PUBLIC HEALTH XXX (2016) 1–9

Available online at www.sciencedirect.com

Public Health

journal homepage: www.elsevier.com/puhe

Original Research

Validity of the 12-item version of the General Health Questionnaire in detecting depression in the general population

A. Lundin [a,b,*], M. Hallgren [b], H. Theobald [c], C. Hellgren [d], M. Torgén [e]

[a] Centre for Epidemiology and Community Medicine, Stockholm County Council, Stockholm, Sweden
[b] Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden
[c] Academic Primary Health Care Centre, Stockholm County Council and Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden
[d] Swedish Council for Higher Education, Stockholm, Sweden
[e] Department of Medical Science, Uppsala University, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form
18 February 2016
Accepted 2 March 2016
Available online xxx

Keywords:
GHQ-12
Sensitivity and specificity
Area under the curve
Self-assessment
Depression
Anxiety

ABSTRACT

Objectives: The 12-item version of the General Health Questionnaire (GHQ-12) is frequently used to measure common mental disorder in public health surveys, but few population-based validations have been made. We validated the GHQ-12 against structured psychiatric interviews of depression using a population-based cohort in Stockholm, Sweden.

Methods: We used a population-based cohort of 484 individuals in Stockholm, Sweden (participation rate 62%). All completed the GHQ-12 and a semi-structured psychiatric interview. Last month DSM-III-R symptoms were used to classify major and minor depression. Three scoring methods for GHQ-12 were assessed, the Standard, Likert and Corrected method. Discriminatory ability was assessed with area under the receiver operating characteristic (ROC) curve.

Results: A total of 9.5% had a major or minor depression. The area under the ROC curve was for the Standard method 0.73 (0.65–0.82), the Likert method 0.80 (0.72–0.87) and the Corrected method 0.80 (0.73–0.87) when using major or minor depression as standard criterion. Adequate sensitivity and specificity for separating those with or without a depressive disorder was reached at ≥12 Likert scored points (80.4 and 69.6%) or ≥6 Corrected GHQ points (78.3 and 73.7%). Sensitivity and specificity was at ≥2 Standard scored points 67.4% and 74.2%.

Conclusion: When scored using the Likert and Corrected methods, the GHQ-12 performed excellently. When scored using the Standard method, performance was acceptable in detecting depressive disorder in the general population. The GHQ-12 appears to be a good proxy for depressive disorder when used in public health surveys.

* *Corresponding author*. Centre for Epidemiology and Community Health, Stockholm County Council, Box 1497, SE 171 29, Solna, Sweden. Tel.: +46 852487604.
E-mail address: andreas.lundin@ki.se (A. Lundin).

http://dx.doi.org/10.1016/j.puhe.2016.03.005
0033-3506/

Introduction

The General Health Questionnaire (GHQ) (60 item version) was developed as a self-administered screening instrument for non-psychotic mental illness in patients.[1] A short version with 12 items (GHQ-12), derived from the non-somatic GHQ-60 is frequently used in epidemiology to monitor population mental health[2–9] and as a screening tool in double-phase sampling.[10–12] The symptoms encompass primarily those for depression syndrome but also anxiety. Each item in the GHQ asks the respondent to rate the degree to which they have experienced a symptom during the last week with four response categories, typically worded: less than usual, no more than usual, rather more than usual, or much more than usual. Typically responses are dichotomized (0-0-1-1) and summed into an index with the range 0–12. The commonly used threshold for distress is three or more using this scoring method,[13] but in population-based studies, which also often use a dichotomized standard-scored GHQ, the thresholds range between 0/1 (in Australia)[6] and 3/4 (in Finland).[5]

While several studies have examined the ability of the GHQ to correctly identify disorders in patient settings,[14,15] few have assessed the questionnaire's performance in population-based samples.[16–19] We know of only two previous studies that attempted to validate the 12-item version of GHQ in a general population of adults.[6,20] Using the 1997 Australian Survey of Health and Well-being, Donath (2001) examined the agreement between the 12-item version of the GHQ with any past month mood or anxiety disorder identified by structured psychiatric interview (the Composite International Diagnostic Interview schedule, CIDI).[6] The author found good overall agreement between the two schedules, but also that a cut-off score of one resulted in a better discriminatory ability than higher cut-offs. Moreover, the Standard scoring method (0, 0, 1, 1) did not perform well compared to the alternative methods sometimes used; specifically, a Likert score (where each item is scored 0, 1, 2, 3)[18,21] intended to measure intensity, and a Corrected scoring method,[19] (where positively phrased items are coded 0-0-1-1 and negatively phrased items 0-1-1-1). The Corrected scoring method was explicitly constructed to also include individuals with chronic mental illness, for whom responses such as 'no more than usual' to the negatively phrased items may be a reasonable answer.[18]

However, the GHQ-12 mean of 0.83 for men and 0.93 for women (for the standard scoring method) in the 1997 Australian Survey of Health and Well-being was low compared to other population means in, for example, England,[22,23] Sweden,[24] Finland[20] and the Netherlands.[2] A healthier population may affect both the optimal cut point and the overall agreement between the GHQ-12 and the interview-based diagnosis. In a Finnish population-based health survey from 2000 which examined the agreement between the Standard GHQ-12 scoring method (male and female means: 1.7 [1.6–1.8] and 1.9 [1.8–2.0]) and CIDI major depression and dysthymia, there was higher overall agreement (AUC = 0.88 for the last two weeks) than in Australia, and also optimal sensitivity and specificity at the 2/3 cut off point.[20] No assessment was made of the Likert and Corrected scoring methods which in the Australian study outperformed the Standard method.

If the GHQ-12 is to be used in general population settings, then its' sensitivity and specificity should be validated in general population studies, rather than using clinical samples. In the current study we investigate the performance of the 12-item interview version of the GHQ in a Swedish adult population sample, and compare the performance of three scoring methods, the Standard, the Likert, and the Corrected method. Because depression includes a spectrum and longitudinal epidemiological studies have shown that those with minor or subclinical depression are more likely to develop major depression,[25] we use both major depression and a wider category including minor depression as outcome. Sensitivity and specificity of the GHQ-12 for distinguishing those with minor or major depression from those without disorder was also examined.

Methods

Population

This study uses data from the 1993 wave of the REBUS study (Swedish acronym for Rehabilitation Needs Survey); a stratified double-phase random sample of originally 18–65 year olds from Stockholm County who were examined in 1970 and followed up with reexaminations in 1980, 1993 and 1996. The screening and sampling procedure has been described previously.[26] Briefly, replicated sampling was used to randomize both in the first phase (stratified periodic systematic sampling), and in the second phase (stratified random sampling). The 1970 examination included a psychiatric interview, a social worker interview, and a thorough medical examination (n = 2578, participation rate = 88.7%). In 1993, as part of a study of risk factors for musculoskeletal illness, participants of working age (41–59 years) who were free from musculoskeletal disorder and severe somatic or psychiatric disorder at baseline were invited to a new examination. A flow chart of the inclusion is provided in Fig. 1. From 1269 possible subjects, 486 were excluded (104 dead or lost in the registers, 119 with musculoskeletal diagnosis, 101 with other serious medical condition [including 17 with mental retardation, two with alcohol use disorders, one with psychosis and one with schizophrenia]). Additionally 136 individuals were excluded because of incomplete information in the 1970 examination and 26 due to incomplete address, leaving 783 eligible participants. Of these, 484 (252 women and 232 men) participated in the examination (62%). The examination took place at the National Institute for Working Life between March 1993 and September 1994 and included six questionnaires on health, working conditions, and psychosocial factors and a semi-structured interview by a psychologist. Mean age in at examination was 48.8 (*SD* 4.5, range 41–59) years. Male gender, younger age and unemployment were significantly associated with non-participation. Telephone interviews with non-participants indicated that lack of time was the main reason for non-participation.[27]

Please cite this article in press as: Lundin A, et al., Validity of the 12-item version of the General Health Questionnaire in detecting depression in the general population, Public Health (2016), http://dx.doi.org/10.1016/j.puhe.2016.03.005