ELSEVIER

Journal of Affective Disorders 46 (1997) 1–13

JOURNAL OF AFFECTIVE DISORDERS

Research report

Application of the Center for Epidemiologic Studies Depression Scale among first-visit psychiatric patients: a new approach to improve its performance

T. Furukawa[a,*], T. Hirai[b], T. Kitamura[c], K. Takahashi[b]

[a] *Department of Psychiatry, Nagoya City University Medical School, Mizuho-cho, Mizuho-ku, Nagoya 467, Japan*
[b] *Musashi Hospital, National Center for Neurology and Psychiatry, Tokyo, Japan*
[c] *National Institute of Mental Health, National Center for Neurology and Psychiatry, Ichikawa, Japan*

Received 4 December 1996; received in revised form 22 April 1997; accepted 22 April 1997

Abstract

Background: Although the Center for Epidemiologic Studies Depression Scale (CES-D) is an internationally popular self-rating scale for depression both in community and clinical settings, extant literature concerning its validity has several shortcomings. The present paper aimed to overcome these problems. *Methods*: We applied newer assessment technology of receiver operating characteristics (ROC) analyses and stratum-specific likelihood ratios (SSLRs) and cross-validated the results in the 'training' and 'testing' data sets of 591 patients representing various clinical settings all over Japan. *Results*: The ROC analyses demonstrated that the CES-D had moderate convergent and discriminant validity to detect major depressive episodes among first-visit psychiatric patients. Selecting single optimal cutoffs, however, failed to arrive at consistent results across various settings. The efficacy of the instrument was most conveniently transportable into clinical practices when converted into SSLRs, which were 0.35 (95%CI: 0.25–0.49) for the score range 0–29, 2.3 (1.8–3.1) for the score range 30–49, and 11.7 (3.1–44.0) for the scores above 50. In addition, the SSLRs proved to be generalizable not only across various clinical settings in our sample but also across psychiatric, primary care and community samples in the published reports. *Conclusion*: Clinicians and clinical epidemiologists can apply the SSLRs of the CES-D to various settings to estimate the probability of suffering from a major depressive episode in a convenient and intuitive manner. © 1997 Elsevier Science B.V.

Keywords: CES-D; Receiver operating characteristics; Stratum-specific likelihood ratios

1. Introduction

The Center for Epidemiologic Studies Depression Scale (CES-D) is a 20-item self-report scale which assesses the frequency/duration of cognitive, affective, behavioral and interpersonal symptoms associated with depression. It was originally developed to measure and detect depressive symptomatology in the community population (Radloff, 1977) but has

[*] Corresponding author. Tel.: +81 52 8538271; fax: +81 52 8520837; e-mail: gba02004@niftyserve.or.jp

0165-0327/97/$17.00 © 1997 Elsevier Science B.V. All rights reserved.
PII S0165-0327(97)00079-7

often been used in general medical settings (Coyne et al., 1994; Parikh et al., 1988; Schulberg et al., 1985; Turk and Okifuji, 1994; Zich et al., 1990) and with psychiatric patients (Craig and Van Natta, 1976, 1979; Faulstich et al., 1986; Hughes et al., 1993; Husaini et al., 1980; Roberts et al., 1990, 1989; Schulberg et al., 1985; Shrout and Yager, 1989; Weissman and Locke, 1975; Weissman et al., 1977). A cutoff score of 16/15 has traditionally been used to distinguish individuals considered to be depressed from those classified as non-depresssed (Comstock and Helsing, 1976; Weissman et al., 1977). The MEDLINE search reveals that the CES-D was used in 120 articles world-wide between the years 1993–1995, a figure next only to the Beck Depression Inventory (Beck et al., 1961) as a self-report measure of depression. In Japan the CES-D is the only self-report depression inventory whose semantic equivalence with the original English version has been ascertained by means of back translation (Shima et al., 1985). The standard cutoff was reported to be optimal with the Japanese sample as well (Shima et al., 1985). Despite this international popularity, the published data concerning the validity of the CES-D to detect depression have several shortcomings.

Firstly, many of the researchers, including the original developers (Craig and Van Natta, 1979; Radloff, 1977) and the Japanese translators (Shima et al., 1985) of the CES-D relied on the case-control method in which a group of typically diagnosable depressive disorder patients is compared with a group of unquestionably healthy subjects. In such a situation the ability of the CES-D to discriminate between the two groups can be overestimated. The failure to include an appropriately broad spectrum of the diseased and non-diseased subjects in the study population can give falsely high sensitivity and specificity: this effect is known as spectrum bias (Ransohoff and Feinstein, 1978).

Second, all the published papers used their whole data set as the ‘training’ set to arrive at certain conclusions but have not attempted to cross-validate the obtained results on a ‘testing’ set. (A ‘training’ set refers to a set of data that is originally used to derive a certain conclusion: a ‘testing’ set is another set of data to which the conclusion derived earlier can be applied and examined as to its transportability.) It is possible that the results derived from the ‘training’ set alone may be overfitting the sample, reflecting only variations within the sample and not variations representative of the underlying population. It is therefore important to cross-validate the obtained results and to test their transportability (Diamond, 1989; Leon et al., 1996).

Thirdly, the CES-D has not been subjected to newer diagnostic technology assessments such as the receiver operating characteristics (ROC) analyses and the stratum-specific likelihood ratios (SSLRs). Validity of a screening instrument has traditionally been expressed in terms of its sensitivity and specificity, which are theoretically independent of the base rate of the target disorder in the population (Yerushalmy, 1947). For an instrument that may take more than two values, however, sensitivity and specificity vary depending on the cutoff, and can therefore not be considered intrinsic to the instrument itself.

A more appropriate method to evaluate the efficacy of a diagnostic test with an ordinal or continuous scale is the ROC analysis (Swets, 1988). The ROC analysis has its origins during World War II, when signal detection theory was applied to radar to characterize how well a radar operator could receive a signal against a noisy background. It has since been an important part of a theory of human detection and recognition behavior. With the publication of the textbook by Swets and Pickett (1982), it has found rapidly increasing application in clinical medicine, especially in radiology and in clinical chemistry (Hanley and McNeil, 1982, 1983). It has been introduced into psychiatric research in the later 1980s by Mari and Williams (1985), Murphy et al. (1987) and Mossman and Somoza (1989). A further enhancement to the ROC analyses that is claimed by some authors as informative, intuitive and practical (Beck, 1986; Dujardin et al., 1994; Peirce and Cornell, 1993; Radack et al., 1986; Sackett et al., 1991) is the use of multi-level or stratum-specific likelihood ratios (SSLRs). To the present authors' knowledge, only three studies have applied the ROC and none has estimated the SSLRs for the CES-D.

The Group for Longitudinal Affective Disorders Study (GLADS) has been conducting a multi-center prospective follow-up study of a broad spectrum of affective disorders, including subthreshold minor depression, mixed anxiety-depression and adjustment disorder with depressed mood under the sponsorship of the Ministry of Health and Welfare, Japan (Furukawa et al., 1995). In the first stage of the collaborative study we examined representative samples of patients visiting the participating centers with a self-report test battery including the CES-D. The present paper focuses on the ability of the CES-D to detect DSM-III-R major depressive episode (major depressive episode of unipolar depression or of bipolar disorder) as determined by a psychiatrist's semi-structured interview among these untreated, first-visit psychiatric patients. We will overcome the above-mentioned shortcomings of the published reports of the CES-D and aim to find the most informative and the most generalizable method to interpret the CES-D scores by first deriving the optimal cutoffs and SSLRs based on the ROC analyses of the training set and then by examining their performances in the testing set as well as in various clinical settings with different base rates and clinical spectrums.

2. Methods

2.1. Patients and procedures

Subjects were 591 patients who constituted representative samples of the first-visit patients to 23 psychiatric hospitals and clinics participating in the GLADS Project during the study period, who had not received any psychotropic medication for the three months preceding their visit, who filled in the self-report test battery including the CES-D, and who were given the DSM-III-R diagnoses by psychiatrists using a semi-structured interview named the Psychiatric Initial Screening for Affective disorders (PISA)(Kitamura, 1992).

The 23 hospitals and clinics included psychiatric departments of 11 university hospitals, 7 general hospitals, 3 mental hospitals and an outpatient clinic, and a psychosomatic department of a university hospital from all over Japan. Each hospital and clinic examined a representative subset of its first-visit patients, selected according to the predetermined rules; in certain centers, a representative subsample meant all the first-visit patients examined by the psychiatrist(s) participating in the GLADS Project; in others, it meant all the first-visit patients on a certain day of the week; in still some others, it meant only the first such patient to show up on a certain day of the week. The selection of these preset rules was left to the individual center as time and human resources varied in each hospital.

The CES-D was scored according to the conventional four-point method; each item was scored between 0-3 and the possible total score ranged between 0–60. The score was considered missing if five or more items had been left unmarked. When one to four items were missing, the remaining subtotal score was multiplied by $20/(20 - \text{number of missing items})$ in order to obtain the total score corresponding to the normal 0–60 score range.

The DSM-III-R diagnoses were made by psychiatrists who were blind to the self-report battery results and who administered the PISA. The PISA lists 33 symptoms corresponding to diagnostic criteria of schizophrenia, mood disorders, anxiety disorders, somatoform disorders, dissociative disorders, organic mental disorders and substance use disorders, and the inter-rater reliability of these psychopathological variables has been reported to range between kappas of 0.71 and 1.00 (median = 0.85)(Furukawa et al., 1995).

2.2. Data analyses

In the following we will use the data from the first half of the patients as a training set to derive the optimal cutoff and SSLRs based on the ROC analyses. The ROC analyses and SSLRs were calculated by a computer program by Peirce and Cornell (1993) which used the nonparametric method described by Hanley and McNeil (1982). The generalizability of these cutoff and SSLR values will then be examined by applying them to the data from the second half of the patients as well as to the data from several clinical settings representative of differing base rates and clinical spectrums.

3. Results

3.1. Training set and ROC analyses

There were 747 patients who made their first visit to the 23 participating centers during the study period, who had not received any psychotropic medication for the preceding three months and who were assigned definitive axes I and/or II diagnoses according to the DSM-III-R by use of the PISA. Of these 747, 156 (20.7%) could not or did not complete the CES-D; in contrast to the 591 patients for whom the CES-D scores were available, these patients overrepresented organic mental disorders, schizophrenia and psychotic disorders not elsewhere classified, and were significantly older (mean±S.D. = 44.9±19.4 vs. 36.9±16.0; $t = 5.29$, df = 745, $p < 0.001$) but there was no significant difference in sex ($\chi^2 = 0.93$, df = 1, $p = 0.33$).

The basic demographic characteristics and the diagnostic classification of the 591 patients are presented in Table 1. The CES-D had internal consistency reliability with Cronbach's alpha coefficient of 0.82. Splitting the total sample into the training set and the testing set according to the date of the patients' visit did not result in any significant difference between the two sets in terms of age ($t = 0.27$, df = 589, $p = 0.79$), sex ($\chi^2 = 0.28$, df = 1, $p = 0.59$) or diagnoses ($\chi^2 = 19.28$, df = 13, $p = 0.11$).

By plotting the sensitivity along the vertical axis and the (1 − specificity) along the horizontal axis corresponding to all the possible cutoffs of the CES-D, we obtain the ROC curve for the training set as shown in Fig. 1. The area under the ROC curve (AUC) is equal to the probability that the test correctly identifies two subjects as normal or abnormal when one is randomly chosen from the normal group and the other is randomly chosen from the abnormal group. For a test that yields no information the AUC is an area under the diagonal, i.e. 0.50. The AUC in the training set was 0.79 (95% CI: 0.74–0.85).

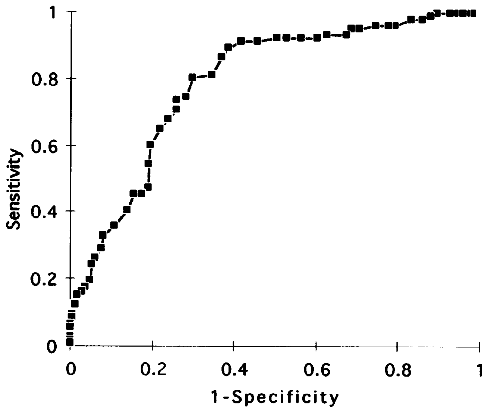

Fig. 1. ROC curve for the CES-D to detect DSM-III-R major depressive episode

In order to evaluate the influence of covariates on the operating characteristics of the CES-D, the ROC analyses were conducted separately for sex and age groups. For males ($n = 131$), the AUC was 0.82

Table 1
Demographic and diagnostic characteristics of the 591 subjects

Characteristic	
Age (mean±S.D.)	36.9±16.0
Sex (%)	Female = 323 (54.7%)
Diagnostic classification (DSM-III-R)	
Disorders usually first evident in infancy, childhood, or adolescence	20 (3.4%)
Organic mental disorders	16 (2.7%)
Psychoactive substance use disorders	17 (2.9%)
Schizophrenia	30 (5.1%)
Delusional disorder	12 (2.0%)
Psychotic disorders not elsewhere classified	19 (3.2%)
Mood disorders	230 (38.9%)
Anxiety disorders	77 (13.0%)
Somatoform disorders	38 (6.4%)
Dissociative disorders	10 (1.7%)
Sleep disorders	29 (4.9%)
Adjustment disorders	40 (6.8%)
Personality disorders	10 (1.7%)
V codes	30 (5.1%)
Others	13 (2.2%)

(95% CI: 0.74–0.89), and for females ($n = 165$), it was 0.79 (95%CI: 0.72–0.86); there was no statistically significant sex difference ($p = 0.98$). Nor was there any difference between the age groups; the AUC was 0.75 (95%CI: 0.66–0.85) for those under age 30 ($n = 125$), 0.84 (95%CI: 0.78–0.91) for those between 30 and 60 ($n = 136$), and 0.78 (95%CI: 0.62–0.95) for those above age 60 ($n = 35$). In the following, we will therefore analyze both sexes and all age groups together.

Some authors suggest that the CES-D is not a specific measure of depression but a general index for ‘demoralization’ (Breslau, 1985; Roberts et al., 1989). In order to test this hypothesis, we examined the ability of the CES-D to detect anxiety disorders, somatoform disorders and adjustment disorders among psychiatric patients. The respective AUCs were 0.36 (95%CI: 0.28–0.44), 0.31 (95%CI: 0.22–0.39), and 0.58 (95%CI: 0.41–0.74).

3.2. Optimal cutoffs

Several methods have been proposed in the literature to obtain the optimal cutoff from the ROC curve. Intuitively and as some authors have actually done (e.g. Roberts et al. (1991), Garrison et al. (1991), van Kammen et al. (1995), and Somervell et al. (1993), the cutoff point closest to the left upper corner of the ROC curve would appear to offer the optimal pair of sensitivity and specificity. Two factors, however, must be taken into account when selecting the optimal cutoff value: the base rate and the risk-benefit ratio between false positives and false negatives. The above-mentioned intuitive method is independent of these two important factors and hence can be misleading.

Kraemer (1992) developed the quality index, denoted $\kappa(w,0)$, which is a form of weighted kappa and measures the agreement between a test and a criterion standard. The choice of an optimal threshold score for a test is determined by the score with the greatest chance-corrected agreement with the criterion standard, i.e., the largest value for

$$\kappa(w,0) = \frac{w[\Pr(D+)\Pr(T-)\kappa(1,0)] + (1-w)[\Pr(D-)\Pr(T+)\kappa(0,0)]}{w[\Pr(D+)\Pr(T-)] + (1-w)[\Pr(D-)\Pr(T+)]} \tag{1}$$

where w is the weight given to sensitivity, Pr(T+) = the probability of having a positive test, Pr(T-) = the probability of having a negative test, Pr(D+) = the probability of having the disease, Pr(D-) = the probability of not having the disease, $\kappa(1,0)$ = (Sensitivity - Pr(T+))/Pr(T-), and $\kappa(0,0)$ = (Specificity - Pr(T-))/Pr(T+). When $w=1$, the equation becomes $\kappa(1,0)$ which is a quality index for sensitivity, and when $w=0$, the equation becomes $\kappa(0,0)$, a quality index for specificity. When $w=0.5$, sensitivity and specificity are equally weighted. Values for $\kappa(w,0)$ when $w=0.2$, 0.5 and 0.8 are plotted against the whole range of the CES-D scores in Fig. 2. As the formula suggests, $\kappa(w,0)$ is dependent on the base rate of the target disorder in the population and we postulated it to be the actual rate observed in the training set. The optimal cutoff is 26/25 when greater emphasis is placed on sensitivity, 31/30 when equal weight is placed on sensitivity and specificity, and 34/33 when specificity is deemed more important.

For the various cutoff values calculated above and for the conventional cutoff of 16/15, the sensitivity, specificity, and agreement with the PISA diagnosis of major depressive episode expressed in Cohen’s kappa are listed in Table 2.

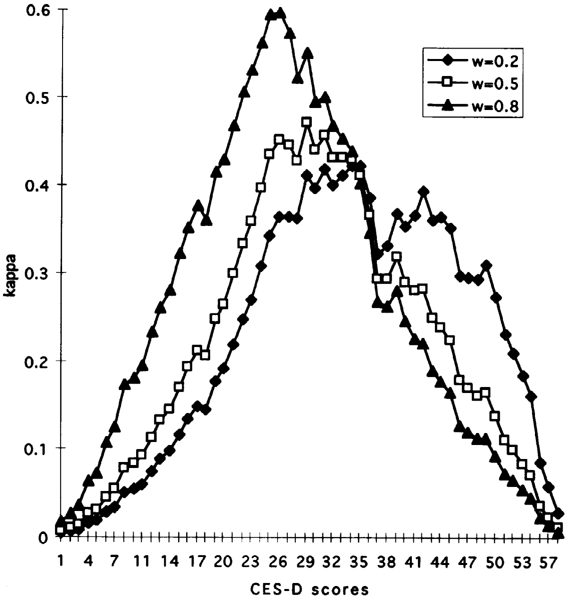

Fig. 2. Quality indices (weighted kappas) for the various cutoff scores of the CES-D

Table 2
Sensitivity, specificity and kappa for various cutoffs

Method	Optimal cutoff	Sensitivity(95% CI)	Specificity(95% CI)	Kappa(95% CI)
When specificity is more important (Kraemer, 1992)	34/33	0.65 (0.55–0.73)	0.78 (0.72–0.83)	0.43 (0.33–0.54)
When sensitivity and specificity are equally weighted (Kraemer, 1992)	31/30	0.74 (0.64–0.81)	0.74 (0.67–0.80)	0.46 (0.36–0.56)
When sensitivity is more important (Kraemer, 1992)	26/25	0.90 (0.82–0.94)	0.62 (0.54–0.68)	0.45 (0.36–0.54)
Traditional cutoff (Radloff, 1977)	16/15	0.95 (0.89–0.98)	0.29 (0.23–0.36)	0.20 (0.13–0.26)

3.3. Stratum-specific likelihood ratios

The ‘optimal’ cutoff values thus vary depending on the risk-benefit ratio and the base rate of the target disorder in the population at hand. In other words, precisely because we need to adjust the cutoff according to the purpose and circumstances of the use of a test, we need the ROC curve plotting all the possible cutoffs in order to evaluate the performance of the test per se. Selecting out a single cutoff means, in terms of the ROC analysis, to draw a line from the point (0,0) to the cutoff and another from the cutoff to the point (1,1), and the AUC is then almost always bound to be smaller than the original AUC. Much information is indeed lost when studies of test performance define sensitivity and specificity in relation to a single cutoff value of a continuous variable (Sox, 1986).

A way to avoid these pitfalls has recently been recommended by several authors (Beck, 1986; Peirce and Cornell, 1993; Radack et al., 1986; Sackett et al., 1991). It is the use of the multi-level or stratum-specific likelihood ratios (SSLRs). A likelihood ratio is a ratio of two probabilities, the probability of a given test result when the disease is present, divided by the probability of the same test result when the disease is absent. The likelihood ratio corresponds to the slope of the tangent of the ROC curve.

The likelihood ratio is used in Bayesian revision of odds where:

$$\text{Prior odds} \times \text{Likelihood ratio} = \text{Posterior odds} \qquad (2)$$

Conversions of odds to probability and probability to odds are performed using the formulae:

$$\text{Probability} = \frac{\text{odds}}{\text{odds} + 1}, \text{ and}$$

$$\text{Odds} = \frac{\text{probability}}{1 - \text{probability}} \qquad (3)$$

Thus the posterior (post-test) probability is greater than the prior (pre-test) probability, if the likelihood ratio is greater than 1.0; the former is equal to the latter, if the likelihood ratio is 1.0; and the former is smaller than the latter, if the likelihood ratio is smaller than 1.0.

Peirce and Cornell (1993) have developed a microcomputer spreadsheet program to arrive at the optimal number of strata of test scores by calculating likelihood ratios specific to different strata along with their 95% confidence intervals. Because with too many strata the likelihood ratios become unstable and degenerate, the following rules of thumb are recommended: (1) to provide sufficient abnormal and normal cases in each stratum to allow the SSLRs to be monotonically related, and (2) to collapse those strata where the SSLRs are close to one another and their 95% CI easily overlap.

The recommended SSLRs thus obtained from the training set were 0.35 (95%CI: 0.25–0.49) for the score range 0–29, 2.3 (1.8–3.1) for the score range 30–49, and 11.7 (3.1–44.0) for the scores above 50. For a given pre-test probability and an SSLR, one can estimate the post-test probability using Fagan’s nomogram (Fagan, 1975) or applying the formulae Eq. (2) and Eq. (3). For example, given the base rate of 36% for a major depressive episode in the training set, those with CES-D score above 50 have a post-test probability of 87% for the target disorder,

Table 3
SSLRs in the testing set and five representative clinical settings

CES-D score range		0–29		30–49		50–60		Hosmer-Lemeshow statistic
	pre-test probability	SSLR (95%CI)	post-test probability	SSLR (95%CI)	post-test probability	SSLR (95%CI)	post-test probability	
Training set ($n=296$)	35.8%	0.35 (0.25–0.49)	16.3%	2.3 (1.8–3.1)	56.2%	11.7 (3.1–44.0)	86.7%	
Testing set ($n=295$)	27.1%	0.66 (0.50–0.87)	19.7%	1.43 (1.09–1.89)	34.7%	5.38 (1.50–19.22)	66.7%	18.16, df = 3, $P=0.0004$
A ($n=51$)	19.6%	0	0%	2.05 (1.39–3.03)	33.3%	∞	100%	2.43, df = 3, $P=0.49$
B ($n=55$)	43.6%	0.54 (0.30–1.00)	29.5%	1.18 (0.65–2.16)	47.7%	∞	100%	4.31, df = 3, $P=0.23$
C ($n=61$)	42.6%	0.48 (0.22–1.00)	26.3%	1.12 (0.72–1.76)	45.4%	∞	100%	5.48, df = 3, $P=0.14$
D ($n=33$)	30.3%	0.43 (0.18–1.06)	15.7%	2.30 (1.14–4.65)	50.0%			0.10, df = 2, $P=0.95$
E ($n=33$)	15.2%	0	0%	9.33 (3.50–24.90)	62.6%	0	0%	7.85, df = 3, $P=0.05$

whereas those with CES-D score below 29 have a post-test probability of 16%. Those with the CES-D score between 30 and 49 have a post-test probability of 56%, one that only allows an indeterminate interpretation and calls for further examination. If the same test is applied to a population where the base rate is 20%, those with CES-D score above 50, between 49-30 and below 30 will have post-test probabilities of 75%, 37% and 8%, respectively. In a population with a base rate of 10%, the post-test probabilities would be 57%, 20% and 4%.

3.4. *Cross-validation of the findings*

We would next like to cross-validate the above findings by applying them to the data from the testing set as well as to those from five clinical settings representing different base rates and clinical spectrums. These were A university hospital psychiatric department with a base rate for major depressive episode of 20%, B general hospital psychiatric department with that of 44%, C psychiatric outpatient clinic with that of 43%, D mental hospital with that of 30% and E university hospital psychosomatic department with that of 15%. The various optimal cutoffs recommended according to Kraemer's quality indices and the tradition were applied to the data from the testing set and the five clinical settings, and the positive predictive values, negative predictive values, sensitivity, specificity and kappa values were calculated (the detailed results are available from the first author upon request).

The positive and negative predictive values are theoretically dependent on the base rate. As expected, some of the observed positive and negative predictive values significantly differed between the training set and that from C psychiatric outpatient clinic or from E university hospital psychosomatic department where the base rates widely differed. The positive and negative predictive values observed in the testing set also tended to be different from those in the training set as there was 9 point difference in the base rate between these two sets, although both the training and the testing sets were taken from the same hospitals.

In contrast, although sensitivity, specificity and kappa values are theoretically dependent on the clinical spectrum, the obtained sensitivity and specificity values were, overall, not statistically significantly different from those in the training set. The kappas ranged between 0.08 and 0.72 (median = 0.34), and only one-tenth of the obtained values were greater than 0.6 and showed satisfactory agreement between the results of the CES-D and the clinical diagnosis of a major depressive episode based on a semi-structured interview.

Table 3 shows the SSLRs for the testing set and the five clinical settings. The goodness-of-fit between the numbers predicted by the training set SSLRs and those actually observed in each stratum was examined with Hosmer-Lemeshow statistic (Lemeshow and Hosmer, 1982). This test answers the question of whether the observed and predicted rate of a major depressive disorder in all intervals jointly agree within the error range expected by chance alone. The SSLRs obtained from the training set resulted in significantly poor goodness-of-fit when applied to the testing set and to the E university hospital psychosomatic department, but resulted in satisfactory fit in the other settings. Table 3 also shows that, despite similar SSLRs, post-test probability of having a major depressive disorder varies greatly from setting to setting, depending on the pre-test probability of having the disorder.

4. Discussion

Approximately one in five of the first-visit patients to the psychiatric hospitals and clinics in the present study could not or did not complete the CES-D. This already low rejection rate should be interpreted with the understanding that the CES-D constituted part of the twelve-page-long self-rating test battery including two other questionnaires. Thus the CES-D appears to be well accepted by psychiatric patients in Japan as well. Swets (1988) suggested that AUCs of 0.5 to 0.7 indicate low test accuracy, 0.7 to 0.9 moderate accuracy, and >0.9 high accuracy. With the AUC of 0.79 or between 0.75–0.84 depending on age and sex in our sample, the CES-D can be said to have moderate accuracy in detecting major depressive episode among first-visit patients in psychiatric settings. As a matter of fact, the figure compares well with such routinely used tests in general medicine as the mean red cell volume to screen for iron-deficiency anemia (AUC = 0.76) and the fasting blood glucose to detect diabetes mellitus (AUC = 0.83) (Erdreich and Lee, 1981), or with short-term prediction of violence (AUC = 0.78)(Mossman, 1994).

To the present authors' knowledge, there have been only three studies which applied the ROC analyses and calculated the AUC values for the CES-D to detect depressive disorder as ascertained by a semi-structured interview. Among community adolescents, the CES-D has been reported to have AUCs between 0.61–0.77 (Garrison et al., 1991) or between 0.83–0.87 (Roberts et al., 1991) when calibrated against DSM-III-R major depression. Among nursing home residents, the reported AUC was 0.85 (Gerety et al., 1994). Thus, the ability of the CES-D to detect major depression in psychiatric settings appears to be comparable to that in community settings. Moreover, in the present sample, the AUCs of the CES-D to detect anxiety disorders or somatoform disorders or adjustment disorders, the three most common forms of psychiatric disorders which could be confused with depression, did not exceed the chance level. It can therefore be concluded that the CES-D has moderately good convergent and divergent validity for detecting major depressive episodes among first-visit psychiatric patients.

What is astonishing, however, is the fact that the optimal cutoffs suggested by the ROC analyses or otherwise in the literature have been wildly varying, even when we restrict our literature search to well designed studies where more than two cutoff scores of the CES-D are examined against some standardized psychiatric interview diagnoses. The conventional cutoff of 16/15, originally adopted as the lower bound of the upper quintile of scores for the general population (Comstock and Helsing, 1976) and termed 'arbitrary' by the developer of the instrument herself (Radloff, 1977), was however found optimal in some community (Katz et al., 1995; Myers and Weissman, 1980) and primary care (Parikh et al., 1988) settings as well as among psychiatric populations (Weissman et al., 1977). Several studies recommended higher cutoffs: based on the ROC analyses, Somervell et al. (1993) found the cutoff of 28/27 to maximize both sensitivity and specificity among American Indian village population; according to Roberts et al. (1991), 24/23 was the optimal cutoff to screen for major depression among American high school students; Cho et al. (1993) selected out 17/16 for Cuban Americans and 26/25 for Puerto Ricans in the community with both sensitivity and specificity around 0.90; Zich et al. (1990) recommended 27/26 with sensitivity of 1.00 and specificity of 0.81 for primary care clinic patients; Schulberg et al. (1985) also recommended 27/26 among newly admitted outpatients at primary medical care centers. Several researchers, however, recommended lower cutoffs: Garrison et al. (1991) found the cutoff of 12/11 to produce the best overall screening characteristics among community adolescent boys; based on calculations of Kraemer's quality indices, Gerety et al. (1994) recommended 13/12 as the optimal cutoff for the case finding of depression in the nursing home. None of these studies, however, have paid adequate attention to the problems of the base rate, the risk-benefit ratio between false positives and false negatives, and the spectrum bias. When these factors are taken into account, our analyses suggest that there are many 'optimal' cutoffs for the CES-D.

As theoretically predicted, the positive and negative predictive values for each recommended cutoff often differed significantly from those of the training set when the base rate was different. Our analyses thus demonstrated, as have been argued by Zarin and Earls (1993) and elsewhere, that applying a fixed 'optimal' cutoff to a population with a substantially different prior probability of the target disorder results in the 'selected' population which are constitutionally different. Failure to recognize the fact that the performance of any diagnostic or prognostic test is affected by the base rate in different clinical settings has been reported to lead to poor decision making or premature dismissal of valid rules in the area of general medicine such as coronary artery disease (Sox et al., 1990) and streptococcal pharyngitis (Poses et al., 1986). The fixed threshold approach cannot accommodate the wide variations actually present in clinical practices (Ruttimann, 1994). The same of course held for psychiatry.

On the contrary, the values for sensitivity and specificity remained largely constant across various clinical settings. This of course is no guarantee that the obtained sensitivity and specificity would apply to samples from primary care settings or from

Table 4
Comparison of SSLRs in the present study and the literature

ES-D score range [a]		0–19		0–29		30–60	
	pre-test probability	SSLR (95%CI)	post-test probability	SSLR (95%CI)	post-test probability	SSLR (95%CI)	post-test probability
Our study ($n=591$)	31.5%	0.22 (0.11–0.43)	9.2%	0.43 (0.28–0.67)	16.5%	3.09 (2.39–4.00)	58.7%
Community mental health center (Schulberg et al., 1985) ($n=269$)	27.5%	0.21 (0.08–0.53)	7.4%	0.64 (0.34–1.20)	19.5%	1.52 (1.28–1.80)	36.6%
Primary medical care center (Schulberg et al., 1985) ($n=294$)	9.2%	0.14 (0.04–0.47)	1.4%	0.71 (0.33–1.53)	6.7%	3.30 (2.41–4.50)	25.1%
Cuban American general population (Cho et al., 1993) ($n=87$)	13.8%	0.18 (0.06–0.54)	2.8%		12.78 (9.2–17.8)	67.2%	
Puerto Rican general population (Cho et al., 1993) ($n=364$)	18.1%	0.08 (0.03–0.18)	1.7%		4.78 (4.19–5.45)	51.4%	

[a] Because these authors provided detailed information only for the scores in the middle range of the CES-D, the strata are cut differently than those selected in our original analyses.

general population, but it appears that we have less of a problem of spectrum bias as far as the psychiatric hospitals and clinics participating in the present study are concerned.

A major problem is the less than satisfactory agreement between the CES-D results, whatever the 'optimal' cutoff selected, and the diagnosis of a major depressive episode according to the PISA semi-structured interview. The agreement measured in Cohen's kappa ranged between 0.08 and 0.72, depending on the cutoff and the clinical setting, with more than two thirds of the values below 0.4, which may be qualified as fair and demanding improvement at best (Kraemer, 1981).

A way to circumvent these problems and to reserve as much information as possible contained in the ROC curve is the use of SSLRs. Merits of the use of SSLRs can be enumerated as follows. Firstly, SSLRs retain as much information as possible that is originally contained in the continuous scale test by deriving multiple level indices instead of reducing the test result into a dichotomous value below or above the cutoff. Secondly, SSLRs, like the sensitivity and specificity but unlike the positive predictive value and negative predictive value and unlike the single 'optimal' cutoff, do not depend on the base rate of the target disorder. In addition, SSLRs are less subject to, although not totally free from, spectrum bias than a single cutoff and its associated sensitivity and specificity, because serious and less serious cases will tend to show up in their corresponding strata and the change in the mix of the more serious vs. less serious patients will have a smaller influence (Dujardin et al., 1994; Feinstein, 1990; Guyatt et al., 1992). Fourthly, SSLRs can be conveniently used at bedside to arrive at the post-test probability from the pre-test probability, using Fagan's nomogram (Fagan, 1975). Taking the base rate or prevalence as the pre-test probability, the post-test probability corresponds to the positive predictive value. SSLRs are thus clinically intuitive. Lastly, SSLRs will come in handy when we perform several tests in sequence, as we usually do in actual practices, because the post-test probability after one test will become the pre-test probability of the next text. We may thus be able to avoid redundant examinations if we stop the tests when we arrive at the post-test probability of, say, 95%.

The SSLRs obtained from our training set appeared to be constantly applicable in different clinical settings in our study. The goodness-of-fit statistic revealed significant difference between the training set and the testing set when the SSLRs calculated from the former were applied to the latter, but this should be interpreted with caution because almost any goodness-of-fit test would result in rejection of the null hypothesis when the sample size is large. Examination of the 95% confidence intervals for the SSLRs suggests that they do mostly overlap. The possible exception could be the E university hospital psychosomatic department. Among the patients visiting this psychosomatic department, the three most frequent diagnoses were mood disorders (27%), anxiety disorder (21%) and V codes (21%) and no organic or schizophrenic disorders were represented. Thus there appear to have been some major differences in spectrums represented at E University and at other sites.

The superior generalizability of SSLRs is further highlighted when our results are compared with those from the literature which provided enough information to calculate SSLRs for the CES-D Table 4). With the possible exception of the score range 20–60 for the Cuban Americans, the reported SSLRs are, to our pleasant surprise, quite overlapping. It is to remember that these authors had originally recommended widely and wildly different cutoffs as described above (27/26 in the study by Schulberg et al. (1985) and 17/16 or 20/19 that by Cho et al. (1993).

It is recommended that clinicians and clinical epidemiologists can use these SSLRs in a convenient and intuitive manner to calculate the post-test probability of suffering from a major depressive episode from its pre-test probability in various clinical and research settings.

Acknowledgements

This study has been undertaken on behalf of the Group for Longitudinal Affective Disorders Study (GLADS) and supported by the Research Grants 3A-6 and 6A-4 for Nervous and Mental Disorders from the Ministry of Health and Welfare, Japan. We thank Professor David L. Streiner, Department of

Psychiatry and of Clinical Epidemiology and Biostatistics, McMaster University, Canada, for his valuable comments on the manuscript.

References

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Arch. Gen. Psychiatry 4, 561–571.

Beck, J.R., 1986. Likelihood ratios: another enhancement of sensitivity and specificity. Arch. Pathol. Lab. Med. 110, 685–686.

Breslau, N., 1985. Depressive symptoms, major depression, and generalized anxiety: a comparison of self-reports on CES-D and results from diagnostic interviews. Psychiatry Res. 15, 219–229.

Cho, M.J., Moscicki, E.K., Narrow, W.E., Rae, D.S., Locke, B.Z., Regier, D.A., 1993. Concordance between two measures of depression in the Hispanic Health and Nutrition Examination Survey. Soc. Psychiatry Psychiatr. Epidemiol. 28, 156–163.

Comstock, G.W., Helsing, K.J., 1976. Symptoms of depression in two communities. Psychol. Med. 6, 551–563.

Coyne, J.C., Fechner-Bates, S., Schwenk, T.L., 1994. Prevalence, nature, and comorbidity of depressive disorders in primary care. Gen. Hosp. Psychiatry 16, 267–276.

Craig, T.J., Van Natta, P.A., 1976. Recognition of depressed affect in hospitalized psychiatric patients: staff and patient perceptions. Dis. Nerv. Syst. 37, 561–566.

Craig, T.J., Van Natta, P.A., 1979. Influence of demographic characteristics on two measures of depressive symptoms. Arch. Gen. Psychiatry 35, 149–154.

Diamond, G.A., 1989. Future imperfect: the limitations of clinical prediction models and limits of clinical prediction. J. Am. Coll. Cardiol. 14, 12A–22A.

Dujardin, B., van den Ende, J., van Gompel, A., Unger, J.-P., van der Stuyft, P., 1994. Likelihood ratios: a real improvement for clinical decision making?. Eur. J. Epidemiol. 10, 29–36.

Erdreich, L.S., Lee, E.T., 1981. Use of relative operating characteristic analysis in epidemiology: a method for dealing with subjective judgement. Am. J. Epidemiol. 114, 649–662.

Fagan, T.J., 1975. Nomogram for Bayes's theorem. New Engl. J. Med. 293, 257.

Faulstich, M.E., Carey, M.P., Ruggiero, L., Enyart, P., Gresham, F., 1986. Assessment of depression in childhood and adolescence: an evaluation of the Center for Epidemiological Studies Depression Scale for Children (CES-DC). Am. J. Psychiatry 143, 1024–1027.

Feinstein, A.R., 1990. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. J. Clin. Epidemiol. 43, 109–113.

Furukawa, T., Takahashi, K., Kitamura, T., Okawa, M., Miyaoka, H., Hirai, T., Ueda, H., Sakamoto, K., Miki, K., Fujita, K., Anraku, K., Yokouchi, T., Mizukawa, R., Hirano, M., Iida, S., Yoshimura, R., Kamei, K., Tsuboi, K., Yoneda, H., Ban, T.A., 1995. The Comprehensive Assessment List for Affective Disorders (COALA): a polydiagnostic, comprehensive, and serial semistructured interview system for affective and related disorders. Acta Psychiatr. Scand. Suppl. 387, 1–36.

Garrison, C.Z., Addy, C.L., Jackson, K.L., McKeown, R.E., Waller, J.L., 1991. The CES-D as a screen for depression and other psychiatric disorders in adolescents. J. Am. Acad. Child Adolesc. Psychiatry 30, 636–641.

Gerety, M.B., Williams, Jr. J.W., Mulrow, C.D., Cornell, J.E., Kadri, A.A., Rosenberg, J., Chiodo, L.K., Long, M., 1994. Performance of case-finding tools for depression in the nursing home: influence of clinical and functional characteristics and selection of optimal threshold scores. J. Am. Geriatr. Soc. 42, 1103–1109.

Guyatt, G.H., Oxman, A.D., Ali, M., Willan, A., McIlroy, W., Patterson, C., 1992. Laboratory diagnosis of iron-deficiency anemia. J. Gen. Int. Med. 7, 145–153.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic curve. Radiology 143, 29–36.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 829–843.

Hughes, D.C., DeMallie, D., Blazer, D.G., 1993. Does age make a difference in the effects of physical health and social support on the outcome of a major depressive episode?. Am. J. Psychiatry 150, 728–733.

Husaini, B.A., Neff, J.A., Harrington, J.B., Hughes, M.D., Stone, R.H., 1980. Depression in rural communities: validating the CES-D scale. J. Comm. Psychol. 8, 20–27.

Katz, R., Stephen, J., Shaw, B.F., Matthew, A., Newman, F., Rosenbluth, M., 1995. The East York Health Needs Study. I: Prevalence of DSM-III-R psychiatric disorder in a sample of Canadian women. Br. J. Psychiatry 166, 100–106.

Kitamura, T., 1992. Psychiatric Initial Screening for Affective disorders (PISA). National Center for Neurology and Psychiatry, Ichikawa.

Kraemer, H.C., 1981. Coping strategies in psychiatric clinical research. J. Consult. Clin. Psychol. 49, 309–319.

Kraemer, H.C., 1992. Evaluating Medical Tests: Objective and Quantitative Guidelines. SAGE Publications, Newbury Park.

Lemeshow, S., Hosmer, D.W.J., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. Am. J. Epidemiol. 115, 92–106.

Leon, A.C., Olfson, M., Weissman, M.M., Portera, L., Sheehan, D.V., 1996. Evaluation of screens for mental disorders in primary care: methodological issues. Psychopharmacol. Bull. 32, 353–361.

Mari, J.D.J., Williams, P., 1985. A comparison of the validity of two psychiatric screening questionnaires (GHQ-12 and SRQ-20) in Brazil, using relative operating characteristic (ROC) analysis. Psychol. Med. 15, 651–659.

Mossman, D., 1994. Assessing predictions of violence: being accurate about accuracy. J. Consult. Clin. Psychol. 62, 783–792.

Mossman, D., Somoza, E., 1989. Maximizing diagnostic information from the dexamethasone suppression test. Arch. Gen. Psychiatry 46, 653–660.

Murphy, J.M., Berwick, D.M., Weinstein, M.C., Borus, J.F., Budman, S.H., Klerman, G.L., 1987. Performance of screening and diagnostic tests: application of receiver operating characteristic analysis. Arch. Gen. Psychiatry 44, 550–555.

Myers, J.K., Weissman, M.M., 1980. Use of a self-report symptom scale to detect depression in a community sample. Am. J. Psychiatry 137, 1081–1084.

Parikh, R.M., Eden, D.T., Price, T.R., Robinson, R.G., 1988. The sensitivity and specificity of the Center for Epidemiologic Studies Depression Scale in screening for post-stroke depression. Int. J. Psychiatry Med. 18, 169–181.

Peirce, J.C., Cornell, R.G., 1993. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med. Decis. Making 13, 141–151.

Poses, R.M., Cebul, R.D., Collins, M., Fager, S.S., 1986. The importance of disease prevalence in transporting clinical prediction rules: the case of streptococcal pharyngitis. Ann. Intern. Med. 105, 586–591.

Radack, K.L., Rouan, G., Hedges, J., 1986. The likelihood ratio: an improved measure for evaluating diagnostic test results. Arch. Pathol. Lab. Med. 110, 689–693.

Radloff, L.S., 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl. Psych. Meas. 1, 385–401.

Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New Engl. J. Med. 299, 926–929.

Roberts, R.E., Lewinsohn, P.M., Seeley, J.R., 1991. Screening for adolescent depression: a comparison of depression scales. J. Am. Acad. Child Adolesc. Psychiatry 30, 58–66.

Roberts, R.E., Rhoades, H.M., Vernon, S.W., 1990. Using the CES-D scale to screen for depression and anxiety: effects of language and ethnic status. Psychiatry Res. 31, 69–83.

Roberts, R.E., Vernon, S.W., Rhoades, H.M., 1989. Effects of language and ethnic status on reliability and validity of the Center for Epidemiologic Studies-Depression Scale with psychiatric patients. J. Nerv. Ment. Dis. 177, 581–592.

Ruttimann, U.E., 1994. Statistical approaches to development and validation of predictive instruments. Crit. Care Clin. 10, 19–35.

Sackett, D.L., Haynes, R.B., Guyatt, G.H. and Tugwell, P., 1991. Clinical Epidemiology: A Basic Science for Clinical Medicine. (2nd edn). Little, Brown and Company, Boston/Toronto/London.

Schulberg, H.C., Saul, M., McClelland, M., Ganguli, M., Christy, W., Frank, R., 1985. Assessing depression in primary medical and psychiatric practices. Arch. Gen. Psychiatry 42, 1164–1170.

Shima, S., Shikano, T., Kitamura, T., Asai, M., 1985. New self-rating scales for depression (in Japanese). Seishin-Igaku (Clinical Psychiatry) 27, 717–723.

Shrout, P.E., Yager, T.J., 1989. Reliability and validity of screening scales: effect of reducing scale length. J. Clin. Epidemiol. 42, 69–78.

Somervell, P.D., Beals, J., Kinzie, J.D., Boehnlein, J., Leung, P., Manson, S.M., 1993. Criterion validity of the Center for Epidemiologic Studies Depression Scale in a population sample from an American Indian village. Psychiatry Res. 47, 255–266.

Sox, H.C.J., 1986. Probability theory in the use of diagnostic tests. Ann. Int. Med. 104, 60–66.

Sox, H.C.J., Hickam, D.H., Marton, H.I., Moses, L., Skeff, K.M., Sox, C.H., Neal, E.A., 1990. Using the patient's history to estimate the probability of coronary artery disease: a comparison of primary and referral practices. Am. J. Med. 89, 7–14.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

Swets, J.A. and Pickett, R.M., 1982. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Academic Press, New York.

Turk, D.C., Okifuji, A., 1994. Detecting depression in chronic pain patients: adequacy of self-reports. Behav. Res. Ther. 32, 9–16.

van Kammen, D.P., Kelley, M.E., Gurklis, J.A., Gilbertson, M.W., Yao, J.K., Peters, J.L., 1995. Behavioral vs biochemical prediction of clinical stability following haloperidol withdrawal in schizophrenia. Arch. Gen. Psychiatry 52, 673–678.

Weissman, M.M., Locke, B.Z., 1975. Comparison of a self-report symptom rating scale (CES-D) with standardized depression rating scales in psychiatric populations. Am. J. Epidemiol. 102, 430–431.

Weissman, M.M., Sholomskas, D., Pottenger, M., Prusoff, B.A., Locke, B.Z., 1977. Assessing depressive symptoms in five psychiatric populations: a validation study. Am. J. Epidemiol. 106, 203–214.

Yerushalmy, J., 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Rep. 62, 1432–1449.

Zarin, D.A., Earls, F., 1993. Diagnostic decision making in psychiatry. Am. J. Psychiatry 150, 197–206.

Zich, J.M., Attkisson, C.C., Greenfield, T.K., 1990. Screening for depression in primary care clinics: the CES-D and the BDI. Int. J. Psychiatry Med. 20, 259–277.