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

David Villarreal-Zegarra, Juan Barrera-Begazo, Sharlyn Otazú-Alfaro, Nikol Mayo-Puchoc, Juan Carlos Bazo-Alvarez, Jeff Huarcaya-Victoria

ABSTRACT

Objectives The Patient Health Questionnaire (PHQ) and Generalised Anxiety Disorder Scale (GAD) are widely used screening tools, but their sensitivity and specificity in low-income and middle-income countries are lower than in high-income countries. We conducted a study to determine the sensitivity and specificity of different versions of these scales in a Peruvian hospital population.

Design Our study has a cross-sectional design.

Setting Our participants are hospitalised patients in a Peruvian hospital. The gold standard was a clinical psychiatric interview following ICD-10 criteria for depression (F32.0, F32.1, F32.2 and F32.3) and anxiety (F41.0 and F41.1).

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.

Results The PHQ-9’s ≥7 cut-off point showed the highest simultaneous sensitivity and specificity when compared 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 with 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.

Conclusions The PHQ and GAD have adequate measurement properties in their different versions. We present specific cut-offs for each version.

BACKGROUND

Until 2019, approximately 280 million people worldwide suffered from depression and 302 million from anxiety. These data reveal that both mental disorders are the most common in the world and lead to the causes of the global burden of mental health disability-adjusted life-years. With the onset of the COVID-19 pandemic, the worldwide prevalence of both disorders increased by around 25%. In Peru, during the COVID-19 pandemic, the prevalence of moderate depressive symptoms also increased by approximately 0.17% in each quarter. However, no population-level evidence has been found about the prevalence of anxious symptomatology or the diagnosis of anxiety in Peru. In this context, the impact of COVID-19 on the prevalence and burden of major depression and anxiety disorders was measured using screening tools. In addition, it was noted that during the pandemic, there was a reduction in the number of mental health service users being seen. Screening tools assist in early diagnosis and intervention that can prevent disease progression and reduce years lost to disability. They are beneficial in contexts with limited mental health professionals providing care to large...
populations, such as in Peru. The opportune identification of people at risk of depression reduces treatment costs and disease burden. Depressive symptom screening is also helpful in national surveys and epidemiological research since, unlike diagnostic instruments, screening measures are typically brief, quick and easy to administer. Internationally, the most used screening instruments for depressive and anxious symptomatology are the Patient Health Questionnaire (PHQ-9), PHQ-8, PHQ-2, Generalised Anxiety Disorder (GAD-7), GAD-2, Depression, Anxiety and Stress Scale-21, Kessler scale-10, Hospital Anxiety and Depression Scale, Five Well-Being Index. Most have been validated in several countries, but only the PHQ and GAD have been validated in the Peruvian context.

In particular, the PHQ versions (PHQ-9, PHQ-8, PHQ-2) and GAD versions (GAD-7, GAD-2) are the most widely used, having extensive evidence of their validity and reliability. However, correctly identifying people at risk of depression or anxiety requires more than internal/externally valid and reliable screening measures; defining an accurate cut-off point for their raw scales (ie, to reach valid interpretations) is also necessary. Such a cut-off point can vary across cultures and subpopulations (eg, general vs clinical), so a local calibration is usually needed. Studies of the different versions of the PHQ and GAD have yielded heterogeneous cut-offs, as they vary between different cultures and populations, such as clinical and general populations. However, several systematic reviews suggest that cut-off 10 is most appropriate for the PHQ-9, PHQ-8, and GAD-7, and cut-offs 2–3 for the PHQ-2 and GAD-2. Furthermore, concerning the PHQ-9 correctness, the summed item score method is the most used compared with the algorithm. However, other forms of correction using diagnostic algorithms are available.

Sensitivity and specificity studies have been barely performed in low-income and middle-income countries. Several of these populations do not count with verified cut-off points from calibration studies (including Peruvian populations), in particular, the inpatient population is particularly vulnerable as they have physical comorbidities that may influence the establishment of cohort points. Therefore, our aim was to determine the optimal cut-off point for the PHQ-9, PHQ-8, PHQ-2, GAD-7 and GAD-2 to discriminate a formal depression and anxiety diagnosis in the Peruvian hospital population. In addition, as secondary objectives, we assessed these scales’ internal structure and reliability.

METHODS
Study design
This study has a cross-sectional design, and we used the Standards for Reporting of Diagnostic Accuracy Studies (STARD 2015).

Participants
The participants were patients from the Liaison Psychiatry Unit of a hospital in Lima, Peru. Psychiatric liaison services provide psychiatric consultation to hospitalised patients with medical or surgical conditions that have a coexisting psychiatric illness or need for psychiatric assessment and management. The total number of participants in our study is similar to the proportion of people who were hospitalised in 2022 in our setting (see online supplemental material 1). The evaluation period started in September 2020 and finished in August 2022. Sampling was non-probabilistic and applied to all participants arriving at the Liaison Psychiatry Unit. The inclusion criteria were that they had complete PHQ-9 and GAD-7 data and were of legal age (>18 years). Participants with missing data were excluded.

The sample size calculation for the PHQ versions was based on an estimated sensitivity of 0.88 and specificity of 0.85, a confidence level of 95%, a prevalence of 6.4% and a drop-out rate of 10%, giving an estimate of 705 participants. The sample size calculation for the GAD versions was based on an estimated sensitivity of 0.83 and specificity of 0.84, a confidence level of 95%, a prevalence of 8.7% and a drop-out rate of 10%, giving an estimate of 694 participants. The web programme based on the paper by Buderer was used to calculate the sample size.

Setting
The Guillermo Almenara Irigoyen National Hospital (HNGAI) was the study site, a highly complex hospital in Lima-Peru (capital city). HNGAI is one of the three largest hospitals of the Social Security system in Peru based on the number of beds (960 hospital beds) and is also a tertiary referral centre for all medical specialties, including psychiatry. It provides healthcare services to 1,547,840 individuals from social insurance. Because it attends to virtually all pathologies, from the simplest to the most complex, it was classified in 2015 as a Specialised Health Institute III-2, the highest level awarded by the Ministry of Health of Peru to hospital establishments.

The Liaison Psychiatry Unit at HNGAI is responsible for consulting with and providing care to patients with depression and anxiety conditions. The Liaison Unit has been recorded in a Google Form to track better the patients treated.

Instruments and variables
PHQ-9, PHQ-8 and PHQ-2
The PHQ is an instrument designed to measure depressive symptoms over the past 2 weeks, according to the diagnostic criteria of the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV), criteria that were retained in the DSM-5. The scale has four response options (0=none, 1=some days, 2=more than half of the...
days, 3=almost every day). The scale had many versions, including the PHQ-9, the full version with nine items and scores ranging from 0 to 27. In Peru, the PHQ-9 had good psychometric properties in terms of structural validity (Comparative fit index [CFI]=0.936; Root Mean Square Error of Approximation [RMSEA]=0.089; Standardized Root Mean Square Residual [SRMR]=0.039), internal consistency (α=ω=0.87) and invariance between age and sex (ΔCFI=0.01). In addition, PHQ-9 had scoring versions related to the DSM-5 indicators, which state that for a case to be positive, there must be at least five depressive symptoms present, and at least one of them must be core depressive symptoms (item 1 and item 2). First, the PHQ-9 algorithm suggests that a symptom is positive if it scores two or more, except the ninth item, suicidal ideation, which is positive if it scores 1 or more. Second, the PHQ-9 adjusted algorithm proposes that a symptom was positive if it scored 1 or more for any of the items in the instrument.

The PHQ-8 was a shortened version of the PHQ-9 without the last item on suicidal ideation. The PHQ-8 was as valuable as the PHQ-9 in detecting cases of major depression. The PHQ-2 is an abbreviated version of the PHQ-9 with only two items, focusing on the first two items related to the core symptoms of depression (anhedonia and depressed mood) and providing scores between 0 and 6. The PHQ-2 was validated in Peru and showed adequate levels of internal consistency (α=0.80).

GAD-7 and GAD-2
The GAD Scale was a Likert-type rating scale with four response options ranging from 0 (not at all) to 3 (almost every day), based on DSM-IV criteria and assesses anxious symptoms during the past 2 weeks. The GAD-7 was the version of the instrument with the original seven items and had a range of scores from 0 to 21. The GAD-7 had good psychometric properties in the Peruvian context for a one-dimensional model (CFI=0.995, Tucker-Lewis index [TLI]=0.999, RMSEA=0.056), adequate internal consistency (ω=0.89) and invariance according to sex (ΔCFI=0.01).

The GAD-2 was adapted from the GAD-7, focusing on the emotional and cognitive expressions of DSM-IV anxiety items 1 and 2. The GAD-2 shows good internal consistency values (ω=0.80) and a relationship with its extended version (r>0.80) in Peruvian context.

Gold standard
The gold standard was an individual clinical psychiatric interview following the criteria of International Classification of Diseases, Tenth Revision, (ICD-10). The clinical assessments were performed by psychiatrists who are members of the Liaison Psychiatry Unit, all of whom have at least 3 years of clinical experience evaluating the psychiatric needs of hospitalised patients. The interview focused on assessing whether the participants had depressive disorder (F32.0, F32.1, F32.2 and F32.3) or anxiety disorder (F41.0 and F41.1), with a duration between 25 and 30 min. The individual clinical psychiatric interview and the psychometric instruments (ie, 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 min (SD=4.5 min), and the order (ie, psychometric instruments before or after the interview) was randomly assigned.

Sociodemographic covariates
Data were collected on sex (male, female), age, marital status (single, married/cohabitant, separated, widowed), educational level (none, elementary, high school, technical, college), currently works (no, yes, retired), living alone (yes, no) and history of psychiatric diagnosis (yes/no). In addition, information was collected on the physical diagnosis of the participants based on the ICD-10.

Statistical analysis
The sociodemographic covariates of the participants were described at frequency and percentage levels. The internal consistency and internal structure analyses were performed with R Studio, with the ‘Lavaan’, ‘Semtools’ and ‘Semplot’ packages (see online supplemental material 2). Sensitivity, specificity and correlation analyses were analysed with Stata V.15 (see online supplemental material 3).

Sensibility and specificity
The PHQ-9, PHQ-8, PHQ-9 algorithm, PHQ-2 were evaluated as diagnostic tests and compared against the gold standard. In addition, the GAD-7 and GAD-2 were scored and compared against the diagnosis of anxiety through the clinical interview (gold standard).

We calculated the positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (+LR), negative LR (−LR) and Youden index. PPV and NPV refer to the proportion of patients correctly diagnosed as positive or negative, respectively. The LR+ is the probability that a person with the disease will test positive given the probability that a person without the disease will test positive. While the LR− is the probability that a person with the disease will test negative given the probability that a person without the disease will test negative. 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. 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.

Internal structure
Confirmatory factor analysis (CFA) was performed considering a one-dimensional model for the PHQ-9,
PHQ-8 and GAD-7. We used the weighted least square mean and variance adjusted estimator and polychoric matrices as it best fits the categorical-ordinal nature of the data. Models were evaluated using a set of goodness-of-fit indices such as CFI and TLI, which must be greater than 0.95 to be considered adequate. In addition, the SRMR and RMSEA at 90% confidence were estimated, which must have values less than 0.08 to be considered adequate. It was impossible to perform a CFA for the PHQ-2 and the GAD-2 because a minimum of three items are required for such analysis.

Internal consistency
We calculated the alpha (α) and McDonald's omega coefficients (ω). Values greater than 0.70 are considered adequate.

Patient and public involvement
No patient involved.

RESULTS
Participants
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 online supplemental material 4). Thus, our study only included 1347 participants (see table 1). Most participants were female (59.4%; n=800), married or living with a partner (57.0%; n=768) and had higher technical or university education (53.5%; n=721). 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.

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%).

Sensibility and specificity
In online supplemental material 5, we provide the values of all cut-off points for the different versions of the PHQ. The cut-off points ≥7 in the PHQ-9 had the best balance between sensitivity and specificity of all the cut-off points evaluated in the various versions of the PHQ, as it obtained a sensitivity of 76.0 (95% CI 71.1 to 80.5) and specificity of 72.1 (95% CI 69.2 to 74.8) (see online supplemental material 6). In addition, the PHQ-9 with a cut-off of ≥10 points (ie, the most used) showed lower levels of sensitivity (54.2; 95% CI 8.7 to 59.6), but higher level of specificity (87.4; 95% CI 85.2 to 89.3), compared with the cut-off point of ≥7.

The algorithm score method for PHQ-9 had low levels of sensitivity (34.7; 95% CI 29.6 to 40.1) but high levels of specificity (93.4; 95% CI 91.7 to 94.8) compared with the raw score method for PHQ-9 with ≥7 cohort points. In contrast, the adjusted algorithm method for PHQ-9 showed slightly higher sensitivity values (78.1; 95% CI 73.3 to 82.5) and better specificity values (66.4; 95% CI 63.4 to 69.3) compared with the raw score method for PHQ-9 with ≥7 cohort points. The raw score for PHQ-9 with cohort point ≥7 showed a better balance between sensitivity and specificity compared with the algorithm method or the algorithm adjusted for PHQ-9.

The best cut-off point found in the PHQ-8 was ≥7 points, as it had a sensitivity of 79.9 (95% CI 75.2 to 84.1), and a specificity of 66.0 (95% CI 63.0 to 69.0) (see online supplemental material 6). The best cut-off point found in the PHQ-2 was ≥2 points, as it had a sensitivity of 84.7 (95% CI 80.4 to 88.4), and a specificity of 55.9 (95% CI 52.8 to 59.0) (see online supplemental material 6).

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 online supplemental material 7. In particular, the cut-off point ≥8 had good performance for GAD-7 with sensitivity values of 53.6 (95% CI 33.9 to 72.5) and specificity of 78.8 (95% CI 76.5 to 81.0) (see online supplemental material 6). The GAD-7’s cut-off point ≥10 (ie, the most used) had lower levels of sensitivity (39.3; 95% CI 21.5 to 59.4), but higher levels of specificity (88.4; 95% CI 86.5 to 90.1), compared with 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 to 88.4) and a specificity of 50.1 (95% CI 47.4 to 52.8) (see online supplemental material 6).

Internal structure
The PHQ-9 one-dimensional model showed adequate goodness-of-fit (χ²=251.9; df=27; CFI=0.974; TLI=0.965; SRMR=0.051; RMSEA (90% CI)=0.079 (0.070 to 0.088)), while the PHQ-8 one-dimensional model reported a similar goodness-of-fit (χ²=209.7; df=20; CFI=0.977; TLI=0.977; SRMR=0.050; RMSEA (90% CI)=0.082 (0.072 to 0.093)). The GAD-7 also showed adequate goodness-of-fit (χ²=122.3; df=14; CFI=0.977; TLI=0.966; SRMR=0.043; RMSEA (90% CI)=0.076 (0.064 to 0.088)).

Reliability
The PHQ-9 (α=0.89; ω=0.86), the PHQ-8 (α=0.88; ω=0.85) and the GAD-7 (α=0.85; ω=0.81) showed optimal internal consistency values. Similarly, the PHQ-2 (α=0.83; ω=0.80) and the GAD-2 (α=0.74; ω=0.70) also showed adequate internal consistency scores. Table 2 shows the raw scores.
<table>
<thead>
<tr>
<th>Table 1</th>
<th>Sociodemographic characteristics (n=1347)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>547</td>
</tr>
<tr>
<td>Women</td>
<td>800</td>
</tr>
<tr>
<td><strong>Age (categories)</strong></td>
<td></td>
</tr>
<tr>
<td>18–29</td>
<td>107</td>
</tr>
<tr>
<td>30–39</td>
<td>164</td>
</tr>
<tr>
<td>40–49</td>
<td>214</td>
</tr>
<tr>
<td>50–59</td>
<td>284</td>
</tr>
<tr>
<td>60–69</td>
<td>294</td>
</tr>
<tr>
<td>70–79</td>
<td>203</td>
</tr>
<tr>
<td>80 to more</td>
<td>81</td>
</tr>
<tr>
<td><strong>Civil status</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>329</td>
</tr>
<tr>
<td>Married or cohabitant</td>
<td>768</td>
</tr>
<tr>
<td>Separated</td>
<td>133</td>
</tr>
<tr>
<td>Widowed</td>
<td>117</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>13</td>
</tr>
<tr>
<td>Elementary school</td>
<td>135</td>
</tr>
<tr>
<td>High school</td>
<td>478</td>
</tr>
<tr>
<td>Technical</td>
<td>246</td>
</tr>
<tr>
<td>University</td>
<td>475</td>
</tr>
<tr>
<td><strong>Currently works</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>330</td>
</tr>
<tr>
<td>Yes</td>
<td>778</td>
</tr>
<tr>
<td>Retired</td>
<td>239</td>
</tr>
<tr>
<td><strong>Living alone</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>99</td>
</tr>
<tr>
<td>No</td>
<td>1248</td>
</tr>
<tr>
<td><strong>History of psychiatric diagnosis</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>388</td>
</tr>
<tr>
<td>No</td>
<td>959</td>
</tr>
<tr>
<td><strong>Diagnosis of depression</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1013</td>
</tr>
<tr>
<td>Yes</td>
<td>334</td>
</tr>
<tr>
<td><strong>Diagnosis of anxiety</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1319</td>
</tr>
<tr>
<td>Yes</td>
<td>28</td>
</tr>
<tr>
<td><strong>Physical illnesses</strong></td>
<td></td>
</tr>
<tr>
<td>A00–B99 Certain infectious and parasitic diseases</td>
<td>109</td>
</tr>
<tr>
<td>C00–D48 Neoplasms, and diseases of the blood and haematopoietic organs and other disorders affecting the mechanism of immunity</td>
<td>348</td>
</tr>
<tr>
<td>E00–E90 Endocrine, nutritional and metabolic diseases</td>
<td>130</td>
</tr>
<tr>
<td>G00–G99 Diseases of the nervous system</td>
<td>96</td>
</tr>
<tr>
<td>H00–H59 Diseases of the eye and adnexa</td>
<td>17</td>
</tr>
</tbody>
</table>

Continued
DISCUSSION
Main findings
We determined the target population’s optimal cut-off points for PHQ scale. The PHQ-9’s ≥7 cut-off point showed the highest sensitivity and specificity when contrasted against a psychiatric diagnosis of depression (gold standard). For a similar contrast, the other optimal cut-off points were: ≥7 for the PHQ-8 and ≥2 for the PHQ-2. In addition, the algorithm scoring or algorithm-adjusted scoring methods for the PHQ-9 had a lower balance between sensitivity and specificity scores than adjusted scoring methods for the PHQ-8. The PHQ-9 raw score scoring method with a cut-off ≥7 contrasted against a psychiatric diagnosis of depression showed the highest sensitivity and specificity when compared against a psychiatric diagnosis of depression. For a similar contrast, the other optimal cut-offs ≥7 had a lower balance between high-sensitivity and specificity scores than adjusted scoring methods for the PHQ-8.

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 confirmed the adequacy of a one-dimensional model for the PHQ-9, PHQ-8 and GAD-7, while all scales showed good internal consistency.

Contrast to literature
At the PHQ-9 level, evidence suggests that the raw score approach is more valuable than diagnostic algorithms, which is consistent with our findings. For the cut-off, different systematic reviews agree that the most commonly used cut-off is ≥10. The optimal cut-off reported in our study was slightly lower than that suggested by the other studies, and two possible factors could explain this difference. First, our population is inpatients in different areas of a high-complexity hospital. Other studies of hospitalised patients with cancer, hospitalised neurology patients and patients with coronary heart disease also found an optimal cut-off between 5 and 7 points. Therefore, hospitalised individuals may be more likely to have depressive symptoms, which may require a lower cut-off on the PHQ-9. Second, several studies in populations from low-income and middle-income countries have reported cut-offs between 5 and 7, for example, Pakistani migrants in the UK, Indian adolescents and primary care in Ethiopia. One reason for the difference in cut-off points between high-income 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. Therefore, factors such as social determinants of health present in such countries may influence cut-off.

Concerning the PHQ-8 and PHQ-9, we found that both scales have similar cut-off points (≥7). Our findings are consistent with a meta-analysis that found that the cut-offs between the two scales are identical; although sensitivity may be minimally reduced with the PHQ-8, specificity is similar between the two scales. The PHQ-8 does not include the item corresponding to suicidal or self-harming ideation, and the use of this version of the PHQ scale may be minimally reduced with the PHQ-8.
is common in the general population, as suicidal ideation is less common in this group. However, at the level of clinical populations, it has been found that omitting this item does not significantly alter the measurement capabilities of the PHQ, as the correlation between the PHQ-8 and PHQ-9 in clinical populations is very close to 1.58

Regarding the GAD-7, our findings are consistent with a meta-analysis that evaluated all possible cut-off points and reported that 28 is the most appropriate for anxiety disorder. It also notes that scores between 7 and 10 points have similar sensitivity and specificity values. Other recent primary studies conducted in hospitalised populations or people with chronic diseases in hospital settings also found optimal cut-offs between 7 and 10 points.69-71

Our results on PHQ-2 were in line with meta-analyses supporting the use of the cut-off of 2 for PHQ-2.35,72 Also, the values most frequently for GAD-2 are cut-off ≥2 and ≥3.18,37,73 The meta-analyses mentioned included studies in general populations (ie, people attending primary care) and people hospitalised for non-communicable or infectious diseases. However, no meta-analyses were found that evaluated cut-off for hospitalised people only. At the level of primary studies, the evidence suggests that cut-offs vary between 2 and 3 points for the PHQ-2 and GAD-2.74,75

Regarding internal validity, a systematic review examined the factor structure of the PHQ-9, noting that the one-dimensional model has been repeatedly confirmed across studies.76 Although several studies evaluated alternative multidimensional models (eg, two-dimensional, three-dimensional or bifactorial models), their dimensions are often highly correlated with each other, so there may be overlapping.76 We did not find systematic reviews on the internal structure of the GAD-7 and the PHQ-8. However, several studies support the one-dimensional model in hospitalised patients for both the PHQ-877 and GAD-7.24,25 In Peru, the GAD-7 and PHQ-9 have shown evidence of a one-dimensional factor structure in different populations, such as the general population,20 pregnant women,31 and university students.32,78 However, no studies have been found evaluating the factor structure of the PHQ-8 in the Peruvian population.

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, 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 10.79,80

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. 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.81 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.

Public health implications

The evaluated instruments are widely used in clinical practice and research to measure symptoms of depression and anxiety, but from today, users will have optimal cut-off points for interpretations. This can help healthcare professionals identify people at risk of depression and anxiety more accurately while informing decisions about their formal diagnosis and consequent treatment. This is especially valuable in hospital environments, where time is crucial.

Our findings are of particular interest to the Peruvian health system, which has clinical practice guidelines for depression that recommend the PHQ-9 as a screening tool in primary care and hospital context.82 Although our results correspond only to a hospital population, our study is the closest approximation to an evaluation of sensitivity and specificity in the Peruvian context, in the absence of similar studies in primary care. On the other hand, there is a lack of national clinical practice guidelines for screening and managing anxiety in Peru. Therefore, our study could contribute to future clinical practice guidelines for GAD.

Although our study found alternative cut-off points to the standard (cut-off≥10) for the PHQ-9 and PHQ-8 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 optimises the use of healthcare resources. Therefore, using 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.

Strengths and limitations

Our results of the study have several strengths. First, to our knowledge, this is the first study in a Peruvian context that evaluates the factorial structure of all PHQ and GAD versions in a hospitalised population. Second, the scales were administered by a team of healthcare professionals with more than 5 years of experience in the clinical assessment of these patients. Third, the sample size was large enough to support all analyses and conclusions. Further, our sample size was larger than other recently published...
Conclusions
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. Also, we present the sensitivity and specificity values of each cut-off point in 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.

Author affiliations
1. Escuela de Medicina, Universidad César Vallejo, Trujillo, Peru
2. Instituto Peruano de Orientación Psicológica, Lima, Peru
3. Research Department of Primary Care and Population Health, University College London (UCL), London, UK
4. Escuela Profesional de Medicina Humana, Universidad Privada San Juan Bautista, Ica, Peru
5. Unidad de Psiquiatría de Enlace, Departamento de Psiquiatría, Hospital Nacional Guillermo Almenara Irigoyen, EsSalud, Lima, Peru

Twitter David Villarreal-Zegarra @davillarrealz

Contributors DV-Z contributed to the conceptualising the study, designing the methodology, developing the software tools, validating the results, conducting formal analyses, curating and managing the data, and contributed to the initial drafting and visualisation of the manuscript. J-B-B contributed to the formal analysis, performed investigations and aided in visualising the findings. S-O-A participated in the investigation phase and contributed to the initial drafting of the manuscript. NM-P engaged in formal analysis, conducted investigations and contributed to the initial drafting of the manuscript. J-C-A contributed to the methodology, conducted investigations, provided critical input for the manuscript in the review and editing stages, and played a supervisory role. JH-V contributed to the conceptualising the study, designing the methodology, developing software tools, validating the results, conducting investigations, managing resources, curating data, project administration responsibilities, participated in reviewing and editing the manuscript, and had responsibility for the overall content as a guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The Hospital Nacional Guillermo Almenara Irigoyen’s Institutional Review Board (Nota No 52 CIEI-GRPA-Essalud-2023) approved the protocol of our study. Throughout the study, the researchers had no access to identifying information about the participants. In addition, participants gave informed consent. All participants were users of the hospital’s Liaison Psychiatry Unit and received psychological or psychiatric care as needed.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The database can be requested from the corresponding author.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
David Villarreal-Zegarra http://orcid.org/0000-0002-2222-4764
Jeff Huarcaya-Victoria http://orcid.org/0000-0003-4525-9545

REFERENCES
2. GBD 2019 Mental Disorders Collaborators. Mental disorders collaborators: global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic


78 Inagaki M, Ohtsuki T, Hara M, et al. TheValidation of the 7-item generalized anxiety disorder scale (GAD-7) in general internal medicine and primary care patients. *BMC Fam Pract* 2012;13:146.


