Journal of Affective Disorders 225 (2018) 503–522

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Review article

The psychometric properties of depression screening tools in primary healthcare settings: A systematic review

Sarira El-Den*, Timothy F. Chen, Yuh-Lin Gan, Eling Wong, Claire L. O'Reilly

Faculty of Pharmacy, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Psychometric
Depression
Screening
Primary healthcare

ABSTRACT

Background: Consensus on a tool for depression screening among adults in primary healthcare (PHC) settings is lacking. This systematic review aimed to explore the psychometric properties of depression screening tools.

Methods: A systematic literature search composed of four terms (screening AND psychometric AND depression AND primary healthcare) was conducted in PubMed, EMBASE, PsycINFO and MEDLINE, between January 1995 through October 2015. Studies that aimed to psychometrically test a depression screening tool among the general adult population in a PHC setting were included. Studies exploring the diagnostic properties of depression screening tools among specific populations were excluded.

Results: Sixty publications, evaluating the psychometric properties of 55 tools or adaptations, were included. Studies were conducted in 24 countries and 18 languages on 48234 adults. The Patient Health Questionnaire-9 was the most evaluated tool with 14 studies evaluating its psychometric properties. Fifty-four studies reported on at least one measure of receiver operating characteristics. Sensitivity and specificity values ranged from 28% to 100% and 43% to 100%, respectively. Cronbach alpha values ranged from 0.56 to 0.94. Other forms of reliability and validity testing were less consistently and commonly reported.

Limitations: The inclusion of studies regardless of methodological quality or design may have limited generalizability, but allowed for a comprehensive and detailed overview of the current literature.

Conclusions: Depression screening tools vary in their psychometric properties. The PHQ-9 was the most extensively psychometrically tested tool. This systematic review may aid PHC professionals in choosing a depression screening tool for universal use as it provides a comprehensive overview of their psychometric properties.

1. Introduction

The World Health Organization reports that depression is "the leading cause of disability worldwide", with over 350 million people affected (World Health Organisation, 2017a). Various guidelines recommend universal depression screening among adults. For example, the United States Preventative Services Task Force "recommends screening for depression in the general adult population" (U.S. Preventive Services Task Force et al., 2016). Depression can lead to suicide which is responsible for over 800,000 deaths yearly, worldwide (World Health Organisation, 2017b). A large proportion of deaths due to suicide occur among people living with affective disorders. For example, over 40% of people who died by suicide between 2011 and 2013 in the United Kingdom had an affective disorder and this percentage has remained relatively consistent over time (The National Confidential Inquiry into Suicide and Homicide by People with Mental Illness, October, 2016). Furthermore, 28% of those who died by suicide had attended or contacted mental health services in the previous 12 month period (The National Confidential Inquiry into Suicide and Homicide by People with Mental Illness, October, 2016). The diagnosis, management and treatment of depression often occurs in primary care. For example, in Australia, 86% of psychotropic medications are prescribed by general practitioners (Australian Bureau of Statistics, 2007). Hence, primary healthcare professionals, such as general practitioners and pharmacists, can play an essential role in the detection, management and treatment of depression in primary care.

Abbreviations: 4DSQ, Four Dimensional Symptom Questionnaire; AUC, Area under the curve; BDI-II, Beck Depression Inventory-II; CBDI, Chinese Beck Depression Inventory; CES-D, Centre for Epidemiological Studies-Depression Scale; CFA, Confirmatory factor analysis; COSMIN, Consensus-based standards for the selection of health measurement instruments; DIF, Differential item functioning; FA, Factor analyses; GHQ-12, General Health Questionnaire-12; GP, General practitioner; K6, Kessler Scale 6; LR, Likelihood ratio; PCA, Principal components analysis; PHQ, Patient health questionnaire; PHQ*, Personal Health Questionnaire; SDDS-PC, Symptom Driven Diagnostic System for Primary Care; WHO-5, World Health Organization–Five Well-Being Index

* Correspondence to: Faculty of Pharmacy, The University of Sydney, Building A15, Camperdown, NSW 2006, Australia.
E-mail address: sarira.el-den@sydney.edu.au (S. El-Den).

http://dx.doi.org/10.1016/j.jad.2017.08.060
Received 30 June 2017; Received in revised form 4 August 2017; Accepted 20 August 2017
Available online 23 August 2017
0165-0327/

Although depression screening is often recommended in guidelines, screening does not confirm a diagnosis (Johns Hopkins Medicine, 2014). Screening is usually conducted using a screening test which is administered to people who are asymptomatic of the disease in question (Johns Hopkins Medicine, 2014). Depending on the results of the test, a person is determined to be at a high or low risk of having the condition and is appropriately referred to "additional testing to determine the presence or absence of disease" (Johns Hopkins Medicine, 2014). An ideal screening test would be able to differentiate those with and without the disease, without any errors (U.S. National Library of Medicine, 2016). However, error-free screening tests are quite rare. Hence, the efficacy of depression screening, in terms of its ability to accurately determine those who are at increased risk of having depression is largely dependent on the screening tool used. Furthermore, among depression screening studies, it is often difficult to determine if a study aims to evaluate screening or diagnosis as the tools and terms applicable to each of these processes are sometimes used interchangeably. Moreover, screening tools are sometimes recommended to be used as measures of depression severity, over time, after a diagnosis has been made, thereby adding to the confusion in terminology (beyondblue, 2016).

It is essential that all measures of health, such as depression screening tools, are reliable and valid to ensure that the results they generate are clinically correct (Mokkink et al., 2010). It is especially important for measurement instruments to yield valid and reliable results in the context of "health-related patient-reported outcomes" as they often involve the measurement of constructs which cannot be measured directly (Mokkink et al., 2010). A broad variety of screening tools for depression are readily available to primary care clinicians. However, it is important to realise that depression screening tools are not all equal in their psychometric properties. Since the effectiveness of depression screening is primarily dependent on the screening tool being used, it is essential to explore the psychometric properties of the screening tools to determine if they are reliable and valid.

There is a wide range of depression screening tools available which vary in length, style, presentation, administration and psychometric properties. Screening tools also vary in their extent of psychometric evaluation. There are numerous studies assessing the reliability and validity of these screening tools; however, there is currently no consensus on one particular screening tool to be used for depression screening across primary healthcare settings. Primary healthcare practitioners may have difficulties in selecting an appropriate screening tool due to the multitude of screening tools that are readily available. For comparisons between various screening studies to be accurate, it would be advantageous for healthcare professionals to be encouraged to use one screening tool that demonstrates sound psychometric properties in primary healthcare settings to ensure the homogeneity of results.

In light of the current evidence, the objectives of this systematic review were:

1. To systematically review the literature surrounding the psychometric properties of depression screening tools in primary healthcare settings.
2. To determine which depression screening tools have sound psychometric properties and should be recommended for use in the general population in primary healthcare settings.

2. Methods

Records were retrieved by searching MEDLINE, PubMed, EMBASE and PsycINFO and through automatic alerts, from 1995 through to October 2015. Limitations were set to only identify records published in English and studies conducted on a human sample. The final literature search strategy was based on combined searches of four concepts and their related terms, using both keywords and mapped subject headings, when possible, depending on the database:

1. Screening or screening tool* or screening test* or screening instrument* or screening scale* AND
2. Validity or reliability or sensitivity or specificity or inter-rater agreement or positive predictive value or negative predictive value or internal consistency or psychometric* AND
3. Depression or unipolar depression or depressive disorder or major depression or major depressive disorder AND
4. Primary healthcare or primary healthcare or primary health or general practice or family practice

A structured inclusion and exclusion checklist was created to ensure the consistency of all studies generated from the literature search. For inclusion in the review, the study had to meet all of the following criteria:

1. The study reported on a tool used to screen for depression.
2. The aim of the study was to evaluate the psychometric properties of a tool, in relation to its ability to screen for depression.
   - The aim specifically contained at least one of the following terms: testing, usefulness, effectiveness, utility, applicability, ability, comparison, reliability, validity, sensitivity, specificity, accuracy, (test/operating) characteristics, or identification.
3. Participants were recruited from a primary healthcare setting (including non-specific, general medicine hospital outpatient settings).
4. All participants included in the analysis were over the age of 18.
5. The publication reported on an original primary research study.

Studies were excluded if:

1. The study reported on the psychometric properties of a tool used to diagnose depression.
2. The study reported on a screening tool(s) for multiple mental illnesses and the psychometric properties of the depression screening subscale or tool were not reported separately.
3. The study only reported on the face or content validity of the depression screening tool.
4. The study aimed to assess the psychometric properties of a depression screening tool in a specific patient population. E.g. oncology patients, geriatrics or specific age and/or gender.
5. Participants had a current psychiatric diagnosis or were taking psychotropic medications, at the time of depression screening.

The purpose of this review was to explore the psychometric properties of depression screening tools to help inform the selection of a screening tool that could be used in the general adult population across primary healthcare settings. For this reason, data collected from specific populations, such as, females only was necessarily excluded as it did not present information that would be relevant to the general adult population which includes females and males.

The Faculty Liaison Librarian, at the University of Sydney Medical Sciences Libraries, was consulted by one of the authors (SE) to ensure the appropriate search terms and databases were used.

2.1. Data extraction

One author (SE) conducted the searches in all four databases. All retrieved records were exported into Endnote (Endnote ×7.0.2, USA) and, both, automatic and manual means were used to identify and remove duplicates. After screening titles and abstracts, for inclusion in the review, the remaining articles were then assessed using the inclusion checklist. All remaining full-text articles were reviewed by either SE or YG using the structured checklist to record eligibility, and were either included, excluded or deferred due to uncertainty. All deferred articles were reviewed and a decision was made by consensus of SE, YG and CO.