Sex-specific Longitudinal Comparison of CES-D and PHQ-9 Depression Scales. A Concordance Analysis using data from the population-based Heinz Nixdorf Recall Study

Miriam Engel (miriam.engel@uk-essen.de)
Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen
https://orcid.org/0000-0003-0161-0858

Markus Peternel
Centre for Urban Epidemiology (CUE), Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen

Karl-Heinz Jöckel
Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen

Sara Schramm
Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen

Uta Slomiany
Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen

Susanne Moebus
Centre for Urban Epidemiology, Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen

Research article

Keywords: Depression, Center for Epidemiologic Studies Depression Scale, Patient Health Questionnaire, Concordance Analysis, Comparison

Posted Date: December 4th, 2019

DOI: https://doi.org/10.21203/rs.2.18099/v1

Abstract

Background: To measure depressive symptoms (DS) in population-based studies, there are two the well-established questionnaires: the Center for Epidemiologic Studies Depression Scale (CES-D) and the Patient Health Questionnaire-9 (PHQ-9). So far, comparisons between both instruments have only been performed using cross-sectional data, and in specific patient groups. Furthermore, comparisons for population-based studies are missing as well as sex-specific analyses. The aim is to evaluate the psychometric properties and concordance of the longitudinal results of CES-D and PHQ-9 in the German population-based Heinz Nixdorf Recall (HNR) study.

Methods: We used data of n=3,084 participants (48.8% men, mean age: 66.8 years). CES-D and PHQ-9 were assessed in the 8th (t8) and 9th (t9) annual postal follow up within two years via questionnaires. DS were defined as CES-D score ≥ 17 and PHQ-9 score ≥ 10. Internal consistency reliability, convergent validity, and agreement between PHQ-9 and CES-D were assessed using respectively Cronbach's alpha, Pearson's correlation, and Cohen's kappa. To analyse DS differences between t8 and t9 we used McNemar's test. Sex-stratified results are presented.

Results: The prevalence of DS at t8 was higher for CES-D (7.8%) than for PHQ-9 (4.4%). The prevalence slightly increased for CES-D (8.1%), as well as for PHQ-9 to 4.5% at t9. Internal consistency of the PHQ-9 and CES-D was good at both times (Cronbach's alpha: CES-D t8 & t9: 0.89, PHQ-9 t8: 0.84; t9: 0.85). Cohen's kappa of agreement between CES-D and PHQ-9 was moderate at both time points (t8: k=0.57; 95% CI: 0.51, 0.63; t9; k=0.58; 95% CI: 0.52, 0.64). For both instruments, the McNemar's test revealed no differences in the proportion of depressives between t8 and t9. We could not observe sex-specific differences of psychometric characteristics, whereas the agreement between the CES-D and PHQ-9 performed slightly higher in men than in women (t8: k=0.61, resp. k=0.55).

Conclusion: The results do not support the superiority of one of the scales. Both scales perform well in this population-based cohort in both sexes and longitudinal approach. The decision, which scale to use may depend on characteristics other than the accuracy of the scale, including feasibility and clinical usability.

Introduction

The mental disorder depression is one of the most common mental illnesses in the world and more than 300 million people suffer from depression worldwide (1, 2). Symptoms include feelings of sadness, hopelessness, loss of pleasure, appetite, weight, self-esteem or self-confidence as well as sleep and concentration disorders that exceed a certain duration, persistence, and intensity. Often depression is associated with co-morbidity. Due to the high prevalence of depression and its comorbidity, the assessment of depressive symptoms is part of the standard program of many epidemiological studies (3, 4).

For an individual medical diagnosis of depressive disorders, a structured and standardized diagnostic interview such as the internationally recognized Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders, 5th ed. (DSM-V) is usually used (5). These interviews are characterized by a comprehensive and time-consuming question structure and therefore generally unfeasible in large population-based studies (6). In epidemiological studies, a large number of self-administrative instruments in the form of psychometric

personality tests are used to detect Major Depressive Disorders (MDD). MDD are determined by the presence of relevant multiple depressive symptoms within a defined time. Self-administered questionnaires are easy to use and cost-effective. Of the many instruments available, the “Center for Epidemiologic Studies Depression Scale (CES-D Scale)" (7) and the “Patient Health Questionnaire (PHQ)" are the most widely used ones in epidemiological studies (8).

It has been shown that these questionnaires performed well as a screening instrument in comparison with the reference standards, a clinical interview (9). However, to make an evidence-based decision for one of the instruments for measuring depressive symptoms in epidemiological studies, it is necessary to precisely analyse the differences and similarities of the instruments as well as to evaluate the psychometric strengths and weaknesses. To compare results between studies, it is also important to know whether these instruments are interchangeably or to what extent the results differ systematically.

The psychometric performance of the CES-D and the PHQ–9 has already been compared in specific patient groups like people with diabetes type 2, multiple sclerosis or systemic sclerosis (10–13). Taken together, the literature shows no relevant differences between the two instruments in terms of reliability, validity as well as differences in specific subpopulations. It is only noted that PHQ–9 is more specific in indicating depression (14), while CES-D detects various parts of depression and some of the other emotions associated with serious illness (15).

However, to best to our knowledge, there are no comparisons between these instruments within the general population. In addition, there are no comparisons with longitudinal data measuring the development of depressive symptoms. Moreover, as it is known that there are large differences in the prevalence and development of depressive symptoms between women and men (16–19) it is useful to examine the psychometric characteristics separately by sex.

We have chosen the CES-D and the PHQ–9 because they are commonly used instruments to assess depression symptomatology in epidemiological studies and they measure different aspects of depression. The PHQ–9 corresponds to major depression and was designed for clinical use. The CES-D has been developed for large epidemiological studies and investigates a variety of aspects of depression.

This study aims to compare the psychometric properties and the concordance of the CES-D and the PHQ–9 in an elderly population using the data set of the German longitudinal, population-based Heinz Nixdorf Recall (HNR) Study.

Methods

Study population

The design of the HNR Study has been described in detail elsewhere (20, 21). Briefly, for baseline examination, 4,814 women and men (49.8% men) aged 45 to 75 years were recruited between 2000 and 2003 from mandatory citizen registries of three large cities (Bochum, Essen, and Mülheim an der Ruhr, Germany). For follow up, participants were invited further two times to the study centre in Essen every five years. In addition, a yearly questionnaire-based postal follow-up was conducted between these examinations. To perform a

longitudinal comparison between the CES-D and the PHQ–9 we used the 8th (hereafter t8) and 9th (t9) follow up year in which both instruments were applied concurrently.

For our analysis, we included 3,084 participants (1,580 women, 1,504 men) who completed both the PHQ–9 and CES-D questionnaires at both time points.

The HNR study is confirmed by the local ethics committees and all participants gave written informed consent before participation.

Assessment of depressive symptoms

We assessed depressive symptoms using the validated tools “Center for Epidemiologic Studies Depression Scale” (CES-D) (7) and the PHQ–9, a sub-module of the “Patient Health Questionnaire (PHQ) “. Both instruments are structured self-administrative scales, in which the participants answer predetermined answer options. The CES-D was always measured immediately before the PHQ–9. The participants were also asked if they are currently in therapy for depression or are taking medication against it. These variables are only used for sensitivity analysis.

Center for Epidemiologic Studies Depression Scale (CES-D)

The CES-D is often used in epidemiological studies in the general population. The CES-D asks for the presence and frequency of symptoms and emotional states in the week before the interview including depressive mood, feelings of guilt or worthlessness, sleep disorders and self-doubt (22, 23). The CES-D is considered as an indicator of depression and is highly correlated with a clinical diagnosis of depression (24).

In the HNR Study, a short version of the CES-D with 15 questions was applied. Answers are given on a 4-point Likert-scale ranging from "less than one day" (0 point) to "5–7 days" (3 points). We calculated a sum score ranged from 0 to 45 points with a higher score indicating more and/or more frequent depressive symptoms. Positively formulated items were coded backward and an average value was calculated over all 15 items. For up to three missing answers, the item value was replaced by the mean value of the answered questions. In the HNR Study, a cutoff point of ≥17 was defined as depression (25, 26).

Patient Health Questionnaire (PHQ)

The Patient Health Questionnaire developed by Spitzer, Kroenke, and Williams (27) is used to screen for MDD with items corresponding to the symptoms identified in the Diagnostic and Statistical Manual (5). In the HNR Study, we used the nine items subscale PHQ–9 which consists of the actual nine criteria of the DSM-V diagnosis for depressive disorders. The PHQ–9 is widely applied in medical settings (8). Participants were asked about the frequency of the emergence of nine different problems or depression criteria over the last two weeks.

There are four possible answers: not at all (0 point), several days (1 point), more than half the days (2 points), and nearly every day (3 points). Total PHQ–9 score ranges from 0 to 27 and are categorized as “none or

minimum" (0–4), "mild" (5–9), "moderate" (10–14), "moderately severe" (15–19), and "severe" (20–27) for depression severity. For up to two missing answers, the item value was replaced by the mean value of the answered questions. We defined a PHQ–9 score ≥ 10 as depression (28).

Basic similarities and differences

The CES-D and PHQ–9 are used to detect depression and depressive symptoms. Both scales are designed as self-administrative questionnaires, brief and easy to assess as well as available in the public domain. Nevertheless, there are differences. The PHQ–9 indicates major depression based on the DSM-IV diagnostic criteria and was developed for clinical use. The CES-D, on the other hand, was developed for large epidemiological studies and measures depressive symptoms with emphasis on the affective component and depressed mood. The CES-D consists of 15 items, whereas the PHQ–9 contains only nine items. Although there is a basic frequency questioning in both scales, the retrospective period differs. The CES-D refers to the last seven days, the PHQ–9 to the last 14 days. The response options are quite similar for both instruments. A four-step scale with increasing frequencies is given.

Demographic Variables

The socio-economic status was assessed in a standardized computer-assisted interview (CAPI) carried out by trained personnel at baseline examination. Education was classified according to the International Standard Classification of Education (ISCED–97) as total years of formal education, combining school and vocational training and was categorized into four groups (≤10 y, 11–13 y, 14–17 y, ≥18 y). Economic activity was categorized into four groups [employed, inactive (e.g. homemaker, but not unemployed), pensioner, and unemployed]. We also recorded if participants were cohabiting with a partner or not. For the classification by subgroups, age was divided into <67 and ≥ 67, as this corresponds to the mean age of the participants.

Statistical Analyses

Cronbach's alpha was calculated for CES-D and PHQ–9 to evaluate the internal consistency based on the correlations between different items on the total scale. It describes the extent to which all the items in a test measure the same concept or construct. Convergent validity of CES-D and PHQ–9 was assessed using Pearson's correlation coefficient to explore the magnitude of the associations between the scales.

Agreement probability is calculated by using the sum of the number of the same classification for both scales, as well as the disagreement probability as the sum of the number of a different classification. The response agreement between PHQ–9 and CES-D with dichotomous cutoffs was evaluated with Cohen's kappa. Cohen's kappa is used to assess the reliability of different measurement methods by quantifying their consistency in placing individuals or items in two or more mutually exclusive categories. We calculated kappa values including 95% confidence intervals (CI) by subtracting/adding the kappa from the value of the 95% CI level (1.96) times the standard error of kappa (29). We interpreted the strength of the agreement as slight (0–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1) (30).

McNemar's test of marginal homogeneity is conducted for the CES-D and PHQ−9, respectively, to test the hypothesis that the proportion of participants with depressive symptoms above the corresponding cutoff is the same at both times. Temporal changes in PHQ−9 and CES-D are represented by the percentage difference from t8 to t9. A change in the depression score is rated 10% or more on the given scale. Temporal changes were classified as an increase, decrease and no change. To assess the conformity of a temporal trend, the proportions of the concordantly and discordantly identified changes are compared.

For a sensitivity analysis, various cutoffs (CES-D score ≥16 to 22) were selected for the CES-D to demonstrate how the agreement and kappa values change. Participants were also stratified according to whether they are currently in therapy for depression. Based on this stratification, the observed agreement and the kappa value were calculated.

Descriptive results are expressed as mean ± SD, percentage (%) or number (n), as appropriate. Results are presented separately by sex. All analysis was performed using SAS 9.4.

Results

The sex-stratified characteristics of the analysed population at t8 (n = 3,084) are shown in table 1, for the entire population as well as for women (n = 1,580) and men (n = 1,504) separately. The average age was 66.8 years at t8. Of the participants, 85.8% lived with a partner and 34.9% completed post-secondary education (≥14 years of education); 45.5% were employed.

The mean score for PHQ−9 and CES-D score was 3.5 resp. 7.1. Women showed a higher score on both scales than men. Figure 1 presents the distribution of CES-D and PHQ−9 at t8 and t9. Both scores were strongly skewed to the right. The mode of PHQ−9 was even 0. For men, the distribution of the PHQ−9 scale was even more right-skewed, than for women. No differences according to the distribution were observable between both time points.

Table 2 shows the prevalence rates. The prevalence of depression at t8 was higher for CES-D (7.8%) than for PHQ−9 (4.4%) The prevalence slightly increased for CES-D (8.1%), as well as for PHQ−9 to 4.5% at t9. The sex-specific analyses admittedly show the well-known observation of higher prevalence in women than in men. However it also reveals (i) an even greater difference between sexes measured by CES-D (men 6.0%, women 9.6%) than by PHQ−9 (3.6%, 5.3%), and (ii) a decreasing prevalence rate in women measured by PHQ−9 (5.3% to 4.9%) whereas it increased when measured by CES-D (9.6% to 10.4%).

Cronbach's alpha showed high reliability for the CES-D (a = 0.89) at both time points (Table 3). Inter-item correlations ranged from 0.16 to 0.67. Similar results were found for PHQ−9 (t8: a = 0.85, t9: a = 0.84). Inter-item correlations ranged from 0.20 to 0.54. No sex-specific differences could be identified. Table 4 depicts a high correlation between CES-D and PHQ−9 (t8: r = 0.84, t9: r = 0.85). The long-term approach between t8 und t9 achieved a correlation of r = 0.71 for both CES-D and PHQ−9.

Table 5 shows the agreement of the two scales at t8 as well as at t9. At t8 both scales classified 113 participants as depressed and 2,818 as non-depressed, resulting in a 92.2% agreement. Kappa coefficient (k = 0.57; 95% CI: 0.51, 0.63) indicated a moderate agreement. Cohen's kappa at t8 was for men slightly higher (k = 0.61; 95% CI: 0.51, 0.70) than for women (k = 0.54; 95% CI: 0.47, 0.63). At t9 the agreement amounted 95.0%

and kappa coefficient of k = 0.58 (95% CI: 0.52, 0.64) for all participants. The agreement, considered separately by sex at t8, did not differ from t9. Figure 2 illustrates the results for the agreement of the two scales showing the kappa coefficient and the 95% confidence interval, divided into different subgroups for both time points. The kappa coefficients for the subgroups ranged between 0.51 and 0.68. The agreement between the CES-D and PHQ–9 can be considered as moderate to substantial within the different subgroups. At both times, the kappa coefficient was greater for men than for women, suggesting that there are more concordant cases for men. It can also be seen that the agreement between the two scales was greater among the younger participants (< 67 years) than among the older participants. Similarly, the agreement among participants with post-secondary education was lower than among participants with less than 14 years of education.

To answer the question of whether there was a higher proportion of participants who were above the cutoff for depression at t9 than at t8, the McNemar's test was performed for each scales. The results (Table 6) show that the proportion of being depressed did not differ significantly (p = 0. 539) according to the CES-D scale. The marginal percentages for being depressed at t8 were 7.8% and 8.1% at t9. For the PHQ–9, this result can be seen even more clearly. At t8 4.4% were depressed and at t9 4.5%. According to the McNemar's test, this leads to a p-value of 0.931, assuming that the proportions did not differ significantly from each other, too. The McNemar's test conducted separately by sex indicates that the proportions of being depressed did not differ between t8 and t9 on both scales.

As described, a relative change of more than 10% between the two time points on one of the scales was regarded as an increase or decrease in the depression score. Table 7 shows the results for analysing these development trends of both scales over time. In total, 75.8% of the participants on both scales identified similar trends (increasing, no change or decreasing) from t8 to t9. A similar pattern was observed among men for 77.1% of participants and 74.5% among women. A proportion of 24.2% of all participants, in contrast, shows a different development trend. Nevertheless, for 0.5% of the participants opposite developments are identified by the two scales.

Sensitivity analysis

By increasing the cutoff for the CES-D, the proportion of depressives decreased accordingly (Table 8). The agreement between the CES-D and PHQ–9 improved with a higher cutoff. The Kappa coefficient improved, too.

Table 9 shows the results stratified by current therapy for depression. The proportion of depressives currently in therapy turn out to be distinctively higher (t8: CES-D 45.2%, PHQ–9 33.3%; t9: 39.8%, 28.5%). Nevertheless, even among the group being not currently in therapy, 3 to 6% were depressed according to CES-D or PHQ–9 (t8: CES-D 5.7%, PHQ–9 2.8%; t9 6.1%; 2.9%). The agreement between the scales was higher among those who were not in therapy. Despite this, the Kappa coefficient at both time points was lower compared to those who were in therapy.

Discussion

This is the first study to compare systematically the psychometric properties of the CES-D and PHQ–9 in the general population in a longitudinal and sex-specific approach. The analyses carried out have shown that both scales achieved concordant results in detecting depression, but also differed from each other in some aspects.

In this study, 4.4% of the participants were identified as depressed with the PHQ–9, slightly lower than the latest WHO prevalence data for depressive disorders in Germany (5.2%) (1). The CES-D, in contrast, classified 8% as depressed in our sample. Women were more likely than men to suffer from depression. This difference was more noticeable in CES-D than in PHQ–9. Each scale performed well and proved to be highly reliable instruments with high Cronbach's alpha values for the total score from 0.84 - 0.89. Concerning internal consistency and convergent validity, our findings match those observed in other studies (10, 11). The degree of concordance of the results regarding a binary classification of participants as depressed or not depressed ranged in a moderate to a substantial level. These results are in line with other studies examining specific patient groups. Zhang et al. (13) compared the CES-D and PHQ–9 in type 2 diabetes patients in China, whereas Milette et al. (12) in patients with systemic sclerosis (SSc) in Canada showing both a moderate agreement between the instruments (k = 0.45; resp. k = 0.49). Their agreement was slightly weaker than in our study.

However, a proportion of the participants were also classified differently from the two scales. It has been shown that the CES-D assesses a participant as depressed rather than PHQ–9. This result is also consistent with other studies in which a higher sensitivity of the CES-D to PHQ–9 was found (31). Nevertheless, in our study some subgroups were more consistent in the agreement between the two scales, for example, men compared to women or participants younger than 67 years compared to even older participants. If the threshold for the cutoff of the CES-D is raised further, as the sensitivity analysis shows, the agreement between CES-D and PHQ–9 improved and the CES-D did not classify as many participants as depressed.

In addition, our study reveals that both scales detected identical tendencies observed with regard to changes in depression over time. Only 0.5% of participants were completely opposed to the development of depressive symptoms identified by the two scales. According to the McNemar's test, the proportions to be depressed did not differ between the two time points on both scales.

There were no sex-specific differences in psychometric characteristics. Cronbach's alpha and Pearson's correlation coefficients were similarly high for men and women. Merely the agreement between the CES-D and the PHQ–9 and the corresponding kappa values show differences between men and women. Men had a higher agreement than women. This may be due to the slightly different domains (e.g. mood, cognition, somatic symptoms) that the two scales are measured. The CES-D measures rather depressed mood and also positive affects, whereas the PHQ–9 measures the criteria of major depression. Furthermore, the CES-D asks about the frequency of depressive symptoms in the last week, while PHQ–9 asks about the frequency in the last two weeks. The shorter time interval captured by the CES-D may detect rather short-term symptoms, including acute problems and stressors that may not indicate major depression. It can be assumed that women are more likely to be in a depressed mood and do not yet meet the criteria for major depression. Therefore, they had an elevated score for the CES-D, but not for the PHQ–9. This may explain the discrepancy between the prevalence determined by CES-D and the PHQ–9 and finally the poorer agreement.

Strengths

There are serval characteristics that are strengths of this study. First, HNR Study is a large representative sample of the general middle and older population followed annually. There were multiple measurements of depressive symptoms available with the widely utilized and well-established scales CES-D and PHQ–9, so the

longitudinal analysis was possible. The sample size allowed the psychometric characteristics of CES-D and PHQ–9 to be properly investigated in the general population. We also stratified by gender to observe gender-specific differences in scales.

Limitations

The study also has a few limitations. The basis for concordance analysis is the dichotomization of the initial continuous scores. This results in a loss of information as it is not possible to determine how close the actual score lies to the cutoff. Another limitation is the lack of clinical confirmation of depression, so we were not able to evaluate our results with a valid diagnosis of depression. This would have allowed testing of CES-D and PHQ–9 sensitivity and specificity.

Future research

A comparison against a gold standard in the form of a structured clinical interview on depressive symptoms can provide information on the diagnostic accuracy of the two instruments.

Conclusion

In summary, we want to highlight the overall good performance of the CES-D and PHQ–9 in assessing depressive symptomatology in the general population. The results do not support the superiority of one scale over the other in terms of psychometric properties. Both scales perform well and their correlations were high, so the decision, which scale to use for a particular purpose may depend on characteristics other than the accuracy of the scale, including feasibility and clinical usability, for example, the PHQ–9 includes fewer items than the CES-D, is shorter and take less time. The PHQ–9 focuses more on the criteria of major depression in clinical practice, whereas the CES-D measures rather different aspects of depression such as depressive mood as well as positive affect, and was not designed for clinical diagnoses.

Declarations

Ethics approval and consent to participate

The HNR Study was approved by the local ethics committees of the Medical Faculty of the University of Duisburg-Essen, and all participants gave written, informed consent prior to participation.

Consent for publication

Not applicable.

Availability of data and materials

The corresponding author has full access to all data in the study and final responsibility for the submission of the article for publication. Due to data security reasons (i.e., data contain potentially participant identifying

information), the HNR Study does not allow sharing data as a public use file. Data requests can also be addressed to recall@uk-essen.de.

Competing interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors thank the Heinz Nixdorf Foundation [Chairman: Martin Nixdorf; Past Chairman: Dr jur. Gerhard Schmidt (†)], for their generous support of this study. Parts of the study were also supported by the German Research Council (DFG) [DFG project: EI 969/2–3, ER 155/6–1;6–2, HO 3314/2–1;2–2;2–3;4–3, INST 58219/32–1, JO 170/8–1, KN 885/3–1, PE 2309/2–1, SI 236/8–1;9–1;10–1,], the German Ministry of Education and Science [BMBF project: 01EG0401, 01GI0856, 01GI0860, 01GS0820_WB2-C, 01ER1001D, 01GI0205], the Ministry of Innovation, Science, Research and Technology, North Rhine-Westphalia (MIWFT-NRW). Furthermore, the study was supported by the deanship of the University Hospital and IFORES of the University Duisburg-Essen, the European Union, the German Competence Network Heart Failure and the Kulturstiftung Essen.

Authors' contributions

Conceptualization: ME, MP, SM; Methodology: ME, SM; KHJ; Formal analysis and investigation: ME, MP; Writing - original draft preparation: ME; Writing - review and editing: ME, MP, KJH, SS, US; SM; Funding acquisition: KHJ, SM; Supervision: SM

Acknowledgements

The authors express their gratitude to all study participants of the Heinz Nixdorf Recall (HNR) Study, the personnel of the HNR study center, the investigative group and all former employees of the HNR study. The authors also thank the Advisory Board of the HNR Study: T. Meinertz, Hamburg, Germany (Chair); C. Bode, Freiburg, Germany; P. J. de Feyter, Rotterdam, Netherlands; B. Güntert, Hall i.T., Austria; F. Gutzwiller, Bern, Switzerland; H. Heinen, Bonn, Germany; O. Hess (†), Bern, Switzerland; B. Klein (†), Essen, Germany; H. Löwel, Neuherberg, Germany; M. Reiser, Munich, Germany; G. Schmidt (†), Essen, Germany; M. Schwaiger, Munich, Germany; C. Steinmüller, Bonn, Germany; T. Theorell, Stockholm, Sweden; and S.N Willich, Berlin, Germany.

References

1.World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates.; 2017.

2.Alonso J, Petukhova M, Vilagut G, Chatterji S, Heeringa S, Ustun TB, et al. Days out of role due to common physical and mental conditions: results from the WHO World Mental Health surveys. Molecular psychiatry.

2011;16(12):1234–46.

3.Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological medicine. 2002;32(6):959–76.

4.Williams JW, Jr., Noel PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? Jama. 2002;287(9):1160–70.

5.American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM–5. Washington, DC American Psychiatric Publ.; 2013.

6.Kessler RC, Ustun TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). International journal of methods in psychiatric research. 2004;13(2):93–121.

7.Radloff LS. The CES-D Scale:A Self-Report Depression Scale for Research in the General Population. Appl Psychol Meas. 1977;1(3):385–401.

8.Kroenke K, Spitzer RL, Williams JB. The PHQ–9: validity of a brief depression severity measure. Journal of general internal medicine. 2001;16(9):606–13.

9.Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire–9 (PHQ–9) for screening to detect major depression: individual participant data meta-analysis. Bmj. 2019;365:l1476.

10.Amtmann D, Kim J, Chung H, Bamer AM, Askew RL, Wu S, et al. Comparing CESD–10, PHQ–9, and PROMIS depression instruments in individuals with multiple sclerosis. Rehabilitation psychology. 2014;59(2):220–9.

11.Khamseh ME, Baradaran HR, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M. Comparison of the CES-D and PHQ–9 depression scales in people with type 2 diabetes in Tehran, Iran. BMC psychiatry. 2011;11:61.

12.Milette K, Hudson M, Baron M, Thombs BD. Comparison of the PHQ–9 and CES-D depression scales in systemic sclerosis: internal consistency reliability, convergent validity and clinical correlates. Rheumatology (Oxford, England). 2010;49(4):789–96.

13.Zhang Y, Ting RZ, Lam MH, Lam SP, Yeung RO, Nan H, et al. Measuring depression with CES-D in Chinese patients with type 2 diabetes: the validity and its comparison to PHQ–9. BMC psychiatry. 2015;15:198.

14.Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. Journal of general internal medicine. 2007;22(11):1596–602.

15.Vilagut G, Forero CG, Barbaglia G, Alonso J. Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-Analysis. PLoS One. 2016;11(5):e0155431.

16.Grigoriadis S, Robinson GE. Gender issues in depression. Annals of clinical psychiatry: official journal of the American Academy of Clinical Psychiatrists. 2007;19(4):247–55.

17.Parker G, Fletcher K, Paterson A, Anderson J, Hong M. Gender differences in depression severity and symptoms across depressive sub-types. J Affect Disord. 2014;167:351–7.

18.Salk RH, Hyde JS, Abramson LY. Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. Psychological bulletin. 2017;143(8):783–822.

19.Kessler RC. Epidemiology of women and depression. J Affect Disord. 2003;74(1):5–13.

20.Schmermund A, Mohlenkamp S, Stang A, Gronemeyer D, Seibel R, Hirche H, et al. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. American heart journal. 2002;144(2):212–8.

21.Stang A, Moebus S, Dragano N, Beck EM, Mohlenkamp S, Schmermund A, et al. Baseline recruitment and analyses of nonresponse of the Heinz Nixdorf Recall Study: identifiability of phone numbers as the major determinant of response. European journal of epidemiology. 2005;20(6):489–96.

22.Clark VA, Aneshensel CS, Frerichs RR, Morgan TM. Analysis of effects of sex and age in response to items on the CES-D scale. Psychiatry research. 1981;5(2):171–81.

23.Fountoulakis KN, Bech P, Panagiotidis P, Siamouli M, Kantartzis S, Papadopoulou A, et al. Comparison of depressive indices: Reliability, validity, relationship to anxiety and personality and the role of age and life events. J Affect Disord. 2007;97(1):187–95.

24.Lewinsohn PM, Seeley JR, Roberts RE, Allen NB. Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. Psychology and aging. 1997;12(2):277–87.

25.Icks A, Albers B, Haastert B, Pechlivanis S, Pundt N, Slomiany U, et al. Risk for high depressive symptoms in diagnosed and previously undetected diabetes: 5-year follow-up results of the Heinz Nixdorf Recall study. PLoS One. 2013;8(2):e56300.

26.Hautzinger M, Bailer M. Allgemeine Depressions Skala. Manual. Göttingen: Beltz-Test-GmbH; 1993.

27.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(18):1737–44.

28.Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ–9): a meta-analysis. CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne. 2012;184(3):E191–6.

29.McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276–82.

30.Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159–74.

31.Beekman AT, Deeg DJ, Van Limbeek J, Braam AW, De Vries MZ, Van Tilburg W. Criterion validity of the Center for Epidemiologic Studies Depression scale (CES-D): results from a community-based sample of older subjects in The Netherlands. Psychological medicine. 1997;27(1):231–5.

Tables

Table 1 Demographic Characteristics

Characteristics	all				men				women			
	mean	SD	n	%	mean	SD	n	%	mean	SD	n	%
			3084	100.0			1504	48.8			1580	51.2
PHQ-9 (missings n=0)	3.5	3.5			3.0	3.4			3.9	3.5		
CES-D (missings n=0)	7.1	6.5			6.4	6.1			7.7	6.8		
Age in years (missings n=0)	66.8	7.3			67.0	7.3			66.7	7.4		
Age group (missings n=0)												
45-54			35	1.1			14	0.9			21	1.3
55-64			1207	39.1			560	37.2			647	41.0
65-75			1419	46.0			727	48.3			692	43.8
75-85			423	13.7			203	13.5			220	13.9
Partnership (missings n=2)			2645	85.8			1409	93.8			1236	78.3
Education in years (missings n=3)												
≤10			270	8.8			58	3.9			212	13.4
11-13			1735	56.3			690	45.9			1045	66.2
14-17			710	23.0			527	35.1			183	11.6
≥18			366	11.9			228	15.2			138	8.8
Employed (missings n=2)												
employed			1403	45.5			813	54.1			590	37.4
inactive o housewife			444	14.4			7	0.5			437	27.7
pensioner			1049	34.0			606	40.3			443	28.1
unemployed			186	6.0			77	5.1			109	6.9

Table 2 Prevalence of Depression Scores

	all		men		women	
	n	%	n	%	n	%
PHQ-9 t8	137	4.4	54	3.6	83	5.3
PHQ-9 t9	138	4.5	60	4.0	78	4.9
CES-D t8	242	7.8	90	6.0	152	9.6
CES-D t9	251	8.1	88	5.9	163	10.4

Table 3 Cronbach's alpha Coefficient of Internal Consistency

	all (n=3084)			men (n=1504)			women (n=1580)		
	n[1]	%	Cronbach's alpha	n	%	Cronbach's alpha	n	%	Cronbach's alpha
PHQ-9 t8	3011	97.6	0.85	1476	98.1	0.85	1535	97.2	0.84
PHQ-9 t9	2992	97.0	0.84	1467	97.5	0.84	1525	96.5	0.84
CES-D t8	2852	92.5	0.89	1411	93.8	0.88	1441	91.2	0.89
CES-D t9	2823	91.5	0.89	1381	91.8	0.89	1442	91.3	0.90

[1] n is slightly different for each scale because not all questionnaires were filled out completely

Table 4 Pearson's Correlation Coefficient between Depression Scales

all				
Scale	PHQ-9 t8	PHQ-9 t9	CES-D t8	CES-D t9
PHQ-9 t8				
PHQ-9 t9	0.71			
CES-D t8	0.84	0.67		
CES-D t9	0.65	0.85	0.71	

men				
Scale	PHQ-9 t8	PHQ-9 t9	CES-D t8	CES-D t9
PHQ-9 t8				
PHQ-9 t9	0.71			
CES-D t8	0.83	0.68		
CES-D t9	0.65	0.84	0.71	

women				
Scale	PHQ-9 t8	PHQ-9 t9	CES-D t8	CES-D t9
PHQ-9 t8				
PHQ-9 t9	0.71			
CES-D t8	0.84	0.66		
CES-D t9	0.65	0.86	0.70	

All correlations are significant at $p < 0.0001$

Table 5 Agreement of depression scores PHQ-9 and CES-D with Cohen's kappa and 95% confidence interval

	CES-D & PHQ-9 both classify participants as						
	non-depressed	depressed	observed Agreement	expected Agreement	Cohen's kappa	95% lower CI	95% Upper CI
t8 all	2818	113	95.0%	88.4%	0.57	0.51	0.63
t9 all	2812	117	95.0%	88.1%	0.58	0.52	0.64
t8 men	1405	45	96.4%	90.9%	0.61	0.51	0.70
t8 women	1413	68	93.7%	86.1%	0.55	0.47	0.63
t9 men	1404	48	96.5%	90.6%	0.63	0.54	0.72
t9 women	1408	69	93.5%	85.8%	0.54	0.47	0.69

Table 6 Contingency tables for McNemar's test of marginal homogeneity for being depressed in t8 vs. t9

			t9						
			CES-D			PHQ-9			
			-	+	total	-	+	total	
t8	CES-D	-	2730	112	2842	(92.2%)			
		+	103	139	242	(7.8%)			
		total	2833	251	3084				
			(91.9%)	(8.1%)					
	PHQ-9	-				2880	67	2947	(95.6%)
		+				66	71	137	(4.4%)
		total				2946	138	3084	
						(95.5%)	(4.5%)		

Table 7 Temporal changes over time

CES-D	PHQ-9	all		men		women	
		n	%	n	%	n	%
↗	↗	184	6.0	79	5.3	105	6.6
→	→	2002	64.9	1028	68.4	974	61.6
↘	↘	151	4.9	53	3.5	98	6.2
↗	→	237	7.7	103	6.8	134	8.5
↗	↘	6	0.2	2	0.1	4	0.3
→	↗	146	4.7	77	5.1	69	4.4
→	↘	148	4.8	69	4.6	79	5.0
↘	↗	8	0.3	4	0.3	4	0.3
↘	→	202	6.5	89	5.9	113	7.2
		3084	100	1504	100.0	1580	100.0

↗ Increase
↘ Decrease
→ No change

Table 8: Sensitivity Analysis: Different Cutoffs for CES-D: Proportion of being depressed and Agreement of PHQ-9 and CES-D with Cohen's kappa and 95% confidence interval

CES-D Cutoff ≥		depressed	observed Agreement	expected Agreement	Cohen's kappa	95% lower CI	95% Upper CI
16	t8	9.4%	93.8%	87.0%	0.53	0.47	0.58
	t9	9.5%	94.0%	86.9%	0.54	0.49	0.60
17	t8	7.9%	95.0%	88.4%	0.57	0.51	0.63
	t9	8.1%	95.0%	88.1%	0.58	0.52	0.64
18	t8	7.1%	95.7%	89.1%	0.60	0.54	0.66
	t9	7.0%	95.7%	89.1%	0.61	0.54	0.67
19	t8	6.1%	96.2%	90.0%	0.62	0.56	0.69
	t9	6.0%	96.5%	90.1%	0.64	0.58	0.71
20	t8	5.3%	96.8%	90.7%	0.65	0.59	0.72
	t9	5.1%	96.7%	90.9%	0.64	0.58	0.71
21	t8	4.7%	97.2%	91.3%	0.68	0.62	0.74
	t9	4.6%	96.9%	91.4%	0.64	0.58	0.71
22	t8	4.1%	97.5%	91.8%	0.69	0.63	0.76
	t9	4.0%	97.1%	91.9%	0.64	0.57	0.71

Table 9: Sensitivity Analysis: Proportion of being depressed and Agreement of PHQ-9 and CES-D with Cohen's kappa and 95% confidence interval by intake of antidepressants or current therapy for depression

Time point	Intake of antidepressants or current therapy for depression	n	depressed CES-D	depressed PHQ-9	observed Agreement	expected Agreement	Cohen's kappa	95% lower CI	95% Upper CI
t8	yes	168	45.2%	33.3%	0.83	0.52	0.66	0.54	0.77
t8	no	2916	5.7%	2.8%	0.94	0.90	0.47	0.40	0.55
t9	yes	186	39.8%	28.5%	0.84	0.54	0.66	0.55	0.77
t9	no	2898	6.1%	2.9%	0.96	0.91	0.50	0.42	0.57

List Of Abbreviations

List of abbreviations

CES-D Center for Epidemiologic Studies Depression Scale

DS Depressive symptoms

DSM-V Diagnostic and Statistical Manual of Mental Disorders, 5th ed

HNR Heinz Nixdorf Recall

MDD Major Depressive Disorders

PHQ-9 Patient Health Questionnaire-9

SD Standard Deviation

Figures

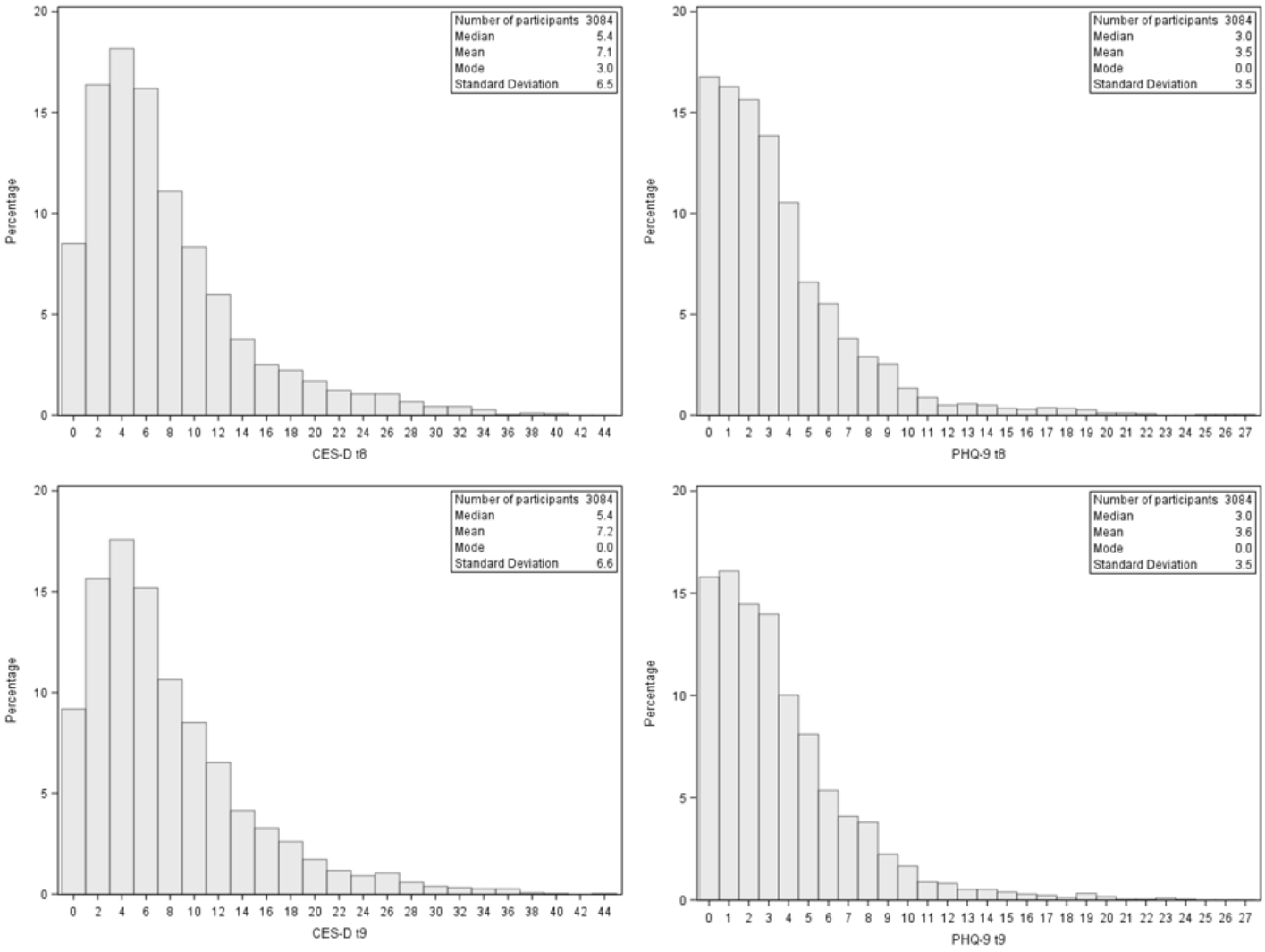

Figure 1

Distribution of Depression Scores

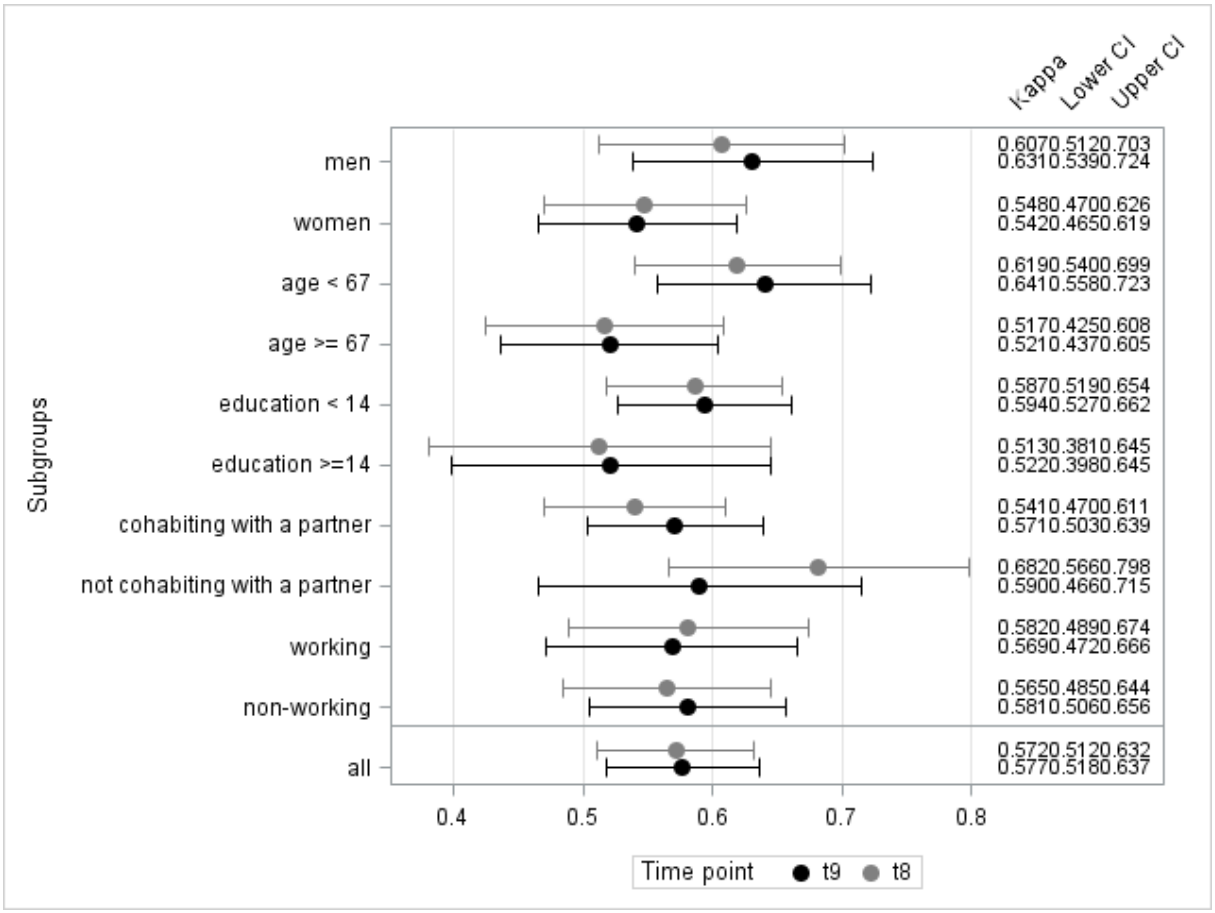

Figure 2

Kappa coefficient with 95% confidence interval (CI) by subgroups