Quality of Life Research (2019) 28:1111–1118
https://doi.org/10.1007/s11136-018-2050-x

REVIEW

How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II)

Michael von Glischinski[1] · Ruth von Brachel[2] · Gerrit Hirschfeld[3]

Accepted: 11 November 2018 / Published online: 19 November 2018

Abstract

Introduction The Beck Depression Inventory revised (BDI-II) is widely used tool to screen for depression. The aim of the present study was to systematically review and synthesize studies that determined optimal cut points for the BDI-II.

Method We identified 27 studies that tried to identify optimal cut points for the BDI-II. Study quality was assessed using QUADAS criteria. Cut points and their variability were analyzed descriptively, via simulation and synthesized with a diagnostic meta-analysis. Analysis was performed on all studies and subgroups based on the setting (psychiatric, somatic, healthy).

Results Cut points identified as optimal ranged from 10 to 25 across all studies. Simulation-based estimations of the variability inherent in studies show that much of the between-study differences may be attributed to random fluctuations. Diagnostic meta-analysis across all studies revealed that a cut point of 14.5 (95% CI 12.75–16.44) is optimal, yielding a sensitivity of 0.86 and a specificity of 0.78. Analyses within the different settings suggest using sample-specific cut points, specifically 18.18 in psychiatric settings, and 12.9 in primary care settings and healthy populations.

Conclusion Most studies aimed at determining optimal cut points fail to acknowledge that reported results are only estimates and subject to random fluctuations resulting in conflicting recommendations for practitioners. Taking into account these fluctuations, we find that practitioners should use different cut points to screen for depression in primary care and healthy populations (a score of 13 and higher indicates depression) and psychiatric settings (a score of 19 and higher indicates depression). Methods to describe this variability and meta-analysis to synthesize findings across studies should be used more widely.

Keywords Depression · Diagnostic utility · Meta-analysis · Beck Depression Inventory

Introduction

Major depressive disorder (MDD) is a pervasive, common condition affecting people of all ages and races [9]. It significantly impairs mental, physical, and social functioning [7, 11]; is highly comorbid with other mental disorders particularly anxiety disorders [3]; and is associated with enormous economic costs [1]. However, a substantial number of patients affected by MDD remains undiagnosed and does not receive proper treatment [16]. The best available reference test for the detection of depressive symptoms is a full psychiatric interview, carried out by a health professional [2]. However, for clinical practice this method is often too time consuming and therefore uneconomic. During the past decades, many standardized self-report measures have been developed, to grade the severity of depressive symptoms and to identify individuals suffering from MDD. A widely used measure of depressive symptoms is the Beck Depression Inventory revised (BDI-II) [15]. Given the wide-use of the BDI-II, several reviews have scrutinized the extensive data on its psychometric performance [4, 27; For a recent review see: 27]. While these agree that the BDI-II exhibits excellent psychometric performance and is able to classify patients as "depressed" vs. "non-depressed," these

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-2050-x) contains supplementary material, which is available to authorized users.

✉ Gerrit Hirschfeld
gerrit.hirschfeld@fh-bielefeld.de

[1] Universität Witten Herdecke, Witten, Germany

[2] Mental Health Research & Treatment Center, Ruhr-Universität Bochum, Bochum, Germany

[3] Faculty of Business Management and Health, Bielefeld University of Applied Sciences, Bielefeld, Germany

studies also noted that different cut points emerged as optimal in different samples. Cut points are not only used to classify patients in clinical practice, but are also used to classify patients' outcomes in clinical trials, e.g., by defining all patients as remitted who attain a post-score below 10 [13]. The aim of the present study is to systematically review studies aimed at identifying optimal cut points for the BDI-II and perform a diagnostic meta-analysis to estimate the sensitivity, specificity, and optimal cut point.

To date, a number of studies tried to establish data-driven optimal cut points. These anchor-based methods require that both the BDI-II as well as a reference test (e.g., structured clinical interviews) are assessed in all study participants [20]. Once both scores are available, receiver-operating-characteristic (ROC)-based methods are used to determine the "optimal" cut point. Usually, the sensitivity and specificity for all available cut points is calculated and the cut point that shows the highest Youden index (sensitivity + specificity) is identified as optimal. However, using this same method in different samples yields highly conflicting findings. In the review by Wang and Gorenstein [27], the cut points determined as optimal range from 7 to 31, with somewhat lower cut points emerging as optimal for samples recruited in somatic hospitals and treatment facilities, and higher cut points in psychiatric treatment facilities. While it makes intuitive sense to use sample-specific cut points to interpret the findings, this severely hampers our ability to integrate findings from disparate studies. Furthermore, these discrepancies between different studies may be due to chance fluctuations. That is, cut points that are optimal within a specific sample do not necessarily emerge as optimal in other samples from the same population or generalize well to the population at large [10].

The aims of the present study are threefold: (1) to describe the study quality of studies that used ROC-based methods to determine cut points for the BDI-II, (2) to describe the proposed cut points and the level of variability inherent to their estimation, and (3) to synthesize the existing studies using diagnostic meta-analysis [23].

Methods

Literature search BDI

We systematically reviewed the existing literature about the BDI-II for studies determining specific cut points. The following criteria had to be met by the individual study to be included: (1) use of ROC methods for determination of cut points; (2) report means and standard deviations of BDI-II for both clinical and non-clinical samples; (3) written in English or German. We also excluded articles that were (1) letters or editorials and/or (2) written in languages other than English or German. We included all studies irrespective of the type of reference standard used to diagnose depression.

The initial search on MEDLINE and PsycInfo (conducted on February 2, 2017) focused on studies about the BDI-II and psychometrics, cut points or cut-offs. We restricted the search to articles which were published between January 1, 1961 and December 31, 2016 (see Online Appendix for Medline syntax). This search yielded 757 results. Furthermore, the reference section of review articles were checked for articles not included in our search results, as well as the reference list of retained articles. The title and abstract of all articles identified by this search strategy were screened by one author (MvG), using the criteria described earlier. Seventy abstracts were also screened by a second author (GH) and yielded identical decisions. Full texts of potentially relevant articles were reviewed by two reviewers (MvG and GH). For studies that potentially collected relevant data (i.e., cut points) but did not report necessary scores or statistics, the authors were contacted via e-mail. Because some studies had been published some decades ago, we only contacted authors of articles published within the last 10 years. A total of 25 authors were contacted, of which 15 responded to our request. Six authors provided the missing information, and nine authors responded that the data were no longer accessible.

Quality assessment of diagnostic accuracy studies

We assessed the methodologic quality of the studies by using the revised Quality Assessment of Diagnostic Accuracy Studies framework [28]. This instrument was developed to facilitate identifying common design problems, such as the lack of blinding. In addition, it highlights methodologic differences between studies, such as differences in the measurement of both reference test and screening methods. Thus, it supports deciding whether a specific study should be included in the synthesis. We discarded the QUADAS-2 item asking whether the index-test used a pre-specified threshold because the studies reviewed here were aimed at determining this threshold.

Data analysis

Data were analyzed in two steps. First, we describe the existing optimal cut points and their variability by performing a simulation study based on the sample characteristics of the studies identified by the systematic review. That is, we used the reported sample size, mean and standard deviation of participants with and without depression to generate samples and calculate the optimal cut point within this generated data. Samples were drawn from a normal distribution with the same parameters (mean and standard deviation) that was found in the study sample. Optimal cut points were

defined as those cut points which maximized the Youden index (sum of sensitivity and specificity). For each of the studies, we simulated 5.000 different samples of patients and controls and calculated the optimal cut point, resulting in 5.000 different cut points for each study. From the 5.000 optimal cut points, we calculated for each study, we calculated the median cut point and the 2.5% and 97.5% quartiles. For details, see the accompanying R-code.

Second, we performed a meta-analysis of the optimal cut points using the multiple cut point model [23] to estimate pooled estimates for sensitivity and specificity, the summary ROC curve and the optimal cut point. Compared to other methods used to estimate the SROC function [18], this method has the advantage of taking into account multiple cut points per study and study heterogeneity, thereby improving the precision at which the model parameters may be estimated. Specifically, we used linear mixed effect models with separate random intercepts and equal random slopes to model the log-transformed diagnostic utility data, assuming an underlying normal distribution of BDI-II scores. Restricted maximum likelihood estimation was used to estimate model parameters. However, the implementation does not allow for formally testing the sources of heterogeneity. That is why we only report separate analyses for different groups of studies. All analyses were performed in the open-source programming language R and the cutpointr package. The raw-data and code to perform the analyses are available as an Online Appendix.

Results

Literature search

Overall, the search strategy identified 768 articles as potentially relevant. Of these, 670 studies were excluded after screening (Fig. 1). The majority of articles were excluded because they only used the BDI-II as an outcome measure rather than reporting on its diagnostic utility or optimal cut points ($n = 528$). Other reasons were that they used the BDI-I ($n = 73$), were written in languages other than English or German ($n = 48$), or were themselves reviews of existing studies ($n = 21$). The remaining 98 articles were assessed for eligibility, resulting in the exclusion of another 71 studies, which either did not determine an optimal cut point ($n = 15$), provided insufficient data ($n = 53$), or were not

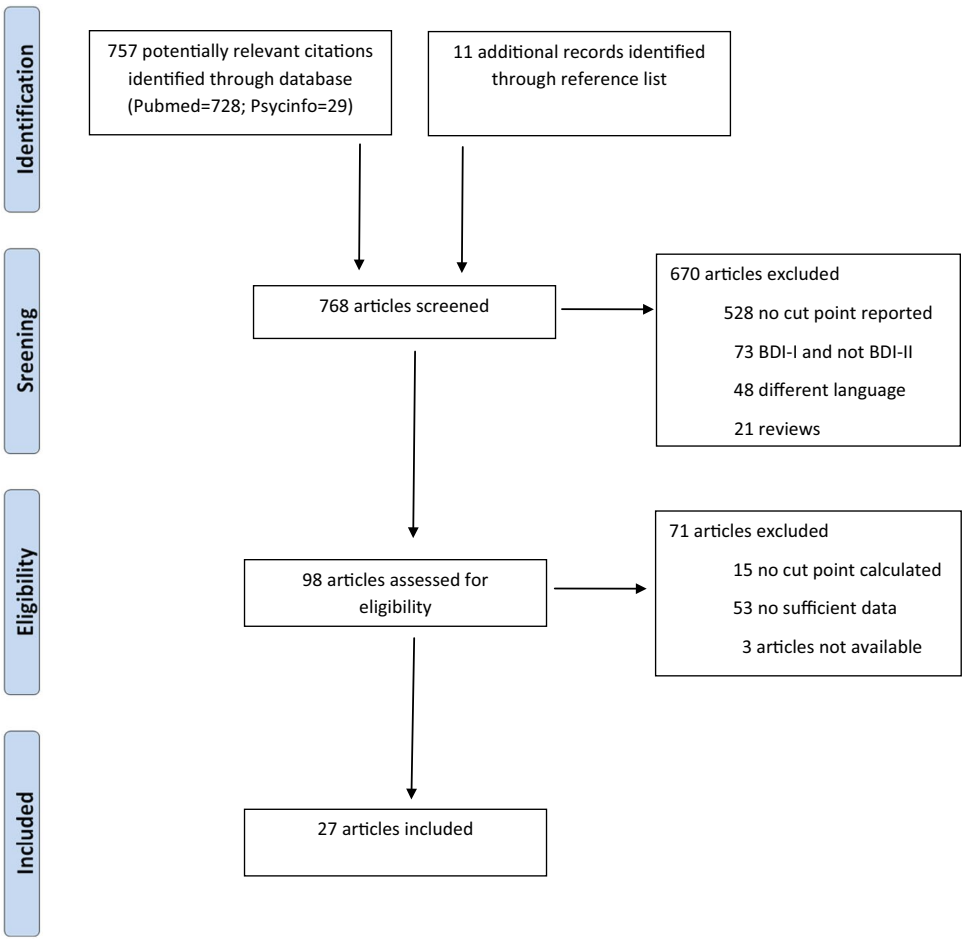

Fig. 1 Flowchart

available ($n = 3$). The remaining 27 articles met inclusion criteria and were included in the analyses. Core information about the studies is given in the Online Appendix. Overall, these studies included 2044 participants that were classified as "depressed" using clinical criteria and 8979 participants classified as "not depressed."

Study quality

With regard to study quality, we found the following positive aspects throughout: all studies showed low risk of bias with regard to patient selection and index testing; in most cases, the reference standard was determined by a (standardized) clinical interview ($n = 25$); the exclusion of study participants was always explained in sufficient detail; and information about participant demographic characteristics (e.g., age, gender), their medical/psychiatric condition, and recruitment setting were provided. However, despite these positive properties across the studies, some limitations have to be reported: the order of testing (BDI-II or reference standard first) was sometimes described only vague ($n = 13$); results of the BDI-II might not always be interpreted without knowledge of the reference standard ($n = 13$); and it was not always clear if all participants received the same reference standard ($n = 2$). This happens when researchers rely on a known-groups validation approach. For example, Osman and colleagues [17] used a known-group validation to compare psychiatric inpatients to high school students. Critically, the healthy participants in such a study mostly are not assessed with the same rigor as the patients, opening the possibility that some of the high school students might also suffer from depression. We classified this study as healthy because the authors aim was to develop cut points to be used in healthy participants. Detailed information concerning the quality assessment of the included studies is reported in Online Appendix A. Furthermore, there were many differences concerning sample characteristics (e.g., healthy, somatic, or psychiatric sample) and recruitment setting (e.g., outpatient or inpatient) between studies, which might be a source of systematic bias. To facilitate comparisons with similar investigations, the studies were grouped by sample recruitment source, resulting in non-clinical ($n = 5$), somatic ($n = 15$), and psychiatric ($n = 7$) samples.

Cut points in included studies

The original authors of the BDI-II recommended the following rules of thumb for the interpretation of their instrument: a score of 0–13 indicates minimal or no depression; 14–19, mild depression; 20–28, moderate depression; and a score of 29–63 indicates severe depression [4]. These cut points might vary, depending on the type of sample and study purpose. Table 1 summarizes the results of included studies which determined cut points for the BDI-II (see Online Appendix A for more information on the study characteristics). Most samples were recruited in outpatient settings (55%), and comprised of English-speaking (74%) medical patients (59%). The remaining studies comprised of several psychiatric (26%) and a few non-clinical (15%) samples. Overall, these 27 studies identified cut points between a range of 10 and 25 as optimal, with a median and mode of each 16. Also when looking at the three samples separately, the optimal cut points were still very variable, with ranges from 10 to 22/7 to 22/and 13.5 to 25 for healthy, somatic, and psychiatric samples.

Figure 2 also indicates the variability inherent in the individual studies. Since this largely depends on the sample size of the control and diseased group, the variability was estimated to be highest in the study by Low and colleagues [14], with 95% of the optimal cut points falling between 8 and 30. While in large-scale studies such as the Leiden Routine Outcome Monitoring Study [21], 95% of simulated cut points would fall between 13 and 15. Overall, the results of the simulation study show that for many studies, the range of resulting optimal cut points encompasses almost the whole range of cut points. Random fluctuations seem to be the most parsimonious explanation for the differences in the reported optimal cut points between studies. However, when taking into account these random fluctuations and constructing confidence intervals for the cut points, one can gage a true optimal cut point. Specifically, when looking at the studies describing somatic samples, the cut point of 14 is entailed in all but one of the ranges and would be thus consistent with the studies.

Diagnostic meta-analysis of cut points

The meta-analysis across all samples showed that the optimal cut point was 14.48 (95% CI 12.75–16.44) with associated sensitivity of 0.86 (95% CI 0.82–0.90) and specificity of 0.78 (95% CI 0.72–0.84; Fig. 3). Performing separate meta-analyses on studies from the three different samples showed that for psychiatric patients a cut point of 18.18 (95% CI 15.35–21.52) would be optimal yielding a sensitivity of 0.87 (95% CI 0.79–0.92) and a specificity of 0.77 (95% CI 0.57–0.90). For somatic samples, a cut point of 12.48 (95% CI 11.55–13.49) would be optimal and yield a sensitivity of 0.88 (95% CI 0.85–0.91) and a specificity of 0.79 (95% CI 0.77–0.82). Finally, in healthy samples, a cut point of 14.06 (95% CI 7.32–20.79) would result in a sensitivity of 0.79 (95% CI 0.52–0.94) and a specificity of 0.76 (95% CI 0.53–0.92). Due to the highly similar and overlapping cut points in the latter two groups, we combined and calculated the meta-analysis. This resulted in an optimal cut point of 12.93 (95% CI 11.93–14.04) that is associated with

Table 1 Overview of ROC-based studies for the BDI-II

Study	Sample			BDI-II[a]			OC	Sen	Spec	AUC	PPV	NPV
	Total	Depressive	Control	Total	Depressive	Control						
Araya (2013)	571	301	270	NA	24.2 (10.2)	13.5 (7.6)	13/14	72.2	64.1	74	69	80
Arnau (2001)	335	31	304	8.74 (9.7)	28 (9.7)	6.7 (7.1)	18	94	92	96	54	99
Carney (2009)	140	36	104	14.1 (10.2)	22 (8.6)	12.0 (9.7)	≥ 17	81	79	84	NR	NR
Chilcot (2008)	40	9	31	12.9 (9.3)	25.6 (7.0)	9.3 (9.3)	≥ 16	89	87	96	88.8	87
de Souza (2010)	50	12	38	NA	26.1 (14.0)	8.8 (8.9)	13/14	100	66	86	48	93
Dolle (2012)	88	24	64	NA	31.6 (9.6)	10.5 (8.9)	≥ 23	88	92	93	81	95
Dutton (2004)	220	65	155	12.6 (10.4)	23.1 (8.7)	8.2 (7.5)	14	88	84	91	70	94
Hayden (2012)	83	15	68	NA	22.9 (5.2)	11.3 (8.4)	13	100	67	86	40.5	100
Hopko (2008)	33	24	9	26.5 (12.5)	32 (9.9)	12.1 (4.7)	22	92	100	NA	100	NR
Huffmann (2010)	131	17	114	9.8 (9.4)	27.5 (9.5)	7.1 (5.9)	≥ 16	88	92	96	62.5	98.1
Jakšić (2013)	314	52	262	10.4 (10.3)	27.6 (11.1)	6.9 (5.6)	15/16	88.46	91.22	96	66.7	97.6
Krefetz (2002)	100	57	43	24.7 (12.5)	25.6 (15.0)	23.9 (12.4)	≥ 24	74	70	78	76	67
Kumar (2002)	100	54	46	NA	32.8 (12.2)	11.0 (10.4)	≥ 21	85	83	92	85	83
Low (2006)	112	6	106	8.0 (7.2)	22.5 (15.1)	7.2 (5.5)	10	100	75	91	18	100
Moullec (2014)	750	42	708	NA	19 (10)	8 (6)	10	83	73	84	16	99
Osman (2008)	167	61	106	NA	23.4 (11.4)	12.5 (10.5)	10	87	57	77	NR	NR
Phan (2015)	56	9	47	NA	32.4 (12.4)	9.4 (8.2)	≥ 13	89	77	95	88.9	76.6
Pietsch (2012)	314	21	293	NA	25.8 (10.1)	7.5 (6.5)	≥ 19	86	93	93	47	99
Plourde (2015)	801	108	693	NA	22 (10)	7 (6)	12	85	79	NR	39	97
Schulte van Maaren (2013)	4474	455	4019	NA	30.8 (10.5)	3.74 (4.7)	13.5	96	96	99	NR	NR
Seignourel (2008)	582	124	458	20.6 (11.8)	31.2 (10.3)	17.7 (10.4)	27	75	75	82	23	97
Shean (2008)	95	17	78	NA	14.8 (6.6)	5.5 (4.2)	10	73	84	NR	47.8	94.2
Strober (2015)	81	11	70	NA	21.8 (6.7)	9.0 (5.2)	14	91	81	93	43.5	98.3
Su (2007)	185	23	162	NA	17 (10.2)	7 (5)	11/12	74	83	84	NR	NR
Subica (2014)	575	233	342	25.3 (12.7)	28 (11.4)	21.2 (13.5)	19	79	54	70	82.6	48.6
Vasegh (2015)	400	38	362	15.2 (12.0)	34.1 (10.7)	13.2 (10.3)	22	87	79	91	NR	NR
Williams (2012)	229	93	136	NA	14.7 (7.4)	6.5 (5.2)	7	95	60	85	62	94

AUC area under curve, *BDI-II* Beck Depression Inventory—second edition, *NPV* negative predictive value, *NR* not reported, *OC* optimal cut point, *PPV* positive predictive value, *Sen* sensitivity, *Spec* specificity

[a]Data are given as mean with standard deviation

a sensitivity of 0.86 (95% CI 0.83–0.89) and a specificity of 0.78 (95% CI 0.74–0.82).

Discussion

The aim of the present study was to systematically review and synthesize the existing data on optimal cut points for the BDI-II. We found 27 studies that supplied relevant data on the diagnostic utility of the BDI-II. Included studies identified cut points between 8 and 20 as "optimal." While overall study quality was good, most authors do not address the fact that the methods they employ are subject to random fluctuations and do not report on the variability of the optimal cut point. The meta-analysis showed that the optimal cut points for psychiatric patients should be 18.18 and 12.9 in healthy and somatic samples. In the following, we first discuss the descriptive results and the meta-analysis before discussing some limitations of the present study.

The descriptive analyses of the cut points revealed two noteworthy results. First, the use of the Youden criterion alone is quite problematic and can result in cut points identified as optimal, that are smaller than the means of the non-depressed group [17, 25]. For example, in the study by Subica and colleagues [25], 575 patients of a psychiatric hospital were classified into two groups: a first group of patients with depression ($n = 342$) and a second group without depression ($n = 233$) using structured clinical interviews. Patients without depression had a mean BDI-II score of 21 and patients with depression a mean of 28. However, the cut point that maximized the Youden index for this classification was 19, i.e., even smaller than the non-depressed group. Thus, if one would use this cut point to interpret BDI-II scores from patients in a psychiatric hospital, the majority of patients

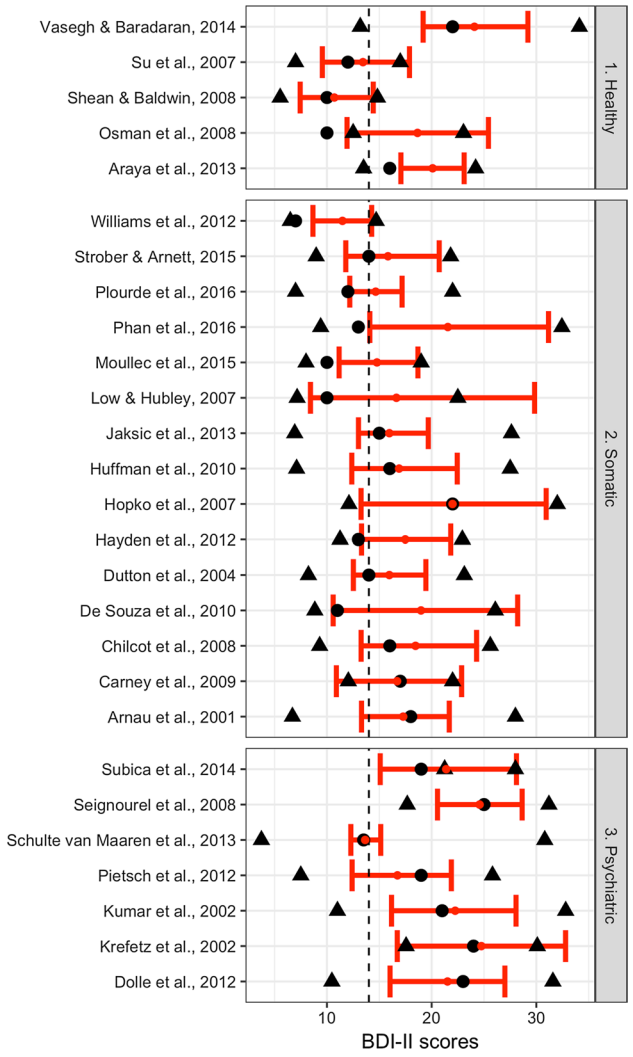

Fig. 2 Variability of cut points for the BDI-II across different studies. Black triangles indicate the means for patients and controls, the black dot indicates the cut point identified as optimal and the red error-bar indicates the range from the 2.5 to 97.5% quartile of the simulated cut points

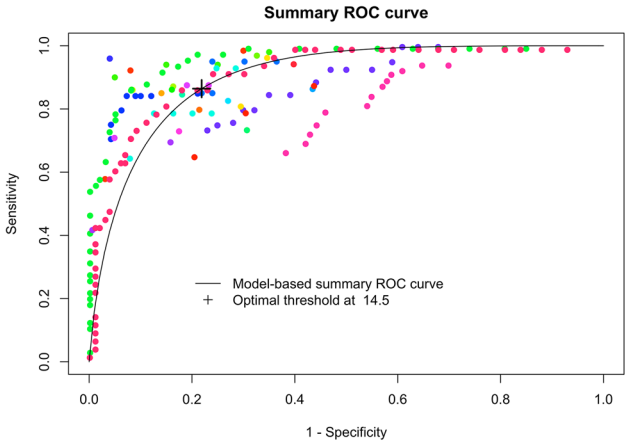

Fig. 3 Summary ROC curve for studies into the diagnostic utility of the BDI-II. Colors indicate the study that contributed the data points. (Color figure online)

without depression would be classified as depressed. While this is mathematically correct, it shows that the maximize Youden criterion may give counter intuitive results in some circumstances. Second, the present results revealed a large heterogeneity in the cut points that were identified as optimal across different studies. While previous reviews have also noted this large disparity [27], the present study is the first to estimate how much of this is due to sampling variability and how much is due to real underlying differences. The results of the simulation study seem to indicate that a lot of this variability is due to sampling variability. That is, one might identify several conflicting cut points as optimal in different samples drawn from a population that is similar to the sample. For example, one study [14] comprised only six patients with depression. If we repeatedly draw samples of six patients, it comes as no surprise that the estimated means in these samples differ widely. Accordingly, the cut points that were identified as optimal in our simulations ranged from 8 to 30 for this particular study. In our view, this result indicates that the findings from individual studies with small sample sizes should be treated with extreme care. Another corollary of the post hoc choice of optimal cut points is that many studies overestimate the true diagnostic utility of a specific measure [10, 12]. For research aimed at establishing optimal cut points for diagnostic methods, it is extremely important to make use of methods to estimate the variability of cut points. Unfortunately, existing guidelines such as STARD also only requires investigating variability in the diagnostic accuracy but not the variability in the cut points identified as optimal [5].

The meta-analysis of somatic and healthy samples revealed the cut points similar to the one proposed by Beck and colleagues [4]. This is remarkable given the fact that only two other empirical studies [6, 24] identified this same cut point as optimal and more than half of the studies identified cut points as optimal that were more than 3 points away from this score. However, based on the results of the present separate meta-analyses, it seems prudent to use different cut points depending on the sample that is being tested. Specifically, one should use much higher cut point in psychiatric samples than in somatic or healthy samples. While this makes intuitively sense and has been echoed by previous authors [22, 27], the present study is the first to offer some empirical support for this separation by inspecting the variability of cut points identified as optimal within individual studies and in a separate meta-analysis. Furthermore, we quantify how much larger optimal cut points need to be in psychiatric samples compared to healthy and somatic

samples. One explanation for the need for higher cutpoints in psychiatric patients is that many symptoms which are assessed as part of the BDI-II, e.g., sleeplessness, trouble with concentration, are also affected by other psychiatric diseases or side effects of psychotropic pharmaceuticals rather than specific symptoms of depression [8]. While more research is needed to test this or alternative explanations, the present results clearly show that the interpretation of BDI-II scores is much improved, if a cut point of 18.18 should be used in psychiatric samples and a cut point of 12.9 in healthy and somatic samples.

However, even cut points identified as optimal in the present meta-analysis are far from perfect. Our results show that a classification based on BDI-II scores alone would miss between 15 and 20% of persons with depression and would wrongfully classify about 20–25% as having a depression even though they do not. From a clinical perspective, this highlights the need for structured clinical interviews when diagnosing depression.

When interpreting the present study, a number of limitations have to be kept in mind. First, we included studies which used different reference tests to classify participants, and that in turn may affect their ability to detect symptoms of depression. In the present study, we decided against performing different analyses for these groups of studies because of the small number of studies in each group, but as more data become available, future meta-analysis may be able to test for these differences. Second, methods to deal with different aspects of diagnostic meta-analyses are currently being developed, each with specific advantages and disadvantages [19]. The implementation we choose [23] makes use of all available data points. However, it does not allow for testing of moderation effects, or the calculation of measures of heterogeneity.

To conclude, the present review identified a large number of studies aimed at establishing optimal cut points for the BDI-II that reported conflicting results. The simulation study we performed showed that these disparate results across studies are due to random fluctuations within studies. Researchers aiming to establish optimal cut points need to take into account random fluctuations in optimal cut points when trying to synthesize different studies [26]. Methods that explicitly model these random fluctuations such as simulations and diagnostic meta-analysis help to decide whether differences between studies reflect a real need for separate cut points or are only due to small sample sizes. Our results support the notion, that researchers using the BDI-II as an outcome measure in clinical studies should define remission as BDI-II scores below 13 since this yields the best classification in healthy and somatic samples. However, the results also indicate that using a cut point of 19 to indicate remission may be better among patients that suffer from comorbid psychiatric illnesses. Given the limitations described above, the latter point needs to be validated in future studies, that directly compare differences in optimal cut points for BDI-II scores in different recruitment settings. For clinicians, our results echo the call for higher cut points to interpret BDI-II scores in psychiatric samples than in somatic and healthy samples [26] in order to maximize its diagnostic utility.

Funding The study was supported by the German Federal ministry for Education and Research (BMBF #01EK1501) and the Witten/Herdecke University, Germany (#IFF2014-14).

References

1. Adachi, Y., Aleksic, B., Nobata, R., Suzuki, T., Yoshida, K., Ono, Y., & Ozaki, N. (2012). Combination use of Beck Depression Inventory and two-question case-finding instrument as a screening tool for depression in the workplace. *British Medical Journal Open, 2*(3), e000596. https://doi.org/10.1136/bmjopen-2011-000596.
2. Balogun, R. A., Turgut, F., Balogun, S. A., Holroyd, S., & Abdel-Rahman, E. M. (2011). Screening for depression in elderly hemodialysis patients. *Nephron Clinical Practice, 118*(2), c72–c77. https://doi.org/10.1159/000320037.
3. Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M. T., … Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine, 46*(16), 3359–3369.
4. Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II: Beck Depression Inventory manual* (2nd ed.). San Antonio: Psychological Corporation.
5. Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. … STARD Group. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ, 351*, h5527.
6. Dutton, G. R., Grothe, K. B., Jones, G. N., Whitehead, D., Kendra, K., & Brantley, P. J. (2004). Use of the Beck Depression Inventory-II with African American primary care patients. *General Hospital Psychiatry, 26*(6), 437–442. https://doi.org/10.1016/j.genhosppsych.2004.06.002.
7. Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., … Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010. *PLoS Medicine, 10*(11), e1001547.
8. Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine, 13*(1), 1.
9. Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry, 75*(4), 336–346.
10. Hirschfeld, G., & do Brasil, P. E. A. A. (2014). A simulation study into the performance of "optimal" diagnostic thresholds in the population: "Large" effect sizes are not enough. *Journal of Clinical Epidemiology, 67*(4), 449–453.
11. Kamenov, K., Caballero, F. F., Miret, M., Leonardi, M., Sainio, P., Tobiasz-Adamczyk, B., … Cabello, M. (2016). Which are the most burdensome functioning areas in depression? A cross-national study. *Frontiers in Psychology, 7*, 1342.
12. Leeflang, M. M. G., Moons, K. G. M., 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. *Clinical Chemistry, 54*(4), 729–737.
13. Lemmens, L., Arntz, A., Peeters, F., Hollon, S. D., Roefs, A., & Huibers, M. J. H. (2015). Clinical effectiveness of cognitive therapy v. interpersonal psychotherapy for depression: Results of a randomized controlled trial. *Psychological Medicine, 45*(10), 2095–2110.
14. Low, G. D., & Hubley, A. M. (2007). Screening for depression after cardiac events using the Beck Depression Inventory-II and the Geriatric Depression Scale. *Social Indicators Research, 82*(3), 527–543. https://doi.org/10.1007/s11205-006-9049-3.
15. McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.). New York: Oxford University.
16. Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: A meta-analysis. *The Lancet, 374*(9690), 609–619.
17. Osman, A., Barrios, F. X., Gutierrez, P. M., Williams, J. E., & Bailey, J. (2008). Psychometric properties of the Beck Depression Inventory-II in nonclinical adolescent samples. *Journal of Clinical Psychology, 64*(1), 83–102. https://doi.org/10.1002/jclp.20433.
18. Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology, 58*(10), 982–990. https://doi.org/10.1016/j.jclinepi.2005.02.022.
19. Reitsma, J. B., Rutjes, A. W. S., Khan, K. S., Coomarasamy, A., & Bossuyt, P. M. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology, 62*(8), 797–806. https://doi.org/10.1016/j.jclinepi.2009.02.005.
20. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109. https://doi.org/10.1016/j.jclinepi.2007.03.012.
21. Schulte-van Maaren, Y. W. M., Carlier, I. V. E., Zitman, F. G., van Hemert, A. M., de Waal, M. W. M., van der Does, A. J. W., … Giltay, E. J. (2013). Reference values for major depression questionnaires: The Leiden Routine Outcome Monitoring Study. *Journal of Affective Disorders, 149*(1–3), 342–349. https://doi.org/10.1016/j.jad.2013.02.009.
22. Seignourel, P. J., Green, C., & Schmitz, J. M. (2008). Factor structure and diagnostic efficiency of the BDI-II in treatment-seeking substance users. *Drug and Alcohol Dependence, 93*(3), 271–278. https://doi.org/10.1016/j.drugalcdep.2007.10.016.
23. Steinhauser, S., Schumacher, M., & Rücker, G. (2016). Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Medical Research Methodology*. https://doi.org/10.1186/s12874-016-0196-1.
24. Strober, L. B., & Arnett, P. A. (2015). Depression in multiple sclerosis: The utility of common self-report instruments and development of a disease-specific measure. *Journal of Clinical and Experimental Neuropsychology, 37*(7), 722–732. https://doi.org/10.1080/13803395.2015.1063591.
25. Subica, A. M., Fowler, J. C., Elhai, J. D., Frueh, B. C., Sharp, C., Kelly, E. L., & Allen, J. G. (2014). Factor structure and diagnostic validity of the Beck Depression Inventory–II with adult clinical inpatients: Comparison to a gold-standard diagnostic interview. *Psychological Assessment, 26*(4), 1106.
26. Wang, D., Tian, L., & Zhao, Y. (2017). Smoothed empirical likelihood for the Youden index. *Computational Statistics & Data Analysis*, *115*, 1–10.
27. Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. *Revista Brasileira de Psiquiatria, 35*(4), 416–431. https://doi.org/10.1590/1516-4446-2012-1048.
28. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., & QUADAS-2 Steering Group (2013). A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology, 66*(10), 1093–1104. https://doi.org/10.1016/j.jclinepi.2013.05.014.