



BMJ Open 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 ^{1,2}, Juan Barrera-Begazo,² Sharlyn Otazú-Alfaro,² Nikol Mayo-Puchoc,² Juan Carlos Bazo-Alvarez,³ Jeff Huarcaya-Victoria ^{4,5}

To cite: Villarreal-Zegarra D, Barrera-Begazo J, Otazú-Alfaro S, *et al*. 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. *BMJ Open* 2023;**13**:e076193. doi:10.1136/bmjopen-2023-076193

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2023-076193>).

Received 01 June 2023
Accepted 21 August 2023



© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Jeff Huarcaya-Victoria;
jeff.huarcaya@upsjb.edu.pe

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 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 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.¹ These data

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Study methods allowed us to establish clinically meaningful cut-off points for Patient Health Questionnaire and Generalised Anxiety Disorder Scale.
- ⇒ Sample size was larger than in other similar studies and large enough to support all analyses and conclusions.
- ⇒ Research findings may not be directly applicable to some hospital or primary care settings due to the specific context of our study population.

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.^{2,3} With the onset of the COVID-19 pandemic, the worldwide prevalence of both disorders increased by around 25%.⁴ In Peru, during the COVID-19 pandemic, the prevalence of moderate depressive symptoms also increased by approximately 0.17% in each quarter.⁵ 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.⁶ In addition, it was noted that during the pandemic, there was a reduction in the number of mental health service users being seen.⁷

Screening tools assist in early diagnosis and intervention that can prevent disease progression and reduce years lost to disability.⁸ 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.^{9–11} Depressive symptom screening is also helpful in national surveys and epidemiological research¹² since, unlike diagnostic instruments, screening measures are typically brief, quick and easy to administer.^{13–14} Internationally, the most used screening instruments for depressive and anxious symptomatology are the Patient Health Questionnaire (PHQ-9),¹⁵ PHQ-8,¹⁶ PHQ-2,¹⁷ Generalised Anxiety Disorder (GAD-7),¹⁸ GAD-2,¹⁸ Depression, Anxiety and Stress Scale-21, Kessler scale-10, Hospital Anxiety and Depression Scale,¹⁹ Five Well-Being Index.¹⁰ Most have been validated in several countries, but only the PHQ and GAD have been validated in the Peruvian context.^{20–21}

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.^{22–24} 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.²⁵ Studies of the different versions of the PHQ and GAD have yielded heterogeneous cut-offs, as they vary between different cultures^{21–26–29} and populations, such as clinical and general populations.^{30–32} However, several systematic reviews suggest that cut-off 10 is most appropriate for the PHQ-9, PHQ-8, and GAD-7,^{33–37} and cut-offs 2–3 for the PHQ-2 and GAD-2.^{35–37} 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.^{38–39}

Sensitivity and specificity studies have been barely performed in low-income and middle-income countries.⁴⁰ 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).⁴¹

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 0.85,³³ a confidence level of 95%, a prevalence of 6.4%^{42–43} 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 0.84,¹⁸ a confidence level of 95%, a prevalence of 8.7%⁴⁴ 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.⁴⁵

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 specialities, including psychiatry (<http://www.essalud.gob.pe/estadistica-institucional/>). 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 responding to consultation requests from different clinical-surgical services at HNGAI.⁴⁶ As part of the evaluation of each patient, in addition to the clinical interview and psychiatric diagnosis, standardised assessments such as the PHQ-9 and GAD-7 are used to ensure adequate monitoring and assess response to the established treatment. Since September 2020, the services provided by the Liaison Unit have 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=no days, 1=some days, 2=more than half of the

days, 3=almost every day).¹⁵ 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 ($\alpha = \omega=0.87$) and invariance between age and sex ($\Delta\text{CFI}<0.01$).²⁰

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.⁴⁷ 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.⁴⁸

The PHQ-8 was a shortened version of the PHQ-9 without the last item on suicidal ideation.¹⁶ The PHQ-8 was as valuable as the PHQ-9 in detecting cases of major depression.⁴⁹ 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 ($\alpha=0.80$).⁵⁰

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.⁵¹ 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]=0.992, RMSEA=0.056), adequate internal consistency ($\omega=0.89$) and invariance according to sex ($\Delta\text{CFI}\leq 0.01$).⁵²

The GAD-2 was adapted from the GAD-7, focusing on the emotional and cognitive expressions of DSM-IV anxiety (items 1 and 2).⁵³ The GAD-2 shows good internal consistency values ($\omega=0.80$) and a relationship with its extended version ($r>0.80$) in Peruvian context.⁵²

Gold standard

The gold standard was an individual clinical psychiatric interview following the criterial 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 5 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-9 adjusted algorithm and 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.⁵⁴ 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.⁵⁵ 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.⁵⁵ 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.⁵⁴ 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 estimator⁵⁶ and polychoric matrices as it best fits the categorical-ordinal nature of the data.⁵⁷ 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.⁵⁸ In addition, the SRMR and RMSEA at 90% confidence were estimated, which must have values less than 0.08 to be considered adequate.⁵⁸ 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.⁵⁹

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%).

Sensitivity 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 ($\chi^2=251.9$; $df=27$; 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 ($\chi^2=202.7$; $df=20$; 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 ($\chi^2=122.3$; $df=14$; CFI=0.977; TLI=0.966; SRMR=0.043; RMSEA (90% CI)=0.076 (0.064 to 0.088)).

Reliability

The PHQ-9 ($\alpha=0.89$; $\omega=0.86$), the PHQ-8 ($\alpha=0.88$; $\omega=0.85$) and the GAD-7 ($\alpha=0.85$; $\omega=0.81$) showed optimal internal consistency values. Similarly, the PHQ-2 ($\alpha=0.83$; $\omega=0.80$) and the GAD-2 ($\alpha=0.74$; $\omega=0.70$) also showed adequate internal consistency scores. [Table 2](#) shows the raw scores.

Table 1 Sociodemographic characteristics (n=1347)

	n	%
Sex		
Men	547	40.6
Women	800	59.4
Age (categories)		
18–29	107	7.9
30–39	164	12.2
40–49	214	15.9
50–59	284	21.1
60–69	294	21.8
70–79	203	15.1
80 to more	81	6.0
Civil status		
Single	329	24.4
Married or cohabitant	768	57.0
Separated	133	9.9
Widowed	117	8.7
Education level		
None	13	1.0
Elementary school	135	10.0
High school	478	35.5
Technical	246	18.2
University	475	35.3
Currently works		
No	330	24.5
Yes	778	57.8
Retired	239	17.7
Living alone		
Yes	99	7.3
No	1248	92.7
History of psychiatric diagnosis		
Yes	388	28.8
No	959	71.2
Diagnosis of depression		
No	1013	75.2
Yes	334	24.8
Diagnosis of anxiety		
No	1319	97.9
Yes	28	2.1
Physical illnesses		
A00–B99 Certain infectious and parasitic diseases	109	8.1
C00–D48 Neoplasms, and diseases of the blood and haematopoietic organs and other disorders affecting the mechanism of immunity	348	25.8
E00–E90 Endocrine, nutritional and metabolic diseases	130	9.7
G00–G99 Diseases of the nervous system	96	7.1
H00–H59 Diseases of the eye and adnexa	17	1.3

Continued

Table 1 Continued

	n	%
H60–H95 Diseases of the ear and mastoid process	17	1.3
I00–I99 Diseases of the circulatory system	111	8.2
J00–J99 Diseases of the respiratory system	107	7.9
K00–K93 Diseases of the gastro-intestinal tract	106	7.9
L00–L99 Diseases of the skin and subcutaneous tissues	74	5.5
M00–M99 Diseases of the musculo-skeletal system and connective tissue	97	7.2
N00–N99 Diseases of the genito-urinary system	97	7.2
O00–O99 Pregnancy, childbirth and puerperium	10	0.7
P00–P96 Certain conditions originating in the perinatal period	0	0.0
Q00–Q99 Congenital malformations, deformities and chromosome anomalies	6	0.4
R00–R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	46	3.4
S00–T98 Trauma, poisoning and certain other consequences of external cause	50	3.7
V01–Y98 External causes of morbidity and mortality	4	0.3
Z00–Z99 Factors influencing health status and contact with healthcare services	90	6.7
U00–U99 Codes for special situations	28	2.1

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 the PHQ-9 raw score scoring method with a cut-off ≥ 7 . 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.

Table 2 Raw scores and internal consistency (n=1347)

	M	SD	Min	Max	α	ω
PHQ-9 score	6.4	5.0	0	27	0.89	0.86
PHQ-8 score	6.1	4.7	0	24	0.88	0.85
PHQ-2 score	1.9	1.6	0	6	0.83	0.80
GAD-7 score	5.1	3.9	0	21	0.85	0.81
GAD-2 score	1.7	1.4	0	6	0.74	0.70

Note: α =classical alpha. ω =McDonald's omega.
GAD, Generalised Anxiety Disorder; PHQ, Patient Health Questionnaire.

Contrast to literature

At the PHQ-9 level, evidence suggests that the raw score approach is more valuable than diagnostic algorithms,³³ which is consistent with our findings. For the cut-off, different systematic reviews agree that the most commonly used cut-off is ≥ 10 .^{33 60} 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,⁶¹ hospitalised neurology patients⁶² and patients with coronary heart disease⁶³ 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,⁶⁴ Indian adolescents⁶⁵ and primary care in Ethiopia.⁶⁶ 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-7.⁶⁷ 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.³⁶ The PHQ-8 does not include the item corresponding to suicidal or self-harming ideation, and the use of this version of the PHQ

is common in the general population, as suicidal ideation is less common in this group.¹⁶ 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 1.⁶⁸

Regarding the GAD-7, our findings are consistent with a meta-analysis that evaluated all possible cut-off points and reported that ≥ 8 is the most appropriate for anxiety disorder.¹⁸ It also notes that scores between 7 and 10 points have similar sensitivity and specificity values.¹⁸ 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 frequented 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.^{21 27} 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 women²¹ and university students.^{52 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

studies'.⁶⁰ Fourth, our study is the first Peruvian study to evaluate the sensitivity and specificity of the PHQ.

Our study has limitations. First, we conducted the study only in a hospital context in a Peruvian city, which limits its applicability to other settings in Peru or other countries. However, it could be used in other Peruvian hospital contexts with similar characteristics, which is relevant because hospital care in Peru (levels II and III of complexity) represents 58.65% of total care.⁸³ Second, the generalisability of our results may be limited because the sampling is not probabilistic, as it does not include other hospitals. However, the hospital where we conducted the study serves 1.1% of all nationally insured EsSalud patients (<http://www.essalud.gob.pe/estadistica-institucional/>). It is also a national referral hospital, which means that people from all over the country are referred to this hospital for treatment. Therefore, the representativeness of the results is ensured. Third, 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 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. Fourth, 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 online supplemental material 7). 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.

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

¹Escuela de Medicina, Universidad César Vallejo, Trujillo, Peru

²Instituto Peruano de Orientación Psicológica, Lima, Peru

³Research Department of Primary Care and Population Health, University College London (UCL), London, UK

⁴Escuela Profesional de Medicina Humana, Universidad Privada San Juan Bautista, Ica, Peru

⁵Unidad de Psiquiatría de Enlace, Departamento de Psiquiatría, Hospital Nacional Guillermo Almenara Irigoyen, EsSalud, Lima, Peru

Twitter David Villarreal-Zegarra @dvillarrealz

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. JB-B contributed to the formal analysis, performed investigations and aided in visualising the findings. SO-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. JCB-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-OlyD-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 responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error 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

- 1 Global Health data Exchange (GHDx), Institute for Health Metrics and Evaluation (IHME). n.d. Available: <https://microdata.worldbank.org/index.php/catalog/ghdx/?page=1&ps=15&repo=ghdx>
- 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

- analysis for the global burden of disease study 2019. *Lancet Psychiatry* 2022;9:137–50.
- 3 GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *Lancet* 2020;396:1204–22.
 - 4 COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. n.d. Available: <https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide>
 - 5 Villarreal-Zegarra D, Reátegui-Rivera CM, Otazú-Alfaro S, et al. Melendez-Torres GJ: estimated impact of the COVID-19 pandemic on the prevalence and treatment of depressive symptoms in Peru: an interrupted time series analysis in 2014–2021. *Soc Psychiatry Psychiatr Epidemiol* 2023;58:1375–85.
 - 6 COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet* 2021;398:1700–1712. .
 - 7 Villarreal-Zegarra D, Segovia-Bacilio P, Paredes-Angeles R, et al. Provision of community mental health care before and during the COVID-19 pandemic: a time series analysis in Peru. *Int J Soc Psychiatry* 2023;207640231185026.
 - 8 Sagan A, McDavid D, Rajan S, et al. European Observatory policy Briefs. In: *Screening: When is it appropriate and how can we get it right?* Copenhagen (Denmark): European Observatory on Health Systems and Policies © World Health Organization 2020 (acting as the host organization for, and secretariat of, the European Observatory on Health Systems and Policies).
 - 9 Iragorri N, Spackman E. Assessing the value of screening tools: reviewing the challenges and opportunities of cost-effectiveness analysis. *Public Health Rev* 2018;39:17.
 - 10 Mulvaney-Day N, Marshall T, Downey Piscopo K, et al. Screening for behavioral health conditions in primary care settings: a systematic review of the literature. *J Gen Intern Med* 2018;33:335–46.
 - 11 Jiao B, Rosen Z, Bellanger M, et al. The cost-effectiveness of PHQ screening and collaborative care for depression in New York City. *PLoS One* 2017;12:e0184210.
 - 12 Ramírez-Bontá F, Vásquez-Vilchez R, Cabrera-Alva M, et al. Mental health data available in representative surveys conducted in latin America and the Caribbean countries: a scoping review [In press]. 2023.
 - 13 Haberer JE, Trabin T, Klinkman M. Furthering the reliable and valid measurement of mental health screening, diagnoses, treatment and outcomes through health information technology. *Gen Hosp Psychiatry* 2013;35:349–53.
 - 14 National Research Council, Division of Behavioral and Social Sciences and Education, Institute of Medicine, et al. *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities*. Washington: National Academies Press, 2009.
 - 15 Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. *JAMA* 1999;282:1737–44.
 - 16 Kroenke K, Strine TW, Spitzer RL, et al. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord* 2009;114:163–73.
 - 17 Kroenke K, Spitzer RL, Williams JBW. The patient health questionnaire-2: validity of a two-item depression screener. *Med Care* 2003;41:1284–92.
 - 18 Plummer F, Manea L, Trepel D, et al. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *Gen Hosp Psychiatry* 2016;39:24–31.
 - 19 Ali G-C, Ryan G, De Silva MJ. Validated screening tools for common mental disorders in low and middle income countries: a systematic review. *PLoS One* 2016;11:e0156939.
 - 20 Villarreal-Zegarra D, Copez-Lonzoy A, Bernabé-Ortiz A, et al. Valid group comparisons can be made with the patient health questionnaire (PHQ-9): a measurement invariance study across groups by demographic characteristics. *PLoS One* 2019;14:e0221717.
 - 21 Zhong Q-Y, Gelaye B, Zaslavsky AM, et al. Diagnostic validity of the generalized anxiety disorder - 7 (GAD-7) among pregnant women. *PLoS ONE* 2015;10:e0125096.
 - 22 Stochl J, Fried EI, Fritz J, et al. Measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment* 2022;29:355–66.
 - 23 Shevlin M, Butter S, McBride O, et al. Measurement invariance of the patient health questionnaire (PHQ-9) and generalized anxiety disorder scale (GAD-7) across four European countries during the COVID-19 pandemic. *BMC Psychiatry* 2022;22:154.
 - 24 Kroenke K, Wu J, Yu Z, et al. Patient health questionnaire anxiety and depression scale: initial validation in three clinical trials. *Psychosom Med* 2016;78:716–27.
 - 25 Urtasun M, Daray FM, Teti GL, et al. Validation and calibration of the patient health questionnaire (PHQ-9) in Argentina. *BMC Psychiatry* 2019;19:291.
 - 26 García-Campayo J, Zamorano E, Ruiz MA, et al. Cultural adaptation into Spanish of the generalized anxiety disorder-7 (GAD-7) scale as a screening tool. *Health Qual Life Outcomes* 2010;8:8.
 - 27 Sawaya H, Atoui M, Hamadeh A, et al. Adaptation and initial validation of the patient health questionnaire - 9 (PHQ-9) and the generalized anxiety disorder - 7 questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample. *Psychiatry Res* 2016;239:245–52.
 - 28 Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008;8:46.
 - 29 Woldetensay YK, Belachew T, Tesfaye M, et al. Validation of the patient health questionnaire (PHQ-9) as a screening tool for depression in pregnant women: Afaan Oromo version. *PLoS One* 2018;13:e0191782.
 - 30 Eack SM, Greeno CG, Lee BJ. Limitations of the patient health questionnaire in identifying anxiety and depression: many cases are undetected. *Res Soc Work Pract* 2006;16:625–31.
 - 31 Lambert SD, Clover K, Pallant JF, et al. Making sense of variations in prevalence estimates of depression in cancer: a co-calibration of commonly used depression scales using Rasch analysis. *J Natl Compr Canc Netw* 2015;13:1203–11.
 - 32 Liu S-I, Yeh Z-T, Huang H-C, et al. Validation of patient health questionnaire for depression screening among primary care patients in Taiwan. *Compr Psychiatry* 2011;52:96–101.
 - 33 Levis B, Benedetti A, Thombs BD, et al. Accuracy of patient health questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:1476.
 - 34 Neupane D, Levis B, Bhandari PM, et al. Selective cutoff reporting in studies of the accuracy of the patient health questionnaire-9 and Edinburgh postnatal depression scale: comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2021;30:e1873.
 - 35 Levis B, Sun Y, He C, et al. Accuracy of the PHQ-2 alone and in combination with the PHQ-9 for screening to detect major depression: systematic review and meta-analysis. *JAMA* 2020;323:2290–300.
 - 36 Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol Med* 2020;50:1368–80.
 - 37 Kroenke K, Spitzer RL, Williams JBW, et al. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen Hosp Psychiatry* 2010;32:345–59.
 - 38 Manea L, Boehnke JR, Gilbody S, et al. Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. *BMJ Open* 2017;7:e015247.
 - 39 Mitchell AJ, Yadegarfar M, Gill J, et al. Case finding and screening clinical utility of the patient health questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. *BJPsych Open* 2016;2:127–38.
 - 40 Mughal AY, Devadas J, Ardman E, et al. A systematic review of validated screening tools for anxiety disorders and PTSD in low to middle income countries. *BMC Psychiatry* 2020;20:338.
 - 41 Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799.
 - 42 Villarreal-Zegarra D, Cabrera-Alva M, Carrillo-Larco RM, et al. Trends in the prevalence and treatment of depressive symptoms in Peru: a population-based study. *BMJ Open* 2020;10:e036777.
 - 43 Hernández-Vásquez A, Vargas-Fernández R, Bendezu-Quispe G, et al. Depression in the Peruvian population and its associated factors: analysis of a national health survey. *J Affect Disord* 2020;273:291–7.
 - 44 Steel Z, Marnane C, Iranpour C, et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *Int J Epidemiol* 2014;43:476–93.
 - 45 Buderer NM. Statistical methodology: I. incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Acad Emerg Med* 1996;3:895–900.
 - 46 Huarcaya-Victoria J, Segura V, Cárdenas D, et al. Analysis of the care provided over a six-month period by the liaison psychiatry unit at a general hospital in Lima. *Revista Colombiana de Psiquiatría (English Ed)* 2022;51:105–12.



- 47 Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the patient health questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *General Hospital Psychiatry* 2015;37:67–75.
- 48 Zuihthoff NPA, Vergouwe Y, King M, *et al.* The patient health questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam Pract* 2010;11:98.
- 49 Shin C, Lee SH, Han KM, *et al.* Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. *Psychiatry Investig* 2019;16:300–5.
- 50 Caycho-Rodríguez T, Barboza-Palomino M, Ventura-León J, *et al.* Spanish translation and validation of a brief measure of anxiety by the COVID-19 in students of health sciences. *Ansiedad y Estrés* 2020;26:174–80.
- 51 Spitzer RL, Kroenke K, Williams JBW, *et al.* A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006;166:1092–7.
- 52 Franco-Jimenez RA, Nuñez-Magallanes A. Psychometric properties of the GAD-7, GAD-2, and GAD-mini in Peruvian college students. *Propósitos y Representaciones* 2022;10:e1437.
- 53 Kroenke K, Spitzer RL, Williams JBW, *et al.* Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Ann Intern Med* 2007;146:317–25.
- 54 Trevethan R. Specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front Public Health* 2017;5:307.
- 55 Ranganathan P, Aggarwal R. Understanding the properties of diagnostic tests - part 2: likelihood ratios. *Perspect Clin Res* 2018;9:99–102.
- 56 Suh Y. The performance of maximum likelihood and weighted least square mean and variance adjusted Estimators in testing differential item functioning with Nonnormal trait distributions. *Structural Equation Modeling: A Multidisciplinary Journal* 2015;22:568–80.
- 57 Holgado-Tello FP, Chacón-MoscOSO S, Barbero-García I, *et al.* Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Qual Quant* 2010;44:153–66.
- 58 Hair JF, Anderson RE, Tatham RL, *et al.* *Análisis multivariante*, 491. Prentice Hall Madrid, 1999.
- 59 McDonald RP. *Test theory: A unified treatment*. New York: Taylor & Francis Group, 1999.
- 60 Costantini L, Pasquarella C, Odone A, *et al.* Screening for depression in primary care with patient health questionnaire-9 (PHQ-9): a systematic review. *J Affect Disord* 2021;279:473–83.
- 61 Hartung TJ, Friedrich M, Johansen C, *et al.* The hospital anxiety and depression scale (HADS) and the 9-item patient health questionnaire (PHQ-9) as screening instruments for depression in patients with cancer. *Cancer* 2017;123:4236–43.
- 62 Sun Y, Kong Z, Song Y, *et al.* The validity and reliability of the PHQ-9 on screening of depression in neurology: a cross sectional study. *BMC Psychiatry* 2022;22:98.
- 63 Gholizadeh L, Shahmansouri N, Heydari M, *et al.* Assessment and detection of depression in patients with coronary artery disease: validation of the Persian version of the PHQ-9. *Contemp Nurse* 2019;55:185–94.
- 64 Husain N, Waheed W, Tomenson B, *et al.* The validation of personal health questionnaire amongst people of Pakistani family origin living in the United Kingdom. *J Affect Disord* 2007;97:261–4.
- 65 Ganguly S, Samanta M, Roy P, *et al.* Patient health questionnaire-9 as an effective tool for screening of depression among Indian adolescents. *J Adolesc Health* 2013;52:546–51.
- 66 Hanlon C, Medhin G, Selamu M, *et al.* Validity of brief screening questionnaires to detect depression in primary care in Ethiopia. *J Affect Disord* 2015;186:32–9.
- 67 Harry ML, Coley RY, Waring SC, *et al.* Evaluating the cross-cultural measurement invariance of the PHQ-9 between American Indian/Alaska native adults and diverse racial and ethnic groups. *J Affect Disord Rep* 2021;4:100121.
- 68 Razykov I, Ziegelstein RC, Whooley MA, *et al.* The PHQ-9 versus the PHQ-8—Is item 9 useful for assessing suicide risk in coronary artery disease patients. *J Psychosom Res* 2012;73:163–8.
- 69 Konkan R, Senormanci O, Guclu O, *et al.* Validity and reliability study for the Turkish adaptation of the generalized anxiety disorder-7 (GAD-7) scale. *Archives of Neuropsychiatry* 2013;50:53–8.
- 70 Gong Y, Zhou H, Zhang Y, *et al.* Validation of the 7-item generalized anxiety disorder scale (GAD-7) as a screening tool for anxiety among pregnant Chinese women. *J Affect Disord* 2021;282:98–103.
- 71 Snijders JTW, van den Oever W, Weerts ZZRM, *et al.* Examining the optimal cutoff values of HADS, PHQ-9 and GAD-7 as screening instruments for depression and anxiety in irritable bowel syndrome. *Neurogastroenterol Motil* 2021;33:e14161.
- 72 Manea L, Gilbody S, Hewitt C, *et al.* Identifying depression with the PHQ-2: a diagnostic meta-analysis. *J Affect Disord* 2016;203:382–95.
- 73 Luo Z, Li Y, Hou Y, *et al.* Adaptation of the two-item generalized anxiety disorder scale (GAD-2) to Chinese rural population: a validation study and meta-analysis. *Gen Hosp Psychiatry* 2019;60:50–6.
- 74 Giuliani M, Gorini A, Barbieri S, *et al.* Examination of the best cut-off points of PHQ-2 and GAD-2 for detecting depression and anxiety in Italian cardiovascular inpatients. *Psychol Health* 2021;36:1088–101.
- 75 Bentley KH, Sakurai H, Lowman KL, *et al.* Validation of brief screening measures for depression and anxiety in young people with substance use disorders. *J Affect Disord* 2021;282:1021–9.
- 76 Lamela D, Soreira C, Matos P, *et al.* Systematic review of the factor structure and measurement invariance of the patient health questionnaire-9 (PHQ-9) and validation of the Portuguese version in community settings. *J Affect Disord* 2020;276:220–33.
- 77 Schantz K, Reighard C, Aikens JE, *et al.* Screening for depression in Andean Latin America: factor structure and reliability of the CES-D short form and the PHQ-8 among Bolivian public hospital patients. *Int J Psychiatry Med* 2017;52:315–27.
- 78 Huarcaya-Victoria J, De-Lama-Morán R, Quiros M, *et al.* Propiedades Psicométricas del patient health questionnaire (PHQ-9) en Estudiantes de Medicina en Lima, Perú. *Rev Neuropsiquiatr* 2020;83:72–8.
- 79 Inagaki M, Ohtsuki T, Yonemoto N, *et al.* Validity of the patient health questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: a cross-sectional study. *General Hospital Psychiatry* 2013;35:592–7.
- 80 Le Hoang Ngoc T, Le M-AT, Nguyen HT, *et al.* Patient health questionnaire (PHQ-9): a depression screening tool for people with epilepsy in Vietnam. *Epilepsy & Behavior* 2021;125:108446.
- 81 Bhandari PM, Levis B, Neupane D, *et al.* Data-driven methods distort optimal cutoffs and accuracy estimates of depression screening tools: a simulation study using individual participant data. *J Clin Epidemiol* 2021;137:137–47.
- 82 Beatrice M-F, Carla M-C, Matilde L-M, *et al.* Clinical practice guideline for the screening and management of the mild depressive episode at the first level of care for the Peruvian social security (Essalud). *Acta Medica Peruana* 2020;37:536–47.
- 83 EsSalud. Análisis Ejecutivo Nacional de a Nivel Las Prestaciones de Salud 2016. In: *Lima: Gerencia central de planeamiento y presupuesto*. EsSalud, 2017.