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, MSc[1,2,3,]*, Carlos G. Forero, PhD[2,1], Alejandra Pinto-Meza, PhD[4,5], Josep M. Haro, MD, PhD[4,5], Ron de Graaf, PhD[6], Ronny Bruffaerts, PhD[7], Viviane Kovess, MD, PhD[8], Giovanni de Girolamo, MD[9], Herbert Matschinger, PhD[10], Montserrat Ferrer, MD, PhD[1,2,11], Jordi Alonso, MD, PhD[1,2,3], on behalf of the ESEMeD Investigators

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

A B S T R A C T

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 Bidemensional 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.

Introduction

The Short-Form 12 Health Survey, Version 1 (SF-12v1), is a generic health-related quality-of-life instrument. It was originally developed in 1994 as a shorter alternative (12 items) to the widely used Short-Form 36 Health Survey (SF-36), for studies in which a 36-item form was too long. The mental health dimension of the SF-36 is composed of the 5 items of the Mental Health Inventory, which has shown good performance in tests of sensitivity (SN) and specificity (SP) relative to other screening tools for depression and other mental disorders [1]. Three of the five items of the Mental Health Inventory are still included in the SF-12 and refer to symptoms related to the diagnostic criteria for common depressive and

anxiety disorders. Furthermore, the questionnaire includes other items regarding functional impairment due to mental problems that are also related to experiences of distress or impairment as a consequence of psychological symptoms, which need to be fulfilled according to both *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) [2] and *International Statistical Classification of Diseases, 10th Revision* (ICD-10) [3] classification systems. As a brief and reliable measure of overall health status, the SF-12 health questionnaire is often included in large population health surveys. Moreover, given its content, the mental component of the SF-12 could serve as a screener of depressive disorders [4] and thus it could be useful for monitoring prevalence [5] and for targeting treatment and prevention [6].

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 Bidemensional Response Process Model algorithm (BRP-12) based on the Item Response Theory (IRT) [9] (Forero et al., 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. Differently 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 BRP-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.

Methods

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.

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.