

Validity of the Sinhala Version of the General Health Questionnaires Item 12 and 30: Using Different Sampling Strategies and Scoring Methods

H T C S Abeysena¹, P L Jayawardana¹, M U P K Peiris², A Rodrigo²

¹Professor, Department of Public Health, ²Senior Lecturer, Department of Psychiatry, Faculty of Medicine, University of Kelaniya, Ragama, Sri Lanka.

ABSTRACT

Objective: To determine the criterion validity of the Sinhala version of the General Health Questionnaire-12 (GHQ-12) and GHQ-30 employing different sampling designs and scoring methods.

Methods: This was a descriptive cross-sectional study including 374 patients. The GHQ-30 was completed by the participants using likert scale and then converted into standard score. The 'Clinical Examination' was done blindly to the GHQ score as the reference standard. Total study sample was considered as a representative sample taken consecutively. Case-reference design included 126 cases and 126 randomly selected controls based on reference standard. Test result-based designs included two groups of positive and negative GHQ, based on the optimal cut-off level. Cut-off levels were determined by using three criteria. In addition stratum specific likelihood (SSLR) ratio also considered.

Results: Applying consecutive sampling design, for the GHQ-12, the optimal cut-off levels were 9/10 using likert score and 2/3 using standard score and for the GHQ-30, 25/26 using likert score and 6/7 using standard score. The optimal cut-off level depends on the different sampling designs employed in addition to criteria for determining cut-off levels. The SSLR of >1 was useful for determining optimal cut-off level. Irrespective of the scoring methods, application of case-reference design tends to be overestimation of the specificity

with high threshold values and test result-based design tends to be overestimation of the sensitivity, compared to consecutive sampling. Using likert scoring method, the sensitivities were higher than standard scoring method.

Conclusions: The optimal cut-off levels depend on the sampling design and the scoring method employed and criteria to determine cut-off levels.


Key Words: Area under the curve, General Health Questionnaire, Likelihood ratio, Primary care, ROC, Validity.

*Correspondence to:

Prof Chrisantha Abeysena,
Professor,
Department of Public Health,
Faculty of Medicine,
University of Kelaniya, Ragama, Sri Lanka.

Article History:

Received: 30-07-2016, Revised: 07-08-2016, Accepted: 25-08-2016

Access this article online	
Website: www.ijmrp.com	Quick Response code 
DOI: 10.21276/ijmrp.2016.2.5.038	

INTRODUCTION

The General Health Questionnaire (GHQ) is widely used to detect short term minor psychiatric disorders. In spite of the original (GHQ) questionnaire, having 60-items, 30-, 28- and 12-item versions also have been derived from it subsequently. All items have a 4 point scoring system. There are four possible methods of scoring the GHQ. The standard scoring (0-0-1-1) and the likert scoring (0-1-2-3) methods are more popular. The GHQ yields an overall total score. The sensitivity and the specificity of the instrument depend on the cut-off level of the total score. The GHQ has been validated in different languages and cultures.^{1,2}

Validity is the extent to which a test measures what it claims to measure. The criterion validity refers to the accuracy of a tool when applied against a reference standard.³ With regards to the GHQ, validity indicates the extent to which the test scores accurately estimate an individual's current psychological state in

relation to the criterion. However, the cut-off scores required to achieving the optimum sensitivity and specificity varies considerably from one setting to another. Also the cut-off levels are varying according to the scoring methods.⁴ Further, the sensitivity and the specificity of a test for a given cut-off level depend on the study design employed. A descriptive cross sectional study using a consecutive or representative sampling is the most appropriate method.⁵ A case-referent design, the criteria for enrolment is based on presence or absence of the disease status. A test result-based design defines the study sample based on positive or negative of the test results.⁵

The Sinhala version of the GHQ-30 has been widely used for research purposes in Sri Lanka. Both Sinhala versions of GHQ-12 and GHQ-30 have been validated in primary care setting using standard scoring method.^{6,7} Thus, the objective of this study is to

determine the criterion validity of the Sinhala version of the GHQ-12 and GHQ-30 for screening minor psychiatric disorders, employing different sampling designs and scoring methods such as standard and Likert.

METHODS

This was a descriptive cross sectional study and details of methods have been published elsewhere.^{6,7} In briefly the study participants included both males and females between the ages of 18 and 75 years, who were also able to read and understand the Sinhala language. All patients (who consented to be in the study) were registered at the Colombo North Teaching Hospital (CNTH) for Out Patient Department (OPD) visits during the study period, and those eligible for inclusion, were recruited until the required sample size was reached. We could not visit to the OPD everyday due to unavoidable circumstances. For a visited day during a morning or afternoon session, around 10 to 15 patients/ day were recruited consecutively.

Sinhala version of the GHQ, which is a self-administered questionnaire, was rated on a four-point likert (0-1-2-3) and then converted into standard (0-0-1-1) scoring method. The minimum possible score using either likert or standard method was zero for both GHQ-12 and GHQ-30. The maximum possible scores using likert and standard scoring methods were 36 and 12 for the GHQ-12, and 90 and 30 for the GHQ-30 respectively. The psychiatrist assessment with regards to the presence or absence of depression/anxiety/social dysfunction was considered as the reference standard which was done blindly to the results of the GHQ score. Analysis was performed in three stages in relation to three sampling designs as follows.

Stage 1: We recruited 374 participants as described above. Of them there were 126 (33.7%) cases diagnosed and 248 (66.3%) controls without minor psychiatric disorders based on reference standard.

Stage 2: Case-reference design – The study sample was divided into two equal size groups according to the reference standard. Therefore 126 cases diagnosed as for the case group and another 126 were selected randomly from the remaining 248 patients without minor psychiatric disorders for the control group.

Sensitivity and specificity of the both instruments were determined for different cut-off values. The optimal cut-off level for the instrument was determined using the Receiver Operation Characteristic (ROC) curves, which were based on the trade-off between sensitivity and 1- specificity. Three criteria were used to determine optimal cut-off value. The first was lowest distance (d) between the point (0, 1) and any point on the ROC curve. The second was Youden index that maximizes the vertical distance from line of equality to the point of the ROC curve. It is the point on the ROC curve which is farthest from line of equality and which is equal to where sum of sensitivity and specificity is maximum. The third was related to the Youden index is the products of sensitivity and specificity. All criteria give equal weight to sensitivity and specificity and impose no ethical, cost, and no prevalence constraints. The ROC curves were constructed separately based on standard score as well as likert score for both GHQ-12 and GHQ-30 for the above two sampling strategies. The area under the curve (AUC) and its 95% confidence interval (CI) were calculated. The statistical analysis was performed using the statistical package 'SPSS Windows version 16'.

Stage 3: Test result-based design – The study sample was divided into two equal size groups according to the optimal cut-off level of the GHQ score determined by the stages 1 and 2 using consecutive sampling. Then equal number of patients with or without psychological morbidities was selected randomly from the remaining patients who scored below or above the optimal cut-off level under each scoring method.

The ethical clearance was obtained from the Ethics Review Committee, Faculty of Medicine, University of Kelaniya and details were reported in previous publications.⁶

RESULTS

The median (Inter quartile range) score of the GHQ-12 and GHQ-30 of the study sample were 1(4) and 3(7) according to the standard score and 9(8) and 22(15) according to the likert score respectively. The standard score distributions were positively skewed and likert scores more closer to a normal distribution.

The GHQ-12 and GHQ-30 scores were categorized into two groups at various cut-off points. Application of case-reference design compared to consecutive sampling design, irrespective of the scoring method and for a given cut-off level, specificity was slightly higher where the cut-off levels were higher. However, for the lower cut-off levels, the specificities were lower in case-reference sampling compared to the consecutive sampling. The sensitivities were not changed across the designs. (Table 1 to 4)

Determination of the optimal cut-off level for GHQ-12

The GHQ-12 was scored in likert scale, the optimal cut-off level of 9/10 was detected both for using consecutive sampling [sensitivity; 81.7% (95% CI 73.9% -88.0%) and specificity 67% (95% CI: 60.7% – 72.8%) and case-reference design [sensitivity; 81.7% (95% CI 73.9% -88.0%) and specificity 66% (95% CI: 56.9% – 78.1%)](by the second and third criteria). However, when determined the cut-off level based on the lowest distance between the point (0,1) and the point on the ROC curve(by the first criteria), it was 10/11 with a sensitivity of 71.4% (95% CI: 62.7% – 79.1%), a specificity of 74.6% (95% CI: 66.1% – 81.9%) for the case-reference design. (Table 1 and 5) The AUC of the ROC curve was 0.80 (95% CI: 75% –85%) for the both sampling designs.

One hundred and eighty five (49.5%) were GHQ-12 positive based on likert score at the cut-off level was 9/10. Therefore another 185 were selected for the control group from the remaining 198 without having psychosocial morbidities. Applying test result-based design the sensitivity, the specificity and the positive likelihood ratio (LR) for minor psychiatric disorders were 84.4% (95% CI: 76.7% – 90.3%), 66.9% (95% CI: 60.7% – 72.7%) and 2.55 (95% CI: 2.1 – 3.1) respectively. (Table 5)

The optimal cut-off level for the GHQ-12 using standard score was 1/2 when applying consecutive sampling (by the first criteria) [sensitivity; 74% (95% CI 65.2% -81.2%) and specificity 71% (95% CI: 65.0% – 76.5%)]. According to the second and third criteria of determining cut-off level the optimal was 2/3 with a sensitivity of 64.3% (95% CI: 55.3% – 72.6%), a specificity of 81.8% (95% CI: 76.5% – 86.4%) for consecutive sampling. For case-reference design it was 1/2 by all the methods with a sensitivity of 74% (95% CI 65.2% -81.2%) and specificity of 73% (95% CI: 64.4% – 80.5%). However for the cut-off level of 1/2 and 2/3, the Youden index was the same.

Where the cut-off level was 1/2 and 2/3 of the GHQ-12 based on standard score, 126(33.7%) and 161(40.9%) were GHQ-12

positive respectively. Therefore another 126 and 161 were selected without having psychosocial morbidities from the remaining 209 and 213 respectively for the test-result based design. According to this design (with cut-off level of 1/2) sensitivity, specificity and positive predictive value for the minor psychiatric disorders were 77.5% (95% CI: 69.0% – 84.6%), 65.7% (95% CI: 58.9% – 72.1%) and 56.5% (95% CI: 48.4 – 64.0) respectively. As shown in Table 5, change the cut-off level to 2/3 sensitivity, specificity and positive predictive value have been improved.

Changing the scoring method from standard score to likert score of the GHQ-12, the sensitivity has been increased from 74% (by criteria 1) or 64.3% (by criteria 2 and 3) to 81.7% and specificity decreased from 71% (by criteria 1) or 82% (by criteria 2 and 3) to 67% when applying consecutive sampling strategy for the specified optimal cut-off levels. The corresponding sensitivity and specificity, applying case-referral design were 74% to 81.7% (by criteria 1 and 3) or 71.4% (by criteria 2) and 73% to 66% (by criteria 1 and 3) or 74.6% (by criteria 2). (Table 1 and 2)

Determination of the optimal cut-off level for GHQ-30

Three optimal cut-off levels were found according to the criterion used. According to the second criteria, the optimal cut-off level was 21/22 for the GHQ-30 using likert score with a sensitivity of 83.3% (95% CI: 75.7% – 89.4%), a specificity of 64.5% (95% CI: 58.2% – 70.5%) when applying consecutive sampling. According to the first criteria it was 24/25 with a sensitivity of 73.8% (95% CI: 65.2% – 78.0%), a specificity of 72.6% (95% CI: 66.6% – 78.0%). According to the third criteria it was 25/26 with a sensitivity of 69.8% (95% CI: 61.0 – 77.7), a specificity of 77% (95% CI: 71.3% – 82.1%). However, the optimal cut-off level was 25/26 (by all three methods) with a sensitivity of 69.8% (95% CI: 61.0% – 77.7%), a specificity of 77% (95% CI: 68.6% – 84.0%) when applying case-reference design. (Table 3) The AUC was 0.80 (0.75 – 0.85) for both sampling designs.

The cut-off level of 21/22, 24/25 and 25/26 of the GHQ-30 based on likert scoring method 193(51.6%), 161(43%) and 145(38.8%) were GHQ-30 positive. According to the GHQ-30, there were 193 (51.6%) with and 181(48.4%) without psychosocial morbidities

based on likert score at the cut-off level of 21/22. Therefore out of 193, 181 patients with and another 181 without psychosocial morbidities were selected for the analysis. Another 161 and 145 patients without having psychosocial morbidities were selected for the control groups at 24/24 and 25/26 cut-off levels respectively. Application of test results-based design at the cut-off level of 21/22, the sensitivity, the specificity and the positive predictive value for the minor psychiatric disorders were 82.6% (95% CI: 74.7% – 88.9%), 64.4% (95% CI: 60.0% – 72.3%) and 55.2% (95% CI: 47.7 – 62.6) respectively. (Table 5) Change the cut-off level to 25/26 all diagnostic indicators have been changed.

The optimal cut-off level was 5/6 (by first and third criteria) for the GHQ-30 according to standard scoring method with a sensitivity of 67.5% (95% CI: 58.5% – 73.5%), a specificity of 80% (95% CI: 74.7% – 85.0%) when applying consecutive sampling. According to the youden index it was 6/7 with a sensitivity of 64.3% (95% CI: 55.3% – 72.6%), a specificity of 83.9% (95% CI: 78.7% – 88.2%). However, the optimal cut-off level was 6/7 (by second and third criteria) with a sensitivity of 64.3% (95% CI: 55.3% – 72.6%), a specificity of 86.5% (95% CI: 79.3% – 91.9%) when applying case-reference design. According to the first criteria (d) it was 5/6 with a sensitivity of 67.5% (95% CI: 58.5% – 75.5%), a specificity of 81.7% (95% CI: 73.9% – 80.1%). (Table 4) The corresponding AUCs of the ROCs were 79% (95% CI: 74% – 84%) and 80% (0.74 – 0.86) respectively. The cut-off level was 5/6 and 6/7 based on standard score, 134(35.8%) and 121(32.3%) respectively were GHQ-30 positive. Therefore another 134 and 121 patients without having psychosocial morbidities were randomly selected for the control groups respectively. Application of test result-based design at 5/6 level, the sensitivity, the specificity and the positive predictive value for minor psychiatric disorders were 79.4% (95% CI: 70.5% – 86.6%), 69.6% (95% CI: 61.8% – 76.6%) and 63.4 (95% CI: 54.7 – 71.6) respectively. As shown in Table 5, change the cut-off levels to 6/7 all diagnostic indicators were improved.

Changing the scoring method from the standard to likert score for the GHQ-30, the sensitivities were increased and the specificities decreased irrespective of applying consecutive sampling or case-reference design. (Table 3 and 4)

Table 1: Sensitivity, Specificity and criteria used for determining cut-off values of GHQ- 12 using different sampling designs, according to various cut-off values based on Likert score

GHQ Score	Consecutive sampling Reference Standard					Case-reference sampling Reference Standard				
	Sn	Sp	d	Y	Sn * Sp	Sn	Sp	d	Y	Sn * Sp
≥6	93.7	28.6	0.717	1.223	0.268	93.7	28.6	0.717	1.223	0.268
≥7	91.3	40.3	0.603	1.316	0.367	91.3	39.7	0.609	1.31	0.362
≥8	87.3	48.8	0.527	1.361	0.426	87.3	47.6	0.539	1.349	0.415
≥9	84.1	57.3	0.456	1.414	0.482	84.1	54.0	0.486	1.381	0.454
≥10	81.7	66.9	0.378	1.486	0.547	81.7	65.9	0.387	1.476	0.538
≥11	71.4	74.2	0.385	1.456	0.529	71.4	74.6	0.382	1.460	0.533
≥12	62.7	81.9	0.416	1.446	0.513	62.7	81.7	0.415	1.444	0.512
≥13	54.0	85.5	0.482	1.395	0.462	54.0	88.9	0.473	1.429	0.480
≥14	50.8	87.9	0.506	1.387	0.446	50.8	88.9	0.504	1.397	0.452
≥15	43.7	90.3	0.571	1.34	0.394	43.7	90.5	0.571	1.342	0.395

Sn – Sensitivity, Sp – Specificity, d – distance between the point (0, 1) and any point on the ROC curve, Y – Youden index

Table 2: Sensitivity, Specificity and criteria used for determining cut-off values of GHQ-12 using different sampling designs, according to various cut-off values based on standard score

GHQ Score	Consecutive sampling Reference Standard					Case-reference sampling Reference Standard				
	Sn	Sp	+		Sn * Sp	Sn	Sp	+		Sn * Sp
			(n=126)	(n=248)				(n=126)	(n=126)	
≥1	88.1	46.4	0.549	1.345	0.409	88.1	43.7	0.575	1.318	0.385
≥2	73.8	71.0	0.391	1.448	0.524	73.8	73.0	0.376	1.468	0.539
≥3	64.3	82.0	0.399	1.463	0.527	64.3	82.5	0.397	1.468	0.530
≥4	57.1	86.7	0.449	1.438	0.495	57.1	88.1	0.445	1.452	0.503
≥5	41.3	92.3	0.592	1.336	0.381	41.3	94.4	0.589	1.357	0.390
≥6	36.5	95.6	0.636	1.321	0.349	36.5	96.8	0.636	1.333	0.353

Sn – Sensitivity, Sp – Specificity, d – distance between the point (0, 1) and any point on the ROC curve, Y – Youden index,

Table 3: Sensitivity, Specificity and criteria used for determining cut-off values of GHQ-30 using different sampling designs, according to various cut-off values based on Likert score

GHQ Score	(Consecutive sampling) Reference Standard					(Case-reference sampling) Reference Standard				
	Sn	Sp	+		Sn * Sp	Sn	Sp	+		Sn * Sp
			(n=126)	(n=248)				(n=126)	(n=126)	
≥15	93.7	28.6	0.716	1.223	0.268	93.7	26.2	0.741	1.199	0.245
≥16	93.7	33.1	0.672	1.268	0.310	93.7	29.4	0.709	1.231	0.275
≥17	93.7	37.5	0.628	1.312	0.351	93.7	35.7	0.646	1.294	0.334
≥18	92.1	43.1	0.574	1.352	0.397	92.1	42.1	0.584	1.342	0.388
≥19	88.1	49.2	0.522	1.373	0.433	88.1	47.6	0.537	1.357	0.419
≥20	85.7	54.8	0.474	1.405	0.469	85.7	52.4	0.497	1.381	0.449
≥21	85.7	60.9	0.416	1.466	0.522	85.7	60.3	0.421	1.460	0.516
≥22	83.3	64.5	0.392	1.478	0.537	83.3	62.7	0.408	1.460	0.522
≥23	81.0	66.1	0.388	1.471	0.535	81.0	65.1	0.397	1.461	0.527
≥24	77.0	68.5	0.390	1.455	0.527	77.0	68.3	0.391	1.453	0.526
≥25	73.8	72.6	0.379	1.464	0.536	73.8	72.2	0.382	1.460	0.533
≥26	69.8	77.0	0.380	1.468	0.537	69.8	77.0	0.380	1.468	0.537
≥27	66.7	77.8	0.400	1.445	0.519	66.7	78.6	0.396	1.453	0.524
≥28	61.9	81.9	0.422	1.438	0.507	61.9	81.7	0.423	1.436	0.506

Sn – Sensitivity, Sp – Specificity, d – distance between the point (0, 1) and any point on the ROC curve, Y – Youden index

Table 4: Sensitivity, Specificity and criteria used for determining cut-off values of GHQ-30 using different sampling designs, according to various cut-off values based on standard score

GHQ Score	(Consecutive sampling) Reference Standard					(Case-reference sampling) Reference Standard				
	Sn	Sp	+		Sn * Sp	Sn	Sp	+		Sn * Sp
			(n=126)	(n=248)				(n=126)	(n=126)	
≥1	92.1	32.7	0.677	1.248	0.301	92.1	31.0	0.694	1.231	0.285
≥2	86.5	46.8	0.549	1.333	0.405	86.5	43.7	0.579	1.302	0.378
≥3	82.5	55.2	0.481	1.377	0.455	82.5	55.6	0.477	1.381	0.459
≥4	77.0	65.7	0.413	1.427	0.506	77.0	65.9	0.411	1.429	0.507
≥5	71.4	74.2	0.385	1.456	0.530	71.4	73.8	0.388	1.452	0.527
≥6	67.5	80.2	0.380	1.477	0.541	67.5	81.7	0.373	1.492	0.551
≥7	64.3	83.9	0.392	1.482	0.539	64.3	86.5	0.382	1.508	0.556
≥8	58.7	86.7	0.434	1.454	0.509	58.7	88.9	0.427	1.476	0.522
≥9	52.4	89.5	0.487	1.419	0.469	52.4	92.1	0.482	1.445	0.483
≥10	47.6	90.3	0.533	1.379	0.430	47.6	92.9	0.529	1.405	0.442
≥11	42.9	92.3	0.576	1.352	0.396	42.9	95.2	0.573	1.381	0.408
≥12	36.5	94.4	0.637	1.309	0.344	36.5	96.0	0.636	1.325	0.350
≥13	34.9	95.2	0.653	1.301	0.332	34.9	97.6	0.651	1.325	0.341
≥14	32.5	97.6	0.675	1.301	0.317	32.5	98.4	0.675	1.309	0.319

Sn – Sensitivity, Sp – Specificity, d – distance between the point (0, 1) and any point on the ROC curve, Y – Youden index

Table 5: Sensitivity, Specificity, PV, LR and Area under the curve of GHQ-12 & GHQ-30 by different sampling designs & scoring methods at the optimal cut off values

GHQ	Scoring Method	Sampling Strategy	Optimal cut-off	Sensitivity (95% CI)	Specificity (95% CI)	PV+ (95% CI)	PV- (95% CI)	LR+ (95% CI)	LR- (95% CI)	AUC (95% CI)	
GHQ 12	Likert	Consecutive	9/10	81.7 (73.9-88.1)	66.9 (60.7-72.8)	55.7 (48.2-63.0)	87.7 (82.3-92.1)	2.47 (2.0-3.0)	0.27 (0.19-0.40)	0.80 (0.75-0.85)	
		Case-reference	9/10	81.7 (73.9-88.1)	65.9 (56.9-74.1)	70.5 (62.4-77.8)	78.3 (69.2-85.7)	2.49 (1.8-3.1)	0.28 (0.19-0.41)	0.80 (0.75-0.85)	
		Case-reference based	10/11	71.4 (62.7-79.1)	74.6 (66.1-81.9)	73.8 (65.0-81.3)	72.3 (63.8-79.8)	2.8 (2.0-3.9)	0.38 (0.29-0.51)	0.80 (0.75-0.85)	
		Test result-based	9/10	84.4 (76.7-90.3)	67.0 (60.7-72.7)	55.7 (48.2-63.0)	89.7 (84.4-93.7)	2.55 (2.1-3.1)	0.23 (0.15-0.35)	-	
	Standard	Consecutive	1/2	74.0 (65.2-81.2)	71.0 (64.9-76.5)	56.5 (48.4-64.0)	84.2 (78.5-88.9)	2.54 (2.0-3.2)	0.37 (0.27-0.5)	0.79 (0.74-0.84)	
		Consecutive	2/3	64.3 (55.3-72.6)	81.8 (76.5-86.4)	64.3 (55.3-72.6)	81.8 (76.5-86.4)	3.54 (2.6-4.8)	0.44 (0.34-0.6)	0.79 (0.74-0.84)	
		Case-reference	1/2	74.0 (65.2-81.2)	73.0 (64.4-80.5)	73.2 (64.6-80.7)	73.6 (65.0-81.0)	2.74 (2.0-3.7)	0.36 (0.26-0.49)	0.79 (0.74-0.85)	
		Test result-based	1/2	77.5 (69.0-84.6)	65.7 (58.9-72.1)	56.5 (48.4-64.0)	83.6 (77.0-89.0)	2.26 (1.83-2.8)	0.34 (0.24-0.48)	-	
	GHQ 30	Likert	Consecutive	21/22	78.6 (69.5-86.1)	69.8 (61.7-77.0)	64.3 (55.3-72.6)	82.5 (74.8-88.7)	2.6 (2.0-3.4)	0.31 (0.21-0.45)	-
			Consecutive	24/25	83.3 (75.6-89.4)	64.5 (58.2-70.5)	54.4 (47.0-61.6)	88.4 (82.8-92.7)	2.35 (1.95-2.8)	0.25 (0.17-0.39)	0.80 (0.75-0.85)
			Consecutive	24/25	73.8 (65.2-78.0)	72.6 (66.6-78.0)	57.8 (49.7-65.5)	84.5 (78.9-89.1)	2.7 (2.1-3.4)	0.36 (0.27-0.50)	0.80 (0.75-0.85)
			Consecutive	25/26	69.8 (61.0-77.7)	77.0 (71.3-82.1)	60.7 (52.2-68.7)	83.4 (77.9-88.0)	3.0 (2.35-3.9)	0.39 (0.3-0.52)	0.80 (0.75-0.85)
Standard		Case-reference	21/22	82.6 (74.7-88.9)	64.4 (60.0-72.3)	55.2 (47.7-62.6)	88.4 (82.8-92.7)	2.46 (2.0-3.0)	0.26 (0.18-0.39)	-	
		Test result-based	24/25	78.8 (70.3-85.8)	66.7 (59.7-73.1)	57.8 (49.7-65.5)	84.5 (77.9-89.7)	2.36 (1.9-2.9)	0.32 (0.2-0.46)	-	
		Test result-based	25/26	77.9 (69.1-85.1)	67.8 (60.4-74.6)	60.7 (52.2-68.7)	82.8 (75.6-88.5)	2.42 (1.9-3.1)	0.33 (0.23-0.47)	-	
		Consecutive	5/6	67.5 (58.5-75.5)	80.2 (74.7-85.0)	63.4 (54.7-71.6)	82.9 (77.5-87.4)	3.4 (2.6-4.5)	0.41 (0.31-0.53)	0.79 (0.74-0.84)	
Standard		Consecutive	6/7	64.3 (55.3-72.6)	83.9 (78.7-88.2)	66.9 (57.8-75.2)	82.2 (76.9-86.7)	4.0 (2.9-5.4)	0.43 (0.33-0.54)	0.79 (0.74-0.84)	
		Case-reference	5/6	67.5 (58.5-75.5)	81.7 (73.9-80.1)	78.7 (69.8-86.0)	71.5 (63.4-78.7)	3.7 (2.5-5.4)	0.40 (0.30-0.52)	0.80 (0.74-0.86)	
		Case-reference	6/7	64.3 (55.3-72.6)	86.5 (79.3-91.9)	82.6 (73.7-89.5)	70.8 (62.9-77.8)	4.76 (3.0-7.5)	0.41 (0.32-0.53)	0.80 (0.74-0.86)	
		Test result-based	5/6	79.4 (70.5-86.6)	66.9 (61.8-76.5)	63.4 (54.7-71.6)	83.6 (76.2-89.4)	2.6 (2.0-3.3)	0.3 (0.2-0.43)	-	
Test result-based	6/7	81.0 (71.9-88.1)	71.8 (63.7-79.0)	66.9 (57.8-75.2)	84.3 (76.6-90.3)	2.9 (2.2-3.8)	0.26 (0.17-0.40)	-			

Positive Predictive Value; PV+, Negative Predictive Value; PV-, Likelihood Ratio Positive; LR+, Likelihood Ratio Negative; LR-, Area Under the Curve; AUC

Table 6: Multilevel likelihood ratios for GHQ-12 and GHQ- 30 based on likert score using consecutive sampling design

GHQ-12 score	Likelihood Ratio	GHQ-30 score	Likelihood Ratio
5	0.44	18	0.67
6	0.20	19	0.43
7	0.47	20	0.00
8	0.38	21	0.62
9	0.25	22	1.50
10	1.41	23	1.67
11	1.13	24	0.80
12	2.42	25	0.91
13	1.13	26	4.00
14	2.96	27	1.20
≥15	3.18	≥28	3.41

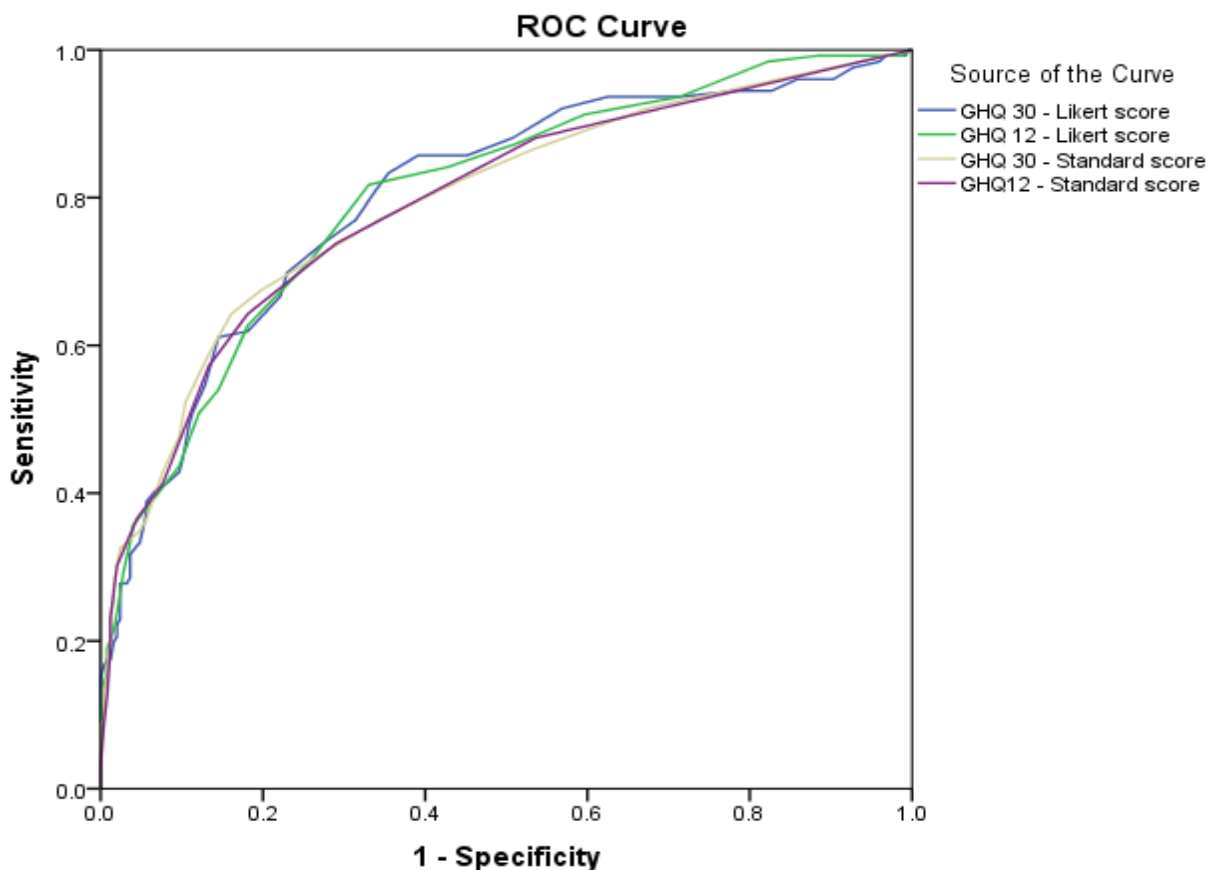


Figure 1: ROC curves for GHQ-12 and GHQ-30 using standard and likert scores in consecutive sampling design

The positive predictive values were higher for case-reference design compared to consecutive sampling. (Table 5) The sensitivities were higher and the specificities lower when applied test-result based design compared to consecutive sampling except for GHQ-30 based on likert score. The likelihood ratios and AUCs were not changed significantly across the sampling design or scoring methods. (Table 5 and Figure 1) The stratum specific likelihood ratio was >1 at the cut-off level of 9/10 for the GHQ-12 and 21/22 and 25/26 for the GHQ-30 based on likert score. However, the stratum 24 and 25, it were <1. (Table 6) Based on the standard scoring method, the cut-off levels were of 2/3 for the GHQ-12 and 6/7 for the GHQ-30 based on the stratum specific likelihood ratio of >1. The case-reference design also showed the agreement of the cut-off levels gained with consecutive sampling based on stratum specific likelihood ratio of >1.

DISCUSSION

Using likert scoring method, the optimal cut-off level based on all three criteria was 9/10 of the GHQ-12, when applied consecutive sampling. The same cut-off level was found as optimal for the case-reference sampling except by the criteria 1. The cut-off level of 9/10 was compatible with the stratum specific LR of >1. For the GHQ-30 it was 21/22 by criteria 2, 24/25 by criteria 1 and 25/26 by criteria 3. For the case-reference design it was 25/26 by the all criteria applied. The cut-off level of 25/26 was compatible with the stratum specific LR of >1. Using the standard scoring method, the optimal cut-off level was 1/2 by criteria 1 and 2/3 by criteria 2 and 3 for the GHQ-12. For the case-reference design, it was 1/2 by criteria 1 and 3. By criteria 2 both 1/2 and 2/3 cut-off levels were optimal. The cut-off level of 2/3 was compatible with the stratum specific LR of >1. For the GHQ-30, it was 5/6 by criteria 1 and 3,

and 6/7 by criteria 2 when applied consecutive sampling design. For the case-reference design it was 5/6 by criteria 1 and 6/7 by the criteria 2 and 3. The cut-off level of 6/7 was compatible with the stratum specific LR of >1 . A study design enrolling consecutive patients considered to be the best method for assessing screening or diagnostic tests. Across the sampling designs stratum specific LR method gave consistent results. The agreement was more between the rule of stratum specific LR >1 criteria and the criteria 2 (Youden index). Youden index is more commonly used criterion because this index reflects the intension to maximize the correct classification rate and is easy to calculate. One Australian study⁴ reported that for the GHQ-12, a cut-off level of 10/11 with a sensitivity of 72.4% and a specificity of 77.4% for a representative sample of Australians using likert score and 0/1 with a sensitivity of 75.4% and a specificity of 70% using standard scoring method. In the present study if the cut-off level was changed to 10/11, the corresponding sensitivity and specificity were 71.4% and 74.2. Irrespective of the criteria for determining the optimal cut-off level our study revealed that it was 9/10 for consecutive sampling. According to the WHO study⁸ the optimal cut-off was 11/12 for GHQ-12 using likert scoring method in a representative sample with a sensitivity of 78.9% and a specificity of 77.4%. This is consistent with the recommendations of the authors of the original GHQ as well.⁹ El-Rufai reported that the optimal cut-off level was 12/13 for GHQ-12 with a sensitivity of 83% and a specificity of 80% in a random sample of United Arab Emirates.¹⁰ In contrast to our study El-Rufaie also reported very high optimal cut off value of 31/32 for the GHQ-30 using likert scoring method with a sensitivity of 93% and a specificity of 86%. Above all studies^{4,8,9,10} had not reported the criteria of determining the optimal cut-off level or stratum specific LRs.

Application of case-reference design tends to be overestimation of the specificity with high threshold values. However for lower threshold values, specificity were lower than corresponding values applied by consecutive sampling irrespective of the scoring methods. In contrast applying test result-based design tends to be overestimation of the sensitivity and underestimation of the specificity compared to consecutive sampling irrespective of the scoring method. However this is not so obvious when applying likert scoring method as where the percentage of positive GHQ-12 and GHQ-30 above the threshold (9/10 and 21/22) of the total population was almost equal to 50%. Therefore, the extent of the bias depends on the prevalence of the condition according to the test (GHQ) results.

Irrespective of the sampling strategies and scoring methods, positive predictive values were higher when applying case-reference design than the other two designs for both GHQ-12 and GHQ-30. This was due to high prevalence of minor psychiatric disorders which was artificially fixed as 50% in the case-reference design.

The likert scoring method tends to be higher sensitivities than using standard scoring method regardless of the version of the GHQ. This is in contrast to other studies which have revealed that there was no difference between the scoring methods.^{4,8} Considering the recommendations by the User's Guide on General Health Questionnaire,⁹ priority is to be given to sensitivity in preference to specificity for the purposes of case detection. However using likert scoring might be cumbersome in a clinical setting especially in GHQ-30 than the standard scoring method.

Our results also showed that the likelihood ratios and AUC which is considered a summary measure of the ability of the GHQ to discriminate between cases and non-cases, have not been change despite applying different designs and scoring methods. However, the cut-off levels and relevant sensitivities and specificities depend on different sampling designs and scoring methods employed, the AUC for all occasions were >0.79 . The AUC which assess overall performance of the test is not dependent on the prevalence of the disease. Equal AUCs of the two tests does not necessarily mean that both the curves are identical, it may cross each other. According to the determined optimal cut-off levels using likert and standard scoring methods, different prevalence of minor psychiatric disorders were found, which range from 35.8% to 51.6%. This inconsistency is further deteriorating the validity of the GHQ 12 or GHQ-30 when applying for another setting. Therefore we recommend larger comparative study to assess the validity of GHQ in terms of optimal cut-off levels, sensitivities, specificities, stratum specific LRs and using different scoring methods.

Using different criteria to determine the optimal cut-off level gave different results. One study reported that youden index (criteria 2) is the best, given equal weigh to sensitivity and specificity.¹¹ We found that stratum specific LRs of >1 is consistent across sampling strategies and scoring methods than the youden index.

For the case-reference designs, we selected all the cases diagnosed by the reference standard. Therefore the sensitivities were not changed. The controls were the representative sample of the controls in the study setting. Therefore the case-reference design employed was considered as a nested case control design in diagnostic area. The nested case control designs are more suitable for when the reference standard is invasive or the new test is costly without compromising the validity.^{12,13} In contrast case-reference design which is analogous to case control studies is more prone to selection and spectrum bias which leads to overestimation of diagnostic indicators.^{5,14} Although the recruitment of our study participants were based on consecutive sampling, which was unlikely to affect the representativeness of the sample as the steps have taken to minimise selection and spectrum bias.

Furthermore, as the psychiatric assessment was carried out blind to the GHQ status, one could assume that the threat to the internal validity of the study would have been minimal. The estimate of prevalence according to the reference standard is dependent on the criteria for 'casesness' that was used for validation. The various methods give different results because of the diversity in defining 'casesness'.¹⁵ The strength of our study was assessing validity of the GHQ using different sampling strategies, scoring methods and criteria to determine optimal cut-off levels using the same primary data, therefore comparability is more obvious.

Our study showed that the validity indicators of the GHQ vary depending on the different sampling designs employed and criteria to determine optimal cut-off levels in addition to the scoring method. Considering all the facts above we conclude that for the GHQ-12, the optimal cut-off values were of 9/10 using likert score and 2/3 using standard score and for the GHQ-30, 25/26 using likert score and 6/7 using standard score. Stratum specific likelihood ratio is a more stable indicator to determine the optimal cut-off value.

ACKNOWLEDGEMENTS

This study was funded by the University of Kelaniya Sri Lanka. We are grateful to the Research and Publication Committee of the University.

REFERENCES

1. Huppert FA, Walters DE, Day NE, Elliott BJ. The factor structure of the General Health Questionnaire (GHQ-30). A reliability study on 6317 community residents. *The British Journal of Psychiatry* 1989;155: 178-185.
2. Jakob KS, Bhugra D, Mann H. The validation of the 12-item General Health Questionnaire among ethnic Indian women living in the United Kingdom. *Psychological Medicine*. 1997;27:1215-1217.
3. Abramson JH, Abramson ZH. *Survey methods in community medicine*. London: Churchill Livingstone; 1990.
4. Donath S. The validity of the 12-item General Health Questionnaire in Australia: a comparison between three scoring methods. *Aust N Z J Psychiatry* 2001; 35:231-235.
5. Knottnerus JA (Edi). *The Evidence Base of Clinical Diagnosis*. London:BMJ Publishing Group;2002.
6. Abeysena C, Peiris U, Jayawardana P,Rodrigo A. Validation of the Sinhala version of 30-item General Health Questionnaires. *International Journal of Collaborative Research on Internal Medicine & Public Health*. 2012;4(7): 1373-1381.
7. Abeysena C, Peiris U, Jayawardana P,Rodrigo A. Validation of the Sinhala version of 12-item General Health Questionnaires. *Journal of Postgraduate Institution of Medicine*. 2014;1(1), E8:1–E8:7. DOI: <http://doi.org/10.4038/jpgim.7859>
8. Goldberg DP et al. Validity of two versions of the GHQ in the WHO study of Mental illness in general health care. *Psycho Med*1997; 27:191 – 197.
9. Goldberg D, Williams P. *A user's guide to the General Health Questionnaire*, Windsor: NFER-Nelson;1991
10. El-rufaie OEF, Daradkeh TW. Validation of the Arabic versions of the Thirty- and Twelve-Item General Health Questionnaires in Primary Care Patients. *British Journal of Psychiatry*. 1996;169:662-664.
11. Perkins NJ, Schisterman EF. The Inconsistency of "Optimal" Cut- points Using Two ROC Based Criteria. *Am J Epidemiol*. 2006;163(7): 670–675.
12. Baker SG, Kramer BS, Srivastava S: Markers for early detection of cancer: statistical guidelines for nested case-control studies. *BMC Med Res Approachol* 2002, 2:4.
13. Pepe MS, Feng Z, Janes H, Bossuyt PM , Potter JD. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. *J Natl Cancer Inst* 2008;100: 1432 – 1438
14. Zhou X-H, Obuchowski NA, Mcclish DK. *Statistical Methods in Diagnostic Medicine*. New York: John Wiley & Sons; 2002.
15. Goldberg DP, Oldhinkel T and Ormel J: Why GHQ threshold varies from one place to another. *Psychol Med* 1998, 28:915-921.

Source of Support: Nil. **Conflict of Interest:** None Declared.

Copyright: © the author(s) and publisher. IJMRP is an official publication of Ibn Sina Academy of Medieval Medicine & Sciences, registered in 2001 under Indian Trusts Act, 1882. This is an open access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cite this article as: H T C S Abeysena, P L Jayawardana, M U P K Peiris, A Rodrigo. Validity of the Sinhala Version of the General Health Questionnaires Item 12 and 30: Using Different Sampling Strategies and Scoring Methods. *Int J Med Res Prof*. 2016; 2(5):180-87.