Assessing psychiatric disorder with a human interviewer or a computer

Glyn Lewis

Abstract

Objective – To compare a self administered computerised assessment of neurotic psychiatric disorder (psychiatric morbidity) with an identical assessment administered by a human interviewer. In particular, to discover whether a computerised assessment overestimates or underestimates the prevalence of psychiatric morbidity in relation to a human interviewer.

Setting – A health centre in south east London, UK.

Subjects – A non-consecutive series of health centre attenders. Complete data were available on 92 subjects.

Design – All subjects received both assessments on the same occasion but were randomised to receive either the computerised assessment first or the human interview first.

Results – The mean total score on the assessment was the same for both methods of administration; computer 8.77 v human 8.69 (95% confidence interval for difference 0.70, 0.87). The correlation between the human and interviewer assessments was 0.91.

Conclusion – Self administered computerised assessments are valid, unbiased measures of psychiatric morbidity. In addition to their use as a research tool, they have potential uses in primary care including screening for psychiatric morbidity and in forming the basis for clinical guidelines.

Methods

A non-consecutive series of subjects who attended a health centre in Bermondsey, south east London were invited to take part. Subjects were selected by receptionist staff if they judged the person would have to wait for some time before seeing the doctor. They were given the computerised assessment and the human interview in random order in a quiet and private room in the health centre and were also asked to complete, by themselves, the “paper and pencil” 12 item General Health Questionnaire (GHQ) and the Hospital Anxiety and Depression Scale. Those who scored 2 or more on the GHQ were described as being above the threshold. The interviewer administered a short sociodemographic questionnaire that included items on sex, age, and social class, classified according to the Goldthorpe and Hope criteria and then divided into computerised assessments of psychiatric morbidity and lack of evidence supporting their clinical efficacy.

There have been reports that in computerised assessments of alcohol consumption respondents admit to consuming larger quantities than they do in similar assessments administered by a human interviewer. Evidence from alcohol purchases is used to suggest that the larger figure is also more accurate. Though other researchers have not found this phenomenon, it raises the issue of whether other questions about socially undesirable characteristics, perhaps including those about mental health, will be relatively over-reported when part of a computer administered questionnaire rather than a questionnaire administered by a human. Greist et al used identical versions of the Diagnostic Interview Schedule, administered either by interviewer or computer, in a sample of psychiatric patients and did not find any bias in the ascertainment of psychiatric diagnoses. Previous studies in Britain have shown that a computerised assessment of psychiatric morbidity showed good agreement with a standardised psychiatric interview administered by a psychiatrist. However, in that study the computerised and human assessments contained different wording and so it was not possible to investigate any possible bias in identifying psychiatric morbidity.

This study was designed to assess the size of any potential bias between an identical assessment of psychiatric morbidity, the revised Clinical Interview Schedule (CIS-R), when it was administered either by a computer or a human interviewer.

British general practitioners are becoming increasingly familiar with computerised methods of assisting practice administration and recording the details of consultations. There is less awareness that computers provide an opportunity to extract information directly from patients by means of self administered questionnaires. Several self administered computerised assessments of psychiatric morbidity have been developed, including one designed in the UK for use in primary care settings which concentrates on rapidly assessing the common neurotic disorders of depression and anxiety (called psychiatric morbidity in this report). Many of the proponents of computerised assessments have been discouraged by the reluctance of clinicians to use such information technology in their work. Among the reasons for this, however, must be included the paucity of data on the validity of computerised psychiatric morbidity assessments.
Table 1  The level of agreement for individual symptoms between the human (H) and computer (C) interview (n = 92).

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Weighted kappa / SEM</th>
<th>Interview</th>
<th>Score &gt; 2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>H</td>
<td>C</td>
</tr>
<tr>
<td>Somatic</td>
<td>0.49 (0.03)</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.63 (0.04)</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>Concentration</td>
<td>0.72 (0.02)</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Depression</td>
<td>0.65 (0.02)</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Irritability</td>
<td>0.72 (0.03)</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Sleep</td>
<td>0.60 (0.05)</td>
<td>53</td>
<td>37</td>
</tr>
<tr>
<td>Worry</td>
<td>0.64 (0.02)</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Depressive ideas</td>
<td>0.70 (0.02)</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Worry</td>
<td>0.66 (0.03)</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.53 (0.03)</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>Phobia</td>
<td>0.48 (0.02)</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Panic</td>
<td>0.69 (0.07)</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Compulsions</td>
<td>0.75 (0.01)</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Obsessions</td>
<td>0.59 (0.02)</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

Results
Ninety seven subjects agreed to take part in the study and complete data was available for 92 subjects. The mean (SD) age of the subjects was 40 (17.7) years, and 81% were women. Twenty per cent were either divorced, separated, or widowed and 48% were in a manual social class. Altogether 12.5% were born outside the UK and 19% had previously consulted a doctor about a mental health problem.

There were no statistically significant differences in the characteristics of those who received the computerised assessment first and those who received it second (table 2). In particular, there was no difference in the proportion who scored above the threshold on the GHQ. The mean total scores on the CIS-R for the first and second assessments were compared and there was no evidence of any significant differences in scores. The order of presentation did not therefore have any influence on total scores.

The mean total score on the computerised assessment was 8.77 (95% confidence interval (CI) 6.89, 10.66) and on the interviewer assessment it was 8.69 (95% CI 6.86, 10.52) (paired t-test; t = 0.2, df = 91, p = 0.8; 95% CI for difference 0.70, 0.87). The correlation between the human and interviewer assessments was 0.91. Subjects were also divided into cases and non-cases on the basis of CIS-R scores (table 3). The index of agreement on positives was 0.66 and the kappa was 0.70 (SD 0.08). Agreement on the individual sections of the CIS-R was also examined (table 1). The mean kappa value across the sections of the CIS-R was 0.63. The possibility of bias within the sections was examined using McNemar's test. In only one of the 14 sections there was any indication of a statistically significant bias, the sleep section (χ² = 13.2; df = 1; p < 0.001), in which the computerised assessment resulted in higher scores.

The reliabilities of the computerised and human CIS-R were estimated using confirmatory factor analysis. The results in table 4 indicate that both measures were more reliable than the questionnaire and had reliabilities of around 0.90. Though computer administration had a higher reliability than the administration by a human interviewer, it is impossible, for technical reasons, to estimate the statistical significance of this difference.

Table 2  Comparison of respondents who were allocated to human and computerised administration for the first assessment.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Computer first</th>
<th>Computer second</th>
<th>Significance test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (%) (95% CI)</td>
<td>81 (73.9, 89.9)</td>
<td>81 (73.9, 89.9)</td>
<td>χ² = 0; df = 1; p = 1.0</td>
</tr>
<tr>
<td>Divorced, separated, or widowed (%) (95% CI)</td>
<td>22 (13.9, 30.9)</td>
<td>17 (9.3, 24.7)</td>
<td>χ² = 0.44; df = 1; p = 1.0</td>
</tr>
<tr>
<td>Born outside UK (%) (95% CI)</td>
<td>12 (5.9, 18.5)</td>
<td>12 (6.0, 19.6)</td>
<td>χ² = 0.006; df = 1; p = 0.94</td>
</tr>
<tr>
<td>In manual occupations (%) (95% CI)</td>
<td>55 (44.6, 65.3)</td>
<td>40 (30.4, 50.6)</td>
<td>χ² = 2.0; df = 1; p = 0.15</td>
</tr>
<tr>
<td>Previously treated for mental health problem (%) (95% CI)</td>
<td>18 (10.5, 26.3)</td>
<td>19 (11.2, 27.2)</td>
<td>χ² = 0.02; df = 1; p = 1.0</td>
</tr>
<tr>
<td>Mean (SD) age</td>
<td>39.9 (18.0)</td>
<td>40.2 (17.6)</td>
<td>t = 0.99; df = 1; p = 0.33</td>
</tr>
<tr>
<td>Above GHQ threshold (%) (95% CI)</td>
<td>45.8 (35.6, 56.0)</td>
<td>42.6 (32.5, 52.7)</td>
<td>χ² = 0.1; df = 1; p = 0.75</td>
</tr>
</tbody>
</table>
Assessing psychiatric disorder with a human interviewer or a computer

Table 3  Agreement on case definition according to method of interviewing.

<table>
<thead>
<tr>
<th>Human</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-case</td>
</tr>
<tr>
<td>Non-case</td>
<td>54</td>
</tr>
<tr>
<td>Case</td>
<td>5</td>
</tr>
</tbody>
</table>

Kappa = 0.70 (SD 0.084).

Discussion

No differences in the ascertainment of psychiatric morbidity were observed when an identical questionnaire was administered either by a human interviewer or by a computer. Though studies enquiring about alcohol intake have led to suggestions that people are more likely to divulge sensitive information to a computer than to another person, this effect was not seen here in assessing psychiatric morbidity. This is consistent with Greist's findings in North America using a different assessment and investigating psychiatric patients rather than primary care attenders. Only one of the 14 sections of the CIS-R showed any evidence of bias, the section on sleep, and it is possible that this was due to chance. Overall, the computerised assessment gave very similar estimates to the human administered assessment. The level of agreement observed between the human and computerised assessments was also similar in magnitude to the results of the study of Lewis et al10 of the agreement between two interviewers administering the CIS-R. This suggests that self administered computerised assessments of psychiatric morbidity, such as the one used here, are as valid as interviewer administered measures in general practice and community settings. It is important to emphasise that the assessment used was designed only to assess neurotic disorders and is therefore suitable for use in primary care and other settings where psychotic disorders are relatively uncommon. The conclusion concerning the validity of computerised assessments therefore applies only to these circumstances.

Only a single interviewer was used here, who was also a trained psychiatric nurse. It is possible that other interviewers, perhaps with a less sympathetic demeanour, may lead to a failure to disclose information about mental health. Indeed one of the attractions of using computerised assessments is the lack of observer bias and the consistency of the assessment in different situations. Even well trained interviewers will inevitably vary in interviewing styles. This might be reflected in the slightly higher reliability observed for the computerised assessment (table 4).

The sample of subjects was not representat-

Table 4 Reliabilities of the different assessments using a factor analysis measurement model.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Score (mean / SD)</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIS-R computer</td>
<td>8.7 (8.8)</td>
<td>0.96</td>
</tr>
<tr>
<td>CIS-R human</td>
<td>8.8 (9.1)</td>
<td>0.86</td>
</tr>
<tr>
<td>GHQ 12</td>
<td>12.3 (6.4)</td>
<td>0.61</td>
</tr>
<tr>
<td>HAD</td>
<td>9.9 (6.6)</td>
<td>0.48</td>
</tr>
</tbody>
</table>

CIS-R = Clinical Interval Schedule (revised); GHQ = General Health Questionnaire; HAD = Hospital Anxiety and Depression Scale.

Information technology is very fashionable just now. Despite these encouraging results it is important that developments in computerised assessments are properly evaluated and that their use does not jeopardise the relationship between doctor and patient. Such relationships influence the non-specific aspects of
treatment as well as patient satisfaction and compliance. Computers cannot replace the clinical acumen of general practitioners but may prove to be a useful adjunct to the treatment of psychiatric morbidity by the primary care team.

This study and GL were funded by the Department of Health. My thanks to Gill Todd for carrying out the assessments and to Dr Richard Donmall for providing facilities and access to patients in the Albion Street Health Centre. I am grateful to Professor Anthony Mann for comments on an earlier draft.