IS THERE GOOD EVIDENCE THAT THE TWO QUESTIONS IN PHQ-2 ARE USEFUL QUESTIONS TO USE IN ORDER TO SCREEN FOR DEPRESSION?

Rachel Wilson¹ & Mark Agius¹,²
¹Clare College Cambridge, Cambridge, UK
²Department of Psychiatry, University of Cambridge, Cambridge, UK

SUMMARY
The PHQ-2 has been presented as an ultra-short screening questionnaire to detect depression in several healthcare settings. This paper will outline its current uses, evaluate its accuracy as a screening method, and suggest potential uses.

Key words: depression - PHQ-2 - screening

* * * * *

INTRODUCTION
Depression is a mental illness characterized by depressed mood and anhedonia (Arroll 2010). Multiple questionnaires and criteria have been proposed which may be used to enable a diagnosis of depression to be made. Such Diagnostic questionnaires include the Composite International Diagnostic Interview (CIDI) (Arroll 2010), the Structured Clinical Interview for DSM-IV (SCID) (Bhana 2015), the Hospital Anxiety and Depression Scale (HAD) (Al-Qadihi 2014), the Hamilton Depression Scale, the Montgomery and Asberg Depression Rating Scale (MADRS), the Beck Depression Inventory (Gjerdingen 2009), as well as the Patient Health Questionnaire (PHQ).

Family physicians often do not identify depressive symptoms; it is reported that over half of the cases of MDD are missed in primary care settings (Arroll 2010, Al-Qadihi 2014, Gjerdingen 2009, Choi 2015, Zhang 2013). This may be because physicians concentrate on primary conditions rather than underlying mental health issues (Margrove 2011). Alternatively, this might be because of a lack of a systematic screen. For example, only 5.9% of staff in cancer, palliative care, or related disciplines in the UK reported using a formal questionnaire to screen for depression, since there was no national guidance on which depression screens to use (Mitchell 2008).

WHAT IS THE PHQ-2?
The PHQ-2 comprises the first two questions of the PHQ-9. The PHQ-9 is a questionnaire made up of 9 questions derived from the DSM-IV criteria for major depressive disorder (MDD) (Arroll 2010). The answers are scored on a Likert scale. A cumulative score of above ten or eleven usually results in a diagnosis of depression, but there are also algorithmic methods of working out PHQ-9 scores. The PHQ-9 has become the preferred diagnostic questionnaire in non-psychiatric settings because it is easy to administer and score (Zuithoff 2010).

The first question of the PHQ-2 asks how often the patient has felt down, depressed, and hopeless over the past two weeks, with answers from 0 (not at all) to 3 (nearly every day). The second question asks how often the patient has lost pleasure or interest in doing things over the same time scale, with the same answers (Arroll 2010). The cumulative score is added up, with threshold score for suspecting the possibility of depression usually of 3 or higher (Arroll 2010, Zhang 2013).

HOW IS PHQ-2 CURRENTLY USED?
The PHQ-2 is not a way of diagnosing depression. This is emphasised in almost all studies that research the accuracy and validity of the questionnaire, however there are some reports in which it is incorrectly used as a definitive diagnosis (Schlosser 2016).

The PHQ-2 can be used to screen patients and find those who might be suffering from depression. These patients should then be given more comprehensive diagnosis questionnaires, such as the PHQ-9 (Arroll 2010, Gjerdingen 2009, Zhang 2013, Corson 2004, Fuchs 2015, Inagaki 2013, Li 2007, Loeb 2015, Thombs 2008).

Unfortunately, the PHQ-2 is not often recommended in healthcare settings around the world. The exception is in the USA. The American Heart Association (AHA) recommends the PHQ-2 followed by PHQ-9 as a systematic screen for depression in cardiovascular patients (Elderon 2011). The US Preventive Services Task Force (USPSTF) recommends routine screening for depression provided that accurate systems are in place (Gjerdingen 2009, Richardson 2010). The American Academy of Paediatrics recommends screening amongst adolescents (Sudhanthat 2015). Despite recommendations for systematic screening, in the USA,
a positive PHQ-2 only currently leads to a PHQ-9 in 5% of cases, instead, physicians tend to use prior knowledge, judgement, and patient history to diagnose depression (Fuchs 2015).

It has been suggested that formal implementation of systematic screening, such as PHQ-2 followed by PHQ-9, could help to decrease the number of missed cases of depression (Loeb 2015). Systematic screening improve rates of referral for further psychiatric care (Sudhanhat 2015). Student-run free clinics (SRFCs) in the USA have successfully implemented the PHQ 2 screen and have identified previously undiagnosed depression (Soltani 2015).

DOES PHQ-2 EFFECTIVELY SCREEN FOR DEPRESSION?

PHQ-2 correlates well with PHQ-9 and with other measures of depression (Al-Qadhi 2014, Yu 2011). Adolescents with scores of above three on PHQ-2 also correlated with functional impairment by the Columbia Impairment Scale, and parent-reported psychosocial impairment (Richardson 2011). These data suggest that PHQ-2 can predict depression.

The accuracy of screening may be measured using sensitivity, specificity, and positive predictive values (PPV). Sensitivity is the probability of testing positive if disease is truly present. Specificity is the probability is the probability of testing negative if disease is not present. PPV is the proportion of positive screens that are true positives.

Sensitivity, specificity, and PPVs all compare one screen to other diagnostic tools, but as there are not definitive diagnostic criteria for depression, it is difficult to calculate and compare PHQ-2 results. In each of the studies in question, PHQ-2 results are compared with different ‘gold-standard’ diagnostic questionnaires, so that data are not comparable between studies. Despite the lack of ability to directly compare accuracy, most studies conclude that the PHQ-2 is both reliable and valid as a screening tool (Arroll 2010, Zhang 2013, Kroenke 2003). It is observed that PHQ-2 scores do remain stable in patients over a long period of time, suggesting that it is a reliable measure (Margrove 2011).

One problem to consider is that specificity may be significantly different depending on age, sex, race, and ethnicity. These differences may be due to different presentation of symptoms. For example, adolescents can present with irritability, rather than low mood (Richardson 2010). However, such differences are not highly clinically significant and may be ignored for the purpose of screening (Li 2007). As a consequence, studies from different populations all over the world do show similar results (Figure 1).

Sensitivity and specificity depend on the threshold score used in evaluating the PHQ-2. When a threshold of three is used, the PHQ-2 is highly specific, with lower sensitivity. Lowering the threshold to two increases sensitivity at the cost of specificity (Arroll 2010, Bhana 2015, Choi 2015, Zuithoff 2010, Inagaki 2013, Thombs 2008, Liu 2016, Suzuki 2016, Thapar 2014). Therefore, selecting the correct threshold for screening, depends on whether the sensitivity or the specificity is considered more valuable.

The main argument for a threshold of two is that sensitivity should be high for PHQ so that patients who could be a danger to themselves are not missed (Inagaki 2013). Specificity does not necessarily have to be high for a screen, since PHQ-9 can rule out false positives. False positives may indicate patients at risk of depression, so lower specificity could actually be quite useful (Richardson 2010).

However, a threshold of three is most commonly used. This is because specificity is considered a more valuable property of screens (Mitchell 2012). A high specificity ensures faster and more accurate screening in busy healthcare settings, thus minimising the number of false positive patients who will have to go through further tests (Bhana 2015, Yu 2011).

Interestingly, it has been found that, when looking at single questions from PHQ-9 as a screen for depression in cancer patients in the UK, the two used in PHQ-2 are neither the most sensitive nor the most specific when compared with DSM-IV depression criteria (Mitchell 2012). They do have high specificity amongst the others, but the PHQ-2 questions have the lowest sensitivity of all the questions except the question that asks about suicidal thoughts. The question that asks about having little energy has been found to be the most sensitive, but with the least specificity. The question about having trouble concentrating on things such as reading is the question with the most specificity (Mitchell 2012) and has been suggested to be used as a supplementary question for the PHQ-2 to increase its PPV (Thapar 2014).

There is an ongoing meta-analysis looking at the diagnostic accuracy of PHQ-2, but the results have not yet been published (Thombs 2014) at time of writing. Generally, PHQ-2 is considered an effective screening tool, but the results of the meta-analysis will give a more definitive answer.
HOW COULD PHQ-2 BE USED IN THE FUTURE?

Screening for depression with PHQ-2 has low PPV in primary care, but high PPV for high-risk samples (Mitchell 2012). This may be because primary care patients are likely to have depressive symptoms because of their health status (Al-Qadhi 2014). For example, health conditions such as the flu may cause anhedonia because of fatigue. This could also work in the opposite way – people may attribute anhedonia to their health status and then answer negatively in a depression screen because of this attribution (Margrove 2011, Inagaki 2013).

APHQ-2 screen could have useful applications in at-risk groups such as adolescents (Zhang 2013, Richardson 2010, Haugen 2016), women who have recently given birth (Gjerdingen 2009), as well as patients with chronic illnesses (Choi 2015, Margrove 2011, Schlosser 2016, Mitchell 2012), rather than the general population in primary care settings.

The PHQ-2 takes less than 2 minutes (Mitchell 2012), so that it can be quickly administered in busy settings (Li 2007, Liu 2016, Thapar 2014), such as in routine check-ups, or in student clinics. The short length makes electronic administration possible, which is helpful for screening for post-partum depression (Gjerdingen 2009), cardiovascular patients (Wang 2015, Elderon 2011), patients who have suffered trauma (Warren 2016), amongst students, or amongst patients with chronic illness (Smits 2015) including haemodialysis (Cusimano 2015).

It is also necessary to take ethics into consideration. In one study, only 80% of new mothers were comfortable with the idea of screening for postpartum depression (Gjerdingen 2009). It may be considered to be unethical to introduce the PHQ-2 into a routine check-up without first asking for permission to screen for depression. On the other hand, there is the risk that, if the patient is informed, stigma against depression may influence them to answer negatively, especially amongst the elderly (Inagaki 2013).

PHQ-2 is a reliable method of screening, but not diagnosing, depression. It has a wide range of applications, however it must be implemented as part of a systematic screen within healthcare services for it to reach its full potential.

Acknowledgements: None.

Conflict of interest: None to declare.

Contribution of individual authors: Mark Agius devised the questions, supervised and corrected the text. Rachel Wilson drafted the text and conducted the literature search.

References


