REVISTA CIENCIAS BIOMÉDICAS

ARTÍCULO DE REVISIÓN

Diagnostic accuracy of scales for depression screening in patients with heart failure: systematic review and meta-analysis

Precisión diagnóstica de las escalas para el tamizaje de la depresión en pacientes con insuficiencia cardíaca: revisión sistemática y meta-análisis

Carlos Arturo Cassiani-Miranda[1], Martin Rueda[2], Paul Anthony Camacho[3]

[1] Faculty of Health Sciences. Medicine program. Mental health and human behavior Research Group – Universidad de Caldas, Manizales, Colombia.
[2] Medical Student. Research Group Neuropsychiatry, Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia.
[3] Neuropsychiatry Research Group - Universidad Autónoma de Bucaramanga, Colombia.

ACCESO ABIERTO

Para citaciones: Cassiani Miranda, C., Rueda, M., & Camacho, A. (2022). Diagnostic accuracy of scales for depression screening in patients with heart failure: systematic review and meta-analysis. Revista Ciencias Biomédicas, 11(2), 127-144. https://doi.org/10.32997/rcb-2022-3934

Recibido: 26 de octubre de 2022
Aprobado: 25 de enero 2022

Autor de correspondencia:
Carlos Arturo Cassiani-Miranda
pcamacho@unab.edu.co

Editor: Inés Benedetti. Universidad de Cartagena-Colombia.

ABSTRACT

Introduction: despite the existence of recommendations for the screening depressive symptoms in patients with cardiovascular disease and heart failure (HF), there are no comparative data regarding the performance of psychometric scales used in patients with HF. This study compares the psychometric performance of screening scales used for depressive symptoms in such patients.

Methods: PRISMA declaration recommendations were used for the systematic review. MEDLINE, EMBASE, Psychology and Behavioral Sciences Collection, SCOPUS, Lilacs, Australasian Medical Index and the African Index from January 2000 to February 2016 were used for the search. The eligible articles were published in any language and they assessed the psychometric properties of screening scales for depressive symptoms in patients with HF. QUADAS-2 criteria was used for quality assessment, and a meta-analysis developed through a hierarchical model obtained the cluster estimations for sensitivity, specificity, likelihood ratio, predictive values, and diagnostic odds ratio (DOR) with 95% confidence intervals.

Results: the initial search identified 1238 citations; only three gathered the inclusion criteria for quantitative assessment. The combined sensitivity and specificity was 56% (95% IC: 45-67%; T^2=0.05) and 98% (95% IC: 96-99%; T^2=0.01) respectively. The area under the curve was 0.92 (95% IC: 0.90-0.94). The variables related with the index test, reference test, Global QUDAS-2 score, and language predicted heterogeneity. Limitations: significant heterogeneity, small number of studies, selective cutoff report, and the lack of a cost-effectiveness analysis.

Conclusions: The GDS-15, HADS-D, PHQ-9, CAT-D and PROMIS scales performed similarly with high specificity values.

Keywords: Screening; depressive disorder; heart failure; systematic review; meta-analysis; diagnostic accuracy.

RESUMEN

Introducción: a pesar de la existencia de recomendaciones para el cribado de síntomas depresivos en pacientes con enfermedad cardiovascular e insuficiencia cardiaca (IC), no existen datos comparativos sobre el rendimiento de las escalas psicométricas utilizadas en pacientes con IC. Este estudio compara el rendimiento psicométrico de las escalas de cribado utilizadas para los síntomas depresivos en dichos pacientesprotocols, the prevalence of febrile neutropenia in these patients has increased.

Métodos: para la revisión sistemática se utilizaron las recomendaciones de la declaración PRISMA. Para la búsqueda se utilizaron MEDLINE, EMBASE, Psychology and Behavioral Sciences Collection, SCOPUS, Lilacs, Australasian Medical Index y el African Index desde enero de 2000 hasta febrero de 2016. Los artículos elegibles se publicaron en cualquier idioma y evaluaron las propiedades psicométricas de las escalas de cribado de síntomas depresivos en pacientes con IC. Se utilizaron los criterios QUADAS-2 para la evaluación de la calidad, y un meta-análisis desarrollado a través de un modelo jerárquico obtuvo las estimaciones agrupadas para la sensibilidad, la especificidad, la razón de verosimilitud, los valores predictivos y la razón de probabilidades de diagnóstico (DOR) con intervalos de confianza del 95%.

Resultados: la búsqueda inicial identificó 1238 citas; sólo tres reunían los criterios de inclusión para la evaluación cuantitativa. La sensibilidad y especificidad combinadas fueron del 56% (IC del 95%: 45-67%; T2=0,05) y del 98% (IC del 95%: 96-99%; T2=0,01) respectivamente. El área bajo la curva fue de 0,92 (IC del 95%: 0,90-0,94). Las variables relacionadas con la prueba índice, la prueba de referencia, la puntuación global QUDAS-2 y el idioma predijeron la heterogeneidad. Limitaciones: Heterogeneidad significativa, pequeño número de estudios, informe de corte selectivo y la falta de un análisis de coste-efectividad.

Conclusión: las escalas GDS-15, HADS-D, PHQ-9, CAT-D y PROMIS se comportaron de forma similar con altos valores de especificidad.

Palabras Clave: Cribado; trastorno depresivo; insuficiencia cardíaca; revisión sistemática; metaanálisis; precisión diagnóstica.

INTRODUCTION

Depression is more common in patients with cardiovascular disease (CVD) - primarily in heart failure (HF) carrier patients - than it is in the general population (1, 2).

According to the instrument used, the prevalence of depression in patients with HF ranges between 11 and 77% (3, 4).

Prospective studies have shown that patients with depression symptoms suffer greater mortality rates from cardiac causes or hospitalization for HF (34% vs. 10.3%; P<0.01), hospitalization for HF (27.4% vs. 9.2%; P = 0.01), all causes of death (27.4% vs. 7.2%; P < 0.01), and prolonged hospital stays (5-8). This data leads us to consider depression as a first-order problem in terms of the comprehensive care provided to patients with HF (9, 10).

Considering the fact that mortality in patients with chronic HF remain extremely high and the global impact of depression in these patients (11), medical associations have suggested that the depression must be evaluated and treated systematically in this patients group (9, 12). Consequently, the American Heart Association (AHA) has published a series of recommendations for the screening of depressive symptoms with CVD (13, 14), whereas other data suggest the screening of depressive symptoms in patients with HF (15).

The gold standard for the diagnostic of a major depressive episode is the clinical interview that evaluates the extent to which a patient complies with the criteria for the diagnostic in DSM-5 (16) or CIE-10 (17). As it is impractical to administer an interview of this type to all patients with CVD, several smaller detection tools have been developed and some of these have been validated specifically in patients with cardiac disease (18).

The number of available questionnaires for the evaluation of the health condition has increased drastically in recent decades (12). As such, the choice of the questionnaire to be used is turning into a major difficulty (12, 19).

The detection tools and the cut-off point established for every one of the scales used in primary attention may not be appropriate for patients with CVD, given that some of the symptoms of cardiac disease may be confused with depressive symptoms (20, 21). The recommendations for depression detection must be determined in each population since the results obtained from patient groups cannot be generalized (22, 23).

The majority of the studies used to evaluate depressive symptomatology in patients with HF are based on self-administered or hetero-administered screening instruments or telephone interviews, and are rarely are based on the clinical diagnostic of depression (24). Only in few studies has the psychometric performance of the same scales been evaluated in patients with HF (25).

The measurement instruments used to evaluate depression in patients with CVD include: Beck's Depression Inventory (Beck Depression Scale-BDI) (26), Center for epidemiological studies-CES D (27, 28), Zung Depression Scale-ZDS (25), Hospital Anxiety Depression Index (HADI) (29), Cardiac Depression Scale-CDS (30), Geriatric Depression Scale-GDS, Hospital Anxiety and Depression Scale-HADS (31), Patient Health Questionnaire-2-PHQ-2 2, and Patient Health Questionnaire-9-PHQ-9 (32).

A suitable psychometric performance of these scales for depression screening in patients with HF has been reported: GDS and HADS (31). The measurements found were: GDS sensitivity 0.810, specificity 0.833, and cut-off 5; HADS sensitivity 0.938, specificity 0.847, and cut-off 7. For its part, PHQ- 9 (33), with a cut-off equal to 10 showed a

sensitivity of 70% and specificity of 92%. Nevertheless, there are no consistent comparative data regarding the psychometric performance of these scales in patients with HF.

There is no consensus on which could be the most useful questionnaire to evaluate the depressive symptomatology in patients with HF, given that diagnostic precision is compromised due to the overlap of some depressive symptoms with the symptoms of cardiac disease (22, 34).

Systematic reviews have been found on the prevalence of depression in cardiac failure (35), and the precision of screening instruments for depression in CVD (22). Delville et al. approached the problem of psychometrics properties specifically in patients with HF; nevertheless, the systematic reviews involve significant methodological deficiencies and they do not comply with international recommendations currently validated for the development of systematic reviews and meta-analyses (36).

It is therefore necessary to carry out a systematic review of the psychometric performance of instruments designed for this goal, in order to issue recommendations based on evidence regarding the most suitable instrument for the screening of depressive symptoms in patients with HF. Consequently, the objective of this systematic review is to compare the psychometric performance of the screening scales used for depressive symptomatology in patients diagnosed with HF.

METHODS

Search strategy

In order to identify the relevant studies of interest to us, we carried out a search in the following databases: MEDLINE, EMBASE, Psychology and Behavioral Sciences Collection, SCOPUS, Lilacs, Australasian Medical Index and the African Index for the January 2000 to February 2016 period; without language restrictions. We chose the 2000-2016 period to ensure that the articles included more recent diagnostic criteria for major depressive disorder of DSM (DSM-IV-TR and DSM-5). The search strategy focused on the terms of the diagnostic test of interest (Screening Tests for Depression) and the clinical disorder that this test attempts to detect (inpatients or outpatients with a diagnostic of heart failure).

The search strategy was first implemented in MEDLINE, and later adjusted for the other databases. The terms "Screening," "Depressive Disorder," and "Hearth Failure" were selected, as well as the MesH terminology range without methodology filters. A manual search was conducted from the list of the articles referenced in full text that would fulfill the inclusion criteria. We traced the citations from the articles included using Google Scholar (37), and conducted a search of abstracts from conferences through the BIOSIS database (http://www.biosis.org/) Meeting Abstracts (www.biomedcentral.com/meetings/), and the Conference Papers Index (www.csa.com/factsheets/cpi-set-c.php). To identify theses and dissertations, we conducted a search using Google Scholar, Networked (NDLTD; http://www.ndltd.org/), and ProQuest (http://www.biosis.org/).

To identify unpublished studies and studies in process, we conducted a search in databases from US Health Services Research Projects in Progress (www.nlm.nih.gov/hsrproj/) and the UK National Research Register (portal.nihr.ac.uk/Pages/NRRArchive.aspx).

Two researchers from the group (CCA and CPC) evaluated the studies for inclusion. In the first instance, they reviewed titles that included keywords. Next, they reviewed all abstracts that included the criteria to be elected. Finally, they read the article in full. If both researchers chose the

same article after having read its title and abstract, this would then be subjected to a review of the full text. Disagreements between the researchers were solved in consensus.

Research selection

Study Type: Systematic review and meta-analysis including studies comparing screening instruments with structured psychiatric interviews.

The methods used were based on the guidelines and recommendations established for Cochrane collaboration (38), as well as the recommendations in the PRISMA declaration (25). The search included studies with validation of scales with reference patterns or observational studies with depression diagnostic that applied screening scales with diagnostic confirmation through structured interview. The eligible articles were those that valuated the psychometric properties of scales for depression screening in patients with HF in any clinical stage and that were published in any language.

In addition to evaluating reliability, validity, and diagnostic precision, the screening instruments in the studies included needed to be compared with standard criteria for major depressive disorder according to the "Diagnostic and Statistical Manual of Mental Disorders" or the "International Classification of Diseases Diagnostic of MDD." These tools provide information with which to construct a 2×2 contingency table, making it possible to calculate sensitivity, specificity, positive predictive value, negative predictive value, and positive and negative likelihood ratios. Participants in the primary studies were inpatients or outpatients of any sex with HF diagnostic at any stage. The outcome of interest was the diagnostic of any depressive disorder.

Studies included whether the depression diagnostic was carried out using reference standards, i.e., through a structured interview such as the Structured Clinical Interview for DSM Disorders (SCID-I), Composite International Diagnostic Interview (CIDI), Mini International Neuropsychiatric Interview (MINI) or any other gold standard psychiatric interview for the scientific community.

The studies in which a depression diagnostic was conducted by a clinician such as an unstructured interview or a comparison between screening instrument and another self-administrated scale were excluded. Those studies that included a mixed population (heart failure and other diagnostics) were included if the outcomes were reported separately, or if the population with HF was greater than 80%.

Data extraction

Standard forms for data collection were used to record the information of interest from the selected articles. The variables extracted were: Sample characteristics (country, outpatient or inpatient setting, age, gender), sample size and MDD proportion according to the reference pattern, information regarding the screening scale used (medication method, language, cut-off, sensitivity, specificity, PPV, NPV, likelihood ratio, area under curve), and reference pattern details. The articles' authors were contacted when it was necessary to clarify information. Data were recorded in contingence tables to calculate the necessary variables to analyze the scales' psychometric properties.

Evaluation of methodological quality and analysis of bias

Quality and risk of bias assessments were carried out in the primary studies using the QUADAS-2 criteria (39). All items from QUADAS-2 regarding risk of bias and outcomes applications were considered. The questions to determine risk of bias and the concern about the applicability of the studies' outcomes were adapted according to the systematic review. In the first domain, the

following question was added: "Patients with cognitive impairment were excluded"? to guarantee greater precision in the detection of depressive symptoms. In second domain, the following question was added: "Was it reported whether the test was self-applied or hetero-applied, if it was hetero-applied who was in charge of the application?".

In the third domain, the following question was added: "Was the application of the reference test conducted by appropriately qualified personnel?" with the objective to guarantee greater accuracy in the depression diagnostic.

Statistical analysis

The Meta-Analysis was implemented in STATA 12 software using MIDAS and METANDI scripts. Sensitivity, specificity and LR+ and LR- were estimated from the proportion of positive or negative tests for subjects who were ill or not ill obtaining the following results for each study and combined studies: the diagnostic odds ratio ($DOR = LR+/LR-$), and confidence interval (CI) 95%. $Cochran\ Q$ and I^2 were explored for preliminary evaluation of heterogeneity of included studies.

A model with random effects was used because the Q test was significant ($p < 0.05$) and $I^2 > 50\%$. Meta-regression was planned for cases where heterogeneity exists ($I^2 > 50\%$.). Moreover, we calculated a Hierarchical Summary Receiver Operating Characteristic (HSROC) curve with logit estimations derived from and specificity of the selected studies to evaluate overall performance of tests through their different thresholds. Cook's Ratio was used to detect influential studies and a dispersion diagram of predicted standardized random effects was created to check outliers. The publication bias was evaluated using Deeks' funnel plot and a $p < 0.05$ was considered to detect publication bias.

RESULTS

Searching process

An initial search identified 1238 citations (1273 prior to discarding duplicates), among which, 35 were preselected by title and abstract. Among the latter, only 17 fulfilled the selection criteria to be read in full text. From these articles, three fulfilled the eligibility criteria required for them to be subjected to quantitative analysis (31, 40, 41). The reasons to exclude these 14 articles were: that they did not use a structured psychiatric interview (N=3) as reference criteria, the mixed population and patients with HF diagnostic they included were less than 85% (N=4), and they used a screening questionnaire (N=3) as reference criteria, the population included was not specifically of patients with HF (N=2), and there were no sensitivity and specificity data (N=2). The selection of the studies is summarized in PRISMA flux diagram (Figure 1) (42). Searching strategy details are shown in Annex 1.

Joint view of the articles included

Table 1 summarizes the characteristics of the articles included: three studies that evaluated six scales to detect depression in patients with HF diagnostic. Haworth JE et al. evaluated the GDS-15 and HADS scales (31). Fischer HF et al, evaluated four scales: HADS, PHQ-9, PROMIS and CAT-D (40), and Poole NA et al. evaluated HADS. Two of these studies were conducted in the United Kingdom (31, 41), and one in Germany (40). Two of these studies were applied to outpatients (31, 40), and one to inpatients (41).

Three studies showed transversal designs using reference patterns with consecutive samples of between 88 (31) and 194 subjects per sample (40). The average age was 43 (41) and 69.9 years (31). All the studies showed a predominance of male participants with a relative frequency of between 79.1% (40) and 85.3% (41).

Figure 1. PRISMA Flow Chart on Search and Selection of Studies Included

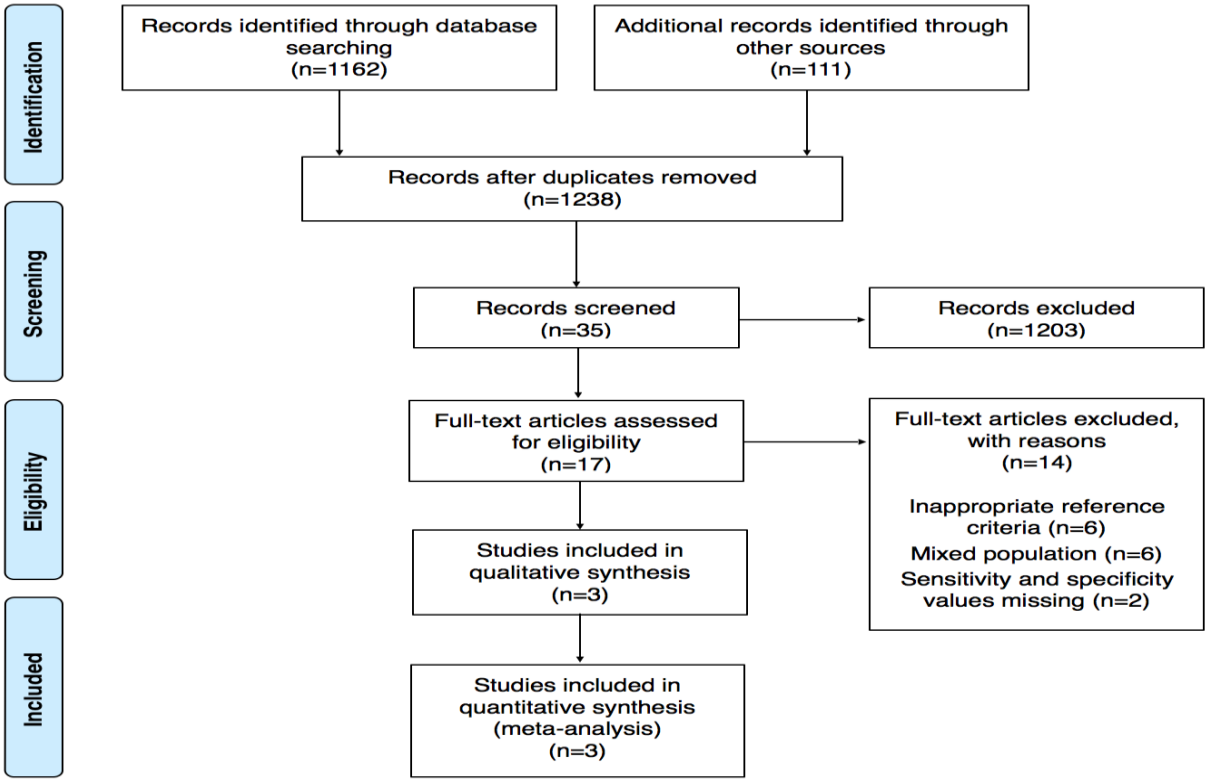


Table 1. Descriptive Characteristics of Included Studies

Study	Sample Characteristics	Diagnostic Standard	Depressed	Instrument	Instrument Characteristics
Fischer HF, et al 2014 (99)	Size: 194 Country: Germany Setting: Ambulatory care Age: x=69.9 SD=10.6 Range 35-88 years Female: 21.1%	SCID-I	13.9%	HADS PHQ-9 PROMIS-D CAT-D	Administration: Interviewer report Language: German
Haworth JE, et al 2007 (33)	Size: 88 Country: United Kingdom Setting: Community Age: x=69.9 SD=7.6 Range 56-92 years Female: N=17	SCID-I	25%	GDS-15 HADS	Administration: Interviewer report Language: English
Poole NA, et al 2006 (100)	Size: 115 Country: United Kingdom Setting: Inpatient care Age: x=43 Range 23-63 years Female: 14.7%	SCID-III-R-np	21%	HADS	Administration: Self-report Language: English

DISH, Depression Interview Structured Hamilton; SCID-I, Structured Clinical Interview for DSM-IV Axis I Disorders; SCID-III-R-np, Structured Clinical Interview for DSM-III-R Non-patient Version; BDI, Beck Depression Inventory; PHQ, Patient Health Questionnaire; HADS, Hospital Anxiety and Depression Scale; PROMIS, Patient Reported Outcome Measurement Information System-Depression-Short Form; CAT-D, Computer-Adaptative Tests-Depression.

The Prevalence of any depressive episode according reference pattern scores was of between 13.9% (40) and 25% (31). The reference pattern used in these three studies was SCID-I (31, 40, 41). In two studies, the index test was hetero-applied and was conducted by trained personnel in the researcher team (31, 40). In one study, the test was self-applied (41).

The cut-off in the test indexes was not the same among the studies and was established from the area under the ROC curve analysis (for GDS-15) (31, 40, 41) and prior studies' recommendations (for HADS-D) (33). The reference test was conducted by trained personnel in the researcher team (41). One study did not specify the profession or training of interviewers (31), whereas the other study test was conducted by a trained psychologist (40).

Evaluation of methodological quality and risk of bias

Figure 2 summarizes the outcomes from the evaluation of the methodological quality and risk of bias of the studies according to QUADAS-2 criteria. Regarding bias risk, neither was classified with low risk of bias in all domains. The only domains classified with uncertain risk were the index test and reference test number 1 of the studies (40). The only domains that revealed low bias risk for all studies were flux and time. According to the applicability criteria, all the studies were classified with low concern about applicability in the three domains.

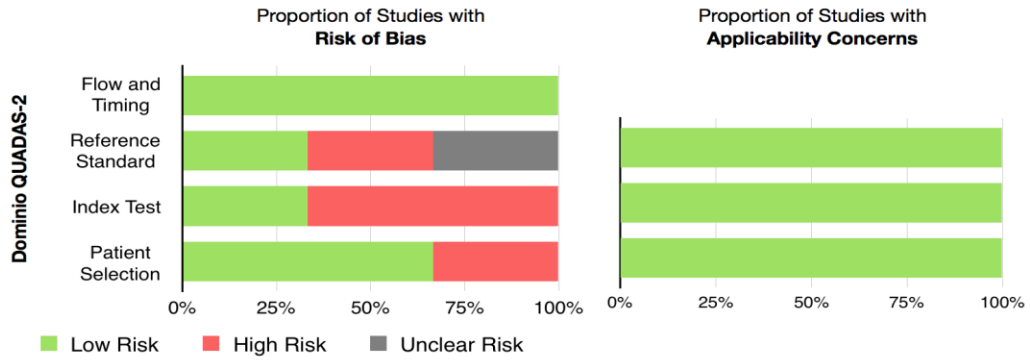
Figure 2. Assessment of Risk of Bias and Methodological Quality of Included Studies

(a) Tabular Presentation

Study	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Fiswcher H F, et al 2014 (99)	☺	☹	?	☺	☺	☺	☺
Haworth J E, et al 2007 (33)	☺	☹	☹	☺	☺	☺	☺
Poole N A, et al 2006 (100)	☹	☺	☺	☺	☺	☺	☺

☺ Low Risk; ☹ High Risk; ? Unclear Risk

(b) Graphic Presentation



Diagnostic properties of studies included

Depression prevalence reported later in the evaluation of 873 subjects with applied scales in the three studies was 29% [IC95%: 25% - 33%; T^2=0.01] with a range of 24% - 38% (Figure 3), did not reveal significant heterogeneity (I^2=27.4%, p=0.23). The data revealed that combined sensitivity of the screening scales for depression was 56% [IC95%: 45% - 67%; T^2=0.05] and combined specificity of 98% [IC95%: 96% - 99%; T^2=0.01]. Positive

likelihood ratio (LR+), negative likelihood ratio (LR-) and totalized DOR were 26.3 [IC95%: 12 - 57], 0.45 [IC95%: 0.35 – 0.58] and 59 [IC95%: 24 -144], respectively. Sensitivity heterogeneity for the evaluated scales revealed a Cochrane Q test of 15.43 (p=0.01,) I^2 = 67.59 [IC95%: 39.58 – 96.01]. The specificity heterogeneity of the evaluated scales revealed a Cochrane Q test of 12.79 (p=0.03,) I^2 = 60.91 [IC95%: 25.93 – 95.88]. Heterogeneity for sensitivity and specificity of evaluated scales were moderately high (figure 4).

Figure 3. Overall Prevalence of Depression of Included Studies

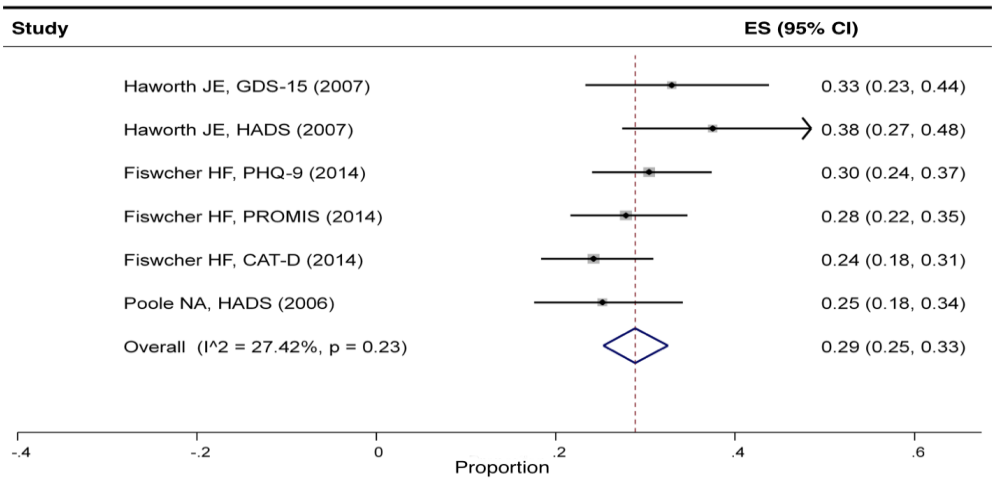
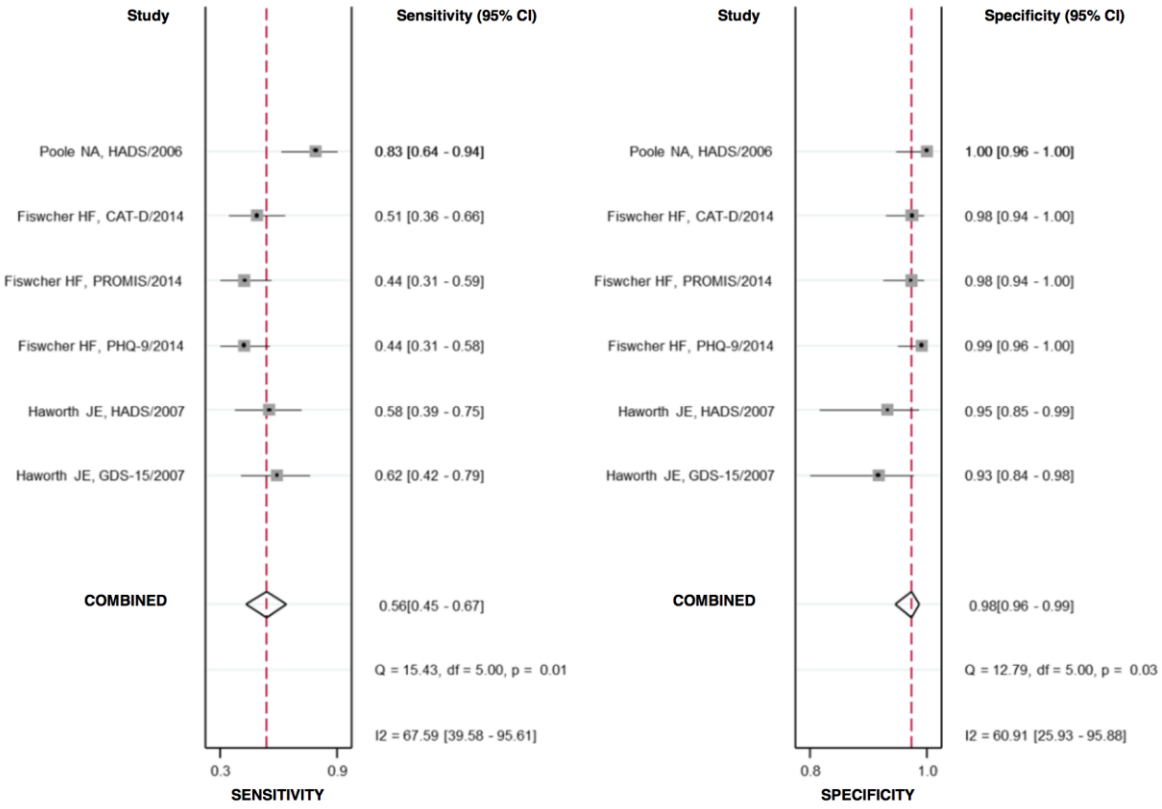
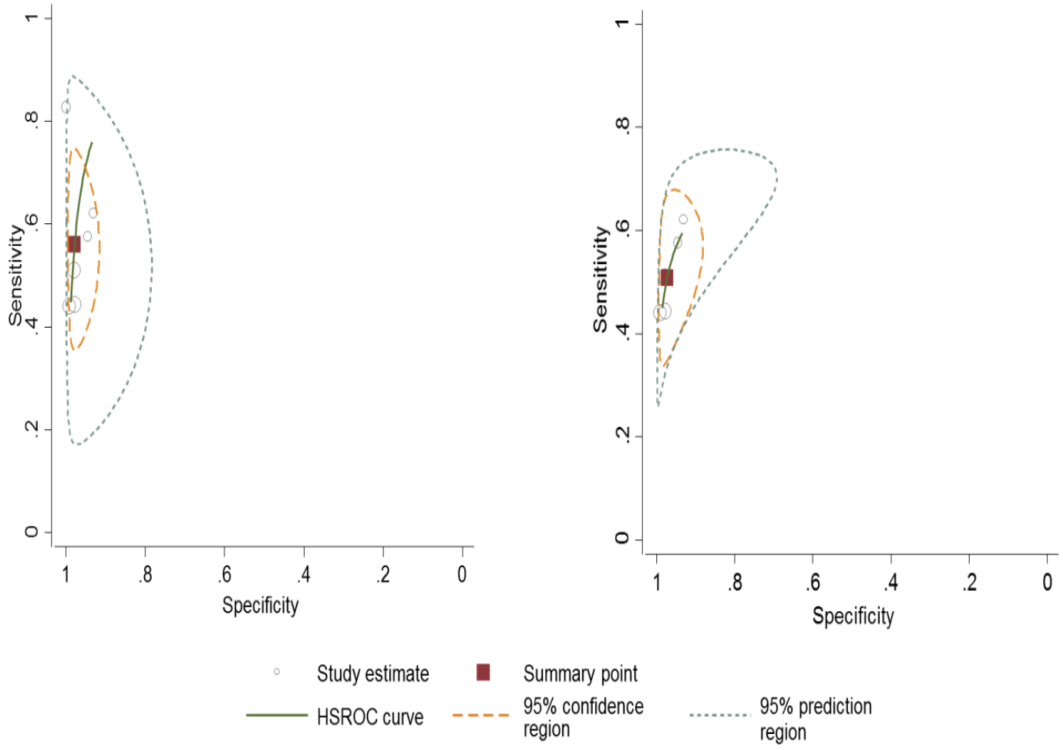


Figure 4. Forest Plots Showing the Sensitivity and Specificity for Individual Analyzed Scales with their 95% Confidence Intervals, Combined Sensitivity and Combined Specificity from All the Included Scales



The HSROC curve has an AUC of 0.92 [IC95%: 0.90 – 0.94], showed a high global psychometric performance for the included scales for a cut-off with sensitivity equal to 56% and specificity equal to 98% (Figure 5 a). The diameter of the circle (study estimate) is proportional to the weight assigned to each of the studies. The summary of sensitivity and specificity is indicated by a red box showing a low variability as a function of specificity. It was observed that the inclusion of the HADS scale in the analysis improved the scales' global discriminatory capacity (Figure 5a). The scales global discriminatory capacity shows an AUC of 0.78 [IC95%: 0.74 – 0.81], due to a reduction of the sensitivity discriminatory capacity (Figure 5 b). The bivariate random effect model was robust to estimate grouped data according goodness of fit and bivariate normality (Figures 6a and 6b). Evaluating Cook's ratio, it was observed that Poole et al.'s study was influential with non-typical behavior with the bigger standard of residuals for sensitivity (Figures 6c and 6d) (41).

Figure 5. HSROC Curve of Depressive Scales for Following Three Studies and Six Evaluated cales



Meta-regression model

A univariate regression model and subgroups analysis showed that the variables related with the index test, reference test, QUADAS-2 global score and language modified the sensitivity estimation while for the specificity, none of the selected variables had a significant effect on heterogeneity (Table 2). However, when analyzing the joint model, it was evidenced that all a priori variables had a significant effect on the observed heterogeneity (Table 3).

Evaluation of publication bias

Deeks' Asymmetry test did not reveal publication biases in the analyzed articles with p = 0.993 (Figure 7).

Figure 6. Graphical Depictions of Residual-Based Goodness-of-Fit, Bivariate Normality, Influence and Outlier Detection Analyses

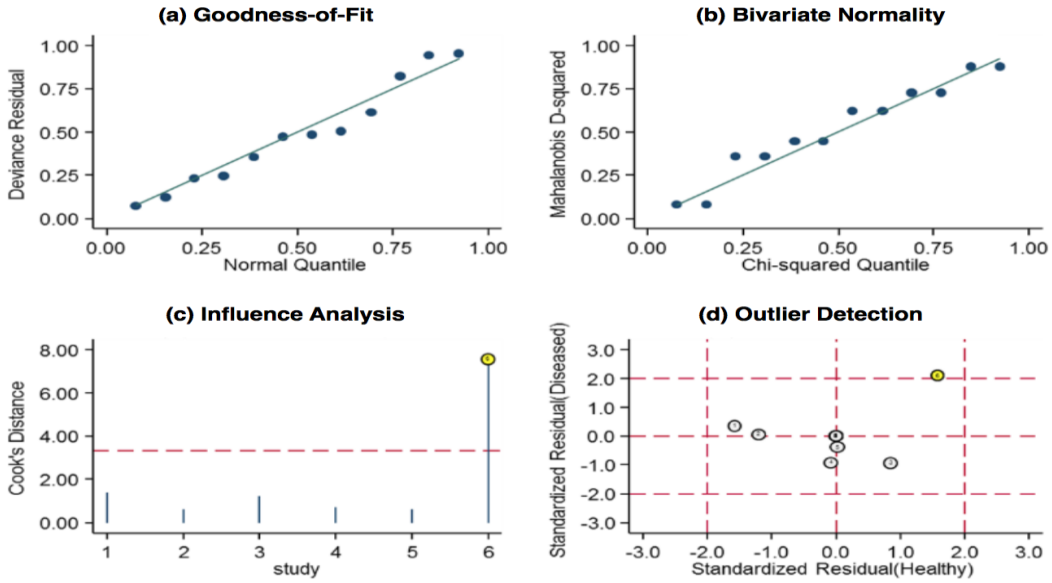


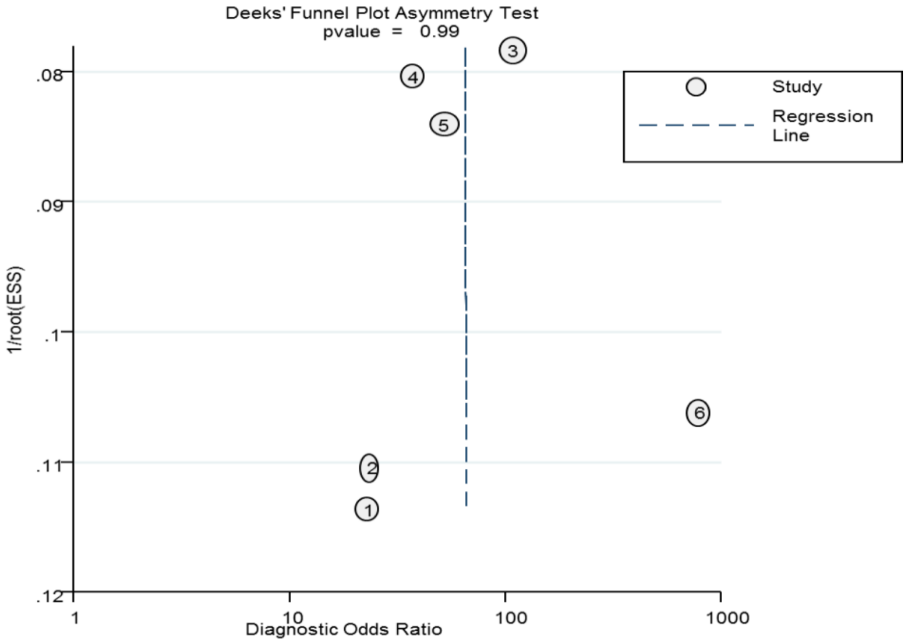
Table 2. Predictive Variables of Heterogeneity for Sensitivity and Specificity

Parameter	Category	N° of Studies	Sensitivity	p1	Specificity	p2
Sample Size	Yes	4	0.55 [0.42 - 0.67]	0.51	0.99 [0.98 - 1.00]	0.99
	No	2	0.60 [0.41 - 0.79]	.	0.94 [0.89 - 0.99]	.
Index	Yes	3	0.46 [0.37 - 0.56]	0.01	0.98 [0.97 - 1.00]	0.52
	No	3	0.67 [0.57 - 0.78]	.	0.97 [0.94 - 1.00]	.
P Reference	Yes	3	0.46 [0.37 - 0.56]	0.01	0.98 [0.97 - 1.00]	0.52
	No	3	0.67 [0.57 - 0.78]	.	0.97 [0.94 - 1.00]	.
QUADAS Score	Yes	3	0.46 [0.37 - 0.56]	0.01	0.98 [0.97 - 1.00]	0.52
	No	3	0.67 [0.57 - 0.78]	.	0.97 [0.94 - 1.00]	.
Language	Yes	3	0.46 [0.37 - 0.56]	0.01	0.98 [0.97 - 1.00]	0.52
	No	3	0.67 [0.57 - 0.78]	.	0.97 [0.94 - 1.00]	.

Table 3. Predictive Variables of Heterogeneity. Joint Model.

Parameter	Category	LRT Chi²	P	I²	I²lo	I²hi
Sample Size	Yes	7.62	0.02	74	42	100
	No	.	.	.	.	.
Index	Yes	9.29	0.01	78	53	100
	No	.	.	.	.	.
P Reference	Yes	9.29	0.01	78	53	100
	No	.	.	.	.	.
QUADAS Score	Yes	9.29	0.01	78	53	100
	No	.	.	.	.	.
Language	Yes	9.29	0.01	78	53	100
	No	.	.	.	.	.

Figure 7. Funnel plot with superimposed regression line



DISCUSSION

This study shows that the 7 scales evaluated in the 3 primary studies show homogeneous diagnostic performance with greater grouped values for specificity than for sensitivity and less variability in the specificity. However, the HADS scale shows different behavior without showing statistically significant differences in global measures of diagnostic accuracy. These results should be analyzed with caution given the significant level of heterogeneity between studies.

Grouped depression prevalence reported with the applied scales in the 3 studies was $29\%\ [IC95\%: 25\% - 33\%;\ T2 = 0.01]$ with a range of $24\% - 38\%$ without showing significant heterogeneity $(I^2 = 27.4\%, p = 0.23)$. The prevalence of any depressive episode according to the reference pattern scores showed values between 13.9% (40) and 25% (31). This behavior in the differences in depression prevalence according to screening tests and psychiatric interview is congruent with the systematic review data about transversal studies where the depression prevalence in patients with heart failure was 33.6% with self-report questionnaires and 19.3% using structured diagnostic interviews such as SCID (used in 3 included studies) (35). Nevertheless, in this meta-analysis, a small variability in the prevalence of major depressive episodes is observed, which can be explained by the clinical context, the way in which the measuring instrument was applied, language, HF clinical stage, sex or sample size.

Heterogeneity, referring to variability between studies, is a key point in the Diagnostic Tests Meta-Analysis (43). Heterogeneity can result from random probability, methodology analytic errors, and/or differences in study designs, protocol, inclusion and exclusion criteria, and diagnostic cut-off (44). It should be noted that DTA studies show greater heterogeneity than intervention studies due to the presence of the cut-off effect (43, 45). The cut-off effect can be explain by the fact that primary studies can use different cut-offs to define positive or negative outcomes from the test (46). The recommendation is that if there is evidence of significant heterogeneity, significant cut-off effect or effective outliers, a SROC curve should be constructed to analyze heterogeneity (43, 47-49).

In this study, I^2 and HSRCO curve visual inspection showed significant heterogeneity with a significant cut-off effect (49) and this was used to calculate a SROC curve using a hierarchical model according to Rutter and Gatsonis' recommendations (50). In fact, the HSROC or bivariate model must be used as standard methods in diagnostic accuracy for meta-analysis studies (38, 48, 49).

In the present case, it was a more appropriate to use a hierarchical model due to the fact that the studies included used different positivity cut-offs for diagnostic scales (38). Another cause to select a hierarchical model to evaluate heterogeneity is that Higgins and Thompson's inconsistence index (I^2) (51) cannot properly identify the variability between studies for dichotomous variables like those used in DTA studies and it can overestimate the found heterogeneity value (49, 52).

According to Leeflang, the authors of DTA systematic reviews, should investigate heterogeneity sources, rather than assess whether heterogeneity exists (38). To do this, it is recommended to carry out a subgroup analysis or a meta-regression analysis (53). Given the characteristics of this review, it was preferred to use a meta-regression analysis.

Bivariate or hierarchical models can be used to evaluate heterogeneity sources. Similarly, some of the groups can be removed from the global analysis (sensitivity analysis), or certain characteristics can be included as covariables in the meta-regression model (38, 43, 54, 55). Given that variables can be selected beforehand by the authors according to clinical and methodological criteria (54), we decided to include as covariables in meta-regression model (according to Cochrane Methods Working Group on Screening and Diagnostic tests recommendations): language, sample size, and 2 questions from QUADAS-2. According to the regression model used to evaluate the heterogeneity sources, it was observed that all the variables included beforehand showed behavior heterogeneity predicted variables. According to the regression model that we used to evaluate the sources of heterogeneity, it was observed that all variables included a priori behaved as predictors of heterogeneity. Nevertheless, these data should be interpreted with caution due to the reduced number of analyzed studies because the observed differences are based on the observation of one or two studies using subgroups. In such circumstances, findings can be coincidental or may be explained by other variables (49).

Strengths and limitations

The main strength of this review was its adherence to internationally recommended methods (38, 56, 57) to raise the investigation question, the identification and selection of eligible studies, the assessment of methodological quality and the risk of bias using QUADAS-2, and the use of strict methods for the quantitative analysis.

An important threat to validating a systematic review is the existence of publication biases. Publication biases occur if the studies with statistically non-significant outcomes are not published, leading to a possibly exaggerated grouped estimate in the systematic review (58, 59). The methods to identify publication biases are compromised in their reliability in diagnostic precision revisions; nevertheless, Deeks et al.'s method has shown less probability of error, which is why it is the favorite for this purpose (60). In this review, the probability of publication bias was reduced due to rigorousness in bibliographic searching and the application of Deeks' asymmetry test that showed the absence of publication biases ($p = 0.993$). On the other hand, the suitable fit of the Meta-regression model used in this study allowed a reliable evaluation of the heterogeneity sources and the identification of outliers.

Nevertheless, there are a number of limitations of the primary studies and in the revision same. First,

the small number of included studies and participants do not allow for significant generalizations.

The comparative precision of the instruments was mainly determined through indirect comparisons, that can lead to confusion due to differences between studies' characteristics and populations (61).

A different cut-off was used for each instrument and we used the optimal cut-off reported in each primary study. The selective report on optimal cut-off can introduce selection biases especially if the sample size is small, as it is in this case (38).

As indicated by the evaluation results using QUADAS-2 criteria, there are a number of methodological biases in most studies as well as key methodological details that have not been reported. However, QUADAS-2 evaluation shows low variability in the methodological quality and all the studies scored with low risk in most stages, except in Haworth et al. who evaluated HADS and GDS. In one of these studies (41), it was not reported whether patients with cognitive impairment were excluded. The latter is a factor that can hinder such patients from correctly understanding the index test and therefore may generate misleading results (62). Only one of these studies (41) established blinding in the application of the reference test and the gold standard. The second study (31) did not establish this, and the last did not report it (40). It is known that a lack of blinding can artificially increase the diagnostic performance on a test (63). In one of these studies (40), it was not reported whether the screening test was self-applied or hetero-applied, generating uncertainty regarding the real answer pattern from those interviewed, because the scales' scores can change significantly depending on whether they are self-applied or hetero-applied (64). In contrast, concern regarding applicability was null in the three studies.

Two variables that may behave as potential sources of heterogeneity were not included in this Meta-analysis: prevalence of MDD in this population and clinical setting (outpatient or inpatient). Moreover, the included studies only evaluated major depression and did not consider minor depression or degrees of severity of depressive disorder. It has been observed that both major and minor depression affect the clinical outcomes of patients with HF (35, 65-69).

Another limitation is the lack of cost-effectiveness analysis in the identification of MDD As such, whether cost also influences the false positives of the depression screening tools in HF is still unclear.

Clinical implications

Given the small number of studies and interviewed subjects, this result cannot be generalized easily to all the patients with depression risk in the population with HF. Due to the fact that in the three studies male participants predominated, this results cannot be generalized to women with depression where prevalence is greater than among men (70) including people with HF (71).

Given the broad diffusion and ease of use of scales as HADS, PHQ-9, GDS, the results of this systematic review confirm its clinical usefulness in the screening of depressive symptoms in patients with heart disease in both inpatient and outpatient settings.

Implications for research

Considering the methodological limitations in primary studies, future studies validating screening instruments should report, in sufficient detail, methodological aspects that allow the assessment of quality methodological criteria when applying QUADAS-2 in systematic reviews. More validation studies are needed to make direct comparisons between the best performing screening scales in patients with HF to really know which is superior.

Validation studies with larger samples are needed for the psychometric properties of scales such as CAT-D and PROMIS-D to show replicable results.

CONCLUSIONS

The psychometric performance of GDS-15, HADS-D, PHQ-9, CAT-D and PROMIS scales are similar in their screening of depressive symptomatology in patients with HF with a high level of specificity. The HADS-D scale has a modifying effect on the global efficacy of the evaluated scales and seems to have a slight advantage as a screening tool with respect to the others due to its higher sensitivity values. Validation studies that directly compare the depression screening scales in patients with HF are required to know which is superior.

AUTHOR'S CONTRIBUTIONS

CMCA: design, search strategy, studies review, data extraction, methodological quality assessment, statistical analysis, article writing. MR: search strategy, studies review, data extraction, methodological quality assessment, article writing. PAC: studies review, methodological quality assessment, statistical analysis, article writing. All authors have approved the final article.

CONFLICTS OF INTEREST: None to declare.

FUNDING: This research has not received specific aid from public sector agencies, the commercial sector or non-profit entities.

REFERENCIAS

1. Ege MR, Yilmaz N, Yilmaz MB. Depression and heart failure. Int J Cardiol. 2012;158(3):474.

2. Moudgil R, Haddad H. Depression in heart failure. Curr Opin Cardiol. 2013;28(2):249-58.

3. Joynt KE, Whellan DJ, O'Connor C M. Why is depression bad for the failing heart? A review of the mechanistic relationship between depression and heart failure. J Card Fail. 2004;10(3):258-71.

4. Konstam V, Moser DK, De Jong MJ. Depression and anxiety in heart failure. J Card Fail. 2005;11(6):455-63.

5. Albert NM, Fonarow GC, Abraham WT, Gheorghiade M, Greenberg BH, Nunez E, et al. Depression and clinical outcomes in heart failure: an OPTIMIZE-HF analysis. Am J Med. 2009;122(4):366-73.

6. O'Connor CM, Abraham WT, Albert NM, Clare R, Gattis Stough W, Gheorghiade M, et al. Predictors of mortality after discharge in patients hospitalized with heart failure: an analysis from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). Am Heart J. 2008;156(4):662-73.

7. Rohyans LM, Pressler SJ. Depressive symptoms and heart failure: examining the sociodemographic variables. Clin Nurse Spec. 2009;23(3):138-44.

8. Sin MK. Personal characteristics predictive of depressive symptoms in Hispanics with heart failure. Issues Ment Health Nurs. 2012;33(8):522-7.

9. Kato N, Kinugawa K, Seki S, Shiga T, Hatano M, Yao A, et al. Quality of life as an independent predictor for cardiac events and death in patients with heart failure. Circ J. 2011;75(7):1661-9.

10. Vaccarino V, Kasl SV, Abramson J, Krumholz HM. Depressive symptoms and risk of functional decline and death in patients with heart failure. J Am Coll Cardiol. 2001;38(1):199-205.

11. Sherwood A, Blumenthal JA, Trivedi R, Johnson KS, O'Connor CM, Adams KF, Jr., et al. Relationship of depression to death or hospitalization in patients with heart failure. Arch Intern Med. 2007;167(4):367-73.

12. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.

13. Jiang W, Kuchibhatla M, Clary GL, Cuffe MS, Christopher EJ, Alexander JD, et al. Relationship between depressive symptoms and long-term mortality in patients with heart failure. Am Heart J. 2007;154(1):102-8.

14. Lichtman JH, Bigger JT, Jr., Blumenthal JA, Frasure-Smith N, Kaufmann PG, Lesperance F, et al. Depression and coronary heart disease: recommendations for screening, referral, and

treatment: a science advisory from the American Heart Association Prevention Committee of the Council on Cardiovascular Nursing, Council on Clinical Cardiology, Council on Epidemiology and Prevention, and Interdisciplinary Council on Quality of Care and Outcomes Research: endorsed by the American Psychiatric Association. Circulation. 2008;118(17):1768-75.

15. Simon GE, Katon WJ, Lin EH, Rutter C, Manning WG, Von Korff M, et al. Cost-effectiveness of systematic depression treatment among people with diabetes mellitus. Arch Gen Psychiatry. 2007;64(1):65-72.

16. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®): American Psychiatric Pub; 2013.

17. Organización Mundial de la Salud. Clasificación Internacional de las Enfermedades (CIE). Trastornos Mentales y del Comportamiento. Criterios Diagnósticos de Investigación. 10 ed. Madrid1992.

18. Elderon L, Whooley MA. Depression and cardiovascular disease. Prog Cardiovasc Dis. 2013;55(6):511-23.

19. Rustad JK, Stern TA, Hebert KA, Musselman DL. Diagnosis and treatment of depression in patients with congestive heart failure: a review of the literature. Prim Care Companion CNS Disord. 2013;15(4).

20. Cully JA, Jimenez DE, Ledoux TA, Deswal A. Recognition and treatment of depression and anxiety symptoms in heart failure. Prim Care Companion J Clin Psychiatry. 2009;11(3):103-9.

21. Eisele M, Blozik E, Stork S, Trader JM, Herrmann-Lingen C, Scherer M. Recognition of depression and anxiety and their association with quality of life, hospitalization and mortality in primary care patients with heart failure - study protocol of a longitudinal observation study. BMC Fam Pract. 2013;14:180.

22. Thombs BD, de Jonge P, Coyne JC, Whooley MA, Frasure-Smith N, Mitchell AJ, et al. Depression screening and patient outcomes in cardiovascular care: a systematic review. JAMA. 2008;300(18):2161-71.

23. U. S. Preventive Services Task Force. Screening for depression: recommendations and rationale. Ann Intern Med. 2002;136(10):760-4.

24. Norra C, Skobel EC, Arndt M, Schauerte P. High impact of depression in heart failure: early diagnosis and treatment options. Int J Cardiol. 2008;125(2):220-31.

25. . !!! INVALID CITATION !!! .

26. Freedland KE, Rich MW, Skala JA, Carney RM, Davila-Roman VG, Jaffe AS. Prevalence of depression in hospitalized patients with congestive heart failure. Psychosom Med. 2003;65(1):119-28.

27. Lesman-Leegte I, Jaarsma T, Sanderman R, Linssen G, van Veldhuisen DJ. Depressive symptoms are prominent among elderly hospitalised heart failure patients. Eur J Heart Fail. 2006;8(6):634-40.

28. Subramanian U, Weiner M, Gradus-Pizlo I, Wu J, Tu W, Murray MD. Patient perception and provider assessment of severity of heart failure as predictors of hospitalization. Heart Lung. 2005;34(2):89-98.

29. Junger J, Schellberg D, Muller-Tasch T, Raupp G, Zugck C, Haunstetter A, et al. Depression increasingly predicts mortality in the course of congestive heart failure. Eur J Heart Fail. 2005;7(2):261-7.

30. Ski CF, Thompson DR, Hare DL, Stewart AG, Watson R. Cardiac Depression Scale: Mokken scaling in heart failure patients. Health Qual Life Outcomes. 2012;10:141.

31. Haworth JE, Moniz-Cook E, Clark AL, Wang M, Cleland JG. An evaluation of two self-report screening measures for mood in an out-patient chronic heart failure population. Int J Geriatr Psychiatry. 2007;22(11):1147-53.

32. Holzapfel N, Zugck C, Muller-Tasch T, Lowe B, Wild B, Schellberg D, et al. Routine screening for depression and quality of life in outpatients with congestive heart failure. Psychosomatics. 2007;48(2):112-6.

33. Hammash MH, Hall LA, Lennie TA, Heo S, Chung ML, Lee KS, et al. Psychometrics of the PHQ-9 as a

measure of depressive symptoms in patients with heart failure. Eur J Cardiovasc Nurs. 2013;12(5):446-53.

34. Sorensenf C, Friis-Hasche E, Haghfelt T, Bech P. Postmyocardial infarction mortality in relation to depression: a systematic critical review. Psychother Psychosom. 2005;74(2):69-80.

35. Rutledge T, Reis VA, Linke SE, Greenberg BH, Mills PJ. Depression in heart failure a meta-analytic review of prevalence, intervention effects, and associations with clinical outcomes. J Am Coll Cardiol. 2006;48(8):1527-37.

36. Delville CL, McDougall G. A systematic review of depression in adults with heart failure: instruments and incidence. Issues Ment Health Nurs. 2008;29(9):1002-17.

37. Bakkalbasi N, Bauer K, Glover J, Wang L. Three options for citation tracking: Google Scholar, Scopus and Web of Science. Biomed Digit Libr. 2006;3:7.

38. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working G. Systematic reviews of diagnostic test accuracy. Ann Intern Med. 2008;149(12):889-97.

39. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

40. Fischer HF, Klug C, Roeper K, Blozik E, Edelmann F, Eisele M, et al. Screening for mental disorders in heart failure patients using computer-adaptive tests. Qual Life Res. 2014;23(5):1609-18.

41. Poole NA, Morgan JF. Validity and reliability of the Hospital Anxiety and Depression Scale in a hypertrophic cardiomyopathy clinic: the HADS in a cardiomyopathy population. Gen Hosp Psychiatry. 2006;28(1):55-8.

42. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg. 2010;8(5):336-41.

43. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005;9(12):1-113, iii.

44. Jones CM, Ashrafian H, Darzi A, Athanasiou T. Guidelines for diagnostic tests and diagnostic accuracy in surgical research. J Invest Surg. 2010;23(1):57-65.

45. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-90.

46. Charoensawat S, Bohning W, Bohning D, Holling H. Meta-analysis and meta-modelling for diagnostic problems. BMC Med Res Methodol. 2014;14:56.

47. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med. 2009;28(21):2653-68.

48. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol. 2008;61(11):1095-103.

49. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 09 0 London: The Cochrane Collaboration. 2010.

50. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865-84.

51. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60.

52. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. Stat Med. 2014;33(16):2701-17.

53. Lee J, Kim KW, Choi SH, Huh J, Park SH. Systematic Review and Meta-Analysis of Studies Evaluating Diagnostic Test Accuracy: A Practical Review for Clinical Researchers-Part II. Statistical Methods of Meta-Analysis. Korean J Radiol. 2015;16(6):1188-96.

54. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med. 2002;21(11):1525-37.

55. Naaktgeboren CA, van Enst WA, Ochodo EA, de Groot JA, Hooft L, Leeflang MM, et al. Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies. J Clin Epidemiol. 2014;67(11):1200-9.

56. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.

57. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4:1.

58. Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. Int J Epidemiol. 2002;31(1):88-95.

59. van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. BMC Med Res Methodol. 2014;14:70.

60. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005;58(9):882-93.

61. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med. 2013;158(7):544-54.

62. Morrell CJ, Curran S, Topping A, Shaik K, Muthukrishnan V, Stephenson J. Identification of depressive disorder among older people in care homes - a feasibility study. Prim Health Care Res Dev. 2011;12(3):255-65.

63. Noyes K, Liu H, Lyness JM, Friedman B. Medicare beneficiaries with depression: comparing diagnoses in claims data with the results of screening. Psychiatr Serv. 2011;62(10):1159-66.

64. Fine TH, Contractor AA, Tamburrino M, Elhai JD, Prescott MR, Cohen GH, et al. Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I major depression module. J Affect Disord. 2013;150(3):1001-7.

65. Fan H, Yu W, Zhang Q, Cao H, Li J, Wang J, et al. Depression after heart failure and risk of cardiovascular and all-cause mortality: a meta-analysis. Prev Med. 2014;63:36-42.

66. Freedland KE, Hesseler MJ, Carney RM, Steinmeyer BC, Skala JA, Davila-Roman VG, et al. Major Depression and Long-Term Survival of Patients With Heart Failure. Psychosom Med. 2016;78(8):896-903.

67. Gathright EC, Goldstein CM, Josephson RA, Hughes JW. Depression increases the risk of mortality in patients with heart failure: A meta-analysis. J Psychosom Res. 2017;94:82-9.

68. Kato N, Kinugawa K, Yao A, Hatano M, Shiga T, Kazuma K. Relationship of depressive symptoms with hospitalization and death in Japanese patients with heart failure. J Card Fail. 2009;15(10):912-9.

69. Schiffer AA, Pelle AJ, Smith OR, Widdershoven JW, Hendriks EH, Pedersen SS. Somatic versus cognitive symptoms of depression as predictors of all-cause mortality and health status in chronic heart failure. J Clin Psychiatry. 2009;70(12):1667-73.

70. Lopez Molina MA, Jansen K, Drews C, Pinheiro R, Silva R, Souza L. Major depressive disorder symptoms in male and female young adults. Psychol Health Med. 2014;19(2):136-45.

71. Gottlieb SS, Khatta M, Friedmann E, Einbinder L, Katzen S, Baker B, et al. The influence of age, gender, and race on the prevalence of depression in heart failure patients. J Am Coll Cardiol. 2004;43(9):1542-9.