PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Sensitivity and specificity of the Patient Health Questionnaire (PHQ-9, PHQ-8, PHQ-2) and General Anxiety Disorder scale (GAD-7, GAD-2) for depression and anxiety diagnosis: A cross-sectional study in a Peruvian hospital population
AUTHORS	Villarreal-Zegarra, David; Barrera-Begazo, Juan; Otazú-Alfaro, Sharlyn; Mayo-Puchoc, Nikol; Bazo-Alvarez, Juan Carlos; Huarcaya-Victoria , Jeff

VERSION 1 – REVIEW

REVIEWER	Benedetti, Andrea McGill University, Departments of Medicine and of Epidemiology, Biostatistics & Occupational Health
REVIEW RETURNED	20-Jun-2023

GENERAL COMMENTS	This paper considers the diagnostic accuracy of several common depresseion and anxiety screening questionnaires in Peru. Improvements are necessary. 1. During what time interval was the gold standard and the screening questionnaires administered? 2. Was no diagnostic interview (e.g. SCID, CIDI) used for the gold standard? 3. Please specify how the ROC curve was used to determinie the optimal cutoff. Oh - I understand - Youden's J was used. Please rephrase, as this statement is very unclear. 4. Comparing the AUC for different cutoff points is not relevant when you have sensitivity and specificity for each cutoff. Please remove this part. 5. The Likelihood ratios are not very informative. Remove. 6. Please provide a flow diagram. It is unclear how you got from 4979 to 1347. Given that you are not comparing instruments, why not include different subjects for different instruments, to increase sample size especially for anxiety? 7. Please read carefully for grammar. 8. How many subjects were truly depressed? Truly anxious? This is crucial information - present in the first paragraph "Participants" in the results.

	9. Just 28 subjects truly had anxiety. It is not reasonable to try to determine the optimal cutoff given so few subjects - small changes in the cutoff will result in big changes in sensitivity. Any differences you see with respect to the standard are likely to be due to chance, and biased. Remove these results. Instead present sensitivity and specificity for all cutoffs. See https://pubmed.ncbi.nlm.nih.gov/33838273/. 10. Youden's J has no clinical significance. Reporting it's value in the text is meaningless. 11. Similarly, any differences in optimal cutoff for the various depression scales are likely due to chance. 12. "However, a meta-analysis found that cutoffs between 8 and 11 showed little difference in sensitivity and specificity [61]." This meta-analysis suffered from selective cutoff reporting. It is unclear what the estimates of sensitivity and specificity mean given that they included different studies at each cutoff. 13. A third reason that cutoffs might differ is chance - especially given small number of actually depressed people. Please discuss. 14. Youden's J is not clinically relelvant. Depending on many factors, one might wish to prioritize sensitivity or specificity, and as such use a different cutoff than that which is optimal via Youden's J (i.e. considering them as equal). Please provide a more nuanced discussion around this point. 15. Please include in the abstract the number of subjects assessed as well as the number truly depressed and truly anxious.

REVIEWER	pitanupong, jarurin Prince of Songkla University, psychiatric
REVIEW RETURNED	27-Jun-2023

GENERAL COMMENTS	Dear Editor, Thank you for the opportunity to review this manuscript. This is an interesting study. However, to make this study more valuable, some points should be revised. 1 Title: Sensitivity and specificity of the Patient Health Questionnaire (PHQ-9, PHQ-8, PHQ-2) and General Anxiety Disorder scale (GAD-7, GAD-2) for depression and anxiety diagnosis: A cross-sectional study in a Peruvian population The title is completed for meaning. 2 Abstract All domain is completed. But Keywords: Depression; Anxiety; Patient Health Questionnaire; Sensitivity and Specificity; Peru should be written in alphabetical order as Anxiety, Depression, Patient Health Questionnaire, Sensitivity and Specificity 3 Introduction or background The content in the introduction section, although some abbreviations are universal, please write the full name of the word and bracket the abbreviation the first time it appears. After that, the author can use abbreviations (such as COVID-19 in line 11, PHQ-9, PHQ-8, PHQ-2, GAD-7, GAD-2 , DASS-21, Kessler-10, HADS-A, HADS, WHO-5). Should check the writing is correct, for example, there is a “ .” in the sentence at line 54.

	What is the reason why countries or low or middle economic countries have a cut-off point value that is different from other high economic countries? If possible, please add research that can be used to explain the reasons for finding new cut off points. In addition, the population of the study was psychiatric patients with physical ailments such as cancer patients, neurology patients and being inpatients. Therefore, it is likely that a review would be required to address the reasons for physical ailments and depression at GAD that required a new cut-off point. 4 Study design (including population, tool, and statistical analysis): The population is patients with medical or surgical conditions that have a co-existing psychiatric illness or need for psychiatric assessment and management. Why not select a population with direct psychiatric patients without comorbid physical illness? So maybe the title needs to be changed? by being consistent with the specific population? 5 Result: Ok 6 Discussion good discussion that include strengths and limitations, however, this study can't be used in physical outpatients with MDD, GAD, please mention. 7 Reference Very more reference, but depend on the rule of journal Finally, please don't give up, but try to rewrite this manuscript to get more valuable research and a better research methodology. Good luck.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Dr. Andrea Benedetti, McGill University
Comments to the Author:
This paper considers the diagnostic accuracy of several common depression and anxiety screening questionnaires in Peru. Improvements are necessary.

1. During what time interval was the gold standard and the screening questionnaires administered?
Reply: Thank for all your comments. We add in the “Instruments and variables” subsection (page 7, paragraph 1):
“The individual clinical psychiatric interview and the psychometric instruments (i.e., PHQ and GAD) were independently applied on the same day, the latter by a mental health nurse or a psychologist and the former by a psychiatrist. The average time between both measurements was 15 minutes (standard deviation = 4.5 minutes), and the order (i.e., psychometric instruments before or after the interview) was randomly assigned.”

2. Was no diagnostic interview (e.g. SCID, CIDI) used for the gold standard?
Reply: We add in the “Strengths and limitations” subsection (page 12, paragraph 2):
“Thirdly, we used an individual psychiatric interview according to the ICD-10 criteria as a gold standard. We were not able to use the Composite International Diagnostic Interview (CIDI) or the Standardised Clinical Assessment (SCID), more typical gold standards, because of the time constraints involved in conducting such interviews. In Peru, health systems are overburdened, and it is not feasible to have lengthy sessions with highly specialised professionals to conduct such structured interviews. However, based on our experience, we believe that a psychiatric interview is a sufficient benchmark in this context.

3. Please specify how the ROC curve was used to determine the optimal cutoff. Oh - I understand - Youden's J was used. Please rephrase, as this statement is very unclear.
Reply: We modified the paragraph in the “statistical analysis” subsection (page 7, paragraph 5):
“The Youden Index is a measure that summarises the performance of a diagnostic test by interpreting it as the probability that the selected cut-off point provides an adequate clinical decision (in terms of sensitivity and specificity), as opposed to the probability that the selected cut-off provides a random decision [55]. The maximum value of the Youden Index was used as a criterion to select the cut-off with the best diagnostic performance for each scale. Values closer to 1 were considered optimal, and those closer to 0 were considered inadequate.”

4. Comparing the AUC for different cutoff points is not relevant when you have sensitivity and specificity for each cutoff. Please remove this part.
Reply: Thank you very much, we have removed the information on AUC.

5. The Likelihood ratios are not very informative. Remove.
Reply: They have been moved from the main manuscript to supplementary material since its reporting is recommended for diagnostic tests (https://www.sciencedirect.com/science/article/pii/S1836955316300583).

6. Please provide a flow diagram. It is unclear how you got from 4979 to 1347. Given that you are not comparing instruments, why not include different subjects for different instruments, to increase sample size especially for anxiety?
Reply: No differences were found in the number of participants who completed the PHQ and GAD questionnaires. That is, those who were excluded did not respond to either instrument. In addition, we feel it is important for the reader to be aware of the reasons for exclusion, so a flow chart is provided in Supplementary Material 4. We modified the participant subsection (page 8, paragraph 5):
“We collected data from 4979 attendances performed within the liaison psychiatry service during the study period. However, some of these attendances were not assessed with PHQ-9 or GAD-7 data (n=3484) or lacked sociodemographic information (n=148) and were eliminated (see supplementary material 4). Thus, our study only included 1347 participants (see Table 1).”

7. Please read carefully for grammar.
Reply: We have thoroughly revised the wording of the manuscript.

8. How many subjects were truly depressed? Truly anxious? This is crucial information - present in the first paragraph "Participants" in the results.
Reply: We add this information to the participant's subsection (page 8, paragraph 5):
“A total of 334 participants (24.8%) were diagnosed with depression, and 28 participants (2.1%) were diagnosed with anxiety, as determined through individual psychiatric interviews conducted based on the ICD-10 criteria.”

9. Just 28 subjects truly had anxiety. It is not reasonable to try to determine the optimal cutoff given so few subjects - small changes in the cutoff will result in big changes in sensitivity. Any differences you see with respect to the standard are likely to be due to chance, and biased. Remove these results. Instead present sensitivity and specificity for all cut-offs. See https://pubmed.ncbi.nlm.nih.gov/33838273/.
Reply: We agree with the reviewer. We modified the results section:
“Because we have a small number of cases with truly anxious people, any changes in the scores of these people could lead to large changes in sensitivity and specificity. Therefore, it is not possible to give an optimal cohort score over the rest, but we present all cohort scores in Supplementary Material 6. In particular, the cut-off point ≥8 had good performance for GAD-7 with sensitivity values of 53.6 (95%CI: 33.9 - 72.5) and specificity of 78.8 (95%CI: 76.5 - 81.0), (see Table 2). The GAD-7’s cut-off

point ≥10 (i.e., the most used) had lower levels of sensitivity (39.3; 95%CI: 21.5 - 59.4), but higher levels of specificity (88.4; 95%CI: 86.5 - 90.1, compared to the cut-off point of ≥8. In addition, the cut-off point for the GAD-2 was ≥2 had a sensitivity of 84.7 (95%CI: 80.4 - 88.4), and a specificity of 50.1 (95%CI: 47.4 - 52.8) (see Table 2)."

We modified discussion section:
"In the case of GAD, the small number of participants with actual anxiety made it impossible to determine an optimal cut-off point. However, we present the sensitivity and specificity of each cut-off point."

We believe that a weakness within the study is the number of participants with gold-standard anxiety, so we have added a limitation (page 12, paragraph 3):
"Fourthly, our study identified a limited number of individuals (n=28) with a diagnosed anxiety condition. Consequently, minor variations in the study cohort could potentially impact the sensitivity or specificity [81]. Nonetheless, we have ensured sufficient statistical power for our analysis based on our sample size calculation. Moreover, all cohort scores on the GAD scale are provided, which can be valuable for future research involving larger numbers of individuals diagnosed with anxiety (refer to Supplementary Material 6)."

Table 2 shows additional cohort cut-off points for the PHQ and GAD scales. In addition, Supplementary Materials 5 and 6 contain all cohort cut-off points for the PHQ and GAD scales, respectively:
● "In supplementary material 5, we provide the values of all cut-off points for the different versions of the PHQ."
● "In supplementary material 6, we present the values of all cut-off points for the different versions of the GAD."

10. Youden's J has no clinical significance. Reporting it's value in the text is meaningless.
Reply: Thank you for the recommendation. We have removed the Youden Index information from the main text in the "Sensitivity and Specificity" subsection (page 8, paragraph 5).

11. Similarly, any differences in optimal cutoff for the various depression scales are likely due to chance.
12. A third reason that cutoffs might differ is chance - especially given small number of actually depressed people. Please discuss.
Reply: We decided to merge the two comments because they touch on the same issue. We consider that with the current evidence, we cannot rule out the use of the cut-off of 10 or more points. We add in the discussion section:
"Our study used the Youden index to determine the optimal cut-off, but it is important to consider that the cut-off may vary depending on the sample size. A recent simulation study found that for large samples of more than 1000 participants, the optimal sensitivity and specificity values can vary by up to approximately 2 points from the optimal cut-off in cross-sectional studies [83]. Therefore, while a sample size calculation was performed to ensure adequate power, we cannot rule out the use of a cut-off of 10 or more for the Peruvian population. However, within the study, we present the sensitivity and specificity found for such a cut-off."

We add in the Public health implications subsection:
"Although our study found alternative cut-off points to the standard (cut-off ≥10) for the PHQ-9, PHQ-8, and GAD-7 questionnaires, it is important to note that in certain contexts, higher specificity values (cut-off ≥10) may be necessary. These higher values enable a more accurate identification of individuals without depression or anxiety, thereby reducing the likelihood of false-positive results. This reduction in false positives is particularly crucial for alleviating the burden on the healthcare system. A

screening tool with high specificity avoids unnecessary diagnoses and optimizes the use of healthcare resources. Therefore, utilizing a cut-off point of 10 or higher for the PHQ-9, PHQ-8, and GAD-7 can facilitate the early and accurate identification of true cases of depression and anxiety, ensuring that resources are appropriately focused on those who need care and treatment."

We add in the Strengths and Limitations's subsection:
"Fourthly, our study identified a limited number of individuals (n=28) with a diagnosed anxiety condition. Consequently, minor variations in the study cohort could potentially impact the sensitivity or specificity [83]. Nonetheless, we have ensured sufficient statistical power for our analysis based on our sample size calculation. Moreover, all cohort scores on the GAD scale are provided, which can be valuable for future research involving larger numbers of individuals diagnosed with anxiety (refer to Supplementary Material 6). "

13. "However, a meta-analysis found that cutoffs between 8 and 11 showed little difference in sensitivity and specificity [61]." This meta-analysis suffered from selective cutoff reporting. It is unclear what the estimates of sensitivity and specificity mean given that they included different studies at each cutoff.
Reply: Thank you very much for your recommendation. We have removed this statement and the reference.

14. Youden's J is not clinically relelvant. Depending on many factors, one might wish to prioritize sensitivity or specificity, and as such use a different cutoff than that which is optimal via Youden's J (i.e. considering them as equal). Please provide a more nuanced discussion around this point.
Reply: We added this information in the discussion section:
"Although our study found alternative cut-off points to the standard (cut-off ≥10) for the PHQ-9, PHQ-8, and GAD-7 questionnaires, it is important to note that in certain contexts, higher specificity values (cut-off ≥10) may be necessary. These higher values enable a more accurate identification of individuals without depression or anxiety, thereby reducing the likelihood of false-positive results. This reduction in false positives is particularly crucial for alleviating the burden on the healthcare system. A screening tool with high specificity avoids unnecessary diagnoses and optimizes the use of healthcare resources. Therefore, utilizing a cut-off point of 10 or higher for the PHQ-9, PHQ-8, and GAD-7 can facilitate the early and accurate identification of true cases of depression and anxiety, ensuring that resources are appropriately focused on those who need care and treatment."

15. Please include in the abstract the number of subjects assessed as well as the number truly depressed and truly anxious.
Reply: We added the number of truly depressed and truly anxious in the abstract:
"Participants: The sample included 1347 participants. A total of 334 participants (24.8%) were diagnosed with depression, and 28 participants (2.1%) were diagnosed with anxiety."

We also add the same information in the results section:
"A total of 334 participants (24.8%) were diagnosed with depression, and 28 participants (2.1%) were diagnosed with anxiety, as determined through individual psychiatric interviews conducted based on the ICD-10 criteria."

Reviewer: 2
Dr. jarurin pitanupong, Prince of Songkla University
Comments to the Author:
Dear Editor,

Thank you for the opportunity to review this manuscript. This is an interesting study. However, to make this study more valuable, some points should be revised.
1 Title: Sensitivity and specificity of the Patient Health Questionnaire (PHQ-9, PHQ-8, PHQ-2) and General Anxiety Disorder scale (GAD-7, GAD-2) for depression and anxiety diagnosis: A cross-sectional study in a Peruvian population
The title is completed for meaning.
Reply: Thank you for all your comments. We understand that there is no action to be taken here.

2 Abstract
All domains are completed. But Keywords: Depression; Anxiety; Patient Health Questionnaire; Sensitivity and Specificity; Peru should be written in alphabetical order as Anxiety, Depression, Patient Health Questionnaire, Sensitivity and Specificity
Reply: The order of the keywords has been changed.

3 Introduction or background
The content in the introduction section, although some abbreviations are universal, please write the full name of the word and bracket the abbreviation the first time it appears. After that, the author can use abbreviations (such as COVID-19 in line 11, PHQ-9, PHQ-8, PHQ-2, GAD-7, GAD-2 , DASS-21, Kessler-10, HADS-A, HADS, WHO-5). Should check the writing is correct, for example, there is a " ." in the sentence at line 54.
Reply: We modified the background section (page 4, paragraph 2):
"Internationally, the most used screening instruments for depressive and anxious symptomatology are the Patient Health Questionnaire (PHQ-9) [14], PHQ-8 [15], PHQ-2 [16], Generalised Anxiety Disorder (GAD-7) [17], GAD-2 [17], Depression, Anxiety and Stress Scale (DASS-21), Kessler scale-10, Hospital Anxiety and Depression Scale (HADS) [18], Five Well-Being Index (WHO-5) [9]. Most have been validated in several countries, but only the PHQ and GAD have been validated in the Peruvian context [19, 20]."

What is the reason why countries or low or middle economic countries have a cut-off point value that is different from other high economic countries? If possible, please add research that can be used to explain the reasons for finding new cut off points.
Reply: We added this paragraph in the subsection of "Contrast to literature" (page 11, paragraph 3):
"Second, several studies in populations from low- and middle-income countries have reported cut-offs between 5 and 7, for example, Pakistani migrants in the UK [65], Indian adolescents [66], and primary care in Ethiopia [67]. One reason for the difference in cut-off points between high and low-income countries may be due to cultural factors, as culturally diverse groups do not achieve invariance between the PHQ-9 and the GAD-7 [68]. Therefore, factors such as social determinants of health present in such countries may influence cut-off."

In addition, the population of the study was psychiatric patients with physical ailments such as cancer patients, neurology patients and being inpatients. Therefore, it is likely that a review would be required to address the reasons for physical ailments and depression at GAD that required a new cut-off point.
Reply: We have added in Table 1, a description of the physical comorbidities of the participants, and the following text in the results section:
"The most common physical morbidities were cardiovascular diseases (n=111; 8.2%), endocrine, nutritional and metabolic diseases (n=130; 9.7%) and neoplasms, diseases of the blood and haematopoietic organs and other diseases affecting the mechanism of immunity (n=348; 25.8%)."

In the method section, we have added:
“In addition, information was collected on the physical morbidities of the participants based on the ICD-10.”

In the discussion section:
“Our study focuses on a hospital-based clinical population with one or more physical morbidities. It is important to consider that our finding of a different cut-off point, equal to or greater than 10 points for PHQ and GAD, may be influenced by the characteristics of this specific population. It is relevant to note that other studies conducted in hospital settings have found cut-off points lower than the recommendation of equal to or greater than 10 [80, 81]. It is important to bear in mind that the cut-off point may vary depending on the reference group and the context in which it is applied.”

4 Study design (including population, tool, and statistical analysis):
The population is patients with medical or surgical conditions that have a co-existing psychiatric illness or need for psychiatric assessment and management. Why not select a population with direct psychiatric patients without comorbid physical illness? So maybe the title needs to be changed? by being consistent with the specific population?
Reply: The study population is a typical population of many Peruvian hospitals, where patients usually have one or more co-morbidities. Therefore, we believe that the title is correct:
“Sensitivity and specificity of the Patient Health Questionnaire (PHQ-9, PHQ-8, PHQ-2) and General Anxiety Disorder scale (GAD-7, GAD-2) for depression and anxiety diagnosis: A cross-sectional study in a Peruvian hospital population”

5 Result:
Ok

6 Discussion
good discussion that include strengths and limitations, however, this study can't be used in physical outpatients with MDD, GAD, please mention.
Reply: We add in the “strengths and limitations” subsection (page 12, paragraph 4):
“Fifth, our study allows us to obtain sensitivity and specificity values for users in inpatient mental health settings; however, our findings are not generalisable to physical outpatients.”

7 Reference
Very more reference, but depend on the rule of journal.
Finally, please don't give up, but try to rewrite this manuscript to get more valuable research and a better research methodology. Good luck.
Reply: Thanks.

Reviewer: 1
Competing interests of Reviewer: None.

Reviewer: 2
Competing interests of Reviewer: Ok

VERSION 2 – REVIEW

REVIEWER	Benedetti, Andrea McGill University, Departments of Medicine and of Epidemiology, Biostatistics & Occupational Health
REVIEW RETURNED	09-Aug-2023

GENERAL COMMENTS	The authors have responded very well to my concerns. However, the abstract should also reflect a more nuanced understanding of "optimal" cutoffs. Please revise.

REVIEWER	pitanupong, jarurin Prince of Songkla University, psychiatric
REVIEW RETURNED	10-Aug-2023

GENERAL COMMENTS	Good job, accept. Thanks.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Dr. Andrea Benedetti, McGill University

Comments to the Author:

The authors have responded very well to my concerns. However, the abstract should also reflect a more nuanced understanding of "optimal" cutoffs. Please revise.

Reply: We have modified the summary:

“The PHQ-9's ≥7 cut-off point showed the highest simultaneous sensitivity and specificity when contrasted against a psychiatric diagnosis of depression. For a similar contrast against the gold standard, the other optimal cut-off points were: ≥7 for the PHQ-8, and ≥2 for the PHQ-2. In particular, the cut-off point ≥8 had good performance for GAD-7 with sensitivity and specificity, and cut-off point ≥10 had lower levels of sensitivity, but higher levels of specificity, compared to the cut-off point of ≥8. Also, we present the sensitivity and specificity values of each cut-off point in PHQ-9, PHQ-8, PHQ-2, GAD-7 and GAD-2. We confirmed the adequacy of a one-dimensional model for the PHQ-9, PHQ-8, and GAD-7, while all PHQ and GAD scales showed good reliability.”

Reviewer: 2

Dr. jarurin pitanupong, Prince of Songkla University

Comments to the Author:

Good job, accept. Thanks

Reply: Thanks.