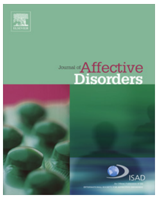
Journal of Affective Disorders 172 (2015) 337–346

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad



Review

Screening for bipolar spectrum disorders: A comprehensive meta-analysis of accuracy studies

André F. Carvalho [a,*], Yemisi Takwoingi [b], Paulo Marcelo G. Sales [a], Joanna K. Soczynska [c], Cristiano A. Köhler [d], Thiago H. Freitas [a], João Quevedo [e,f], Thomas N. Hyphantis [g], Roger S. McIntyre [c,h], Eduard Vieta [I]

[a] Translational Psychiatry Research Group, Faculty of Medicine, Federal University of Ceara, Fortaleza, CE, Brazil
[b] Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK
[c] Mood Disorders Psychopharmacology Unit, University of Toronto, Toronto, ON, Canada
[d] Memory Research Laboratory, Brain Institute (ICe), Federal University of Rio Grande do Norte (UFRN), Natal, RN, Brazil
[e] Laboratory of Neurosciences, Graduate Program in Health Sciences, Health Sciences Unit, University of Southern Santa Catarina, Criciúma, SC, Brazil
[f] Center for Experimental Models in Psychiatry, Department of Psychiatry and Behavioral Sciences, The University of Texas Medical School at Houston, Houston, TX, USA
[g] Department of Psychiatry, University of Ioaninna, Ioaninna, Greece
[h] Departments of Psychiatry and Pharmacology, University of Toronto, Toronto, ON, Canada
[I] Bipolar Disorders Unit, Clinical Institute of Neurosciences, Hospital Clinic, IDIBAPS, University of Barcelona, CIBERSAM, Barcelona, Catalonia, Spain

ARTICLE INFO

Article history:
Received 2 October 2014
Accepted 14 October 2014
Available online 23 October 2014

Keywords:
Bipolar disorder
Screening
Accuracy studies
Systematic review
Meta-analysis

ABSTRACT

Background: Bipolar spectrum disorders are frequently under-recognized and/or misdiagnosed in various settings. Several influential publications recommend the routine screening of bipolar disorder. A systematic review and meta-analysis of accuracy studies for the bipolar spectrum diagnostic scale (BSDS), the hypomania checklist (HCL-32) and the mood disorder questionnaire (MDQ) were performed.
Methods: The Pubmed, EMBASE, Cochrane, PsycINFO and SCOPUS databases were searched. Studies were included if the accuracy properties of the screening measures were determined against a DSM or ICD-10 structured diagnostic interview. The QUADAS-2 tool was used to rate bias.
Results: Fifty three original studies met inclusion criteria (N=21,542). At recommended cutoffs, summary sensitivities were 81%, 66% and 69%, while specificities were 67%, 79% and 86% for the HCL-32, MDQ, and BSDS in psychiatric services, respectively. The HCL-32 was more accurate than the MDQ for the detection of type II bipolar disorder in mental health care centers (P=0.018). At a cutoff of 7, the MDQ had a summary sensitivity of 43% and a summary specificity of 95% for detection of bipolar disorder in primary care or general population settings.
Limitations: Most studies were performed in mental health care settings. Several included studies had a high risk of bias.
Conclusions: Although accuracy properties of the three screening instruments did not consistently differ in mental health care services, the HCL-32 was more accurate than the MDQ for the detection of type II BD. More studies in other settings (for example, in primary care) are necessary.

Contents

* Corresponding author. Tel./fax: +558532617227.
E-mail address: andrefc7@terra.com.br (A.F. Carvalho).

http://dx.doi.org/10.1016/j.jad.2014.10.024
0165-0327/© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The diagnosis of bipolar disorders in most circumstances is not straightforward and requires a careful assessment of its longitudinal course. Almost three-third of individuals with bipolar disorders report having received a misdiagnosis at least once, while a proper diagnosis takes on average 10 years from the initiation of affective symptoms (Drancourt et al., 2013; Lish et al., 1994). Evidences also indicate that bipolar disorder is prevalent and frequently under-recognized in primary care (Cerimele et al., 2014; Culpepper, 2014). Furthermore, depressive symptoms and episodes more frequently predominate in the longitudinal course of bipolar disorders (Judd et al., 2003); this results in a significant proportion of individuals with BD being misdiagnosed as having unipolar depression (Hirschfeld and Vornik, 2004). These patients misdiagnosed as having major depressive disorder are more likely to receive antidepressant monotherapy (Matza et al., 2005) which may result in manic switches, cycle acceleration, and possibly heightened suicidality (Bond et al., 2008; Ghaemi et al., 2004; Undurraga et al., 2012).

The use of self-report screening instruments for bipolar disorder that are both time- and cost-effective may aid in the timely recognition of this illness. In the last several years four self-report questionnaires have been developed to screen for bipolar spectrum disorders, namely the mood disorders questionnaire (MDQ) (Hirschfeld et al., 2000), the bipolar spectrum diagnostic scale (BSDS) (Ghaemi et al., 2005), the hypomanic checklist (HCL-32) (Angst et al., 2005) and the mood swings questionnaire/survey (MSQ/MSS) (Parker et al., 2008; Parker et al., 2006). These screening tools are readily available for clinical use. Briefly, the MDQ screens for a lifetime history of (hypo) mania with 13 yes/no questions reflecting DSM-IV criteria. These questions are followed by a single yes/no question asking whether the symptoms clustered in the same period. The final question evaluates the level of impairment resulting from the symptoms. The MDQ developers recommended a cut-off score of seven endorsed symptoms that co-occurred and caused at least moderate impairment. (Hirschfeld et al., 2000) The BSDS consists of two parts. The first part is a paragraph containing 19 statements describing several manifestations of bipolar disorder. Each affirmatively checked sentence is counted as 1 point. The second part of the BSDS is a single multiple-choice question asking respondents how well the paragraph describes their behavior (very well or almost perfect – 6 points; fairly well – 4 points; to some degree but not in most respects – 2 points; not really at all – 0 points). In the initial study, a cut-off point of 13 yielded the best balance of sensitivity/specificity (Ghaemi et al., 2005). In the HCL-32, after a brief introduction, the respondent is instructed to think of a period when he/she was in a “high” state and answer 32 yes/no questions about their mood and behavior during that period. Each ‘yes’ response is scored 1, whereas each ‘no’ answer is scored as 0. In the initial study, the authors suggested a cut-off score of 14 (Angst et al., 2005).

Notwithstanding several influential publications recommend the routine screening in clinical practice (Anderson et al., 2012; Chessick and Dimidjian, 2010; Frye, 2011; Loganathan et al., 2010), concerns have been raised regarding the validity and applicability of these screening tools (Phelps and Ghaemi, 2006; Zimmerman, 2012; Zimmerman et al., 2010). Phelps and Ghaemi (2006) used previously published data on sensitivity and specificity of the MDQ and BSDS to estimate positive and negative predictive values at varying prevalence levels using Bayesian statistical concepts. At lower prevalence or low prior clinical probability (for example, in primary care), high negative predictive values were verified indicating that both instruments effectively rule out bipolar disorders. However, in these contexts the positive predictive value significantly dropped resulting in a higher number of ‘false positives’.

The BSDS, HCL-32 and MDQ have been the most extensively investigated screening tools for bipolar spectrum disorders in accuracy studies and epidemiological surveys. Therefore, the overarching aims of this report were to perform a systematic review and meta-analysis to evaluate and compare the diagnostic accuracy of these three screening tools in different clinical settings. Our secondary objective was to investigate the effect of pre-defined potential sources of heterogeneity on estimates of test performance.

2. Method

2.1. Search strategy and selection of studies

Studies were identified through three methods. First, we conducted comprehensive computerized literature searches in five bibliographical databases – MEDLINE, EMBASE, Cochrane CENTRAL, PsycInfo and SCOPUS – from inception to January 9th, 2014. Search strings are provided in the Supplementary material S1 that accompanies the online edition of this article. Second, this search strategy was augmented through tracking citations of included articles in Google Scholar. Finally, references of relevant reviews were examined to identify potentially relevant studies (see references in the Supplementary material). No language restrictions were applied. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement (Liberati et al., 2009) and the Cochrane Handbook for Diagnostic Test Accuracy Reviews (Macaskill et al., 2010). Two investigators screened title/abstracts for potential eligibility. Disagreements were resolved through consensus. References selected for full-text review were evaluated by two independent raters. Disagreements were resolved by discussion.

We included studies in which the diagnostic accuracy of the MDQ, the BSDS or the HCL-32 was investigated in general adult psychiatric populations, primary care or in community-derived samples with validated structured psychiatric interviews for the DSM-IV or DSM-IV-TR as reference standards. Studies were excluded if they: involved perinatal/postpartum specific populations; involved child and adolescent samples; did not use a validated structured interview as reference standard; did not provide data for deriving a two-by-two table (FP- false positives; FN- false negatives; TN – true negatives and TP – true positives) even after corresponding authors were contacted for additional data.

2.2. Data extraction and quality assessment

Using a structured spreadsheet, data on the following characteristics were extracted: author, publication year, study design, setting, sample size, reference standard, version of the screening instrument and data for two-by-two tables. We appraised the quality of included studies by using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool (Whiting et al., 2011). Briefly, the QUADAS-2 tool consists of four domains: patient selection, index test, reference standard, and flow and timing. The four domains are assessed for risk of bias and the first three domains are also assessed for concerns regarding applicability. We developed guidance tailored to this review on how to appraise and interpret each signaling question within the domains. Two authors (JKS and THF) extracted data and assessed each included study according to the QUADAS-2 criteria. Inter-rater agreement of the QUADAS-2 assessment was excellent (overall Kappa value$=$0.81). Disagreements were resolved through consensus.

2.3. Meta-analyses

To explore variation in diagnostic accuracy between studies, we plotted estimates of the observed sensitivities and specificities for each test in forest plots and in receiver-operating characteristic (ROC) space using data for a single cut-off from each study. Each summary ROC (SROC) curve shows the expected trade-off between sensitivity and specificity across studies using different cut-off scores for each instrument. Sensitivity refers to a test's ability to correctly identify individuals with a given disorder, and is computed as the number of individuals with the disorder that were classified as test positives (i.e. TP) divided by the total number of individuals with the disorder. Specificity refers to a test's ability to identify those without the disorder, and is computed as the number of individuals without the disorder who were classified as test negatives (i.e. TN) divided by the total number of individuals without the disorder. We analyzed data from studies conducted in different clinical settings separately (categorized as mental health care settings and primary care/general population settings). We considered bipolar disorder in general and then performed separate analyses for bipolar disorder type II and bipolar disorder not otherwise specified (NOS) where data were available.

Since studies used different cut-offs to define a positive screen for each test, whenever sufficient data were available we performed meta-analyses using hierarchical summary ROC (HSROC) models. The HSROC model includes random effects parameters that allow for variation in accuracy and cut-off between studies. The model also includes a shape parameter that allows accuracy to vary with cut-offs thus enabling asymmetry in the shape of the SROC curve. If a study reported sensitivity and specificity at multiple cut-offs, the optimum cut-off (as defined by the authors based on the most adequate balance between sensitivity and specificity) was selected. Thus, only a pair of sensitivity and specificity from each study was included in a meta-analysis. To enable estimation of the average operating point (summary sensitivity and specificity) for each test at a specific cut-off, we restricted meta-analysis to studies that reported the cut-off. Whenever few studies were available, we simplified the HSROC model by assuming a symmetrical SROC curve or fixed effects for the accuracy and/or threshold parameters.

We compared the diagnostic accuracy of the three instruments obtained from all included studies (indirect comparison), and then performed additional analyses restricted to studies that made head-to-head comparisons (i.e. applied two of the instruments to the same participants). We made test comparisons by adding a covariate for test type to the HSROC model to assess the effect of test type on the accuracy, cut-off and/or shape parameters of the model. Since summary sensitivities and specificities are only clinically interpretable when the studies included in a meta-analysis use a common cut-off, we estimated sensitivity at points on the SROC curves corresponding to the lower quartile, median and upper quartile of the specificities observed in the studies included in the meta-analysis. In addition, whenever the estimated SROC curves had the same shape, we calculated the relative diagnostic odds ratio (RDOR) as a summary of the relative accuracy of two screening instruments. To assess the statistical significance of differences in test accuracy, likelihood ratio tests were used for comparisons of models with and without covariate terms.

To investigate heterogeneity in the diagnostic accuracy of each instrument, we added potential sources of heterogeneity as a covariate to the HSROC model (meta-regression). We a priori considered the following variables: language of the instrument (Asian versus non-Asian); two signaling questions from the patient selection domain of the QUADAS-2 tool that reflect patient recruitment ('Was a consecutive or random sample of patients enrolled?') and study design ('Was a case control design avoided?'). Finally, the percentage of bipolar disorder type II/NOS in each included study (categorized as either $<$ or $\geq$ median values of included studies) was also investigated as a potential source of heterogeneity in test performance for detection of bipolar disorder in general.

All HSROC analyses were performed using the NLMIXED procedure in the SAS software (version 9.2; SAS Institute, Cary, North Carolina) (Macaskill, 2004). We used Review Manager (version 5.2; Copenhagen; The Nordic Cochrane Center, The Cochrane Collaboration, 2012) to generate forest plots and SROC plots. Significance level was set at $\alpha=0.05$.

3. Results

3.1. Selection of studies

Fig. 1 summarizes the study selection process. After examining a total of 541 titles and abstracts (371 after removal of duplicates), we selected 84 unique references for further consideration. We excluded 31 of the retrieved articles (reasons for exclusion are

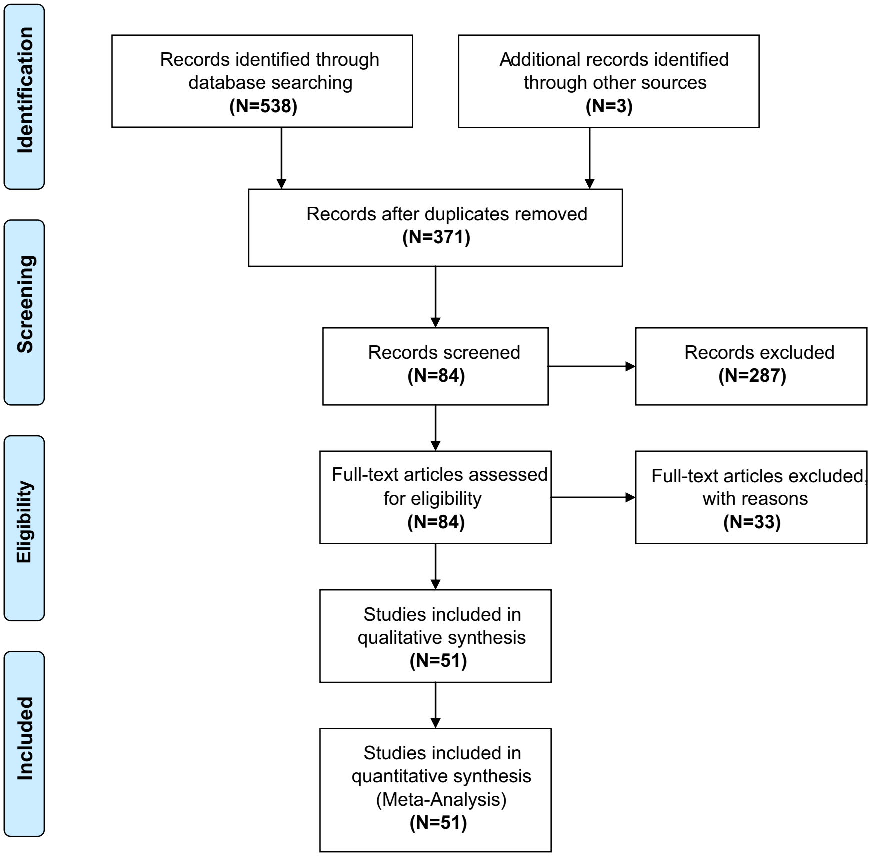

Identification
Records identified through database searching (N=538)
Additional records identified through other sources (N=3)
Screening
Records after duplicates removed (N=371)
Records screened (N=84)
Records excluded (N=287)
Eligibility
Full-text articles assessed for eligibility (N=84)
Full-text articles excluded, with reasons (N=33)
Included
Studies included in qualitative synthesis (N=51)
Studies included in quantitative synthesis (Meta-Analysis) (N=51)

Fig. 1. PRISMA flowchart of study selection.

provided in Supplementary Table S1). A total of 53 original studies (5566 bipolar disorder cases; 21,543 patients) met our inclusion criteria. Descriptive characteristics of the included studies are provided in Supplementary Table S2.

3.2. Assessment of bias of included studies

Supplementary Table S3 shows the overall risk of bias and applicability concerns for the 53 included studies. A large proportion of studies (19 studies; 35.8%) showed a high risk of bias in the 'patient selection domain' and also gave high applicability concern in the same domain; 11 of the studies used a case-control design and did not enroll a consecutive sample of patients. Overall, most studies had a high risk of bias in at least one QUADAS-2 domain (29 studies; 54.7%).

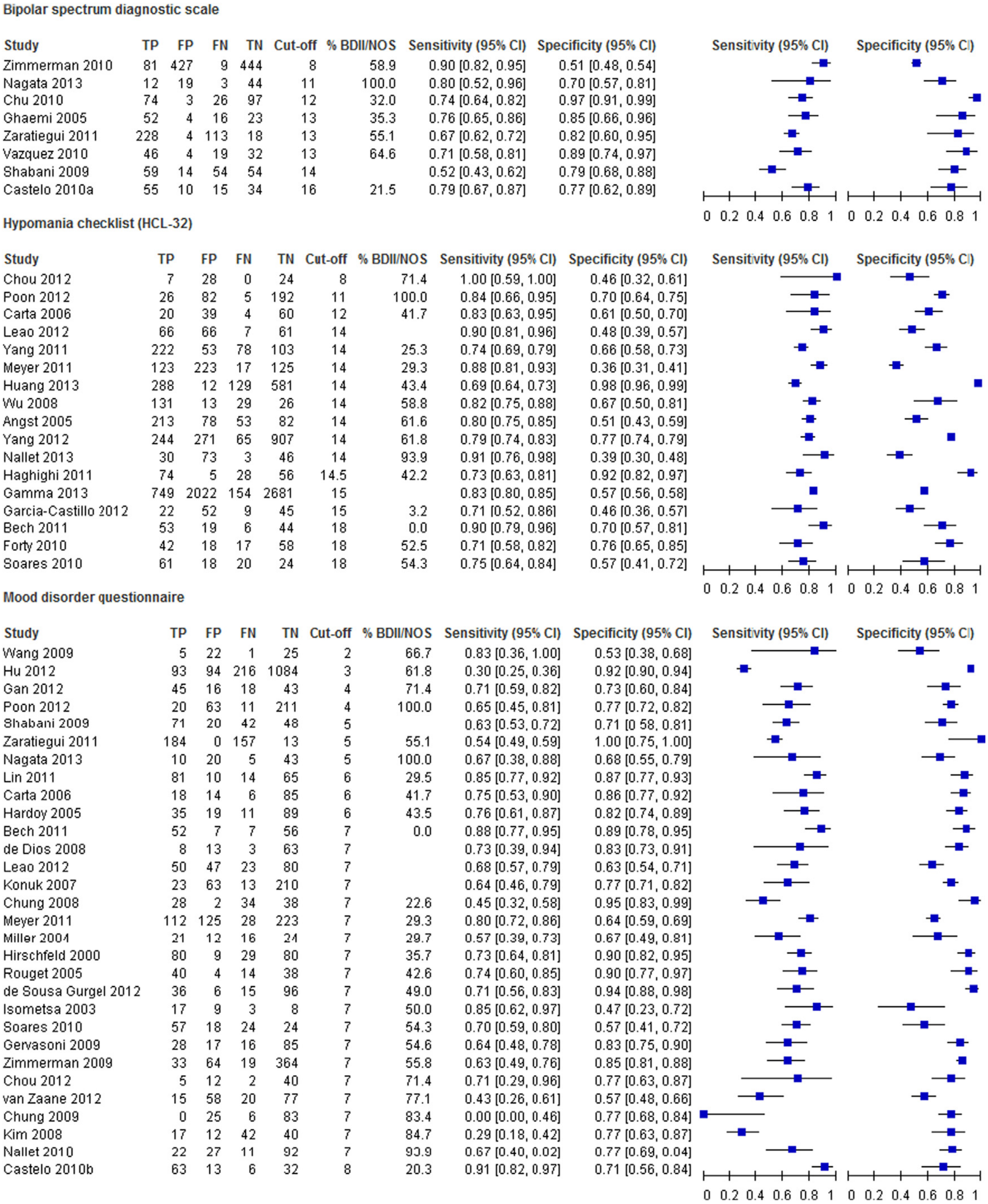
3.3. Detection of any type of bipolar disorder (BD type I, BD type II or BD NOS)

3.3.1. Comparison of the BSDS, HCL-32 and MDQ for the detection of bipolar disorder in the mental health care setting

The studies reported different optimal cut-off scores for each of the screening instruments (see Supplementary Table S4). Overall, 44 studies (5021 cases; 17,451 participants) were performed in mental health services (Fig. 2). Table 1 summarizes the sensitivities and specificities for the BSDS, MDQ and HCL-32 at specific cut-offs for which data were available for a separate meta-analysis of each instrument at a common cut-off. At the developer recommended cut-offs of 14, 13, and 7 for the HCL-32, BSDS and MDQ respectively, the summary sensitivities were 81% (95% CI 77–85%), 69% (95% CI 63–74%) and 66% (95% CI 57–73%); the corresponding summary specificities were 67% (95% CI 47–82%), 79% (95% CI 72–84%) and 86% (95% CI 74–93%).

Using all available studies in an indirect comparison (i.e. unrestricted to head-to-head studies), we compared the test performance of the MDQ (30 studies), the BSDS (8 studies) and the HCL-32 (17 studies). The shape of the SROC curves significantly differed ($p=0.002$) as well as the accuracy of the screening instruments ($p=0.029$). Because the shape of the SROC curve for each instrument was different and asymmetric, this implies that the accuracy of each instrument varies with cut-off. Fig. 3 presents the SROC curves for the three instruments. The BSDS curve is consistently above the MDQ curve in the region containing most of the observed data. The HCL-32 curve is above the MDQ and BSDS curve at higher values of specificity, but the curve then crosses both the MDQ and the BSDS curves and accuracy is lower at lower values of specificity. This is also evident in Supplementary Table S5, which shows the sensitivities estimated from the curves at quartiles of the observed specificities in included studies. Using quartiles of the observed prevalence from the included studies, the table also shows the clinical implications of using each of the instruments in a hypothetical cohort of 100 patients. For example, out of a cohort of 100 patients with a bipolar disorder prevalence of 18%, and assuming a specificity of 77%, the sensitivities of MDQ, HCL-32 and BSDS of 70%, 78% and 78% would miss 4, 4, and 5 cases respectively, while 19 of those without bipolar disorder would be false positives.

In direct comparisons, three studies compared the BSDS (469 cases; 622 patients) to the MDQ (469 cases; 613 patients). Cut-offs differed between studies for the BSDS, and the direction of the differences in sensitivity and specificity were inconsistent. Two of the studies reported higher sensitivity and contrasting specificity for the BSDS (at cut-offs 11 and 13) compared to the MDQ at a cut-off of 5; the third study reported lower sensitivity and higher specificity at a cut-off of 14 for the BSDS and a cut-off of 5 for the MDQ. Eight studies (448 cases; 1572 patients) directly compared MDQ and HCL-32 (Supplementary Fig. S3). Despite differences in cut-offs, the results



Bipolar spectrum diagnostic scale

Study	TP	FP	FN	TN	Cut-off	% BDII/NOS	Sensitivity (95% CI)	Specificity (95% CI)
Zimmerman 2010	81	427	9	444	8	58.9	0.90 [0.82, 0.95]	0.51 [0.48, 0.54]
Nagata 2013	12	19	3	44	11	100.0	0.80 [0.52, 0.96]	0.70 [0.57, 0.81]
Chu 2010	74	3	26	97	12	32.0	0.74 [0.64, 0.82]	0.97 [0.91, 0.99]
Ghaemi 2005	52	4	16	23	13	35.3	0.76 [0.65, 0.86]	0.85 [0.66, 0.96]
Zaratiegui 2011	228	4	113	18	13	55.1	0.67 [0.62, 0.72]	0.82 [0.60, 0.95]
Vazquez 2010	46	4	19	32	13	64.6	0.71 [0.58, 0.81]	0.89 [0.74, 0.97]
Shabani 2009	59	14	54	54	14		0.52 [0.43, 0.62]	0.79 [0.68, 0.88]
Castelo 2010a	55	10	15	34	16	21.5	0.79 [0.67, 0.87]	0.77 [0.62, 0.89]

Hypomania checklist (HCL-32)

Study	TP	FP	FN	TN	Cut-off	% BDII/NOS	Sensitivity (95% CI)	Specificity (95% CI)
Chou 2012	7	28	0	24	8	71.4	1.00 [0.59, 1.00]	0.46 [0.32, 0.61]
Poon 2012	26	82	5	192	11	100.0	0.84 [0.66, 0.95]	0.70 [0.64, 0.75]
Carta 2006	20	39	4	60	12	41.7	0.83 [0.63, 0.95]	0.61 [0.50, 0.70]
Leao 2012	66	66	7	61	14		0.90 [0.81, 0.96]	0.48 [0.39, 0.57]
Yang 2011	222	53	78	103	14	25.3	0.74 [0.69, 0.79]	0.66 [0.58, 0.73]
Meyer 2011	123	223	17	125	14	29.3	0.88 [0.81, 0.93]	0.36 [0.31, 0.41]
Huang 2013	288	12	129	581	14	43.4	0.69 [0.64, 0.73]	0.98 [0.96, 0.99]
Wu 2008	131	13	29	26	14	58.8	0.82 [0.75, 0.88]	0.67 [0.50, 0.81]
Angst 2005	213	78	53	82	14	61.6	0.80 [0.75, 0.85]	0.51 [0.43, 0.59]
Yang 2012	244	271	65	907	14	61.8	0.79 [0.74, 0.83]	0.77 [0.74, 0.79]
Nallet 2013	30	73	3	46	14	93.9	0.91 [0.76, 0.98]	0.39 [0.30, 0.48]
Haghighi 2011	74	5	28	56	14.5	42.2	0.73 [0.63, 0.81]	0.92 [0.82, 0.97]
Gamma 2013	749	2022	154	2681	15		0.83 [0.80, 0.85]	0.57 [0.56, 0.58]
Garcia-Castillo 2012	22	52	9	45	15	3.2	0.71 [0.52, 0.86]	0.46 [0.36, 0.57]
Bech 2011	53	19	6	44	18	0.0	0.90 [0.79, 0.96]	0.70 [0.57, 0.81]
Forty 2010	42	18	17	58	18	52.5	0.71 [0.58, 0.82]	0.76 [0.65, 0.85]
Soares 2010	61	18	20	24	18	54.3	0.75 [0.64, 0.84]	0.57 [0.41, 0.72]

Mood disorder questionnaire

Study	TP	FP	FN	TN	Cut-off	% BDII/NOS	Sensitivity (95% CI)	Specificity (95% CI)
Wang 2009	5	22	1	25	2	66.7	0.83 [0.36, 1.00]	0.53 [0.38, 0.68]
Hu 2012	93	94	216	1084	3	61.8	0.30 [0.25, 0.36]	0.92 [0.90, 0.94]
Gan 2012	45	16	18	43	4	71.4	0.71 [0.59, 0.82]	0.73 [0.60, 0.84]
Poon 2012	20	63	11	211	4	100.0	0.65 [0.45, 0.81]	0.77 [0.72, 0.82]
Shabani 2009	71	20	42	48	5		0.63 [0.53, 0.72]	0.71 [0.58, 0.81]
Zaratiegui 2011	184	0	157	13	5	55.1	0.54 [0.49, 0.59]	1.00 [0.75, 1.00]
Nagata 2013	10	20	5	43	5	100.0	0.67 [0.38, 0.88]	0.68 [0.55, 0.79]
Lin 2011	81	10	14	65	6	29.5	0.85 [0.77, 0.92]	0.87 [0.77, 0.93]
Carta 2006	18	14	6	85	6	41.7	0.75 [0.53, 0.90]	0.86 [0.77, 0.92]
Hardoy 2005	35	19	11	89	6	43.5	0.76 [0.61, 0.87]	0.82 [0.74, 0.89]
Bech 2011	52	7	7	56	7	0.0	0.88 [0.77, 0.95]	0.89 [0.78, 0.95]
de Dios 2008	8	13	3	63	7		0.73 [0.39, 0.94]	0.83 [0.73, 0.91]
Leao 2012	50	47	23	80	7		0.68 [0.57, 0.79]	0.63 [0.54, 0.71]
Konuk 2007	23	63	13	210	7		0.64 [0.46, 0.79]	0.77 [0.71, 0.82]
Chung 2008	28	2	34	38	7	22.6	0.45 [0.32, 0.58]	0.95 [0.83, 0.99]
Meyer 2011	112	125	28	223	7	29.3	0.80 [0.72, 0.86]	0.64 [0.59, 0.69]
Miller 2004	21	12	16	24	7	29.7	0.57 [0.39, 0.73]	0.67 [0.49, 0.81]
Hirschfeld 2000	80	9	29	80	7	35.7	0.73 [0.64, 0.81]	0.90 [0.82, 0.95]
Rouget 2005	40	4	14	38	7	42.6	0.74 [0.60, 0.85]	0.90 [0.77, 0.97]
de Sousa Gurgel 2012	36	6	15	96	7	49.0	0.71 [0.56, 0.83]	0.94 [0.88, 0.98]
Isometsa 2003	17	9	3	8	7	50.0	0.85 [0.62, 0.97]	0.47 [0.23, 0.72]
Soares 2010	57	18	24	24	7	54.3	0.70 [0.59, 0.80]	0.57 [0.41, 0.72]
Gervasoni 2009	28	17	16	85	7	54.6	0.64 [0.48, 0.78]	0.83 [0.75, 0.90]
Zimmerman 2009	33	64	19	364	7	55.8	0.63 [0.49, 0.76]	0.85 [0.81, 0.88]
Chou 2012	5	12	2	40	7	71.4	0.71 [0.29, 0.96]	0.77 [0.63, 0.87]
van Zaane 2012	15	58	20	77	7	77.1	0.43 [0.26, 0.61]	0.57 [0.48, 0.66]
Chung 2009	0	25	6	83	7	83.4	0.00 [0.00, 0.46]	0.77 [0.68, 0.84]
Kim 2008	17	12	42	40	7	84.7	0.29 [0.18, 0.42]	0.77 [0.63, 0.87]
Nallet 2010	22	27	11	92	7	93.9	0.67 [0.40, 0.02]	0.77 [0.69, 0.04]
Castelo 2010b	63	13	6	32	8	20.3	0.91 [0.82, 0.97]	0.71 [0.56, 0.84]

Fig. 2. Forest plot of BSDS, HCL-32 and MDQ for detection of bipolar disorder in mental health settings. The plot shows study specific estimates of sensitivity and specificity (with 95% confidence intervals) at a specific cut-off. The studies are ordered according to cut-off and % BDII/NOS. Where % BDII/NOS is blank, the information was not reported by the study. % BDII/NOS = percentage of bipolar cases that were bipolar disorder type II or not otherwise specified; FN = false negative; FP = false positive; TN = true negative; TP = true positive.

from the eight studies were consistent with the HCL-32 showing higher sensitivity and lower specificity than the MDQ. However, the curves for the two instruments lie close together and there was no evidence of a difference in accuracy ($p = 0.21$).

3.3.2. *Comparison of accuracies of the BSDS, HCL-32 and MDQ in the primary care or general population*

Five studies (240 BD cases; 3321 participants) evaluated the BSDS (one study), the HCL-32 (one study) and the MDQ (four studies) in the primary care setting or general population (see Supplementary Fig. S2 that follows the online version of this article). One study directly compared the BSDS to the HCL-32 for the detection of bipolar depression in a primary care sample with depression (29 cases; 576 patients) (Smith et al., 2011). This study reported a higher sensitivity and lower specificity for the BSDS at a cutoff of 12 compared to the HCL-32 at a cutoff of 18 (Supplementary Fig. S2). A meta-analysis comparing the three instruments in these settings was not possible due to limited data. Four studies (all with an optimum cutoff of 7) investigated the accuracy of the MDQ in the general population or primary care setting (182 cases; 2169 patients/participants). Supplementary Fig. S3 depicts the SROC curve of these four studies.

Summary sensitivity and specificity (95% CI) were 43.0% (11–81%) and 95% (45–100%) respectively.

3.3.3. Detection of bipolar disorder type II

Seventeen studies evaluated the BSDS (3 studies; 59 cases; 392 patients), HCL-32 (5 studies; 518 cases; 2430 patients) and MDQ (11 studies; 395 cases; 2774 patients) for detection of BD type II (Fig. 4). Two studies were comparative: one compared the HCL-32 and MDQ, and the other compared the BSDS and MDQ in the same population. All 17 studies were performed in a mental health care setting.

Table 1
Summary diagnostic characteristics of BSDS, HCL-32 and MDQ for detection of any type of bipolar disorder in mental health center and primary care or general population settings, according to test cut-off.

Instrument	Cut-off	*N*	Cases	Patients	Sensitivity (95% CI)	Specificity (95% CI)
Mental health setting						
BSDS	13	3	113	559	69% (63–74%)	86% (74–93%)
HCL-32	14	9	1845	4807	81% (77–85%)	67% (47–82%)
MDQ	6	3	165	447	81% (73–88%)	85% (79–89%)
	7	19	969	3220	66% (57–73%)	79% (72–84%)
Primary care or general population setting						
MDQ	7	4	182	2169	43% (11–81%)	95% (45–100%)

We compared the test performance of the BSDS, HCL-32 and MDQ. Fig. 5 presents the SROC curves for the three instruments. The BSDS was not significantly more accurate than the MDQ with an RDOR (95% CI) of 1.7 (0.8–3.8, $p=0.19$). However, there was evidence that the accuracy of the HCL-32 was superior to that of the MDQ with an RDOR of 2.0 (1.1 to 3.4, $p=0.018$). Supplementary Table S6 shows the sensitivities estimated from the curves at the median specificity obtained from the included studies. For example, given a cohort of 100 patients with a 15% prevalence of BD type II and a fixed specificity of 69%, the MDQ, HCL-32 and BSDS (with sensitivities of 68%, 81% and 78%, respectively) would miss 5, 3 and 3 cases respectively, while 26 of those without type II BD would be false positives.

3.3.4. Detection of bipolar disorder not otherwise specified

Two studies (30 cases; 264 patients) reported the diagnostic accuracy of the MDQ for detection of BD NOS (see Supplementary Fig. S4). Both studies used a cut-off of 7 and were conducted in a mental health setting (de Sousa Gurgel et al., 2012; Kim et al., 2008). The sensitivities were 29% (10–56%) (Kim et al., 2008) and 69% (39–91%) (de Sousa Gurgel et al., 2012), and the corresponding specificities were 77% (67–85%) and 80% (72–86%).

3.4. Assessment of heterogeneity

The results of investigations of heterogeneity are summarized in Supplementary Table S7 for the three instruments in a mental health care setting. Because few studies evaluated the BSDS, we were unable

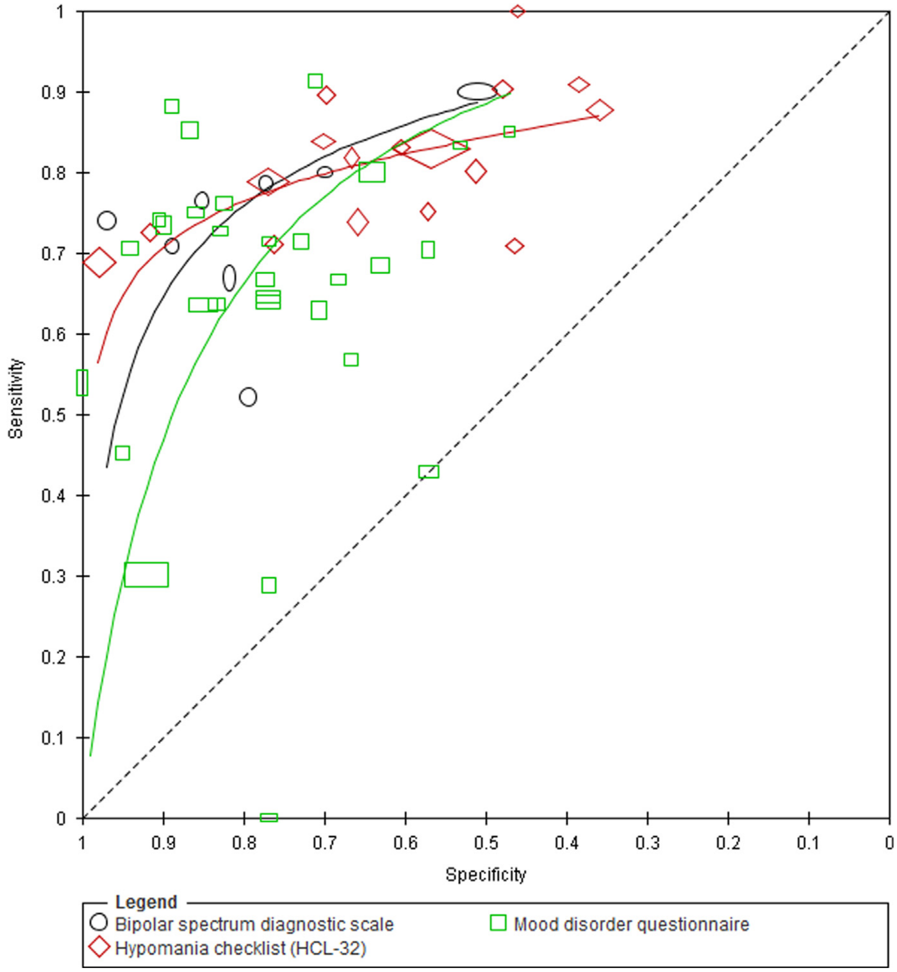
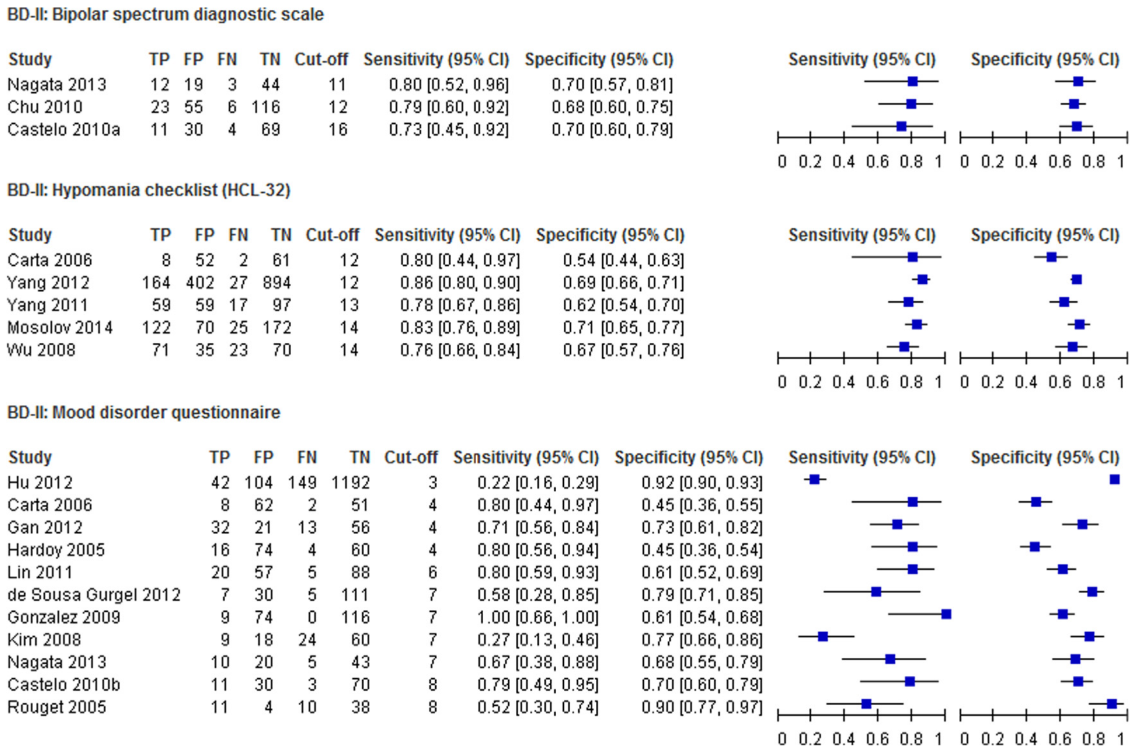


Fig. 3. Summary ROC plots of the BSDS, HCL-32 and MDQ for detection of bipolar disorder in mental health center setting. For each test, each symbol represents the pair of sensitivity and specificity from a study. The size of the symbols is scaled according to the sample size of the study. Plotted curves are restricted to the range of specificities for each instrument.



BD-II: Bipolar spectrum diagnostic scale

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Nagata 2013	12	19	3	44	11	0.80 [0.52, 0.96]	0.70 [0.57, 0.81]
Chu 2010	23	55	6	116	12	0.79 [0.60, 0.92]	0.68 [0.60, 0.75]
Castelo 2010a	11	30	4	69	16	0.73 [0.45, 0.92]	0.70 [0.60, 0.79]

BD-II: Hypomania checklist (HCL-32)

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Carta 2006	8	52	2	61	12	0.80 [0.44, 0.97]	0.54 [0.44, 0.63]
Yang 2012	164	402	27	894	12	0.86 [0.80, 0.90]	0.69 [0.66, 0.71]
Yang 2011	59	59	17	97	13	0.78 [0.67, 0.86]	0.62 [0.54, 0.70]
Mosolov 2014	122	70	25	172	14	0.83 [0.76, 0.89]	0.71 [0.65, 0.77]
Wu 2008	71	35	23	70	14	0.76 [0.66, 0.84]	0.67 [0.57, 0.76]

BD-II: Mood disorder questionnaire

Study	TP	FP	FN	TN	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)
Hu 2012	42	104	149	1192	3	0.22 [0.16, 0.29]	0.92 [0.90, 0.93]
Carta 2006	8	62	2	51	4	0.80 [0.44, 0.97]	0.45 [0.36, 0.55]
Gan 2012	32	21	13	56	4	0.71 [0.56, 0.84]	0.73 [0.61, 0.82]
Hardoy 2005	16	74	4	60	4	0.80 [0.56, 0.94]	0.45 [0.36, 0.54]
Lin 2011	20	57	5	88	6	0.80 [0.59, 0.93]	0.61 [0.52, 0.69]
de Sousa Gurgel 2012	7	30	5	111	7	0.58 [0.28, 0.85]	0.79 [0.71, 0.85]
Gonzalez 2009	9	74	0	116	7	1.00 [0.66, 1.00]	0.61 [0.54, 0.68]
Kim 2008	9	18	24	60	7	0.27 [0.13, 0.46]	0.77 [0.66, 0.86]
Nagata 2013	10	20	5	43	7	0.67 [0.38, 0.88]	0.68 [0.55, 0.79]
Castelo 2010b	11	30	3	70	8	0.79 [0.49, 0.95]	0.70 [0.60, 0.79]
Rouget 2005	11	4	10	38	8	0.52 [0.30, 0.74]	0.90 [0.77, 0.97]

Fig. 4. Forest plot of HCL-32 and MDQ for detection of bipolar disorder type II. The plot shows study specific estimates of sensitivity and specificity (with 95% confidence intervals) at a specific cut-off. All studies were performed in a mental health setting. The studies are ordered according to cut-off and study name. FN=false negative; FP=false positive; TN=true negative; TP=true positive.

to perform meta-regression analyses to assess heterogeneity in the diagnostic accuracy of this instrument.

For each instrument, we examined the distribution of the percentage of BD cases that were BD-II/NOS. The median percentage (interquartile range) for the BSDS (7 studies), MDQ (26 studies) and HCL-32 (15 studies) were 55% (32%, 65%), 53% (29%, 62%) and 54% (36%, 71%) respectively. For the HCL-32, there was no evidence of a difference in diagnostic accuracy between studies with a percentage of BD-II/NOS above or below the median percentage ($p=0.34$). Conversely, for the MDQ, there was strong evidence ($p<0.001$) of a difference in diagnostic accuracy between the two groups of studies – studies with a percentage of BD-II/NOS above the median showed lower accuracy compared to studies below the median with an RDOR (95% CI) of 0.29 (0.15–0.59).

For both the MDQ and HCL-32, there was no evidence of a difference in diagnostic accuracy between Asian and non-Asian studies with $p=$ 0.13 and $p=0.16$, respectively. For the two QUADAS-2 (Whiting et al., 2011) signaling questions, we grouped ‘no’ and ‘unclear’ responses as one subgroup because our interest was in how the ‘yes’ subgroup (indicating low risk of bias) would compare to the ‘no’ or ‘unclear’ subgroups (indicating high or unclear risk of bias). For the MDQ, there was evidence to suggest a difference in accuracy between studies that enrolled a consecutive or random sample of patients compared to studies that did not or were unclear ($p=0.03$). However, there was no evidence of a difference in the accuracy of the HCL-32 ($p=0.11$). For the MDQ and HCL-32, there was no evidence of a difference in accuracy between studies that used a case control design and those that did not or were unclear, with $p=0.31$ and $p=0.29$ respectively.

4. Discussion

In this meta-analysis, we determined the accuracy properties of the BSDS, HCL-32 and MDQ for the screening of bipolar spectrum disorders in psychiatric settings. However, the diagnostic properties of each instrument varies with cut-offs. At a cut-off of 7 the specificity of the MDQ seemed higher than that of the HCL-32 at a cut-off of 14, while the sensitivity of the HCL-32 was higher. This finding was further supported by studies that compared both instruments in the same population, even though cut-offs differed between studies.

For the detection of type II BD, the HCL-32 was significantly more accurate than the MDQ. Differences in the characteristics of the instruments could explain these findings. The MDQ includes a series of questions derived from the DSM-IV criteria for a manic episode (Hirschfeld et al., 2000). Since its development, the MDQ has been validated in psychiatric settings across a multitude of cultures worldwide. Some investigators raised initial concerns that the psychometric properties of the MDQ would be less satisfactory for the detection of type II BD (Benazzi, 2003; Mago, 2001). Subsequently, other reports indicated that the MDQ had lower accuracy for the detection of more subtle BD cases (for example, type II BD) (Hardoy et al., 2005; Weber Rouget et al., 2005). Hypomania presents in certain circumstances a ‘bright’ side specifically in patients who are more elated/active and less irritable/risk-taking (Brand et al., 2011; Gamma et al., 2008). Hypomanic individuals on the ‘bright side’ may rate themselves as more stress-tolerant and self-efficacious (Brand et al., 2011). Therefore, hypomanic episodes are prone to significant recall bias because a significant proportion of individuals may not perceive themselves as ‘abnormal’ when experiencing hypomanic symptoms and/or episodes. This may explain why the MDQ, which exclusively evaluates self-reported (hypo) manic symptoms, is less accurate than the HCL-32 for detection of type II BD. As a result of this perceived limitation of the MDQ, the HCL-32 (Angst et al., 2005) and the BSDS (Ghaemi et al., 2005) were developed to improve the detection of less exuberant bipolar spectrum disorders (e.g. type II and NOS). Developers of the HCL-32 attempted to develop an instrument to screen for bipolar spectrum disorders

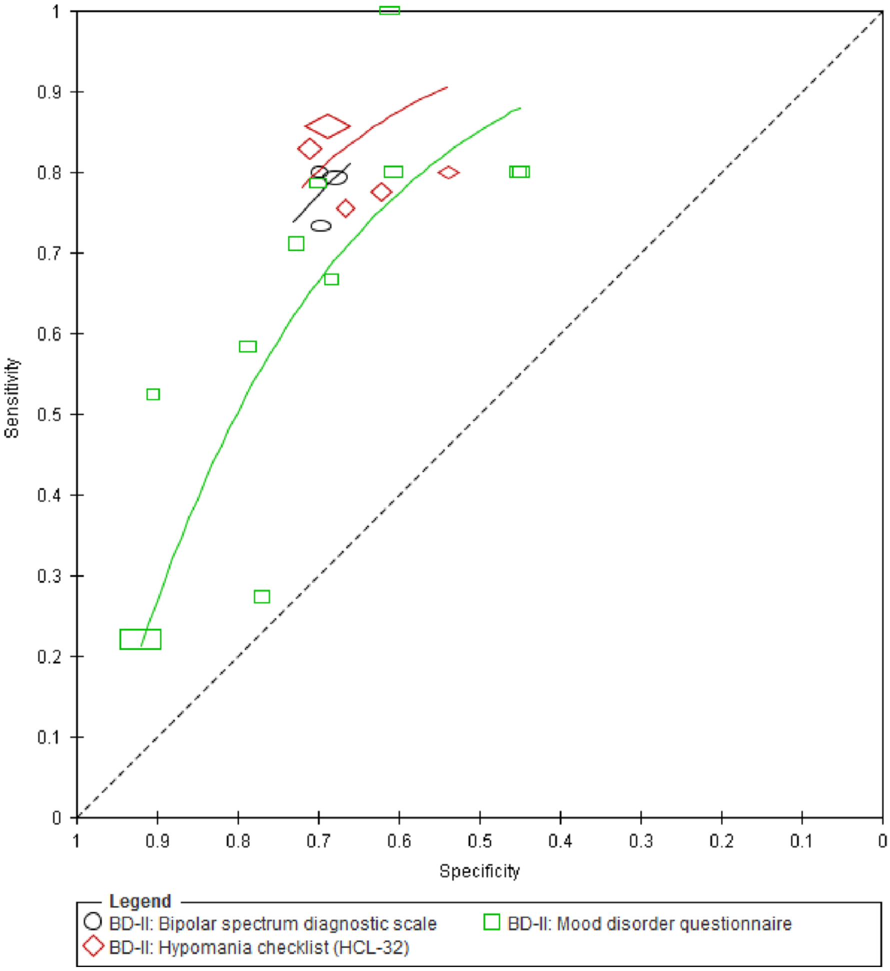


Fig. 5. Summary ROC plot of the BSDS, HCL-32 and MDQ for detection of bipolar disorder type II in mental health center setting. For each test, each symbol represents the pair of sensitivity and specificity from a study. The size of the symbols is scaled according to the sample size of the study. Plotted curves are restricted to the range of specificity for each instrument.

among patients in current depressive episodes through priming the respondent to the cyclical nature of BD and including more (hypo) manic manifestations (Angst et al., 2005), while the BSDS describes clinical manifestations of BD (including depressive symptoms) and emphasizes mood swings.

We identified significant sources of heterogeneity in our meta-analyses. First, the percentage of type II/NOS BD cases included in each study appeared to affect estimates of the performance of the MDQ. Conversely, the proportion of type II/NOS BD cases did not affect the diagnostic accuracy of the HCL-32. This analysis provide further support that the HCL-32 is more accurate than the MDQ for the detection of 'softer' (Angst and Marneros, 2001) BD cases. Second, categorization of studies into 'Asian' versus 'non-Asian' did not explain heterogeneity in study results for the HCL-32 or the MDQ. We performed these analyses because previous reports found lower sensitivity for the MDQ in Asian samples (Chung et al., 2008; Hu et al., 2012; Kim et al., 2008). Specifically, the impairment question of the MDQ seemed to explain its lower sensitivity in Asian cultures as an alternative scoring procedure eliminating this MDQ criteria restored the sensitivity of the instrument (Chung et al., 2008; Kim et al., 2008). We did not find evidence to support lower accuracy of the MDQ in Asian populations. Third, we found that a 'No' or 'Unclear' answer to the QUADAS-2 signaling question 'Was a consecutive or random sample of patients enrolled?' had a significant effect on test accuracy. Taking into account that several studies were performed in tertiary mental health care centers, selection of non-random/non-consecutive samples (and consequently prior knowledge of case status) may over-estimate the accuracy properties of a screening tool. Although case-control studies are prone to bias, the QUADAS-2 signaling question 'Was a case-control design avoided?' did not affect the accuracy of the MDQ or the BSDS. However, these results should be interpreted with caution as relatively few studies were rated as either a 'No' or 'Unclear' response to this question.

A recent systematic review indicated that BD may be a prevalent and under-recognized mental disorder in primary care (Cerimele et al., 2014). The authors found a lower prevalence for BD when structured diagnostic interviews were used (range 0.5–4.3%) compared to a screening instrument (7.6–9.8%). This finding highlights the possibility that a positive screen for BD may include a high number of false positive cases. Our meta-analysis indicated that the MDQ has a lower sensitivity for the detection of BD in primary care or general population settings compared to psychiatric settings. However, the instrument retained a high specificity in these settings. However, both sensitivity and specificity were subject to substantial uncertainty due to the small number of studies and between-study variation in estimates of test performance. A single study compared the BSDS to the HCL-32 for the detection of BD among primary care patients with depression (Smith et al., 2011). Accordingly, in this study both instruments had low positive predictive values (0.3 and 0.5, respectively). Evidence thus far indicates that these instruments may have lower sensitivity for the detection of BD in these settings compared to mental health centers. Several complex factors may contribute to this finding, notably prior-knowledge of disease status in mood disorder clinics.

4.1. Strengths and limitations

The main strengths of this review include the use of internationally recommended methods for study identification and selection, quality assessment and meta-analysis. Furthermore, this meta-analysis included a large number of studies and participants. Nevertheless, there were limitations. First, the comparative accuracy of the three instruments was determined mainly through indirect comparisons. Indirect comparisons are prone to confounding due to differences in study and population characteristics (Takwoingi et al., 2013). However, for the detection of any type of bipolar disorder, we also performed a direct comparison of the HCL-32 and MDQ, and results were consistent with the indirect comparison. Second, several different cut-offs were used for each instrument and we used the optimal cut-off that was reported in each study for our analyses. Selective reporting of optimal cut-offs can introduce bias if the selection is data driven but the bias is minimized in large studies (Leeflang et al., 2008). Because the median sample size in our review was 164 (interquartile range 122 to 363), we expect any bias to be minimal even if some of the included studies used a data driven approach to select the optimum cut-off. Furthermore, we compared the accuracy of the three instruments across the range of cut-offs by performing HSROC analyses. Third, the methodological quality of many of the included studies was limited. We investigated the effect of two relevant items of the QUADAS-2 tool on test performance. Fourth, included articles used the DSM-IV criteria as the reference standard. The DSM-5 introduced important changes in the taxonomy of mood disorders. Thus, the summary accuracy properties obtained in this review may be different considering DSM-5 criteria as reference standard. Finally, the DSM-5 field trials revealed that the inter-rater reliabilities for type I BD (kappa=0.56) and especially for type II BD (kappa=0.40) are not optimal (Freedman et al., 2013). Therefore, the accuracy of screening instruments should be interpreted considering intrinsic limitations of the 'gold standard' (i.e., the reliability of a DSM-based structured psychiatric interview).

4.2. Clinical implications

This review indicates that the accuracy of the BSDS, HCL-32 and MDQ are cut-off dependent. The instruments should not be considered case-finding tools, because a substantial proportion of patients who screen positive for BD do not actually have the disorder (Zimmerman, 2014). Therefore, a confirmatory diagnostic interview should follow a positive screen. Furthermore, a higher frequency of BD II/NOS amongst BD cases has a negative impact on the accuracy of the MDQ. For the detection of type II BD, the HCL-32 is superior to the MDQ. A meta-analysis of test accuracy provides a relevant first-step in test evaluation but other factors should be carefully considered (Leeflang et al., 2013). For example, cost-effectiveness analyses assessing the cost implications of false positives associated with the use of BD screening measures is important. However, it should be noted that the cost-effectiveness of case identification is complex to model and requires a number of assumptions concerning probabilities assigned in the BD treatment care pathway, and explicit values of treatment outcomes (Menzin et al., 2009; Valenstein et al., 2001). A previous cost effectiveness analysis indicated that a one-time administration of the MDQ in primary care patients with a major depressive episode would result in significant reductions in 5-year costs to managed-care plans (Menzin et al., 2009). Finally, well-designed randomized controlled trials (RCT) of BD screening will provide evidence related to patient outcomes. To our knowledge, no RCT has evaluated the effectiveness of BD screening on patient outcomes. Finally, a relevant clinical implication for improving the screening of bipolar disorder among patients with a major depressive episode would be a better accuracy in treatment prescription as antidepressant monotherapy may be associated with a heightened risk of hospitalization due to mania (Pacchiarotti et al., 2013) and is clearly not recommended for type I BD patients (Vieta, In press; Viktorin et al., 2014). Therefore, a better discrimination between unipolar and bipolar I depression would potentially result in improved outcomes and a reduced risk of iatrogeny.

4.3. Implications for research

Screening tools for BD have been used in large-scale epidemiological surveys as proxies to estimate the prevalence of BD in primary care (Cerimele et al., 2014) and in the general population (Hirschfeld et al., 2003). This review provides evidence that researchers should clearly differentiate a positive screen for BD due to the number of false positives associated with BD screening. There were few studies of the BSDS in a mental health setting compared to studies of the HCL-32 and the MDQ. The limited evidence from primary care and general population settings indicate that the sensitivity of the MDQ is lower in these settings than in mental health center settings. Future studies should investigate the diagnostic properties of the three screening instruments in primary care.

5. Conclusions

Screening instruments for BD have elevated specificities indicating that these scales would effectively screen out a large proportion of true negatives. However, a positive screen should be confirmed by a clinical diagnostic evaluation for BD. The accuracy properties of the MDQ and HCL-32 are supported by a larger evidence base than those of the BSDS. The HCL-32 is more accurate for the detection of type II BD than the MDQ in mental health care settings.

Role of funding source

Not applicable.

Conflict of interest

AFC received speaker honoraria from Abbott, GlaxoSmithKline, Lundbeck, Pfizer and Ache. Dr. Takwoingi, Mrs. Soczynska, Dr. Sales, Dr. Freitas, Dr. Quevedo and Dr. Hyphantis report no relationships with competing financial interests. Dr. McIntyre has received grant support or honoraria from or has served in scientific advisory boards for Astra Zeneca, Bristol Myers-Squibb, France Foundation, GlaxoSmithKline, Janssen-Ortho, Eli Lilly, Organon, Lundbeck, Merck, Pfizer, CME Outfitters, Optum Health, I3CME and Physicians' Postgraduate Press. Dr. Vieta has received grant support or honoraria from or has served on advisory boards for Alexza, Almirall, AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Ferrer, Forest, GlaxoSmithKline, Janssen, Jazz, Johnson & Johnson, Lundbeck, Merck-Sharpe & Dohme, Novartis, Organon, Otsuka, Pfizer, Richter, Sanofi, Servier, Schering Plough, Shire, unovion, Takeda, Teva, United Biosource Corporation, and Wyeth.

Acknowledgments

AFC and JQ are recipients of research fellowship awards from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; Brazil). CAK receives a postdoctoral fellowship award from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES; Brazil).

Appendix A. Supplementary Material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jad.2014.10.024.

References

Anderson, I.M., Haddad, P.M., Scott, J., 2012. Bipolar disorder (Clinical Research Ed.). BMJ 345, e8508.

Angst, J., Adolfsson, R., Benazzi, F., Gamma, A., Hantouche, E., Meyer, T.D., Skeppar, P., Vieta, E., Scott, J., 2005. The HCL-32: towards a self-assessment tool for hypomanic symptoms in outpatients. J. Affect. Disord. 88, 217–233.

Angst, J., Marneros, A., 2001. Bipolarity from ancient to modern times: conception, birth and rebirth. J. Affect. Disord. 67, 3–19.

Benazzi, F., 2003. Improving the mood disorder questionnaire to detect bipolar II disorder. Can. J. Psychiatry Rev. Can. Psychiatr. 48, 770–771.

Bond, D.J., Noronha, M.M., Kauer-Sant'Anna, M., Lam, R.W., Yatham, L.N., 2008. Antidepressant-associated mood elevations in bipolar II disorder compared with bipolar I disorder and major depressive disorder: a systematic review and meta-analysis. J. Clin. Psychiatry 69, 1589–1601.

Brand, S., Gerber, M., Puhse, U., Holsboer-Trachsler, E., 2011. 'Bright side' and 'dark side' hypomania are associated with differences in psychological functioning, sleep and physical activity in a non-clinical sample of young adults. J. Affect. Disord. 131, 68–78.

Cerimele, J.M., Chwastiak, L.A., Dodson, S., Katon, W.J., 2014. The prevalence of bipolar disorder in general primary care samples: a systematic review. Gen. Hosp. Psychiatry 36, 19–25.

Chessick, C.A., Dimidjian, S., 2010. Screening for bipolar disorder during pregnancy and the postpartum period. Arch. Women's Ment. Health 13, 233–248.

Chung, K.F., Tso, K.C., Cheung, E., Wong, M., 2008. Validation of the Chinese version of the mood disorder questionnaire in a psychiatric population in Hong Kong. Psychiatry Clin. Neurosci. 62, 464–471.

Culpepper, L., 2014. Misdiagnosis of bipolar depression in primary care practices. J. Clin. Psychiatry 75, e05.

de Sousa Gurgel, W., Reboucas, D.B., Negreiros de Matos, K.J., Carneiro, A.H., Gomes de Matos e Souza, F., 2012. Brazilian Portuguese validation of mood disorder questionnaire. Compr. Psychiatry 53, 308–312.

Drancourt, N., Etain, B., Lajnef, M., Henry, C., Raust, A., Cochet, B., Mathieu, F., Gard, S., Mbailara, K., Zanouy, L., Kahn, J.P., Cohen, R.F., Wajsbrot-Elgrabli, O., Leboyer, M., Scott, J., Bellivier, F., 2013. Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment. Acta. Psychiatr. Scand. 127, 136–144.

Freedman, R., Lewis, D.A., Michels, R., Pine, D.S., Schultz, S.K., Tamminga, C.A., Gabbard, G.O., Gau, S.S., Javitt, D.C., Oquendo, M.A., Shrout, P.E., Vieta, E., Yager, J., 2013. The initial field trials of DSM-5: new blooms and old thorns. Am. J. Psychiatry 170, 1–5.

Frye, M.A., 2011. Clinical practice. Bipolar disorder—a focus on depression. N. Engl. J. Med. 364, 51–59.

Gamma, A., Angst, J., Ajdacic-Gross, V., Rossler, W., 2008. Are hypomanics the happier normals? J. Affect. Disord. 111, 235–243.

Ghaemi, S.N., Miller, C.J., Berv, D.A., Klugman, J., Rosenquist, K.J., Pies, R.W., 2005. Sensitivity and specificity of a new bipolar spectrum diagnostic scale. J. Affect. Disord. 84, 273–277.

Ghaemi, S.N., Rosenquist, K.J., Ko, J.Y., Baldassano, C.F., Kontos, N.J., Baldessarini, R.J., 2004. Antidepressant treatment in bipolar versus unipolar depression. Am. J. Psychiatry 161, 163–165.

Hardoy, M.C., Cadeddu, M., Murru, A., Dell'Osso, B., Carpiniello, B., Morosini, P.L., Calabrese, J.R., Carta, M.G., 2005. Validation of the Italian version of the "mood disorder questionnaire" for the screening of bipolar disorders. Clin. Pract. Epidemiol. Ment. Health: CP & EMH 1, 8.

Hirschfeld, R., Williams, J., Spitzer, R., Calabrese, J., Flynn, L., Keck Jr., P., Lewis, L., McElroy, S., Post, R., Rapport, D., Russell, J., Sachs, G., Zajecka, J., 2000. Development and validation of a screening instrument for bipolar spectrum disorder: the mood disorder questionnaire. Am. J. Psychiatry 157, 1873–1875.

Hirschfeld, R.M., Calabrese, J.R., Weissman, M.M., Reed, M., Davies, M.A., Frye, M.A., Keck Jr., P.E., Lewis, L., McElroy, S.L., McNulty, J.P., Wagner, K.D., 2003. Screening for bipolar disorder in the community. J. Clin. Psychiatry 64, 53–59.

Hirschfeld, R.M., Vornik, L.A., 2004. Recognition and diagnosis of bipolar disorder. J. Clin. Psychiatry 65 (Suppl 15), S5–S9.

Hu, C., Xiang, Y.T., Wang, G., Ungvari, G.S., Dickerson, F.B., Kilbourne, A.M., Lai, K.Y., Si, T.M., Fang, Y.R., Lu, Z., Yang, H.C., Hu, J., Chen, Z.Y., Huang, Y., Sun, J., Wang, X.P., Li, H.C., Zhang, J.B., Chiu, H.F., 2012. Screening for bipolar disorder with the mood disorders questionnaire in patients diagnosed as major depressive disorder – the experience in China. J. Affect. Disord. 141, 40–46.

Judd, L.L., Schettler, P.J., Akiskal, H.S., Maser, J., Coryell, W., Solomon, D., Endicott, J., Keller, M., 2003. Long-term symptomatic status of bipolar I vs. bipolar II disorders. Int. J. Neuropsychopharmacol./Off. Sci. J. Coll. Int. Neuropsychopharmacol. (CINP) 6, 127–137.

Kim, B., Wang, H.R., Son, J.I., Kim, C.Y., Joo, Y.H., 2008. Bipolarity in depressive patients without histories of diagnosis of bipolar disorder and the use of the mood disorder questionnaire for detecting bipolarity. Compr. Psychiatry 49, 469–475.

Leeflang, M.M., Deeks, J.J., Takwoingi, Y., Macaskill, P., 2013. Cochrane diagnostic test accuracy reviews. Syst. Rev. 2, 82.

Leeflang, M.M., Moons, K.G., Reitsma, J.B., Zwinderman, A.H., 2008. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin. Chem. 54, 729–737.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gotzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration (Clinical Research Ed.). BMJ 339, b2700.

Lish, J.D., Dime-Meenan, S., Whybrow, P.C., Price, R.A., Hirschfeld, R.M., 1994. The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. J. Affect. Disord. 31, 281–294.

Loganathan, M., Lohano, K., Roberts, R.J., Gao, Y., El-Mallakh, R.S., 2010. When to suspect bipolar disorder. J. Fam. Pract.59, 682–688.

Macaskill, P., 2004. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. J. Clin. Epidemiol. 57, 925–932.

Macaskill, P., Gatsonis, C., Deeks, J.J., Harbord, R.M., Takwoingi, Y., 2010. Chapter 10: analysing and presenting results (Available from:). In: Deeks, J.J., Bossuyt, P.M., Gatsonis, C. (Eds.), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, Available from: http://srdta.cochrane.org/.

Mago, R., 2001. Bipolar disorder questionnaire. Am. J. Psychiatry 158, 1743–1744.

Matza, L.S., Rajagopalan, K.S., Thompson, C.L., de Lissovoy, G., 2005. Misdiagnosed patients with bipolar disorder: comorbidities, treatment patterns, and direct treatment costs. J. Clin. Psychiatry 66, 1432–1440.

Menzin, J., Sussman, M., Tafesse, E., Duczakowski, C., Neumann, P., Friedman, M., 2009. A model of the economic impact of a bipolar disorder screening program in primary care. J. Clin. Psychiatry 70, 1230–1236.

Pacchiarotti, I., Bond, D.J., Baldessarini, R.J., Nolen, W.A., Grunze, H., Licht, R.W., Post, R.M., Berk, M., Goodwin, G.M., Sachs, G.S., Tondo, L., Findling, R.L., Youngstrom, E.A., Tohen, M., Undurraga, J., Gonzalez-Pinto, A., Goldberg, J.F., Yildiz, A., Altshuler, L.L., Calabrese, J.R., Mitchell, P.B., Thase, M.E., Koukopoulos, A., Colom, F., Frye, M.A., Malhi, G.S., Fountoulakis, K.N., Vazquez, G., Perlis, R.H., Ketter, T.A., Cassidy, F., Akiskal, H., Azorin, J.M., Valenti, M., Mazzei, D.H., Lafer, B., Kato, T., Mazzarini, L., Martinez-Aran, A., Parker, G., Souery, D., Ozerdem, A., McElroy, S.L., Girardi, P., Bauer, M., Yatham, L.N., Zarate, C.A., Nierenberg, A.A., Birmaher, B., Kanba, S., El-Mallakh, R.S., Serretti, A., Rihmer, Z., Young, A.H., Kotzalidis, G.D., MacQueen, G.M., Bowden, C.L., Ghaemi, S.N., Lopez-Jaramillo, C., Rybakowski, J., Ha, K., Perugi, G., Kasper, S., Amsterdam, J.D., Hirschfeld, R.M., Kapczinski, F., Vieta, E., 2013. The international society for bipolar disorders (ISBD) task force report on antidepressant use in bipolar disorders. Am. J. Psychiatry 170, 1249–1262.

Parker, G., Fletcher, K., Barrett, M., Synnott, H., Breakspear, M., Hyett, M., Hadzi-Pavlovic, D., 2008. Screening for bipolar disorder: the utility and comparative properties of the MSS and MDQ measures. J. Affect. Disord. 109, 83–89.

Parker, G., Hadzi-Pavlovic, D., Tully, L., 2006. Distinguishing bipolar and unipolar disorders: an isomer model. J. Affect. Disord. 96, 67–73.

Phelps, J.R., Ghaemi, S.N., 2006. Improving the diagnosis of bipolar disorder: predictive value of screening tests. J. Affect. Disord. 92, 141–148.

Smith, D.J., Griffiths, E., Kelly, M., Hood, K., Craddock, N., Simpson, S.A., 2011. Unrecognised bipolar disorder in primary care patients with depression. Br. J. Psychiatry: J. Ment. Sci. 199, 49–56.

Takwoingi, Y., Leeflang, M.M., Deeks, J.J., 2013. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann. Intern. Med. 158, 544–554.

Undurraga, J., Baldessarini, R.J., Valenti, M., Pacchiarotti, I., Vieta, E., 2012. Suicidal risk factors in bipolar I and II disorder patients. J. Clin. Psychiatry 73, 778–782.

Valenstein, M., Vijan, S., Zeber, J.E., Boehm, K., Buttar, A., 2001. The cost-utility of screening for depression in primary care. Ann. Intern. Med. 134, 345–360.

Vieta, E., 2014. Antidepressants in bipolar I disorder: never as monotherapy. Am J. Psychiatry 171, 1023–1026.

Viktorin, A., Lichtenstein, P., Thase, M.E., Larsson, H., Lundholm, C., Magnusson, P.K., Landen, M., 2014. The risk of switch to mania in patients with bipolar disorder during treatment with an antidepressant alone and in combination with a mood stabilizer. Am. J. Psychiatry 171, 1067–1073.

Weber Rouget, B., Gervasoni, N., Dubuis, V., Gex-Fabry, M., Bondolfi, G., Aubry, J.M., 2005. Screening for bipolar disorders using a French version of the mood disorder questionnaire (MDQ). J. Affect. Disord. 88, 103–108.

Whiting, P.F., Rutjes, A.W., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M., Sterne, J.A., Bossuyt, P.M., 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155, 529–536.

Zimmerman, M., 2012. Misuse of the mood disorders questionnaire as a case-finding measure and a critique of the concept of using a screening scale for bipolar disorder in psychiatric practice. Bipolar Disord. 14, 127–134.

Zimmerman, M., 2014. Screening for bipolar disorder: confusion between case-finding and screening. Psychother. Psychosom. 83, 259–262.

Zimmerman, M., Galione, J.N., Chelminski, I., Young, D., Ruggero, C.J., 2010. Performance of the bipolar spectrum diagnostic scale in psychiatric outpatients. Bipolar Disord. 12, 528–538.