathway. Specificity of the PHQ-2 was equivalent to that previously seen in the early postacute period. For the PHQ-9, specificity was lower but acceptable. However, previously recommended PHQ-2 and PHQ-9 cutoffs resulted in lower sensitivity than ideal. Supporting previous literature, ideal cut points seen for the HADS-D were lower than traditionally recommended cut points.

This study provides diagnostic accuracy information for depression measures not previously evaluated in a stroke population. The AUC of the BDI-II (0.89) was the highest of all measures examined and was within the range seen for the BDI at different time points over the first 18 months post-stroke (AUCs 0.86–0.93). The K-10 performed equivalently to the depression-specific measures and correlates strongly with the PHQ-2, PHQ-9, and the BDI-II. However, the DT demonstrated the lowest AUC of all measures. It did not demonstrate a sensitivity–specificity combination that met with guidelines for a good depression screening tool and had the weakest correlations with the other scales. Although factor analysis of the K-10 has identified a depression specific factor, the DT addresses global distress in a single item, potentially capturing a range of nondepressive states of emotional distress (reducing sensitivity). Alternatively, the item’s presentation format may have impacted the results: patients with stroke have been found to be less likely to correctly complete visual analog scales than healthy control subjects, and visual analog mood scales have been found to lack sensitivity when used with patients with stroke. Further work is required to determine the role for the DT in stroke care.

Study strengths included a sample that covered a long period after stroke; inclusion of patients with multiple strokes; use of a strong criterion standard; examination of multiple instruments contemporaneously; and use of an appropriate range of accuracy statistics and receiver operating characteristic comparisons. Study limitations included small sample size (leading to low numbers of depression cases) resulting in low precision of the accuracy statistics; no formal assessment of cognitive and functional status; nonrandomization of test order potentially resulting in test fatigue; and having the same assessor administer the Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition and the screening questionnaires, potentially contributing to observed agreement. A higher rate of major depression was found in our sample (18.1%) than seen in previous studies (12.5%–15.8%). It may be that those with depression symptoms were more likely to volunteer. Comparison with published data also suggests the present sample may have participants of a younger age and a higher ratio of patients having multiple strokes than seen in the region’s stroke population. It may not be possible to generalize our results to other populations and replication studies are needed.

Several clinical implications arise from our results. Findings support recommendations advocating use of the PHQ-2, PHQ-9, and HADS in patients with stroke able to complete a
Table 2. Sensitivity, Specificity, Likelihood Ratios, and Posttest Probabilities at Recommended and Ideal Cut Points

<table>
<thead>
<tr>
<th>Categorical</th>
<th>Test Positive/SCID</th>
<th>Test Negative/SCID</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Likelihood Ratios</th>
<th>Posttest Probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;1†</td>
<td>10/13</td>
<td>37/59</td>
<td>0.77 (0.46–0.95)</td>
<td>0.63 (0.49–0.75)</td>
<td>2.06 (1.4–2.9)</td>
<td>0.37 (0.1–1.0)</td>
</tr>
<tr>
<td>&gt;2*</td>
<td>9/13</td>
<td>49/59</td>
<td>0.69 (0.39–0.91)</td>
<td>0.83 (0.71–0.92)</td>
<td>4.08 (2.8–6.0)</td>
<td>0.37 (0.1–1.0)</td>
</tr>
<tr>
<td>&gt;3‡</td>
<td>8/13</td>
<td>54/59</td>
<td>0.62 (0.32–0.86)</td>
<td>0.92 (0.81–0.97)</td>
<td>7.26 (4.7–11.2)</td>
<td>0.42 (0.1–1.2)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;6†</td>
<td>11/13</td>
<td>37/59</td>
<td>0.85 (0.55–0.98)</td>
<td>0.63 (0.49–0.75)</td>
<td>2.27 (1.7–3.1)</td>
<td>0.25 (0.07–0.9)</td>
</tr>
<tr>
<td>&gt;8‡</td>
<td>10/13</td>
<td>44/59</td>
<td>0.77 (0.46–0.95)</td>
<td>0.75 (0.62–0.85)</td>
<td>3.03 (2.2–4.2)</td>
<td>0.31 (0.1–0.9)</td>
</tr>
<tr>
<td>&gt;9*</td>
<td>9/13</td>
<td>46/59</td>
<td>0.69 (0.39–0.91)</td>
<td>0.78 (0.65–0.88)</td>
<td>3.14 (2.1–4.6)</td>
<td>0.39 (0.2–1.0)</td>
</tr>
<tr>
<td>HADS-D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;5‡</td>
<td>12/13</td>
<td>40/59</td>
<td>0.92 (0.64–1.00)</td>
<td>0.68 (0.54–0.79)</td>
<td>2.87 (2.3–3.6)</td>
<td>0.11 (0.02–0.8)</td>
</tr>
<tr>
<td>&gt;7*</td>
<td>8/13</td>
<td>49/59</td>
<td>0.62 (0.32–0.86)</td>
<td>0.83 (0.71–0.92)</td>
<td>3.63 (2.3–5.7)</td>
<td>0.46 (0.2–1.1)</td>
</tr>
<tr>
<td>BDI-II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;11†‡</td>
<td>12/13</td>
<td>42/59</td>
<td>0.92 (0.64–1.00)</td>
<td>0.71 (0.58–0.82)</td>
<td>3.20 (2.60–4.0)</td>
<td>0.11 (0.02–0.7)</td>
</tr>
<tr>
<td>&gt;13*</td>
<td>11/13</td>
<td>44/59</td>
<td>0.85 (0.55–0.98)</td>
<td>0.75 (0.62–0.85)</td>
<td>3.33 (2.5–4.4)</td>
<td>0.21 (0.05–0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distress scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>&gt;1†</td>
<td>13/13</td>
<td>19/58</td>
<td>1.00 (0.75–1.00)</td>
<td>0.33 (0.21–0.46)</td>
<td>1.49 (1.0–2.2)</td>
</tr>
<tr>
<td></td>
<td>&gt;3*</td>
<td>10/13</td>
<td>37/59</td>
<td>0.69 (0.39–0.91)</td>
<td>0.57 (0.43–0.70)</td>
<td>1.61 (1.0–2.5)</td>
</tr>
<tr>
<td></td>
<td>&gt;17†</td>
<td>11/13</td>
<td>35/59</td>
<td>0.85 (0.55–0.98)</td>
<td>0.59 (0.46–0.72)</td>
<td>2.08 (1.5–2.8)</td>
</tr>
<tr>
<td></td>
<td>&gt;19*</td>
<td>10/13</td>
<td>41/59</td>
<td>0.77 (0.46–0.95)</td>
<td>0.69 (0.56–0.81)</td>
<td>2.52 (1.8–3.6)</td>
</tr>
<tr>
<td></td>
<td>&gt;25‡</td>
<td>7/13</td>
<td>56/59</td>
<td>0.54 (0.25–0.81)</td>
<td>0.95 (0.86–0.99)</td>
<td>10.59 (6.4–17.6)</td>
</tr>
<tr>
<td></td>
<td>HADS-Total</td>
<td>&gt;10*</td>
<td>12/13</td>
<td>37/59</td>
<td>0.92 (0.64–1.00)</td>
<td>0.63 (0.49–0.75)</td>
</tr>
<tr>
<td></td>
<td>&gt;14‡</td>
<td>11/13</td>
<td>44/59</td>
<td>0.85 (0.55–0.98)</td>
<td>0.75 (0.62–0.85)</td>
<td>3.33 (2.5–4.4)</td>
</tr>
</tbody>
</table>

SCID indicates Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; PHQ, Patient Health Questionnaire; HADS-D, Hospital Anxiety and Depression Scale-Depression; BDI, Beck Depression Inventory; DT, Distress Thermometer; K-10, Kessler-10.

*Previously recommended cutoff.
†Highest sum of specificity and sensitivity.
‡Highest sum of specificity and sensitivity.

screening tool. However, the PHQ-9 diagnostic algorithm (providing a categorical diagnosis of major depressive syndrome) had high specificity but unacceptably low sensitivity and cannot be recommended. In a stroke population, it can be argued that screening should optimize detection of major depression cases by using assessments with maximum sensitivity because the clearly outlined negative effects of depression outweigh the consequences of conducting further assessments on patients with no depression. The 21-item BDI-II demonstrated best performance with regard to AUC and sensitivity and specificity at selected cut points; however, it was not significantly superior to the other measures except for the DT. These relatively small differences would not outweigh the time-saving benefits of using shorter measures in routine screening. The performance of the K-10 indicated that, although described as a psychological distress measure, it has good accuracy as a MDE screener in this population. This is important because many patients with stroke in Australia will be managed by general practitioners who are increasingly using this scale in clinical practice.

Although results support potential use of these measures for depression screening and case-finding in stroke services, future research should use larger and representative stroke samples to enable subgroup analyses (eg, for age, gender, disability, and ethnicity) and to confirm ideal cut points. Further work is required to determine the accuracy of screening instruments to detect important related constructs in patients with stroke, for example, anxiety and global distress. Appropriate measures for patients with communication and cognitive impairments should be developed and validated. Finally, the impact of routine depression screening and feedback on clinician behavior and patient outcomes using appropriate screening measures needs further examination in intervention studies.

Acknowledgments
We thank Megan Alston and the Community Stroke Team; Dr Michael Pollack and the Hunter Stroke Service; and the Hunter Medical Research Institute Research Register for assistance with participant recruitment; and Dr Alison Beck for data collection.

Sources of Funding
Funding provided by a bequest from Jennie Thomas through Hunter Medical Research Institute.

Disclosures
None.
References


Depression Screening in Stroke: A Comparison of Alternative Measures With the Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (Major Depressive Episode) as Criterion Standard

Alyna Turner, John Hambridge, Jennifer White, Gregory Carter, Kerrie Clover, Louise Nelson and Maree Hackett

Stroke. published online February 23, 2012;
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/early/2012/02/23/STROKEAHA.111.643296