The Mental Component of the Short-Form 12 Health Survey (SF-12) as a Measure of Depressive Disorders in the General Population: Results with Three Alternative Scoring Methods

Gemma Vilagut, MSc1,2,3,*, Carlos G. Forero, PhD2,1, Alejandra Pinto-Meza, PhD4,5, Josep M. Haro, MD, PhD4,5, Ron de Graaf, PhD6, Ronny Bruffaerts, PhD7, Viviane Kovess, MD, PhD8, Giovanni de Girolamo, MD9, Herbert Matschinger, PhD10, Montserrat Ferrer, MD, PhD1,2,11, Jordi Alonso, MD, PhD1,2,3, on behalf of the ESEMeD Investigators

1Health Services Research Unit, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain; 2CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain; 3Pompeu Fabra University (UPF), Barcelona, Spain; 4Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain; 5CIBER en Salud Mental (CIBERSAM), Madrid, Spain; 6Netherlands Institute of Mental Health and Addictions, Utrecht, The Netherlands; 7Katholieke Universiteit Leuven, Leuven, Belgium; 8París Descartes University and Department of Epidemiology, EHESP, Paris, France; 9IRC – St. John of God, Clinical Research Centre, Brescia, Italy; 10Institute of Social Medicine, Occupational Health and Public Health, University of Leipzig, Leipzig, Germany; 11Universitat Autònoma de Barcelona (UAB), Barcelona, Spain.

ABSTRACT

Objectives: To evaluate the performance of the Mental Component of the Short-Form 12 Health Survey, Version 1(SF-12v1), as a screening measure of depressive disorders. Methods: Data come from the European Study of the Epidemiology of Mental Disorders (ESEMeD), a cross-sectional survey carried out on representative samples of 21,425 individuals from the noninstitutionalized adult general population of six European countries (response rate = 61.2%). The SF-12 was administered and scored according to three algorithms: the “original” method (mental component summary of SF-12 [MCS-12]), the RAND-12 (RAND-12 Mental Health Composite [RAND-12 MHC]), and the Bidimensional Response Process Model 12 mental health score (BRP-12 MHS), based on a two-factor Item Response Theory graded response model. Thirty-day and 12-month depressive disorders (major depressive episode or dysthymia) were assessed with the Composite International Diagnostic Interview, Version 3.0, by using Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition criteria. Receiver operating characteristic curves analysis was carried out, and optimal cutoff points maximizing balance between sensitivity (SN) and specificity (SP) were chosen for the three methods. Results: Prevalence of 30-day and 12-month depressive disorders in the overall sample was 1.5% and 4.4%, respectively. The area under the curve for 30-day depressive disorders was 0.92, and it decreased to 0.85 for 12-month disorders, regardless of the scoring method. Optimal cutoff for 30-day depressive disorders was 45.6 (SN = 0.86; SP = 0.88) for the MCS-12, 44.5 for the RAND-12 MHC (SN = 0.87, SP = 0.86), and 40.2 for the BRP-12 MHS (SN = 0.87, SP = 0.87). The selected 12-month cutoffs for MCS-12 and RAND-12 MHC were between 4.2 and 5.8 points below the general population means of each country, with SN range 0.67 to 0.78 and SP range 0.77 to 0.87. Conclusions: The SF-12 yielded acceptable results for detecting both active and recent depressive disorders in general population samples, suggesting that the questionnaire could be used as a useful screening tool for monitoring the prevalence of affective disorders and for targeting treatment and prevention.

Keywords: depressive disorders, diagnostic accuracy, health-related quality-of-life, mental health screening, mental disorders, screening, SF-12.

Copyright © 2013, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.
The performance of the SF-12 for the assessment of mental disorders in the general population, however, has been scarcely studied. To our knowledge, only one study has assessed its diagnostic accuracy, reporting an area under the curve (AUC) of 0.92 to predict major depression [4]. Even though these authors provide cutoff points of the SF-12 for the Australian general population, the question is whether these cutoff points are adequate for Europe.

The items of the SF-12 were selected to reproduce the two summary measures, Physical Component Summary (PCS) and Mental Component Summary (MCS), of the SF-36. The SF-12 scoring method proposed by Ware et al. [7] assumes that each item contributes to both physical component summary and mental component summary (MCS-12) and that these two measures are uncorrelated. This scoring method, however, does not necessarily optimize the information contained within the items. Alternative scoring methods for the SF-12 have been proposed: the RAND-12 Health Status Inventory [8] and the Bidimensional Response Process Model algorithm (BRP-12) based on the Item Response Theory (IRT) [9] (Forero, under review). The scores derived from the RAND-12 represent composite estimates of the corresponding RAND-36 Health Status Inventory Physical Health Composite and Mental Health Composite (MHC). The RAND-36 and RAND-12 algorithms avoid item ambiguities by letting the items load on just one factor, but factors are allowed to correlate by means of an oblique rotation method. Thus, in the RAND-12, six of the items contribute to the Physical Health Composite and the remaining ones to the MHC dimensions [8]. Conversely, the BRP-12 scoring is based on a two-factor IRT Graded Response Model [10] directly applied to the SF-12 items, where all items are allowed to load on both dimensions, much in the fashion of the SF-12 MCS-12, and the correlation between both dimensions is set to 0. Different from the classical SF-12 and the RAND-12 models, BRP-12 scores do not serve as a surrogate measure for the 36-item versions and all information is extracted from the 12 items. Scores obtained with weighted combinations of the same items are expected to be highly correlated; however, different weights have great impact on model reliability. As an IRT model, the BRP-12 mental health score (MHS) obtains a set of weights that maximizes reliability [11]. In our case, the RAND-12 MHS is more reliable than the MCS-12 and the RAND-12 MHC in terms of model-based internal consistency reliability (the proportion of observed variance attributable to the factor model underlying the score) [12], with a value of 0.77 for the BRP-12 MHS, as compared with the obtained values of 0.66 for the MCS-12 and 0.67 for the RAND-12 MHC. It is not clear, however, whether the screening accuracy of the instrument for depressive disorders differs according to the scoring method [9] (Forero, under review).

In the present study, we aimed to evaluate the performance of the SF-12 to detect depressive disorders in the general population. We compared classification abilities of three scoring methods (MCS-12, RAND-12 MHC, and BRP-12 MHS) by using data from a representative sample from the general population of six European Countries [13]. Results were obtained both for the whole European sample and by country. In addition, we aimed to estimate the best cutoff point for each of the proposed methods for screening purposes of depressive disorders in Europe.

The methods used for data collection have been described elsewhere [13]. Briefly, a stratified, multistage, clustered area probability sample of noninstitutionalized adult population (aged 18 years or older) in Belgium, France, Germany, Italy, The Netherlands, and Spain was selected. The questions were administered by trained lay interviewers at the respondent’s house between January 2001 and August 2003 by using computer-assisted personal interview techniques. The total sample size achieved was 21,425 individuals, with an overall weighted response rate of 61.2%, ranging from 45.9% in France to 78.6% in Spain.

**Measures**

**Mental Disorders**

Mental disorders were assessed by using version 3.0 of the World Health Organization Composite International Diagnostic Interview (CIDI 3.0) [14], a fully structured lay administered diagnostic interview designed to assess the presence of most common mental disorders following the definitions and criteria of both the DSM-IV [2] and the ICD-10 [3] classification systems. Here we consider the DSM-IV diagnostics of common disorders of the depressive spectrum (major depression episode or dysthymia).

We assessed whether respondents fulfilled criteria for these disorders any time in the previous 30 days (30-day disorders) and 12 months (12-month disorders). We decided to look at the two recall periods to determine whether the SF-12 questionnaire was sensitive to both active and recent episodes, even though the disorder may not be present anymore at the interview time.

A clinical reappraisal study with blinded clinical follow-up interviews using the Structured Clinical Interview for DSM-IV [15] in several surveys (France, Italy, Spain, US) found generally good concordance between diagnoses based on the CIDI 3.0 and those based on the Structured Clinical Interview for DSM-IV [16].

The SF-12

SF-12v1 was used because version 2 was not available when the ESEMeD surveys were designed. The standard form, with a recall period of 4 weeks in most of the items but three, was administered to all respondents. We focus on the mental summary measure of the SF-12, which has been obtained following two already available scoring methods: the “original” MCS-12 scores proposed by Ware et al. [7] and the RAND-12 MHC proposed by Hays [8].

The MCS-12 score is calculated by using US-derived item weights for response categories following recommendations from the authors of the instrument (which were done after having assessed the equivalence between country-specific and US weights in nine countries including most of the countries evaluated here) when international comparisons are to be conducted [17,18]. The weights to be applied are the coefficients of a linear regression model that was estimated on a representative sample of the US general population to predict the MCS of the SF-36 from a set of dummy variables defining all but one item response categories of each of the 12 items of the SF-12. The RAND-12 MHC, in turn, also applies response category weights that were obtained from one-parameter IRT models on each of the eight RAND-36 scales. Moreover, additional scoring weights, obtained from a linear regression model of the RAND-36 MHC composite on six IRT-weighted items that contribute to the mental score, were applied to each item. Both the MCS-12 and the RAND-12 MHC use norm-based scoring, where the mental summary measures have a mean of 50 and an SD of 10 in the US general population and scores greater (lower) than 50 reflect better (worse) mental health status than the US general population.

An alternative scoring method for the SF-12 has been proposed [9] (Forero et al., under review) and is applied here, the BRP-12.

**Sample Description**

Data come from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project, a cross-sectional survey conducted in six European countries to study the prevalence and correlates of mental disorders.

**Methods**

**Value in Health** 16 (2013) 564–573
This is an IRT scoring algorithm based on the application of a multidimensional Graded Response Model to SF-12 items [10], where items define two uncorrelated latent traits (physical and mental) to explain the individual response process. The distinctive feature of the BRP-12 model is that all items are indicators of both dimensions, so that item responses depend on the individual location on both health factors. This model yields good results in terms of fit (root mean square error of approximation = 0.057; comparative fit index = 0.95; Tucker Lewis index = 0.94) and overall model-based internal-consistency reliability [12] of 0.77 for the MHS. Model estimation of the BRP-12 was conducted by using Marginal Maximum Likelihood via the Expectation-Maximization algorithm. Individual scores were computed as Expected a Posteriori [19] estimates of the latent trait (i.e., mental health factor) person parameter of each individual and transformed into T scores, with an average of 50 and an SD of 10 in our sample. Higher scores are indicative of better mental health status.

Statistical analysis
The 30-day and 12-month prevalence of depressive disorders was estimated for the overall sample and for each country separately. Mean values and 95% confidence interval (CI) of the mental scores of the SF-12 using the three scoring methods were also obtained for the overall sample and for each country.

The discriminant validity of the three scoring methods for each target diagnosis was assessed with the receiver operating characteristic (ROC) curves that graph the SN against the false-positive rate (1 – SP) of every possible value of the scale. Non-parametric methods were used to calculate the AUCs and corresponding 95% CIs of each scoring method for each diagnosis of interest. The AUC evaluates discrimination ability and can be interpreted as the probability that a randomly selected case according to the CIDI will score higher on the mental component of the SF-12 than will a randomly selected noncase [20]. AUC values range from 0.5 to 1.0. It takes the value 0.5 when no discrimination exists, that is, the scale is performing at a chance level, and 1.0 when there is perfect discrimination. It has been suggested that AUC values from 0.5 to 0.7 represent rather low discrimination ability, between 0.7 and 0.9 correspond to moderate discrimination ability, and values greater than 0.9 are considered as high [21].

Screening cutoff points on each of the three scoring methods were selected to detect 30-day and 12-month depressive disorders. The best cutoff point was selected on the basis of empirical Youden Index, that is, the observed value that maximizes SN + SP – 1. This index has been suggested as a summary measure that assumes that both SN and SP are equally important [22].

The SN (the percentage of true cases correctly classified by a certain cut point of the SF-12), SP (the percentage of true noncases correctly classified) [22], positive likelihood ratio (the probability of a person who has the disease testing positive divided by the probability of a person who does not have the disease testing positive), and negative likelihood ratio (the probability of a person who has the disease testing negative divided by the probability of a person who does not have the disease testing negative) were obtained for the selected cutoff points [23]. The diagnostic odds ratio (DOR) and the AUC of the SF-12 measures dichotomized at the selected thresholds to predict the presence of the mental disorders of interest were also calculated. Although the AUC was originally developed to study the association between a continuous predictor and a dichotomous outcome, the AUC can be used when the predictor is a dichotomous variable, in which special case AUC equals (SN + SP)/2. The AUC has been selected over the more popular individual-level concordance statistic Cohen’s Kappa (κ) because the latter depends on prevalence and consequently is often low in situations in which there appears to be high agreement between low-prevalence measures [24].

The analysis regarding 12-month depressive disorders was stratified by country. Results for 30-day diagnostic were carried out only for the overall sample because of the low number of individuals with active depressive disorders within each of the countries.

All analyses were conducted in SAS™ software, version 9.0, of the SAS system for windows [25], and SUDDAAN software, version 10.0 [26]. Individuals were weighted to account for the different probabilities of selection as well as to restore age and gender distribution of the population within each country and the relative dimension of the population across countries. The

---

Fig. 1 – Thirty-day and 12-month prevalence of depressive disorders* in the total ESEMeD sample and by country. 12-month depressive disorders (in black), 30-day depressive disorders (in white).

*Depressive disorders: major depression episode or dysthymia based on DSM-IV criteria as assessed with the CIDI 3.0.
standard errors were estimated by taking into account the complex sample design by using the Taylor series linearization method.

Results

The mean age of the sample was 47.1 (SD = 29.3) years, with most individuals in the category of 35 to 49 years of age and 20% of the sample older than 65 years. Fifty-two per cent of them were female, and 67% were married or cohabiting. Over one-third of the sample had received more than 12 years of education [13].

Depressive disorders in the past 30 days were present in 1.5% of the sample, with values across countries ranging from 1.1% (Germany) to 1.9% (Spain). The prevalence of 12-month depressive disorders in the total sample was 4.4%, and there were significant differences across countries, from 6.5% in France to 3.4% in Italy and 3.3% in Germany. A prevalence of around 5% was found in Belgium and The Netherlands and 4.4% in Spain (Fig. 1).

Fig. 2 – Mean values (95% CI) of the mental component of the SF-12 in the ESEMeD sample, according to scoring method* and country. ESEMeD sample. MCS-12 \( ^{(1)} \) (black diamonds), RAND-12 MHC \( ^{(2)} \) (grey squares), BRP-12 MHS \( ^{(3)} \) (white circles).

* MCS-12 and RAND-12 MHC use norm-based scoring with Mean = 50 and Standard deviation = 10 in the US General Population; BRP-12 scores were transformed to T metric with a mean = 50 and Standard Deviation = 10 in the ESEMeD sample.

\( ^{(1)} \) MCS-12: Original Mental Component Summary.

\( ^{(2)} \) RAND-12 MHC: RAND-12 Mental Health Composite.

\( ^{(3)} \) BRP-12 MHS: Bidimensional Response Process model Mental Health Score.

Table 1 – Areas under the curve (AUCs) for the mental component of the SF-12v1 against depressive disorders*, according to scoring method and country.

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Country</th>
<th>AUC (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MCS-12</td>
</tr>
<tr>
<td>30-d depressive disorders</td>
<td>Overall sample</td>
<td>0.93 (0.01)</td>
</tr>
<tr>
<td></td>
<td>Overall sample</td>
<td>0.85 (0.01)</td>
</tr>
<tr>
<td></td>
<td>Belgium</td>
<td>0.83 (0.02)</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>0.81 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>0.88 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Italy</td>
<td>0.86 (0.02)</td>
</tr>
<tr>
<td></td>
<td>The Netherlands</td>
<td>0.86 (0.02)</td>
</tr>
<tr>
<td></td>
<td>Spain</td>
<td>0.84 (0.02)</td>
</tr>
</tbody>
</table>

BRP-12 MHS, Bidimensional Response Process Model Mental Health Score; CIDI 3.0, Composite International Diagnostic Interview Version 3.0; DSM-IV, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; MCS-12, Original Mental Component Summary; RAND-12 MHC, RAND-12 Mental Health Composite; SE, standard error; SF-12v1, Short-Form 12 Health Survey Version 1.

* Depressive disorders: major depression episode or dysthymia based on DSM-IV criteria as assessed with the CIDI 3.0.
(worst) scores (mean = 52.6). Conversely, Belgium, Germany, and The Netherlands showed the highest scores, with mean values near 54.4. Scores for the RAND-12 MHC were significantly lower than for MCS-12 in most of the countries, with a mean value of 52.9 for the overall sample. The BRP-12 scores were transformed into the T metric (mean = 50, SD = 10).

The rank order correlations among the three scoring methods ranged between 0.89 and 0.97, the correlation between the MCS-12 and the RAND-12 MHC being the lowest and the correlation between the BRP-12 and the RAND-12 being the highest. These results were consistent across countries (results not shown but available under request).

Table 1 presents the AUC for the three scoring methods against 30-day and 12-month depressive disorders for the overall sample. The AUC value for 30-day depressive disorders was high, above 0.92 according to the three scoring methods, while it decreased to a moderate value of 0.85 for 12-month depressive disorders. There were no statistically significant differences between the AUCs of the three scoring methods for both the 30-day and the 12-month assessments. With regard to the comparison between countries, the AUC for 12-month depressive disorders was higher in Germany (around 0.89), Italy, and The Netherlands (0.86 for the three methods) than in Belgium (around 0.83) and France (above 0.82 for the three scoring methods).

As shown in Table 2, the best screening cut point for MCS-12 to evaluate 30-day depressive disorders was 45.6, with SN = 0.86 (0.02) and SP = 0.88 (0.003), which lead to a positive likelihood ratio of 6.9 (95% CI 6.5–7.3) and a negative likelihood ratio of 0.2 (95% CI 0.1–0.2). The cutoff point of 44.5 for the RAND-12 MHC and 40.2 for the BRP-12 provided similar values of SN and SP. A score of 48.9 was chosen as the best screening cutoff point for the MCS-12 against 12-month depressive disorders, with SN = 0.74 and SP = 0.83. The best cutoff point chosen for the RAND-12 MHC was 47.9, with an SN of 0.77 and an SP of 0.79 for the whole sample. For the BRP-12 score, the selected cutoff point that provided best trade-off between SN and SP was 41.9, with similar SN as the MCS-12 (0.73) but slightly higher SP (0.84).

There were statistically significant differences across countries with regard to the sensitivities and specificities achieved by the suggested cutoff point to screen for 12-month depressive disorders according to the three scoring methods (Table 3). Sensitivities for the MCS-12 ranged from 0.67 in The Netherlands to 0.78 in Germany, and SP values ranged from 0.79 in France and Spain to 0.87 in The Netherlands.

Discussion

This study evaluated the screening performance of the mental component of the SF-12 to identify individuals with depressive disorders, using data from representative samples of the adult populations of six European countries. The performance of the original scoring method of the SF-12, proposed by Ware et al. [7], was compared with that of two alternative scoring methods, the RAND-12 [6] and the BRP-12 [9] (Forero et al., under review). The results show AUCs around 0.92 for 30-day disorders, accuracy comparable to that of other widely used specific screening measures [27,28]. The AUC for 12-month depressive disorders, as expected, is a little lower but still high (around 0.85). This suggests that beyond the usual applications of the SF-12, this short questionnaire could be a useful screening tool for the assessment of affective disorders in general population surveys, such as Health Interview Surveys, in which this instrument is often included.

No statistically significant differences are found across the three scoring methods that have been compared in terms of accuracy to detect depressive disorders. The scoring method suggested by Farivar et al. [29] has been additionally tested providing similar results to the other three, with an AUC value of 0.93 (0.01). This
### Table 2 – Accuracy of the SF-12v1 for depressive disorders* (30-d and 12-mo) at selected cutoff points according to scoring method, in the ESEMeD sample.

<table>
<thead>
<tr>
<th>Scoring method and cut point</th>
<th>Prevalence SF-12</th>
<th>Prevalence depressive disorders</th>
<th>Sensitivity (SN)</th>
<th>Specificity (SP)</th>
<th>LR+</th>
<th>Low</th>
<th>Upp</th>
<th>LR−</th>
<th>Low</th>
<th>Upp</th>
<th>Design-adjusted odds ratio</th>
<th>Design-adjusted AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>%</td>
<td>%</td>
<td>%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>30-d depressive disorders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCS-12 ≤ 45.6</td>
<td>13.5</td>
<td>0.3</td>
<td>1.5</td>
<td>0.1</td>
<td>0.86</td>
<td>0.02</td>
<td>0.88</td>
<td>0.003</td>
<td>6.9</td>
<td>6.5</td>
<td>7.3</td>
<td>0.2</td>
</tr>
<tr>
<td>RAND-12</td>
<td>15.3</td>
<td>0.4</td>
<td>1.5</td>
<td>0.1</td>
<td>0.87</td>
<td>0.02</td>
<td>0.86</td>
<td>0.004</td>
<td>6.1</td>
<td>5.8</td>
<td>6.4</td>
<td>0.2</td>
</tr>
<tr>
<td>MHC ≤ 44.5</td>
<td>14.2</td>
<td>0.4</td>
<td>1.5</td>
<td>0.1</td>
<td>0.87</td>
<td>0.02</td>
<td>0.87</td>
<td>0.003</td>
<td>6.6</td>
<td>6.3</td>
<td>7.0</td>
<td>0.2</td>
</tr>
<tr>
<td>BRP-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHS ≤ 40.2</td>
<td>19.5</td>
<td>0.4</td>
<td>4.3</td>
<td>0.2</td>
<td>0.74</td>
<td>0.02</td>
<td>0.83</td>
<td>0.004</td>
<td>4.3</td>
<td>4.1</td>
<td>4.5</td>
<td>0.3</td>
</tr>
<tr>
<td>MHC ≤ 48.9</td>
<td>23.1</td>
<td>0.5</td>
<td>4.3</td>
<td>0.2</td>
<td>0.77</td>
<td>0.02</td>
<td>0.79</td>
<td>0.005</td>
<td>3.7</td>
<td>3.5</td>
<td>3.9</td>
<td>0.3</td>
</tr>
<tr>
<td>BRP-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHS ≤ 41.9</td>
<td>18.0</td>
<td>0.4</td>
<td>4.3</td>
<td>0.2</td>
<td>0.73</td>
<td>0.02</td>
<td>0.85</td>
<td>0.004</td>
<td>4.7</td>
<td>4.5</td>
<td>5.0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Note. Estimates and standard errors (SEs) or 95% confidence interval.

AUC, area under the curve; BRP-12 MHS, Bidimensional Response Process Model Mental Health Score; CIDI 3.0, Composite International Diagnostic Interview, Version 3.0; DSM-IV, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; ESEMeD, European Study of the Epidemiology of Mental Disorders; Low, lower limit 95% confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio; MCS-12, Original Mental Component Summary; RAND-12 MHC, RAND-12 Mental Health Composite; SE, standard error; SF-12v1, Short-Form 12 Health Survey, Version 1; Upp, upper limit 95% confidence interval.

* Depressive disorders: major depression episode or dysthymia based on DSM-IV criteria as assessed with the CIDI 3.0.

† Prevalence according to the SF-12 with the selected cutoff point.

‡ AUC for a dichotomous predictor (1 = below or equal to the cutoff point; 0 = above the cutoff point).
Table 3 – Accuracy of the SF-12v1 for depressive disorders* (30-d and 12-mo) at selected cutoff points according to scoring method and country, in the ESEMeD sample.

<table>
<thead>
<tr>
<th>Scoring method and cut point</th>
<th>Country</th>
<th>Prevalence SF-12</th>
<th>Prevalence depressive disorders</th>
<th>Sensitivity (SN)</th>
<th>Specificity (SP)</th>
<th>LR+ Low Up</th>
<th>LR− Low Up</th>
<th>Design-adjusted odds ratio</th>
<th>Design-adjusted AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCS-12 ≤ 48.9</td>
<td>Belgium</td>
<td>17.1</td>
<td>5.3</td>
<td>0.6</td>
<td>0.68</td>
<td>0.05</td>
<td>0.86</td>
<td>0.01</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>23.6</td>
<td>6.5</td>
<td>0.6</td>
<td>0.70</td>
<td>0.04</td>
<td>0.80</td>
<td>0.01</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>15.7</td>
<td>3.3</td>
<td>0.4</td>
<td>0.78</td>
<td>0.04</td>
<td>0.86</td>
<td>0.01</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>Italy</td>
<td>20.0</td>
<td>3.4</td>
<td>0.3</td>
<td>0.77</td>
<td>0.04</td>
<td>0.82</td>
<td>0.01</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>The Netherlands</td>
<td>15.5</td>
<td>5.1</td>
<td>0.6</td>
<td>0.67</td>
<td>0.05</td>
<td>0.87</td>
<td>0.01</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>Spain</td>
<td>23.3</td>
<td>4.4</td>
<td>0.3</td>
<td>0.74</td>
<td>0.03</td>
<td>0.79</td>
<td>0.01</td>
<td>3.5</td>
</tr>
<tr>
<td>RAND-12</td>
<td>Belgium</td>
<td>19.7</td>
<td>5.3</td>
<td>0.6</td>
<td>0.71</td>
<td>0.05</td>
<td>0.83</td>
<td>0.01</td>
<td>4.2</td>
</tr>
<tr>
<td>MHC ≤ 47.9</td>
<td>France</td>
<td>26.2</td>
<td>6.5</td>
<td>0.6</td>
<td>0.72</td>
<td>0.04</td>
<td>0.77</td>
<td>0.01</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>19.5</td>
<td>3.3</td>
<td>0.4</td>
<td>0.82</td>
<td>0.04</td>
<td>0.83</td>
<td>0.01</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>Italy</td>
<td>24.8</td>
<td>3.4</td>
<td>0.3</td>
<td>0.79</td>
<td>0.04</td>
<td>0.77</td>
<td>0.01</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>The Netherlands</td>
<td>21.8</td>
<td>5.1</td>
<td>0.6</td>
<td>0.76</td>
<td>0.05</td>
<td>0.81</td>
<td>0.01</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>Spain</td>
<td>25.3</td>
<td>4.4</td>
<td>0.3</td>
<td>0.78</td>
<td>0.03</td>
<td>0.77</td>
<td>0.01</td>
<td>3.4</td>
</tr>
<tr>
<td>BRP-12 MHS ≤ 41.9</td>
<td>Belgium</td>
<td>16.4</td>
<td>5.3</td>
<td>0.6</td>
<td>0.66</td>
<td>0.05</td>
<td>0.86</td>
<td>0.01</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>23.3</td>
<td>6.5</td>
<td>0.6</td>
<td>0.70</td>
<td>0.04</td>
<td>0.80</td>
<td>0.01</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>14.8</td>
<td>3.3</td>
<td>0.4</td>
<td>0.80</td>
<td>0.04</td>
<td>0.88</td>
<td>0.01</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>Italy</td>
<td>18.9</td>
<td>3.4</td>
<td>0.3</td>
<td>0.77</td>
<td>0.04</td>
<td>0.83</td>
<td>0.01</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>The Netherlands</td>
<td>17.5</td>
<td>5.1</td>
<td>0.6</td>
<td>0.70</td>
<td>0.05</td>
<td>0.85</td>
<td>0.01</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>Spain</td>
<td>17.1</td>
<td>4.4</td>
<td>0.3</td>
<td>0.71</td>
<td>0.04</td>
<td>0.85</td>
<td>0.01</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Note. Estimates and standard errors (SEs) or 95% confidence interval.
AUC, area under the curve; BRP-12 MHS, Bidimensional Response Process Model Mental Health Score; CIDI 3.0, Composite International Diagnostic Interview, Version 3.0; DSM-IV, Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition; ESEMeD, European Study of the Epidemiology of Mental Disorders; Low, lower limit 95% confidence interval; LR+, positive likelihood ratio; LR−, negative likelihood ratio; MCS-12, Original Mental Component Summary; RAND-12 MHC, RAND-12 Mental Health Composite; SE, standard error; SF-12v1, Short-Form 12 Health Survey, Version 1; Upp, upper limit 95% confidence interval.

*Depressive disorders: major depression episode or dysthymia based on DSM-IV criteria as assessed with the CIDI 3.0.

†Prevalence according to the SF-12 with the selected cutoff point.

‡AUC for a dichotomous predictor (1 = below or equal to the cutoff point; 0 = above the cutoff point).
method, however, has not been included in this article because its rationale was close to that of the RAND-12 MHC, with a spearman correlation of 0.97 between them and both scoring methods assuming that the mental and physical dimensions are correlated.

This study has some limitations that must be considered when interpreting the results. First, the SF-12 was interviewer-administered, while most typical form of the questionnaire is self-administration. Several studies found that respondents tend to report better health status in interviewer-administered surveys than in self-administered interviews [30–32]. Second, measurement invariance across countries has not been evaluated. However, average scores by country are consistent with previous results for which no measurement bias regarding diagnostics was found [33]. In addition, model fit for the BRP-12 is adequate, which provides partial support to believe that the SF-12 does not show serious noninvariance. Even though noninvariance in all possible subgroups cannot be entirely ruled out by an isolated fit index, it would be more likely if fit indexes were so poor as to suggest that a different kind of model, affecting a subsample of substantial size, could be underlying the full sample.

Despite these limitations, we are confident that the discriminant capacity for 30-day depressive disorders observed in our study is high, and it is similar to that of a study carried out in Australia [4], which reported an AUC of 0.92 (95% CI 0.90–0.93) for both the MCS-12 and the RAND-12 MHC to assess depression. In addition, in that same study, a score of 45 or less was recommended as the best screening cutoff for depression, with an SN of 0.87 and an SP over 0.80. Gill et al.’s [4] results were based on CIDI responses using ICD-10 criteria with a 4-week recall period. This ensured the effective diagnosis of an active depressive episode at the moment of the interview, which makes detection more likely. It is important to note that when we use 12-month DSM-IV criteria, the AUCs obtained for depressive disorders are still high, showing that the mental component of the questionnaire is sensitive to the presence of a recent disorder. Therefore, the SF-12 might be adequate to discriminate among individuals with disorders even though the pathology is not present anymore at the time of the interview, or it is present with subthreshold symptomatology.

Our results are in the higher range of previous studies conducted on the MCS of the SF-36. In a study carried out on a representative sample of the US general population using the SF-36 [34], an AUC of 0.77 was reported on the MCS of the SF-36 for the assessment of depressive disorders (major depression and/or dysthymia), with an SN of 0.74 and an SP of 0.81 using the cutoff point of 42. In other studies, conducted on groups of patients with the SF-36 [35–37], the AUC of detecting any mental disorder was in the 0.82 to 0.91 range, with SN values between 0.73 and 1.0 and specificities between 0.64 and 0.90. It is important to note, however, that in one of these studies [35] the presence of disorder was assessed with the Centre for Epidemiological Studies Depression (CES-D) Scale, a self-reported measure with 20 items related to depressive symptomatology that does not follow standard diagnostic criteria such as the ICD or the DSM like the CIDI.

Results observed in our study are within the range of other specific psychopathology scales based on symptoms that are commonly used in general population epidemiologic surveys such as the 12-item version of the General Health Questionnaire (GHQ-12) or the CES-D Scale. In a review of validation studies using the 12-item version of the GHQ-12 to evaluate its screening ability for common mental disorders, Goldberg et al. [38] reported SN values ranging from 0.67 to 0.93 (median = 0.84) and specificities ranging from 0.59 to 0.91 (median = 0.79). In the same study, results were presented from primary care samples obtained in 15 centers around the world, showing AUCs ranging from 0.83 to 0.95. Similarly, screening performance of the CES-D against depressive disorders has been found to be moderate to high in several studies, with AUC values ranging between 0.74 and 0.94 [39–47].

In our study, the discriminant capacity of the SF-12 to detect individuals with depressive disorders is significantly greater than its ability to detect anxiety disorders (with AUC values of 0.7 for both 12-month and 30-day recall periods). This is consistent with data reported by Gill et al. [4] and may be explained by the fact that the SF-12 contains two items related to symptoms in the affective disorder spectrum, while only one of the symptoms relate to anxiety disorders. Further work would be required to assess to what extent the discriminant capacity of the SF-12 is due to these two items. The discriminant capacity of the questionnaire for anxiety disorders, however, increases substantially when we do not take into account the disorders of social phobia and specific phobia in the count of anxiety disorders, achieving AUC values of 0.77 for 12-month disorders and 0.83 for 30-day disorders. We therefore do not recommend SF-12 as a screener of anxiety disorders.

The optimal cutoff point for the mental component of the SF-12 to screen for 30-day depressive disorders in Europe is 45.6 for the MCS-12 score. This cutoff point is substantially higher than the score of 42 or less suggested for the mental component of the SF-36 in the US general population to detect depressive disorders according to the Diagnostic Interview Schedule [34]. However, both values are about 8 points below the respective general population mean (i.e., 53.7 in Europe and 50 in the United States), and the accuracy achieved in the EU sample with the suggested cutoff point (SN = 0.86, SP = 0.88) is better than that achieved in the US sample (SN = 0.74, SP = 0.81). These performance results imply that when administered to the general European population, the SF-12 will miss only 14% of the true cases (SN = 0.86), while it will identify as depressed 12% of the individuals who would not comply with diagnostic criteria (SP = 0.88).

The ESEMeD sample was selected to be representative of the adult general population of each country. In addition, poststratification weights were used to restore the slight differences encountered in the distributions of age, gender, and region within each country as compared with those of the general population [13]. Thus, we are reasonably convinced that the cutoff point suggested here is adequate for its use in the six countries studied. Further research, however, is needed to determine whether this cutoff point could be adequately used in other countries.

The cutoff point suggested for the RAND-12 MHC to assess 30-day depressive disorders is 44.5, while for the BRP-12 MHS it is 40.2. Similar to what has been described for the MCS-12, these cutoff points are 8.4 and 9.8 points below the corresponding general population mean, respectively. Such consistency would support the use of the cutoff points described in our study. Nevertheless, it is important to stress that these cutoff points are not directly comparable, because the MCS-12 and RAND-12 MHC are norm-based scores with a mean of 50 in the US general population, and mean scores in our sample are higher (mean MCS-12 = 52.9, mean RAND-12 MHC = 52.9), while BRP-12 MHS are T-scored to have a mean of 50 in our sample. To be able to compare them, specific methods of scale linking based on equipercentile curves or linear linking should be applied, which goes beyond the scope of this article. Moreover, comparisons of the three scoring methods using equipercentile curves have already been applied previously on the same sample (Forero et al., under review).

Conclusions

Results of this large study show that in spite of not having been developed as a screening questionnaire, the mental component of the SF-12 performs adequately as a measure of active and recent depressive disorders in the adult general population. The SF-12 is sensitive enough to active and recent depressive disorders, and its accuracy is at a similar level as that of other screening measures specifically developed for the detection of
depressive disorders, such as the CES-D scale. Thus, the SF-12 can be confidently used as a screening measure for these disorders for prevalence monitoring and targeting treatment from general population surveys, such as health interview surveys, in which this instrument is usually included.

For depressive disorder detection purposes, and whenever US norms are used to obtain the MCS-12 in Europe, the cutoff point of 45.6 is recommended over the US general population SF-36 cutoff point of 42. The European cutoff provides specific results due to better fit to population characteristics.

Acknowledgments

ESEMeD is carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey. We thank the WMH staff for assistance with instrumentation, fieldwork, and data analysis. A complete list of WMH funding support and publications can be found at http://www.hcp.med.harvard.edu/wmh.

Source of financial support: Gemma Vilagut was supported by “Fondo De Investigación Sanitaria. ISCIII” (ECA07/059). Carlos García Forero was supported by “ Ministerio de Ciencia e Innovación” FSE grant (JCI-2009-05486). This work was supported by a grant from Instituto de Salud Carlos III FEDER (PI08/90724).

The ESEMeD project (http://www.epremed.org) was funded by the European Commission (Contracts QLG5-1999-01042; SANCO 2004123; EHAC 20081308), the Piedmont Region (Italy), Fondo de Investigación Sanitaria, Instituto de Salud Carlos III, Spain (FIS 00/0028-02), Ministerio de Ciencia y Tecnología, Spain (SAF 2000-158-CE), Departament de Salut, Generalitat de Catalunya, Spain, and other local agencies and by an unrestricted educational grant from GlaxoSmithKline.

REFERENCES


