Dynamic shortening of psychometric questionnaires: A focus on PHQ-9

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 248 (2024) 30–37

Procedia Computer Science

www.elsevier.com/locate/procedia

International Society for Research on Internet Interventions 12th Scientific Meeting (ISRII 12)

Dynamic Shortening of Psychometric Questionnaires: A Focus on PHQ-9

Zahraa Abdulhussein[a,*], Pepijn van de Ven[a,b]

[a] *Department of Electronic and Computer Engineering, University of Limerick, Limerick V94 T9PX, Ireland*
[b] *Health Research Institute, University of Limerick, V94 T9PX, Ireland*

Abstract

The Patient Health Questionnaire-9 (PHQ-9) is a widely utilized assessment tool designed to screen for and measure the severity of depression. Recent research has explored the viability of shorter versions of the questionnaire, such as the PHQ-DEP-4 and PHQ-2, particularly for screening depressed individuals before clinical trials. However, these short tests use a fixed number of questions. This study investigated whether there is merit in posing a dynamic number of questions in screening individuals for depressive symptomatology. To address this, our study focuses on creating a short and dynamic version of the PHQ-9. Our approach utilizes dynamic thresholds tailored to individual responses to predict depression. Optimal thresholds are determined by minimizing the number of questions while simultaneously maximizing the Youden index. Results indicate the Dynamic PHQ-9 model demonstrated superior discrimination ability achieving highest Youden index, sensitivity, and specificity.

Keywords: Depression; PHQ-9; PHQ-DEP-4; PHQ-2; Dynamic tests; Adaptive tests;

1. Introduction

The Patient Health Questionnaire-9 (PHQ-9) is a widely utilised assessment tool designed to screen for and measure the severity of depression. It consists of nine questions as shown in Table 1, each with response options ranging from 0 to 3. This test offers a comprehensive view of an individual's depressive symptoms. A response of 0 signifies the absence of symptoms ("Not at all"), while 1 indicates symptoms experienced "Several days", 2 denotes "More than half the days", and 3 suggests symptoms occurring "Nearly every day". This structured format allows healthcare professionals to systematically assess intensity of depressive symptoms experienced by patients. The sum of responses per patient yields the PHQ score, a numerical representation of depression severity. Scores of 5, 10, 15, and 20 denote varying degrees of depression: mild, moderate, moderately severe, and severe, respectively [1]. For binary

* Corresponding author.
E-mail address: Zahraa.Abdulhussein@ul.ie

1877-0509 © 2024 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2024.10.360

classification purposes, numerous studies advocate for utilizing a cutoff PHQ-9 score ≥ 10 to differentiate between depressed and non-depressed individuals [1][2][3], which is also the criterion applied in our study.

Table 1. PHQ-9 Questions.

Over the last 2 weeks, how often have you been bothered by the following problems:
phq1: Little interest in doing things?
phq2: Feeling down, depressed, or hopeless?
phq3: Trouble sleeping or sleeping too much?
phq4: Feeling tired or having little energy?
phq5: Poor appetite or overeating?
phq6: Feeling bad about yourself?
phq7: Trouble concentrating on things?
phq8: Moving or speaking slowly or too fast?
phq9: Thought you would be better off dead?

While the PHQ-9 offers valuable insights into the severity of depression, recent research has explored the viability of shorter versions of the questionnaire, such as the PHQ-DEP-4 and PHQ-2 [4], particularly for screening depressed individuals before clinical trials. These short tests use a fixed number of questions. The present study investigated whether there is merit in posing a dynamic number of questions to improve the performance of the screening instrument and/or reduce the question burden.

To achieve the objectives of this study, we adopted a dynamic approach to shortening the PHQ-9 questionnaire. The methodology involves iterating through the PHQ-9 questions in their standard order and making classification decisions based on cumulative scores. The model evaluates whether to continue asking questions or to stop based on predefined thresholds, aiming to classify individuals as either depressed or non-depressed. A grid search was employed to determine the optimal thresholds, with performance assessed using the Youden index, a metric chosen for its ability to balance sensitivity and specificity. The model was tested on a large dataset from the National Health and Nutrition Examination Survey (NHANES), covering the years 2005 to 2020, and its performance was compared to that of fixed-length questionnaires such as the PHQ-2 and PHQ-DEP-4.

2. Related Works

The PHQ-DEP-4, a condensed version of the PHQ-9 comprising four questions aimed at predicting depression, was developed using the optimal test assembly method and item response theory [4]. This abbreviated test consists of questions phq1, phq2, phq6, and phq8. In the original study [4], the authors used data from 7,850 participants across 20 primary studies conducted in multiple countries with the PHQ-9 completed in English. Using a cutoff ≥ 10 for the PHQ-9 and evaluating various cutoffs for the PHQ-DEP-4, the authors found the optimal cutoff for the latter to be ≥ 4. Comparative analysis revealed that the PHQ-DEP-4 exhibited a sensitivity of 0.788 and a specificity of 0.837, while the PHQ-9 demonstrated a sensitivity of 0.761 and a specificity of 0.866. These findings suggest that the PHQ-DEP-4 is a valid short form, preserving essential information compared to the full-length PHQ-9, with minimal loss.

A meta-analysis conducted in 2022 evaluated the diagnostic accuracy of the PHQ-DEP-4 compared to the PHQ-9 across a large database of 34,698 participants from 75 primary studies worldwide [5]. Utilizing a cutoff ≥ 10 for the PHQ-9 and various cutoffs for PHQ-DEP-4, the study found the optimal cutoff for the PHQ-DEP-4 to be ≥ 4, consistent with previous research. Across different diagnostic interviews (semi-structured, fully structured, and MINI), the PHQ-DEP-4 demonstrated sensitivities of 0.88, 0.68, and 0.80, respectively, and specificities of 0.79, 0.85, and 0.83, respectively. While the PHQ-9 demonstrated sensitivities of 0.88, 0.64, and 0.73, respectively, and specificities of 0.85, 0.89, and 0.89, respectively. The study concluded that the sensitivity of the PHQ-DEP-4 was non-inferior to that of the PHQ-9, with slightly lower specificity.

The PHQ-2 comprises two questions, PHQ1 and PHQ2, designed to assess the cardinal symptoms of depression. Extensive research underscores the validity of the PHQ-2 in identifying depression, particularly in flagging individuals for further evaluation [6][7][8]. A 2020 meta-analysis [9] assessed the accuracy of the PHQ-2 alone and in combination with the PHQ-9 for detecting major depression in a sample of 44,318 participants, including 4,572 with major depression. Sensitivity for a PHQ-2 score ≥ 2 followed by a PHQ-9 score ≥ 10 (0.82) did not significantly differ from sensitivity for PHQ-9 scores ≥ 10 alone (0.86), while the specificity of the combined approach was marginally higher (0.87 vs. 0.85). Utilizing the combined PHQ-2 and PHQ-9 approach was estimated to reduce the number of participants needing to complete the full PHQ-9 by 57%. Investigation into various cutoff scores for the PHQ-2 has been conducted to optimize diagnostic accuracy. Common cutoffs include scores of ≥ 2 or ≥ 3, with higher cutoffs typically associated with increased specificity but reduced sensitivity [10][11][12][13][14][15][16].

An open issue within the current literature on abbreviated versions of the PHQ-9, such as the PHQ-DEP-4 and PHQ-2, is that these tests employ a fixed number of questions, which may not fully capture the nuances of individual depressive symptomatology. While these shorter forms have demonstrated validity, they do not adapt to individual responses, potentially limiting their accuracy. Our study addresses this gap by introducing a dynamic approach that adjusts the number of questions based on cumulative scores, aiming to maximize diagnostic accuracy while minimizing the number of questions needed. This contributes to the field by providing a more tailored screening tool that can better accommodate individual differences in symptom presentation, thereby potentially improving both the precision of depression diagnosis and the efficiency of screening processes.

3. Methods

Following on from previous endeavours to reduce the PHQ-9 questionnaire length, as described in section 2, this paper proposes a dynamic approach to the reduction of the number of questions asked before a decision is made. In line with common terminology in machine learning, we will refer to the decision algorithm as the model. The approach iterates through the PHQ-9 questions in their normal order and allows for three decisions to be made after every question asked based on the cumulative score up to that point: if it exceeds a higher cumulative threshold, the respondent is deemed depressed; if it is below or equal to a lower cumulative threshold, the respondent is deemed to be non-depressed; for any other scores, the next question will be asked and the score will again be assessed against the thresholds. This section delves further into this methodology and also details the origin and characteristics of the data used to illustrate the model's performance.

3.1. Data

The dataset originates from the CDC surveys conducted in the United States, known as the National Health and Nutrition Examination Survey (NHANES)[17]. These surveys involve the administration of multiple questionnaires and tests to survey participants on an annual basis. Specifically, the PHQ-9 data utilized in our analysis spans the years 2005 to 2020. The dataset comprises 49,461 records encompassing demographics information, PHQ-9 items, and PHQ-10 item reporting on. To focus solely on PHQ-9 items, we excluded the PHQ-10 item from our analysis. Subsequently, records with responses indicating refusal to answer (coded as 7) or uncertainty (coded as 9) were removed, alongside any records containing missing values. This process resulted in a dataset containing 44,749 records devoid of missing values. Table 2 shows the characteristics of the dataset.

3.2. The Dynamic Questionnaire Model

The model operates by sequentially processing each PHQ item and applying distinct lower and upper cumulative thresholds to the sum of responses for the PHQ items processed up to that time. If the sum of an individual's responses is less than or equal to the lower cumulative threshold, they are classified as non-depressed, and no further questions are asked. If the sum exceeds the upper cumulative threshold, they are classified as depressed, and no further questions are asked. Those who are not classified as either, are further evaluated until all questions are answered. Any respondent who is not ruled in or out after all questions have been answered, will be labelled as depressed.

Fig 1 illustrates this process in a flow chart, showcasing the model's dynamic nature. The flowchart utilizes several key notations to represent the operational process of the model. Firstly, 'PHQ_items' denote the responses provided to each question in the PHQ test. The parameter 'N' represents the total number of PHQ items being processed sequentially. 'upper_th' and 'lower_th' signify the upper and lower cumulative thresholds, respectively. Additionally, 'pos_st' and 'neg_st' denote tune-able parameters for positive and negative cases, respectively. These parameters dictate the starting point from which the classification process begins, thereby optimizing model performance.

Table 2. Characteristics of NHANES data.

Characteristics	Number(%)	Characteristics	Number(%)
Sample	44,535	Female	22,652 (0.50)
Age, years	48.0±18.6	Marital Status	
Ethnicity		Married	17574 (0.39)
Mexican American	6853 (0.15)	Widowed	2695 (<0.1)
Other Hispanic	4278 (<0.1)	Divorced	3751 (<0.1)
Non-Hispanic White	18114 (0.40)	Separated	1159 (<0.1)
Non-Hispanic Black	10107 (0.22)	Never married	1159 (<0.1)
Other Race	5183 (0.11)	Living with partner	2842 (<0.1)
Education Level – Youth		Refused	19 (<0.1)
9th Grade	50 (<0.1)	Don't know	1 (<0.1)
10th Grade	107 (<0.1)	Missing	9892 (0.22)
11th Grade	416 (<0.1)	Family Income - Annual	
12th Grade, No Diploma	166 (<0.1)	$0 to $4,999	169 (<0.1)
High School Graduate	735 (<0.1)	$5,000 to $9,999	254 (<0.1)
GED or Equivalent	47 (<0.1)	$10,000 to $14,999	392 (<0.1)
More than high school	567 (<0.1)	$15,000 to $19,999	364 (<0.1)
Less Than 9th Grade	42 (<0.1)	$20,000 to $24,999	373 (<0.1)
Missing	42404 (0.95)	$25,000 to $34,999	611 (0.1)
Education Level – Adults 20+		$35,000 to $44,999	478 (0.1)
Less Than 9th Grade	3994 (<0.1)	$45,000 to $54,999	440 (<0.1)
9-11th Grade (Includes 12th grade with no diploma)	5670 (0.12)	$55,000 to $64,999	271 (<0.1)
High School Grad/GED or Equivalent	9799 (0.22)	$65,000 to $74,999	254 (<0.1)
Some College or AA degree	12747 (0.28)	$75,000 and over	980 (<0.1)
College Graduate or above	9774 (0.21)	Over $20,000	67 (<0.1)
Refused	10 (<0.1)	Under $20,000	27 (<0.1)
Don't Know	22 (<0.1)	Refused	27 (<0.1)
Missing	2519 (<0.1)	Don't know	56 (<0.1)
PHQ-9 ≥ 10	3,935 (<0.1)	Missing	39772 (0.89)

Within the iterative process, the model begins with the response of the first question (phq1). Subsequent iterations gradually incorporate additional responses, extending to 'phq2', 'phq3', and so forth until reaching the chosen maximum number of items (N). If no prediction is made after processing all items, the individual is classified as depressed, terminating the process. The decisions node represented by the second and forth diamond shapes determine the order of checking, based on whether positive or negative cases are assessed first. This design accommodates discrepancies in starting points for positive and negative cases.

For example, when 'pos_st' is set to 1 and 'neg_st' to 2, it checks for positive cases using the upper threshold ('upper_th') for 'phq1'. If positive, the process concludes; otherwise, it advances to the next question. Subsequently, in the second iteration, the model checks for negative cases using the lower threshold for both 'phq1' and 'phq2.' If not negative, it then proceeds to assess for positive cases using the upper threshold, continuing this iterative process until a determination is reached.

The lower cumulative threshold is a list of 9 items where each item represents a sum of responses for each of the preceding PHQ questions. For example, consider a cumulative threshold list such as [0, 1, 4, 4, 5, 6, 7, 8, 9]. In this list, the first value (0) represents the threshold for phq1, the second value (1) represents the cumulative threshold for the phq1 and phq2 items, the third value (4) represents the cumulative threshold for phq1, phq2, and phq3 items, and so forth. So, for this example and assuming the algorithm applies the threshold from the first question onwards, a person responding Not at all (0) to question 1, would have been deemed non-depressed after this first question. A person who responded Several Days (1) to question 1 and Not at all to question 2 would be deemed non-depressed after answering these two questions. For the upper cumulative threshold, the threshold list [2, 4, 6, 8, 10, 12, 14, 16, 18] was employed. This threshold aims to identify extreme cases of depression early in the process.

To determine the optimal lower threshold, a grid search method was employed. Initially, all possible combinations of cumulative thresholds were generated, each consisting of a list of nine values. Subsequently, the model's performance was assessed using each threshold, with the optimal one selected based on the highest Youden index

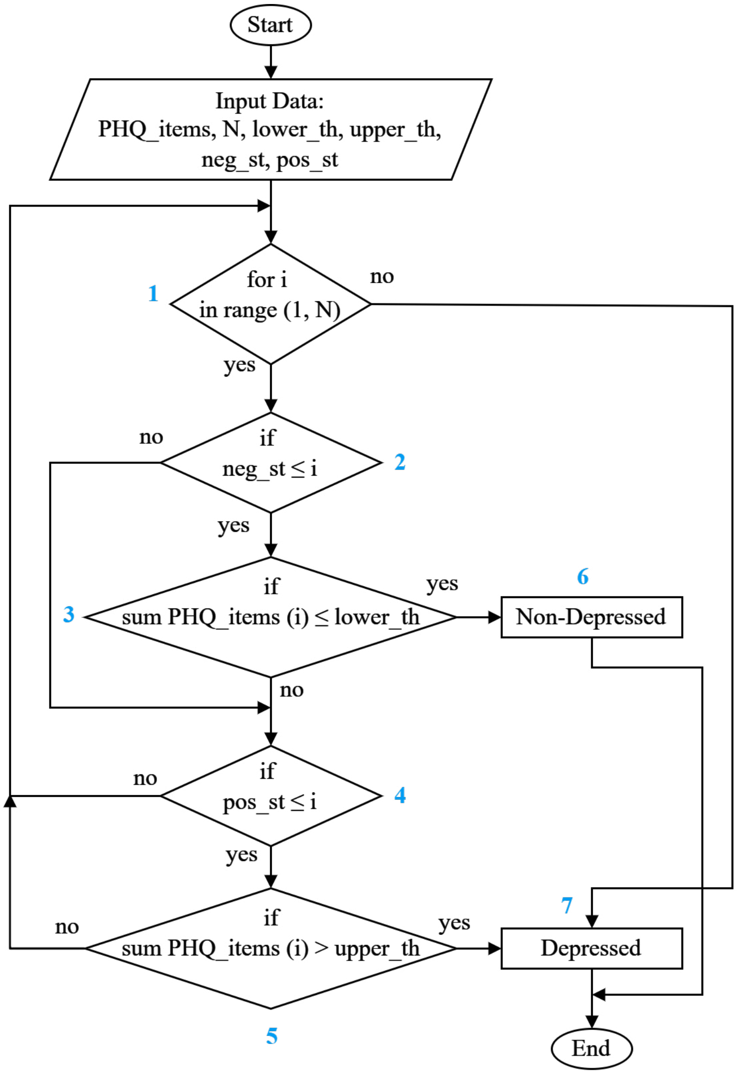

Fig. 1. Dynamic Questionnaire Model Flow Chart

and the lowest average number of questions. The Youden index was chosen as it provides a balanced measure of a test's performance by accounting for both sensitivity and specificity, making it particularly useful in evaluating the effectiveness of diagnostic tests. This index helps ensure that the selected threshold not only maximizes the model's ability to correctly identify true positives (sensitivity) but also minimizes the likelihood of false positives (specificity). Eq. 1 shows the formula for calculating the Youden index. Results of the dynamic approach were compared with the PHQ-2 and PHQ-DEP-4. To allow these reference standards to yield the best results possible, the optimal thresholds for discriminating between positive and negative cases for these instruments were determined. For the PHQ-DEP-4 model, the optimal threshold is determined to be 4, and for the PHQ-2 model, the optimal threshold is 2.

$$Youden\ Index = Sensitivity + Specificity - 1 \tag{1}$$

4. Results

The dynamic model was applied to the items in PHQ-9, PHQ-DEP-4, and PHQ-2, resulting in three dynamic models. Subsequently, these dynamic models were compared with the original forms of these instruments, labelled as the static PHQ-DEP-4 and PHQ-2, as shown in Table 3. Regarding the starting point parameter for both negative and positive cases, experimentation revealed that initiating the classification of positive cases from the first question (phq1) onwards and classifying negative cases from the second question (phq2) onwards yielded the highest performance and lowest average number of questions across all models. The performance metrics of these models are detailed in Table 3.

Table 3. Results.

Model's Name	Youden Index	Sensitivity	Specificity	Accuracy	Average # of questions	Cumulative Thresholds for negative cases
Dynamic PHQ-9	0.95	0.97	0.98	0.98	3.50	[0, 0, 0, 0, 2, 2, 4, 6, 9]
Static PHQ-DEP-4	0.82	0.85	0.96	0.95	4	
Dynamic PHQ-DEP-4	0.79	0.88	0.91	0.91	2.27	[0, 1, 1, 2]
Static PHQ-2	0.77	0.90	0.86	0.87	2	
Dynamic PHQ-2	0.69	0.97	0.72	0.74	1.96	[0, 0]

The Dynamic PHQ-9 model demonstrated superior discrimination ability, achieving a Youden Index of 0.95, along with notable sensitivity (0.97) and specificity (0.98). This model attained the highest accuracy of 0.98 with an average of 3.50 questions. Fig.2 expresses these numbers in terms of correct and incorrect classifications.

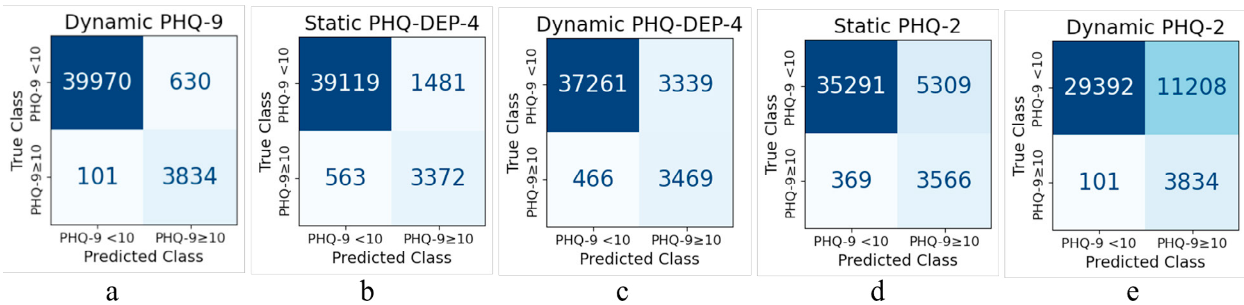

Fig. 2. Confusion Matrices

Fig. 2a shows that the dynamic PHQ-9 model misclassifies 630 individuals as depressed, which is fewer than in any of the other models. Shortening the PHQ-2 dynamically has a large negative effect on false positives as demonstrated by the 11,208 false positives in Fig. 2.e. This outcome arises because the model classifies individuals as non-depressed only if they respond with a zero to both 'phq1' and 'phq2', as indicated by the cumulative threshold list ([0, 0]). Consequently, all other individuals are deemed depressed by the end of the evaluation. Looking at false negatives, Fig. 2a and 2e show that the dynamic PHQ-9 and dynamic PHQ-2 models misclassifies 101 individuals as non-depressed, whereas the static PHQ-DEP-4 (Fig. 2b) and static PHQ-2 (Fig. 2d) model misclassifies 563 and 369 depressed individuals respectively.

Furthermore, the Dynamic model with PHQ-9 maintains high performance with an average of 3.50 questions. While the Dynamic and Static PHQ-DEP-4 models also perform well, requiring an average of 2.27 and 4 questions respectively, they do not reach the same level of performance as the Dynamic PHQ-9. Similarly, the Dynamic and Static PHQ-2 models, with the lowest average number of questions at 1.96 and 2 respectively, do not achieve the same performance level as the Dynamic PHQ-9.

Despite the promising performance of the dynamic PHQ-9 model, several limitations should be acknowledged. One significant limitation is the model's reliance on the cumulative thresholds derived from a single dataset, which may affect its generalizability to other populations. Additionally, while the model effectively reduces the number of questions asked, there is a risk of misclassification, particularly for individuals with atypical response patterns that do not align well with the predefined thresholds. Furthermore, the computational intensity required to determine the optimal thresholds through grid search poses challenges for scalability, particularly in real-time clinical settings.

5. Conclusion

In summary, the dynamic model introduced here offers a strategy for depression screening using the PHQ assessment, but this approach is expected to be applicable to any questionnaire with ordinal responses. The model processes each PHQ item sequentially, applying distinct lower and upper cumulative thresholds to classify respondents as either depressed or non-depressed. The determination of optimal thresholds is achieved through grid search, aiming to maximize the Youden index while minimizing the average questions asked. The dynamic model was evaluated against the PHQ-DEP-4 and PHQ-2 models, with the dynamic PHQ-9 model demonstrating the highest performance. Specifically, the dynamic PHQ-9 model achieved a Youden index of 0.95, sensitivity of 0.97, and specificity of 0.98, underscoring its potential as a valuable screening tool for depression. The dynamic model not only enhances classification accuracy but also reduces the number of questions required for evaluation. This streamlined approach efficiently identifies depressed and non-depressed individuals early in the assessment process.

6. Future Work

Further investigation is essential to evaluate the broader applicability of the dynamic model, particularly in using clinician diagnoses as a benchmark rather than relying solely on PHQ-9 scores. This shift could potentially enhance the robustness and reliability of the model in real-world clinical settings. Additionally, the current study utilized only one dataset from the U.S.; extending this research to include datasets from other regions would provide a more comprehensive understanding of the model's limitations and generalizability.

Moreover, the present dynamic model focuses primarily on reducing the length of the test for each individual. Future work could explore the possibility of dynamically reordering PHQ items instead of maintaining a static order, which may further improve the model's efficiency. Lastly, the method used to generate dynamic thresholds involved evaluating all possible threshold combinations, leading to significant computational demands. Future research could benefit from employing statistical methods, such as probabilistic measures, to estimate optimal thresholds more efficiently, thereby reducing computation time and cost.

Acknowledgements

This work has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI 18/CRT/6049. Also, we would like to thank Darragh Glavin for the thoughtful comments and recommendations on this research.

References

[1] Kroenke, K., Spitzer, and R. L. (2002). "The PHQ-9: A new depression diagnostic and severity measure." *Psychiatric Annals*, **32** (9): 509-515.

[2] Levis B, Benedetti A, and Thombs B D. (2019) "Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis" *BMJ* **365** :l1476

[3] Kroenke K, Spitzer RL, Williams JB. (2001) "The PHQ-9: validity of a brief depression severity measure." *J Gen Intern Med* **16** :606-13.

[4] Ishihara M, Harel D, Levis B, et al. (2019)"Shortening self-report mental health symptom measures through optimal test assembly methods: Development and validation of the Patient Health Questionnaire-Depression-4." *Depress Anxiety.* **36** (1):82-92.

[5] Daphna Harel, Brooke Levis, Ying Sun, et al. (2022) "External validation of a shortened screening tool using individual participant data meta-analysis: A case study of the Patient Health Questionnaire-Dep-4" *Methods*, **204**:300-311.

[6] Kroenke, Kurt MD, Spitzer, Robert L. MD, Williams Janet B. W. DSW. (2003) "The Patient Health Questionnaire-2: Validity of a Two-Item Depression Screener." *Medical Care*, **41** (11):1284-1292.

[7] Bernd Löwe, Kurt Kroenke, and Kerstin Gräfe. (2005) "Detecting and monitoring depression with a two-item questionnaire (PHQ-2)" *Journal of Psychosomatic Research*, **58** (2):163-171.

[8] Li, C., Friedman, B., Conwell, Y. and Fiscella, K. (2007), "Validity of the Patient Health Questionnaire 2 (PHQ-2) in Identifying Major Depression in Older People." *Journal of the American Geriatrics Society*, **55**: 596-602.

[9] Levis B, Sun Y, He C, et al. (2020) "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*. **323** (22):2290-2300.

[10] Giuliani, M., Gorini, A., Barbieri, S., Veglia, F., and Tremoli, E. (2021). "Examination of the best cut-off points of PHQ-2 and GAD-2 for detecting depression and anxiety in Italian cardiovascular inpatients." *Psychology and Health,* **36** (9):1088–1101.

[11] Ghazisaeedi, M., Mahmoodi, H., Arpaci, I. et al. (2022)“Validity, Reliability, and Optimal Cut-off Scores of the WHO-5, PHQ-9, and PHQ-2 to Screen Depression Among University Students in Iran.” *Int J Ment Health Addiction* **20**:1824–1833.

[12] Megan V. Smith, Nathan Gotman, Haiqun Lin, Kimberly A. (2010) “Yonkers, Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women?”, *General Hospital Psychiatry*, **32**:544-548.

[13] Gigantesco, A., Palumbo, G., Cena, L., Camoni, L., Trainini, A., Stefana, A. and Mirabella, F. (2022), “A Brief Depression Screening Tool for Perinatal Clinical Practice: The Performance of the PHQ-2 Compared with the PHQ-9.” *J Midwifery Womens Health.* **67**: 586-592.

[14] Bizu Gelaye, Ijala Wilson, Hanna Y. Berhane, Negussie Deyessa, Yonas Bahretibeb, Dawit Wondimagegn, Teshome Shibre Kelkile, Yemane Berhane, Jesse R. Fann, Michelle A. Williams, (2016) “Diagnostic validity of the Patient Health Questionnaire-2 (PHQ-2) among Ethiopian adults”, *Comprehensive Psychiatry*, **70**:216-221.

[15] Orlando Scoppetta, Carlos Arturo Cassiani-Miranda, Karen Nicolle Arocha-Díaz, Diego Fernando Cabanzo-Arenas, Adalberto Campo-Arias. (2021) “Validity of the patient health questionnaire-2 (PHQ-2) for the detection of depression in primary care in Colombia”, *Journal of Affective Disorders*, **278**:576-582.

[16] Chagas, M. H. N., Crippa, J. A. S., Loureiro, S. R., Hallak, J. E. C., Meneses-Gaya, C. de, Machado-de-Sousa, J. P., … Tumas, V. (2011). “Validity of the PHQ-2 for the screening of major depression in Parkinson’s disease: Two questions and one important answer.” *Aging and Mental Health*, **15** (7):838–843.

[17] Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centres for Disease Control and Prevention, 2005-2020, https://wwwn.cdc.gov/nchs/nhanes/Default.aspx