Pergamon

European Journal of Cancer, Vol. 35, No. 1, pp. 79–85, 1999

Printed in Great Britain
0959-8049/99/$ - see front matter

PII: S0959-8049(98)00308-6

Original Paper

Are We Using Appropriate Self-report Questionnaires for Detecting Anxiety and Depression in Women with Early Breast Cancer?

A. Hall,[1] R. A'Hern[2] and L. Fallowfield[1]

[1]CRC Psychosocial Oncology Group, Department of Oncology, UCL Medical School, Bland Sutton Institute, 48 Riding House St, London W1P 7PL; and [2]Department of Computing Science, Royal Marsden Hospital Trust, Mulberry House, Dovehouse St, London SW3 6JJ, U.K.

The aim of this prospective study was to identify the psychiatric morbidity associated with the diagnosis and treatment of early breast cancer. At each of five time points, 269 women were interviewed using a shortened version of the Present State Examination (PSE) and 266 completed self-assessment questionnaires, the Hospital and Anxiety Depression Scale (HADS) and the Rotterdam Symptom Checklist (RSCL). This paper compares the ability of the questionnaires to detect psychiatric morbidity with that of the PSE. The majority of women who experienced anxiety and/or depression did so within 3 months of their initial surgery. The clinical interview identified anxiety disorder in 132 of 266 women (49.6%) and depressive illness in 99/266 (37.2%) during the first 3 months. Using the recommended threshold of ≥ 11 for caseness, the sensitivities for both tests were very low at 24.2% (HADS anxiety) and 14.1% (HADS depression) and 30.6% (RSCL psychological distress scale). Lowering the threshold value to ≥ 7 on the HADS improved the sensitivity to 72% for the anxiety subscale, but it remained low at 37.4% for the depression subscale. A threshold of ≥ 7 for the RSCL scale raised sensitivity to 66.7%. Lowering the threshold values raised the sensitivity of both the instruments but decreased their specificity: the lower the threshold, the greater the number of women who were identified as false positives which would increase the work load for clinic staff if used as a screening tool. Given that the HADS was inadequate in discriminating for depressive illness, it was not surprising that its use as a unitary scale with a threshold value as low as 12 resulted in a sensitivity of only 42.7%. In the light of these findings, we question the use of both the HADS and the RSCL as suitable research or screening instruments for detection of psychological morbidity in early breast cancer. © **1999 Elsevier Science Ltd. All rights reserved.**

Key words: HADS, RSCL, PSE, anxiety, depression, breast cancer
Eur J Cancer, Vol. 35, No. 1, pp. 79–85, 1999

INTRODUCTION

EVERY YEAR in the U.K., around 24 000 women are diagnosed with breast cancer. There has been extensive research into the psychosocial sequelae of the diagnosis and treatment of this disease providing ample evidence that a significant minority of women are at increased risk of developing anxiety and depression [1, 2]. Systematic studies show that between 20 and 35% of women with breast cancer, irrespective of their stage of disease, have measurable psychiatric morbidity [3, 4]. In research settings, questionnaires are used frequently to assess quality of life of patients entered into clinical trials, so it is important to determine how reliably they detect differences between groups. In clinical practice, detection rates of this morbidity have been shown to be low [5]. Patients may be reluctant to disclose how they are feeling and doctors and nurses may be unwilling to enquire or lack the appropriate communication skills. Accurate methods of detection of psychiatric distress that are applicable to busy clinics and acceptable

Correspondence to A. Hall, Centre for Medical and Dental Education, St Bartholomews and the Royal London School of Medicine and Dentistry, Robin Brook Centre, West Smithfield, London, EC1A 7BE, U.K. e-mail: a.hall@mds.qmw.ac.uk
Received 23 Mar. 1998; revised 29 Jul. 1998; accepted 4 Aug. 1998.

to patients would enable doctors and specialist nurses to concentrate their efforts on women who need their help and thus use their time more efficiently. Two questionnaires are widely used as screening instruments to measure psychological morbidity for patients with cancer in both research and clinical settings: the Hospital Anxiety and Depression Scale (HADS) [6] and the Rotterdam Symptom Checklist (RSCL) [7].

The HADS is intended for use with physically ill patients and unlike other psychiatric self-report questionnaires, does not contain somatic symptoms which may be attributable to disease and/or treatment. The 14-item instrument comprises two 7-item subscales designed to discriminate between anxiety and depression. All items refer to symptoms experienced during the previous week. The authors [6] recommended using a threshold of 11 to include all probable cases of anxiety disorder and depressive illness and a threshold of 8–10 to identify all possible cases of the two disorders.

The RSCL was developed primarily as an instrument to measure symptoms reported by patients with cancer participating in clinical research. It is a multidimensional scale for use in assessment of quality of life and contains two subscales: one of these is an 8-item scale assessing psychological distress during the previous three days. The recommended threshold for this scale is 11, according to personal communication with one of the authors, cited by Hopwood and associates [8].

The few studies that have now been published measuring the performance of these questionnaires as screening tools in different populations of patients with cancer, have reported somewhat equivocal findings. Razavi and colleagues reported that the French version of the HADS scale failed to discriminate between anxiety and depression when measured against a psychiatric interview [9]. They suggested summation of both scales to provide a unitary 14-item psychological distress scale and tested its ability to screen for psychiatric morbidity on a sample of in-patients with various cancer sites. Application of receiver operating characteristic (ROC) analysis [10] found that the optimal thresholds on the unitary scale were 19 for major depressive disorder and 13 for depressive illness plus adjustment disorder.

In a study of patients with a variety of cancers, using paired combinations of questionnaires including the HADS and the RSCL, Ibbotson and colleagues [11] concluded that both instruments performed well. In this study, psychiatric diagnosis was made following the administration of the Psychiatric Assessment Schedule [12], using DSM III criteria [13]. The HADS was used as a unitary scale with 14 identified as the optimal threshold value using ROC analysis; the study did not test the performance of the HADS subscales. However, when patients were classified as disease-free, having stable disease or with progressive disease, some differences in the performances of the instruments emerged. The HADS performed well in patients who were disease-free or whose disease was judged stable, but was not sensitive or specific enough for patients with progressive disease. Only the RSCL performed well in those who had progressive disease. In patients who were receiving active treatment, both questionnaires were effective as screening instruments.

Patients with advanced breast cancer, in a study by Hopwood and colleagues [8] completed the HADS and the RSCL and were interviewed using the Clinical Interview Schedule [14]. The HADS was able to discriminate between anxiety and depression. Optimal sensitivity and specificity were found to be in line with the recommended threshold of 11 for both the HADS anxiety and depression subscales and 11 on the RSCL psychological distress scale. However, the misclassification rate for the HADS depression subscale was 25%, compared with 12% for the anxiety subscale and 21% on the RSCL. When these thresholds were applied to a large outpatient sample of women with advanced breast cancer [15], the authors found that although both questionnaires identified very similar percentages of cases, there was considerable discrepancy in the makeup of the two groups: concordance in identifying cases was only 43%.

A study conducted by Ramirez and colleagues [16] focused specifically on women with early breast cancer. The primary aim was to assess the ability of the pre-operative HADS score to detect current mood disorder and to predict disorder at 3 and 12 months postoperatively, measured against a psychiatric interview. The HADS was used as a unitary 14-item scale; no results were reported about the ability of the HADS subscales to discriminate between anxiety and depression. The application of ROC curve analysis demonstrated an optimum threshold of 11 on the unitary scale to achieve a sensitivity of 70% in identifying psychiatric disorder either pre-operatively or at 3 and 12 months postoperatively. The performance of the HADS in identifying concurrent mood disorder at 3 and 12 months was poor. The authors concluded that the performance of a single pre-operative HADS score in identifying mood disorder in the year after diagnosis was superior to using the HADS to detect concurrent mood disorder at intervals throughout that year.

Thus, there appears to be uncertainty, not only about which questionnaire most accurately identifies patients with anxiety and depression, but also about what threshold values should be used. In this paper we compare the performance of the HADS and the RSCL in detecting anxiety and depression with that found following a standardised clinical interview.

PATIENTS AND METHODS

Sample

Two hundred and sixty-nine women treated for early breast cancer in the south of England were recruited to a large Cancer Research Campaign funded study designed to assess psychological outcomes of different treatment policies. Detailed results from this study have been reported elsewhere [4]. All the women were aged less than 75 years and had stage I or stage II disease. Of the 115 women who were treated by local excision, 107 (93%) received adjuvant radiotherapy. Adjuvant chemotherapy was given to 20/269 (7%) of the sample and 155/182 (85%) of postmenopausal women were given tamoxifen.

Assessments

The women were interviewed in their own homes by trained interviewers at five different time points: postoperatively (within 3 weeks of surgery), at 3 months, and at 1, 2 and 3 years after initial diagnosis and treatment, using a shortened version of the Present State Examination (PSE) [17]. Self-assessment questionnaires including the HADS and the RSCL were left with the women to complete and return in a pre-paid envelope immediately following the interview. Only results for the postoperative and 3 month assessment are reported here. Results for the other assessments have been previously published elsewhere [18].

The PSE is a standardised semi-structured interview designed to assess present mental state to which psychiatric diagnostic criteria (DSM III) were applied. Depressive illness was diagnosed if women had reported depressed mood for at least the four preceding weeks accompanied by at least four of the following symptoms: changes in weight or appetite; disturbances of sleep pattern; energy retardation; loss of interest; feelings of guilt or worthlessness; loss of libido; diminished concentration; suicidal ideation. Anxiety state was diagnosed if women had experienced generalised persistent anxiety for at least 4 weeks accompanied by symptoms from the following categories: motor tension, autonomic symptoms, apprehensive expectation and problems with concentration.

The interviewers were blind to the questionnaire scores when PSE ratings were made. All interviews were tape recorded in order to enable rate–rerate and interrater tests for reliability. Checks of rate–rerate reliability were performed on a random sample of 241 (30%) interviews. Any discrepancies were checked by an independent rater. Ratings showed little evidence of drift. Interrater reliability checks for each of the three interviewers were also performed on a further 81 of 269 (30%) tapes and showed good overall concordance between interviewers. All reliability coefficients were found to be within satisfactory limits using a weighted Kappa statistic [19].

Analysis

Sensitivity and specificity. The *sensitivity* of a test refers to its ability to identify the proportion of true cases of affective disorder in a given population (number of true cases identified/number of true cases identified plus number of false negatives). The higher the sensitivity, the lower the rate of false negatives (those who are incorrectly identified as non-cases). The *specificity* of a test refers to its ability to identify the proportion of true non-cases (number of true non-cases identified/number of true non-cases identified plus number of false positives). The higher the specificity, the lower the rate of false positives (those who are incorrectly identified as true cases). Changing the threshold values for a test alters both sensitivity and specificity. Lowering the threshold value increases sensitivity but decreases specificity so that there are fewer patients assessed as not distressed when in fact they are (false negatives), but more who are assessed as distressed when they are not (false positives). Raising the value has the reverse effect, decreasing sensitivity but increasing specificity.

Misclassification rate. The misclassification rate (MR) refers to the number of patients who are identified by the questionnaire as either false negatives or false positives and is calculated from the ratio false negatives + false positives/total number in the sample.

Positive predictive value. The positive predictive value (PPV) [20] of a test refers to the probability of a score at or above the chosen cut-off point being a true case. It is calculated from the ratio of correctly identified cases/total number of patients who are identified as true cases. Thus, it takes into account only true cases and false positives, not those patients who are false negatives.

RESULTS

Sample

Results reported elsewhere [18] showed that, in the 3 year duration of the study, the majority of the 269 women who experienced anxiety and/or depression had done so within the first 3 months following diagnosis and treatment. To maximise the number of available cases on which to test the performance of the questionnaires, data were therefore combined from the postoperative and 3 month assessments. Postoperatively and/or at 3 months, 266 (99%) women completed the HADS and 267 (99%) completed the RSCL. The data set includes all the women who were identified by the PSE as cases of anxiety and/or depression and who completed the questionnaires at one time point or the other, but not both, so that all patients are represented once only.

The thresholds selected to apply to the data for the HADS anxiety and depression subscales and for the RSCL were: ≥ 11, the recommended published score for caseness; ≥ 8, which is the value that takes into account borderline cases; and ≥ 7 which was the optimal value for the RSCL identified by Ibbotson and colleagues [11]. The thresholds selected for the HADS as a unitary scale were ≥ 19, ≥ 14 and ≥ 11 to enable comparison of results with the findings of other studies cited above.

Prevalence

One hundred and thirty-two of 266 (49.6%) women were assessed as anxious and 99/266 (37.2%) as depressed by the clinical interview (PSE), postoperatively or at the 3 month interview.

Using a threshold value of 11 for the HADS, 36/266 (13.5%) of the sample were classified as anxious and 20/266 (7.5%) as depressed. Reducing the value to 8 increased cases of anxiety to 105 (39.4%) and of depression to 44 (16.5%). At a threshold of 7, 122 (45.8%) of the women were anxious and 49 (18.4%) were depressed.

On the psychological distress subscale of the RSCL, 49/267 (18.3%), 86/267 (32.2%) and 115/267 (43.1%) of women were identified as distressed using thresholds of 11, 8 and 7, respectively.

Performance of questionnaires compared with clinical interview (PSE)

(i) HADS anxiety and depression subscales. Using the recommended threshold for caseness of 11, sensitivity for the anxiety subscale was 24.2% and specificity was 97% (Table 1). Thus, of the 132 women diagnosed as anxious by the PSE, 32 were identified by the HADS but 100 were not (false negatives). Of the 134 women identified as not anxious by the PSE, 130 were similarly classified by the HADS. On the depression subscale, sensitivity was 14.1% and specificity 98.2%. Fourteen women were identified as depressed by both the PSE and the HADS but the latter failed to detect 85 who were false negatives. Of the 167 women who were found not to be depressed by the PSE, 161 were similarly identified by the HADS.

Applying a threshold value of 8 raised sensitivity on the anxiety subscale to 63.6%, so that 84 of 132 women were rated anxious on both the HADS and the PSE but 48 were false negatives according to the HADS. Specificity was 84.3%: 113 of 134 were identified as not anxious on the two measures and 21 women were false positives. On the depression subscale, sensitivity was 33.3% and specificity 93.4%. Thirty-three women were identified as depressed using both tests but another 66 were false negatives using the HADS. Both tests identified 156 women as non-depressed but the HADS assessed 11 who were false positives.

Table 1. *Screening performance of HADS and RSCL measured against clinical interview (PSE)*

Threshold		*n*	Sensitivity %	*n* on PSE	*n* false neg.	Specificity %	*n* on PSE	*n* false pos.
HADS								
≥11	Anxiety	266	24.2	132	100	97.0	134	4
	Depression	266	14.1	99	85	98.2	167	6
≥8	Anxiety	266	63.6	132	48	84.3	134	21
	Depression	266	33.3	99	66	93.4	167	11
≥7	Anxiety	266	72.0	132	37	79.9	134	27
	Depression	266	37.4	99	62	92.8	167	12
RSCL								
Psychological distress								
≥11		267	30.6	144	100	95.9	123	5
≥8		267	52.1	144	69	91.1	123	11
≥7		267	66.7	144	48	84.6	123	19

HADS, Hospital and Anxiety Depression Scale; RSCL, Rotterdam Symptom Checklist; PSE, Present State Examination.

Table 2. *Screening performance of HADS unitary scale measured against clinical interview (PSE)*

HADS threshold	*n*	Sensitivity (%)	*n* on PSE	*n* false neg.	Specificity %	*n* on PSE	*n* false pos.
≥19	266	20.3	143	103	98.4	123	3
≥14	266	40.6	143	74	94.3	123	8
≥12	266	42.7	143	61	92.7	123	9

Abbreviations as in Table 1 footnote.

Lowering the threshold value to 7 improved sensitivity to 72% on the anxiety subscale: 95/132 were rated as anxious on both measures and 37 were false negatives on the HADS. Specificity was 79.9% so that the HADS and the PSE were both identifying 107 women as non-anxious but 27 women were false positives according to the HADS. Sensitivity on the depression subscale remained low at 37.4% so that only 37 women were identified as depressed on both measures and 62 women were false negatives. Specificity remained high at 92.8%, so that the two measures both identified 155 women as non-cases and 12 women were false positives according to the HADS.

(ii) HADS unitary scale. The application of a threshold of 19 resulted in a sensitivity of 20.3% (Table 2). Thus, of the 143 women identified as suffering from psychological distress (anxiety and/or depression) by the PSE, 40 were identified by the HADS but 103 were false negatives. Lowering the threshold value to 14 raised sensitivity to 40.6%; 69 of 143 women were distressed according to both the HADS and the PSE but the HADS did not identify 74 who were distressed according to the PSE. Further reducing the threshold to 12 made little difference to the sensitivity of the HADS at 42.7%, with 61 of 143 women false negatives. The specificities at all three thresholds were high at 98.4%, 94.3% and 92.7%, respectively, so that the numbers of distressed women identified by the HADS but not the PSE (false positives) were very low.

(iii) RSCL psychological distress scale. Using a threshold value of 11 on the psychological distress scale, sensitivity for the RSCL was 30.6% and specificity was 95.5% (Table 1). Thus 44 women were deemed to be distressed on both the RSCL and the PSE but an additional 100 patients were false negatives assessed by the RSCL. Both the RSCL and the PSE identified 118 of 123 women as non-distressed and 5 women were false positives using the questionnaire. Lowering the value to 8 resulted in a sensitivity of 52.1% and specificity of 91.1, with 75 women identified as cases by both PSE and RSCL and 112 as non-cases. The number of women who were false negatives was 69 and 11 were false positives. Applying the threshold of 7 resulted in a sensitivity rate of 66.7% and a specificity of 84.6%. The RSCL and the PSE both identified 96 women as distressed and 104 who were not distressed, but 48 women were false negatives and 19 were false positives.

Positive predictive values (PPV) and misclassification rate

Table 3 shows that PPVs for both the HADS and the RSCL were high for all three chosen thresholds: the highest PPV was 90% for the RSCL applying a threshold of 11 and the lowest was 75% on the HADS depression subscale using a value of 8. Table 3 also shows that the misclassification rates according to the HADS were 39.1% and 34.2% for anxiety and depression, respectively, using the threshold of 11. At a threshold of 8, rates fell to 25.9% for anxiety and 28.9% for depression. Applying the threshold of 7, misclassification for anxiety was 24.1% but 45.9% for depression. The MR for the RSCL psychological distress scale using the threshold value of 11 was a little lower at 29.9%. The MR rose to 39.1% when the threshold of 8 was used and fell again to 25.1% at a value of 7.

Table 3. *Positive predictive values (PPV) and misclassification rates (MR) for HADS and RSCL*

	HADS anxiety	HADS depression	RSCL
Threshold ≥11			
PPV (*n*)	89% (36)	82% (17)	90% (49)
MR	39.1%	34.2%	29.9%
Threshold ≥8			
PPV (*n*)	80% (105)	75% (44)	87% (86)
MR	25.9%	28.9%	39.1%
Threshold ≥7			
PPV (*n*)	78% (122)	76% (49)	83% (115)
MR	24.1%	45.9%	25.1%

Abbreviations as in Table 1 footnote.

Table 4. *Comparison of rates of morbidity, using different thresholds, between HADS* (n = 266) *and RSCL* (n = 267)

	≥ 11	≥ 8	≥ 7
HADS			
Total cases	40 (15%)	109 (41%)	125 (47%)
Anxiety	23 (9%)	65 (24%)	76 (29%)
Depression	4 (2%)	4 (2%)	3 (1%)
Both	13 (5%)	40 (15%)	46 (17%)
RSCL			
Total cases	49 (18%)	85 (32%)	114 (43%)
Both HADS & RSCL	33	75	98
HADS only	7	10	16
RSCL only	16	34	27
Concordance	59%	52%	64%

Abbreviations as in Table 1 footnote.

Performance of HADS subscales against RSCL

Table 4 shows that at the postoperative or 3 month assessment, there was no significant difference in the prevalence of morbidity detected by the two questionnaires, using threshold values of 11, 8 or 7. The threshold of 11 identified 15% of the sample with the HADS and 18% with the RSCL but the proportion of cases increased to 47% and 43% when the threshold of 7 was applied. Using the thresholds of 11, 8 and 7, concordance between the two questionnaires in identifying cases was 59%, 52% and 64%, respectively.

DISCUSSION

In a clinical trial setting, the misclassification rate of an instrument must be low in order to detect reliably clinically important differences in outcomes between trial arms. In a clinical setting, the sensitivity of an instrument is of primary importance, as it is important to keep rates of false negatives as low as possible. Choosing the optimal threshold value for a test entails a trade-off between sensitivity and specificity. Raising the sensitivity threshold decreases specificity, which means that an increasing number of patients will be false positives, resulting in an additional burden for clinic staff. In practice, in a clinical setting, threshold values may have to be determined by the resources available to the staff. The PPV is of significance clinically as it refers to the probability that a detected case is a true case.

In this study, the prevalence of psychiatric disorder detected by both the HADS and the RSCL was so low at the recommended value of 11 that it comes as no surprise that sensitivities for the two instruments were unacceptably low at this threshold. The rates of anxiety and depression identified were no higher than one would expect to find in a general population sample. What emerges from the analysis of data in this study is that the HADS subscales, using threshold values of 7, were reasonably accurate in screening for cases of anxiety, but not of depression. Given this difference in identifying different types of morbidity, it was not surprising that the use of the HADS as a unitary scale had unacceptably low sensitivity. In order to raise this sensitivity, the threshold would have to be set so low that it would inevitably identify large numbers of women who were false positives. These findings, therefore, do not support those of either Razavi and colleagues [9] or Ibbotson and colleagues [11].

To some extent, our findings are in concordance with those of Ramirez and colleagues [16], who also found that the HADS performed poorly in identifying concurrent mood disorder. To achieve sensitivities of 77% at 3 and 12 months postoperatively, unitary scale thresholds had to be lowered to 5 and 6, respectively, which substantially reduced specificities to 52% and 66%. Ramirez and colleagues also found that the potential for the pre-operative HADS unitary score to predict mood disorder in the year after diagnosis was influenced by age. Among women aged < 50 years ($n = 30$), sensitivity was high at 90%, using the threshold value of 11 but the false positive rate was 40%. Among women aged > 50 years ($n = 61$) who experienced mood disorder, the sensitivity of the HADS fell to 57%, although the false positive rate was low at 3%. In our study, no such difference was found in the ability of the HADS to detect psychiatric morbidity between women aged > 50 years and < 50 years (data not shown).

It is worth noting, however, that the above studies may have overestimated the performance of the screening instruments because thresholds were chosen that optimised sensitivity and specificity in the same data set that were used to assess the tests' performance. Thus, the ability of these instruments to detect psychological distress may be poorer than their results would suggest. Hopwood and colleagues [8, 15] used the more stringent procedure of determining optimal threshold values on one patient population and then applying these values to a second, different population. Nonetheless, the results reported here do not replicate their finding that 11 is the recommended threshold value for both HADS subscales and the RSCL psychological distress scale.

There may have been differences between this study and those cited above due to differences in the populations being sampled. This study, like that of Ramirez and colleagues [16] explored morbidity in a sample of women who were comparatively homogeneous in terms of disease, stage of disease and treatment. All the women had early stage breast cancer and had only recently been diagnosed and completed their surgical treatment. Virtually all of them were receiving adjuvant therapy. Similarly, the sample of Hopwood and colleagues [15] was homogeneous as all patients had advanced cancer of the breast though not all were receiving active or palliative treatments. The samples used in the Razavi and colleagues [9, 10] and Ibbotson and colleagues [11] studies were more heterogeneous, as patients had cancers in a variety of sites, were in different phases of the course of the disease and were not all receiving treatment.

The RSCL was a more accurate screening instrument for overall psychological distress than the unitary HADS scale but none the less was associated with only a 67% sensitivity using a threshold value of 7. Once again, raising sensitivity to an acceptable 75–80% would greatly increase the number of cases who were false positives. Ibbotson and colleagues [11] had to lower the threshold value to 7 in order to achieve acceptable sensitivity and specificity. This threshold of 7 is considerably lower than the recommended value of 11. Concordance between the RSCL and the HADS in identifying the same cases was somewhat higher than the results reported by Hopwood and colleagues [15]. Nonetheless, at a threshold of 11, of the 56 cases identified by the HADS or the RSCL, only 33 were detected by both questionnaires, giving a concordance of 59%. At a threshold of 7, 98 of 141 cases were detected by both questionnaires, a concordance of 64%.

The PPVs for both questionnaires at all thresholds ranged from 75 to 90%, meaning that there was a high probability that cases detected by either instrument were true cases. The

misclassification rate for both instruments was rather high: at a threshold value of 7, one in four women was either a false positive or a false negative on the HADS anxiety subscale and the RSCL psychological distress scale. Almost half of the women were misclassified using the HADS depression subscale.

The discrepancy between identification of 'caseness' according to the HADS and that of the PSE requires explanation. The anxiety subscale was composed of items taken from the anxiety section of the PSE itself, so a reasonable degree of concordance would be expected between the two instruments in the identification of cases of anxiety. No information is given on how the items on the depression subscale were selected, other than that they were largely based on the anhedonic state. Although the internal consistency of the two subscales was high, no factor analysis was conducted to establish which of a range of possible items would best discriminate for anxiety and depression. Data were collected to test the internal consistency of the scales on only 50 patients and reliability tested on a further 50. These were very small samples, yielding only totals of 12 out of 100 cases of depression and 16 out of 100 cases of anxiety. Therefore it is possible that the depression items selected for the depression subscale, in particular, are inappropriate for screening in a cancer population. Only one study [8] has reported that the HADS has the ability to discriminate between the two types of morbidity.

One possible explanation for the findings could be that the interviewers using the PSE identified too many cases. Psychiatric morbidity from the whole sample of women in this study has been reported in detail elsewhere [4]. Rates of anxiety and depression were very similar to those found in comparable studies of women with early breast cancer. Among the women who completed both questionnaires after the postoperative interview only, the PSE identified 102/253 (40.3%) of patients with an anxiety disorder and 68/253 (26.9%) with depressive illness. The interviewers were well trained and experienced and both interrater and rate–rerate reliabilities were satisfactory. Clinical interviews using schedules such as the PSE to which DSM criteria are applied are considered as a gold standard because they use rigorous criteria for making a diagnosis of anxiety and depression. However, no interview schedule is completely reliable: Dean and colleagues [12] reported only 61% concordance in cases of psychiatric illness between the Psychiatric Assessment Schedule and the Present State Examination.

The discrepancies we found between caseness as identified by the interview schedule and the screening questionnaires may in part be explained by the different time frames employed. The PSE reviews how the patient has been feeling in the past 4 weeks. The HADS refers to the past week and the RSCL only to the previous 3 days. Therefore the PSE may have been identifying quite large numbers of women who experienced adjustment disorder but whose symptoms had abated or started to resolve within the last few days, as reflected in their questionnaire scores. Adjustment order, by definition, would be resolved by the end of the first postoperative year and indeed, cases of anxiety and depression had fallen to 66 and 49, respectively, by this time [4]. However, the HADS anxiety and depression subscales achieved sensitivities of only 61% and 45% respectively at a threshold of 7, and sensitivity at 45% on the unitary scale at a threshold value of 14 was little different to what was achieved during the first 3 months. The RSCL sensitivity of 69% at a threshold of 7 at 1 year was nearly identical to that achieved at the earlier time point. Therefore it seems unlikely that the performance of the questionnaires was affected by the presence of adjustment disorder, reflected in the application of differing time frames.

Table 5. *Number of patients needed for a two-arm clinical trial to detect* 30% *difference in outcome, with* 90% *power and two-sided* 5% *significance level*

		Sensitivity (%)					
		40	60	80	90	95	100
Specificity (%)	70	10 500	1270	460	310	260	220
	80	2400	680	310	230	200	180
	90	930	400	230	180	160	140
	95	620	320	190	150	140	120
	100	430	250	160	130	120	112

There are implications for sample size if instruments with poor sensitivity and specificity are used in clinical trial settings. This is illustrated by hypothetical examples in Table 5, which show the numbers of patients needed for a two-arm clinical trial, designed to detect a 30% difference in outcome with 90% power and a two-sided 5% significance level. If sensitivity and specificity were both 100%, then 112 patients would be required to detect such a difference. If however, the sensitivity of the instrument was only 60% and the specificity was 90%, then 400 patients would be needed. If the levels of sensitivity and specificity were similar to those reported in this paper (40 and 95%, respectively), then 620 patients would be needed to detect a 30% difference in outcome, using the HADS unitary scale with a threshold of 14. This is approximately six times as many patients as would be required if a clinical interview was used. A corollary is that instruments with poor sensitivity and specificity will fail to detect clinically important differences in psychological outcome in clinical trials.

In view of these findings, doubts must be raised about whether these instruments are sufficiently accurate for research and screening purposes. We are currently recommending the GHQ-12 [21] which is a well validated self-report questionnaire for the detection of psychological distress. The questionnaire does not include any somatic items, which may be attributable to disease and/or treatment. Patients are asked to evaluate their present psychological state relative to their normal functioning. In a large cross-cultural study, Goldberg and colleagues [22] found that overall, the GHQ-12 achieved a sensitivity of 83.4% and specificity of 76.3%. These values were raised to 84.65 and 89.3%, respectively, for the U.K. sample, with a threshold of 3/4 indicating a high probability of a clinical level of distress. The questionnaire is easy to score and the questions themselves serve as useful probes in clinic settings for doctors and nurses to ascertain with their patients whether some form of psychosocial intervention may be appropriate.

1. Hall A, Fallowfield L. Psychological outcome of treatment for early breast cancer: a review. *Stress Med* 1989, **5**, 167–175.
2. Spiegel D. Psychosocial aspects of breast cancer treatment. *Semin Oncol* 1997, **1**(1), 36–47.
3. Maunsell E, Brisson J, Deschenes L. Psychological distress after initial treatment for breast cancer: a comparison of partial and total mastectomy. *J Clin Epidemiol* 1989, **42**, 765–771.

4. Fallowfield L, Hall A, Maguire P, Baum M. Psychological outcomes of different treatment policies in women with early breast cancer outside a clinical trial. *Br Med J* 1990, **301**, 575–580.
5. Maguire P. Improving the recognition and treatment of affective disorders in cancer patients. In Granville-Grossman K, ed. *Recent Advances in Clinical Psychology*, Vol. 2. Edinburgh, Churchill Livingstone, 15–30.
6. Zigmond A, Snaith R. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand* 1983, **67**, 361–370.
7. De Haes J, Van Knippenberg F, Nejit J. Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. *Br J Cancer* 1990, **62**, 1034–1038.
8. Hopwood P, Howell A, Maguire P. Screening for psychiatric morbidity in patients with advanced breast cancer: validation of two self-report questionnaires. *Br J Cancer* 1991, **64**, 353–356.
9. Razavi D, Delavaux N, Farvacques C, Robaye E. Screening for adjustment disorders and major depressive disorders in cancer in-patients. *Br J Psychiatry* 1990, **156**, 79–83.
10. Murphy J, Berwick D, Weinstein M, Borys G, Budman S, Klerman G. Performance of screening and diagnostic tests: application of receiver operating characteristic analysis. *Arch Gen Psychiatr* 1987, **44**, 550–555.
11. Ibbotson T, Maguire P, Selby P, Priestman T, Wallace L. Screening for anxiety and depression in cancer patients: the effects of disease and treatment. *Eur J Cancer* 1994, **30A**, 37–40.
12. Dean C, Surtees P, Sashidharan P. Comparison of research diagnostic systems in an Edinburgh community sample. *Br J Psychiat* 1983, **142**, 247–256.
13. American Psychiatric Association on Nomenclature and Statistics. *Diagnostic and Statistical Manual of Mental Disorders* 1980. Washington DC, American Psychiatric Association.
14. Goldberg D, Cooper B, Eastwood M, Kedward H, Shepherd M. A standardised psychiatric interview for use in community surveys. *Br J Prevent Soc Med* 1970, **24**, 18.
15. Hopwood P, Howell A, Maguire P. Psychiatric morbidity in patients with advanced cancer of the breast: prevalence measured by two self-rating questionnaires. *Br J Cancer* 1991, **64**, 349–352.
16. Ramirez A, Richards M, Jarrett S, Fentiman I. Can mood disorder in women with breast cancer be identified preoperatively? *Br J Cancer* 1995, **72**, 1509–1512.
17. Wing J, Cooper J, Sartorius N. *Measurement and Classification of Psychiatric Symptoms*. Cambridge, Cambridge University Press, 1974.
18. Fallowfield L, Hall A, Maguire P, Baum M, A'Hern R. A question of choice: results of a prospective 3-year follow-up study of women with breast cancer. *The Breast* 1994, **3**, 202–208.
19. Cohen J. A coefficient of agreement for nominal scales. *Educ Psych Meas* 1960, **20**, 37–46.
20. Vecchio T. Predictive value of a single diagnostic test in unselected populations. *New Engl J Med* 1996, **274**, 1171.
21. Goldberg D, Williams P. *A Users Guide to the General Health Questionnaire*. Windsor, NFER-Nelson, 1988.
22. Goldberg D, Gater R, Sartorius N, *et al*. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol Med* 1997, **27**, 191–197.

Acknowledgement—The authors would like to thank the Cancer Research Campaign for funding this study.