

1 **Comparison of the accuracy of the 7-item HADS Depression subscale and 14-item total**
2 **HADS for screening for major depression: a systematic review and individual participant**
3 **data meta-analysis**

4
5 **Abstract**

6 The 7-item Hospital Anxiety and Depression Scale Depression subscale (HADS-D) and
7 the total score of the 14-item HADS (HADS-T) are both used for major depression screening.
8 Compared to the HADS-D, the HADS-T includes anxiety items and requires more time to
9 complete. We compared the screening accuracy of the HADS-D and HADS-T for major
10 depression detection. We conducted an individual participant data meta-analysis and fit bivariate
11 random-effects models to assess diagnostic accuracy among participants with both HADS-D and
12 HADS-T scores. We identified optimal cutoffs, estimated sensitivity and specificity with 95%
13 confidence intervals (CIs), and compared screening accuracy across paired cutoffs via two-stage
14 and individual-level models. We used a 0.05 equivalence margin to assess equivalency in
15 sensitivity and specificity. 20,700 participants (2,285 major depression cases) from 98 studies
16 were included. Cutoffs of ≥ 7 for the HADS-D (sensitivity 0.79 [0.75, 0.83], specificity 0.78
17 [0.75, 0.80]) and ≥ 15 for the HADS-T (sensitivity 0.79 [0.76, 0.82], specificity 0.81 [0.78,
18 0.83]) minimized the distance to the top-left corner of the receiver operating characteristic curve.
19 Across all sets of paired cutoffs evaluated, differences of sensitivity between HADS-T and
20 HADS-D ranged from -0.05 to 0.01 (0.00 at paired optimal cutoffs), and differences of
21 specificity were within 0.03 for all cutoffs (0.02 to 0.03). The pattern was similar among
22 outpatients, although the HADS-T was slightly (not non-equivalently) more specific among

23 inpatients. The accuracy of HADS-T was equivalent to the HADS-D for detecting major
24 depression. In most settings, the shorter HADS-D would be preferred.

25 **Keywords:** HADS-D, HADS-T, individual participant data meta-analysis, depression
26 screening, diagnostic accuracy

27 **Public significance statements:**

28 The present study suggests that the accuracy of 14-item Hospital Anxiety and Depression Scale
29 (HADS-D) and the 7-item HADS Depression subscale (HADS-D) are equivalent for detecting
30 major depression. Using the 7-item HADS-D for depression screening instead of the full 14-item
31 HADS-T has minimal influence on performance of the measure but would reduce patient and
32 participant burden in most clinical and research settings.

33
34
35

36 © 2022, American Psychological Association. This paper is not the copy of record and may
37 not exactly replicate the final, authoritative version of the article. Please do not copy or cite
38 without authors' permission. The final article will be available, upon publication, via its
39 DOI: 10.1037/pas0001181

40 The 14-item Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983)
41 was developed to facilitate the identification of anxiety disorders and major depression in people
42 with a physical illness. The HADS includes two subscales. The 7-item Depression subscale
43 (HADS-D) was designed to assess continuous depressive symptoms and for depression
44 screening, whereas the 7-item Anxiety subscale (HADS-A) was designed to assess and screen for
45 anxiety (Zigmond & Snaith, 1983). Both HADS-D and full HADS total scores (HADS-T) have
46 been used to screen for major depression (Mitchell, Meader, & Symonds, 2010; Vodermaier &
47 Millman, 2011). The HADS-T takes more time to complete and includes anxiety items not
48 specific to depression. Some have suggested, though, that anxiety symptoms should be
49 considered when assessing depression (Schatzberg, 2019). Furthermore, previous reviews have
50 provided some preliminary evidence that HADS-T may perform better than the HADS-D
51 (Mitchell, Meader, & Symonds, 2010; Vodermaier & Millman, 2011).

52 Commonly used HADS-D cutoff thresholds of ≥ 8 for “possible” depression and ≥ 11 for
53 “probable” depression were established in the original validation study, which included only 100
54 participants (11 depression cases) (Zigmond & Snaith, 1983). A recent individual participant
55 data meta-analysis (IPDMA) on HADS-D accuracy to screen for major depression (101 studies;
56 25,574 participants; 2,549 major depression cases) found that a cutoff of ≥ 7 maximized
57 combined sensitivity and specificity across reference standards; standard cutoffs of ≥ 8 and ≥ 11
58 were less sensitive but more specific (Wu, Levis, Sun, et al., 2021). There is not a standard cutoff
59 for screening to detect major depression with the HADS-T.

60 Two previous meta-analyses, both done with studies of cancer patients, have indirectly
61 compared the HADS-D and HADS-T for detecting major depression (Mitchell et al., 2010;
62 Vodermaier & Millman, 2011). Both searched through October 2009 for eligible studies. One

63 evaluated 9 studies that used the HADS-D with a cutoff of 8 or greater and 6 studies that used
64 the HADS-T with a cutoff of 15 (number of participants not reported) (Mitchell et al., 2010),
65 whereas the other included 2-5 studies each in analyses of HADS-D cutoffs of 7, 9, and 11 and
66 HADS-T cutoffs of 15, 17, 19 and 20 (470 to 872 participants per analysis) (Vodermaier &
67 Millman, 2011). Both meta-analyses suggested that the HADS-T may perform better than the
68 HADS-D, but there was a high level of uncertainty due to indirect comparisons between
69 participants from different studies that reported HADS-D and HADS-T results, the small number
70 of total participants, and possible selective outcome reporting bias (Levis et al., 2017; Neupane
71 et al., 2021; Rice & Thombs, 2016; Thombs et al., 2011; Thombs & Rice, 2016) since not all
72 primary studies reported results from the same cutoffs.

73 Using the full 14-item HADS-T for depression screening would be warranted if it is
74 sufficiently more accurate than the shorter 7-item HADS-D to justify the additional time and
75 patient burden involved. We previously assessed the accuracy of the HADS-D using IPDMA
76 (Wu, Levis, Sun, et al., 2021). IPDMA involves a standard systematic review, followed by
77 synthesis of original research data from primary studies, rather than extracting summary data
78 (Riley, Lambert, & Abo-Zaid, 2010). In that IPDMA, we found that diagnostic accuracy of
79 HADS-D was not significantly different for any cutoffs across reference standards based on
80 participant characteristics, including age, sex, cancer diagnosis, country human development
81 index levels, participant recruitment settings, or the study's risk of bias ratings (Wu et al., 2021).
82 In the present study, we included studies from the HADS-D IPDMA where HADS-T scores were
83 provided or could be calculated from individual item scores. Our objectives were to (1) directly
84 compare screening accuracy of the HADS-T and HADS-D for major depression detection using
85 the same participant data across all studies regardless of reference standard, and (2) replicate the

86 comparison among studies that used a semi-structured diagnostic interview [e.g., Structured
87 Clinical Interview for the DSM (SCID) (First, 1995)] as a reference standard, since semi-
88 structured interviews more closely reflect the actual diagnostic process than fully-structured
89 interviews.

90 **Methods**

91 The present study used a subset of studies and participants from our previously conducted
92 HADS-D IPDMA (Wu, Levis, Sun, et al., 2021) for which HADS-T scores were also available.
93 Analyses of HADS-D and HADS-T diagnostic accuracy were conducted according to the
94 HADS-D IPDMA methods (Wu, Levis, Sun, et al., 2021) with the addition of analyses to
95 directly compare HADS-D and HADS-T accuracy.

96 **Dataset eligibility**

97 For the main HADS-D meta-analysis, datasets from articles in any language were eligible
98 for inclusion if (1) they included diagnostic classification for current Major Depressive Disorder
99 (MDD) or Major Depressive Episode (MDE) using Diagnostic and Statistical Manual of Mental
100 Disorders (DSM) (American Psychiatric Association, 1987; 1994; 2000; 2013) or International
101 Classification of Diseases (ICD) (World Health Organization (WHO), 1992) criteria based on a
102 validated semi-structured or fully structured interview; (2) they included total scores for the
103 HADS-D; (3) the diagnostic interview and HADS-D were administered within two weeks of
104 each other, because DSM and ICD major depression diagnostic criteria specify that symptoms
105 must have been present in the last two weeks; (4) participants were ≥ 18 years of age and not
106 recruited from youth or psychiatric settings; and (5) participants were not recruited because they
107 were identified as having symptoms of depression, since screening is done to identify previously
108 unrecognized cases. We focused on MDD and MDE because major guidelines on depression

109 screening have focused on screening for major depression but have not considered screening for
110 less severe conditions, such as dysthymia or persistent depressive disorder, for which treatment
111 options and effectiveness are much less well delineated (Joffres et al., 2013; National
112 Collaborating Centre for Mental Health (UK), 2010; Siu & US Preventive Services Task Force,
113 2016). Consistent with this, few primary studies collect or report diagnostic status for dysthymia
114 or persistent depressive disorder. Datasets where not all participants were eligible were included
115 if primary data allowed selection of eligible participants. For the present study, we only included
116 primary datasets from the HADS-D IPDMA that also provided HADS-T scores or item scores to
117 calculate HADS-T scores.

118 **Search strategy and study selection**

119 A medical librarian searched Medline, Medline In-Process & Other Non-Indexed Citations
120 and PsycINFO via OvidSP, and Web of Science via ISI Web of Knowledge from inception to
121 October 25, 2018 using a peer-reviewed (McGowan, Sampson, Salzwedel, Cogo, Foerster, &
122 Lefebvre, 2016) search strategy (Supplementary Methods A). We also reviewed reference lists of
123 relevant reviews and queried contributing authors about non-published studies. Search results
124 were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication,
125 unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada) for
126 tracking search results.

127 Pairs of investigators independently reviewed titles and abstracts for eligibility. If either
128 deemed a study potentially eligible, full-text review was done by two investigators,
129 independently, with disagreements resolved by consensus, consulting a third investigator when
130 necessary. Translators were consulted for languages other than those for which team members
131 were fluent.

132 **Data contribution, extraction, and synthesis**

133 Authors of eligible datasets were invited to contribute de-identified primary data. We
134 emailed corresponding authors of eligible primary studies at least three times, as necessary. If we
135 did not receive a response, we emailed co-authors and attempted to contact corresponding
136 authors by phone.

137 Diagnostic interview and country were extracted from published reports by pairs of
138 investigators independently, with disagreements resolved by consensus. Countries were
139 categorized as “very high”, “high” or “low-medium” development based on the United Nation’s
140 Human Development Index (HDI) for the country for the year of the study publication. The HDI
141 is a statistical composite index that includes indicators of life expectancy, education, and income
142 (United Nations Development Programme, 2020). Participant-level data included age, sex,
143 participant recruiting setting, HADS-D scores, HADS-T scores, and major depression status
144 (case or non-case). For defining major depression, we considered MDD or MDE based on the
145 DSM or ICD. If more than one was reported, we prioritized MDE over MDD (because screening
146 would attempt to detect depressive episodes and further interview would determine if the episode
147 is related to MDD, bipolar disorder or persistent depressive disorder). We also prioritized DSM
148 over ICD because most studies use DSM criteria.

149 Individual participant data were converted to a standard format and synthesized into a
150 single dataset with study-level data. We compared published participant characteristics and
151 diagnostic accuracy estimates with results from raw datasets and resolved any discrepancies in
152 consultation with primary study investigators.

153 **Risk of Bias Assessment**

154 Risk of bias of included studies was assessed by two investigators independently using
155 the QUality Assessment of Diagnostic Accuracy Studies-2 tool (QUADAS-2; Supplementary
156 Methods B) (Whiting et al., 2011). Any discrepancies were resolved via consensus with a third
157 investigator involved as necessary. Risk of bias was coded at both study and participant levels
158 since some classifications (e.g., the time between index test and reference standard) may have
159 differed among participants from the same study. The QUADAS-2 results were used to describe
160 the risk of bias of each included study.

161 **Statistical Analyses**

162 To compare the screening accuracy of the HADS-D and HADS-T across relevant cutoffs to
163 detect major depression, we first estimated overall sensitivity and specificity for HADS-D and
164 HADS-T by combining all studies regardless of reference standard. Reference standards used in
165 primary studies included semi-structured interviews (e.g., SCID (First, 1995)), fully structured
166 interviews (the Mini International Neuropsychiatric Interview (MINI) excluded) (e.g., Composite
167 International Diagnostic Interview (CIDI) (Robins et al., 1988)), and the MINI (Lecrubier et al.,
168 1997; Sheehan et al., 1997). Different types of reference standards have different design and
169 performance characteristics (Levis, Benedetti, et al., 2019; Levis et al., 2020; Wu, Levis,
170 Ioannidis, et al., 2021; Wu, Levis, Sun, et al., 2020), and estimates of sensitivity and specificity
171 differ by type (Negeri, et al., 2021; Levis, Benedetti, et al., 2019; Levis et al., 2020; Wu, Levis,
172 Sun, et al., 2021). It is reasonable to assume, though, that differences in sensitivity and
173 specificity between HADS-D and HADS-T accuracy among the same participants are not
174 associated with reference standard type, since in each primary study the HADS-D and HADS-T
175 were compared to the same reference standard. Thus, our main analysis included all studies
176 regardless of reference standard.

177 Separately, as a sensitivity analysis, to ensure that results would not differ by clinical
178 interview, we repeated all analyses for only studies that used a semi-structured interview as the
179 reference standard. Semi-structured interviews (e.g., SCID (First, 1995), Schedules for Clinical
180 Assessment in Neuropsychiatry (WHO, 1994), Schedule for Affective Disorders and
181 Schizophrenia (Endicott & Spitzer, 1987), and Monash Interview for Liaison Psychiatry (Clarke,
182 Smith, Herrman, & McKenzie, 1998)) are intended to be administered by experienced
183 diagnosticians and are considered to more closely reflect clinical diagnostic procedures than fully
184 structured interviews or the MINI (Brugha, Bebbington, & Jenkins, 1999; Brugha, Jenkins, Taub,
185 Meltzer, & Bebbington, 2001; Nosen & Woody, 2008). We did not conduct additional sensitivity
186 analyses with fully structured interviews or the MINI.

187 Overall and separately, for studies that used a semi-structured reference standard, for all
188 possible cutoffs 0-21 of the HADS-D and 0-42 of the HADS-T, we fitted bivariate random-
189 effects models via Gauss-Hermite quadrature (Riley, Dodd, Craig, Thompson, & Williamson,
190 2008). This is a two-stage meta-analytic approach that models sensitivity and specificity
191 simultaneously and accounts for the correlation between them and the precision of estimates
192 within studies. We also constructed empirical receiver operating characteristic (ROC) plots based
193 on pooled sensitivity and specificity estimates and calculated area under the curves (AUC) for
194 the two tests.

195 To investigate heterogeneity across studies, overall and for studies with a semi-structured
196 reference standard, we generated forest plots for the differences in sensitivity and specificity
197 estimates between the HADS-D and HADS-T for the optimal cutoffs based on pooled results.
198 We also quantified heterogeneity at the optimal cutoffs for the HADS-D and HADS-T by
199 reporting the estimated variances of the random effects for the differences in the HADS-D and

200 HADS-T sensitivity and specificity (τ^2) (Fagerland, Lydersen, & Laake, 2014; Higgins &
201 Thompson, 2002).

202 To compare the diagnostic accuracy of the HADS-D and HADS-T, using the analyses
203 that pooled across reference standards and within semi-structured reference standard category,
204 we first calculated the differences of the AUCs with 95% confidence intervals (CIs). Second, we
205 compared the ROC plots visually to determine if one measure consistently perform better than
206 the other across cutoffs. Third, we compared differences in sensitivity and specificity for optimal
207 cutoffs and other cutoffs close to the optimal cutoff to determine if there were differences and the
208 magnitude of any differences. To do this, we identified the optimal cutoff that minimized the
209 values of the distance to the top-left corner of the ROC curves (NCSS, 2017) for both HADS-D
210 and HADS-T and a set of other cutoffs that were close to the optimal cutoff. The distance to the
211 top-left corner of the ROC curve for each cutoff value is calculated by $d =$
212 $\sqrt{(1-\text{Sensitivity})^2 + (1-\text{Specificity})^2}$ (NCSS, 2017). Since there is no *a priori* method to align
213 cutoffs on the HADS-D and HADS-T that perform most similarly in terms of sensitivity and
214 specificity, we did this based on examination of results and consensus among investigators.
215 Then, we compared the sensitivity and specificity between the HADS-D and HADS-T for pairs
216 of optimal cutoffs and four other pairs of cutoffs close to the optimal; the interval between
217 cutoffs for HADS-T was 2 instead of 1 because HADS-T doubled the length and the total score
218 of HADS-D. For all cutoffs on the HADS-D and HADS-T, 95% CIs for the differences between
219 HADS-D and HADS-T sensitivity and specificity were constructed via a cluster bootstrap
220 approach (Van der Leeden, Busing, & Meijer, 1997; Van der Leeden, Meijer, & Busing, 2008)
221 with resampling at the study and subject level. For each comparison, we ran 1000 iterations of
222 the bootstrap. For each bootstrap iteration, the bivariate random-effects model was fitted to the

223 HADS-D and HADS-T data, and the pooled sensitivities and specificities were computed
224 separately, as described above, for all cutoffs of HADS-D and HADS-T.

225 In addition to comparing the HADS-D and HADS-T with pooling of study-level results,
226 as a sensitivity analysis, we compared sensitivity and specificity of the HADS-D and HADS-T
227 across cutoffs via an individual-level analysis. For the individual-level analysis, for each pair of
228 matched HADS-D and HADS-T cutoffs, we fitted a linear mixed model with the difference
229 between the HADS-D and HADS-T screening results as the outcome. The screening result is
230 dichotomous, either positive = 1 or negative = 0. If the HADS-T screening result was positive
231 (which was 1), but HADS-D was negative (which was 0), the outcome, i.e., the difference
232 between HADS-T and HADS-D results, was $1 - 0 = 1$; if both screening results were positive or
233 negative, the outcome was 0 ($1 - 1$ or $0 - 0$); and if the HADS-T screening result was negative,
234 but HADS-D was positive, the outcome was -1 ($0 - 1 = -1$). This model modeled the differences
235 in sensitivity and specificity simultaneously and included random effects both at the study level.
236 From this model, for each set of HADS-D and HADS-T paired cutoffs, we estimated the
237 difference in sensitivity and specificity between the two tests and associated CIs. These CIs from
238 the bootstrap approach and individual-level analysis allowed us to test whether the sensitivity
239 and specificity of the HADS-T is equivalent to that of the HADS-D based on a pre-specified
240 equivalence margin of $\delta = 0.05$ (Walker & Nowacki, 2011), as we have done in previous studies
241 (Harel et al., 2021; Ishihara et al., 2019; Wu, Levis, Riehm, et al., 2020).

242 As a sensitivity analysis, we compared accuracy of HADS-D and HADS-T results
243 stratified by subgroups based on inpatient and outpatient care settings (we planned to conduct
244 sensitivity analysis in each participant recruit setting, separately, but we were able to do this only
245 for inpatient and outpatient medical settings because there were too few participants from non-

246 medical and mixed inpatient/outpatient settings). In addition, we conducted a subgroup analysis
247 only among patients from cancer studies because meta-analyses (Mitchell et al., 2010;
248 Vodermaier & Millman, 2011) of studies from cancer care settings reported that the HADS-T
249 may perform better than the HADS-D in those settings. We did not conduct the sensitivity
250 analysis to assess whether inclusion of published results from the eligible studies that did not
251 provide raw data influenced results because we did this in the main HADS-D IPDMA and found
252 no differences (Wu et al., 2021).

253 To examine whether measurement differences across participant characteristics,
254 including country, may have influenced our results, we assessed whether sensitivity and
255 specificity differed for the HADS-D based on these characteristics, and then, we re-examined
256 HADS-D and HADS-T differences for any variables where differences were found. To assess
257 possible influences on sensitivity and specificity, we conducted one-stage meta-regressions. In
258 the first step, we repeated the analysis that we did in the main HADS-D IPDMA by interacting
259 all subgrouping variables (age [measured continuously], sex [reference category = female]),
260 country HDI level [reference category = very high], cancer diagnosis [reference category = no],
261 participant recruiting setting [reference category = inpatient specialty care], interactions of
262 QUADAS-2 signaling item responses [reference category = low risk] with logit (sensitivity) and
263 logit (1 – specificity) of the HADS-D (Wu et al., 2021). We conducted these analyses separately
264 by reference standards (semi-structured interview, fully structured interview, MINI), since these
265 types of interviews have been shown to identify different individuals (Wu et al., 2021). In the
266 second step, we added country/language variables to the model (Germany, Spain, Lithuania,
267 Norway, Korea, Japan [reference category = English speaking countries]). These models were
268 restricted to the subset of the studies from countries with more than 500 participants that had

269 complete data for all relevant variables and used a semi-structured interview or the MINI (there
270 were not enough data for the studies that used a fully structured reference standard). Country
271 HDI level was dropped from the model because all countries included in this analysis had very
272 high HDI. For any variables that were found to be associated with the sensitivity or specificity
273 across all cutoffs, we compared accuracy of HADS-D and HADS-T results stratified by
274 subgroups based on these variables.

275 All analyses were run in R (R version R 3.5.0 (R Core Team, 2020) and R Studio
276 version 1.1.423 (RStudio Team, 2020)) using the lme4 package (Bates, Maechler, Bolker, &
277 Walker, 2015).

278 **Registration and Protocol**

279 The main HADS-D IPDMA was registered in PROSPERO (CRD42015016761), and a
280 protocol was published (Thombs et al., 2016). The present study was not included in the protocol
281 for the main HADS-D IPDMA, but a separate protocol was developed and posted online prior to
282 initiating the study (<https://osf.io/438ak/>).

283 **Data Availability**

284 Data contribution agreements with primary study authors do not include permission to
285 make their data publicly available, although the dataset used in this study will be archived
286 through a McGill University repository (Borealis,
287 <https://borealisdata.ca/dataverse/depressdproject/>). The R codes used for the analysis will be
288 made publicly available through the same repository. Requests to access the dataset to verify
289 study results but not for other purposes can be sent to the corresponding authors via the “Access
290 Dataset” function on the repository website.

291 **Results**

292 **Search Results and Inclusion of Primary Data**

293 For the main HADS-D IPDMA, of 14,465 unique titles and abstracts identified from the
294 database search, 13,895 were excluded after title and abstract review and 330 after full-text
295 (Supplementary Table A), leaving 240 eligible articles with data from 165 unique participant
296 samples (Supplementary Figure A). Of the 165 unique samples, 93 (56%) contributed data (66%
297 of eligible participants). In addition, authors of included studies contributed data from 10 studies
298 that were unpublished or did not come up in the search, for a total of 103 HADS-D datasets
299 contributed to our IPDMA. Five studies without HADS individual item scores or separate total
300 scores for the HADS-D and HADS-T were excluded from the present study (see Supplementary
301 Table B2). Thus, 20,700 participants (2,285 major depression cases) from 98 studies were
302 analyzed (91% of 22,755 participants from the 103 HADS-D datasets). Included study
303 characteristics are shown in Supplementary Table B1. Characteristics of eligible studies that did
304 not provide data, including the five studies excluded because they only provided HADS-D or
305 HADS-T total scores, are shown in Supplementary Table B2.

306 Of 98 included studies, 58 used semi-structured interviews to assess major depression
307 (10,311 participants), including 54 that used the SCID (9,676 participants); 31 used the MINI
308 (7,445 participants); and 9 used other. Participant characteristics are shown in Table 1.

309 Supplementary Table C shows QUADAS-2 ratings for included studies. There were only
310 11 studies with “low” risk of bias rating across all QUADAS-2 domains.

311 **Comparison of Screening Accuracy Between the HADS-D and HADS-T**

312 ROC plots comparing sensitivity and specificity estimates for all cutoffs between the
313 HADS-D (0-21) and HADS-T (0-42) among all included studies are shown in Figure 1. A large
314 part of the plots for the HADS-D and HADS-T were overlapping. The HADS-T performed better

315 than HADS-D at some cutoffs, but this pattern was not consistent across cutoffs. The AUCs for
316 the HADS-D and HADS-T were similar among all studies (0.853 versus 0.872). We also
317 compared the ROCs among studies that used a semi-structured reference standard and found a
318 similar pattern (Supplementary Figure B).

319 Based on the pooled sensitivity and specificity across all HADS-D and HADS-T cutoffs,
320 among all studies, the cutoff that minimized the values of the distance to the top-left corner of
321 the ROC curves was ≥ 7 for the HADS-D (sensitivity [95% CI] = 0.79 [0.75, 0.83], specificity
322 [95% CI] = 0.78 [0.75, 0.80]) and ≥ 15 for the HADS-T (sensitivity [95% CI] = 0.79 [0.76,
323 0.82], specificity [95% CI] = 0.81 [0.78, 0.83]) (Table 2).

324 The comparison of sensitivity and specificity between the HADS-D and HADS-T for the
325 optimal cutoffs (HADS-D ≥ 7 vs. HADS-T ≥ 15) and other cutoffs close to the optimal cutoffs (\geq
326 5 vs. ≥ 11 ; ≥ 6 vs. ≥ 13 ; ≥ 8 vs. ≥ 17 ; ≥ 9 vs. ≥ 19 ; ≥ 10 vs. ≥ 21 ; and ≥ 11 vs. ≥ 23 are presented
327 in Table 2. Overall, for the pairs of optimal cutoffs or other cutoffs close to the optimal, the
328 differences in sensitivity and specificity between HADS-D and HADS-T using the bootstrapping
329 approach across all 98 primary studies were small. Precision of estimates was high, and the
330 width of 95% CIs ranged from 5% to 9% for sensitivity and 2% to 4% for specificity across all
331 cutoffs examined. For sensitivity, the differences of HADS-T – HADS-D for all pairs of cutoffs
332 were not statistically significant (the differences were between -0.05 and 0.01, CIs were within
333 or overlapped with the range of -0.05 and 0.05). Therefore, at five pairs of optimal cutoffs or
334 other cutoffs close to the optimal, the sensitivity of the HADS-T was equivalent to that of the
335 HADS-D; the equivalency was indeterminant on the other two pairs, based on the pre-specified
336 equivalence margin of $\delta = 0.05$. For specificity, estimates of HADS-T were equivalent to HADS-
337 D for all seven pairs of cutoffs (the differences of HADS-T – HADS-D were between 0.02 and

338 0.03; CIs were all within -0.05 and 0.05). Relevant results among studies that used a semi-
339 structured reference standard were consistent with overall estimates (Supplementary Table D1).

340 The comparison of results via individual-level analysis are presented in Table 3. For each
341 pair of matched HADS-D and HADS-T cutoffs, the differences in sensitivity and specificity
342 between the two tests were similar to those from the bivariate random-effects models. This was
343 also true among studies that used a semi-structured reference standard (Supplementary Table
344 D2).

345 Among participants in inpatient care settings (Table 4a; 8,827 participants from 38
346 studies), the comparison results of HADS-T – HADS-D in sensitivity were similar to the overall
347 estimates; the differences in specificity were slightly larger than overall estimates, however, the
348 95% CIs generally overlapped with -0.05 and 0.05 and were classified as indeterminate to
349 equivalency, with one exception (HADS-D \geq 6 vs. HADS-T \geq 13) for which HADS-T specificity
350 was greater than for the HADS-D. The comparison results among participants in outpatient care
351 settings (Table 4b; 9,547 participants from 54 studies) and participants from studies done in
352 cancer care settings (Supplementary Table E; 5608 participants from 23 studies) were similar to
353 overall estimates. Within the semi-structured reference standard category, similar patterns were
354 found (Supplementary Tables D3 and D4).

355 The meta-regression results indicated no significant differences in sensitivity and
356 specificity were found for any individual participant characteristics or risk of bias ratings
357 (Supplementary Table F1-F3). After adding the country/language variables to the model, the
358 sensitivity and specificity of HADS-D was invariant based on all variables across reference
359 standards except that specificity estimates of the HADS-D were associated with Germany and
360 Spain among studies that used a semi-structured reference standard; specifically, the HADS-D

361 had lower specificity among participants from Germany and Spain compared to studies done
362 with participants from English speaking countries (Supplementary Table G1-G2).

363 Therefore, we conducted subgroup analysis of our comparisons of HADS-D and HADS-T
364 accuracy for participants from Germany or Spain. For each pair of matched HADS-D and
365 HADS-T cutoffs among participants from Germany (Supplementary Table H1), the comparison
366 results of HADS-T – HADS-D in sensitivity and specificity were similar to the overall estimates;
367 among participants from Spain (Supplementary Table H2), differences in specificity were
368 slightly larger than overall estimates, however, the 95% CIs all overlapped with -0.05 and 0.05
369 and were classified as indeterminate to equivalent, and differences in sensitivity were similar to
370 the overall estimates.

371 A forest plot of the differences of sensitivity and specificity estimates for HADS-D ≥ 7 vs.
372 HADS-T ≥ 15 across all studies is shown in Figure 2. At the optimal cutoffs, there was low
373 heterogeneity in the differences between HADS-D and HADS-T across the 98 studies with
374 estimated inter-study heterogeneity (τ^2) < 0.01 for sensitivity and < 0.01 for specificity. The
375 forest plot of the differences of sensitivity and specificity estimates at optimal cutoffs for the
376 HADS-D and HADS-T among studies that used a semi-structured reference standard is shown in
377 Supplementary Figure C.

378 Discussion

379 We assessed the equivalency of screening accuracy of the HADS-D and HADS-T across
380 all cutoffs to detect major depression and compared accuracy across paired optimal cutoffs and
381 other cutoffs close to the optimal cutoffs to test whether the HADS-T is superior to HADS-D for
382 major depression detection. There were two main findings. First, among all 98 included studies
383 the values of the distance to the top-left corner of the ROC curves (Riley et al., 2008) were

384 minimized at a HADS-D cutoff ≥ 7 (sensitivity = 0.79, specificity = 0.78) and at a HADS-T
385 cutoff ≥ 15 (sensitivity = 0.79, specificity = 0.81). Second, at paired optimal cutoffs and six other
386 cutoffs close to the optimal cutoffs, the HADS-D was similarly accurate compared to the HADS-
387 T overall and among studies that used a semi-structured reference standard.

388 Overall, for all 98 primary studies, across all sets of paired cutoffs, the sensitivity and
389 specificity of the HADS-T were classified as equivalent to that of the HADS-D based on the pre-
390 specified equivalency margin. Although the HADS-T was slightly more specific (range 0.02 to
391 0.03), all the 95% CIs for differences in sensitivity and specificity of HADS-T – HADS-D were
392 within or overlapped with the range of -0.05 and 0.05. When we analyzed data separately among
393 studies that used a semi-structured reference standard, differences in sensitivity and specificity
394 between the HADS-D and HADS-T were similar to the overall estimates.

395 Furthermore, similar to overall estimates, there were no substantive differences in
396 performance between the HADS-D and HADS-T in detecting major depression among medical
397 outpatients. Among inpatients, the HADS-T and HADS-D were also equivalent in sensitivity.
398 The HADS-T performed slightly better than HADS-D in terms of specificity, and equivalency
399 was indeterminate based on the pre-specified equivalence margin, except for one pair of cutoffs.
400 This finding is possibly related to the greater presence of anxiety symptoms in inpatients versus
401 outpatients and its relationship to depression (Schatzberg, 2019).

402 Previous conventional meta-analyses of results from cancer patients (Mitchell et al.,
403 2010; Vodermaier & Millman, 2011) suggested that the HADS-T may perform better than the
404 HADS-D, but that conclusion was highly uncertain given the limitations of the samples and
405 methods. Through our IPDMA, with its large dataset and more rigorous comparison methods
406 including both bivariate random-effects models and individual-level models, a two-level

407 bootstrap approach (Fagerland et al., 2014; Higgins & Thompson, 2002), and subgroup analysis,
408 we found there was no consistent evidence that the HADS-T is superior to HADS-D for major
409 depression detection, including in cancer care settings. In addition, we did not identify any
410 differences between HADS-D and HADS-T accuracy that were associated with individual
411 participant characteristics or countries. Therefore, in research and clinical general practice, using
412 the full 14-item HADS-T for depression screening would likely result in no to minimal gain in
413 screening accuracy but would add unnecessary burden to patients compared to the 7-item
414 HADS-D.

415 To our knowledge, this is the first meta-analysis that directly compared the HADS-D and
416 HADS-T for screening for depression using the same large individual participant dataset for both
417 screening tools. Strengths of this study included the large overall sample size and high precision
418 of estimates of differences, the ability to compare results for HADS-D and HADS-T across all
419 cutoffs from all studies, and the ability to assess screening accuracy overall and by inpatient and
420 outpatient subgroups. There are also limitations to consider. First, for the full IPDMA data,
421 primary data from 72 of 165 published eligible datasets (44% of datasets, 34% of participants)
422 were not included, and only those datasets with complete data for all individual HADS item
423 scores (91% of available data) were included in this study. Nonetheless, this sample was much
424 larger than the few primary studies that have previously compared the HADS-D and HADS-T.
425 Second, we did not conduct analyses restricted to studies with “low” risk of bias ratings across
426 QUADAS-2 domains. However, in sensitivity analysis in this study and in our main IPDMA on
427 the HADS-D (Wu, et al., 2021), risk of bias ratings were not associated with screening accuracy.
428 Third, the present study used a subset of studies and participants from our previously conducted
429 HADS-D IPDMA (Wu, et al., 2021). This IPDMA project was designed to assess the accuracy

430 of the HADS-D for detecting major depression. Diagnoses of other mental disorders, including,
431 anxiety disorders, were not collected in most of the included primary studies. Thus, we were not
432 able to evaluate the sensitivity and specificity of the HADS-D, HADS-Anxiety, or HADS-T for
433 detecting mental disorders generally. Forth, we did not record inter-rated reliability for risk of
434 bias ratings; however, all ratings were done by trained reviewers and any disagreements were
435 addressed by consensus, including a third investigator as necessary.

436 **Conclusions**

437 In summary, this study found that sensitivity and specificity of the HADS-T were not
438 superior to the HADS-D for detecting major depression in a large individual participant dataset.
439 Using the 7-item HADS-D for depression screening instead of the full 14-item HADS-T has
440 minimal influence on performance of the measure but would reduce patient and participant
441 burden in clinical and research settings. Both HADS-D and HADS-T have only modest
442 screening ability and discussion of their exact indications for use and related caveats are beyond
443 the scope of this article. However, there were no substantive differences in performance between
444 the HADS-D and HADS-T in detecting major depression among medical outpatients, although
445 there was a slight advantage in specificity of indeterminate equivalency for the HADS-T among
446 medical inpatients, for whom adding the anxiety items of HADS-A may improve accuracy.

447

448 **Ethical Approval:** As this study involved secondary analysis of anonymized previously
449 collected data, the Research Ethics Committee of the Jewish General Hospital declared that this
450 project did not require research ethics approval. However, for each included dataset, we
451 confirmed that the original study received ethics approval and that all patients provided informed
452 consent.

453 **REFERENCES**

- 454 *Akechi, T., Okuyama, T., Sugawara, Y., Shima, Y., Furukawa, T. A., & Uchitomi, Y. (2006).
455 Screening for depression in terminally ill cancer patients in Japan. *Journal of Pain &*
456 *Symptom Management* , 31(1), 5-12.
- 457 *Al-Asmi, A., Dorvlo, A. S., Burke, D. T., Al-Adawi, S., Al-Zaabi, A., Al-Zadjali, H. A., ... &
458 Al-Adawi, S. (2012). The detection of mood and anxiety in people with epilepsy using
459 two-phase designs: experiences from a tertiary care centre in Oman. *Epilepsy Research*,
460 98(2-3), 174–181.
- 461 American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders:
462 DSM-III. 3rd ed. Washington, DC: American Psychiatric Association.
- 463 American Psychiatric Association. (1994) Diagnostic and statistical manual of mental disorders:
464 DSM-IV. 4th ed. Washington (DC): American Psychiatric Association.
- 465 American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders:
466 DSM-IV-TR. 4th ed, Text Revision. Washington (DC): American Psychiatric
467 Association.
- 468 American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders:
469 DSM-5. 5th ed. Washington (DC): American Psychiatric Association.
- 470 *Amoozegar, F., Patten, S. B., Becker, W. J., Bulloch, A. G., Fiest, K. M., Davenport, W. J., ...
471 & Jette, N. (2017). The prevalence of depression and the accuracy of depression
472 screening tools in migraine patients. *General Hospital Psychiatry*, 48, 25-31.
- 473 Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects
474 Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.
475 doi:10.18637/jss.v067.i01

476 *Bayón-Pérez, C., Hernando, A., Álvarez-Comino, M., Cebolla, S., Serrano, L., Gutiérrez, F., ...
477 & Pulido, F. (2016). Toward a comprehensive care of HIV patients: finding a strategy to
478 detect depression in a Spanish HIV cohort. *AIDS Care*, 28, 834 - 841.

479 *Beck, K. R., Tan, S. M., Lum, S. S., Lim, L. E., & Krishna, L. K. (2016). Validation of the
480 emotion thermometers and hospital anxiety and depression scales in Singapore:
481 Screening cancer patients for distress, anxiety and depression. *Asia-Pacific Journal of*
482 *Clinical Oncology*, 12(2), e241–e249.

483 *Beraldi, A., Baklayan, A., Hoster, E., Hiddemann, W., & Heussner, P. (2014). Which
484 questionnaire is most suitable for the detection of depressive disorders in haemato-
485 oncological patients? Comparison between HADS, CES-D and PHQ-9. *Oncology*
486 *Research and Treatment*, 37,108.

487 *Bernstein, C. N., Zhang, L., Lix, L. M., Graff, L. A., Walker, J. R., Fisk, J. D., ... & CIHR
488 Team in Defining the Burden and Managing the Effects of Immune-mediated
489 Inflammatory Disease. (2018). The validity and reliability of screening measures for
490 depression and anxiety disorders in inflammatory bowel disease. *Inflammatory Bowel*
491 *Diseases*, 24(9), 1867-1875.

492 *Braeken, A. P. B. M., Lechner, L., Houben, R. M. A., Van Gils, F. C. J. M., & Kempen, G. I. J.
493 M. (2011). Psychometric properties of the Screening Inventory of Psychosocial Problems
494 (SIPP) in Dutch cancer patients treated with radiotherapy. *European Journal of Cancer*
495 *Care*, 20(3), 305-314.

496 Brugha, T. S., Bebbington, P. E., & Jenkins, R. (1999). A difference that matters: comparisons of
497 structured and semi-structured psychiatric diagnostic interviews in the general

498 population. *Psychological Medicine*, 29(5), 1013-1020.
499 doi:10.1017/S0033291799008880

500 Brugha, T. S., Jenkins, R., Taub, N., Meltzer, H., & Bebbington, P. E. (2001). A general
501 population comparison of the Composite International Diagnostic Interview (CIDI) and
502 the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychological
503 Medicine*, 31(6), 1001-1013. doi:10.1017/S0033291701004184

504 *Bunevicius, A., Peceliuniene, J., Mickuviene, N., Valius, L., & Bunevicius, R. (2007).
505 Screening for depression and anxiety disorders in primary care patients. *Depression and
506 Anxiety*, 24(7), 455–460.

507 *Bunevicius, A., Staniute, M., Brozaitiene, J., & Bunevicius, R. (2012). Diagnostic accuracy of
508 self-rating scales for screening of depression in coronary artery disease patients. *Journal
509 of Psychosomatic Research*, 72(1), 22–25.

510 *Butnorienė, J., Bunevicius, A., Norkus, A., & Bunevicius, R. (2014). Depression but not
511 anxiety is associated with metabolic syndrome in primary care based community sample.
512 *Psychoneuroendocrinology*, 40, 269–276.

513 *Can, C., Cimilli, C., Ozenli, Y., Ergor, G., Aysevener, E. O., Unek, T., & Astarcioglu, I. (2018).
514 Quality of life and psychiatric disorders before and one year after liver
515 transplantation. *Journal of Clinical and Analytical Medicine*, 9(5), 396-401.

516 *Chen, C. K., Tsai, Y. C., Hsu, H. J., Wu, I. W., Sun, C. Y., Chou, C. C., ... & Wang, L. J.
517 (2010). Depression and suicide risk in hemodialysis patients with chronic renal failure.
518 *Psychosomatics*, 51(6), 528–528.e6.

519 *Cheung, G., Patrick, C., Sullivan, G., Cooray, M., & Chang, C. L. (2012). Sensitivity and
520 specificity of the Geriatric Anxiety Inventory and the Hospital Anxiety and Depression

521 Scale in the detection of anxiety disorders in older people with chronic obstructive
522 pulmonary disease. *International Psychogeriatrics*, 24(1), 128–136.

523 Clarke, D. M., Smith, G. C., Herrman, H. E., & McKenzie, D. P. (1998). Monash Interview for
524 Liaison Psychiatry (MILP) - Development, reliability, and procedural validity.
525 *Psychosomatics*, 39(4), 318-328. doi:10.1016/S0033-3182(98)71320-9

526 *Consoli, S. M., Rolhion, S., Martin, C., Ruel, K., Cambazard, F., Pellet, J., & Misery, L. (2006).
527 Low levels of emotional awareness predict a better response to dermatological treatment
528 in patients with psoriasis. *Dermatology*, 212(2), 128–136.

529 *Costa-Requena, G., Ballester Arnal, R., & Gil, F. (2013). Perceived social support in Spanish
530 cancer outpatients with psychiatric disorder. *Stress and Health: Journal of the*
531 *International Society for the Investigation of Stress*, 29(5), 421–426.

532 *Cukor, D., Coplan, J., Brown, C., Friedman, S., Newville, H., Safier, M., ... & Kimmel, P. L.
533 (2008). Anxiety disorders in adults treated by hemodialysis: a single-center
534 study. *American Journal of Kidney Diseases*, 52(1), 128-136.

535 *da Rocha e Silva, C. E. D. R., Brasil, M. A. A., Do Nascimento, E. M., de Bragança Pereira, B.,
536 & André, C. (2013). Is poststroke depression a major depression?. *Cerebrovascular*
537 *Diseases*, 35(4), 385-391.

538 *De Souza, J., Jones, L. A., & Rickards, H. (2010). Validation of self-report depression rating
539 scales in Huntington's disease. *Movement Disorders*, 25(1), 91-96.

540 *de la Torre, A. Y., Oliva, N., Echevarrieta, P. L., Pérez, B. G., Caporusso, G. B., Titaro, A.
541 J., ... & Daray, F. M. (2016). Major depression in hospitalized Argentine general medical
542 patients: Prevalence and risk factors. *Journal of Affective Disorders*, 197, 36–42.

543 *de Oliveira, G. N., Lessa, J. M., Gonçalves, A. P., Portela, E. J., Sander, J. W., & Teixeira, A.
544 L. (2014). Screening for depression in people with epilepsy: comparative study among
545 neurological disorders depression inventory for epilepsy (NDDI-E), hospital anxiety and
546 depression scale depression subscale (HADS-D), and Beck depression inventory (BDI).
547 *Epilepsy & Behavior, 34*, 50–54.

548 *Douven, E., Schievink, S. H., Verhey, F. R., van Oostenbrugge, R. J., Aalten, P., Staals, J., &
549 Köhler, S. (2016). The Cognition and Affect after Stroke - a Prospective Evaluation of
550 Risks (CASPER) study: rationale and design. *BMC Neurology, 16*, 65.

551 *Dorow, M., Stein, J., Pabst, A., Weyerer, S., Werle, J., Maier, W., ... & Riedel-Heller, S. G.
552 (2018). Categorical and dimensional perspectives on depression in elderly primary care
553 patients—Results of the AgeMooDe study. *International Journal of Methods in*
554 *Psychiatric Research, 27*(1), e1577.

555 *Drabe, N., Zwahlen, D., Büchi, S., Moergeli, H., Zwahlen, R. A., & Jenewein, J. (2008).
556 Psychiatric morbidity and quality of life in wives of men with long-term head and neck
557 cancer. *Psycho-Oncology, 17*(2), 199–204.

558 Endicott, J., & Spitzer, R. L. (1987). [Schedule for Affective Disorders and Schizophrenia
559 (SADS)]. *Acta psychiatrica Belgica, 87*(4), 361-516.

560 *Fábregas, B. C., Moura, A. S., Ávila, R. E., Faria, M. N., Carmo, R. A., & Teixeira, A. L.
561 (2014). Sexual dysfunction and dissatisfaction in chronic hepatitis C patients. *Revista da*
562 *Sociedade Brasileira de Medicina Tropical, 47*(5), 564–572.

563 Fagerland, M. W., Lydersen, S., & Laake, P. (2014). Recommended tests and confidence
564 intervals for paired binomial proportions. *Statistics in Medicine, 33*(16), 2850-2875.
565 doi:10.1002/sim.6148

566 *Ferentinos, P., Paparrigopoulos, T., Rentzos, M., Zouvelou, V., Alexakis, T., & Evdokimidis, I.
567 (2011). Prevalence of major depression in ALS: comparison of a semi-structured
568 interview and four self-report measures. *Amyotrophic Lateral Sclerosis*, 12(4), 297-302..

569 *Fiest, K. M., Patten, S. B., Wiebe, S., Bulloch, A. G., Maxwell, C. J., Jette, N. (2014).
570 Validating screening tools for depression in epilepsy. *Epilepsia*, 55(10), 1642-1650.

571 First, M. B. (1995). *Structured Clinical Interview for the DSM (SCID)*. New York (NY): John
572 Wiley & Sons, Inc.

573 *Fischer, H. F., Klug, C., Roeper, K., Blozik, E., Edelmann, F., Eisele, M., ... & Herrmann-
574 Lingen, C. (2014). Screening for mental disorders in heart failure patients using
575 computer-adaptive tests. *Quality of Life Research*, 23(5), 1609–1618.

576 *Gagnon, N., Flint, A. J., Naglie, G., & Devins, G. M. (2005). Affective correlates of fear of
577 falling in elderly persons. *The American Journal of Geriatric Psychiatry*, 13(1), 7–14.

578 *Gandy, M., Sharpe, L., Perry, K. N., Miller, L., Thayer, Z., Boserio, J., & Mohamed, A. (2012).
579 Assessing the efficacy of 2 screening measures for depression in people with epilepsy.
580 *Neurology*, 79(4), 371–375.

581 *Golden, J., Conroy, R. M., & O'Dwyer, A. M. (2007). Reliability and validity of the Hospital
582 Anxiety and Depression Scale and the Beck Depression Inventory (Full and FastScreen
583 scales) in detecting depression in persons with hepatitis C. *Journal of Affective Disorders*,
584 100(1-3), 265–269.

585 *Gould, K. R., Ponsford, J. L., Johnston, L., & Schönberger, M. (2011). Predictive and
586 associated factors of psychiatric disorders after traumatic brain injury: a prospective
587 study. *Journal of Neurotrauma*, 28(7), 1155–1163.

588 *Grassi, L., Sabato, S., Rossi, E., Marmai, L., & Biancosino, B. (2009). Affective syndromes and
589 their screening in cancer patients with early and stable disease: Italian ICD-10 data and
590 performance of the Distress Thermometer from the Southern European Psycho-Oncology
591 Study (SEPOS). *Journal of Affective Disorders, 114*(1-3), 193–199.

592 *Hahn, D., Reuter, K., & Härter, M. (2006). Screening for affective and anxiety disorders in
593 medical patients - comparison of HADS, GHQ-12 and Brief-PHQ. *GMS Psycho-Social
594 Medicine, 3*.

595 Harel, D., Levis, B., Ishihara, M., Levis, A. W., Vigod, S. N., Howard, L. M., . . . DEPRESsion
596 Screening Data (DEPRESSD) EPDS Collaboration. (2021). Shortening the Edinburgh
597 postnatal depression scale using optimal test assembly methods: Development of the
598 EPDS-Dep-5. *Acta Psychiatrica Scandinavica, 143*(4), 348-362. doi:10.1111/acps.13272

599 *Härter, M., Woll, S., Wunsch, A., Bengel, J., & Reuter, K. (2005). Screening for mental
600 disorders in cancer, cardiovascular and musculoskeletal diseases. *Social Psychiatry and
601 Psychiatric Epidemiology, 41*, 56-62.

602 *Hartung, T. J., Friedrich, M., Johansen, C., Wittchen, H. U., Faller, H., Koch, U., Brähler,
603 E., ... & Mehnert, A. (2017). The Hospital Anxiety and Depression Scale (HADS) and
604 the 9-item Patient Health Questionnaire (PHQ-9) as screening instruments for depression
605 in patients with cancer. *Cancer, 123*(21), 4236–4243.

606 Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis.
607 *Statistics In Medicine, 21*(11), 1539-1558. doi:10.1002/sim.1186

608 *Hitchon, C. A., Zhang, L., Peschken, C. A., Lix, L. M., Graff, L. A., Fisk, J. D., Patten, S. B.,
609 Bolton, J., Sareen, J., El-Gabalawy, R., Marriott, J., Bernstein, C. N., & Marrie, R. A.

610 (2020). Validity and Reliability of Screening Measures for Depression and Anxiety
611 Disorders in Rheumatoid Arthritis. *Arthritis Care & Research*, 72(8), 1130–1139.

612 *Honarmand, K., & Feinstein, A. (2009). Validation of the Hospital Anxiety and Depression
613 Scale for use with multiple sclerosis patients. *Multiple Sclerosis*, 15(12), 1518–1524.

614 *Huey, N. S., Guan, N. C., Gill, J. S., Hui, K. O., Sulaiman, A. H., & Kunagasundram, S. (2018).
615 Core Symptoms of Major Depressive Disorder among Palliative Care Patients.
616 *International Journal of Environmental Research and Public Health*, 15(8), 1758.

617 Ishihara, M., Harel, D., Levis, B., Levis, A. W., Riehm, K. E., Saadat, N., . . . Thombs, B. D.
618 (2019). Shortening self-report mental health symptom measures through optimal test
619 assembly methods: Development and validation of the Patient Health Questionnaire-
620 Depression-4. *Depression and Anxiety*, 36(1), 82-92. doi:10.1002/da.22841

621 *Jackson, M. L., Tolson, J., Schembri, R., Bartlett, D., Rayner, G., Lee, V. V., & Barnes, M.
622 (2021). Does continuous positive airways pressure treatment improve clinical depression
623 in obstructive sleep apnea? A randomized wait-list controlled study. *Depression and*
624 *Anxiety*, 38(5), 498–507.

625 *Jang, J. E., Kim, S. W., Kim, S. Y., Kim, J. M., Park, M. H., Yoon, J. H., . . . & Yoon, J. S.
626 (2013). Religiosity, depression, and quality of life in Korean patients with breast cancer:
627 a 1-year prospective longitudinal study. *Psycho-Oncology*, 22(4), 922–929.

628 Joffres, M., Jaramillo, A., Dickinson, J., Lewin, G., Pottie, K., Shaw, E., Gorber, S. C., &
629 Tonelli, M. (2013). Recommendations on screening for depression in adults. *CMAJ*:
630 *Canadian Medical Association Journal*, 185(9), 775–782.
631 <https://doi.org/10.1503/cmaj.130403>

- 632 *Juliao, M., Barbosa, A., Oliveira, F., & Nunes, B. (2013). Prevalence and factors associated
633 with desire for death in patients with advanced disease: results from a Portuguese cross-
634 sectional study. *Psychosomatics*, 54(5), 451–457.
- 635 *Kang, H. J., Stewart, R., Kim, J. M., Jang, J. E., Kim, S. Y., Bae, K. Y., ... & Yoon, J. S.
636 (2013). Comparative validity of depression assessment scales for screening poststroke
637 depression. *Journal of Affective Disorders*, 147(1-3), 186–191.
- 638 *Keller, M., Sommerfeldt, S., Fischer, C., Knight, L., Riesbeck, M., Lowe, B., ... & Lehnert, T.
639 (2004). Recognition of distress and psychiatric morbidity in cancer patients: a multi-
640 method approach. *Annals of Oncology*, 15(8),1243.
- 641 *Kjaergaard, M., Arfwedson Wang, C. E., Waterloo, K., & Jorde, R. (2014). A study of the
642 psychometric properties of the Beck Depression Inventory-II, the Montgomery and
643 Åsberg Depression Rating Scale, and the Hospital Anxiety and Depression Scale in a
644 sample from a healthy population. *Scandinavian Journal of Psychology*, 55(1), 83–89.
- 645 *Kugaya, A., Akechi, T., Okuyama, T., Nakano, T., Mikami, I., Okamura, H., & Uchitomi, Y.
646 (2000). Prevalence, predictive factors, and screening for psychologic distress in patients
647 with newly diagnosed head and neck cancer. *Cancer*, 88(12), 2817–2823.
- 648 *Lambert, S. D., Clover, K., Pallant, J. F., Britton, B., King, M. T., Mitchell, A. J., Carter, G.
649 (2015). Making Sense of Variations in Prevalence Estimates of Depression in Cancer: A
650 Co-Calibration of Commonly Used Depression Scales Using Rasch Analysis. *Journal of*
651 *the National Comprehensive Cancer Network*, 13(10),1203.
- 652 *Law, M., Naughton, M. T., Dhar, A., Barton, D., & Dabscheck, E. (2014). Validation of two
653 depression screening instruments in a sleep disorders clinic. *Journal of Clinical Sleep*
654 *Medicine*, 10(6), 683–688.

655 Lecrubier, Y., Sheehan, D. V., Weiller, E., Amorim, P., Bonora, I., Sheehan, K. H., . . . Dunbar,
656 G. C. (1997). The Mini International Neuropsychiatric Interview (MINI). A short
657 diagnostic structured interview: Reliability and validity according to the CIDI. *European*
658 *Psychiatry*, *12*(5), 224-231. doi:10.1016/S0924-9338(97)83296-8

659 *Lee, Y., Wu, Y. S., Chien, C. Y., Fang, F. M., & Hung, C. F. (2016). Use of the Hospital
660 Anxiety and Depression Scale and the Taiwanese Depression Questionnaire for screening
661 depression in head and neck cancer patients in Taiwan. *Neuropsychiatric Disease and*
662 *Treatment*, *12*, 2649–2657.

663 *Lee, C. Y., Lee, Y., Wang, L. J., Chien, C. Y., Fang, F. M., & Lin, P. Y. (2017). Depression,
664 anxiety, quality of life, and predictors of depressive disorders in caregivers of patients
665 with head and neck cancer: A six-month follow-up study. *Journal of Psychosomatic*
666 *Research*, *100*, 29–34.

667 *Lees, R.A., Stott, D.J., Quinn, T.J., & Broomfield, N.M. (2014). Feasibility and Diagnostic
668 Accuracy of Early Mood Screening to Diagnose Persisting Clinical Depression/Anxiety
669 Disorder after Stroke. *Cerebrovascular Diseases*, *37*, 323 - 329.

670 Levis, B., Benedetti, A., Levis, A. W., Ioannidis, J. P. A., Shrier, I., Cuijpers, P., . . . Thombs, B.
671 D. (2017). Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A
672 Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient
673 Health Questionnaire-9 Depression Screening Tool. *American Journal of Epidemiology*,
674 *185*(10), 954-964. doi:10.1093/aje/kww191

675 Levis, B., Benedetti, A., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., . . . Thombs, B. D.
676 (2018). Probability of major depression diagnostic classification using semi-structured

677 versus fully structured diagnostic interviews. *British Journal of Psychiatry*, 212(6), 377-
678 385. doi:10.1192/bjp.2018.54

679 Levis, B., Benedetti, A., Thombs, B. D., Akena, D. H., Arroll, B., Ayalon, L., . . . DEPRESSion
680 Screening Data (DEPRESSD) Collaboration. (2019). Accuracy of Patient Health
681 Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant
682 data meta-analysis. *BMJ*, 365. doi:10.1136/bmj.11476

683 Levis, B., McMillan, D., Sun, Y., He, C., Rice, D. B., Krishnan, A., . . . Thombs, B. D. (2019).
684 Comparison of major depression diagnostic classification probability using the SCID,
685 CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: An
686 individual participant data meta-analysis. *International Journal of Methods in Psychiatric
687 Research*, 28(4). doi:10.1002/mpr.1803

688 Levis, B., Negeri, Z., Sun, Y., Benedetti, A., Thombs, B. D., & DEPRESSion Screening Data
689 (DEPRESSD) EPDS Group. (2020). Accuracy of the Edinburgh Postnatal Depression
690 Scale (EPDS) for screening to detect major depression among pregnant and postpartum
691 women: systematic review and meta-analysis of individual participant data. *BMJ*, 371.
692 doi:10.1136/bmj.m4022

693 *Loosman, W. L., Siegert, C. E., Korzec, A., & Honig, A. (2010). Validity of the Hospital
694 Anxiety and Depression Scale and the Beck Depression Inventory for use in end-stage
695 renal disease patients. *The British Journal of Clinical Psychology*, 49(Pt 4), 507–516.

696 *Love, A. W., Kissane, D. W., Bloch, S., & Clarke, D. (2002). Diagnostic efficiency of the
697 Hospital Anxiety and Depression Scale in women with early stage breast cancer. *The
698 Australian and New Zealand Journal of Psychiatry*, 36(2), 246–250.

- 699 *Love, A. W., Grabsch, B., Clarke, D. M., Bloch, S., & Kissane, D. W. (2004). Screening for
700 depression in women with metastatic breast cancer: a comparison of the Beck Depression
701 Inventory Short Form and the Hospital Anxiety and Depression Scale. *The Australian
702 and New Zealand journal of psychiatry*, 38(7), 526–531.
- 703 *Lowe, B., Gräfe, K., Zipfel, S., Spitzer, R. L., Herrmann-Lingen, C., Witte, S., & Herzog, W.
704 (2003). Detecting panic disorder in medical and psychosomatic outpatients: comparative
705 validation of the Hospital Anxiety and Depression Scale, the Patient Health
706 Questionnaire, a screening question, and physicians' diagnosis. *Journal of Psychosomatic
707 Research*, 55(6), 515–519.
- 708 *Marrie, R. A., Zhang, L., Lix, L. M., Graff, L. A., Walker, J. R., Fisk, J. D., ... & Bernstein, C.
709 N. (2018). The validity and reliability of screening measures for depression and anxiety
710 disorders in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 20, 9–15.
- 711 *Massardo, L., Bravo-Zehnder, M., Calderón, J., Flores, P., Padilla, O., Aguirre, J. M., ... &
712 González, A. (2015). Anti-N-methyl-D-aspartate receptor and anti-ribosomal-P
713 autoantibodies contribute to cognitive dysfunction in systemic lupus erythematosus.
714 *Lupus*, 24(6), 558–568.
- 715 *Matsuoka, Y., Nishi, D., Nakajima, S., Yonemoto, N., Hashimoto, K., Noguchi, H., ... & Kim,
716 Y. (2009). The Tachikawa cohort of motor vehicle accident study investigating
717 psychological distress: Design, methods and cohort profiles. *Social Psychiatry and
718 Psychiatric Epidemiology*, 44(4), 333–340.
- 719 *McFarlane, A. C., Browne, D., Bryant, R. A., O'Donnell, M., Silove, D., Creamer, M., &
720 Horsley, K. (2009). A longitudinal analysis of alcohol consumption and the risk of
721 posttraumatic symptoms. *Journal of Affective Disorders*, 118(1-3), 166–172.

722 McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016).
723 PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of*
724 *Clinical Epidemiology*, *75*, 40-46.

725 *Meyer, A., Wollbrück, D., Täschner, R., Singer, S., Ehrensperger, C., Danker, H., ... &
726 Schwarz, R. (2008). Psychological status and morbidity of the spouses of laryngectomy
727 patients. *Zeitschrift für Klinische Psychologie und Psychotherapie: Forschung und*
728 *Praxis*, *37*, 172.

729 *Michopoulos, I., Douzenis, A., Gournellis, R., Christodoulou, C., Kalkavoura, C.,
730 Michalopoulou, P.G., ... & Lykouras, L. (2010). Major depression in elderly medical
731 inpatients in Greece, prevalence and identification. *Aging Clinical and Experimental*
732 *Research*, *22*, 148-151.

733 Mitchell, A. J., Meader, N., & Symonds, P. (2010). Diagnostic validity of the Hospital Anxiety
734 and Depression Scale (HADS) in cancer and palliative settings: A meta-analysis. *Journal*
735 *of Affective Disorders*, *126*(3), 335-348. doi:10.1016/j.jad.2010.01.067

736 National Collaborating Centre for Mental Health (UK). (2010). *Depression in Adults with a*
737 *Chronic Physical Health Problem: Treatment and Management*. British Psychological
738 Society (UK). <http://www.ncbi.nlm.nih.gov/books/NBK82916/>

739 NCSS. (2017). One ROC curve and cutoff analysis, Chapter 546. Retrieved September 20, 2021,
740 from <https://www.ncss.com/software/ncss/roc-curves-ncss/>

741 Negeri, Z. F., Levis, B., Sun, Y., He, C., Krishnan, A., Wu, Y., Thombs, B. D. . . . DEPRESSion
742 Screening Data (DEPRESSD) PHQ Group. (2021). Accuracy of the Patient Health
743 Questionnaire-9 (PHQ-9) for screening to detect major depression: an updated systematic

744 review and individual participant data meta-analysis. *BMJ*, 375:n2183. doi:
745 <https://doi.org/10.1136/bmj.n2183>

746 Neupane, D., Levis, B., Bhandari, P. M., Thombs, B. D., Benedetti, A., & DEPRESSion
747 Screening Data (DEPRESSD) Collaboration. (2021). Selective cutoff reporting in studies
748 of the accuracy of the PHQ-9 and EPDS depressions screening tools: comparison of
749 results based on published cutoffs versus all cutoffs using individual participant data
750 meta-analysis. *International Journal of Epidemiology*, 30(3), e1873. doi:
751 10.1002/mpr.1873.

752 Nosen, E., & Woody, S. R. (2008). Chapter 8: Diagnostic assessment in research. In *Handbook*
753 *of Research Methods in Abnormal and Clinical Psychology* (ed. D. McKay), pp. 109-124.
754 Sage: Thousand Oaks.

755 *O'Rourke, S., MacHale, S., Signorini, D., Dennis, M. (1998). Detecting psychiatric morbidity
756 after stroke: comparison of the GHQ and the HAD Scale. *Stroke*, 29, 980-985.

757 *Öztürk, A., Deveci, E., Bağcıoğlu, E., Atalay, F. & Serdar, Z. (2013). Anxiety, depression,
758 social phobia, and quality of life in Turkish patients with acne and their relationships with
759 the severity of acne. *Turkish Journal of Medical Sciences*, 43 (4), 660-666.

760 *Patel, D., Sharpe, L., Thewes, B., Rickard, J., Schnieden, V., & Lewis, C. (2010). Feasibility of
761 using risk factors to screen for psychological disorder during routine breast care nurse
762 consultations. *Cancer Nursing*, 33(1), 19–27.

763 *Patel, D., Sharpe, L.A., Thewes, B., Bell, M.L., & Clarke, S.J. (2011). Using the Distress
764 Thermometer and Hospital Anxiety and Depression Scale to screen for psychosocial
765 morbidity in patients diagnosed with colorectal cancer. *Journal of Affective Disorders*,
766 131(1-3), 412-6.

767 *Patten, S. B., Burton, J. M., Fiest, K. M., Wiebe, S., Bulloch, A. G., Koch, M., ... & Jetté, N.
768 (2015). Validity of four screening scales for major depression in MS. *Multiple Sclerosis*,
769 *21*(8), 1064–1071.

770 *Pedroso, V. S., Vieira, É. L., Brunoni, A. R., Lauterbach, E. C., Teixeira, A. L. (2016).
771 Psychopathological evaluation and use of the Hospital Anxiety and Depression Scale in a
772 sample of Brazilian patients with post-stroke depression. *Archives of Clinical Psychiatry*
773 *(São Paulo)*, *43*, 147-50.

774 *Phan, T., Carter, O., Adams, C., Waterer, G., Chung, L. P., Hawkins, M., ... & Strobel, N.
775 (2016). Discriminant validity of the Hospital Anxiety and Depression Scale, Beck
776 Depression Inventory (II) and Beck Anxiety Inventory to confirmed clinical diagnosis of
777 depression and anxiety in patients with chronic obstructive pulmonary disease. *Chronic*
778 *Respiratory Disease*, *13*(3), 220–228.

779 *Pintor, L., Fuente, E.D., Peri, J.M., Pérez-Villa, F., & Roig, E. (2006). Evaluación psiquiátrica
780 transversal en pacientes candidatos a un trasplante cardíaco. *Psiquiatría Biológica*, *13*,
781 122-126.

782 *Prisnie, J. C., Fiest, K. M., Coutts, S. B., Patten, S. B., Atta, C. A., Blaikie, L., ... & Jetté, N.
783 (2016). Validating screening tools for depression in stroke and transient ischemic attack
784 patients. *International Journal of Psychiatry in Medicine*, *51*(3), 262–277.

785 R Core Team. (2020). R: A language and environment for statistical computing. R Foundation
786 for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

787 RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA
788 <http://www.rstudio.com/>.

789 *Reme, S. E., Lie, S. A., & Eriksen, H. R. (2014). Are 2 questions enough to screen for
790 depression and anxiety in patients with chronic low back pain?. *Spine*, *39*(7), E455–E462.

791 Rice, D. B., & Thombs, B. D. (2016). Risk of Bias from Inclusion of Currently Diagnosed or
792 Treated Patients in Studies of Depression Screening Tool Accuracy: A Cross-Sectional
793 Analysis of Recently Published Primary Studies and Meta-Analyses. *Plos One*, *11*(2).
794 doi:10.1371/journal.pone.0150067

795 Riley, R. D., Dodd, S. R., Craig, J. V., Thompson, J. R., & Williamson, P. R. (2008). Meta-
796 analysis of diagnostic test studies using individual patient data and aggregate data.
797 *Statistics in Medicine*, *27*(29), 6111-6136. doi:10.1002/sim.3441

798 Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant
799 data: rationale, conduct, and reporting. *BMJ*, *340*. doi:10.1136/bmj.c221

800 Robins, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., . . . Towle, L. H.
801 (1988). The Composite International Diagnostic Interview: an epidemiologic instrument
802 suitable for use in conjunction with different diagnostic systems and in different cultures.
803 *Archives Of General Psychiatry*, *45*(12), 1069-1077.

804 *Rooney, A. G., McNamara, S., Mackinnon, M., Fraser, M., Rampling, R., Carson, A., & Grant,
805 R. (2013). Screening for major depressive disorder in adults with cerebral glioma: an
806 initial validation of 3 self-report instruments. *Neuro-Oncology*, *15*(1), 122-129.

807 *Ryan, D. A., Gallagher, P., Wright, S., Cassidy, E. M. (2012). Sensitivity and specificity of the
808 Distress Thermometer and a two-item depression screen (Patient Health Questionnaire-2)
809 with a 'help' question for psychological distress and psychiatric morbidity in patients with
810 advanced cancer. *Psycho-Oncology*, *21*(12), 1275.

811 *Sanchez-Gistau, V., Sugranyes, G., Baillés, E., Carreño, M., Donaire, A., Bargalló, N., &
812 Pintor, L. (2012). Is major depressive disorder specifically associated with mesial
813 temporal sclerosis?. *Epilepsia*, 53(2), 386–392.

814 *Sánchez, R., Peri, J. M., Baillés, E., Bastidas, A., Pérez-Villa, F., Bulbena, A., & Pintor, L.
815 (2012). Evaluación de psicopatología, afrontamiento y apoyo familiar en el cumplimiento
816 de pautas médicas en los 12 meses posteriores a un trasplante cardiaco. *Psiquiatría*
817 *Biológica*, 19, 1-5.

818 *Sánchez, R., Baillés, E., Peri, J.M., Bastidas, A., Pérez-Villa, F., Bulbena, A., & Pintor, L.
819 (2014). Cross-sectional psychosocial evaluation of heart transplantation candidates.
820 *General Hospital Psychiatry*, 36(6), 680-5.

821 *Saracino, R.M., Weinberger, M.I., Roth, A.J., Hurria, A., & Nelson, C.J. (2017). Assessing
822 depression in a geriatric cancer population. *Psycho-Oncology*, 26, 1484 - 1490.

823 Schatzberg, A. F. (2019). Scientific Issues Relevant to Improving the Diagnosis, Risk
824 Assessment, and Treatment of Major Depression. *American Journal of Psychiatry*,
825 176(5), 342-347. doi:10.1176/appi.ajp.2019.19030273

826 *Schellekens, M.P., van den Hurk, D.G., Prins, J.B., Molema, J., van der Drift, M.A., &
827 Speckens, A.E. (2016). The suitability of the Hospital Anxiety and Depression Scale,
828 Distress Thermometer and other instruments to screen for psychiatric disorders in both
829 lung cancer patients and their partners. *Journal of Affective Disorders*, 203, 176-183 .

830 *Schwarzbold, M.L., Diaz, A.P., Nunes, J.C., Sousa, D.S., Hohl, A., Guarnieri, R., ... & Walz,
831 R. (2014). Validity and screening properties of three depression rating scales in a
832 prospective sample of patients with severe traumatic brain injury. *Revista brasileira de*
833 *psiquiatria*, 36, 206-12 .

834 *Senturk, V., Stewart, R.J., & Sağduyu, A. (2007). Screening for mental disorders in leprosy
835 patients: comparing the internal consistency and screening properties of HADS and
836 GHQ-12. *Leprosy Review*, 78(3), 231-42 .

837 Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Janavs, J., Weiller, E., Keskiner, A., . . . Dunbar,
838 G. C. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI)
839 according to the SCID-P and its reliability. *European Psychiatry*, 12(5), 232-241.
840 doi:10.1016/S0924-9338(97)83297-X

841 *Sia, A.D., Williams, L.J., Pasco, J.A., Jacka, F.N., Brennan-Olsen, S.L., & Veerman, J.L.
842 (2018). The Population Mean Mood Predicts The Prevalence of Depression in an
843 Australian Context. *Australian & New Zealand Journal of Psychiatry*, 52, 461 - 472.

844 *Simard, S., & Savard, J. (2015). Screening and comorbidity of clinical levels of fear of cancer
845 recurrence. *Journal of Cancer Survivorship*, 9, 481-491.

846 *Singer, S., Danker, H., Dietz, A., Hornemann, B., Koscielny, S., Oeken, J., ... & Krauss, O.
847 (2008). Screening for mental disorders in laryngeal cancer patients: a comparison of 6
848 methods. *Psycho-Oncology*, 17.

849 *Singer, S., Kuhnt, S., Götze, H., Hauss, J., Hinz, A., Liebmann, A., Krauss, O., Lehmann, A., &
850 Schwarz, R. (2009). Hospital anxiety and depression scale cutoff scores for cancer
851 patients in acute care. *British Journal of Cancer*, 100(6), 908–912.

852 Siu, A. L., & US Preventive Services Task Force (USPSTF). (2016). Screening for Depression in
853 Adults: US Preventive Services Task Force Recommendation Statement. *JAMA*, 315(4),
854 380–387. <https://doi.org/10.1001/jama.2015.18392>

855 *Soyseth, T. S., Lund, M. B., Bjortuft, O., Heldal, A., Søyseth, V., Dew, M. A., ... & Malt, U. F.
856 (2016). Psychiatric disorders and psychological distress in patients undergoing evaluation
857 for lung transplantation: a national cohort study. *General Hospital Psychiatry*, 42, 67–73.

858 *Stafford, L., Berk, M., & Jackson, H. J. (2007). Validity of the Hospital Anxiety and
859 Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients
860 with coronary artery disease. *General Hospital Psychiatry*, 29(5), 417–424.

861 *Stafford, L., Judd, F., Gibson, P., Komiti, A., Quinn, M., & Mann, G. B. (2014). Comparison of
862 the hospital anxiety and depression scale and the center for epidemiological studies
863 depression scale for detecting depression in women with breast or gynecologic cancer.
864 *General Hospital Psychiatry*, 36(1), 74–80.

865 *Stone, J., Townend, E., Kwan, J., Haga, K., Dennis, M. S., & Sharpe, M. (2004). Personality
866 change after stroke: Some preliminary observations. *Journal of Neurology, Neurosurgery*
867 *& Psychiatry*, 75(12), 1708–1713.

868 *Sultan, S., Luminet, O., & Hartemann, A. (2010). Cognitive and anxiety symptoms in screening
869 for clinical depression in diabetes: a systematic examination of diagnostic performances
870 of the HADS and BDI-SF. *Journal of Affective Disorders*, 123(1-3), 332–336.

871 Thombs, B. D., Arthurs, E., El-Baalbaki, G., Meijer, A., Ziegelstein, R. C., & Steele, R. J.
872 (2011). Risk of bias from inclusion of patients who already have diagnosis of or are
873 undergoing treatment for depression in diagnostic accuracy studies of screening tools for
874 depression: systematic review. *BMJ*, 343. doi:10.1136/bmj.d4825

875 Thombs, B. D., Benedetti, A., Kloda, L. A., Levis, B., Azar, M., Riehm, K. E., . . . Tonelli, M.
876 (2016). Diagnostic accuracy of the Depression subscale of the Hospital Anxiety and
877 Depression Scale (HADS-D) for detecting major depression: protocol for a systematic

878 review and individual patient data meta-analyses. *BMJ Open*, 6(4). doi:10.1136/bmjopen-
879 2016-011913

880 Thombs, B. D., & Rice, D. B. (2016). Sample sizes and precision of estimates of sensitivity and
881 specificity from primary studies on the diagnostic accuracy of depression screening tools:
882 a survey of recently published studies. *International Journal of Methods in Psychiatric
883 Research*, 25(2), 145-152. doi:10.1002/mpr.1504

884 *Tiringer, I., Simon, A., Herrfurth, D., Suri, I., Szalai, K., & Veress, A. (2008). Occurrence of
885 anxiety and depression disorders after acute cardiac events during hospital rehabilitation.
886 Application of the Hospital Anxiety and Depression Scale as a screening instrument.
887 *Psychiatria Hungarica*, 23(6), 430–443.

888 *Tung, K. Y., Cheng, K. S., Lee, W. K., Kwong, P. K., Chan, K. W., Law, A. C., & Lo, W. T.
889 (2015). Psychiatric Morbidity in Chinese Adults with Type 1 Diabetes in Hong Kong.
890 *East Asian Archives of Psychiatry*, 25(3), 128–136.

891 *Turner, A., Hambridge, J., White, J., Carter, G., Clover, K., Nelson, L., & Hackett, M. (2012).
892 Depression screening in stroke: a comparison of alternative measures with the structured
893 diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth
894 edition (major depressive episode) as criterion standard. *Stroke*, 43(4), 1000–1005.

895 United Nations Development Programme. (2020). Human Development Reports.
896 <http://hdr.undp.org/en/content/human-development-index-hdi>. Accessed September 20,
897 2021.

898 van der Leeden R, Busing FMTA, Meijer E. (1997). *Bootstrap methods for two-level models*.
899 *Technical report PRM 97-04*. Leiden University, Department of Psychology: Leiden, The
900 Netherlands.

901 van der Leeden R, Meijer E, Busing FMTA (2008). Chapter 11: Resampling multilevel models.
902 In *Handbook of Multilevel Analysis* (ed. J. Leeuw, E. Meijer), pp. 401-433. Springer:
903 New York.

904 Vodermaier, A., & Millman, R. D. (2011). Accuracy of the Hospital Anxiety and Depression
905 Scale as a screening tool in cancer patients: a systematic review and meta-analysis.
906 *Supportive Care in Cancer, 19*(12), 1899-1908. doi:10.1007/s00520-011-1251-4

907 *Walker, J., Postma, K., McHugh, G. S., Rush, R., Coyle, B., Strong, V., & Sharpe, M. (2007).
908 Performance of the Hospital Anxiety and Depression Scale as a screening tool for major
909 depressive disorder in cancer patients. *Journal of Psychosomatic Research, 63*(1), 83–91.

910 *Walterfang, M. A., O'Donovan, J., Fahey, M. C., & Velakoulis, D. (2007). The neuropsychiatry
911 of adrenomyeloneuropathy. *CNS Spectrums, 12*(9), 696–701.

912 *Wong, L. Y., Yiu, R. L., Chiu, C. K., Lee, W. K., Lee, Y. L., Kwong, P. K., & Lo, W. T.
913 (2015). Prevalence of Psychiatric Morbidity in Chinese Subjects with Knee Osteoarthritis
914 in a Hong Kong Orthopaedic Clinic. *East Asian Archives of Psychiatry, 25*(4), 150–158.

915 Walker, E., & Nowacki, A. S. (2011). Understanding Equivalence and Noninferiority Testing.
916 *Journal of General Internal Medicine, 26*(2), 192-196. doi:10.1007/s11606-010-1513-8

917 Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., . . .
918 QUADAS-2 Group. (2011). QUADAS-2: A Revised Tool for the Quality Assessment of
919 Diagnostic Accuracy Studies. *Annals of Internal Medicine, 155*(8), 529-U104.
920 doi:10.7326/0003-4819-155-8-201110180-00009

921 World Health Organization (WHO). (1992). *The ICD-10 classification of mental and*
922 *behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health
923 Organization.

924 World Health Organization. (1994). *Schedules for clinical assessment in neuropsychiatry:*
925 *version 2*. World Health Organization.

926 Wu, Y., Levis, B., Ioannidis, J. P. A., Benedetti, A., Thombs, B. D., & DEPRESSion Screening
927 Data (DEPRESSD) Collaboration. (2020). Probability of Major Depression Classification
928 Based on the SCID, CIDI, and MINI Diagnostic Interviews: A Synthesis of Three
929 Individual Participant Data Meta-Analyses. *Psychotherapy and Psychosomatics*, *90*(1),
930 28-40. doi:10.1159/000509283

931 Wu, Y., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., . . . Thombs, B. D. (2020).
932 Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review
933 and individual participant data meta-analysis. *Psychological Medicine*, *50*(8), 1368-1380.
934 doi:10.1017/S0033291719001314

935 Wu, Y., Levis, B., Sun, Y., He, C., Krishnan, A., Neupane, D., . . . DEPRESSion Screening Data
936 (DEPRESSD) HADS Group. (2021). Accuracy of the Hospital Anxiety and Depression
937 Scale Depression subscale (HADS-D) to screen for major depression: systematic review
938 and individual participant data meta-analysis. *BMJ*, *373*. doi:10.1136/bmj.n972

939 Wu, Y., Levis, B., Sun, Y., Krishnan, A., He, C., Riehm, K. E., . . . Thombs, B. D. (2020).
940 Probability of major depression diagnostic classification based on the SCID, CIDI and
941 MINI diagnostic interviews controlling for Hospital Anxiety and Depression Scale -
942 Depression subscale scores: An individual participant data meta-analysis of 73 primary
943 studies. *Journal of Psychosomatic Research*, *129*. doi:10.1016/j.jpsychores.2019.109892

944 *Yamashita, A., Noguchi, H., Hamazaki, K., Sato, Y., Narisawa, T., Kawashima, Y., . . . &
945 Matsuoka, Y. J. (2017). Serum polyunsaturated fatty acids and risk of psychiatric

946 disorder after acute coronary syndrome: A prospective cohort study. *Journal of Affective*
947 *Disorders*, 218, 306–312.

948 Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta*
949 *Psychiatrica Scandinavica*, 67(6), 361–370. [doi: 10.1111/j.1600-0447.1983.tb09716.x](https://doi.org/10.1111/j.1600-0447.1983.tb09716.x)

950

951 *Studies that included in the IPDMA

952

953

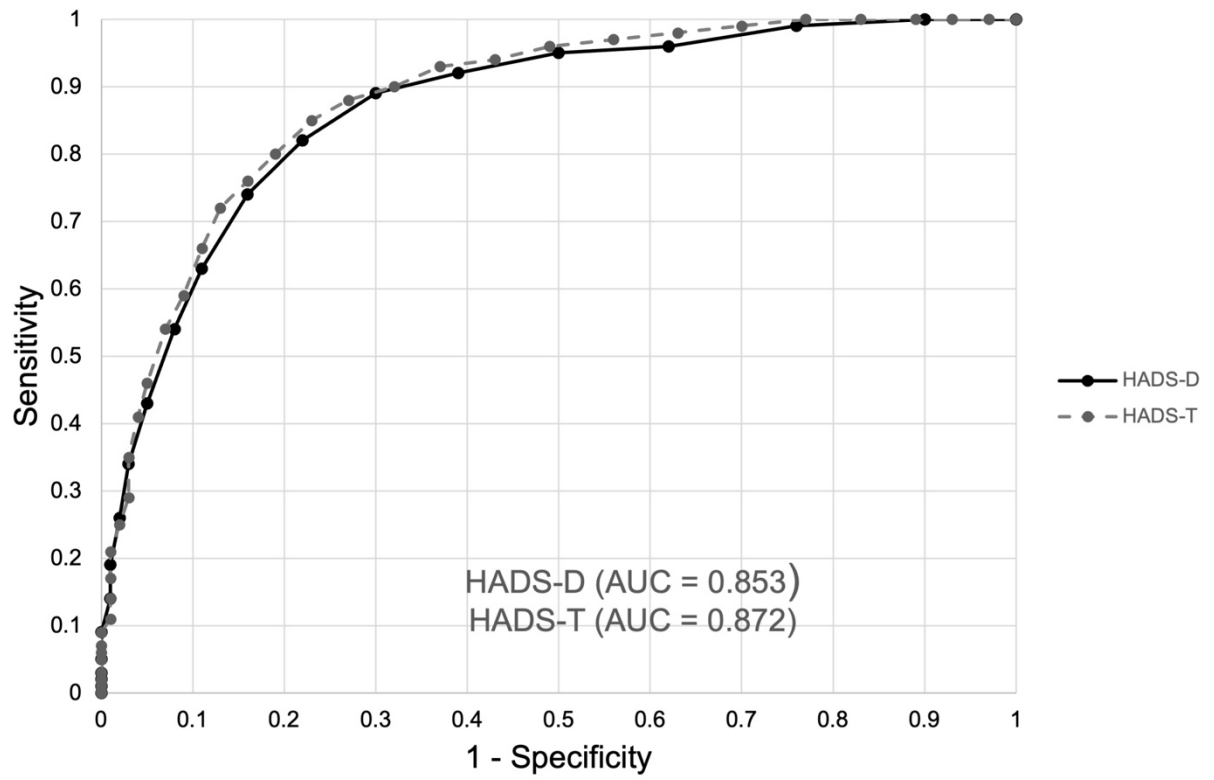


Fig 1. ROC curve for HADS-D and HADS-T across all studies.

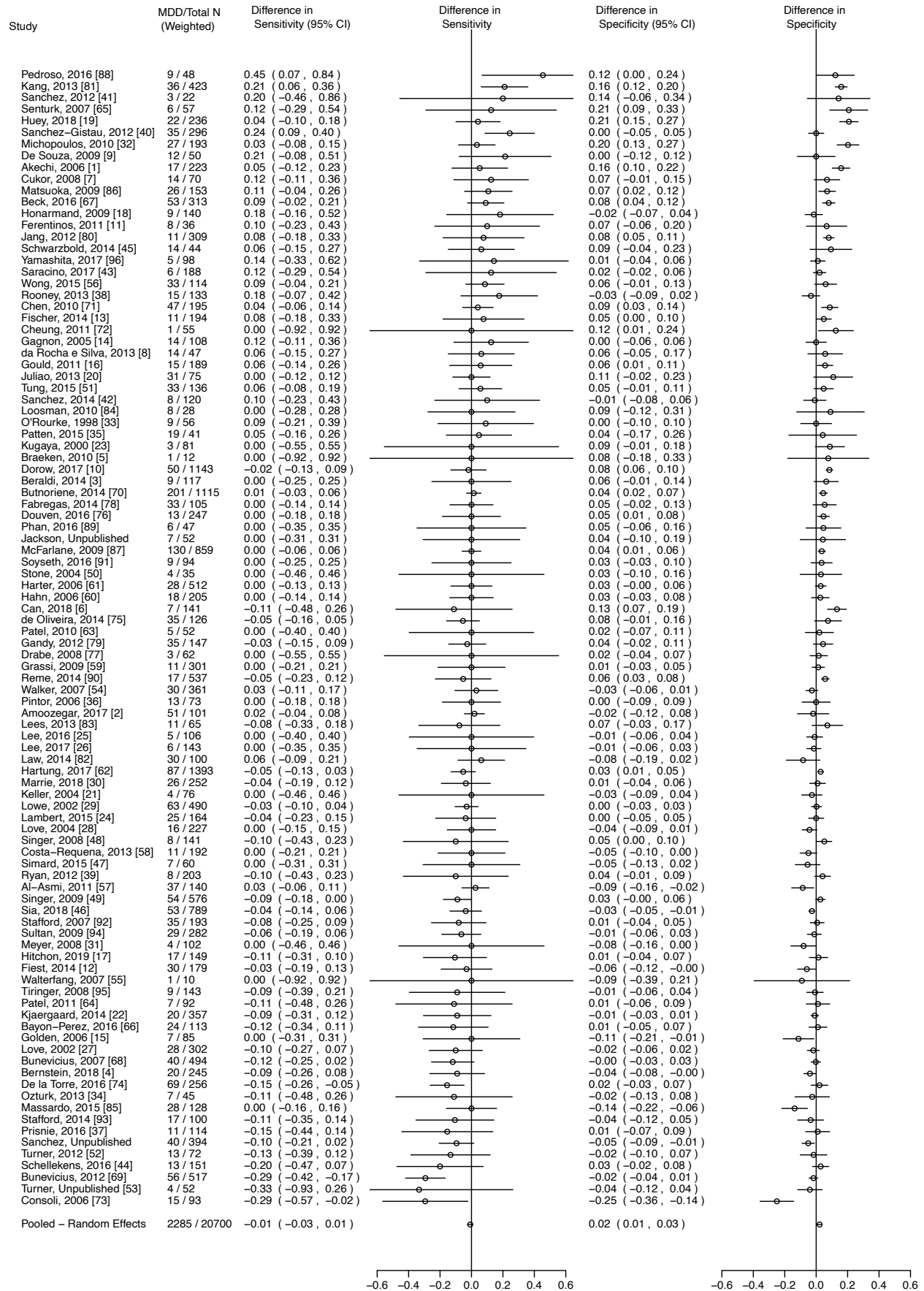


Fig 2. Forest plots of the difference in sensitivity and specificity estimates at the optimal cutoff (HADS-D: ≥ 7 ; HADS-T: ≥ 15) between HADS-D and HADS-T across all studies^a (N Studies = 98^b; N Participants = 20,700; N major depression = 2,285)^c

^a τ^2 for the difference of sensitivity and specificity were both < 0.001 .

^b References for all included studies are marked with an asterisk in the reference list. The reference numbers refer to Supplementary Material References.

^c The studies were sorted by the sum of difference in sensitivity and difference in specificity in descending order.

Table 1. Participant data by subgroups^a

Participant Subgroup	N Studies	N Participants	N (%) Major Depression
All participants	98	20,700	2,285 (11)
Participants not currently diagnosed with a mental disorder or receiving treatment for a mental health problem	38	6,995	495 (7)
Age <60	92	11,795	1,452 (12)
Age ≥60	92	8,741	779 (9)
Women	96	11,111	1,342 (12)
Men	89	9,494	911 (10)
Very high country human development index	90	20,088	2,130 (11)
High country human development index	8	612	155 (25)
Participants diagnosed with cancer^b	27	5,767	433 (8)
Inpatient specialty care	38	8,827	1,047 (12)
Outpatient specialty care	54	9,547	1,072 (11)
Non-medical	7	1,908	116 (6)
Inpatient/outpatient mixed	3	418	50 (12)

^a Some variables were coded at the study level, while others were coded at the participant level. Thus, number of studies does not always add up to the total number.

^b The statistics here were from individual-level variable of cancer diagnosis, slight different from what we used in the subgroup analysis which based on the study-level care setting variable.

Table 2. Comparison of sensitivity and specificity estimates between HADS-D and HADS-T for pairs of optimal cutoffs and cutoffs close to the optimal cutoffs across all studies

HADS-D ^a					HADS-T					HADS-T – HADS-D			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.90	(0.87, 0.92)	0.61	(0.58, 0.64)	11	0.91	(0.89, 0.93)	0.63	(0.60, 0.66)	0.01	(-0.01, 0.04)	0.02	(-0.00, 0.04)
6	0.86	(0.82, 0.88)	0.70	(0.67, 0.73)	13	0.86	(0.83, 0.88)	0.73	(0.70, 0.75)	0.00	(-0.03, 0.03)	0.03	(0.01, 0.05)
7 ^b	0.79	(0.75, 0.83)	0.78	(0.75, 0.80)	15 ^c	0.79	(0.76, 0.82)	0.81	(0.78, 0.83)	0.00	(-0.05, 0.02)	0.03	(0.01, 0.04)
8	0.70	(0.66, 0.74)	0.84	(0.82, 0.86)	17	0.70	(0.66, 0.74)	0.87	(0.85, 0.89)	0.00	(-0.05, 0.04)	0.03	(0.01, 0.04)
9	0.60	(0.55, 0.64)	0.89	(0.87, 0.91)	19	0.58	(0.54, 0.61)	0.91	(0.9, 0.93)	-0.02	(-0.07, 0.02)	0.02	(0.01, 0.03)
10	0.50	(0.45, 0.54)	0.92	(0.91, 0.94)	21	0.45	(0.41, 0.49)	0.95	(0.94, 0.95)	-0.05	(-0.10, -0.01)	0.03	(0.01, 0.03)
11	0.39	(0.35, 0.43)	0.95	(0.94, 0.96)	23	0.34	(0.31, 0.37)	0.97	(0.96, 0.97)	-0.05	(-0.10, -0.01)	0.02	(0.01, 0.03)

^a N Studies = 98; N Participants = 20,700; N major depression = 2,285

^b The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-D.

^c The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-T.

CI: confidence interval

Table 3. Comparison of sensitivity and specificity estimates between HADS-D and HADS-T for pairs of optimal cutoffs and cutoffs close to the optimal cutoffs across all studies via individual-level model

HADS-D^a	HADS-T	HADS-T – HADS-D	
Cutoff	Cutoff	Sensitivity	Specificity
5	11	0.02 (-0.00, 0.03)	0.01 (-0.00, 0.03)
6	13	0.01 (-0.01, 0.03)	0.03 (0.01, 0.04)
7 ^b	15 ^c	0.00 (-0.02, 0.03)	0.02 (0.01, 0.04)
8	17	0.00 (-0.03, 0.03)	0.03 (0.02, 0.04)
9	19	-0.02 (-0.05, 0.01)	0.03 (0.02, 0.04)
10	21	-0.05 (-0.08, -0.02)	0.03 (0.02, 0.03)
11	23	-0.05 (-0.09, -0.02)	0.02 (0.02, 0.03)

^a N Participants = 20,700; N major depression = 2,285

^b The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-D.

^c The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-T.

Table 4a. Comparison of sensitivity and specificity estimates between HADS-D and HADS-T for pairs of optimal cutoffs and cutoffs close to the optimal cutoffs among participants recruited from inpatient care settings

HADS-D ^a					HADS-T					HADS-T – HADS-D			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.90	(0.87, 0.93)	0.55	(0.49, 0.60)	11	0.90	(0.87, 0.92)	0.62	(0.56, 0.68)	0.00	(-0.03, 0.03)	0.07	(0.04, 0.11)
6	0.86	(0.83, 0.89)	0.64	(0.58, 0.69)	13	0.85	(0.81, 0.88)	0.72	(0.67, 0.77)	-0.01	(-0.07, 0.02)	0.08	(0.06, 0.12)
7 ^b	0.80	(0.75, 0.83)	0.73	(0.68, 0.78)	15 ^{cd}	0.79	(0.74, 0.82)	0.81	(0.76, 0.85)	-0.01	(-0.08, 0.02)	0.08	(0.05, 0.11)
8	0.73	(0.68, 0.78)	0.80	(0.76, 0.84)	17	0.69	(0.64, 0.74)	0.87	(0.83, 0.9)	-0.04	(-0.11, 0.03)	0.07	(0.04, 0.09)
9	0.63	(0.58, 0.69)	0.86	(0.82, 0.89)	19	0.59	(0.54, 0.64)	0.91	(0.88, 0.93)	-0.04	(-0.14, 0.01)	0.05	(0.03, 0.07)
10	0.55	(0.49, 0.61)	0.90	(0.87, 0.93)	21	0.46	(0.41, 0.51)	0.95	(0.92, 0.96)	-0.09	(-0.19, -0.03)	0.05	(0.03, 0.06)
11	0.45	(0.39, 0.51)	0.93	(0.91, 0.95)	23	0.36	(0.32, 0.41)	0.97	(0.95, 0.98)	-0.09	(-0.18, -0.02)	0.04	(0.02, 0.05)

^a N Studies = 38; N Participants = 8,827; N major depression = 1,047

^b The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-D.

^c The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-T.

^d On this cutoff of HADS-T, the model convergence code was 0 when using the default optimizer in glmer, but there were meaningful CIs.

CI: confidence interval

Table 4b. Comparison of sensitivity and specificity estimates between HADS-D and HADS-T for pairs of optimal cutoffs and cutoffs close to the optimal cutoffs among participants recruited from outpatient care settings

HADS-D ^a					HADS-T					HADS-T – HADS-D			
Cutoff	Sensitivity	95% CI	Specificity	95% CI	Cutoff	Sensitivity	95% CI	Specificity	95% CI	Sensitivity	95% CI	Specificity	95% CI
5	0.91	(0.87, 0.94)	0.63	(0.60, 0.67)	11	0.92	(0.89, 0.95)	0.62	(0.59, 0.66)	0.01	(-0.02, 0.04)	-0.01	(-0.03, 0.01)
6	0.87	(0.82, 0.91)	0.72	(0.69, 0.75)	13	0.88	(0.84, 0.91)	0.72	(0.69, 0.75)	0.01	(-0.02, 0.05)	0.00	(-0.01, 0.02)
7 ^b	0.82	(0.75, 0.86)	0.79	(0.76, 0.81)	15 ^c	0.81	(0.76, 0.84)	0.80	(0.77, 0.82)	-0.01	(-0.07, 0.04)	0.01	(-0.01, 0.03)
8	0.71	(0.65, 0.77)	0.85	(0.83, 0.87)	17	0.73	(0.67, 0.78)	0.86	(0.84, 0.88)	0.02	(-0.04, 0.07)	0.01	(-0.00, 0.03)
9	0.60	(0.54, 0.66)	0.90	(0.88, 0.91)	19	0.59	(0.53, 0.65)	0.91	(0.90, 0.92)	-0.01	(-0.08, 0.04)	0.01	(0.00, 0.03)
10	0.49	(0.43, 0.55)	0.93	(0.91, 0.94)	21	0.45	(0.39, 0.52)	0.94	(0.93, 0.95)	-0.04	(-0.11, 0.02)	0.01	(0.00, 0.03)
11	0.38	(0.32, 0.44)	0.95	(0.94, 0.96)	23	0.34	(0.29, 0.39)	0.96	(0.95, 0.97)	-0.04	(-0.10, 0.01)	0.01	(0.00, 0.02)

^a N Studies = 54; N Participants = 9,547; N major depression = 1,072

^b The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-D.

^c The cutoff minimizes the values of the distance to the top-left corner of the ROC curves for HADS-T.

CI: confidence interval