

BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open

Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2016-015247
Article Type:	Research
Date Submitted by the Author:	23-Nov-2016
Complete List of Authors:	Manea, Laura; University of York, Health Sciences Boehnke, Jan; University of York Gilbody, Simon; The University of York, Department of Health Sciences Moriarty, Andrew; University of York, Health Sciences McMillan, Dean; University of York, Department of Health Sciences
Primary Subject Heading:	Mental health
Secondary Subject Heading:	Diagnostics
Keywords:	Depression & mood disorders < PSYCHIATRY, Screening, PHQ-9, diagnostic meta-analysis, allegiance effect

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 **Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A**
2 **systematic review and meta-analysis**

3
4 Laura Manea MMedSci MRCPsych*, Jan Boehnke PhD, Simon Gilbody DPhil FRCPsych
5 FRSA, Andrew Moriarty MSc, Dean McMillan PhD

6
7 *Corresponding Author

8 Hull York Medical School and Department of Health Sciences, ARRC Building, University
9 of York, YO10 5DD

10 Email: laura.manea@york.ac.uk

1
2
3 25 **Abstract**
4

5 26 **Objectives** To investigate whether an authorship effect is found that leads to better
6
7 27 performance in studies conducted by the original developers of the PHQ-9 (non-independent
8
9 28 studies).

10
11 29 **Design** Systematic review with random effects bivariate diagnostic meta-analysis. Search
12
13 30 strategies included electronic databases, examination of reference lists, and forward citation
14
15 31 searches.

16
17 32 **Inclusion criteria** Included studies provided sufficient data to calculate the diagnostic
18
19 33 accuracy of the PHQ-9 against a gold standard diagnosis of major depression using the
20
21 34 algorithm or the summed item scoring method at cut-off point 10.

22
23 35 **Data extraction** Descriptive information, methodological quality criteria, and 2×2
24
25 36 contingency tables.

26
27 37 **Results**
28

29
30 38 Seven non-independent and twenty independent studies reported the diagnostic performance
31
32 39 of the PHQ-9 using the algorithm scoring method. Pooled diagnostic odds ratio (DOR) for
33
34 40 the first group was 64.40, and 15.05 for independent studies group. The allegiance status was
35
36 41 a significant predictor of DOR variation ($p < 0.0001$).

37
38 42 Five non-independent studies and twenty-six independent studies reported the performance of
39
40 43 the PHQ-9 at recommended cut-off point of 10. Pooled DOR for the non-independent group
41
42 44 was 49.31, and 24.96 for the independent studies. The allegiance status was a significant
43
44 45 predictor of DOR variation ($P = 0.015$).

45
46 46 Some potential alternative explanations for the observed authorship effect including
47
48 47 differences in study characteristics and quality were found, though it is not clear how some of
49
50 48 them account for the observed differences
51

52
53 50 **Conclusions**
54

55
56 51 Non-independent studies reported better performance of the PHQ-9. Allegiance status was
57
58 52 predictive of variation in the DOR. Based on the observed differences between independent
59
60

1
2
3 53 and non-independent studies we were unable to conclude or exclude that allegiance effects
4 54 are present in studies examining the diagnostic performance of the PHQ-9. This study
5
6 55 highlights the need for future meta-analyses of diagnostic validation studies of psychological
7
8 56 measures to evaluate the impact of researcher allegiance in the primary studies.
9
10
11 57
12
13 58
14

15 59 **Strengths and limitations of this study**
16
17
18 60
19

- 20 61 • An original study—the first meta-analysis of diagnostic validation studies of
21 62 psychological measures to evaluate the impact of researcher allegiance.
22
23 63 • Using rigorous methodology—strict inclusion/exclusion and quality assessment
24 64 criteria.
25
26 65 • We found that the allegiance effect was a significant predictor of the variation of the
27 66 diagnostic odds ratio in the meta-regression analysis.
28
29
30 67 • Substantial variability observed in methodological quality of included studies.
31
32 68 • Based on the observed methodological differences between the independent and non-
33 69 independent studies we were unable to conclude or exclude that allegiance effects are
34
35 70 present in studies examining the diagnostic performance of the PHQ-9.
36
37 71
38
39 72
40
41 73
42
43
44 74
45
46 75
47
48
49 76
50
51 77
52
53
54 78
55
56 79
57
58
59
60

1
2
3 80 Research on allegiance effects has a long tradition in psychotherapy research. In this context
4 81 *allegiance* describes the phenomenon that researchers and clinicians who developed a
5
6 82 treatment approach or are for other reasons invested in it tend to find larger effect sizes in
7
8 83 favour of their treatment than for comparison groups. (Luborsky et al., 2006) This finding has
9
10 84 been extensively replicated (Dragioti, Dimoliatis, & Evangelou, 2015; Munder, Brüttsch,
11
12 85 Leonhart, Gerger, & Barth, 2013) and is also robust when the quality of research is controlled
13
14 86 for. Researcher allegiance is subject of on-going debates about the design of efficacy studies
15
16 87 as well as implications for policy. (Dragioti et al., 2015; McLeod, 2010; Winter, 2010)
17
18 88 Researcher allegiance is also discussed widely in the literature on experimental as well as
19
20 89 evaluation research. (Staines & Cleland, 2007) Since the motivational underpinnings of
21
22 90 allegiance effects are potentially far more ingrained into human behaviour and decision
23
24 91 making than previously thought (e.g., (Markman & Hirt, 2002)), they may occur commonly
25
26 92 in clinical research in general.

27
28 93 Although it has been suggested that allegiance effects may play a role in the validation of
29
30 94 psychological screening and case-finding tools (e.g., O'Shea et al., in press), systematic
31
32 95 evaluations of this hypothesis are rare and studies that acknowledge potential allegiance
33
34 96 effects in such studies mainly come from forensic psychology and psychiatry backgrounds.
35
36 97 (Blair, Marcus, & Boccaccini, 2008; Lilienfeld & Jones, 2008; Singh, Grann, & Fazel, 2013;
37
38 98 Walters, 2009) Diagnostic validation studies are geared at establishing the sensitivity and
39
40 99 specificity of a screening or case finding tool, which is used in practice to differentiate cases
41
42 100 from non-cases or to decide about whether further assessment or treatment is indicated or will
43
44 101 be offered An allegiance effect in such studies would be seen in systematically higher
45
46 102 sensitivities or specificities if the original author(s) is(are) part of the team of such a study.
47
48 103 Such a bias would have a deleterious affect on practice through promising over-optimistic
49
50 104 accuracy of the screening or case finding tool or in evaluating the cost-effectiveness of the
51
52 105 measure in a screening or case-finding context.

53
54 106 The depression module of the Patient Health Questionnaire (PHQ-9) is a widely used
55
56 107 depression-screening instrument in non-psychiatric settings. The PHQ-9 was developed by a
57
58 108 team of researchers, with its development underwritten by an educational grant from Pfizer
59
60 109 US Pharmaceuticals. (Kroenke, Spitzer, & Williams, 2001) The PHQ-9 can be scored using
110
111 110 different methods, including an algorithm based on DSM-IV criteria and a cut-off based on
112
113 111 summed-item scores. The psychometric properties of these two approaches have been
114
115 112 summarised in two recently published meta-analyses. (Manea, Gilbody, & McMillan, 2015;

1
2
3 113 Moriarty, Gilbody, McMillan, & Manea, 2015)The goal of the current review is to
4 114 investigate, based on an established database of PHQ-9 diagnostic validation studies (Manea
5 115 et al., 2015; Moriarty et al., 2015), whether an allegiance effect is found that leads to an
6 116 increased sensitivity and specificity in studies that were conducted by researchers closely
7 117 connected to the original developers of the instrument.

118 METHODS

119 *Study Selection*

120 Similar search strategies were used in both systematic reviews. (For full details please see
121 Manea et al. (2014) and Moriarty et al. (2015)). Embase, MEDLine and PSYCHInfo were
122 searched from 1999 (when the PHQ-9 was first developed) to August 2013 (Manea et al.,
123 2015) and September 2013 (Moriarty et al., 2015) respectively, using the terms “PHQ-9”,
124 “PHQ”, “PHQ\$” and “patient health questionnaire”. The reference lists of studies fitting the
125 inclusion criteria were manually searched and a reverse citation search in Web of Science
126 was performed. Authors of unpublished studies were contacted and conference abstracts were
127 reviewed in an attempt to minimise publication bias.

128 The following inclusion-exclusion criteria were used:

129 *Population:* Adult population. *Instrument:* Studies that used the PHQ-9. *Comparison*
130 *(reference standard):* The accuracy of the PHQ-9 had to be assessed against a recognised
131 gold-standard instrument for the diagnosis of either Diagnostic and Statistical Manual (DSM)
132 or International Classification of Disease (ICD) criteria for major depression. Studies were
133 included if the diagnoses were made using a standardised diagnostic structured interview
134 schedule (e.g. Mini International Neuropsychiatric Interview (MINI), Structured Clinical
135 Interview for DSM Disorders (SCID)). Unguided clinician diagnoses with no reference to a
136 standard structured diagnostic schedule or comparisons of the PHQ-9 with other self-report
137 measures were excluded. Studies were also excluded if the target diagnosis was not major
138 depressive disorder (MDD, e.g. any depressive disorder). *Outcome:* Studies had to report
139 sufficient information to calculate a 2*2 contingency table for the algorithm or the
140 recommended cut-off point 10. *Study design:* Any design. *Additional criterion:* We avoided
141 double counting of evidence by ensuring that only one study of those that reported
142 overlapping datasets in different journals were included in the meta-analysis. Citations with

1
2
3 143 overlapping samples were examined to establish whether they contained information relevant
4 144 to the research question that was not contained in the included report.

5
6
7 145 *Quality assessment*

8
9
10 146 Quality assessment was performed using the QUADAS-2 tool, a tool for evaluating the risk
11 147 of bias and applicability of primary diagnostic accuracy studies when conducting diagnostic
12 148 systematic reviews. (Whiting et al., 2011) It covers the areas of: patient selection, index test,
13 149 reference standard and flow and timing. (Mann, Hewitt, & Gilbody, 2009) This tool was
14 150 adapted for the two reviews and quality assessments were carried out by two independent
15 151 reviewers for all studies included in the reviews.

16
17
18
19
20 152 *Data synthesis and statistical analysis*

21
22
23 153 We constructed 2x2 tables for cut-off point 10 (Moriarty et al., 2015) and the algorithm
24 154 scoring method (Manea et al., 2015) Pooled estimates of sensitivity, specificity,
25 155 positive/negative likelihood ratios, and diagnostic odds ratios were calculated using random
26 156 effects bivariate meta-analysis. (Reitsma et al., 2005) Summary Receiver Operator
27 157 Characteristic curves (sROC) were constructed using the bivariate model to produce a 95%
28 158 confidence ellipse within ROC space. (Walter, 2002) Each data point in the summary ROC
29 159 space represents a separate study, unlike a traditional ROC plot, which explores the effect of
30 160 varying thresholds on sensitivity and specificity in a single study.

31
32
33 161 We undertook a meta-regression analysis of logit diagnostic odds ratio using research
34 162 allegiance as covariate in the meta-regression model. (Lijmer, Bossuyt, & Heisterkamp,
35 163 2002; S. G. Thompson & Higgins, 2002) Analyses were conducted using STATA version 12,
36 164 with the metan, metandi and metareg user-written commands.

37
38
39
40
41 165 *Allegiance Rating*

42
43
44 166 We rated authorship on a paper of any of the developers of the PHQ-9 - Kurt Kroenke, MD,
45 167 Robert L Spitzer, MD, and Janet B W Williams – as an indicator of potential allegiance. We
46 168 also rated as evidence of allegiance as acknowledged collaborations with the developers of
47 169 the PHQ-9, even if they were not listed as co-authors or if the authors acknowledged funding
48 170 from Pfizer to conduct the study.

49
50
51
52
53
54
55
56 171

1
2
3 172 RESULTS
4
5
6 173
7

8 174 **Overview of included studies**
9

10 175 31 studies reported the diagnostic properties of the PHQ-9 at cut-off point 10 and were
11 included in this analysis (Moriarty et al., 2015) 27 studies were included in the algorithm
12 review (Manea et al., 2015). The study selection flowcharts can be found in Appendix 1
13 (figures 1 and 2). The characteristics of these studies are reported in tables 1 and 2 and the
14 results of the methodological assessment are presented in tables 3 and 4.
15
16
17
18

19 180 **Algorithm scoring method**
20
21
22 181

23
24 182 Descriptive characteristics
25
26

27 183 The descriptive characteristics of the included studies are presented in table 1. Seven
28 individual studies that reported the diagnostic performance of the PHQ-9 using the algorithm
29 scoring method were co-authored by the original developers of the PHQ-9 (Diez-Quevedo,
30 Rangil, Sanchez-Planell, Kroenke, & Spitzer, n.d.; Gräfe, Zipfel, Herzog, & Löwe, 2004;
31 Löwe et al., 2004; Spitzer, Kroenke, & Williams, 1999; Thekkumpurath et al., 2011),
32 specifically acknowledged one of the developers and support by an educational grant from
33 Pfizer US (Muramatsu et al., 2007), or were co-authored by the first author of a previous
34 study that had also been co-authored by one of the developers (Navinés et al., 2012). Twenty
35 independent studies reported the diagnostic properties of the PHQ-9 using the algorithm
36 scoring method.
37
38
39
40
41
42
43
44
45
46

47 194 Three (43%, 3/7) of the non-independent studies were conducted exclusively in hospital
48 settings (Diez-Quevedo et al., n.d.; Navinés et al., 2012; Thekkumpurath et al., 2011). The
49 remaining four studies (67%, 4/7) were conducted in different settings or non-exclusively
50 hospital settings: one in primary care (Spitzer et al., 1999) and three in mixed settings:
51 psycho-somatic walk in clinics and family practices (Gräfe et al., 2004)¹, outpatient clinics
52
53
54
55

56
57 ¹ This study provided separate estimates for the two settings in which it was conducted; therefore
58 separate psychometric estimates were generated for each sample for both algorithm scoring method and
59 summed items scoring method at cut-off point 10 (see below).
60

1
2
3 199 and family practices (Löwe et al., 2004) and primary care and hospital settings (Muramatsu et
4 al., 2007). In the independent group, thirteen (65%, 13/20) studies were conducted in
5 200
6 201 hospital settings (Eack, Greeno, & Lee, 2006; Fann et al., n.d.; Gelaye et al., 2013; Hyphantis
7
8 202 et al., 2011; Inagaki et al.; Khamseh et al., 2011; Persoons, Luyckx, Desloovere,
9
10 203 Vandenberghe, & Fischler, n.d.; Picardi et al., 2005; Stafford, Berk, & Jackson, 2007;
11
12 204 Thombs, Ziegelstein, & Whooley, 2008; A. W. Thompson et al., 2011; Turner et al., 2012;
13
14 205 van Steenbergen-Weijnenburg et al., 2010). Of the remaining seven studies, six were
15
16 206 conducted in primary care settings (Arroll et al., 2010; Ayalon, Goldfracht, & Bech, 2010;
17
18 207 Henkel et al., 2004; Lamers et al., 2008; Lotrakul, Sumrithe, & Saipanish, 2008; Zuithoff et
19
20 208 al., 2010) and one in a community sample (Gjerdingen, Crow, McGovern, Miner, & Center,
21
22 209 2009).

23
24 210 In both groups (independent and non-independent studies), the majority of studies validated a
25
26 211 translated version of the PHQ-9. Two of the studies authored by developers (28%, 2/7)
27
28 212 (Spitzer et al., 1999; Thekkumpurath et al., 2011), and eight (40%, 8/20) independent studies
29
30 213 (Arroll et al., 2010; Eack et al., 2006; Fann et al., n.d.; Gjerdingen et al., 2009; Stafford et al.,
31
32 214 2007; Thombs et al., 2008; A. W. Thompson et al., 2011; Turner et al., 2012) were conducted
33
34 215 in English.

35
36 216 The mean prevalence of major depressive disorder in the group of studies co-authored by
37
38 217 PHQ-9 developers was 13.4 (range 6.1% - 29.2%); in the independent group it was 15.5%
39
40 218 (range 3.9% - 32.4%). The mean age of patients in the PHQ-9 developers group was 45.75;
41
42 219 all but one study had a mean age in the range of 40 to 50 years. In the independent group the
43
44 220 mean age was 54.6 (range 29.3 – 75.0), with almost half (8) of the studies reporting a mean
45
46 221 age of over 60. The percentage of females in the PHQ-9 developers was 56.8% (range 28.6%
47
48 222 - 67.8%) and in the independent group was 59.1 (18% -100%).

49
50
51
52
53
54
55
56
57
58
59
60

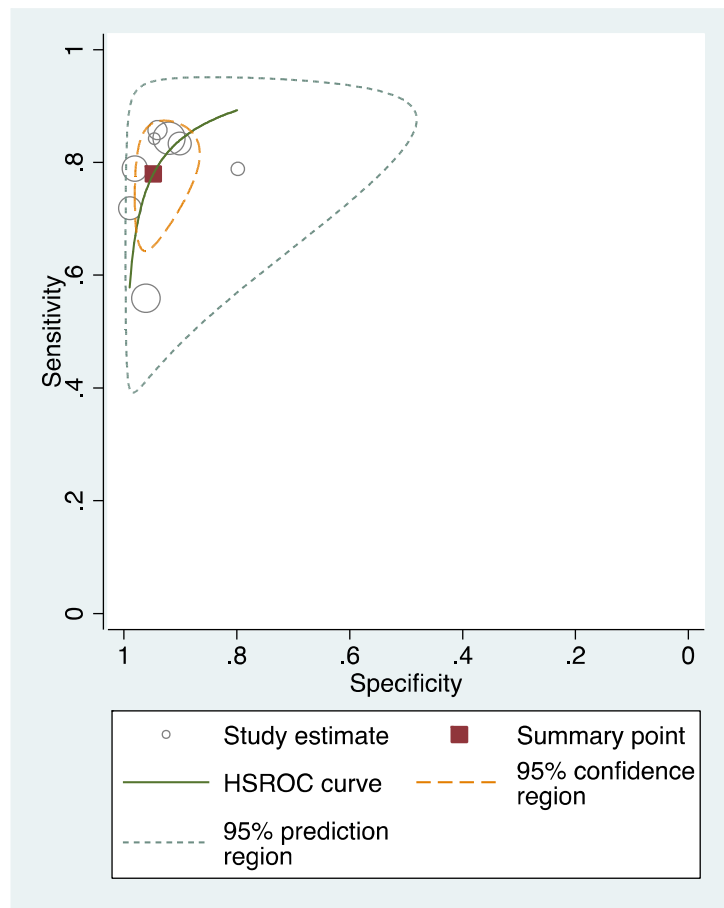
223
224 All of the non-independent studies used a self-reported PHQ-9, whereas in 7 independent
225
226 225 studies (30%, 6/20) the PHQ-9 was administered by a researcher (Ayalon et al., 2010; Fann et
227
228 226 al., n.d.; Gelaye et al., 2013; Gjerdingen et al., 2009; Hyphantis et al., 2011; Inagaki et al.).
229
230 227 Apart from Muramatsu et al. (2007) all of the non-independent studies used the SCID as a
231
232 228 gold standard; the independent studies used a wider range of gold standards including SCAN,
233
234 229 CIDI, MINI, and C-DIS, though the SCID was also frequently used by the independent
235
236 230 studies as well (45%, 9/20 studies).

1
2
3 231 Four out of the seven non-independent studies (57%) did not include a conflict of interests
4 232 statement (Diez-Quevedo et al., n.d.; Gräfe et al., 2004; Muramatsu et al., 2007; Spitzer et al.,
5 233 1999). Also, four (57%) of the non-independent studies acknowledged funding from Pfizer
6 234 (Gräfe et al., 2004; Löwe et al., 2004; Muramatsu et al., 2007; Spitzer et al., 1999). Only one
7 235 study (Muramatsu et al., 2007) acknowledged the collaboration with one of the developers of
8 236 the PHQ-9.

9
10
11
12
13 237 Of the independent studies, twelve (60%) did not include a conflict of interests statement
14 238 (Eack et al., 2006; Fann et al., n.d.; Gelaye et al., 2013; Gjerdingen et al., 2009; Henkel et al.,
15 239 n.d.; Hyphantis et al., 2011; Lamers et al., 2008; Lotrakul et al., 2008; Persoons et al., n.d.;
16 240 Picardi et al., 2005; Stafford et al., 2007; A. W. Thompson et al., 2011). It appears that newer
17 241 studies were more likely to include a conflict of interest statement, which may reflect a recent
18 242 change in reporting. Funding was acknowledged by most studies (18/20) and most received
19 243 funding from academic or/and health research institutions. Two studies received funding
20 244 from pharmaceutical companies – Lundbeck (Ayalon et al., 2010) and Pfizer (Persoons et al.,
21 245 n.d.) and one study acknowledged that Pfizer Italia provided the Italian version of PHQ-9 and
22 246 gave the authors permission to use it (Picardi et al., 2005).

23
24
25
26
27
28
29
30
31 247 Diagnostic test accuracy

32
33 248 Pooled sensitivity and specificity was calculated separately for the independent and non-
34 249 independent studies. Pooled sensitivity for the non-independent studies of the PHQ-9 was
35 250 0.77 (95% CI = 0.70 – 0.84), pooled specificity was 0.94 (95% CI = 0.90 – 0.97) and the
36 251 pooled diagnostic odds ratio was 64.40 (95% CI = 34.15 – 121.43). Heterogeneity was high
37 252 ($I^2 = 78.9\%$). Figure 1 represents the summary ROCs for this set of studies.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

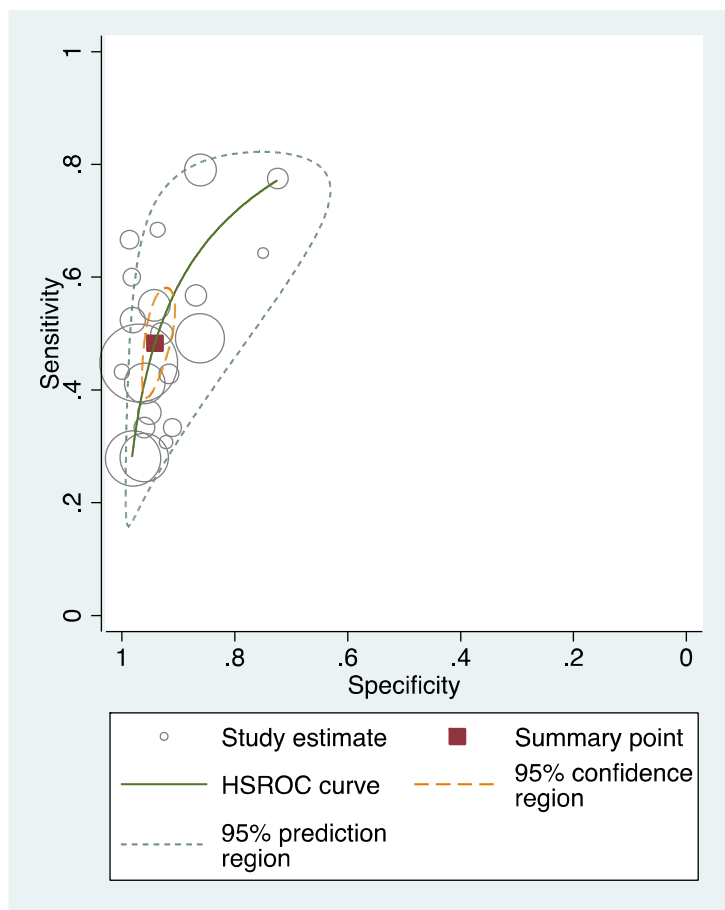


253

254 Figure 1. PHQ-9 algorithm scoring method summary ROC plot of diagnosis of major
 255 depressive disorder in non-independent studies. Pooled sensitivity and specificity using a bi-
 256 variate meta-analysis.

257

258 Pooled sensitivity for the independent studies was lower compared to the developer authored
 259 studies group at 0.48 (95% CI = 0.41 – 0.91); whereas pooled specificity was the same at
 260 0.94 (95% CI = 0.91 – 0.95). The pooled diagnostic odds ratio was approximately four times
 261 lower at 15.05 (95% CI = 11.03 – 20.52) (see figure 2 or sROC). Heterogeneity was
 262 substantial at $I^2 = 68.1\%$.



263

264 Figure 2. PHQ-9 algorithm scoring method summary ROC plot of diagnosis of major
 265 depressive disorder in independent studies. Pooled sensitivity and specificity using a bi-
 266 variate meta-analysis.

267

268

269 The meta-regression analysis for algorithm studies with independent status as the predictor of
 270 the diagnostic odds ratio showed that independent status was a significant predictor of the
 271 diagnostic odds ratio ($p < 0.0001$) and explained a substantial amount of the observed
 272 heterogeneity (51.54%).

273

274 Quality assessment

1
2
3 275 The results of the quality assessment using QUADAS-2 are given in table 3 for the studies
4 276 reporting on the diagnostic performance of the algorithm scoring method. In the patient
5 277 selection domain, more of the independent studies (65%, 13/20) than the non-independent
6 278 (29%, 2/7) met the criterion for consecutive referrals. There were no marked differences on
7
8 279 the other two criteria in this domain (avoid case-control design, avoid inappropriate
9 280 exclusions). In the index test domain, the proportion of studies reporting that the PHQ-9 was
10 281 conducted blind to the reference test was comparable between the two groups. There were
11 282 differences in this domain for those studies using a translated version of the test. All non-
12 283 English non-independent studies (5/5) used an appropriately translated version of the PHQ-9;
13 284 whereas just over a half of the independent studies reported this (55%, 6/11). However, the
14 285 majority of both sets of studies did not report details of psychometric properties of the
15 286 translated version. For the reference test domain, nearly all studies in both groups were rated
16 287 as using a reference test that would correctly classify the condition. While most studies
17 288 conducted by the developers of the PHQ-9 reported that the reference test was interpreted
18 289 blind to the PHQ-9 score (86%, 6/7), this was reported in only 60% (12/20) of the
19 290 independent studies.

20
21 291 The two sets of studies that used translated versions of the reference test were broadly
22 292 comparable. There was a slight indication that the non-independent studies were more likely
23 293 to use an appropriately translated version of the reference test and report data on the
24 294 psychometric properties of the translated version, though the numbers for the translated
25 295 comparison are very low. There were, however, some more notable differences on the flow
26 296 and timing domain. Most of the studies conducted by the developers ensured that the time
27 297 between the index and reference test was under two weeks (86%, 6/7) in comparison to 70%
28 298 (14/20) of the independent studies. More non-independent studies met the criterion for 'all
29 299 participants included in the analysis' (57%, 4/7) than the independent studies (25%).

30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46 300

47 301 **Summed items scoring method (cut-off point 10)**

48
49
50 302

51 303 Descriptive characteristics

52
53
54 304 Table 2 presents the sample characteristics of the thirty-one PHQ-9 validation studies that
55 305 reported the psychometric properties of the PHQ-9 at cut-off point 10. Five of these studies
56
57
58
59
60

1
2
3 306 were co-authored by the original developers of the instrument or acknowledged collaboration
4 307 (Gräfe et al., 2004; Kroenke et al., 2001; Thekkumpurath et al., 2011; Williams et al., 2005)
5 308 or were co-authored by the first author of a previous study that had also been co-authored by
6 309 one of the developers (Navinés et al., 2012). Twenty-six studies were conducted by
7 310 independent researchers.
8
9
10
11

12 311
13
14 312 Three (60%, 3/5) of the non-independent studies (Navinés et al., 2012; Thekkumpurath et al.,
15 313 2011; Williams et al., 2005) and eleven independent studies (42%, 11/26) (Chagas et al.,
16 314 2013; Elderon, Smolderen, Na, & Whooley, 2011; Fann et al., n.d.; Gelaye et al., 2013;
17 315 Hyphantis et al., 2011; Khamseh et al., 2011; Rooney et al., 2013; Stafford et al., 2007;
18 316 Thombs et al., 2008; Watnick, Wang, Demadura, & Ganzini, 2005; Zhang et al., 2013) were
19 317 conducted in hospital settings.
20
21
22
23
24
25

26 318
27
28 319 Three (60%, 3/5) non-independent studies (Kroenke et al., 2001; Thekkumpurath et al., 2011;
29 320 Williams et al., 2005) and thirteen independent studies (13/26) (Adewuya, Ola, & Afolabi,
30 321 2006; Arroll et al., 2010; Elderon et al., 2011; Fann et al., n.d.; Fine et al., 2013; Gilbody,
31 322 Richards, Brealey, & Hewitt, 2007; Gjerdingen et al., 2009; Phelan et al., 2010; Rooney et
32 323 al., 2013; Sidebottom, Harrison, Godecker, & Kim, 2012; Stafford et al., 2007; Thombs et al.,
33 324 2008; Watnick et al., 2005), were conducted in English.
34
35
36
37
38
39

40 325
41 326 The mean prevalence of major depressive disorder in the group of studies authored by PHQ-9
42 327 developers was 13.2% (range 6.1% - 33.5%) and in the independent group was 16.1% (range
43 328 2.5% - 43.2%). The mean age of patients in the PHQ-9 developers group studies was 48.1
44 329 (range 41.9 -61.0) and in the 26 independent studies that reported these data was 49.1 (range
45 330 23.0 – 78.0). The percentage of females in the PHQ-9 developers studies that reported these
46 331 data (Gräfe et al., 2004; Kroenke et al., 2001; Navinés et al., 2012; Thekkumpurath et al.,
47 332 2011) was 56.3% (range 28.6% – 67.8%) and in the independent group was 64.9 % (range
48 333 12% -100%).
49
50
51
52
53
54

55 334
56
57
58
59
60

1
2
3 335 Three of the non-independent studies used the self-reported mode of administration and two
4 336 of them did not specify how the PHQ-9 was administered. In 9 independent studies (34%,
5 337 9/26) the PHQ-9 was administered by the researcher (de Lima Osório, Vilela Mendes,
6 338 Crippa, & Loureiro, 2009; Fann et al., n.d.; Fine et al., 2013; Gelaye et al., 2013; Gjerdingen
7 339 et al., 2009; Hyphantis et al., 2011; Patel et al., 2008; Phelan et al., 2010; Sidebottom et al.,
8 340 2012). All studies authored by developers used SCID as a gold standard; the independent
9 341 studies used a wider range of gold standards including SCAN, CIDI, MINI, CIS-R, C-DIS,
10 342 though the SCID was used in half of the studies (50%, 13/26 studies).

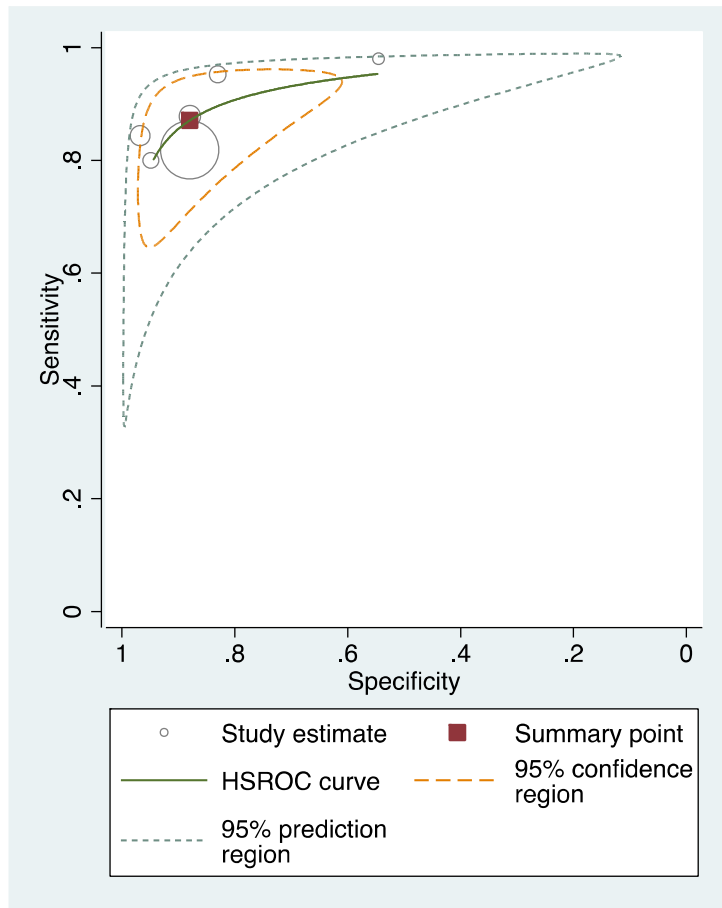
11 343 Three non-independent studies (60%) did not include a conflict of interests statement (Gräfe
12 344 et al., 2004; Kroenke et al., 2001; Williams et al., 2005). Two of these studies (Gräfe et al.,
13 345 2004; Kroenke et al., 2001) acknowledged funding from Pfizer. None of the non-independent
14 346 studies acknowledged collaboration or authorship of one of the developers of the PHQ-9.

15 347 Of the independent studies, thirteen (42%) did not include a conflict of interests statement
16 348 (Adewuya et al., 2006; Arroll et al., 2010; Azah et al., 2005; de Lima Osório et al., 2009;
17 349 Fann et al., n.d.; Gelaye et al., 2013; Gjerdingen et al., 2009; Hyphantis et al., 2011; Liu et
18 350 al., 2011; Lotrakul et al., 2008; Stafford et al., 2007; Watnick et al., 2005; Wittkamp et al.,
19 351 2009). Similar to the algorithm studies, the newer studies were more likely to include a
20 352 conflict of interest statement. Funding was acknowledged by most studies (27/31) and most
21 353 received funding from academic or/and health research institutions. One study (Gilbody et
22 354 al., 2007) acknowledged that the last author involved in the development of one of the
23 355 instruments (CORE-OM), 'but does not gain financially from its use'. One study (Elderon et
24 356 al., 2011) acknowledged funding from industry, AHA Pharmaceuticals Roundtable, but stated
25 357 that 'the funding organisations had no role in the design or conduct of the study, collection,
26 358 management, analysis or interpretation of data; or preparation, review or approval of the
27 359 manuscript. Fine et al., 2013 disclosed that the last author had financial and consulting
28 360 interests (Pfizer was not cited as one of them).

29 361

30 362 Diagnostic test accuracy

31 363 Pooled sensitivity for the studies linked to the developers of the PHQ-9 was 0.87 (95% CI =
32 364 0.77 – 0.93), pooled specificity was 0.87 (95% CI = 0.76 – 0.94) and the pooled diagnostic
33 365 odds ratio was 49.31 (95% CI = 25.74 – 94.48) – see table 5. Heterogeneity was moderate (I^2
34 366 = 55.1%). Figure 4 represents the summary ROCs for this group.

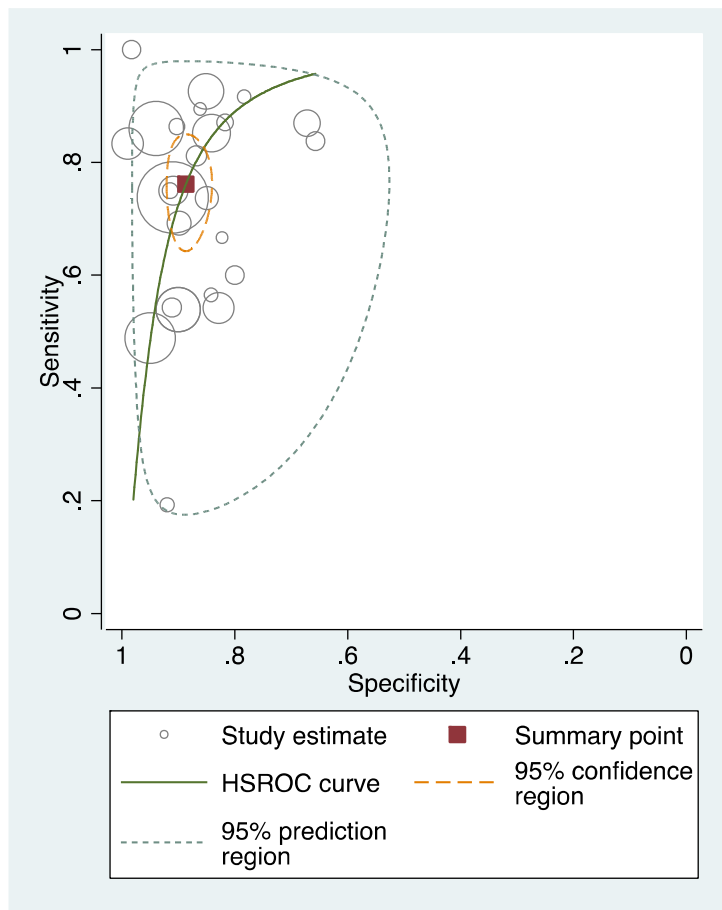


367

368 Figure 3. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot of
 369 diagnosis of major depressive disorder in non-independent studies. Pooled sensitivity and
 370 specificity using a bi-variate meta-analysis.

371

372 Pooled sensitivity for the independent studies was 0.76 (95% CI, 0.67 – 0.83), pooled
 373 specificity was 0.88 95% CI (0.85 – 0.91) and the pooled diagnostic odds ratio was 24.96
 374 (95% CI 14.81 – 42.08), approximately half that of the non-independent studies (table 5).
 375 Heterogeneity was high at $I^2 = 81.5\%$. Figure 5 represents the summary ROCs for this group.



376

377 Figure 4. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot of
 378 diagnosis of major depressive disorder in independent studies. Pooled sensitivity and
 379 specificity using a bi-variate meta-analysis.

1
2
3 380 The meta-regression for the studies using a cut-off point of 10 with allegiance status of the
4 381 predictor showed that allegiance status was a significant predictor of the diagnostic odds ratio
5 382 (P = 0.015) and explained 18.95% of observed heterogeneity.
6
7
8

9 383

10
11 384 Quality assessment
12

13
14 385 The results of the quality assessment using the QUADAS-2 are given in table 4. For the
15 386 patient selection domain, the two groups of studies were broadly comparable on two items
16 387 (consecutive or random sample, avoid case-control design). However, all of the studies from
17 388 the non-independent studies were rated as avoiding inappropriate exclusions (5/5) in contrast
18 389 to 58% (15/26) of the independent studies.
19
20
21
22

23 390

24
25 391 On the index test domain, there were a number of differences between the two groups of
26 392 studies. More of the independent studies (81%, 21/26) reported that the PHQ-9 was
27 393 interpreted blind to the reference test compared to 60% (3/5) of the studies conducted by the
28 394 developers of the PHQ-9. All (5/5) of the studies from the PHQ-9 developers were rated as
29 395 pre-specifying the threshold on the PHQ-9 compared to 73% (19/26) of the independent
30 396 studies. The two sets of studies were broadly comparable in terms of two items from the
31 397 reference test domain (correctly classify target condition, reference test interpreted blind).
32 398 Only one of the studies from the developers of the PHQ-9 used a translated version of the
33 399 index test or reference test, so it is not possible to comment on differences between the two
34 400 sets of studies in terms of these items from the index or reference test domains. For the flow
35 401 and timing domain, the two groups of studies were broadly comparable for two of the criteria
36 402 (interval of two weeks or less, all participants receive same reference test). However, fewer
37 403 than half of the independent studies met the criterion for 'all participants included in the
38 404 analysis' (42%, 11/26); whereas all of the studies by the developers of the PHQ-9 met this
39 405 criterion.
40
41
42
43
44
45
46
47
48
49

50 406

51
52
53 407 **Discussion**
54

55
56 408 This is to our knowledge the first systematic examination of a possible 'allegiance' or
57 409 authorship effect in the validation of screening or case finding psychological instrument for a
58
59
60

1
2
3 410 common mental health disorder. We reviewed diagnostic validation studies of the PHQ-9, a
4 411 widely used depression screening-instrument. We found that non-independent studies
5 412 reported higher sensitivity paired with similar specificity compared to studies conducted by
6 413 independent researchers. When entered as a covariate in meta-regression analyses,
7 414 independence status was predictive of variation in the DOR for both the algorithm scoring
8 415 method and the summed-item scoring method at a cut-off point of 10.
9
10
11
12

13
14 416

15
16 417 Previous research has proposed several possible explanations for the allegiance effect (Blair
17 418 et al., 2008; Lilienfeld & Jones, 2008; Singh et al., 2013). One possibility is the advertent bias
18 419 that may serve to inflate the performance of a test when evaluated by those who have
19 420 developed it. However, before concluding that the differences are due to this, it is important
20 421 to explore and rule out alternative explanations. First, it is possible that any observed
21 422 differences are a result of differences in study characteristics of the two sets of studies (e.g.,
22 423 setting, clinical population). Secondly, differences in the methodological quality of the
23 424 studies may also account for any differences. These possibilities are examined below.
24
25
26
27
28
29

30 425

31
32
33 426 Difference in study characteristics as potential alternative explanations
34

35 427 The two sets of studies were broadly comparable in terms of gender and the prevalence of
36 428 depression, so these variables are unlikely to offer an explanation for the differences. While
37 429 there were some indications from both sets of comparisons that the PHQ-9 may have been
38 430 researcher-administered more often in the independent studies, it is not immediately clear
39 431 how this would lead to lowered diagnostic performance.
40
41
42
43

44 432

45
46 433 The diagnostic meta-analyses of the PHQ-9 (Manea et al., 2015; Moriarty et al., 2015) have
47 434 shown that the sensitivity and DOR of the PHQ-9 tends to be lower in hospital settings for
48 435 both algorithm and summed-item scoring methods. Whilst the fact that proportionally more
49 436 independent algorithm studies were conducted in secondary care could explain the lower
50 437 sensitivity and DOR values in the algorithm studies, in the studies that reported the cut-off
51 438 point of 10 this would not be the case as proportionally more studies authored by developers
52 439 were conducted in hospital settings.
53
54
55
56
57
58
59
60

440

441 Similarly, differences in the proportions of studies using translated versions of the PHQ-9 are
442 also unlikely to offer an obvious explanation of the difference in diagnostic performance,
443 because in the algorithm set of studies more of the non-independent studies used a translated
444 version of the test, but the proportions were in the opposite direction for the studies using a
445 cut off of 10. A similar conclusion is also likely to apply to the age of the samples. There
446 were more older adults studies in the independent than non-independent studies in the
447 algorithm comparison. Depression could be more difficult to identify in older adults due to
448 physical co-morbidities that may present with similar symptomatology to depression and
449 could account for the lower diagnostic performance in the independent studies. However, the
450 independent samples in the studies that reported the psychometric properties at cut-off point
451 10 had younger samples than the non-independent studies, so this would not support this
452 interpretation.

453

454 The SCID was used as the gold standard in nearly all of the non-independent studies. The fact
455 that some independent studies used other gold standards could potentially explain the poorer
456 psychometric properties of the PHQ-9 in these studies. The SCID is often regarded as the
457 most valid of the available semi-structured interviews used in depression diagnostic validity
458 studies as the reference standard. If we assume that this is the case and, furthermore, that the
459 PHQ-9 is an accurate method of screening for depression, then the PHQ-9 may be more
460 likely to agree with the SCID than other reference standards.

461

462 Differences in methodological quality as potential alternative explanations

463 The quality of the studies was evaluated using the QUADAS-2. Although there were several
464 potential methodological differences between the two groups of studies from the algorithm
465 papers, not all of these offer obvious explanations of the observed differences and some are
466 unlikely as explanations. For example, more of the studies from the developers of the PHQ-9
467 ensured that the reference test was interpreted blind to the index test. This is unlikely to
468 account for the observed differences, because a lack of blinding is typically associated with
469 artificially increased diagnostic performance, which is in the opposite direction to the pattern
470 of results observed here. The impact of some other differences is less clear-cut. For example,

1
2
3 471 a higher number of the independent studies met the criterion for consecutive referrals. For
4 472 this to provide an explanation of the of the observed differences, the non-consecutive nature
5 473 of the referrals in the studies by those who had developed the PHQ-9 would need to have led
6 474 to the over-inclusion of true positives or under-inclusion of false negatives given that these
7
8 475 studies tended to report higher sensitivity relative to the independent studies (and vice versa
9 476 for the independent studies). It is not immediately obvious how this would occur. The studies
10 477 by the developers of the PHQ-9 were more likely to have met the criterion of ‘included all
11 478 participants in the analysis’. It is possible that the greater loss of participants from the
12 479 independent studies may have artificially reduced the observed diagnostic accuracy, though,
13 480 again, it is not immediately obvious how this would have affected the true positive and false
14 481 negative rates. Although there is not an obvious explanation of how these differences in
15 482 methodological quality could account for the observed differences in diagnostic performance,
16 483 it is important to recognise that they cannot on that basis be ruled out.
17
18
19
20
21
22
23
24
25
26
27

484

28 485 There are, however, two differences in methodological quality among the algorithm studies
29 486 that are clearer potential alternative explanations. The higher rate of appropriate translations
30 487 among the studies conducted by the developers of the PHQ-9 is potentially important,
31 488 because lower diagnostic estimates may be expected from studies that have poorly translated
32 489 versions of the index test. In the flow and timing domain, more of the studies by the
33 490 developers of the PHQ-9 ensured that there was a less than two-week interval between the
34 491 index and reference test. This is consistent with lower diagnostic performance in the
35 492 independent studies: as the interval increases it is likely that depression status may change
36 493 and this would lead to lower levels of agreement between the index test and the reference
37 494 test.
38
39
40
41
42
43
44
45
46
47

495

48 496 There were also differences on some quality assessment items between the two sets of studies
49 497 in the summed item scoring method comparison. The threshold was reported as pre-specified
50 498 in all of the studies by the developers of the PHQ-9 in contrast to approximately three
51 499 quarters of the independent studies. On the face of it, this is unlikely to explain the observed
52 500 differences, because the use of a pre-specified cut-off point is likely to be associated with
53 501 lower not higher diagnostic test performance. One possibility, however, is that studies that
54 502 performed poorly at this cut-off point were less likely to be reported by those who had
55
56
57
58
59
60

1
2
3 503 developed the measure. As discussed in more detail in the limitations section, we were unable
4 504 to explore this possibility through the use of formal tests for publication bias.
5
6

7 505
8

9
10 506 All non-independent studies avoided inappropriate exclusions compared to approximately
11 507 half of the independent studies. While this is a potential alternative explanation of the
12 508 differences it is not immediately obvious how this would explain the differences in diagnostic
13 509 performance between the two sets of studies. Fewer than half of the independent studies met
14 510 the criterion for 'all participants included in the analysis' in contrast to all of the studies by
15 511 the developers of the PHQ-9 met this criterion, but again this difference should if at usually
16 512 work against the inclusive studies, not those excluding cases. More of the independent studies
17 513 reported that the PHQ-9 was interpreted blind to the reference test. This does offer a potential
18 514 explanation, because the absence of blinding may artificially inflate diagnostic accuracy.
19
20
21
22
23
24

25 515
26

27 516 **Limitations**

28
29
30 517 The results of this review need to be viewed in the light of the limitations of the primary
31 518 studies that contributed to the review and the review itself. An important consideration is to
32 519 establish whether any observed differences between the diagnostic performance of the
33 520 independent and non-independent studies are better accounted for by study characteristic or
34 521 methodological differences. Caution, however, is needed in interpreting any differences,
35 522 because of the small number of non-independent studies in both the algorithm and cut-off 10
36 523 comparisons. The small number of non-independent studies also meant that we were also
37 524 unable to explore the potential role of publication bias in the independent and non-
38 525 independent studies. At least 10 studies are required to use standard methods of examining
39 526 publication bias, but the number of non-independent studies in both the algorithm and cut-off
40 527 10 comparisons were fewer than this.
41
42
43
44
45
46
47
48

49 528
50

51 529
52
53

54 530 **Conclusions and implications for further research.**

55
56
57
58
59
60

1
2
3 531 The aims of the review was to investigate whether an allegiance effect is found that leads to
4 532 an increased diagnostic performance in diagnostic validation studies that were conducted by
5 533 teams connected to the original developers of the PHQ-9. Our analyses showed that
6 534 diagnostic studies conducted by independent researchers had lower sensitivity paired with
7 535 similar specificity compared to studies that were classified as non-independent. This
8 536 conclusion held for both the algorithm and cut-off 10 studies. We explored a range of
9 537 possible alternative explanations for the observed allegiance effect including both differences
10 538 in study characteristics and study quality. A number of potential differences were found,
11 539 though for some of these it is not clear how they would necessarily account for the observed
12 540 differences. However, there were a number of differences that offered potential alternative
13 541 explanations unconnected to allegiance effects. These included the greater use of the SCID in
14 542 the studies rated as non-independent in both the algorithm and the cut-off 10 studies. In the
15 543 algorithm studies, the studies rated as non-independent were also more likely to use an
16 544 appropriate translation of the PHQ-9 and were also more likely to ensure that the index and
17 545 reference test were conducted within two weeks of each other, both of which may be
18 546 associated with an improvement in observed diagnostic performance of an instrument. The
19 547 majority of studies in both meta-analyses did not provide clear statements about potential
20 548 conflict of interest and/or funding, however the newer studies were more likely to provide
21 549 such statements, which may reflect increasing transparency in this area of research.

22 550

23 551 We cannot, therefore, conclude that allegiance effects are present in studies examining the
24 552 diagnostic performance of the PHQ-9; but nor can we rule them out. Conflicts of interest are
25 553 an important area of investigation in medical and behavioural research, particularly due to
26 554 concerns about trial results being influenced by industry sponsorship. Future diagnostic
27 555 validity in this area should as a matter of routine present clear statements about potential
28 556 conflicts of interest and funding, particularly relating to the development of the instrument
29 557 under evaluation. Future meta-analyses of diagnostic validation studies of psychological
30 558 measures should routinely evaluate the impact of researcher allegiance in the primary studies
31 559 examined in the meta-analysis.

32 560

33 561

1
2
3 562 **Contributors** LM led on all stages of the review and is the guarantor. We used an established
4 563 database of diagnostic validation studies of the PHQ-9 (Manea et al., 2015; Moriarty et al.,
5 564 2015) SG provided expert advice on methodology and approaches to assessment of the
6 565 evidence base. AM carried out the literature searches, screened the studies, extracted data and
7 566 assessed the quality of the included studies for one of the systematic reviews (Moriarty et al.,
8 567 2015) . LM carried out the literature searches, screened the studies, extracted data and
9 568 assessed the quality of the included studies for the other systematic review (Manea et al.,
10 569 2015), analysed the data for both systematic reviews and drafted the report. JB involved in
11 570 the development of the study, wrote the introduction section of the review and contributed to
12 571 the production of the final report. DM supervised the quality assessment, methodology and
13 572 approaches to evidence synthesis, provided senior advice and support throughout and
14 573 contributed to the production of the final report. All parties were involved in drafting and/or
15 574 commenting on the report.

16 575

17 576 **Competing interests** None declared.

18 577

19 578 **Provenance and peer review** Not commissioned; externally peer reviewed.

20 579

21 580 **Data sharing statement** No additional data are available.

22 581

23 582 REFERENCES

- 24 583 1. Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health
25 584 questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university
26 585 students. *Journal of Affective Disorders*, 96(1–2), 89–93.
27 586 <http://doi.org/10.1016/j.jad.2006.05.021>
- 28 587 2. (2010). Validation of PHQ-2 and PHQ-9 to Screen for Major Depression in the Primary
29 588 Care Population. *The Annals of Family Medicine*, 8(4), 348–353.
30 589 <http://doi.org/10.1370/afm.1139>

- 1
2
3 590 3. Ayalon, L., Goldfracht, M., & Bech, P. (2010). "Do you think you suffer from
4 591 depression?" Reevaluating the use of a single item question for the screening of
5 592 depression in older primary care patients. *International Journal of Geriatric Psychiatry*,
6 593 25(5), 497–502. <http://doi.org/10.1002/gps.2368>
- 7
8
9
10 594 4. Azah, M. N. N., Shah, M. E. M., Juwita, S., Bahri, I. S., Rushidi, W. M. W. M., & Jamil,
11 595 Y. M. (2005). Validation of the Malay Version Brief Patient Health Questionnaire
12 596 (PHQ-9) among Adult Attending Family Medicine Clinics. *International Medical*
13 597 *Journal*.
- 14
15
16
17
18 598 5. Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is There an Allegiance Effect for
19 599 Assessment Instruments? Actuarial Risk Assessment as an Exemplar. *Clinical*
20 600 *Psychology: Science and Practice*, 15(4), 346–360. <http://doi.org/10.1111/j.1468->
21 601 2850.2008.00147.x
- 22
23
24
25 602 6. Chagas, M. H. N., Tumas, V., Rodrigues, G. R., Machado-de-Sousa, J. P., Filho, A. S.,
26 603 Hallak, J. E. C., & Crippa, J. A. S. (2013). Validation and internal consistency of Patient
27 604 Health Questionnaire-9 for major depression in Parkinson's disease. *Age and Ageing*,
28 605 42(5), 645–649. <http://doi.org/10.1093/ageing/aft065>
- 29
30
31
32
33 606 7. de Lima Osório, F., Vilela Mendes, A., Crippa, J. A., & Loureiro, S. R. (2009). Study of
34 607 the Discriminative Validity of the PHQ-9 and PHQ-2 in a Sample of Brazilian Women
35 608 in the Context of Primary Health Care. *Perspectives in Psychiatric Care*, 45(3), 216–
36 609 227. <http://doi.org/10.1111/j.1744-6163.2009.00224.x>
- 37
38
39
40 610 8. Validation and utility of the patient health questionnaire in diagnosing mental disorders in
41 611 1003 general hospital Spanish inpatients. *Psychosomatic Medicine*, 63(4), 679–86.
42 612 Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11485122>
- 43
44
45
46 613 9. Dragioti, E., Dimoliatis, I., & Evangelou, E. (2015). Disclosure of researcher allegiance in
47 614 meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal.
48 615 *BMJ Open*, 5(6), e007206–e007206. <http://doi.org/10.1136/bmjopen-2014-007206>
- 49
50
51
52 616 10. Eack, S. M., Greeno, C. G., & Lee, B.-J. (2006). Limitations of the Patient Health
53 617 Questionnaire in Identifying Anxiety and Depression: Many Cases Are Undetected.
54 618 *Research on Social Work Practice*, 16(6), 625–631.
55 619 <http://doi.org/10.1177/1049731506291582>
- 56
57
58
59
60

- 1
2
3 620 11. Elderon, L., Smolderen, K. G., Na, B., & Whooley, M. A. (2011). Accuracy and prognostic
4 621 value of American Heart Association: recommended depression screening in patients
5 622 with coronary heart disease: data from the Heart and Soul Study. *Circulation*.
6 623 *Cardiovascular Quality and Outcomes*, 4(5), 533–40.
7 624 <http://doi.org/10.1161/CIRCOUTCOMES.110.960302>
- 8
9
10
11
12 625 12. Fann, J. R., Bombardier, C. H., Dikmen, S., Esselman, P., Warms, C. A., Pelzer, E., ...
13 626 Temkin, N. (n.d.). Validity of the Patient Health Questionnaire-9 in assessing depression
14 627 following traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, 20(6),
15 628 501–11. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16304487>
- 16
17
18
19
20 629 13. Fine, T. H., Contractor, A. A., Tamburrino, M., Elhai, J. D., Prescott, M. R., Cohen, G.
21 630 H., ... Calabrese, J. R. (2013). Validation of the telephone-administered PHQ-9 against
22 631 the in-person administered SCID-I major depression module. *Journal of Affective*
23 632 *Disorders*, 150(3), 1001–1007. <http://doi.org/10.1016/j.jad.2013.05.029>
- 24
25
26
27 633 14. Gelaye, B., Williams, M. A., Lemma, S., Deyessa, N., Bahretibeb, Y., Shibre, T., ...
28 634 Andrew Zhou, X. H. (2013). Validity of the patient health questionnaire-9 for depression
29 635 screening and diagnosis in East Africa. *Psychiatry Research*, 210(2).
30 636 <http://doi.org/10.1016/j.psychres.2013.07.015>
- 31
32
33
34 637 15. Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in
35 638 medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-
36 639 analysis. *Journal of General Internal Medicine*, 22, 1596–1602.
37 640 <http://doi.org/10.1007/s11606-007-0333-y>
- 38
39
40
41
42 641 16. Gjerdingen, D., Crow, S., McGovern, P., Miner, M., & Center, B. (2009). Postpartum
43 642 depression screening at well-child visits: validity of a 2-question screen and the PHQ-9.
44 643 *Annals of Family Medicine*, 7(1), 63–70. <http://doi.org/10.1370/afm.933>
- 45
46
47
48 644 17. Gräfe, K., Zipfel, S., Herzog, W., & Löwe, B. (2004). Screening psychischer Störungen
49 645 mit dem “Gesundheitsfragebogen für Patienten (PHQ-D)“. *Diagnostica*, 50(4), 171–181.
50 646 <http://doi.org/10.1026/0012-1924.50.4.171>
- 51
52
53
54 647 18. Henkel, V., Mergl, R., Kohnen, R., Allgaier, A.-K., Möller, H.-J., & Hegerl, U. (n.d.).
55 648 Use of brief depression screening tools in primary care: consideration of heterogeneity
56 649 in performance in different patient groups. *General Hospital Psychiatry*, 26(3), 190–8.
- 57
58
59
60

- 1
2
3 650 <http://doi.org/10.1016/j.genhosppsy.2004.02.003>
4
5 651 19. Hyphantis, T., Kotsis, K., Voulgari, P. V., Tsifetaki, N., Creed, F., & Drosos, A. A.
6
7 652 (2011). Diagnostic accuracy, internal consistency, and convergent validity of the Greek
8
9 653 version of the patient health questionnaire 9 in diagnosing depression in rheumatologic
10
11 654 disorders. *Arthritis Care & Research*, 63(9), 1313–1321.
12
13 655 <http://doi.org/10.1002/acr.20505>
14
15 656 20. Inagaki, M., Ohtsuki, T., Yonemoto, N., Kawashima, Y., Saitoh, A., Oikawa, Y., ...
16
17 657 Yamada, M. Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general
18
19 658 internal medicine primary care at a Japanese rural hospital: a cross-sectional study.
20
21 659 *General Hospital Psychiatry*, 35(6), 592–7.
22
23 660 <http://doi.org/10.1016/j.genhosppsy.2013.08.001>
24
25 661 21. Khamseh, M. E., Baradaran, H. R., Javanbakht, A., Mirghorbani, M., Yadollahi, Z., &
26
27 662 Malek, M. (2011). Comparison of the CES-D and PHQ-9 depression scales in people
28
29 663 with type 2 diabetes in Tehran, Iran. *BMC Psychiatry*, 11(1), 61.
30
31 664 <http://doi.org/10.1186/1471-244X-11-61>
32
33 665 22. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief
34
35 666 depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–13.
36
37 667 Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11556941>
38
39 668 23. Lamers, F., Jonkers, C. C. M., Bosma, H., Penninx, B. W. J. H., Knottnerus, J. A., & van
40
41 669 Eijk, J. T. M. (2008). Summed score of the Patient Health Questionnaire-9 was a reliable
42
43 670 and valid method for depression screening in chronically ill elderly patients. *Journal of*
44
45 671 *Clinical Epidemiology*, 61(7), 679–687. <http://doi.org/10.1016/j.jclinepi.2007.07.018>
46
47 672 24. Lijmer, J. G., Bossuyt, P. M. M., & Heisterkamp, S. H. (2002). Exploring sources of
48
49 673 heterogeneity in systematic reviews of diagnostic tests. *Statistics in Medicine*, 21(11),
50
51 674 1525–1537. <http://doi.org/10.1002/sim.1185>
52
53 675 25. Lilienfeld, S. O., & Jones, M. K. (2008). Allegiance Effects in Assessment: Unresolved
54
55 676 Questions, Potential Explanations, and Constructive Remedies. *Clinical Psychology:*
56
57 677 *Science and Practice*, 15(4), 361–365. <http://doi.org/10.1111/j.1468-2850.2008.00148.x>
58
59 678 26. Liu, S.-I., Yeh, Z.-T., Huang, H.-C., Sun, F.-J., Tjung, J.-J., Hwang, L.-C., ... Yeh, A.
60
679 W.-C. (2011). Validation of Patient Health Questionnaire for depression screening

- 1
2
3 680 among primary care patients in Taiwan. *Comprehensive Psychiatry*, 52(1), 96–101.
4 681 <http://doi.org/10.1016/j.comppsy.2010.04.013>
5
6
7 682 27. Lotrakul, M., Sumrithe, S., & Saipanish, R. (2008). Reliability and validity of the Thai
8 683 version of the PHQ-9. *BMC Psychiatry*, 8(1), 46. [http://doi.org/10.1186/1471-244X-8-](http://doi.org/10.1186/1471-244X-8-46)
9 684 46
10
11
12
13 685 28. Löwe, B., Spitzer, R. L., Gräfe, K., Kroenke, K., Quenter, A., Zipfel, S., ... Herzog, W.
14 686 (2004). Comparative validity of three screening questionnaires for DSM-IV depressive
15 687 disorders and physicians' diagnoses. *Journal of Affective Disorders*, 78(2), 131–40.
16 688 Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/14706723>
17
18
19
20 689 29. Luborsky, L., Diguier, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., ...
21 690 Schweizer, E. (2006). The Researcher's Own Therapy Allegiances: A "Wild Card" in
22 691 Comparisons of Treatment Efficacy. *Clinical Psychology: Science and Practice*, 6(1),
23 692 95–106. <http://doi.org/10.1093/clipsy.6.1.95>
24
25
26
27
28 693 30. Manea, L., Gilbody, S., & McMillan, D. (2015). A diagnostic meta-analysis of the Patient
29 694 Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression.
30 695 *General Hospital Psychiatry*, 37(1), 67–75.
31 696 <http://doi.org/10.1016/j.genhosppsy.2014.09.009>
32
33
34
35 697 31. Mann, R., Hewitt, C. E., & Gilbody, S. M. (2009). Assessing the quality of diagnostic
36 698 studies using psychometric instruments: applying QUADAS. *Social Psychiatry and*
37 699 *Psychiatric Epidemiology*, 44(4), 300–307. <http://doi.org/10.1007/s00127-008-0440-z>
38
39
40
41 700 32. Markman, K. D., & Hirt, E. R. (2002). Social Prediction and the "Allegiance Bias."
42 701 *Social Cognition*, 20(1), 58–86. <http://doi.org/10.1521/soco.20.1.58.20943>
43
44
45 702 33. McLeod, J. (2010). Taking allegiance seriously—implications for research policy and
46 703 practice. *European Journal of Psychotherapy and Counselling*. Retrieved from
47 704 <http://www.tandfonline.com/doi/abs/10.1080/13642531003637791?journalCode=rejp20>
48 705 #.WC8nwktA-ao.mendeley
49
50
51
52
53 706 34. Moriarty, A. S., Gilbody, S., McMillan, D., & Manea, L. (2015). Screening and case
54 707 finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a
55 708 meta-analysis. *General Hospital Psychiatry*, 37(6), 567–576.
56 709 <http://doi.org/10.1016/j.genhosppsy.2015.06.012>
57
58
59
60

- 1
2
3 710 35. Munder, T., Brüttsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher
4 711 allegiance in psychotherapy outcome research: An overview of reviews. *Clinical*
5 712 *Psychology Review*, 33(4), 501–511. <http://doi.org/10.1016/j.cpr.2013.02.002>
- 6
7
8 713 36. Muramatsu, K., Miyaoka, H., Kamijima, K., Muramatsu, Y., Yoshida, M., Otsubo, T., &
9 714 Gejyo, F. (2007). The patient health questionnaire, Japanese version: validity according
10 715 to the mini-international neuropsychiatric interview-plus. *Psychological Reports*, 101(3
11 716 Pt 1), 952–60. <http://doi.org/10.2466/pr0.101.3.952-960>
- 12
13
14 717 37. Navinés, R., Castellví, P., Moreno-España, J., Gimenez, D., Udina, M., Cañizares, S., ...
15 718 Martín-Santos, R. (2012). Depressive and anxiety disorders in chronic hepatitis C
16 719 patients: Reliability and validity of the Patient Health Questionnaire. *Journal of Affective*
17 720 *Disorders*, 138(3), 343–351. <http://doi.org/10.1016/j.jad.2012.01.018>
- 18
19
20 721 38. Patel, V., Araya, R., Chowdhary, N., King, M., Kirkwood, B., Nayak, S., ... Weiss, H. A.
21 722 (2008). Detecting common mental disorders in primary care in India: a comparison of
22 723 five screening questionnaires. *Psychological Medicine*, 38(2).
23 724 <http://doi.org/10.1017/S0033291707002334>
- 24
25
26 725 39. Persoons, P., Luyckx, K., Desloovere, C., Vandenberghe, J., & Fischler, B. (n.d.).
27 726 Anxiety and mood disorders in otorhinolaryngology outpatients presenting with
28 727 dizziness: validation of the self-administered PRIME-MD Patient Health Questionnaire
29 728 and epidemiology. *General Hospital Psychiatry*, 25(5), 316–23. Retrieved from
30 729 <http://www.ncbi.nlm.nih.gov/pubmed/12972222>
- 31
32
33 730 40. Phelan, E., Williams, B., Meeker, K., Bonn, K., Frederick, J., LoGerfo, J., & Snowden,
34 731 M. (2010). A study of the diagnostic accuracy of the PHQ-9 in primary care elderly.
35 732 *BMC Family Practice*, 11(1), 63. <http://doi.org/10.1186/1471-2296-11-63>
- 36
37
38 733 41. Picardi, A., Adler, D. A., Abeni, D., Chang, H., Pasquini, P., Rogers, W. H., & Bungay,
39 734 K. M. (2005). Screening for depressive disorders in patients with skin diseases: a
40 735 comparison of three screeners. *Acta Dermato-Venereologica*, 85(5), 414–9.
41 736 <http://doi.org/10.1080/00015550510034966>
- 42
43
44 737 42. Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., &
45 738 Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces
46 739 informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*,

- 1
2
3 740 58(10), 982–990. <http://doi.org/10.1016/j.jclinepi.2005.02.022>
4
5 741 43. Rooney, A. G., McNamara, S., Mackinnon, M., Fraser, M., Rampling, R., Carson, A., &
6
7 742 Grant, R. (2013). Screening for major depressive disorder in adults with cerebral glioma:
8
9 743 an initial validation of 3 self-report instruments. *Neuro-Oncology*, *15*(1), 122–129.
10
11 744 <http://doi.org/10.1093/neuonc/nos282>
12
13 745 44. Sidebottom, A. C., Harrison, P. A., Godecker, A., & Kim, H. (2012). Validation of the
14
15 746 Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. *Archives of*
16
17 747 *Women's Mental Health*, *15*(5), 367–374. <http://doi.org/10.1007/s00737-012-0295-x>
18
19 748 45. Singh, J. P., Grann, M., & Fazel, S. (2013). Authorship Bias in Violence Risk
20
21 749 Assessment? A Systematic Review and Meta-Analysis. *PLoS ONE*, *8*(9), e72484.
22
23 750 <http://doi.org/10.1371/journal.pone.0072484>
24
25 751 46. Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-
26
27 752 report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of
28
29 753 Mental Disorders. Patient Health Questionnaire. *JAMA*, *282*(18), 1737–44. Retrieved
30
31 754 from <http://www.ncbi.nlm.nih.gov/pubmed/10568646>
32
33 755 47. Stafford, L., Berk, M., & Jackson, H. J. (2007). Validity of the Hospital Anxiety and
34
35 756 Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients
36
37 757 with coronary artery disease. *General Hospital Psychiatry*, *29*(5), 417–424.
38
39 758 <http://doi.org/10.1016/j.genhosppsych.2007.06.005>
40
41 759 48. Staines, G. L., & Cleland, C. M. (2007). Bias in meta-analytic estimates of the absolute
42
43 760 efficacy of psychotherapy. *Review of General Psychology*, *11*(4), 329–347.
44
45 761 <http://doi.org/10.1037/1089-2680.11.4.329>
46
47 762 49. Thekkumpurath, P., Walker, J., Butcher, I., Hodges, L., Kleiboer, A., O'Connor, M., ...
48
49 763 Sharpe, M. (2011). Screening for major depression in cancer outpatients: the diagnostic
50
51 764 accuracy of the 9-item patient health questionnaire. *Cancer*, *117*(1), 218–27.
52
53 765 <http://doi.org/10.1002/cncr.25514>
54
55 766 50. Thombs, B. D., Ziegelstein, R. C., & Whooley, M. A. (2008). Optimizing detection of
56
57 767 major depression among patients with coronary artery disease using the patient health
58
59 768 questionnaire: data from the heart and soul study. *Journal of General Internal Medicine*,
60
769 *23*(12), 2014–7. <http://doi.org/10.1007/s11606-008-0802-y>

- 1
2
3 770 51. Thompson, A. W., Liu, H., Hays, R. D., Katon, W. J., Rausch, R., Diaz, N., ... Vickrey,
4 771 B. G. (2011). Diagnostic accuracy and agreement across three depression assessment
5 772 measures for Parkinson's disease. *Parkinsonism & Related Disorders*, *17*(1), 40–45.
6 773 <http://doi.org/10.1016/j.parkreldis.2010.10.007>
7
8
9
10 774 52. Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be
11 775 undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559–1573.
12 776 <http://doi.org/10.1002/sim.1187>
13
14
15
16 777 53. Turner, A., Hambridge, J., White, J., Carter, G., Clover, K., Nelson, L., & Hackett, M.
17 778 (2012). Depression screening in stroke: a comparison of alternative measures with the
18 779 structured diagnostic interview for the diagnostic and statistical manual of mental
19 780 disorders, fourth edition (major depressive episode) as criterion standard. *Stroke; a*
20 781 *Journal of Cerebral Circulation*, *43*(4), 1000–5.
21 782 <http://doi.org/10.1161/STROKEAHA.111.643296>
22
23
24
25
26
27 783 54. van Steenberg-Weijnenburg, K. M., de Vroege, L., Ploeger, R. R., Brals, J. W.,
28 784 Vloedveld, M. G., Veneman, T. F., ... van der Feltz-Cornelis, C. M. (2010). Validation
29 785 of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized
30 786 outpatient clinics. *BMC Health Services Research*, *10*(1), 235.
31 787 <http://doi.org/10.1186/1472-6963-10-235>
32
33
34
35
36 788 55. Walter, S. D. (2002). Properties of the summary receiver operating characteristic (SROC)
37 789 curve for diagnostic test data. *Statistics in Medicine*, *21*(9), 1237–1256.
38 790 <http://doi.org/10.1002/sim.1099>
39
40
41
42 791 56. Walters, G. D. (2009). The Psychological Inventory of Criminal Thinking Styles and
43 792 Psychopathy Checklist: Screening version as incrementally valid predictors of
44 793 recidivism. *Law and Human Behavior*, *33*(6), 497–505. [http://doi.org/10.1007/s10979-](http://doi.org/10.1007/s10979-008-9167-3)
45 794 [008-9167-3](http://doi.org/10.1007/s10979-008-9167-3)
46
47
48
49 795 57. Watnick, S., Wang, P.-L., Demadura, T., & Ganzini, L. (2005). Validation of 2
50 796 depression screening tools in dialysis patients. *American Journal of Kidney Diseases :*
51 797 *The Official Journal of the National Kidney Foundation*, *46*(5), 919–24.
52 798 <http://doi.org/10.1053/j.ajkd.2005.08.006>
53
54
55
56
57 799 58. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J.
58
59
60

- 1
2
3 800 B., ... QUADAS-2 Group. (2011). QUADAS-2: a revised tool for the quality
4 801 assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, 155(8), 529–36.
5 802 <http://doi.org/10.7326/0003-4819-155-8-201110180-00009>
6
7
8
9 803 59. Williams, L. S., Brizendine, E. J., Plue, L., Bakas, T., Tu, W., Hendrie, H., & Kroenke, K.
10 804 (2005). Performance of the PHQ-9 as a screening tool for depression after stroke.
11 805 *Stroke; a Journal of Cerebral Circulation*, 36(3), 635–8.
12 806 <http://doi.org/10.1161/01.STR.0000155688.18207.33>
13
14
15
16 807 60. Winter, D. A. (2010, May 7). Editorial. Routledge. Retrieved from
17 808 [http://www.tandfonline.com/doi/abs/10.1080/13642531003637726?needAccess=true&j](http://www.tandfonline.com/doi/abs/10.1080/13642531003637726?needAccess=true&journalCode=rejp20#.WC8nUSu6MP8.mendeley)
18 809 [ournalCode=rejp20#.WC8nUSu6MP8.mendeley](http://www.tandfonline.com/doi/abs/10.1080/13642531003637726?needAccess=true&journalCode=rejp20#.WC8nUSu6MP8.mendeley)
19
20
21
22 810 61. Wittkamp, K., van Ravesteijn, H., Baas, K., van de Hoogen, H., Schene, A., Bindels, P.,
23 811 ... van Weert, H. (2009). The accuracy of Patient Health Questionnaire-9 in detecting
24 812 depression and measuring depression severity in high-risk groups in primary care.
25 813 *General Hospital Psychiatry*, 31(5), 451–459.
26 814 <http://doi.org/10.1016/j.genhosppsy.2009.06.001>
27
28
29
30
31 815 62. Zhang, Y., Ting, R., Lam, M., Lam, J., Nan, H., Yeung, R., ... Chan, J. C. N. (2013).
32 816 Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong
33 817 Chinese subjects with type 2 diabetes. *Journal of Affective Disorders*, 151(2), 660–666.
34 818 <http://doi.org/10.1016/j.jad.2013.07.014>
35
36
37
38
39 819 63. Zuithoff, N. P., Vergouwe, Y., King, M., Nazareth, I., van Wezep, M. J., Moons, K. G., &
40 820 Geerlings, M. I. (2010). The Patient Health Questionnaire-9 for detection of major
41 821 depressive disorder in primary care: consequences of current thresholds in a
42 822 cross-sectional study. *BMC Family Practice*, 11(1), 98. [http://doi.org/10.1186/1471-](http://doi.org/10.1186/1471-2296-11-98)
43 823 [2296-11-98](http://doi.org/10.1186/1471-2296-11-98)
44
45
46
47
48 824

Table 1: Descriptive characteristics of algorithm studies (Manea et al., 2014)

Study	Sample characteristics (Country, setting, age, sex)	Sample size and % depressed	PHQ-9 characteristics	Diagnostic standard	a) Conflict of interest (COI) declaration b) Funding c) Relationship with original developers
Diez-Quevedo et al. (2001)	Country: Spain Setting: Medical and surgical tertiary hospitals Age (yrs): M=43 (SD=14.2) Female: 45.6%	N = 1003 Depressed: 8.2%	Administration: Self-report Language: Spanish	DSM-III-R SCID	a) a) No COI declaration b) Funding acknowledged (academic institutions) c) Not acknowledged
Gräfe et al. (2004)	Country: Germany	N = 528	Language: German	DSM-IV	a) No COI declaration b) Acknowledged funding from Pfizer c) Not acknowledged

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Setting: Depressed:
psychosomatic 29.2%
walk-in clinics psychosomatic Administration: SCID
and family patients; self-report
practices 6.16% medical patients

Age (yrs): M =
41.9 (SD =
13.8)

Female: 67.8%

Lowe et al. (2004)

Country: N = 501 Administration: DSM-IV
Germany Self-report

- a) COI declaration 'This study was supported by unrestricted restricted grants from Pfizer Germany and from the medical faculty of the University of Heidelberg Germany, and there are no COI.'
- b) Acknowledged funding from Pfizer and academic institution
- c) Not acknowledged

Setting: Depressed: Language: SCID
Outpatient 13.2% German
clinics and family practices

1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14	Muramatsu et al.				
15	(2007)	Country: Japan	N = 131	Administration: Self-report	DSM-IV
16					
17					
18					
19		Setting: Primary care and general hospital	Depressed: 28.2%	Language: Japanese	MINI
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30	Navinés et al. (2012)	Country: Spain	N = 500	Administration: Self-report	DSM-IV
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					

Age (yrs): M =
41.7 (SD =
13.8)

Female: 67.1%

- a) No COI declaration
- b) Acknowledged funding from Pfizer
- c) Acknowledged one of the developers of the PHQ-9:
'The authors acknowledge Dr R L Spitzer'

Setting:

Primary care
and general
hospital

Depressed:
28.2%

Language:
Japanese

MINI

Age (yrs): M =
43.3 (SD =
16.4)

Female: 59.5%

- a) All authors declared that they had no COI.
- b) Role of funding source declared
- c) Not acknowledged

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Setting:
General hospital
(patients with chronic HCV)
Depressed: 6.4%
Language: Spanish
SCID

Age (yrs): M = 43.4 (SD = 10.2)
Female: 28.6%

Spitzer et al. (1999)

Country: US
N = 3000 (585 received SCID)
Administration: Self-report
DSM-III-R

- a) No COI declaration
- b) Acknowledged funding from Pfizer. 'Drs Spitzer and Williams receive honoraria and consulting money from Pfizer Inc, which has supported this work.'
- c) N/A

Setting:
Primary care
Depressed: 10%
Language: English
SCID

Age (yrs): M = 46 (SD = 17.2)
Female: 66%

1						
2						
3						
4						
5						
6						
7						a) COI declaration: 'Supported by Cancer
8	Thekkumpurath et al.	Country: UK	N = 782	Administration:	DSM-IV	Research UK'
9	(2010)			Not stated		b) As in a)
10						c) Not acknowledged
11		Setting:				
12		Hospital	Depressed:	Language:	SCID	
13		(cancer	6.3% (of the	English		
14		patients)	whole sample)			
15						
16		Age (yrs): M =				
17		61				
18						
19		Female: 63%				
20						
21						
22						
23						a) COI declaration: 'The project was funded by an
24						Investigator's Initiated Research Grant from Lundbeck
25						International given to Dr Liat Ayalon. Lundbeck
26						International had no other involvement in the project
27						concept of design or in this paper. Per Bech has
28				Administration:		occasionally over the past 3 years until August 2008
29	Ayalon et al. (2010)	Country: Israel	N = 153	Researcher	DSM-IV	received funding from and has been speaker or
30				administered		member of advisory boards for pharmaceutical
31						companies with an interest in the drug treatment of
32						affective disorders (Astra-Zeneca, Lilly, H. Lundbeck
33						A/S, Lundbeck Foundation and Organon). '
34						b) Acknowledged funding from Lundbeck International
35						
36						
37						
38						
39						
40						
41						
42						
43						
44						
45						
46						
47						
48						
49						

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

	Age (yrs): M = 75 (SD = 8.1)	Depressed: 3.9 %	Language: Hebrew	SCID	
	Female: 40.5 %				
	Country: US	N = 50	Administration: Self-report	DSM-IV	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Eack et al. (2006)	Setting: Community mental health centers for children	Depressed: 28%	Language: English	SCID	
	Age (yrs): M = 39.20 (SD 9.63)				
	Female: 100%				
Fann et al. (2005)	Country: US	N = 135	Administration: Telephone-administered	DSM-IV	a) No COI declaration b) Funding acknowledged (academic institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Setting:
Trauma
hospital
(inpatients with
traumatic brain
injury)

Depressed:
16.3%

Language:
English

SCID

Age (yrs): M =
42 (SD=17.9)

Female: 29.1%

Gelaye et al. (2011)

Country:
Ethiopia

N = 363

Administration:
Researcher-
administered

DSM-IV

- a) No COI declaration
- b) Funding acknowledged (academic /health
research institutions)

Setting:
General
hospital

Depressed:
12.6%

Language:
Amharic

SCAN

Age (yrs): 34.9
(SD=11.6)

Female: 63.1 %

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Gjerdingen et al. (2009)	Country: US	N = 438	Administration: Telephone or self-report	DSM-IV	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
	Setting: Community	Depressed: 4.6%	Language: English	SCID	
	Age (yrs): M = 29.3				
	Female: 100%				
Henkel et al. (2004)	Country: Germany	N = 448	Administration: self-report	DSM-IV	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
	Setting: primary care	Depressed: 10%		CIDI	
	Age (yrs): not reported		Language: German		
	Female: 74%				

1					
2					
3					
4					
5					
6					
7					a) No COI declaration
8					b) No funding acknowledgement
9	Hyphantis et al.	Country:	N = 213	Administration:	DSM-IV
10	(2011)	Greece		Researcher administered	
11					
12		Setting:			
13		Hospital –	Depressed:	Language:	MINI
14		rheumatology	32.4%	Greek	
15		patients			
16					
17		Age (yrs): M =			
18		54.2 (SD =			
19		13.5)			
20					
21		Female: 74%			
22					
23					
24					a) COI declaration: ‘The authors declare that they have no competing interests.’
25			N = 104 out of	Administration:	DSM-IV
26	Inagaki et al. (2013)	Country: Japan	511 received	Researcher	
27			MINI	administered	b) Funding acknowledged (academic /health research institutions)
28					
29					
30		Setting:			
31		General	Depressed:	Language:	MINI
32		hospital	7.4%	Japanese	
33					
34					
35		Age whole			
36		sample (yrs): M			
37		= 73.5 (SD =			
38		12.3)			
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

	Female: 59.3%				
Khamseh et al. (2011)	Country: Iran	N = 185	Administration: Self report	DSM-IV	a) COI declaration: The authors declared no competing interests b) Funding acknowledged (academic /health research institutions)
	Setting: Diabetes clinic	Depressed: 43.2%	Language: Persian	SCID	
	Age (yrs): M = 56.17 (SD = 9.60)				
	Female: 51.9%				
Lamers et al. (2008)	Country: Netherlands	N = 713	Administration: Self report	DSM-IV	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
	Setting: Primary care (elderly)	Depressed: 10.7%	Language: Dutch	MINI	
	Age (yrs): M = 71.4 (SD = 6.90)				
	Female: 48.2%				

1						
2						
3						
4						
5						
6						
7					a) No COI declaration	
8	Lotrakul et al. (2008)	Country: Thailand	N = 279	Administration: Self report	DSM-IV	b) Funding acknowledged (academic /health research institutions)
9						
10					MINI	
11						
12						
13		Setting: Primary care	Depressed: 6.8%	Language: Thai		
14						
15						
16						
17		Age (yrs): M =				
18		45.0 (SD =				
19		14.30)				
20						
21		Female: 73.7%				
22						
23						
24	Persoons et al.	Country:	N = 268 (97	Administration:	DSM-IV	a) No COI declaration
25	(2003)	Belgium	received	Self-report		b) Funding acknowledged (academic /health
26			MINI)			research institutions) and Pfizer Belgium
27						
28						
29		Setting:				
30		Hospital	Depressed:	Language:	MINI	
31		(otolaryngology	16.5%	Dutch		
32		patients)				
33						
34		Age (yrs): M =				
35		48.2 (SD =				
36		12.9)				
37						
38						
39						
40						
41						
42						
43						
44						
45						
46						
47						
48						
49						

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

	Female: 65.6%				
Picardi et al. (2005)	Country: Italy	N = 141	Administration: Self-report	DSM-IV	a) No COI declaration b) Funding acknowledged (academic /health research institutions). Acknowledged Pfizer Italia SRL for providing the Italian version of the PHQ-9 and for permission to use it.
	Setting: Hospital (dermatology inpatients)	Depressed: 8.5%	Language: Italian	SCID	
	Age (yrs): M = 37.5				
	Female: 56%				
Stafford et al. (2007)	Country: Australia	N = 193	Administration: Self-report	DSM-IV	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
	Setting: Hospital (cardiology patients)	Depressed: 18%	Language: English	MINI	
	Age (yrs): M = 64.1 (SD = 10.3)				

1					
2					
3					
4					
5					
6					
7		Female: 66%			
8					
9					
10		Country: US	N = 1024	Administration: Not stated	DSM
11					
12					
13					a) COI declaration "None disclosed"
14					b) Funding acknowledged (academic/health research institutions)
15		Setting: Hospital			
16		(outpatients with coronary heart disease)	Depressed: 22%	Language: English	C-DIS
17					
18	Thombs et al. (2008)				
19					
20					
21					
22		Age (yrs): M =			
23		67 (SD = 11)			
24					
25					
26		Female: 18%			
27					
28					
29					a) No COI declaration
30	Thompson et al.				b) Funding acknowledged (academic/health research institutions)
31	(2010)	Country: US	N = 214	Administration: Self administered	DSM-IV
32					
33					SCID
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

	Setting: Patients with Parkinson Disease	Depressed: 14%	Language: English		
	Age (yrs): 72.5 (SD = 9.6)				
	Female: 42%				
Turner et al. (2012)	Country: Australia	N = 72	Administration: Self administered	DSM-IV	a) COI declaration: Disclosures 'None'. b) Funding acknowledged (academic/health research institutions)
	Setting: Stroke patients	Depressed: 18%	Language: English	SCID	
	Age (yrs): 66.7 (SD = 13.1)				
	Female: 47.2%				
van Steenberg- Weijnenburg (2010)	Country: Netherlands	N = 197	Administration: Self administered	DSM-IV	a) COI declaration: 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions) - 'this had no influence on the content of this article'.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Setting:
Diabetes
patients
Depressed:
18.8%
Language:
Dutch
SCID
Age (yrs): M =
61.8 (SD =
13.6)
Female: 48.7%

Country:
Netherlands
N = 1338
Administration:
Self-report
DSM-IV

- a) COI declaration ‘The authors declare that they have no competing interests.’
- b) Funding acknowledged (academic/health research institutions).

Zuithoff et al.
(2010)

Setting:
Primary care
Depressed:
13%
Language:
Dutch
CIDI
Age (yrs): M =
51 (sd = 16.7)
Female: 63%

Table 2: Descriptive characteristics of the summed items scoring method studies cut-off point 10 (Moriarty et al, 2015)

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	a) Conflict of interest (COI) declaration b) Funding c) Relationship with original developers
13. Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Mean age: 41.9 (SD = 13.8) Female: 67.8%	N = 528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Administration: self-report Language: German Cut-offs: 10 to 14	DSM-IV SCID	c) No COI declaration d) Acknowledged funding from Pfizer e) Not acknowledged
16. Kroenke et al. (2001)	Country: USA Setting: Primary care Mean age: 46 (SD=17) Female: 66%	N = 580 7.1% MDD	Administration: Self-report Language: English Cut-offs: 9 to 15	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) N/A
22. Navinés et al.	Country: Spain	N = 500	Administration: Self-	DSM-IV	a) All authors declared

(2012)	Setting: General hospital (patients with chronic HCV) Mean age: 43.4 (SD = 10.2) Female: 28.6%	6.4% MDD	report Language: Spanish Cut-offs: 10	SCID	that they had no COI. b) Role of funding source declared c) Not acknowledged
29. Thekkumpurath et al. (2010)	Country: UK Setting: Hospital (cancer patients) Mean age: 61 Female: 63%	N = 782 6.3% MDD (of the whole sample)	Administration: Not stated Language: English Cut-offs: 5 to 10	DSM-IV SCID	c) COI declaration: 'Supported by Cancer Research UK' d) As in a) e) Not acknowledged
33. Williams et al. (2005)	Country: USA Setting: Secondary care (Post-stroke) Mean age: Unclear Female: Unclear	N = 316 33.5% MDD	Administration: Unclear Language: English Cut-offs: 10	DSM-IV SCID	c) No COI declaration d) Funding acknowledged (academic institutions) e) Not acknowledged

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	d) Conflict of interest declaration e) Funding
-------	------------------------	-----------------------	-----------------------	---------------------	---

				d	
1. Adewuya et al. (2006)	Country: Nigeria Setting: community (students) Mean age: 24.8 (15-40) Female: 41.2%	N = 512 2.5% MDD	Administration: Self-report Language: English Cut-offs: 8 to 12	DSM-IV MINI	a) No COI declaration b) No funding declaration
2. Arroll et al. (2010)	Country: New Zealand Setting: Primary care Mean age: 49 (17-99) Female: 61%	N = 2642 6.2% MDD	Administration: Not stated Language: English Cut-offs: 8,10,12,15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
3. Azah et al. (2005)	Country: Malaysia Setting: Primary care Mean age: 38.7 (18-79) Female: 61.7%	N = 180 16.6% MDD	Administration: Self-report Language: Malay Cut-offs: 5 to 12	DSM-IV CIDI	b) No COI declaration c) Funding acknowledged (academic /health research institutions)
4. Chagas et al. (2013)	Country: Brazil	N = 84	Administration: self-report	DSM-IV SCID	a) COI declaration "None declared"

	Setting: Secondary care Mean age: Not stated Female: 52.7%	25.5% MDD	Language: Brazilian Cut-offs: 7 to 10		b) Funding acknowledged (academic/health research institutions)
6. de Lima Osorio et al. (2009)	Country: Brazil Setting: Primary care Mean age: Unclear Female: 100%	N = 177 34% MDD	Administration: research assistants Language: Brazilian Portuguese Cut-offs: 10 to 15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)
7. Elderon et al. (2011)	Country: USA Setting: Secondary care Mean age: Unclear Female: 18%	N = 1022 18.3% MDD	Administration: self-report Language: English Cut-offs: 10	C-DIS	a) COI declaration – ‘No disclosures’ b) Funding acknowledged (academic institutions and industry – AHA Pharmaceuticals Roundtable) – ‘The funding organisations had no role in the design or conduct of the study, collection, management, analysis or interpretation of

					data; or preparation, review or approval of the manuscript.'
8. Fann et al. (2005)	Country: US Setting: Trauma hospital (inpatients with traumatic brain injury) Mean age: 42 (SD=17.9) Female: 29.1%	N = 135 16.3% MDD	Administration: Telephone-administered Language: English Cut-offs: 10	DSM-IV SCID	b) No COI declaration c) Funding acknowledged (academic institutions)
9. Fine et al. (2013)	Country: USA Setting: Primary care (Ohio Army National Guard) Mean age: 31 (17-60) Female: 12%	N = 498 21.5% MDD	Administration: Telephone-administered Language: English Cut-offs: 10,15	DSM-IV SCID-I	a) COI – last author disclosed financial and consulting interests (Pfizer not one of them). All other authors declared that they have no COI. b) Funding acknowledged – DoD Medical Research. “The sponsor had no role in study design, data collection, analysis,

					interpretation of results, report writing or manuscript submission.
10. Gelaye et al. (2013)	Country: Ethiopia Setting: General hospital Mean age: 34.9 (SD=11.6) Female: 63.1 %	N = 363 12.6% MDD	Administration: Researcher-administered Language: Amharic Cut-offs: 9 to 11	DSM-IV SCAN	c) No COI declaration d) Funding acknowledged (academic /health research institutions)
11. Gilbody et al. (2007)	Country: UK Setting: Primary care Mean age: 42.5 (SD 13.6) Female: 77%	N = 96 37.5 MDD	Administration: Not stated Language: English Cut-offs: 9 to 13	DSM-IV SCID	a) COI declaration – last author involved in the development of one of the instruments (CORE-OM), 'but does not gain financially from its use. b) Funding acknowledged (academic /health research institutions)
12. Gjerdingen et	Country: USA	N = 438	Administration:	DSM-IV	c) No COI declaration

al. (2009)	Setting: Community Mean age: 29.3 Female: 100%	4.6% MDD	Telephone or self-report Language: English Cut-offs: 10	SCID	d) Funding acknowledged (academic /health research institutions)
14. Hyphantis et al. (2011)	Country: Greece Setting: Hospital – rheumatology patients Mean age: 54.2 (SD = 13.5) Female: 74%	N = 213 32.4% MDD	Administration: Researcher administered Language: Greek Cut-offs: 4 to 16	DSM-IV MINI	c) No COI declaration d) No funding acknowledgement
15. Khamseh et al. (2011)	Country: Iran Setting: Outpatient diabetic clinic Mean age: 56.1 (SD=9.6) Female: 51.8%	N = 185 43.2% MDD	Administration: Self-report Language: Persian Cut-offs: 10,13	DSM-IV SCID	c) COI declaration: The authors declared no competing interests d) Funding acknowledged (academic /health research institutions)

19. Liu et al. (2011)	Country: Taiwan Setting: Primary care Mean age: Not specified Female: 60.9%	N = 1532 3.3% MDD	Administration: Self-report Language: Chinese version Cut-offs: 9 to 11	SCAN	a) a) No COI declaration b) Funding acknowledged (academic /health research institutions)
20. Lotrakul et al. (2008)	Country: Thailand Setting: Primary care Mean age: 45.0 (SD = 14.30) Female: 73.7%	N = 279 6.8% MDD	Administration: Self report Language: Thai Cut-offs: 7 to 15	DSM-IV MINI	c) No COI declaration d) Funding acknowledged (academic /health research institutions)
23. Patel et al. (2008)	Country: India Setting: Primary care Mean age: 37.5 (18-83) Female: 56.4%	N = 299 4.3% MDD	Administration: Face-to-face interview Language: Not specified Cut-offs: 7 to 15	CIS-R	a) COI declaration – No Declaration of Interest b) Funding acknowledged (academic /health research institutions)
24. Phelan et al. (2010)	Country: USA Setting: Primary care	N = 71 12% MDD	Administration: Research assistant	DSM-IV SCID	a) COI declaration – No competing interests

	(elderly) Mean age: 78 (SD=7) Female: 62%		Language: English Cut-offs: 8 to 12		b) Funding acknowledged (academic /health research institutions) . "The funder had no role in the study design, methods, data collection, analysis or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication.
25. Rooney et al. (2013)	Country: UK Setting: Secondary care (glioma) Mean age: 54.2 (SD=12.3) Female: 42.6%	N = 129 13.5% MDD	Administration: Self-report Language: English Cut-offs: 8 to 11	DSM-IV SCID	a) COI declaration "The authors declare that they have no COI" b) Funding acknowledged (academic/health research institutions)
26. Sherina et al. (2012)	Country: Malaysia Setting: Primary care	N= 146 21.2% MDD	Administration: Self-report	CIDI	a) COI declaration "The authors declare that they

	Mean age: 30.9 (18-81) Female: 100%		Language: Malay Cut-offs: 10		have no competing interests" b) Funding acknowledged (academic/health research institutions)
27. Sidebottom et al. (2012)	Country: USA Setting: Community (prenatal) Mean age: 23 (SD=5.5) Female: 100%	N = 745 3.6% MDD	Administration: Interview Language: English Cut-offs: 10	DSM-IV SCID	a) COI declaration "The authors declare that they have no financial COI" b) Funding acknowledged (academic/health research institutions)
28. Stafford et al. (2007)	Country: Australia Setting: Secondary care (cardiac procedures) Mean age: 64.14 (38-91) Female: 19.2%	N = 193 18.1% MDD	Administration: Self-report Language: English Cut-offs: 10	DSM-IV MINI	b) No COI declaration c) Funding acknowledged (academic/health research institutions)
30. Thombs et al. (2008)	Country: US Setting: Hospital (outpatients with coronary heart disease) Mean age: 67 (SD = 11)	N = 1024 22% MDD	Administration: Not stated Language: English Cut-offs: 7 to 10	DSM C-DIS	b) COI declaration "None disclosed" b) Funding acknowledged (academic/health research institutions)

	Female: 18%				
32. Watnick et al. (2005)	Country: USA Setting: Secondary care (dialysis) Mean age: 63 (SD=15) Female: 32.3%	N = 62 19% MDD	Administration: Self-report Language: English Cut-offs: 10	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
34. Wittkamp et al. (2009)	Country: Netherlands Setting: Primary care Mean age: 49.8 Female: 66.7%	N = 664 12.3% MDD	Administration: Self-report Language: Not specified Cut-offs: 10 and 15	DSM-IV SCIDI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
35. Zhang et al. (2013)	Country: Hong Kong Setting: Secondary care (diabetic outpatients) Mean age: 55.1 (SD=9.5) Female: 40.8%	N = 99 23.2% MDD	Administration: Self-report Language: Chinese version Cut-offs: 15	DSM-IV MINI	a) COI declaration – last author acknowledged financial COI. The other authors declare that they have no competing interests. b)) Funding

					acknowledged (academic/health research institutions)
36. Zuithoff et al. (2010)	Country: Netherlands Setting: Primary care Age (yrs): M = 51 (sd = 16.7) Female: 63%	N = 1338 Depressed: 13%	Administration: Self- report Language: Dutch	DSM-IV CIDI	b) COI declaration "The authors declare that they have no competing interests. b) Funding acknowledged (academic/health research institutions)

Table 3: Quality assessment of included studies in the algorithm meta-analysis (Manea et al., 2014)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Study	Patient selection: Consecutive or random sample	Patient selection: Avoid case-control / avoid artificially inflated base rate	Patient selection: Avoided inappropriate exclusions	Patient selection: Overall risk of bias	Index test: PHQ-9 interpreted blind to reference test	Index test: If translated, appropriate translation	Index test: If translated, psychometric properties reported	Index test: Overall risk of bias
Diez-Quevedo et al. (2001)	✗	✓	✗	High	?	✓	✓	Unclear
Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	✓	Unclear
Lowe et al. (2004)	✗	✓	✓	High	✓	✓	✓	Low
Muramatsu et al. (2007)	?	✓	?	Unclear	✓	✓	?	Unclear
Navines et al. (2012)	✓	✓	✓	Low	✓	✓	?	Unclear
Spitzer et al. (1999)	✗	✓	✓	High	✓	n/a	n/a	Low
Thekkumpurath et al. (2010)	✗	✗	✓	High	✓	n/a	n/a	Low
Arroll et al. (2010)	✓	✓	✓	Low	✓	n/a	n/a	Low
Ayalon et al. (2010)	?	✓	✓	Unclear	?	✓	?	Unclear
Eack et al. (2006)	?	✓	?	Unclear	?	n/a	n/a	Unclear
Fann et al. (2005)	✓	✗	✗	High	✓	n/a	n/a	Low
Gelaye et al. (2013)	?	✗	?	High	✓	✓	?	Unclear
Gjerdengen et al. (2009)	✓	✓	✓	Low	?	n/a	n/a	Unclear
Henkel et al. (2004)	✓	✓	✓	Low	?	n/a	n/a	Unclear
Hyphantis et al. (2011)	✓	✓	✗	High	✓	?	?	Unclear

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Inagaki et al. (2013)	✓	✗	✓	High	✓	?	?	Unclear
Khamseh et al. (2011)	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. (2008)	✓	✗	✗	High	✓	?	?	Unclear
Lotrakul et al. (2008)	✗	✓	?	High	✓	✓	?	Unclear
Persoons et al. (2003)	✓	✓	✓	Low	✓	✓	n/a	Low
Picardi et al. (2005)	✓	✓	✓	Low	✓	?	?	Unclear
Stafford et al. (2007)	✓	✓	✓	Low	✓	n/a	n/a	Low
Thombs et al. (2008)	✗	✓	?	Unclear	?	n/a	n/a	Unclear
Thomson et al. (2011)	?	✓	✓	Unclear	?	n/a	n/a	Unclear
Turner et al. (2012)	✓	✓	✓	Low	✓	n/a	n/a	Low
Van Steenberg-Wijnenburg (2010)	?	✓	✓	Unclear	?	?	?	Unclear
Zuithoff et al. (2010)	✓	✓	✓	Low	✓	✓	?	Unclear

✓ = criterion met; ✗ = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

¹If studies reported multiple cut-off points, 'threshold pre-specified' is coded as not applicable.

Table 3: Quality assessment of included studies in the algorithm meta-analysis (Manea et al., 2015) (continued)

Study	Reference test: Reference test correctly classifies target condition	Reference test: Reference test interpreted blind to PHQ-9	Reference test: If translated, appropriate translation	Reference test: If translated, psychometric properties reported	Reference test: Overall risk of bias	Flow / timing: Interval of two weeks or less	Flow / timing: All participants receive same reference test	Flow / timing: All participants included in analysis?	Flow / timing: Overall risk of bias
Diez-Quevedo et al. (2001)	✓	✓	✓	?	Unclear	✓	✓	✓	Low
Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
Lowe et al. (2004)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Muramatsu et al. (2007)	✓	✓	✓	✓	Low	✓	✓	?	Unclear
Navines et al. (2012)	✓	✓	?	?	Unclear	✓	✓	✓	Low
Spitzer et al. (1999)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✗	High
Arroll et al. (2010)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Ayalon et al. (2010)	✓	?	✓	?	Unclear	?	✓	✓	Unclear
Eack et al. (2006)	✓	?	n/a	n/a	Unclear	?	✓	?	Unclear
Fann et al. (2005)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Gelaye et al. (2013)	✓	✓	✓	✓	Low	✓	✓	✗	High
Gjerdingen et al. (2009)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Henkel et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High

Table 4: Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis (Moriarty et al., 2015)

Hyphantis et al. (2011)	✓	✓	?	?	Unclear	✓	✓	✗	High
Inagaki et al. (2013)	✓	✓	✓	?	Unclear	✓	✓	✗	High
Khamsseh et al. (2011)	✓	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. (2008)	✓	✓	?	?	Unclear	?	✓	✗	High
Lotrakul et al. (2008)	✓	✓	✓	✓	Low	?	✓	✗	High
Persoons et al. (2003)	✓	✓	?	?	Unclear	✓	✓	✓	Low
Picardi et al. (2005)	✓	✓	✓	?	Unclear	✓	✓	✗	High
Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Thombs et al. (2008)	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
Thompson et al. (2011)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Turner et al. (2012)	✓	?	n/a	n/a	Unclear	?	✓	✗	High
Van Steenberghe-Wijnenburg (2010)	✓	✗	?	?	High	✓	✓	✗	High
Zuithoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

✓ = criterion met; ✗ = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

Study	Patient selection: Consecutive or random sample	Patient selection: Avoid case-control / avoid artificially inflated base rate	Patient selection: Avoided inappropriate exclusions	Patient selection: Overall risk of bias	Index test: PHQ-9 interpreted blind to reference test	Index test: Was a threshold pre-specified?	Index test: If translated, appropriate translation	Index test: If translated, psychometric properties reported	Index test: Overall risk of bias
13. Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	✓	✓	Unclear
16. Kroenke et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
22. Navinés et al. (2012)	✓	✓	✓	Low	✓	✓	✓	?	Unclear
29. Thekkumpurath et al. (2010)	✗	✗	✓	High	✓	✓	n/a	n/a	Low
33. Williams et al. (2005)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
1. Adewuya et al. (2006)	✓	✓	✗	Unclear	✓	✓	n/a	n/a	Low
2. Arroll et al. (2010)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
3. Azah et al. (2005)	✓	✗	?	High	✓	✓	✓	✓	Low
4. Chagas et al. (2013)	✓	✓	✓	Low	✓	✓	✓	✓	Low
6. de Lima Osorio et al. (2009)	✓	✗	✓	High	?	✗	n/a	n/a	High
7. Elderon et al.	✓	✓	✓	Low	✓	✓	n/a	n/a	Low

(2011)

8. Fann et al. (2005)	✓	✗	✗	High	✓	✓	n/a	n/a	Low
9. Fine et al. (2013)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
10. Gelaye et al. (2013)	?	✗	?	High	✓	✗	✓	?	High
11. Gilbody et al. (2007)	?	✓	?	Unclear	✓	✓	n/a	n/a	Low
12. Gjerdingen et al. (2009)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
14. Hyphantis et al. (2011)	✓	✗	✓	High	✓	✓	?	?	Unclear
15. Khamseh et al. (2011)	✓	✓	?	Unclear	✓	✓	✓	?	Unclear
19. Liu et al. (2011)	✓	✓	?	Unclear	✓	✗	✓	?	High
20. Lotrakul et al. (2008)	✗	✓	?	Unclear	✓	✓	✓	?	Unclear
23. Patel et al. (2008)	✓	✓	✓	Low	✓	✓	?	?	Unclear
24. Phelan et al. (2010)	✗	✓	✓	High	✓	✗	n/a	n/a	High
25. Rooney et al. (2013)	✓	✓	✓	Low	?	✗	n/a	n/a	High
26. Sherina et al. (2012)	✓	✓	✗	High	✓	✓	✓	✓	Low
27. Sidebottom	✓	✓	✓	Low	✓	✓	n/a	n/a	Low

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

et al. (2012)									
28. Stafford et al. (2007)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
30. Thombs et al. (2008)	✗	✓	?	High	✓	?	n/a	n/a	Unclear
32. Watnick et al. (2005)	?	✗	✓	High	✓	✓	n/a	n/a	Low
34. Wittkampf et al. (2009)	✓	✓	✓	Low	✓	?	n/a	n/a	Unclear
35. Zhang et al. (2013)	✓	✓	?	Unclear	?	✓	?	?	Unclear
36. Zuithoff et al. (2010)	✓	✓	✓	Low	✓	✓	✓	?	Unclear

Table 4: Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis (Moriarty et al., 2015)

Study	Reference test: Reference test correctly classifies target condition	Reference test: Reference test interpreted blind to PHQ-9	Reference test: If translated, appropriate translation	Reference test: If translated, psychometric properties reported	Reference test: Overall risk of bias	Flow / timing: Interval of two weeks or less	Flow / timing: All participants receive same reference test	Flow / timing: All participants included in analysis?	Flow / timing: Overall risk of bias
13. Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
16. Kroenke et	✓	✓	n/a	n/a	Low	✓	✓	✓	Low

al. (2011)									
22. Navinés et al. (2012)	✓	✓	?	?	Unclear	✓	✓	✓	Low
29. Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
33. Williams et al. (2005)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
<hr/>									
1. Adewuya et al. (2006)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
2. Arroll et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
3. Azah et al. (2005)	✓	✓	✓	✓	Low	✓	✓	X	High
4. Chagas et al. (2013)	✓	✓	?	?	Unclear	✓	✓	X	High
6. de Lima Osorio et al. (2009)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
7. Elderon et al. (2011)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
8. Fann et al. (2005)	✓	?	n/a	n/a	Unclear	✓	✓	X	High
9. Fine et al. (2013)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
10. Gelaye et al. (2013)	✓	✓	✓	✓	Low	✓	✓	X	High

1										
2										
3										
4										
5										
6										
7	11. Gilbody et al. (2007)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
8										
9	12. Gjerdingen et al. (2009)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
10										
11	14. Hyphantis et al. (2011)	✓	✓	?	?	Unclear	✓	✓	✗	High
12										
13	15. Khamseh et al. (2011)	✓	✓	✓	?	Unclear	✓	✓	?	Unclear
14										
15	19. Liu et al. (2011)	✓	✓	✓	✓	Low	✓	✓	?	Unclear
16										
17	20. Lotrakul et al. (2008)	✓	✓	✓	✓	Low	?	✓	✗	High
18										
19	23. Patel et al. (2008)	✓	✓	✓	?	Unclear	?	✓	✗	High
20										
21	24. Phelan et al. (2010)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
22										
23	25. Rooney et al. (2013)	✓	?	n/a	n/a	Unclear	?	✓	✗	High
24										
25	26. Sherina et al. (2012)	✓	✓	✓	✓	Low	✓	✓	✓	Low
26										
27	27. Sidebottom et al. (2012)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
28										
29	28. Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
30										
31	30. Thombs et al. (2008)	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
32										
33	32. Watnick et al. (2005)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
34										
35	34. Wittkamp et al. (2009)	✓	✓	n/a	n/a	Low	?	✓	✗	High
36										
37										
38										
39										
40										
41										
42										
43										
44										
45										
46										
47										
48										
49										

35. Zhang et al. (2013)	✓	?	✓	✓	Unclear	✗	✓	✗	High
36. Zuithoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

Table 5. Pooled estimates of diagnostic properties of the PHQ-9 at cut-off point 10 and using algorithm scoring method in the non-independent vs independent studies groups

Settings	No of studies	No of patients	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive likelihood ratio (95% CI)	Pooled negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)	Heterogeneity: I ²
Manea et al, 2014 SR – RA group	7	4,065	0.77 (0.70 – 0.84)	0.94 (0.90 – 0.97)	14.97 (8.39 – 26.71)	0.23 (0.17 - 0.31)	64.40 (34.15 – 121.43)	78.9%
Manea et al, 2014 SR Independent studies	21	9,900	0.48 (0.41 – 0.91)	0.94 (0.91 – 0.95)	8.26 (6.15 – 11.09)	0.54 (0.48 – 0.62)	15.05 (11.03 – 20.52)	68.1%
Moriarty et al., 2015 SR – RA group	5	6,188	0.87 (0.77 – 0.93)	0.87 (0.76 – 0.94)	7.24 (3.74 – 14.03)	0.14 (0.08 - 0.25)	49.31 (25.74 – 94.48)	55.1%
Moriarty et al., 2015 SR Independent studies	26	13,164	0.76 (0.67 – 0.83)	0.88 (0.85 – 0.91)	6.72 (5.06 – 8.92)	0.26 (0.19 - 0.37)	24.96 (14.81 – 42.08)	81.5%

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

For peer review only

Appendix 1

Figure 1: PRISMA flowchart - search and selection of included diagnostic accuracy studies for the systematic review of studies reporting diagnostic accuracy of the PHQ-9 at using the summed items scoring method (Manea et al, 2014)

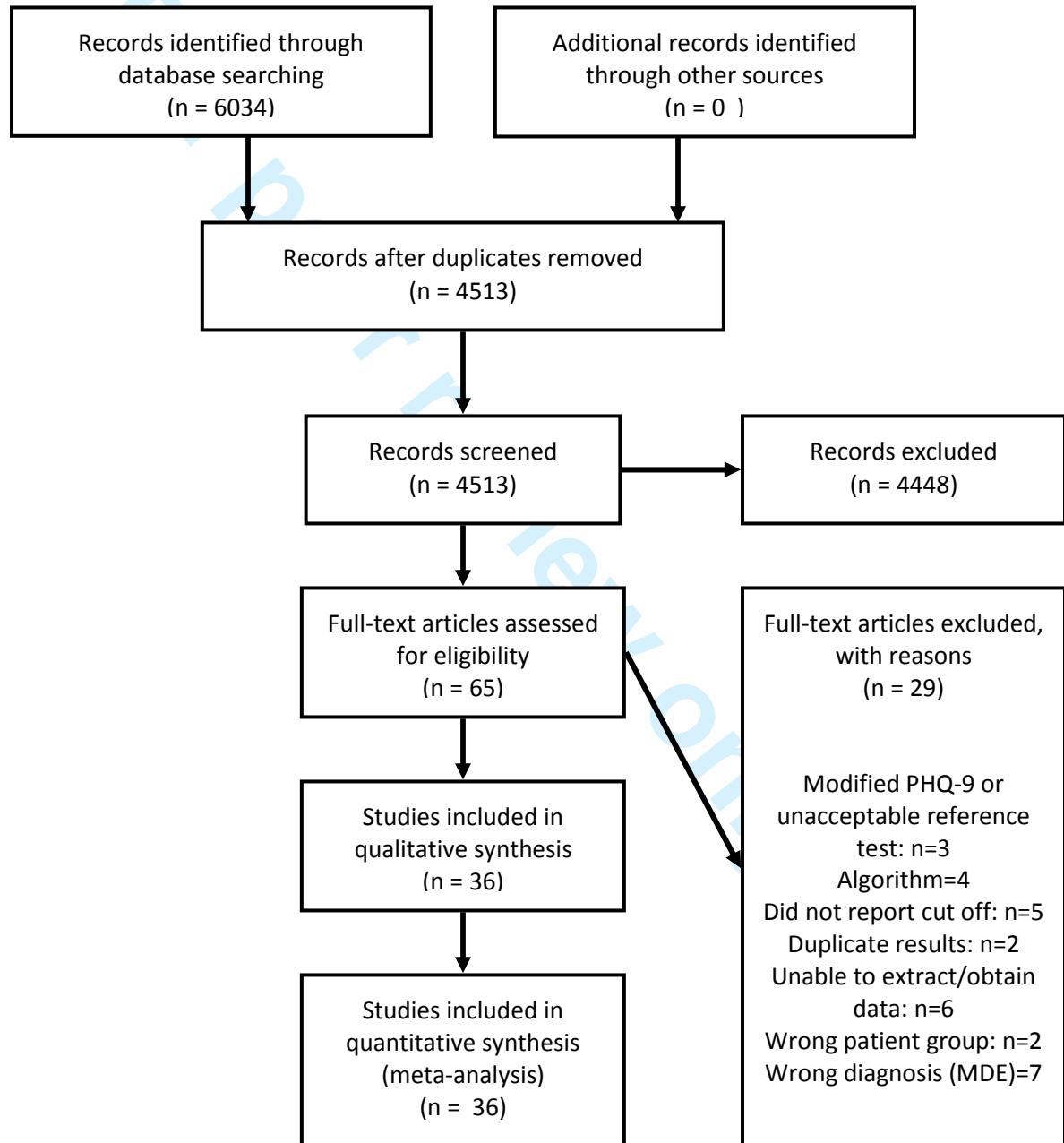
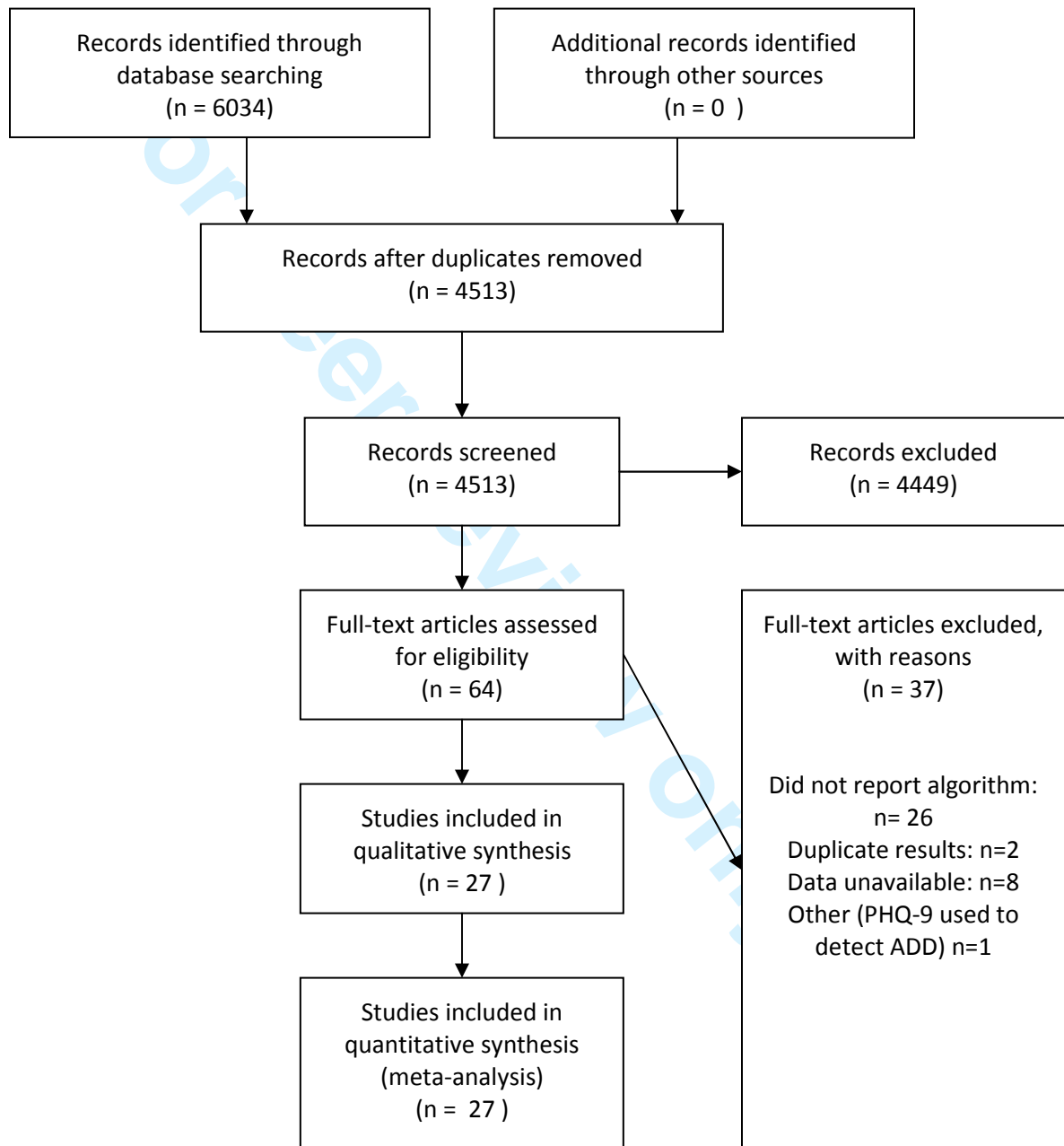


Figure 2: PRISMA flowchart - search and selection of included diagnostic accuracy studies for the systematic review of studies reporting diagnostic accuracy of the PHQ-9 at using the algorithm scoring method (Moriarty et al., 2015)





PRISMA 2009 Checklist

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Available online (see Manea et al., 2015; Moriarty et al., 2015)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5-6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6



PRISMA 2009 Checklist

Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6, 21
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Appendix
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Tables 1 and 2
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Tables 3 and 4
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Table 5
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Tables 3 and 4
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	11 and 17
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	17-21
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	21-22
FUNDING			



PRISMA 2009 Checklist

Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	23
---------	----	--	----

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Page 2 of 2

For peer review only

BMJ Open

Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2016-015247.R1
Article Type:	Research
Date Submitted by the Author:	08-May-2017
Complete List of Authors:	Manea, Laura; University of York, Health Sciences Boehnke, Jan Rasmus; University of York Gilbody, Simon; The University of York, Department of Health Sciences Moriarty, Andrew; University of York, Health Sciences McMillan, Dean; University of York, Department of Health Sciences
Primary Subject Heading:	Mental health
Secondary Subject Heading:	Diagnostics
Keywords:	Depression & mood disorders < PSYCHIATRY, Screening, PHQ-9, diagnostic meta-analysis, allegiance effect

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1 **Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis**

2

3 Laura Manea MMedSci MRCPsych*, Jan R. Boehnke PhD, Simon Gilbody DPhil FRCPsych FRSA, Andrew S. Moriarty MRes, Dean
4 McMillan PhD

5

6 *Corresponding Author

7 Hull York Medical School and Department of Health Sciences, ARRC Building, University of York, YO10 5DD

8 Email: laura.manea@york.ac.uk

9

10

11

12

13

14

15

For peer review only

1
2
3
4
5 166
7 178
9 1810
11 1912
13 2014
15 2116
17 2218
19 23 **Abstract**

20 24 **Objectives** To investigate whether an authorship effect is found that leads to better performance in studies conducted by the original developers
21 25 of the PHQ-9 (allegiant studies).

22 26 **Design** Systematic review with random effects bivariate diagnostic meta-analysis. Search strategies included electronic databases, examination
23 27 of reference lists, and forward citation searches.

24 28 **Inclusion criteria** Included studies provided sufficient data to calculate the diagnostic accuracy of the PHQ-9 against a gold standard diagnosis
25 29 of major depression using the algorithm or the summed item scoring method at cut-off point 10.

26 30 **Data extraction** Descriptive information, methodological quality criteria, and 2×2 contingency tables.

27 31 **Results**

1
2
3
4
5 32 Seven allegiant and twenty independent studies reported the diagnostic performance of the PHQ-9 using the algorithm scoring method. Pooled
6
7 33 diagnostic odds ratio (DOR) for the allegiant group was 64.40, and 15.05 for non-allegiant studies group. The allegiance status was a significant
8
9 34 predictor of DOR variation ($p < 0.0001$).

10
11 35 Five allegiant studies and twenty-six non-allegiant studies reported the performance of the PHQ-9 at recommended cut-off point of 10. Pooled
12
13 36 DOR for the allegiant group was 49.31, and 24.96 for the non-allegiant studies. The allegiance status was a significant predictor of DOR
14
15 37 variation ($P = 0.015$).

16
17 38 Some potential alternative explanations for the observed authorship effect including differences in study characteristics and quality were found,
18
19 39 though it is not clear how some of them account for the observed differences

20
21 40

22 23 41 **Conclusions**

24
25
26 42 Allegiant studies reported better performance of the PHQ-9. Allegiance status was predictive of variation in the DOR. Based on the observed
27
28 43 differences between independent and non-independent studies we were unable to conclude or exclude that allegiance effects are present in
29
30 44 studies examining the diagnostic performance of the PHQ-9. This study highlights the need for future meta-analyses of diagnostic validation
31
32 45 studies of psychological measures to evaluate the impact of researcher allegiance in the primary studies.

33
34 46

35
36 47

37 38 48 **Strengths and limitations of this study**

39
40 49

- 1
2
3
4
5 50 a) An original study—the first meta-analysis of diagnostic validation studies of psychological measures to evaluate the impact of researcher
6 allegiance.
7 51
8 b) Using rigorous methodology—strict inclusion/exclusion and quality assessment criteria.
9 52
10 53 c) We found that the allegiance effect was a significant predictor of the variation of the diagnostic odds ratio in the meta-regression
11 analysis.
12 54
13 55 d) Substantial variability observed in methodological quality of included studies.
14
15 56 e) Based on the observed methodological differences between the independent and non-independent studies we were unable to conclude or
16
17 57 exclude that allegiance effects are present in studies examining the diagnostic performance of the PHQ-9.
18
19 58
20
21 59
22
23 60
24
25 61
26
27 62
28
29 63
30
31 64
32
33 65
34
35 66
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1
2
3
4
5 67 Research on allegiance effects has a long tradition in psychotherapy research. In this context *allegiance* describes the phenomenon that
6
7 68 researchers and clinicians who developed a treatment approach or are for other reasons invested in it tend to find larger effect sizes in favour of
8
9 69 their treatment than for comparison groups. [1] This finding has been extensively replicated [2], [3] and is also robust when the quality of
10
11 70 research is controlled for. Researcher allegiance is subject of on-going debates about the design of efficacy studies as well as implications for
12
13 71 policy. [2], [4], [5] Researcher allegiance is also discussed widely in the literature on experimental as well as evaluation research. [6] Since the
14
15 72 motivational underpinnings of allegiance effects are potentially far more ingrained into human behaviour and decision making than previously
16
17 73 thought (e.g., [7]), they may occur commonly in clinical research in general.

18 74 Although it has been suggested that allegiance effects may play a role in the validation of psychological screening and case-finding tools (e.g.,
19
20 75 O'Shea et al., in press), systematic evaluations of this hypothesis are rare and studies that acknowledge potential allegiance effects in such
21
22 76 studies mainly come from forensic psychology and psychiatry backgrounds. [8]–[11] Diagnostic validation studies are geared at establishing the
23
24 77 sensitivity and specificity of a screening or case finding tool, which is used in practice to differentiate cases from non-cases or to decide about
25
26 78 whether further assessment or treatment is indicated or will be offered. An allegiance effect in such studies would be seen in systematically
27
28 79 higher sensitivities or specificities if the original author(s) is (are) part of the team of such a study. Such a bias would have a deleterious affect on
29
30 80 practice through promising over-optimistic accuracy of the screening or case finding tool or in evaluating the cost-effectiveness of the measure
31
32 81 in a screening or case-finding context.

33 82 The depression module of the Patient Health Questionnaire (PHQ-9) is a widely used depression-screening instrument in non-psychiatric
34
35 83 settings. The PHQ-9 was developed by a team of researchers, with its development underwritten by an educational grant from Pfizer US
36
37 84 Pharmaceuticals. [12] The PHQ-9 can be scored using different methods, including an algorithm based on DSM-IV criteria and a cut-off based
38
39 85 on summed-item scores. The psychometric properties of these two approaches have been summarised in two recently published meta-analyses.
40
41 86 [13], [14] The goal of the current review is to investigate, based on an established database of PHQ-9 diagnostic validation studies [13], [14],
42
43
44
45
46
47
48
49

1
2
3
4
5 87 whether an allegiance effect is found that leads to an increased sensitivity and specificity in studies that were conducted by researchers closely
6
7 88 connected to the original developers of the instrument.

8
9 89 METHODS

10
11 90 *Study Selection*

12
13
14 91 Similar search strategies were used in both systematic reviews. (For full details please see Manea et al. (2014) and Moriarty et al. (2015)).
15
16 92 Embase, MEDLine and PSYCHInfo were searched from 1999 (when the PHQ-9 was first developed) to August 2013 [13] and September 2013
17
18 93 [14] respectively, using the terms “PHQ-9”, “PHQ”, “PHQ\$” and “patient health questionnaire”. The search strategy is presented in Appendix 2.
19
20 94 The reference lists of studies fitting the inclusion criteria were manually searched and a reverse citation search in Web of Science was
21
22 95 performed. Authors of unpublished studies were contacted and conference abstracts were reviewed in an attempt to minimise publication bias.

23
24 96 The following inclusion-exclusion criteria were used:

25
26 97 *Population:* Adult population. *Instrument:* Studies that used the PHQ-9. *Comparison (reference standard):* The accuracy of the PHQ-9 had to be
27
28 98 assessed against a recognised gold-standard instrument for the diagnosis of either Diagnostic and Statistical Manual (DSM) or International
29
30 99 Classification of Disease (ICD) criteria for major depression. Studies were included if the diagnoses were made using a standardised diagnostic
31
32 100 structured interview schedule (e.g. Mini International Neuropsychiatric Interview (MINI), Structured Clinical Interview for DSM Disorders
33
34 101 (SCID)). Unguided clinician diagnoses with no reference to a standard structured diagnostic schedule or comparisons of the PHQ-9 with other
35
36 102 self-report measures were excluded. Studies were also excluded if the target diagnosis was not major depressive disorder (MDD, e.g. any
37
38 103 depressive disorder). *Outcome:* Studies had to report sufficient information to calculate a 2*2 contingency table for the algorithm or the
39
40 104 recommended cut-off point 10. *Study design:* Any design. *Additional criterion:* We avoided double counting of evidence by ensuring that only

1
2
3
4
5 105 one study of those that reported overlapping datasets in different journals were included in the meta-analysis. Citations with overlapping samples
6 106 were examined to establish whether they contained information relevant to the research question that was not contained in the included report.

7
8
9 107 *Quality assessment*

10
11 108 Quality assessment was performed using the QUADAS-2 tool, a tool for evaluating the risk of bias and applicability of primary diagnostic
12 109 accuracy studies when conducting diagnostic systematic reviews. [15] It covers the areas of: patient selection, index test, reference standard and
13 110 flow and timing. [16] This tool was adapted for the two reviews and quality assessments were carried out by two independent reviewers for all
14 111 studies included in the reviews.

15
16
17
18
19 112 *Data synthesis and statistical analysis*

20
21 113 We constructed 2x2 tables for cut-off point 10 [14] and the algorithm scoring method [13] Pooled estimates of sensitivity, specificity,
22 114 positive/negative likelihood ratios, and diagnostic odds ratios were calculated using random effects bivariate meta-analysis. [17] Heterogeneity
23 115 was assessed using I^2 for the diagnostic odds ratio, an estimate of the proportion of study variability that is due to between-study variability
24 116 rather than sampling error. We considered values of $\geq 50\%$ to indicate substantial heterogeneity.[18] Summary Receiver Operator Characteristic
25 117 curves (sROC) were constructed using the bivariate model to produce a 95% confidence ellipse within ROC space. [19] Each data point in the
26 118 summary ROC space represents a separate study, unlike a traditional ROC plot, which explores the effect varying thresholds on sensitivity and
27 119 specificity in a single study.

28
29
30 120 We undertook a meta-regression analysis of logit diagnostic odds ratio using research allegiance as covariate in the meta-regression model. [20],
31 121 [21] Analyses were conducted using STATA version 12, with the metan, metandi and metareg user-written commands.

32
33
34
35
36
37
38 122 *Allegiance Rating*

1
2
3
4
5 123 We rated authorship on a paper if any of the developers of the PHQ-9 - Kurt Kroenke, MD, Robert L Spitzer, MD, and Janet B W Williams – as
6
7 124 an indicator of potential allegiance. We also rated as evidence of allegiance as acknowledged collaborations with the developers of the PHQ-9,
8
9 125 even if they were not listed as co-authors or if the authors acknowledged funding from Pfizer to conduct the study.

10
11 126

12 127 RESULTS

13
14
15
16 128

17 18 129 **Overview of included studies**

19
20
21 130 31 studies reported the diagnostic properties of the PHQ-9 at cut-off point 10 or above and were included in this analysis. [14] 27 studies were
22
23 131 included in the algorithm review [13]. The study selection flowcharts can be found in Appendix 1 (figures 1 and 2). The characteristics of these
24
25 132 studies are reported in tables 1 and 2 and the results of the methodological assessment are presented in tables 3 and 4.

26 27 133 **Algorithm scoring method**

28
29 134

30 31 135 Descriptive characteristics

32
33
34 136 The descriptive characteristics of the included studies are presented in table 1. Seven individual studies that reported the diagnostic performance
35
36 137 of the PHQ-9 using the algorithm scoring method were co-authored by the original developers of the PHQ-9 [22]–[26], specifically
37
38 138 acknowledged one of the developers and support by an educational grant from Pfizer US [27], or were co-authored by the first author of a

1
2
3
4
5 139 previous study that had also been co-authored by one of the developers [28]. Twenty non-allegiant studies reported the diagnostic properties of
6
7 140 the PHQ-9 using the algorithm scoring method.

8
9 141

10
11 142 Three (43%, 3/7) of the allegiant studies were conducted exclusively in hospital settings [22], [26], [28]. The remaining four studies (67%, 4/7)
12
13 143 were conducted in different settings or non-exclusively hospital settings: one in primary care [25] and three in mixed settings: psycho-somatic
14
15 144 walk in clinics and family practices [23]¹, outpatient clinics and family practices [24] and primary care and hospital settings [27]. In the non-
16
17 145 allegiant group, thirteen (65%, 13/20) studies were conducted in hospital settings [29]–[41]. Of the remaining seven studies, six were conducted
18
19 146 in primary care settings [42]–[47] and one in a community sample [48].

20
21 147 In both groups (non-allegiant and allegiant studies), the majority of studies validated a translated version of the PHQ-9. Two of the studies
22
23 148 authored by developers (28%, 2/7) [25], [26], and eight (40%, 8/20) allegiant studies [29], [30], [37]–[40], [42], [48] were conducted in English.

24
25 149 The mean prevalence of major depressive disorder in the group of allegiant studies was 13.4 % (range 6.1% - 29.2%); in the non-allegiant group
26
27 150 it was 15.5% (range 3.9% - 32.4%). The mean age of patients in the PHQ-9 developers group was 45.7; all but one study had a mean age in the
28
29 151 range of 40 to 50 years. In the non-allegiant group the mean age was 54.6 (range 29.3 – 75.0), with almost half (8) of the studies reporting a
30
31 152 mean age of over 60. The percentage of females in the PHQ-9 developers was 56.8% (range 28.6% - 67.8%) and in the non-allegiant group was
32
33 153 59.1 (18% -100%).

34 154
35
36
37
38

39
40
41
42
43
44
45
46
47
48
49
¹ This study provided separate estimates for the two settings in which it was conducted; therefore separate psychometric estimates were generated for each sample for both algorithm scoring method and summed items scoring method at cut-off point 10 (see below).

1
2
3
4
5 155 All allegiant studies used a self-reported PHQ-9, whereas in 7 non-allegiant studies (30%, 6/20) the PHQ-9 was administered by a researcher
6 156 [30]–[33], [43], [48]. Apart from Muramatsu et al. (2007) all allegiant studies used the SCID as a gold standard; the non-allegiant studies used a
7
8 157 wider range of gold standards including SCAN, CIDI, MINI, and C-DIS, though the SCID was also frequently used by the independent studies
9 158 as well (45%, 9/20 studies).

10
11
12 159 Four out of the seven allegiant studies (57%) did not include a conflict of interests statement [22], [23], [25], [27]. Also, four (57%) of the
13 160 allegiant studies acknowledged funding from Pfizer [23]–[25], [27]. Only one study [27] acknowledged the collaboration with one of the
14 161 developers of the PHQ-9.

15
16
17 162 Of the non-allegiant studies, twelve (60%) did not include a conflict of interests statement [29]–[32], [35]–[37], [39], [45], [46], [48], [49]. It
18 163 appears that newer studies were more likely to include a conflict of interest statement, which may reflect a recent change in reporting. Funding
19 164 was acknowledged by most studies (18/20) and most received funding from academic or/and health research institutions. Two studies received
20 165 funding from pharmaceutical companies – Lundbeck [43] and Pfizer [35] and one study acknowledged that Pfizer Italia provided the Italian
21 166 version of PHQ-9 and gave the authors permission to use it [36].

22
23
24
25
26
27 167 Diagnostic test accuracy

28
29
30 168 Pooled sensitivity and specificity was calculated separately for the non-allegiant and allegiant studies. Pooled sensitivity for the allegiant studies
31 169 of the PHQ-9 was 0.77 (95% CI = 0.70 – 0.84), pooled specificity was 0.94 (95% CI = 0.90 – 0.97), and the pooled diagnostic odds ratio was
32 170 64.40 (95% CI = 34.15 – 121.43). Heterogeneity was high ($I^2 = 78.9\%$). Figure 1 represents the summary ROCs for this set of studies.

33
34
35
36 171

37
38 172
39
40
41
42
43
44
45

1
2
3
4
5 173 -----
6
7
8 174 Figure 1. PHQ-9 algorithm scoring method summary ROC plot for the diagnosis of major depressive disorder in allegiant studies (Panel A) and
9 175 non-allegiant studies (Panel B). Pooled sensitivity and specificity estimates using a bi-variate meta-analysis (*HSROC* hierarchical receiver-
10 176 operating characteristic).
11
12
13 177 -----
14

15
16 178
17
18 179
19
20 180 Pooled sensitivity for the non-allegiant studies was lower compared to the developer authored studies group at 0.48 (95% CI = 0.41 – 0.91),
21 181 pooled specificity was the same at 0.94 (95% CI = 0.91 – 0.95). The pooled diagnostic odds ratio was approximately four times lower at 15.05
22 182 (95% CI = 11.03 – 20.52) (see figure 1). Heterogeneity was substantial at $I^2 = 68.1\%$.
23
24
25

26 183

27
28 184

29
30
31 185 The meta-regression analysis for algorithm studies with non-allegiant status as the predictor of the diagnostic odds ratio showed that non-
32 186 allegiant status was a significant predictor of the diagnostic odds ratio ($p < 0.0001$) and explained a substantial amount of the observed
33 187 heterogeneity (51.5%).
34
35

36
37 188

38
39 189 Quality assessment
40
41
42
43
44
45

1
2
3
4
5 190 The results of the quality assessment using QUADAS-2 are given in table 3 for the studies reporting on the diagnostic performance of the
6
7 191 algorithm scoring method. In the patient selection domain, more non-allegiant studies (65%, 13/20) than allegiant (29%, 2/7) met the criterion
8
9 192 for consecutive referrals. There were no marked differences on the other two criteria in this domain (avoid case-control design, avoid
10 193 inappropriate exclusions). In the index test domain, the proportion of studies reporting that the PHQ-9 was conducted blind to the reference test
11 194 was comparable between the two groups. There were differences in this domain for those studies using a translated version of the test. All non-
12 195 English allegiant studies (5/5) used an appropriately translated version of the PHQ-9; whereas just over a half of the non-allegiant studies
13 196 reported this (55%, 6/11). However, the majority of both sets of studies did not report details of psychometric properties of the translated
14 197 version. For the reference test domain, nearly all studies in both groups were rated as using a reference test that would correctly classify the
15 198 condition. While most allegiant studies reported that the reference test was interpreted blind to the PHQ-9 score (86%, 6/7), this was reported in
16 199 only 60% (12/20) of the non-allegiant studies.

17
18
19
20
21
22 200 The two sets of studies that used translated versions of the reference test were broadly comparable. There was a slight indication that the
23 201 allegiant studies were more likely to use an appropriately translated version of the reference test and report data on the psychometric properties
24 202 of the translated version, though the numbers for the translated comparison are very low. There were, however, some more notable differences
25 203 on the flow and timing domain. Most allegiant studies ensured that the time between the index and reference test was under two weeks (86%,
26 204 6/7) in comparison to 70% (14/20) of the non-allegiant studies. More allegiant studies met the criterion for 'all participants included in the
27 205 analysis' (57%, 4/7) than non-allegiant studies (25%).
28
29
30
31
32

33 206

34
35
36 207 **Summed items scoring method (cut-off point 10 or above)**

37
38 208

39
40
41
42
43
44
45

1
2
3
4
5 209 Descriptive characteristics
6

7
8 210 Table 2 presents the sample characteristics of the thirty-one PHQ-9 validation studies that reported the psychometric properties of the PHQ-9 at
9 211 cut-off point 10 or above. Five of these studies were co-authored by the original developers of the instrument or acknowledged collaboration
10 212 [12], [23], [26], [50] or were co-authored by the first author of a previous study that had also been co-authored by one of the developers [28].
11 213 Twenty-six studies were conducted by independent researchers.
12
13
14

15 214

16
17 215 Three (60%, 3/5) allegiant studies [26], [28], [50] and eleven non-allegiant studies (42%, 11/26) [30]–[32], [34], [37], [38], [51]–[55] were
18 216 conducted in hospital settings.
19
20

21 217

22
23
24 218 Three (60%, 3/5) allegiant studies [12], [26], [50] and thirteen non-allegiant studies (13/26) [30], [37], [38], [42], [48], [52]–[54], [56]–[60], were
25 219 conducted in English.
26
27

28 220

29
30
31 221 The mean prevalence of major depressive disorder in the allegiant group was 13.2% (range 6.1% - 33.5%) and in the non-allegiant group was
32 222 16.1% (range 2.5% - 43.2%). The mean age of patients in the allegiant group studies was 48.1 (range 41.9 -61.0) and in the 26 non-allegiant
33 223 studies that reported these data was 49.1 (range 23.0 – 78.0). The percentage of females in the allegiant studies that reported these data [12],
34 224 [23], [26], [28] was 56.3% (range 28.6% – 67.8%) and in the non-allegiant group was 64.9 % (range 12% -100%).
35
36
37

38 225
39
40
41
42
43
44
45

1
2
3
4
5 226 Three allegiant studies used the self-reported mode of administration and two of them did not specify how the PHQ-9 was administered. In 9
6
7 227 non-allegiant studies (34%, 9/26) the PHQ-9 was administered by the researcher [30]–[32], [48], [57], [59]–[62]. All allegiant studies used SCID
8
9 228 as a gold standard; the non-allegiant studies used a wider range of gold standards including SCAN, CIDI, MINI, CIS-R, C-DIS, though the SCID
10 229 was used in half of the studies (50%, 13/26 studies).

11
12 230 Three allegiant studies (60%) did not include a conflict of interests statement [12], [23], [50]. Two of these studies [12], [23] acknowledged
13 231 funding from Pfizer. None of the allegiant studies acknowledged collaboration or authorship of one of the developers of the PHQ-9.

14
15
16
17 232 Of the non-allegiant studies, thirteen (42%) did not include a conflict of interests statement [30]–[32], [37], [42], [46], [48], [54], [56], [61],
18 233 [63]–[65]. Similar to the algorithm studies, the newer studies were more likely to include a conflict of interest statement. Funding was
19 234 acknowledged by most studies (27/31) and most received funding from academic or/and health research institutions. One study [58]
20 235 acknowledged that the last author involved in the development of one of the instruments (CORE-OM), ‘but does not gain financially from its
21 236 use’. One study [52] acknowledged funding from industry, AHA Pharmaceuticals Roundtable, but stated that ‘the funding organisations had no
22 237 role in the design or conduct of the study, collection, management, analysis or interpretation of data; or preparation, review or approval of the
23 238 manuscript. Fine et al., 2013 disclosed that the last author had financial and consulting interests (Pfizer was not cited as one of them).
24
25
26
27
28
29

30 240 *Diagnostic test accuracy*

31
32 241 Pooled sensitivity of allegiant studies was 0.87 (95% CI = 0.77 – 0.93), pooled specificity was 0.87 (95% CI = 0.76 – 0.94), and the pooled
33 242 diagnostic odds ratio was 49.31 (95% CI = 25.74 – 94.48) – see table 5. Heterogeneity was moderate ($I^2 = 55.1\%$). Figure 2 represents the
34 243 summary ROCs for this group.
35
36
37
38
39
40
41
42
43
44
45

1
2
3
4
5 246

6 -----
7 247 Figure 2. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot for diagnosis of major depressive disorder in allegiant
8 248 studies (panel A) and non-allegiant studies (panel B). Pooled sensitivity and specificity using a bi-variate meta-analysis (*HSROC* hierarchical
9 249 receiver-operating characteristic).
10 -----

11 250

12 -----
13
14 251 Pooled sensitivity of non-allegiant studies was 0.76 (95% CI, 0.67 – 0.83), pooled specificity was 0.88 95% CI (0.85 – 0.91), and the pooled
15 252 diagnostic odds ratio was 24.96 (95% CI 14.81 – 42.08), approximately half that of the allegiant studies (table 2). Heterogeneity was high at $I^2 =$
16
17 253 81.5 %. Figure 2 represents the summary ROCs for this group.
18
19

20 254
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

1
2
3
4
5 255 The meta-regression for the studies using a cut-off point of 10 or above with allegiance status of the predictor showed that allegiance status was
6 256 a significant predictor of the diagnostic odds ratio ($P = 0.015$) and explained 19.0% of observed heterogeneity.

7
8
9 257

10
11 258 *Quality assessment*

12
13
14 259 The results of the quality assessment using the QUADAS-2 are given in table 4. For the patient selection domain, the two groups of studies were
15 260 broadly comparable on two items (consecutive or random sample, avoid case-control design). However, all allegiant studies were rated as
16 261 avoiding inappropriate exclusions (5/5) in contrast to 58% (15/26) of the non-allegiant studies.

17
18
19
20 262

21
22 263 On the index test domain, there were a number of differences between the two groups of studies. More of the non-allegiant studies (81%, 21/26)
23 264 reported that the PHQ-9 was interpreted blind to the reference test compared to 60% (3/5) of the allegiant studies. All (5/5) allegiant studies were
24 265 rated as pre-specifying the threshold on the PHQ-9 compared to 73% (19/26) of the non-allegiant studies. The two sets of studies were broadly
25 266 comparable in terms of two items from the reference test domain (correctly classify target condition, reference test interpreted blind). Only one
26 267 allegiant study used a translated version of the index test or reference test, so it is not possible to comment on differences between the two sets of
27 268 studies in terms of these items from the index or reference test domains. For the flow and timing domain, the two groups of studies were broadly
28 269 comparable for two of the criteria (interval of two weeks or less, all participants receive same reference test). However, fewer than half of the
29 270 non-allegiant studies met the criterion for 'all participants included in the analysis' (42%, 11/26); whereas all allegiant studies met this criterion.

30
31
32
33
34
35
36
37 271

38
39 272 **Discussion**

1
2
3
4
5 273 This is to our knowledge the first systematic examination of a possible ‘allegiance’ or authorship effect in the validation of screening or case
6
7 274 finding psychological instrument for a common mental health disorder. We reviewed diagnostic validation studies of the PHQ-9, a widely used
8
9 275 depression screening-instrument. We found that allegiant studies reported higher sensitivity paired with similar specificity compared to non-
10
11 276 allegiant studies. When entered as a covariate in meta-regression analyses, allegiance status was predictive of variation in the DOR for both the
12
13 277 algorithm scoring method and the summed-item scoring method at a cut-off point of 10 or above.
14
15 278

16
17 279 Previous research has proposed several possible explanations for the allegiance effect [9]–[11]. One possibility is the advertent bias that may
18
19 280 serve to inflate the performance of a test when evaluated by those who have developed it. However, before concluding that the differences are
20
21 281 due to this, it is important to explore and rule out alternative explanations. First, it is possible that any observed differences are a result of
22
23 282 differences in study characteristics of the two sets of studies (e.g., setting, clinical population). Secondly, differences in the methodological
24
25 283 quality of the studies may also account for any differences. These possibilities are examined below.
26
27 284

28 285 Difference in study characteristics as potential alternative explanations

29
30
31 286 The two sets of studies were broadly comparable in terms of gender and the prevalence of depression, so these variables are unlikely to offer an
32
33 287 explanation for the differences. While there were some indications from both sets of comparisons that the PHQ-9 may have been researcher-
34
35 288 administered more often in the independent studies, it is not immediately clear how this would lead to lowered diagnostic performance.
36
37 289

1
2
3
4
5 290 The diagnostic meta-analyses of the PHQ-9 [13], [14] have shown that the sensitivity and DOR of the PHQ-9 tends to be lower in hospital
6
7 291 settings for both algorithm and summed-item scoring methods. Whilst the fact that proportionally more non-allegiant algorithm studies were
8
9 292 conducted in secondary care could explain the lower sensitivity and DOR values in the algorithm studies, in the studies that reported the cut-off
10
11 293 point of or above this would not be the case as proportionally more allegiant studies were conducted in hospital settings.

12
13 294 Similarly, differences in the proportions of studies using translated versions of the PHQ-9 are also unlikely to offer an obvious explanation of the
14
15 295 difference in diagnostic performance, because in the algorithm set of studies more of the allegiant studies used a translated version of the test,
16
17 296 but the proportions were in the opposite direction for the studies using a cut off of 10 or above. We tested this by carrying out a sensitivity
18
19 297 analysis restricting the sample to English studies and studies with adequate translation. The allegiance effect was still predictive of DOR
20
21 298 variation between allegiance and non-allegiance studies variation in both algorithm ($p = 0.00$) and summed item scoring at cut-off point of 10
22
23 299 meta-analyses ($p = 0.02$).

24
25 300 A similar conclusion is also likely to apply to the age of the samples. There were more older adults studies in the non-allegiant than allegiant
26
27 301 studies in the algorithm comparison. Depression could be more difficult to identify in older adults due to physical co-morbidities that may
28
29 302 present with similar symptomatology to depression and could account for the lower diagnostic performance in the non-allegiant studies.
30
31 303 However, the non-allegiant samples in the studies that reported the psychometric properties at cut-off point 10 or above had younger samples
32
33 304 than the allegiant studies, so this would not support this interpretation.

34
35 306 The SCID was used as the gold standard in nearly all allegiant studies. The fact that some non-allegiant studies used other gold standards could
36
37 307 potentially explain the poorer psychometric properties of the PHQ-9 in these studies. The SCID is often regarded as the most valid of the
38
39 308 available semi-structured interviews used in depression diagnostic validity studies as the reference standard. If we assume that this is the case
40
41 309 and, furthermore, that the PHQ-9 is an accurate method of screening for depression, then the PHQ-9 may be more likely to agree with the SCID
42
43
44
45

1
2
3
4
5 310 than other reference standards. However, when we carried out a sensitivity analysis restricting the sample to SCID only studies the allegiance
6
7 311 effect was still predictive of DOR variation between allegiance and non-allegiance studies variation in both algorithm ($p = 0.01$) and summed
8
9 312 item scoring at cut-off point of 10 reviews ($p = 0.02$).

10 313

11 314

12 13 14 15 16 315 Differences in methodological quality as potential alternative explanations

17
18 316 The quality of the studies was evaluated using the QUADAS-2. Although there were several potential methodological differences between the
19
20 317 two groups of studies from the algorithm papers, not all of these offer obvious explanations of the observed differences and some are unlikely as
21
22 318 explanations. For example, more allegiant studies ensured that the reference test was interpreted blind to the index test. This is unlikely to
23
24 319 account for the observed differences, because a lack of blinding is typically associated with artificially increased diagnostic performance, which
25
26 320 is in the opposite direction to the pattern of results observed here. The impact of some other differences is less clear-cut. For example, a higher
27
28 321 number of the non-allegiant studies met the criterion for consecutive referrals. For this to provide an explanation of the of the observed
29
30 322 differences, the non-consecutive nature of the referrals in the studies by those who had developed the PHQ-9 would need to have led to the over-
31
32 323 inclusion of true positives or under-inclusion of false negatives given that these studies tended to report higher sensitivity relative to the non-
33
34 324 allegiant studies (and vice versa for the independent studies). It is not immediately obvious how this would occur. The allegiant studies were
35
36 325 more likely to have met the criterion of 'included all participants in the analysis'. It is possible that the greater loss of participants from the non-
37
38 326 allegiant studies may have artificially reduced the observed diagnostic accuracy, though, again, it is not immediately obvious how this would
39
40 327 have affected the true positive and false negative rates. Although there is not an obvious explanation of how these differences in methodological
41
42 328 quality could account for the observed differences in diagnostic performance, it is important to recognise that they cannot on that basis be ruled
43
44 329 out.

1
2
3
4
5 330

6
7
8 331 There are, however, two differences in methodological quality among the algorithm studies that are clearer potential alternative explanations.
9 332 The higher rate of appropriate translations among the allegiant studies is potentially important, because lower diagnostic estimates may be
10 333 expected from studies that have poorly translated versions of the index test. In the flow and timing domain, more allegiant studies ensured that
11 334 there was a less than two-week interval between the index and reference test. This is consistent with lower diagnostic performance in the non-
12 335 allegiant studies: as the interval increases it is likely that depression status may change and this would lead to lower levels of agreement between
13 336 the index test and the reference test.

14
15
16
17
18 337

19
20
21 338 There were also differences on some quality assessment items between the two sets of studies in the summed item scoring method comparison.
22 339 The threshold was reported as pre-specified in all allegiant studies in contrast to approximately three quarters of the non-allegiant studies. On the
23 340 face of it, this is unlikely to explain the observed differences, because the use of a pre-specified cut-off point is likely to be associated with lower
24 341 not higher diagnostic test performance. One possibility, however, is that studies that performed poorly at this cut-off point were less likely to be
25 342 reported by those who had developed the measure. As discussed in more detail in the limitations section, we were unable to explore this
26 343 possibility through the use of formal tests for publication bias.

27
28
29
30
31
32 344

33
34 345 All allegiant studies avoided inappropriate exclusions compared to approximately half of the non-allegiant studies. While this is a potential
35 346 alternative explanation of the differences it is not immediately obvious how this would explain the differences in diagnostic performance
36 347 between the two sets of studies. Fewer than half of the non-allegiant studies met the criterion for 'all participants included in the analysis', in
37 348 contrast to all of the allegiant studies met this criterion, but again this difference should usually work against the inclusive studies, not those

1
2
3
4
5 349 excluding cases. More of the non-allegiant studies reported that the PHQ-9 was interpreted blind to the reference test. This does offer a potential
6
7 350 explanation, because the absence of blinding may artificially inflate diagnostic accuracy.
8

9 351

10
11 352 **Limitations**

12
13
14 353 The results of this review need to be viewed in the light of the limitations of the primary studies that contributed to the review and the review
15
16 354 itself. An important consideration is to establish whether any observed differences between the diagnostic performance of the non-allegiant and
17
18 355 allegiant studies are better accounted for by study characteristic or methodological differences. Caution, however, is needed in interpreting any
19
20 356 differences, because of the small number of allegiant studies in both the algorithm and cut-off 10 or above comparisons. The small number of
21
22 357 allegiant studies also meant that we were also unable to explore the potential role of publication bias in the non-allegiant and allegiant studies. At
23
24 358 least 10 studies are required to use standard methods of examining publication bias, but the number of allegiant studies in both the algorithm and
25
26 359 cut-off 10 or above comparisons were fewer than this.
27

28
29 360

30 361

31
32 362 **Conclusions and implications for further research.**

33
34 363 The aims of the review was to investigate whether an allegiance effect is found that leads to an increased diagnostic performance in diagnostic
35
36 364 validation studies that were conducted by teams connected to the original developers of the PHQ-9. Our analyses showed that diagnostic studies
37
38 365 conducted by independent/non-allegiant researchers had lower sensitivity paired with similar specificity compared to studies that were classified
39
40 366 as allegiant. This conclusion held for both the algorithm and cut-off 10 or above studies. We explored a range of possible alternative
41
42
43
44
45

1
2
3
4
5 367 explanations for the observed allegiance effect including both differences in study characteristics and study quality. A number of potential
6
7 368 differences were found, though for some of these it is not clear how they would necessarily account for the observed differences. However, there
8
9 369 were a number of differences that offered potential alternative explanations unconnected to allegiance effects. In the algorithm studies, the
10 370 studies rated as allegiant were also more likely to use an appropriate translation of the PHQ-9 and were also more likely to ensure that the index
11 371 and reference test were conducted within two weeks of each other, both of which may be associated with an improvement in observed diagnostic
12 372 performance of an instrument. The majority of studies in both meta-analyses did not provide clear statements about potential conflict of interest
13 373 and/or funding, however the newer studies were more likely to provide such statements, which may reflect increasing transparency in this area of
14 374 research.

18
19 375

20
21 376 We cannot, therefore, conclude that allegiance effects are present in studies examining the diagnostic performance of the PHQ-9; but nor can we
22 377 rule them out. Conflicts of interest are an important area of investigation in medical and behavioural research, particularly due to concerns about
23 378 trial results being influenced by industry sponsorship. Future diagnostic validity in this area should as a matter of routine present clear statements
24 379 about potential conflicts of interest and funding, particularly relating to the development of the instrument under evaluation. Future meta-
25 380 analyses of diagnostic validation studies of psychological measures should routinely evaluate the impact of researcher allegiance in the primary
26 381 studies examined in the meta-analysis.

27 382

28 383

29 384 **Contributors** LM led on all stages of the review and is the guarantor. We used an established database of diagnostic validation studies of the
30 385 PHQ-9 [13], [14] SG provided expert advice on methodology and approaches to assessment of the evidence base. AM carried out the literature

1
2
3
4
5 386 searches, screened the studies, extracted data and assessed the quality of the included studies for one of the systematic reviews (Moriarty et al.,
6
7 387 2015) . LM carried out the literature searches, screened the studies, extracted data and assessed the quality of the included studies for the other
8
9 388 systematic review (Manea et al., 2015), analysed the data for both systematic reviews and drafted the report. JB was involved in the development
10 389 of the study, wrote the introduction section of the review and contributed to the production of the final report. DM supervised the quality
11
12 390 assessment, methodology and approaches to evidence synthesis, provided senior advice and support throughout and contributed to the
13
14 391 production of the final report. All parties were involved in drafting and/or commenting on the report.
15

16 392

17
18 393 **Competing interests** None declared.
19

20
21 394

22
23 395 **Provenance and peer review** Not commissioned; externally peer reviewed.
24

25
26 396

27
28 397 **Data sharing statement** No additional data are available.
29

30
31 398

32
33 399 REFERENCES
34

- 35
36 400 1. [1] L. Luborsky, L. Diguier, D. A. Seligman, R. Rosenthal, E. D. Krause, S. Johnson, G. Halperin, M. Bishop, J. S. Berman, and E.
37 401 Schweizer, "The Researcher's Own Therapy Allegiances: A 'Wild Card' in Comparisons of Treatment Efficacy," *Clin. Psychol. Sci.*
38 402 *Pract.*, vol. 6, no. 1, pp. 95–106, May 2006.
39
40
41
42
43
44
45

- 1
2
3
4
5 403 [2] E. Dragioti, I. Dimoliatis, and E. Evangelou, "Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of
6
7 404 psychotherapy: a systematic appraisal," *BMJ Open*, vol. 5, no. 6, pp. e007206–e007206, Jun. 2015.
- 8
9 405 [3] T. Munder, O. Brüttsch, R. Leonhart, H. Gerger, and J. Barth, "Researcher allegiance in psychotherapy outcome research: An overview of
10
11 406 reviews," *Clin. Psychol. Rev.*, vol. 33, no. 4, pp. 501–511, Jun. 2013.
- 12
13 407 [4] D. A. Winter, "Editorial." Routledge, 07-May-2010.
- 14
15
16 408 [5] J. McLeod, "Taking allegiance seriously—implications for research policy and practice," *Eur. J. Psychother. Couns.*, May 2010.
- 17
18 409 [6] G. L. Staines and C. M. Cleland, "Bias in meta-analytic estimates of the absolute efficacy of psychotherapy.," *Rev. Gen. Psychol.*, vol. 11,
19
20 410 no. 4, pp. 329–347, 2007.
- 21
22 411 [7] K. D. Markman and E. R. Hirt, "Social Prediction and the 'Allegiance Bias,'" *Soc. Cogn.*, vol. 20, no. 1, pp. 58–86, Feb. 2002.
- 23
24
25 412 [8] G. D. Walters, "The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening version as incrementally
26
27 413 valid predictors of recidivism.," *Law Hum. Behav.*, vol. 33, no. 6, pp. 497–505, 2009.
- 28
29 414 [9] J. P. Singh, M. Grann, and S. Fazel, "Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis," *PLoS*
30
31 415 *One*, vol. 8, no. 9, p. e72484, Sep. 2013.
- 32
33 416 [10] P. R. Blair, D. K. Marcus, and M. T. Boccaccini, "Is There an Allegiance Effect for Assessment Instruments? Actuarial Risk Assessment
34
35 417 as an Exemplar," *Clin. Psychol. Sci. Pract.*, vol. 15, no. 4, pp. 346–360, Oct. 2008.
- 36
37 418 [11] S. O. Lilienfeld and M. K. Jones, "Allegiance Effects in Assessment: Unresolved Questions, Potential Explanations, and Constructive
38
39 419 Remedies," *Clin. Psychol. Sci. Pract.*, vol. 15, no. 4, pp. 361–365, Oct. 2008.
- 40
41
42
43
44
45
46
47
48
49

- 1
2
3
4
5 420 [12] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: validity of a brief depression severity measure.," *J. Gen. Intern. Med.*, vol.
6 421 16, no. 9, pp. 606–13, Sep. 2001.
- 8
9 422 [13] L. Manea, S. Gilbody, and D. McMillan, "A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring
10 423 method as a screen for depression," *Gen. Hosp. Psychiatry*, vol. 37, no. 1, pp. 67–75, Jan. 2015.
- 13 424 [14] A. S. Moriarty, S. Gilbody, D. McMillan, and L. Manea, "Screening and case finding for major depressive disorder using the Patient
14 425 Health Questionnaire (PHQ-9): a meta-analysis," *Gen. Hosp. Psychiatry*, vol. 37, no. 6, pp. 567–576, Nov. 2015.
- 17 426 [15] P. F. Whiting, A. W. S. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. G. Leeflang, J. A. C. Sterne, P. M. M.
18 427 Bossuyt, and QUADAS-2 Group, "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.," *Ann. Intern.
20 428 Med.*, vol. 155, no. 8, pp. 529–36, Oct. 2011.
- 23 429 [16] R. Mann, C. E. Hewitt, and S. M. Gilbody, "Assessing the quality of diagnostic studies using psychometric instruments: applying
24 430 QUADAS," *Soc. Psychiatry Psychiatr. Epidemiol.*, vol. 44, no. 4, pp. 300–307, Apr. 2009.
- 27 431 [17] J. B. Reitsma, A. S. Glas, A. W. S. Rutjes, R. J. P. M. Scholten, P. M. Bossuyt, and A. H. Zwinderman, "Bivariate analysis of sensitivity
28 432 and specificity produces informative summary measures in diagnostic reviews," *J. Clin. Epidemiol.*, vol. 58, no. 10, pp. 982–990, Oct.
30 433 2005.
- 33 434 [18] University of York. NHS Centre for Reviews and Dissemination., *Systematic reviews : CRD's guidance for undertaking reviews in health
34 435 care*. CRD, University of York, 2009.
- 37 436 [19] S. D. Walter, "Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data," *Stat. Med.*, vol. 21, no.
38 437 9, pp. 1237–1256, May 2002.

- 1
2
3
4
5 438 [20] J. G. Lijmer, P. M. M. Bossuyt, and S. H. Heisterkamp, “Exploring sources of heterogeneity in systematic reviews of diagnostic tests,”
6
7 439 *Stat. Med.*, vol. 21, no. 11, pp. 1525–1537, Jun. 2002.
- 8
9 440 [21] S. G. Thompson and J. P. T. Higgins, “How should meta-regression analyses be undertaken and interpreted?,” *Stat. Med.*, vol. 21, no. 11,
10
11 441 pp. 1559–1573, Jun. 2002.
- 12
13 442 [22] C. Diez-Quevedo, T. Rangil, L. Sanchez-Planell, K. Kroenke, and R. L. Spitzer, “Validation and utility of the patient health questionnaire
14
15 443 in diagnosing mental disorders in 1003 general hospital Spanish inpatients,” *Psychosom. Med.*, vol. 63, no. 4, pp. 679–86.
- 16
17
18 444 [23] K. Gräfe, S. Zipfel, W. Herzog, and B. Löwe, “Screening psychischer Störungen mit dem “Gesundheitsfragebogen für Patienten (PHQ-
19
20 445 D)“,” *Diagnostica*, vol. 50, no. 4, pp. 171–181, Oct. 2004.
- 21
22 446 [24] B. Löwe, R. L. Spitzer, K. Gräfe, K. Kroenke, A. Quenter, S. Zipfel, C. Buchholz, S. Witte, and W. Herzog, “Comparative validity of
23
24 447 three screening questionnaires for DSM-IV depressive disorders and physicians’ diagnoses,” *J. Affect. Disord.*, vol. 78, no. 2, pp. 131–40,
25
26 448 Feb. 2004.
- 27
28 449 [25] R. L. Spitzer, K. Kroenke, and J. B. Williams, “Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study.
29
30 450 Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire,” *JAMA*, vol. 282, no. 18, pp. 1737–44, Nov. 1999.
- 31
32 451 [26] P. Thekkumpurath, J. Walker, I. Butcher, L. Hodges, A. Kleiboer, M. O’Connor, L. Wall, G. Murray, K. Kroenke, and M. Sharpe,
33
34 452 “Screening for major depression in cancer outpatients: the diagnostic accuracy of the 9-item patient health questionnaire,” *Cancer*, vol.
35
36 453 117, no. 1, pp. 218–27, Jan. 2011.
- 37
38 454 [27] K. Muramatsu, H. Miyaoka, K. Kamijima, Y. Muramatsu, M. Yoshida, T. Otsubo, and F. Gejyo, “The patient health questionnaire,
39
40 455 Japanese version: validity according to the mini-international neuropsychiatric interview-plus,” *Psychol. Rep.*, vol. 101, no. 3 Pt 1, pp.

- 1
2
3
4
5 456 952–60, Dec. 2007.
6
7
8 457 [28] R. Navinés, P. Castellví, J. Moreno-España, D. Gimenez, M. Udina, S. Cañizares, C. Diez-Quevedo, M. Valdés, R. Solà, and R. Martín-
9 458 Santos, “Depressive and anxiety disorders in chronic hepatitis C patients: Reliability and validity of the Patient Health Questionnaire,” *J.*
10 459 *Affect. Disord.*, vol. 138, no. 3, pp. 343–351, May 2012.
11
12
13 460 [29] S. M. Eack, C. G. Greeno, and B.-J. Lee, “Limitations of the Patient Health Questionnaire in Identifying Anxiety and Depression: Many
14 461 Cases Are Undetected.,” *Res. Soc. Work Pract.*, vol. 16, no. 6, pp. 625–631, Nov. 2006.
15
16
17 462 [30] J. R. Fann, C. H. Bombardier, S. Dikmen, P. Esselman, C. A. Warm, E. Pelzer, H. Rau, and N. Temkin, “Validity of the Patient Health
18 463 Questionnaire-9 in assessing depression following traumatic brain injury.,” *J. Head Trauma Rehabil.*, vol. 20, no. 6, pp. 501–11.
19
20
21 464 [31] B. Gelaye, M. A. Williams, S. Lemma, N. Deyessa, Y. Bahretibeb, T. Shibre, D. Wondimagegn, A. Lemenhe, J. R. Fann, A. Vander
22 465 Stoep, and X.-H. H. Andrew Zhou, “Validity of the patient health questionnaire-9 for depression screening and diagnosis in East Africa,”
23 466 *Psychiatry Res.*, vol. 210, no. 2, pp. 653–661, Dec. 2013.
24
25
26
27 467 [32] T. Hyphantis, K. Kotsis, P. V. Voulgari, N. Tsifetaki, F. Creed, and A. A. Drosos, “Diagnostic accuracy, internal consistency, and
28 468 convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders,”
29 469 *Arthritis Care Res. (Hoboken)*, vol. 63, no. 9, pp. 1313–1321, Sep. 2011.
30
31
32
33 470 [33] M. Inagaki, T. Ohtsuki, N. Yonemoto, Y. Kawashima, A. Saitoh, Y. Oikawa, M. Kurosawa, K. Muramatsu, T. A. Furukawa, and M.
34 471 Yamada, “Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural
35 472 hospital: a cross-sectional study.,” *Gen. Hosp. Psychiatry*, vol. 35, no. 6, pp. 592–7, Jan. 2013.
36
37
38
39 473 [34] M. E. Khamseh, H. R. Baradaran, A. Javanbakht, M. Mirghorbani, Z. Yadollahi, and M. Malek, “Comparison of the CES-D and PHQ-9
40
41
42
43
44
45
46
47
48
49

- 1
2
3
4
5 474 depression scales in people with type 2 diabetes in Tehran, Iran,” *BMC Psychiatry*, vol. 11, no. 1, p. 61, Dec. 2011.
- 6
7
8 475 [35] P. Persoons, K. Luyckx, C. Desloovere, J. Vandenberghe, and B. Fischler, “Anxiety and mood disorders in otorhinolaryngology
9 476 outpatients presenting with dizziness: validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology.,”
10 477 *Gen. Hosp. Psychiatry*, vol. 25, no. 5, pp. 316–23.
- 11
12
13 478 [36] A. Picardi, D. A. Adler, D. Abeni, H. Chang, P. Pasquini, W. H. Rogers, and K. M. Bungay, “Screening for depressive disorders in
14 479 patients with skin diseases: a comparison of three screeners.,” *Acta Derm. Venereol.*, vol. 85, no. 5, pp. 414–9, 2005.
- 15
16
17 480 [37] L. Stafford, M. Berk, and H. J. Jackson, “Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to
18 481 screen for depression in patients with coronary artery disease,” *Gen. Hosp. Psychiatry*, vol. 29, no. 5, pp. 417–424, Sep. 2007.
- 19
20
21 482 [38] B. D. Thombs, R. C. Ziegelstein, and M. A. Whooley, “Optimizing detection of major depression among patients with coronary artery
22 483 disease using the patient health questionnaire: data from the heart and soul study.,” *J. Gen. Intern. Med.*, vol. 23, no. 12, pp. 2014–7, Dec.
23 484 2008.
- 24
25
26
27 485 [39] A. W. Thompson, H. Liu, R. D. Hays, W. J. Katon, R. Rausch, N. Diaz, E. L. Jacob, S. D. Vassar, and B. G. Vickrey, “Diagnostic
28 486 accuracy and agreement across three depression assessment measures for Parkinson’s disease,” *Parkinsonism Relat. Disord.*, vol. 17, no.
29 487 1, pp. 40–45, Jan. 2011.
- 30
31
32
33 488 [40] A. Turner, J. Hambridge, J. White, G. Carter, K. Clover, L. Nelson, and M. Hackett, “Depression screening in stroke: a comparison of
34 489 alternative measures with the structured diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth edition
35 490 (major depressive episode) as criterion standard.,” *Stroke.*, vol. 43, no. 4, pp. 1000–5, Apr. 2012.
- 36
37
38
39 491 [41] K. M. van Steenbergen-Weijnenburg, L. de Vroege, R. R. Ploeger, J. W. Brals, M. G. Vloedveld, T. F. Veneman, L. Hakkaart-van Roijen,
40
41
42
43
44
45

- 1
2
3
4
5 492 F. F. Rutten, A. T. Beekman, and C. M. van der Feltz-Cornelis, "Validation of the PHQ-9 as a screening instrument for depression in
6 diabetes patients in specialized outpatient clinics," *BMC Health Serv. Res.*, vol. 10, no. 1, p. 235, Dec. 2010.
7 493
- 8
9 494 [42] B. Arroll, F. Goodyear-Smith, S. Crengle, J. Gunn, N. Kerse, T. Fishman, K. Falloon, and S. Hatcher, "Validation of PHQ-2 and PHQ-9
10 to Screen for Major Depression in the Primary Care Population," *Ann. Fam. Med.*, vol. 8, no. 4, pp. 348–353, Jul. 2010.
11 495
- 12
13 496 [43] L. Ayalon, M. Goldfracht, and P. Bech, "'Do you think you suffer from depression?': Reevaluating the use of a single item question for
14 the screening of depression in older primary care patients," *Int. J. Geriatr. Psychiatry*, vol. 25, no. 5, pp. 497–502, May 2010.
15 497
- 16
17 498 [44] V. Henkel, R. Mergl, R. Kohnen, A.-K. Allgaier, H.-J. Möller, and U. Hegerl, "Use of brief depression screening tools in primary care:
18 consideration of heterogeneity in performance in different patient groups," *Gen. Hosp. Psychiatry*, vol. 26, no. 3, pp. 190–198, May 2004.
19 499
- 20
21 500 [45] F. Lamers, C. C. M. Jonkers, H. Bosma, B. W. J. H. Penninx, J. A. Knottnerus, and J. T. M. van Eijk, "Summed score of the Patient
22 Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients," *J. Clin. Epidemiol.*,
23 501
24 vol. 61, no. 7, pp. 679–687, Jul. 2008.
25 502
- 26
27 503 [46] M. Lotrakul, S. Sumrithe, and R. Saipanish, "Reliability and validity of the Thai version of the PHQ-9," *BMC Psychiatry*, vol. 8, no. 1, p.
28 46, Dec. 2008.
29 504
- 30
31 505 [47] N. P. Zuithoff, Y. Vergouwe, M. King, I. Nazareth, M. J. van Wezep, K. G. Moons, and M. I. Geerlings, "The Patient Health
32 Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study,"
33 506
34 *BMC Fam. Pract.*, vol. 11, no. 1, p. 98, Dec. 2010.
35 507
- 36
37 508 [48] D. Gjerdingen, S. Crow, P. McGovern, M. Miner, and B. Center, "Postpartum depression screening at well-child visits: validity of a 2-
38 question screen and the PHQ-9," *Ann. Fam. Med.*, vol. 7, no. 1, pp. 63–70, 2009.
39 509
40
41
42
43
44
45

- 1
2
3
4
5 510 [49] V. Henkel, R. Mergl, R. Kohnen, A.-K. Allgaier, H.-J. Möller, and U. Hegerl, “Use of brief depression screening tools in primary care:
6 consideration of heterogeneity in performance in different patient groups.,” *Gen. Hosp. Psychiatry*, vol. 26, no. 3, pp. 190–8.
7 511
- 8
9 512 [50] L. S. Williams, E. J. Brizendine, L. Plue, T. Bakas, W. Tu, H. Hendrie, and K. Kroenke, “Performance of the PHQ-9 as a screening tool
10 for depression after stroke.,” *Stroke.*, vol. 36, no. 3, pp. 635–8, Mar. 2005.
11 513
- 12
13 514 [51] M. H. N. Chagas, V. Tumas, G. R. Rodrigues, J. P. Machado-de-Sousa, A. S. Filho, J. E. C. Hallak, and J. A. S. Crippa, “Validation and
14 internal consistency of Patient Health Questionnaire-9 for major depression in Parkinson’s disease,” *Age Ageing*, vol. 42, no. 5, pp. 645–
15 515 649, Sep. 2013.
16 516
- 17
18
19 517 [52] L. Elderon, K. G. Smolderen, B. Na, and M. A. Whooley, “Accuracy and prognostic value of American Heart Association: recommended
20 depression screening in patients with coronary heart disease: data from the Heart and Soul Study.,” *Circ. Cardiovasc. Qual. Outcomes*,
21 518 vol. 4, no. 5, pp. 533–40, Sep. 2011.
22 519
- 23
24
25 520 [53] A. G. Rooney, S. McNamara, M. Mackinnon, M. Fraser, R. Rampling, A. Carson, and R. Grant, “Screening for major depressive disorder
26 in adults with cerebral glioma: an initial validation of 3 self-report instruments,” *Neuro. Oncol.*, vol. 15, no. 1, pp. 122–129, Jan. 2013.
27 521
- 28
29 522 [54] S. Watnick, P.-L. Wang, T. Demadura, and L. Ganzini, “Validation of 2 depression screening tools in dialysis patients.,” *Am. J. Kidney*
30 523 *Dis.*, vol. 46, no. 5, pp. 919–24, Nov. 2005.
31 524
- 32
33 524 [55] Y. Zhang, R. Ting, M. Lam, J. Lam, H. Nan, R. Yeung, W. Yang, L. Ji, J. Weng, Y.-K. Wing, N. Sartorius, and J. C. N. Chan,
34 “Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes,” *J.*
35 525 *Affect. Disord.*, vol. 151, no. 2, pp. 660–666, Nov. 2013.
36 526
- 37
38
39 527 [56] A. O. Adewuya, B. A. Ola, and O. O. Afolabi, “Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression
40
41
42
43
44
45

- 1
2
3
4
5 528 amongst Nigerian university students,” *J. Affect. Disord.*, vol. 96, no. 1–2, pp. 89–93, Nov. 2006.
6
7
8 529 [57] T. H. Fine, A. A. Contractor, M. Tamburrino, J. D. Elhai, M. R. Prescott, G. H. Cohen, E. Shirley, P. K. Chan, T. Goto, R. Slembariski, I.
9 530 Liberzon, S. Galea, and J. R. Calabrese, “Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I
10 531 major depression module,” *J. Affect. Disord.*, vol. 150, no. 3, pp. 1001–1007, Sep. 2013.
11
12
13 532 [58] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt, “Screening for depression in medical settings with the Patient Health Questionnaire
14 533 (PHQ): A diagnostic meta-analysis,” *J. Gen. Intern. Med.*, vol. 22, no. 11, pp. 1596–1602, Oct. 2007.
15
16
17 534 [59] E. Phelan, B. Williams, K. Meeker, K. Bonn, J. Frederick, J. LoGerfo, and M. Snowden, “A study of the diagnostic accuracy of the PHQ-
18 535 9 in primary care elderly,” *BMC Fam. Pract.*, vol. 11, no. 1, p. 63, Dec. 2010.
19
20
21 536 [60] A. C. Sidebottom, P. A. Harrison, A. Godecker, and H. Kim, “Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal
22 537 depression screening,” *Arch. Womens. Ment. Health*, vol. 15, no. 5, pp. 367–374, Oct. 2012.
23
24
25 538 [61] F. de Lima Osório, A. Vilela Mendes, J. A. Crippa, and S. R. Loureiro, “Study of the Discriminative Validity of the PHQ-9 and PHQ-2 in
26 539 a Sample of Brazilian Women in the Context of Primary Health Care,” *Perspect. Psychiatr. Care*, vol. 45, no. 3, pp. 216–227, Jul. 2009.
27
28
29 540 [62] V. Patel, R. Araya, N. Chowdhary, M. King, B. Kirkwood, S. Nayak, G. Simon, and H. A. Weiss, “Detecting common mental disorders in
30 541 primary care in India: a comparison of five screening questionnaires,” *Psychol. Med.*, vol. 38, no. 2, Feb. 2008.
31
32
33 542 [63] M. N. N. Azah, M. E. M. Shah, S. Juwita, I. S. Bahri, W. M. W. M. Rushidi, and Y. M. Jamil, “Validation of the Malay Version Brief
34 543 Patient Health Questionnaire (PHQ-9) among Adult Attending Family Medicine Clinics,” *Int. Med. J.*, 2005.
35
36
37 544 [64] S.-I. Liu, Z.-T. Yeh, H.-C. Huang, F.-J. Sun, J.-J. Tjung, L.-C. Hwang, Y.-H. Shih, and A. W.-C. Yeh, “Validation of Patient Health
38 545 Questionnaire for depression screening among primary care patients in Taiwan,” *Compr. Psychiatry*, vol. 52, no. 1, pp. 96–101, Jan. 2011.
39
40
41
42
43
44
45
46
47
48
49

- 1
2
3
4
5 546 [65] K. Wittkamp, H. van Ravesteijn, K. Baas, H. van de Hoogen, A. Schene, P. Bindels, P. Lucassen, E. van de Lisdonk, and H. van Weert,
6 547 “The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary
7 548 care,” *Gen. Hosp. Psychiatry*, vol. 31, no. 5, pp. 451–459, Sep. 2009.
8
9
10
11 549
12
13 550
14
15
16 551
17
18 552
19
20
21 553
22
23 554
24
25
26 555
27
28 556
29
30
31 557
32
33 558
34
35
36 559
37
38 560
39
40
41
42
43
44
45
46
47
48
49

For peer review only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Table 1: Descriptive characteristics of algorithm studies (Manea et al., 2014)					
Study	Sample characteristics	Sample size and % depressed	PHQ-9 characteristics	Diagnostic standard	a) Conflict of interest (COI) declaration b) Funding c) Relationship with original developers
	(Country, setting, age, sex)				
Diez-Quevedo et al. (2001)	Country: Spain Setting: Medical and surgical tertiary hospitals Age (yrs): M=43 (SD=14.2) Female: 45.6%	N = 1003 Depressed: 8.2%	Administration: Self-report Language: Spanish	DSM-III-R SCID	a) No COI declaration b) Funding acknowledged (academic institutions) c) Not acknowledged
Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Age (yrs): M = 41.9 (SD = 13.8) Female: 67.8%	N = 528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Language: German Administration: self-report	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) Not acknowledged

Lowe et al. (2004)	Country: Germany Setting: Outpatient clinics and family practices Age (yrs): M = 41.7 (SD = 13.8) Female: 67.1%	N = 501 Depressed: 13.2%	Administration: Self-report Language: German	DSM-IV SCID	a) COI declaration 'This study was supported by unrestricted restricted grants from Pfizer Germany and from the medical faculty of the University of Heidelberg Germany, and there are no COI.' b) Acknowledged funding from Pfizer and academic institution c) Not acknowledged
Muramatsu et al. (2007)	Country: Japan Setting: Primary care and general hospital Age (yrs): M = 43.3 (SD = 16.4) Female: 59.5%	N = 131 Depressed: 28.2%	Administration: Self-report Language: Japanese	DSM-IV MINI	a) No COI declaration b) Acknowledged funding from Pfizer c) Acknowledged one of the developers of the PHQ-9: 'The authors acknowledge Dr R L Spitzer'
Navinés et al. (2012)	Country: Spain Setting: General hospital (patients with chronic HCV) Age (yrs): M = 43.4 (SD = 10.2) Female: 28.6%	N = 500 Depressed: 6.4%	Administration: Self-report Language: Spanish	DSM-IV SCID	a) All authors declared that they had no COI. b) Role of funding source declared c) Not acknowledged

Spitzer et al. (1999)	Country: US Setting: Primary care Age (yrs): M = 46 (SD = 17.2) Female: 66%	N = 3000 (585 received SCID) Depressed: 10%	Administration: Self-report Language: English	DSM-III-R SCID	a) No COI declaration b) Acknowledged funding from Pfizer. 'Drs Spitzer and Williams receive honoraria and consulting money from Pfizer Inc, which has supported this work.' c) N/A
Thekkumpurath et al. (2010)	Country: UK Setting: Hospital (cancer patients) Age (yrs): M = 61 Female: 63%	N = 782 Depressed: 6.3% (of the whole sample)	Administration: Not stated Language: English	DSM-IV SCID	a) COI declaration: 'Supported by Cancer Research UK' b) As in a) c) Not acknowledged
Ayalon et al. (2010)	Country: Israel Age (yrs): M = 75 (SD = 8.1) Female: 40.5 %	N = 153 Depressed: 3.9 %	Administration: Researcher administered Language: Hebrew	DSM-IV SCID	a) COI declaration: 'The project was funded by an Investigator's Initiated Research Grant from Lundbeck International given to Dr Liat Ayalon. Lundbeck International had no other involvement in the project concept of design or in this paper. Per Bech has occasionally over the past 3 years until August 2008 received funding from and has been speaker or member of advisory boards for pharmaceutical companies with an interest in the drug treatment of affective disorders (Astra-Zeneca, Lilly, H. Lundbeck A/S, Lundbeck Foundation and Organon).' b) Acknowledged funding from Lundbeck International

Eack et al. (2006)	Country: US Setting: Community mental health centers for children Age (yrs): M = 39.20 (SD 9.63) Female: 100%	N = 50 Depressed: 28%	Administration: Self-report Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Fann et al. (2005)	Country: US Setting: Trauma hospital (inpatients with traumatic brain injury) Age (yrs): M = 42 (SD=17.9) Female: 29.1%	N = 135 Depressed: 16.3%	Administration: Telephone-administered Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

<p>Gelaye et al. (2011)</p>	<p>Country: Ethiopia</p> <p>Setting: General hospital</p> <p>Age (yrs): 34.9 (SD=11.6)</p> <p>Female: 63.1 %</p>	<p>N = 363</p> <p>Depressed: 12.6%</p>	<p>Administration: Researcher-administered</p> <p>Language: Amharic</p>	<p>DSM-IV SCAN</p>	<p>a) No COI declaration b) Funding acknowledged (academic /health research institutions)</p>
<p>Gjerdingen et al. (2009)</p>	<p>Country: US</p> <p>Setting: Community</p> <p>Age (yrs): M = 29.3</p> <p>Female: 100%</p>	<p>N = 438</p> <p>Depressed: 4.6%</p>	<p>Administration: Telephone or self-report</p> <p>Language: English</p>	<p>DSM-IV SCID</p>	<p>a) No COI declaration b) Funding acknowledged (academic /health research institutions)</p>
<p>Henkel et al. (2004)</p>	<p>Country: Germany</p> <p>Setting: primary care</p> <p>Age (yrs): not reported</p> <p>Female: 74%</p>	<p>N = 448</p> <p>Depressed: 10%</p>	<p>Administration: self-report</p> <p>Language: German</p>	<p>DSM-IV CIDI</p>	<p>a) No COI declaration b) Funding acknowledged (academic /health research institutions)</p>

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	Hyphantis et al. (2011)	Country: Greece Setting: Hospital – rheumatology patients Age (yrs): M = 54.2 (SD = 13.5) Female: 74%	N = 213 Depressed: 32.4%	Administration: Researcher administered Language: Greek	DSM-IV MINI	a) No COI declaration b) No funding acknowledgement
16 17 18 19 20 21 22 23 24 25 26 27	Inagaki et al. (2013)	Country: Japan Setting: General hospital Age whole sample (yrs): M = 73.5 (SD = 12.3) Female: 59.3%	N = 104 out of 511 received MINI Depressed: 7.4%	Administration: Researcher administered Language: Japanese	DSM-IV MINI	a) COI declaration: ‘The authors declare that they have no competing interests.’ b) Funding acknowledged (academic /health research institutions)
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49	Khamseh et al. (2011)	Country: Iran Setting: Diabetes clinic Age (yrs): M = 56.17 (SD = 9.60) Female: 51.9%	N = 185 Depressed: 43.2%	Administration: Self report Language: Persian	DSM-IV SCID	a) COI declaration: The authors declared no competing interests b) Funding acknowledged (academic /health research institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Lamers et al. (2008)	Country: Netherlands Setting: Primary care (elderly) Age (yrs): M = 71.4 (SD = 6.90) Female: 48.2%	N = 713 Depressed: 10.7%	Administration: Self report Language: Dutch	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Lotrakul et al. (2008)	Country: Thailand Setting: Primary care Age (yrs): M = 45.0 (SD = 14.30) Female: 73.7%	N = 279 Depressed: 6.8%	Administration: Self report Language: Thai	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Persoons et al. (2003)	Country: Belgium Setting: Hospital (otolaryngology patients) Age (yrs): M = 48.2 (SD = 12.9) Female: 65.6%	N = 268 (97 received MINI) Depressed: 16.5%	Administration: Self-report Language: Dutch	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic /health research institutions) and Pfizer Belgium

Picardi et al. (2005)	Country: Italy Setting: Hospital (dermatology inpatients) Age (yrs): M = 37.5 Female: 56%	N = 141 Depressed: 8.5%	Administration: Self-report Language: Italian	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions). Acknowledged Pfizer Italia SRL for providing the Italian version of the PHQ-9 and for permission to use it.
Stafford et al. (2007)	Country: Australia Setting: Hospital (cardiology patients) Age (yrs): M = 64.1 (SD = 10.3) Female: 66%	N = 193 Depressed: 18%	Administration: Self-report Language: English	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
Thombs et al. (2008)	Country: US Setting: Hospital (outpatients with coronary heart disease) Age (yrs): M = 67 (SD = 11) Female: 18%	N = 1024 Depressed: 22%	Administration: Not stated Language: English	DSM C-DIS	a) COI declaration "None disclosed" b) Funding acknowledged (academic/health research institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Thompson et al. (2010)	Country: US Setting: Patients with Parkinson Disease Age (yrs): 72.5 (SD = 9.6) Female: 42%	N = 214 Depressed: 14%	Administration: Self administered Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
Turner et al. (2012)	Country: Australia Setting: Stroke patients Age (yrs): 66.7 (SD = 13.1) Female: 47.2%	N = 72 Depressed: 18%	Administration: Self administered Language: English	DSM-IV SCID	a) COI declaration: Disclosures 'None'. b) Funding acknowledged (academic/health research institutions)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	van Steenberg-Weijnenburg (2010)	Country: Netherlands Setting: Diabetes patients Age (yrs): M = 61.8 (SD = 13.6) Female: 48.7%	N = 197 Depressed: 18.8%	Administration: Self administered Language: Dutch	DSM-IV SCID	a) COI declaration: 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions) - 'this had no influence on the content of this article'.
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49	Zuithoff et al. (2010)	Country: Netherlands Setting: Primary care Age (yrs): M = 51 (sd = 16.7) Female: 63%	N = 1338 Depressed: 13%	Administration: Self-report Language: Dutch	DSM-IV CIDI	a) COI declaration 'The authors declare that they have no competing interests.' b) Funding acknowledged (academic/health research institutions).

Table 2: Descriptive characteristics of the summed items scoring method studies cut-off point 10 (Moriarty et al, 2015)

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	Conflict of interest (COI) declaration Funding c) Relationship with original developers
13. Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Mean age: 41.9 (SD = 13.8) Female: 67.8%	N = 528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Administration: self-report Language: German Cut-offs: 10 to 14	DSM-IV SCID	No COI declaration Acknowledged funding from Pfizer Not acknowledged
16. Kroenke et al. (2001)	Country: USA Setting: Primary care Mean age: 46 (SD=17) Female: 66%	N = 580 7.1% MDD	Administration: Self-report Language: English Cut-offs: 9 to 15	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) N/A
22. Navinés et al. (2012)	Country: Spain Setting: General hospital (patients with chronic HCV) Mean age: 43.4 (SD = 10.2) Female: 28.6%	N = 500 6.4% MDD	Administration: Self-report Language: Spanish Cut-offs: 10	DSM-IV SCID	a) All authors declared that they had no COI. b) Role of funding source declared c) Not acknowledged
29. Thekkumpurath et al. (2010)	Country: UK Setting: Hospital (cancer patients) Mean age: 61 Female: 63%	N = 782 6.3% MDD (of the whole sample)	Administration: Not stated Language: English Cut-offs: 5 to 10	DSM-IV SCID	c) COI declaration: 'Supported by Cancer Research UK' d) As in a) e) Not acknowledged
33. Williams et al. (2005)	Country: USA	N = 316	Administration: Unclear	DSM-IV SCID	a) No COI declaration b) Funding

561

	Setting: Secondary care (Post-stroke) Mean age: Unclear Female: Unclear	33.5% MDD	Language: English Cut-offs: 10		acknowledged (academic institutions) c) Not acknowledged
1. Adewuya et al. (2006)	Country: Nigeria Setting: community (students) Mean age: 24.8 (15-40) Female: 41.2%	N = 512 2.5% MDD	Administration: Self-report Language: English Cut-offs: 8 to 12	DSM-IV MINI	a) No COI declaration b) No funding declaration
2. Arroll et al. (2010)	Country: New Zealand Setting: Primary care Mean age: 49 (17-99) Female: 61%	N = 2642 6.2% MDD	Administration: Not stated Language: English Cut-offs: 8,10,12,15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
3. Azah et al. (2005)	Country: Malaysia Setting: Primary care Mean age: 38.7 (18-79) Female: 61.7%	N = 180 16.6% MDD	Administration: Self-report Language: Malay Cut-offs: 5 to 12	DSM-IV CIDI	b) No COI declaration c) Funding acknowledged (academic /health research institutions)
4. Chagas et al. (2013)	Country: Brazil	N = 84	Administration: self-report	DSM-IV SCID	a) COI declaration "None declared"

	Setting: Secondary care Mean age: Not stated Female: 52.7%	25.5% MDD	Language: Brazilian Cut-offs: 7 to 10		b) Funding acknowledged (academic/health research institutions)
6. de Lima Osorio et al. (2009)	Country: Brazil Setting: Primary care Mean age: Unclear Female: 100%	N = 177 34% MDD	Administration: research assistants Language: Brazilian Portuguese Cut-offs: 10 to 15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)
7. Elderon et al. (2011)	Country: USA Setting: Secondary care Mean age: Unclear Female: 18%	N = 1022 18.3% MDD	Administration: self-report Language: English Cut-offs: 10	C-DIS	a) COI declaration – ‘No disclosures’ b) Funding acknowledged (academic institutions and industry – AHA Pharmaceuticals Roundtable) – ‘The funding organisations had no role in the design or conduct of the study, collection, management, analysis or interpretation of data; or preparation, review or approval of the manuscript.’
8. Fann et al. (2005)	Country: US Setting: Trauma hospital (inpatients with traumatic	N = 135 16.3% MDD	Administration: Telephone-administered Language: English	DSM-IV SCID	b) No COI declaration c) Funding acknowledged (academic

	brain injury) Mean age: 42 (SD=17.9) Female: 29.1%		Cut-offs: 10		institutions)
9. Fine et al. (2013)	Country: USA Setting: Primary care (Ohio Army National Guard) Mean age: 31 (17-60) Female: 12%	N = 498 21.5% MDD	Administration: Telephone-administered Language: English Cut-offs: 10,15	DSM-IV SCID-I	a) COI – last author disclosed financial and consulting interests (Pfizer not one of them). All other authors declared that they have no COI. b) Funding acknowledged – DoD Medical Research. ‘The sponsor had no role in study design, data collection, analysis, interpretation of results, report writing or manuscript submission.
10. Gelaye et al. (2013)	Country: Ethiopia Setting: General hospital Mean age: 34.9 (SD=11.6) Female: 63.1 %	N = 363 12.6% MDD	Administration: Researcher-administered Language: Amharic Cut-offs: 9 to 11	DSM-IV SCAN	c) No COI declaration d) Funding acknowledged (academic /health research institutions)
11. Gilbody et al.	Country: UK	N = 96	Administration: Not	DSM-IV	a) COI declaration –

(2007)	Setting: Primary care Mean age: 42.5 (SD 13.6) Female: 77%	37.5 MDD	stated Language: English Cut-offs: 9 to 13	SCID	last author involved in the development of one of the instruments (CORE-OM), 'but does not gain financially from its use. b) Funding acknowledged (academic /health research institutions)
12. Gjerdingen et al. (2009)	Country: USA Setting: Community Mean age: 29.3 Female: 100%	N = 438 4.6% MDD	Administration: Telephone or self-report Language: English Cut-offs: 10	DSM-IV SCID	c) No COI declaration d) Funding acknowledged (academic /health research institutions)
14. Hyphantis et al. (2011)	Country: Greece Setting: Hospital – rheumatology patients Mean age: 54.2 (SD = 13.5) Female: 74%	N = 213 32.4% MDD	Administration: Researcher administered Language: Greek Cut-offs: 4 to 16	DSM-IV MINI	c) No COI declaration d) No funding acknowledgement
15. Khamseh et al. (2011)	Country: Iran Setting: Outpatient diabetic clinic Mean age: 56.1 (SD=9.6)	N = 185 43.2% MDD	Administration: Self-report Language: Persian Cut-offs: 10,13	DSM-IV SCID	c) COI declaration: The authors declared no competing interests d) Funding acknowledged (academic /health

					research institutions)
19. Liu et al. (2011)	<p>Female: 51.8%</p> <p>Country: Taiwan</p> <p>Setting: Primary care</p> <p>Mean age: Not specified</p> <p>Female: 60.9%</p>	<p>N = 1532</p> <p>3.3% MDD</p>	<p>Administration: Self-report</p> <p>Language: Chinese version</p> <p>Cut-offs: 9 to 11</p>	SCAN	<p>a) a) No COI declaration</p> <p>b) Funding acknowledged (academic /health research institutions)</p>
20. Lotrakul et al. (2008)	<p>Country: Thailand</p> <p>Setting: Primary care</p> <p>Mean age: 45.0 (SD = 14.30)</p> <p>Female: 73.7%</p>	<p>N = 279</p> <p>6.8% MDD</p>	<p>Administration: Self report</p> <p>Language: Thai</p> <p>Cut-offs: 7 to 15</p>	DSM-IV MINI	<p>c) No COI declaration</p> <p>d) Funding acknowledged (academic /health research institutions)</p>
23. Patel et al. (2008)	<p>Country: India</p> <p>Setting: Primary care</p> <p>Mean age: 37.5 (18-83)</p> <p>Female: 56.4%</p>	<p>N = 299</p> <p>4.3% MDD</p>	<p>Administration: Face-to-face interview</p> <p>Language: Not specified</p> <p>Cut-offs: 7 to 15</p>	CIS-R	<p>a) COI declaration – No Declaration of Interest</p> <p>b) Funding acknowledged (academic /health research institutions)</p>
24. Phelan et al. (2010)	<p>Country: USA</p> <p>Setting: Primary care (elderly)</p> <p>Mean age: 78 (SD=7)</p> <p>Female: 62%</p>	<p>N = 71</p> <p>12% MDD</p>	<p>Administration: Research assistant</p> <p>Language: English</p> <p>Cut-offs: 8 to 12</p>	DSM-IV SCID	<p>a) COI declaration – No competing interests</p> <p>b) Funding acknowledged (academic /health research institutions) . ‘The funder had no role in the study design, methods,</p>

					data collection, analysis or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication.
25. Rooney et al. (2013)	Country: UK Setting: Secondary care (glioma) Mean age: 54.2 (SD=12.3) Female: 42.6%	N = 129 13.5% MDD	Administration: Self-report Language: English Cut-offs: 8 to 11	DSM-IV SCID	a) COI declaration "The authors declare that they have no COI" b) Funding acknowledged (academic/health research institutions)
26. Sherina et al. (2012)	Country: Malaysia Setting: Primary care Mean age: 30.9 (18-81) Female: 100%	N= 146 21.2% MDD	Administration: Self-report Language: Malay Cut-offs: 10	CIDI	a) COI declaration "The authors declare that they have no competing interests" b) Funding acknowledged (academic/health research institutions)
27. Sidebottom et al. (2012)	Country: USA Setting: Community (prenatal) Mean age: 23 (SD=5.5) Female: 100%	N = 745 3.6% MDD	Administration: Interview Language: English Cut-offs: 10	DSM-IV SCID	b) COI declaration "The authors declare that they have no financial COI" b) Funding acknowledged (academic/health research institutions)
28. Stafford et al. (2007)	Country: Australia Setting: Secondary care (cardiac procedures)	N = 193 18.1% MDD	Administration: Self-report Language: English	DSM-IV MINI	b) No COI declaration c) Funding acknowledged (academic/health

	Mean age: 64.14 (38-91) Female: 19.2%		Cut-offs: 10		research institutions)
30. Thombs et al. (2008)	Country: US Setting: Hospital (outpatients with coronary heart disease) Mean age: 67 (SD = 11) Female: 18%	N = 1024 22% MDD	Administration: Not stated Language: English Cut-offs: 7 to 10	DSM C-DIS	b) COI declaration "None disclosed" b) Funding acknowledged (academic/health research institutions)
32. Watnick et al. (2005)	Country: USA Setting: Secondary care (dialysis) Mean age: 63 (SD=15) Female: 32.3%	N = 62 19% MDD	Administration: Self-report Language: English Cut-offs: 10	DSM-IV SCID	b) No COI declaration c) Funding acknowledged (academic/health research institutions)
34. Wittkamp et al. (2009)	Country: Netherlands Setting: Primary care Mean age: 49.8 Female: 66.7%	N = 664 12.3% MDD	Administration: Self-report Language: Not specified Cut-offs: 10 and 15	DSM-IV SCIDI	No COI declaration b) Funding acknowledged (academic/health research institutions)
35. Zhang et al. (2013)	Country: Hong Kong Setting: Secondary care (diabetic outpatients) Mean age: 55.1 (SD=9.5)	N = 99 23.2% MDD	Administration: Self-report Language: Chinese version Cut-offs: 15	DSM-IV MINI	COI declaration – last author acknowledged financial COI. The other authors declare that they have no competing interests.) Funding acknowledged

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

	Female: 40.8%				(academic/health research institutions)
36. Zuithoff et al. (2010)	Country: Netherlands Setting: Primary care Age (yrs): M = 51 (sd = 16.7) Female: 63%	N = 1338 Depressed: 13%	Administration: Self-report Language: Dutch	DSM-IV CIDI	b) COI declaration "The authors declare that they have no competing interests." b) Funding acknowledged (academic/health research institutions)

562
563

Table 3: Quality assessment of included studies in the algorithm meta-analysis (Manea et al., 2014)

Study	Patient selection:	Patient selection:	Patient selection:	Patient selection:	Index test:	Index test:	Index test:	Index test:
	Consecutive or random sample	Avoid case-control / avoid artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias	PHQ-9 interpreted blind to reference test	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias
Allegiant studies								
Diez-Quevedo et al. (2001)	✗	✓	✗	High	?	✓	✓	Unclear
Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	✓	Unclear

1									
2									
3									
4									
5	Lowe et al.	✘	✓	✓	High	✓	✓	✓	Low
6	(2004)								
7									
8	Muramatsu et	?	✓	?	Unclear	✓	✓	?	Unclear
9	al. (2007)								
10									
11	Navines et al.	✓	✓	✓	Low	✓	✓	?	Unclear
12	(2012)								
13									
14	Spitzer et al.	✘	✓	✓	High	✓	n/a	n/a	Low
15	(1999)								
16									
17	Thekkumpurath	✘	✘	✓	High	✓	n/a	n/a	Low
18	et al. (2010)								
19									
20									
21	Non-allegiant studies								
22									
23	Arroll et al.	✓	✓	✓	Low	✓	n/a	n/a	Low
24	(2010)								
25									
26	Ayalon et al. (?	✓	✓	Unclear	?	✓	?	Unclear
27	2010)								
28									
29	Eack et al.	?	✓	?	Unclear	?	n/a	n/a	Unclear
30	(2006)								
31									
32	Fann et al.	✓	✘	✘	High	✓	n/a	n/a	Low
33	(2005)								
34									
35	Gelaye et al.	?	✘	?	High	✓	✓	?	Unclear
36	(2013)								
37									
38	Gjerdingen et	✓	✓	✓	Low	?	n/a	n/a	Unclear
39									
40									
41									
42									
43									
44									
45									
46									
47									
48									
49									

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

al. (2009)

Henkel et al. (2004)	✓	✓	✓	Low	?	n/a	n/a	Unclear
Hyphantis et al. (2011)	✓	✓	✗	High	✓	?	?	Unclear
Inagaki et al. (2013)	✓	✗	✓	High	✓	?	?	Unclear
Khamseh et al. (2011)	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al (2008)	✓	✗	✗	High	✓	?	?	Unclear
Lotrakul et al. (2008)	✗	✓	?	High	✓	✓	?	Unclear
Persoons et al. (2003)	✓	✓	✓	Low	✓	✓	n/a	Low
Picardi et al. (2005)	✓	✓	✓	Low	✓	?	?	Unclear
Stafford et al. (2007)	✓	✓	✓	Low	✓	n/a	n/a	Low
Thombs et al. (2008)	✗	✓	?	Unclear	?	n/a	n/a	Unclear
Thomson et	?	✓	✓	Unclear	?	n/a	n/a	Unclear

For peer review only

al. (2011)									
Turner et al. (2012)	✓	✓	✓	Low	✓	n/a	n/a	Low	
Van Steenberg-Wijnenburg (2010)	?	✓	✓	Unclear	?	?	?	Unclear	
Zuithoff et al. (2010)	✓	✓	✓	Low	✓	✓	?	Unclear	

✓ = criterion met; ✗ = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

564

Table 3: Quality assessment of included studies in the algorithm meta-analysis (Manea et al., 2014) (continued)

Study	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:
	Reference test correctly classifies target condition	Reference test interpreted blind to PHQ-9	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Interval of two weeks or less	All participants receive same reference test	All participants included in analysis?	Overall risk of bias
Allegiant studies									
Diez-Quevedo et al. (2001)	✓	✓	✓	?	Unclear	✓	✓	✓	Low

1										
2										
3										
4										
5	Gräfe et al.	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
6	(2004)									
7										
8	Lowe et al.	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
9	(2004)									
10										
11	Muramatsu et	✓	✓	✓	✓	Low	✓	✓	?	Unclear
12	al. (2007)									
13										
14	Navines et al.	✓	✓	?	?	Unclear	✓	✓	✓	Low
15	(2012)									
16										
17	Spitzer et al.	✓	✓	n/a	n/a	Low	✓	✓	✗	High
18	(1999)									
19										
20	Thekkumpurath	✓	✓	n/a	n/a	Low	?	✓	✗	High
21	et al. (2010)									
22										
23										
24	Non-allegiant studies									
25										
26	Arroll et al.	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
27	(2010)									
28										
29	Ayalon et al. (✓	?	✓	?	Unclear	?	✓	✓	Unclear
30	2010)									
31										
32	Eack et al.	✓	?	n/a	n/a	Unclear	?	✓	?	Unclear
33	(2006)									
34										
35	Fann et al.	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
36	(2005)									
37										
38	Gelaye et al.	✓	✓	✓	✓	Low	✓	✓	✗	High
39										
40										
41										
42										
43										
44										
45										
46										
47										
48										
49										

(2013)

Gjerdingen et al. (2009)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Henkel et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Hyphantis et al. (2011)	✓	✓	?	?	Unclear	✓	✓	✗	High
Inagaki et al. (2013)	✓	✓	✓	?	Unclear	✓	✓	✗	High
Khamseh et al. (2011)	✓	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. (2008)	✓	✓	?	?	Unclear	?	✓	✗	High
Lotrakul et al. (2008)	✓	✓	✓	✓	Low	?	✓	✗	High
Persoons et al. (2003)	✓	✓	?	?	Unclear	✓	✓	✓	Low
Picardi et al. (2005)	✓	✓	✓	?	Unclear	✓	✓	✗	High
Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Thombs et al.	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low

Table 4: Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis (Moriarty et al., 2015)

(2008)

Thompson et al. (2011)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Turner et al. (2012)	✓	?	n/a	n/a	Unclear	?	✓	✗	High
Van Steenberg-Wijenburg (2010)	✓	✗	?	?	High	✓	✓	✗	High
Zuithoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

✓ = criterion met; ✗ = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

565

566

567

568

569

570

571

572

Study	Patient selection:	Patient selection:	Patient selection:	Patient selection:	Index test:	Index test:	Index test:	Index test:	Index test:
	Consecutive or random sample	Avoid case-control / avoid artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias	PHQ-9 interpreted blind to reference test	Was a threshold pre-specified?	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias
Allegiant studies									
13. Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	✓	✓	Unclear
16. Kroenke et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
22. Navinés et al. (2012)	✓	✓	✓	Low	✓	✓	✓	?	Unclear
29. Thekkumpurath et al. (2010)	✗	✗	✓	High	✓	✓	n/a	n/a	Low
33. Williams et al. (2005)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
Non-allegiant studies									
1. Adewuya et	✓	✓	✗	Unclear	✓	✓	n/a	n/a	Low

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

al. (2006)										
2. Arroll et al. (2010)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	
3. Azah et al. (2005)	✓	✗	?	High	✓	✓	✓	✓	Low	
4. Chagas et al. (2013)	✓	✓	✓	Low	✓	✓	✓	✓	Low	
6. de Lima Osorio et al. (2009)	✓	✗	✓	High	?	✗	n/a	n/a	High	
7. Elderon et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	
8. Fann et al. (2005)	✓	✗	✗	High	✓	✓	n/a	n/a	Low	
9. Fine et al. (2013)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear	
10. Gelaye et al. (2013)	?	✗	?	High	✓	✗	✓	?	High	
11. Gilbody et al.	?	✓	?	Unclear	✓	✓	n/a	n/a	Low	

1
2
3
4
5 (2007)
6
7
8 12. Gjerdingen et
9 al. (2009) ✓ ✓ ✓ Low ? ✓ n/a n/a Unclear
10
11 14. Hyphantis et
12 al. (2011) ✓ ✗ ✓ High ✓ ✓ ? ? Unclear
13
14 15. Khamseh et
15 al. (2011) ✓ ✓ ? Unclear ✓ ✓ ✓ ? Unclear
16
17
18 19. Liu et al.
19 (2011) ✓ ✓ ? Unclear ✓ ✗ ✓ ? High
20
21
22 20. Lotrakul et
23 al. (2008) ✗ ✓ ? Unclear ✓ ✓ ✓ ? Unclear
24
25
26 23. Patel et al.
27 (2008) ✓ ✓ ✓ Low ✓ ✓ ? ? Unclear
28
29
30 24. Phelan et al.
31 (2010) ✗ ✓ ✓ High ✓ ✗ n/a n/a High
32
33
34 25. Rooney et al.
35 (2013) ✓ ✓ ✓ Low ? ✗ n/a n/a High
36
37
38 26. Sherina et al. ✓ ✓ ✗ High ✓ ✓ ✓ ✓ Low
39
40
41
42
43
44
45

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

	(2012)									
27.	Sidebottom et al. (2012)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
28.	Stafford et al. (2007)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
30.	Thombs et al. (2008)	✗	✓	?	High	✓	?	n/a	n/a	Unclear
32.	Watnick et al. (2005)	?	✗	✓	High	✓	✓	n/a	n/a	Low
34.	Wittkamp et al. (2009)	✓	✓	✓	Low	✓	?	n/a	n/a	Unclear
35.	Zhang et al. (2013)	✓	✓	?	Unclear	?	✓	?	?	Unclear
36.	Zuithoff et al. (2010)	✓	✓	✓	Low	✓	✓	✓	?	Unclear

573

Table 4: Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis (Moriarty et al., 2015) (continued)

Study	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:
	Reference test	Reference test	If translated,	If translated, psychometric	Overall risk of	Interval of two	All participants	All participants	Overall risk of

	correctly classifies target condition	interpreted blind to PHQ-9	appropriate translation	properties reported	bias	weeks or less	receive same reference test	included in analysis?	bias
Allegiant studies									
13. Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
16. Kroenke et al. (2011)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
22. Navinés et al. (2012)	✓	✓	?	?	Unclear	✓	✓	✓	Low
29. Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
33. Williams et al. (2005)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
Non-allegiant studies									
1. Adewuya et al. (2006)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
2. Arroll et al.	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear

(2010)

3. Azah et al.
(2005)

✓

✓

✓

✓

Low

✓

✓

✗

High

4. Chagas et al.
(2013)

✓

✓

?

?

Unclear

✓

✓

✗

High

6. de Lima
Osorio et al.
(2009)

✓

?

n/a

n/a

Unclear

?

✓

✓

Unclear

7. Elderon et al.
(2011)

✓

✓

n/a

n/a

Low

✓

✓

✓

Low

8. Fann et al.
(2005)

✓

?

n/a

n/a

Unclear

✓

✓

✗

High

9. Fine et al.
(2013)

✓

?

n/a

n/a

Unclear

?

✓

✓

Unclear

10. Gelaye et
al. (2013)

✓

✓

✓

✓

Low

✓

✓

✗

High

11. Gilbody et
al. (2007)

✓

✓

n/a

n/a

Low

?

✓

✓

Unclear

12. Gjerdingen

✓

?

n/a

n/a

Unclear

✓

✓

✗

High

1
2
3
4
5 et al. (2009)
6

7
8 14. Hyphantis
9 et al. (2011)

✓ ✓ ? ? Unclear ✓ ✓ ✗ High

10
11 15. Khamseh et
12 al. (2011)

✓ ✓ ✓ ? Unclear ✓ ✓ ? Unclear

13
14
15 19. Liu et al.
16 (2011)

✓ ✓ ✓ ✓ Low ✓ ✓ ? Unclear

17
18
19 20. Lotrakul et
20 al. (2008)

✓ ✓ ✓ ✓ Low ? ✓ ✗ High

21
22
23 23. Patel et al.
24 (2008)

✓ ✓ ✓ ? Unclear ? ✓ ✗ High

25
26
27 24. Phelan et
28 al. (2010)

✓ ✓ n/a n/a Low ✓ ✓ ✓ Low

29
30
31 25. Rooney et
32 al. (2013)

✓ ? n/a n/a Unclear ? ✓ ✗ High

33
34
35 26. Sherina et
36 al. (2012)

✓ ✓ ✓ ✓ Low ✓ ✓ ✓ Low

37
38 27. Sidebottom
39

✓ ✓ n/a n/a Low ✓ ✓ ✗ High

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

et al. (2012)

28. Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
30. Thombs et al. (2008)	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
32. Watnick et al. (2005)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
34. Wittkamp et al. (2009)	✓	✓	n/a	n/a	Low	?	✓	✗	High
35. Zhang et al. (2013)	✓	?	✓	✓	Unclear	✗	✓	✗	High
36. Zuithoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

574

575 **Table 5. Pooled estimates of diagnostic properties of the PHQ-9 at cut-off point 10 and using algorithm scoring method in the non-independent vs**
576 **independent studies groups**

577

Settings	No of studies	No of patients	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive likelihood ratio (95% CI)	Pooled negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)	Heterogeneity: I ²
----------	---------------	----------------	----------------------	----------------------	---	---	--------------------------------	-------------------------------

Manea et al, 2014 SR – RA group	7	4,065	0.77 (0.70 – 0.84)	0.94 (0.90 – 0.97)	14.97 (8.39 – 26.71)	0.23 (0.17 - 0.31)	64.40 (34.15 – 121.43)	78.9%
Manea et al, 2014 SR Independent studies	21	9,900	0.48 (0.41 – 0.91)	0.94 (0.91 – 0.95)	8.26 (6.15 – 11.09)	0.54 (0.48 – 0.62)	15.05 (11.03 – 20.52)	68.1%
Moriarty et al., 2015 SR – RA group	5	6,188	0.87 (0.77 – 0.93)	0.87 (0.76 – 0.94)	7.24 (3.74 – 14.03)	0.14 (0.08 - 0.25)	49.31 (25.74 – 94.48)	55.1%
Moriarty et al., 2015 SR Independent studies	26	13,164	0.76 (0.67 – 0.83)	0.88 (0.85 – 0.91)	6.72 (5.06 – 8.92)	0.26 (0.19 - 0.37)	24.96 (14.81 – 42.08)	81.5%

578

579

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

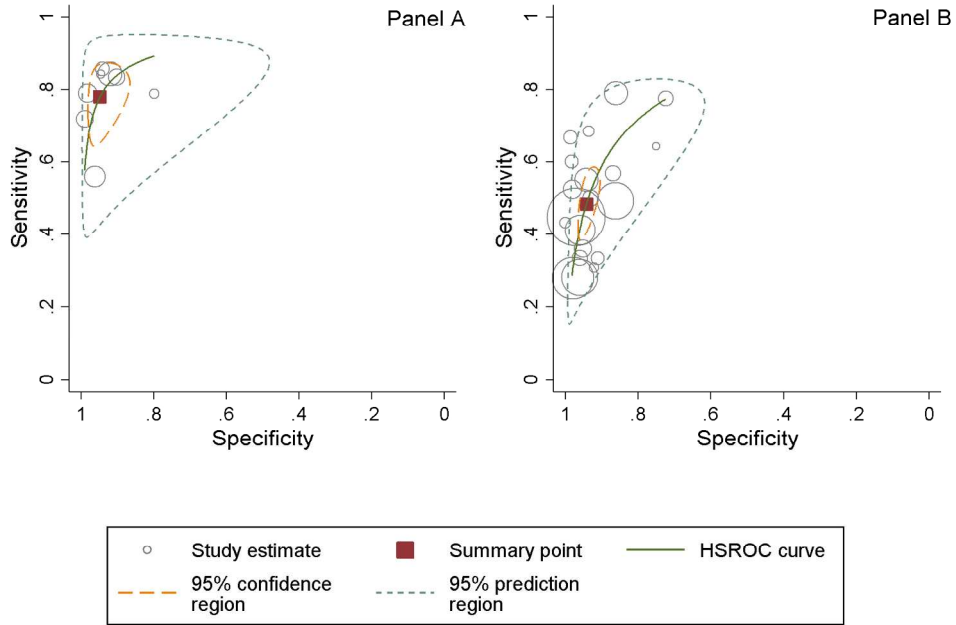
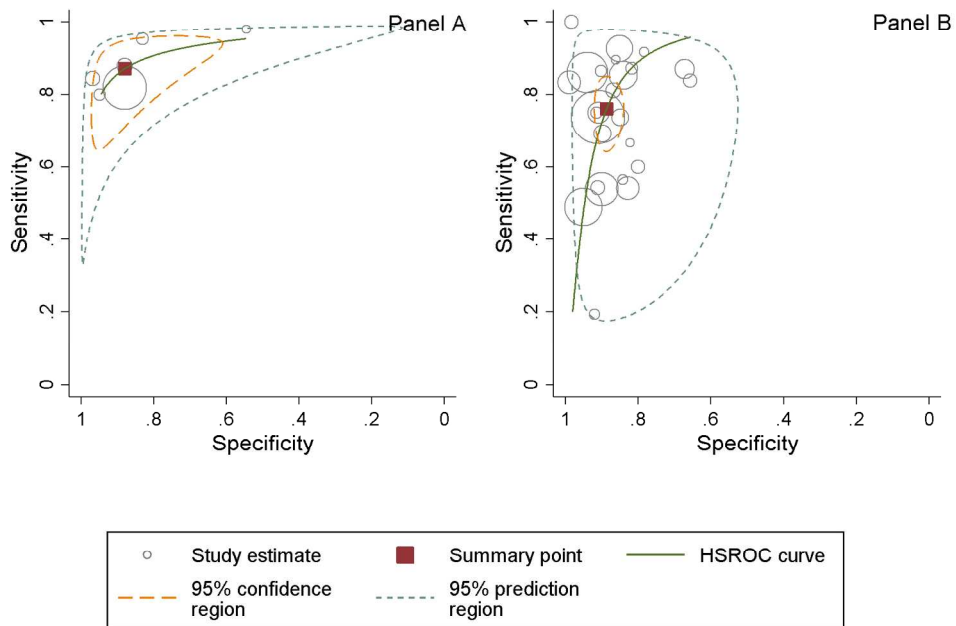


Figure 1. PHQ-9 algorithm scoring method summary ROC plot for the diagnosis of major depressive disorder in allegiant studies (Panel A) and non-allegiant studies (Panel B). Pooled sensitivity and specificity estimates using a bi-variate meta-analysis (HSROC hierarchical receiver-operating characteristic).

169x123mm (300 x 300 DPI)

For peer review only



Caption : Figure 2. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot for diagnosis of major depressive disorder in allegiant studies (panel A) and non-allegiant studies (panel B). Pooled sensitivity and specificity using a bi-variate meta-analysis (HSROC hierarchical receiver-operating characteristic).

169x123mm (300 x 300 DPI)

Appendices to: Manea L, Boehnke JR, Gilbody S, Moriarty AS, McMillan D, Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. Manuscript submitted for publication at BMJOpen.

Appendix 1

Figure 1: PRISMA flowchart - search and selection of included diagnostic accuracy studies for the systematic review of studies reporting diagnostic accuracy of the PHQ-9 at using the summed items scoring method (Manea et al, 2014)

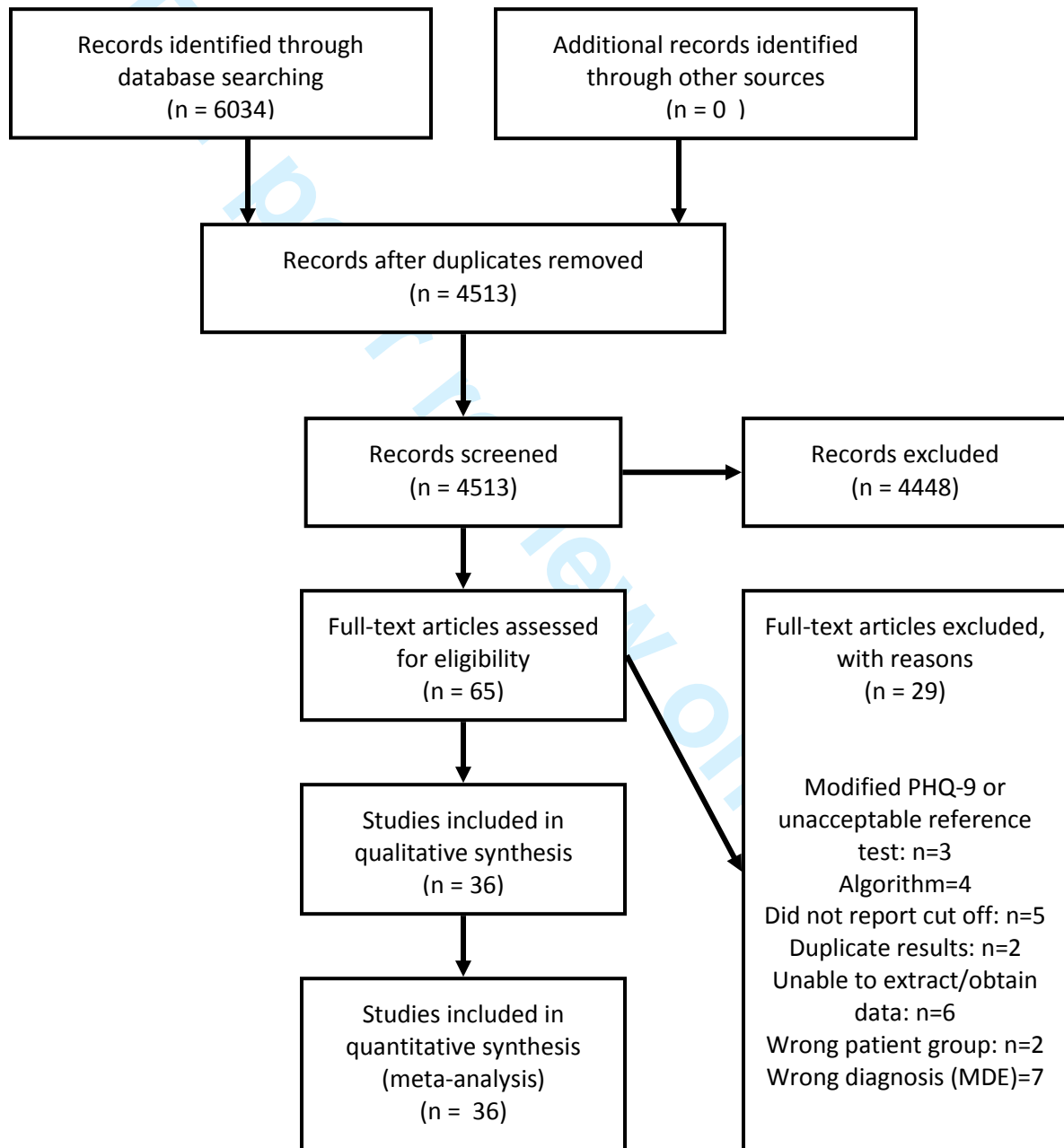
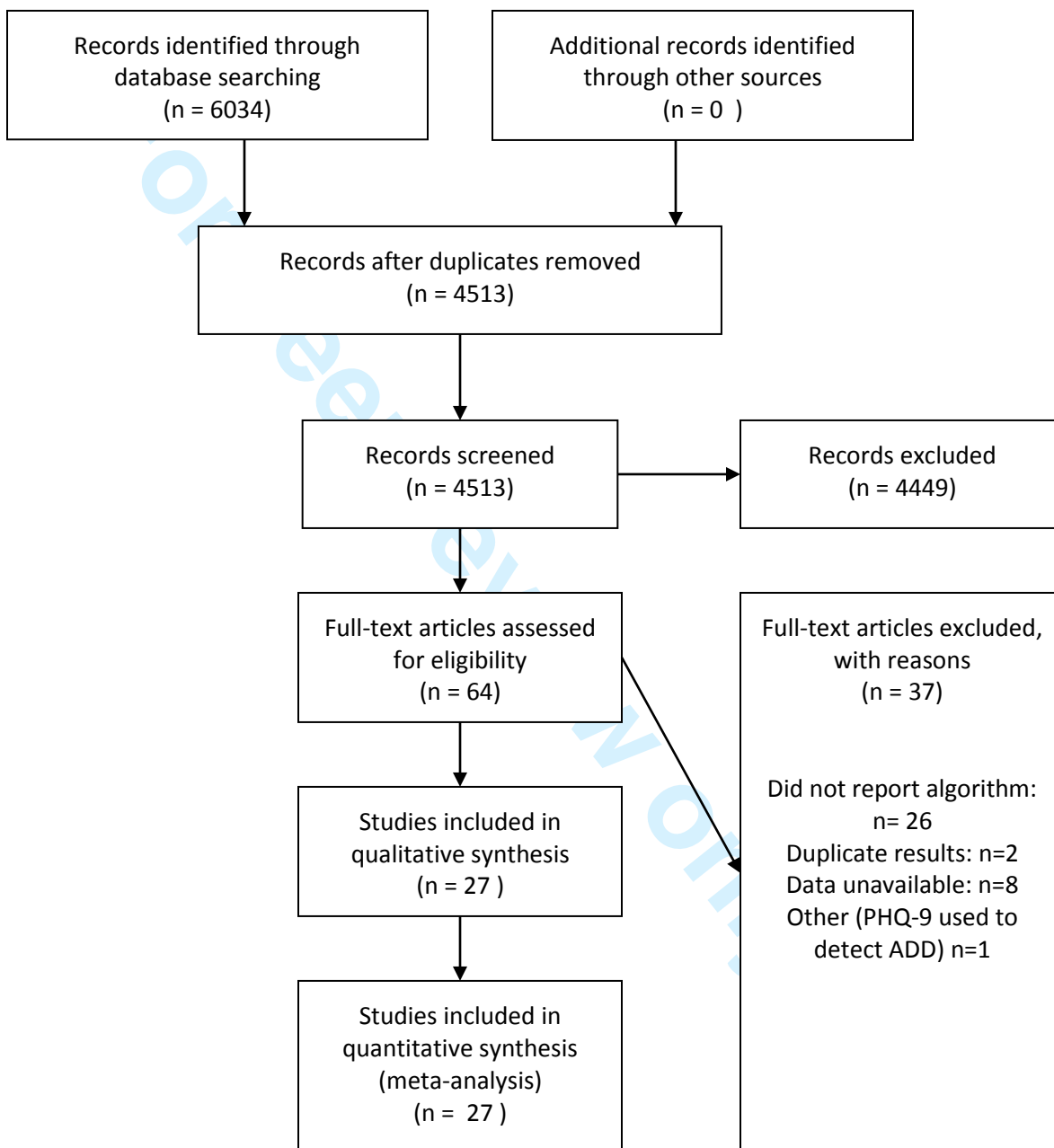


Figure 2: PRISMA flowchart - search and selection of included diagnostic accuracy studies for the systematic review of studies reporting diagnostic accuracy of the PHQ-9 at using the algorithm scoring method (Moriarty et al., 2015)



Appendix 2: Search terms used in Embase, MEDLINE and PsycINFO

(phq adj5 “9”).ti,ab.
(phq adj5 item\$).ti,ab.
(patient health questionnaire adj5 “9”).ti,ab.
(patient health questionnaire adj5 item\$).ti,ab.
(prime md adj5 “9”).ti,ab.
(prime md adj5 item\$).ti,ab.

For peer review only



PRISMA 2009 Checklist

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Available online (see Manea et al., 2015; Moriarty et al., 2015)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5-6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6



PRISMA 2009 Checklist

Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6, 21
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Appendix
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Tables 1 and 2
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Tables 3 and 4
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Table 5
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Tables 3 and 4
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	11 and 17
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	17-21
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	21-22
FUNDING			



PRISMA 2009 Checklist

4	Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	23
---	---------	----	--	----

7 *From:* Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA
8 Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

9 For more information, visit: www.prisma-statement.org.

10 Page 2 of 2

For peer review only

BMJ Open

Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2016-015247.R2
Article Type:	Research
Date Submitted by the Author:	21-Jul-2017
Complete List of Authors:	Manea, Laura; University of York, Health Sciences Boehnke, Jan Rasmus; University of York Gilbody, Simon; The University of York, Department of Health Sciences Moriarty, Andrew; University of York, Health Sciences McMillan, Dean; University of York, Department of Health Sciences
Primary Subject Heading:	Mental health
Secondary Subject Heading:	Diagnostics
Keywords:	Depression & mood disorders < PSYCHIATRY, Screening, PHQ-9, diagnostic meta-analysis, allegiance effect

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

1 **Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis**

2

3 Laura Manea MMedSci MRCPsych*, Jan R. Boehnke PhD, Simon Gilbody DPhil FRCPsych FRSA, Andrew S. Moriarty MRes, Dean
4 McMillan PhD

5

6 *Corresponding Author

7 Hull York Medical School and Department of Health Sciences, ARRC Building, University of York, YO10 5DD

8 Email: laura.manea@york.ac.uk

9

10

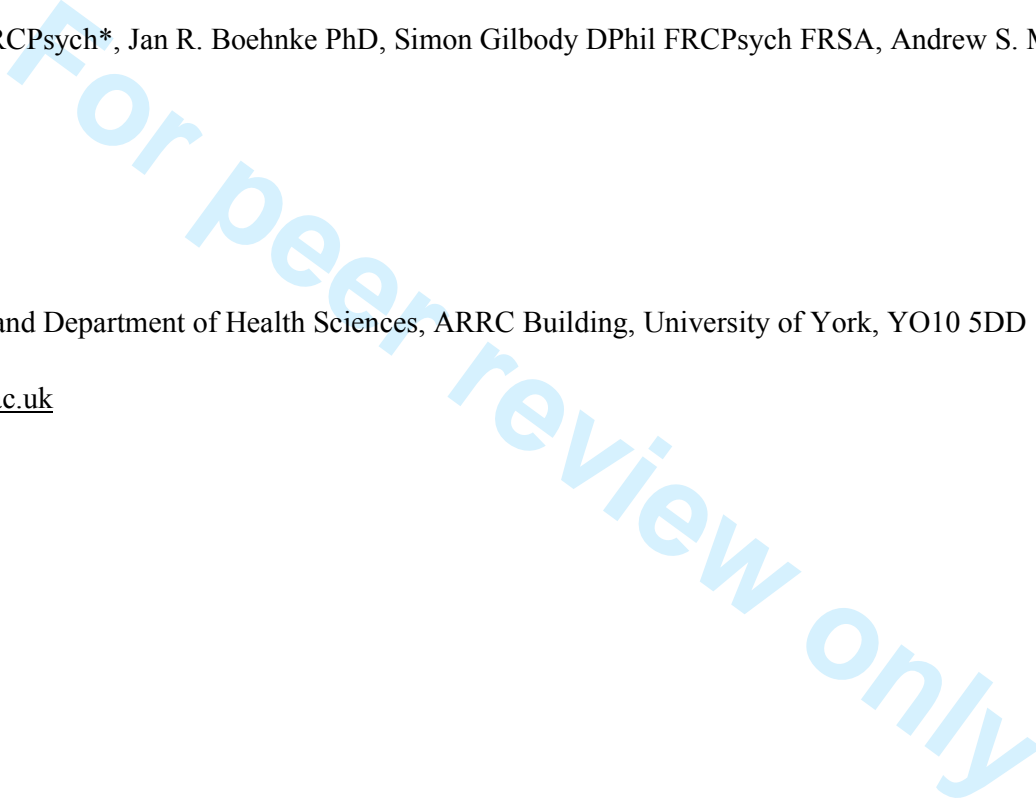
11

12

13

14

15



1
2
3
4
5 166
7 178
9 1810
11 1912
13 2014
15 2116
17 2218
19 23 **Abstract**

20 24 **Objectives** To investigate whether an authorship effect is found that leads to better performance in studies conducted by the original developers
21 25 of the PHQ-9 (allegiant studies).

22 26 **Design** Systematic review with random effects bivariate diagnostic meta-analysis. Search strategies included electronic databases, examination
23 27 of reference lists, and forward citation searches.

24 28 **Inclusion criteria** Included studies provided sufficient data to calculate the diagnostic accuracy of the PHQ-9 against a gold standard diagnosis
25 29 of major depression using the algorithm or the summed item scoring method at cut-off point 10.

26 30 **Data extraction** Descriptive information, methodological quality criteria, and 2×2 contingency tables.

27 31 **Results**

1
2
3
4
5 32 Seven allegiant and twenty independent studies reported the diagnostic performance of the PHQ-9 using the algorithm scoring method. Pooled
6
7 33 diagnostic odds ratio (DOR) for the allegiant group was 64.40, and 15.05 for non-allegiant studies group. The allegiance status was a significant
8
9 34 predictor of DOR variation ($p < 0.0001$).

10
11 35 Five allegiant studies and twenty-six non-allegiant studies reported the performance of the PHQ-9 at recommended cut-off point of 10. Pooled
12
13 36 DOR for the allegiant group was 49.31, and 24.96 for the non-allegiant studies. The allegiance status was a significant predictor of DOR
14
15 37 variation ($P = 0.015$).

16
17 38 Some potential alternative explanations for the observed authorship effect including differences in study characteristics and quality were found,
18
19 39 though it is not clear how some of them account for the observed differences

20
21 40

22 23 41 **Conclusions**

24
25
26 42 Allegiant studies reported better performance of the PHQ-9. Allegiance status was predictive of variation in the DOR. Based on the observed
27
28 43 differences between independent and non-independent studies we were unable to conclude or exclude that allegiance effects are present in
29
30 44 studies examining the diagnostic performance of the PHQ-9. This study highlights the need for future meta-analyses of diagnostic validation
31
32 45 studies of psychological measures to evaluate the impact of researcher allegiance in the primary studies.

33
34 46

35
36 47

37 38 48 **Strengths and limitations of this study**

39
40 49

- 1
2
3
4
5 50 a) An original study—the first meta-analysis of diagnostic validation studies of psychological measures to evaluate the impact of researcher
6 allegiance.
7 51
8 52 b) Using rigorous methodology—strict inclusion/exclusion and quality assessment criteria.
9
10 53 c) We found that the allegiance effect was a significant predictor of the variation of the diagnostic odds ratio in the meta-regression
11 analysis.
12 54
13 55 d) Substantial variability observed in methodological quality of included studies.
14
15 56 e) Based on the observed methodological differences between the independent and non-independent studies we were unable to conclude or
16 exclude that allegiance effects are present in studies examining the diagnostic performance of the PHQ-9.
17 57
18 58
19
20
21 59
22
23 60
24
25
26 61
27
28 62
29
30
31 63
32
33 64
34
35
36 65
37
38 66
39
40
41
42
43
44
45
46
47
48
49

1
2
3
4
5 67 Research on allegiance effects has a long tradition in psychotherapy research. In this context *allegiance* describes the phenomenon that
6
7 68 researchers and clinicians who developed a treatment approach or are for other reasons invested in it tend to find larger effect sizes in favour of
8
9 69 their treatment than for comparison groups. [1] This finding has been extensively replicated [2], [3] and is also robust when the quality of
10
11 70 research is controlled for. Researcher allegiance is subject of on-going debates about the design of efficacy studies as well as implications for
12
13 71 policy. [2], [4], [5] Researcher allegiance is also discussed widely in the literature on experimental as well as evaluation research. [6] Since the
14
15 72 motivational underpinnings of allegiance effects are potentially far more ingrained into human behaviour and decision making than previously
16
17 73 thought (e.g., [7], they may occur commonly in clinical research in general.

18 74 Although it has been suggested that allegiance effects may play a role in the validation of psychological screening and case-finding tools (e.g.,
19
20 75 O'Shea et al., in press), systematic evaluations of this hypothesis are rare and studies that acknowledge potential allegiance effects in such
21
22 76 studies mainly come from forensic psychology and psychiatry backgrounds. [8]–[11] Diagnostic validation studies are geared at establishing the
23
24 77 sensitivity and specificity of a screening or case finding tool, which is used in practice to differentiate cases from non-cases or to decide about
25
26 78 whether further assessment or treatment is indicated or will be offered. An allegiance effect in such studies would be seen in systematically
27
28 79 higher sensitivities or specificities if the original author(s) is (are) part of the team of such a study. Such a bias would have a deleterious affect on
29
30 80 practice through promising over-optimistic accuracy of the screening or case finding tool or in evaluating the cost-effectiveness of the measure
31
32 81 in a screening or case-finding context.

33 82 The depression module of the Patient Health Questionnaire (PHQ-9) is a widely used depression-screening instrument in non-psychiatric
34
35 83 settings. The PHQ-9 was developed by a team of researchers, with its development underwritten by an educational grant from Pfizer US
36
37 84 Pharmaceuticals. [12] The PHQ-9 can be scored using different methods, including an algorithm based on DSM-IV criteria and a cut-off based
38
39 85 on summed-item scores. The psychometric properties of these two approaches have been summarised in two recently published meta-analyses.
40
41 86 [13], [14] The goal of the current review is to investigate, based on an established database of PHQ-9 diagnostic validation studies [13], [14],
42
43
44
45
46
47
48
49

1
2
3
4
5 87 whether an allegiance effect is found that leads to an increased sensitivity and specificity in studies that were conducted by researchers closely
6
7 88 connected to the original developers of the instrument.

8
9 89 METHODS

10
11 90 *Study Selection*

12
13
14 91 Similar search strategies were used in both systematic reviews. (For full details please see Manea et al. (2014) and Moriarty et al. (2015)).
15
16 92 Embase, MEDLine and PSYCHInfo were searched from 1999 (when the PHQ-9 was first developed) to August 2013 [13] and September 2013
17
18 93 [14] respectively, using the terms “PHQ-9”, “PHQ”, “PHQ\$” and “patient health questionnaire”. The search strategy is presented in Appendix 1.
19
20 94 The reference lists of studies fitting the inclusion criteria were manually searched and a reverse citation search in Web of Science was
21
22 95 performed. Authors of unpublished studies were contacted and conference abstracts were reviewed in an attempt to minimise publication bias.

23
24 96 The following inclusion-exclusion criteria were used:

25
26 97 *Population:* Adult population. *Instrument:* Studies that used the PHQ-9. *Comparison (reference standard):* The accuracy of the PHQ-9 had to be
27
28 98 assessed against a recognised gold-standard instrument for the diagnosis of either Diagnostic and Statistical Manual (DSM) or International
29
30 99 Classification of Disease (ICD) criteria for major depression. Studies were included if the diagnoses were made using a standardised diagnostic
31
32 100 structured interview schedule (e.g. Mini International Neuropsychiatric Interview (MINI), Structured Clinical Interview for DSM Disorders
33
34 101 (SCID)). Unguided clinician diagnoses with no reference to a standard structured diagnostic schedule or comparisons of the PHQ-9 with other
35
36 102 self-report measures were excluded. Studies were also excluded if the target diagnosis was not major depressive disorder (MDD, e.g. any
37
38 103 depressive disorder). *Outcome:* Studies had to report sufficient information to calculate a 2*2 contingency table for the algorithm or the
39
40 104 recommended cut-off point 10. *Study design:* Any design. *Additional criterion:* We avoided double counting of evidence by ensuring that only

1
2
3
4
5 105 one study of those that reported overlapping datasets in different journals were included in the meta-analysis. Citations with overlapping samples
6 106 were examined to establish whether they contained information relevant to the research question that was not contained in the included report.

7
8
9 107 *Quality assessment*

10
11 108 Quality assessment was performed using the QUADAS-2 tool, a tool for evaluating the risk of bias and applicability of primary diagnostic
12 109 accuracy studies when conducting diagnostic systematic reviews. [15] It covers the areas of: patient selection, index test, reference standard and
13 110 flow and timing. [16] This tool was adapted for the two reviews and quality assessments were carried out by two independent reviewers for all
14 111 studies included in the reviews.

15
16
17
18
19 112 *Data synthesis and statistical analysis*

20
21 113 We constructed 2x2 tables for cut-off point 10 [14] and the algorithm scoring method [13] Pooled estimates of sensitivity, specificity,
22 114 positive/negative likelihood ratios, and diagnostic odds ratios were calculated using random effects bivariate meta-analysis. [17] Heterogeneity
23 115 was assessed using I^2 for the diagnostic odds ratio, an estimate of the proportion of study variability that is due to between-study variability
24 116 rather than sampling error. We considered values of $\geq 50\%$ to indicate substantial heterogeneity.[18] Summary Receiver Operator Characteristic
25 117 curves (sROC) were constructed using the bivariate model to produce a 95% confidence ellipse within ROC space. [19] Each data point in the
26 118 summary ROC space represents a separate study, unlike a traditional ROC plot, which explores the effect varying thresholds on sensitivity and
27 119 specificity in a single study.

28
29
30 120 We undertook a meta-regression analysis of logit diagnostic odds ratio using research allegiance as covariate in the meta-regression model. [20],
31 121 [21] Analyses were conducted using STATA version 12, with the metan, metandi and metareg user-written commands.

32
33
34
35
36
37
38 122 *Allegiance Rating*

1
2
3
4
5 123 We rated authorship on a paper if any of the developers of the PHQ-9 - Kurt Kroenke, MD, Robert L Spitzer, MD, and Janet B W Williams – as
6 124 an indicator of potential allegiance. We also rated as evidence of allegiance as acknowledged collaborations with the developers of the PHQ-9,
7 125 even if they were not listed as co-authors or if the authors acknowledged funding from Pfizer to conduct the study.
8
9
10

11 126

12 127 RESULTS

13 128

14 129 **Overview of included studies**

15
16 130 31 studies reported the diagnostic properties of the PHQ-9 at cut-off point 10 or above and were included in this analysis. [14] 27 studies were
17 131 included in the algorithm review [13]. The study selection flowcharts can be found in Appendix 2 (figures 1 and 2). The characteristics of these
18 132 studies are reported in tables 1 and 2 and the results of the methodological assessment are presented in tables 3 and 4.
19
20

21 133 **Algorithm scoring method**

22 134

23 135 Descriptive characteristics

24 136 The descriptive characteristics of the included studies are presented in table 1. Seven individual studies that reported the diagnostic performance
25 137 of the PHQ-9 using the algorithm scoring method were co-authored by the original developers of the PHQ-9 [22]–[26], specifically
26 138 acknowledged one of the developers and support by an educational grant from Pfizer US [27], or were co-authored by the first author of a
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

1
2
3
4
5 139 previous study that had also been co-authored by one of the developers [28]. Twenty non-allegiant studies reported the diagnostic properties of
6
7 140 the PHQ-9 using the algorithm scoring method.

8
9 141

10
11 142 Three (43%, 3/7) of the allegiant studies were conducted exclusively in hospital settings [22], [26], [28]. The remaining four studies (67%, 4/7)
12
13 143 were conducted in different settings or non-exclusively hospital settings: one in primary care [25] and three in mixed settings: psycho-somatic
14
15 144 walk in clinics and family practices [23]¹, outpatient clinics and family practices [24] and primary care and hospital settings [27]. In the non-
16
17 145 allegiant group, thirteen (65%, 13/20) studies were conducted in hospital settings [29]–[41]. Of the remaining seven studies, six were conducted
18
19 146 in primary care settings [42]–[47] and one in a community sample [48].

20
21 147 In both groups (non-allegiant and allegiant studies), the majority of studies validated a translated version of the PHQ-9. Two of the studies
22
23 148 authored by developers (28%, 2/7) [25], [26], and eight (40%, 8/20) allegiant studies [29], [30], [37]–[40], [42], [48] were conducted in English.

24
25 149 The mean prevalence of major depressive disorder in the group of allegiant studies was 13.4 % (range 6.1% - 29.2%); in the non-allegiant group
26
27 150 it was 15.5% (range 3.9% - 32.4%). The mean age of patients in the PHQ-9 developers group was 45.7; all but one study had a mean age in the
28
29 151 range of 40 to 50 years. In the non-allegiant group the mean age was 54.6 (range 29.3 – 75.0), with almost half (8) of the studies reporting a
30
31 152 mean age of over 60. The percentage of females in the PHQ-9 developers was 56.8% (range 28.6% - 67.8%) and in the non-allegiant group was
32
33 153 59.1 (18% -100%).

34 154
35
36
37
38

39
40
41
42
43
44
45
46
47
48
49
¹ This study provided separate estimates for the two settings in which it was conducted; therefore separate psychometric estimates were generated for each sample for both algorithm scoring method and summed items scoring method at cut-off point 10 (see below).

1
2
3
4
5 155 All allegiant studies used a self-reported PHQ-9, whereas in 7 non-allegiant studies (30%, 6/20) the PHQ-9 was administered by a researcher
6
7 156 [30]–[33], [43], [48]. Apart from Muramatsu et al. (2007) all allegiant studies used the SCID as a gold standard; the non-allegiant studies used a
8
9 157 wider range of gold standards including SCAN, CIDI, MINI, and C-DIS, though the SCID was also frequently used by the independent studies
10
11 158 as well (45%, 9/20 studies).

12
13 159 Four out of the seven allegiant studies (57%) did not include a conflict of interests statement [22], [23], [25], [27]. Also, four (57%) of the
14
15 160 allegiant studies acknowledged funding from Pfizer [23]–[25], [27]. Only one study [27] acknowledged the collaboration with one of the
16
17 161 developers of the PHQ-9.

18
19 162 Of the non-allegiant studies, twelve (60%) did not include a conflict of interests statement [29]–[32], [35]–[37], [39], [44]–[46], [48]. It appears
20
21 163 that newer studies were more likely to include a conflict of interest statement, which may reflect a recent change in reporting. Funding was
22
23 164 acknowledged by most studies (18/20) and most received funding from academic or/and health research institutions. Two studies received
24
25 165 funding from pharmaceutical companies – Lundbeck [43] and Pfizer [35] and one study acknowledged that Pfizer Italia provided the Italian
26
27 166 version of PHQ-9 and gave the authors permission to use it [36].

28 167 Diagnostic test accuracy

29
30 168 Pooled sensitivity and specificity was calculated separately for the non-allegiant and allegiant studies. Pooled sensitivity for the allegiant studies
31
32 169 of the PHQ-9 was 0.77 (95% CI = 0.70 – 0.84), pooled specificity was 0.94 (95% CI = 0.90 – 0.97), and the pooled diagnostic odds ratio was
33
34 170 64.40 (95% CI = 34.15 – 121.43). Heterogeneity was high ($I^2 = 78.9\%$). Figure 1 represents the summary ROCs for this set of studies.
35

36 171

37
38 172
39
40
41
42
43
44
45

1
2
3
4
5 173 -----
6
7
8 174 Figure 1. PHQ-9 algorithm scoring method summary ROC plot for the diagnosis of major depressive disorder in allegiant studies (Panel A) and
9 175 non-allegiant studies (Panel B). Pooled sensitivity and specificity estimates using a bi-variate meta-analysis (*HSROC* hierarchical receiver-
10 176 operating characteristic).
11
12
13 177 -----
14

15
16 178
17
18 179
19
20 180 Pooled sensitivity for the non-allegiant studies was lower compared to the developer authored studies group at 0.48 (95% CI = 0.41 – 0.91),
21 181 pooled specificity was the same at 0.94 (95% CI = 0.91 – 0.95). The pooled diagnostic odds ratio was approximately four times lower at 15.05
22 182 (95% CI = 11.03 – 20.52) (see figure 1). Heterogeneity was substantial at $I^2 = 68.1\%$.
23
24
25

26 183

27
28 184

29
30
31 185 The meta-regression analysis for algorithm studies with non-allegiant status as the predictor of the diagnostic odds ratio showed that non-
32 186 allegiant status was a significant predictor of the diagnostic odds ratio ($p < 0.0001$) and explained a substantial amount of the observed
33 187 heterogeneity (51.5%).
34
35

36
37 188

38
39 189 Quality assessment
40
41
42
43
44
45

1
2
3
4
5 190 The results of the quality assessment using QUADAS-2 are given in table 3 for the studies reporting on the diagnostic performance of the
6
7 191 algorithm scoring method. In the patient selection domain, more non-allegiant studies (65%, 13/20) than allegiant (29%, 2/7) met the criterion
8
9 192 for consecutive referrals. There were no marked differences on the other two criteria in this domain (avoid case-control design, avoid
10 193 inappropriate exclusions). In the index test domain, the proportion of studies reporting that the PHQ-9 was conducted blind to the reference test
11 194 was comparable between the two groups. There were differences in this domain for those studies using a translated version of the test. All non-
12 195 English allegiant studies (5/5) used an appropriately translated version of the PHQ-9; whereas just over a half of the non-allegiant studies
13 196 reported this (55%, 6/11). However, the majority of both sets of studies did not report details of psychometric properties of the translated
14 197 version. For the reference test domain, nearly all studies in both groups were rated as using a reference test that would correctly classify the
15 198 condition. While most allegiant studies reported that the reference test was interpreted blind to the PHQ-9 score (86%, 6/7), this was reported in
16 199 only 60% (12/20) of the non-allegiant studies.

17
18
19
20
21
22 200 The two sets of studies that used translated versions of the reference test were broadly comparable. There was a slight indication that the
23 201 allegiant studies were more likely to use an appropriately translated version of the reference test and report data on the psychometric properties
24 202 of the translated version, though the numbers for the translated comparison are very low. There were, however, some more notable differences
25 203 on the flow and timing domain. Most allegiant studies ensured that the time between the index and reference test was under two weeks (86%,
26 204 6/7) in comparison to 70% (14/20) of the non-allegiant studies. More allegiant studies met the criterion for 'all participants included in the
27 205 analysis' (57%, 4/7) than non-allegiant studies (25%).

28
29
30
31
32
33 206

34
35
36 207 **Summed items scoring method (cut-off point 10 or above)**

37
38 208
39
40
41
42
43
44
45

1
2
3
4
5 209 Descriptive characteristics
6

7
8 210 Table 2 presents the sample characteristics of the thirty-one PHQ-9 validation studies that reported the psychometric properties of the PHQ-9 at
9 211 cut-off point 10 or above. Five of these studies were co-authored by the original developers of the instrument or acknowledged collaboration
10 212 [12], [23], [26], [49] or were co-authored by the first author of a previous study that had also been co-authored by one of the developers [28].
11 213 Twenty-six studies were conducted by independent researchers.
12
13
14

15 214

16
17 215 Three (60%, 3/5) allegiant studies [26], [28], [49] and eleven non-allegiant studies (42%, 11/26) [30]–[32], [34], [37], [38], [50]–[54] were
18 216 conducted in hospital settings.
19
20

21 217

22
23
24 218 Three (60%, 3/5) allegiant studies [12], [26], [49] and thirteen non-allegiant studies (13/26) [30], [37], [38], [42], [48], [51]–[53], [55]–[59], were
25 219 conducted in English.
26
27

28 220

29
30
31 221 The mean prevalence of major depressive disorder in the allegiant group was 13.2% (range 6.1% - 33.5%) and in the non-allegiant group was
32 222 16.1% (range 2.5% - 43.2%). The mean age of patients in the allegiant group studies was 48.1 (range 41.9 -61.0) and in the 26 non-allegiant
33 223 studies that reported these data was 49.1 (range 23.0 – 78.0). The percentage of females in the allegiant studies that reported these data [12],
34 224 [23], [26], [28] was 56.3% (range 28.6% – 67.8%) and in the non-allegiant group was 64.9 % (range 12% -100%).
35
36
37

38 225
39
40
41
42
43
44
45

1
2
3
4
5 226 Three allegiant studies used the self-reported mode of administration and two of them did not specify how the PHQ-9 was administered. In 9
6
7 227 non-allegiant studies (34%, 9/26) the PHQ-9 was administered by the researcher [30]–[32], [48], [56], [58]–[61]. All allegiant studies used SCID
8
9 228 as a gold standard; the non-allegiant studies used a wider range of gold standards including SCAN, CIDI, MINI, CIS-R, C-DIS, though the SCID
10 229 was used in half of the studies (50%, 13/26 studies).

11
12 230 Three allegiant studies (60%) did not include a conflict of interests statement [12], [23], [49]. Two of these studies [12], [23] acknowledged
13 231 funding from Pfizer. None of the allegiant studies acknowledged collaboration or authorship of one of the developers of the PHQ-9.

14
15
16
17 232 Of the non-allegiant studies, thirteen (42%) did not include a conflict of interests statement [30]–[32], [37], [42], [46], [48], [53], [55], [60],
18 233 [62]–[64]. Similar to the algorithm studies, the newer studies were more likely to include a conflict of interest statement. Funding was
19 234 acknowledged by most studies (27/31) and most received funding from academic or/and health research institutions. One study [57]
20 235 acknowledged that the last author involved in the development of one of the instruments (CORE-OM), ‘but does not gain financially from its
21 236 use’. One study [51] acknowledged funding from industry, AHA Pharmaceuticals Roundtable, but stated that ‘the funding organisations had no
22 237 role in the design or conduct of the study, collection, management, analysis or interpretation of data; or preparation, review or approval of the
23 238 manuscript. Fine et al., 2013 disclosed that the last author had financial and consulting interests (Pfizer was not cited as one of them).

24
25
26
27
28
29
30 240 *Diagnostic test accuracy*

31
32 241 Pooled sensitivity of allegiant studies was 0.87 (95% CI = 0.77 – 0.93), pooled specificity was 0.87 (95% CI = 0.76 – 0.94), and the pooled
33 242 diagnostic odds ratio was 49.31 (95% CI = 25.74 – 94.48) – see table 5. Heterogeneity was moderate ($I^2 = 55.1\%$). Figure 2 represents the
34 243 summary ROCs for this group.

35
36
37
38 244

39
40 245

1
2
3
4
5 246

6 -----
7 247 Figure 2. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot for diagnosis of major depressive disorder in allegiant
8 248 studies (panel A) and non-allegiant studies (panel B). Pooled sensitivity and specificity using a bi-variate meta-analysis (*HSROC* hierarchical
9 249 receiver-operating characteristic).
10 -----

11 250

12 -----
13
14 251 Pooled sensitivity of non-allegiant studies was 0.76 (95% CI, 0.67 – 0.83), pooled specificity was 0.88 95% CI (0.85 – 0.91), and the pooled
15 252 diagnostic odds ratio was 24.96 (95% CI 14.81 – 42.08), approximately half that of the allegiant studies (table 2). Heterogeneity was high at $I^2 =$
16
17 253 81.5 %. Figure 2 represents the summary ROCs for this group.
18
19

20 254
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45

1
2
3
4
5 255 The meta-regression for the studies using a cut-off point of 10 or above with allegiance status of the predictor showed that allegiance status was
6 256 a significant predictor of the diagnostic odds ratio ($P = 0.015$) and explained 19.0% of observed heterogeneity.

7
8
9 257

10
11 258 *Quality assessment*

12
13
14 259 The results of the quality assessment using the QUADAS-2 are given in table 4. For the patient selection domain, the two groups of studies were
15 260 broadly comparable on two items (consecutive or random sample, avoid case-control design). However, all allegiant studies were rated as
16 261 avoiding inappropriate exclusions (5/5) in contrast to 58% (15/26) of the non-allegiant studies.

17
18
19
20 262

21
22 263 On the index test domain, there were a number of differences between the two groups of studies. More of the non-allegiant studies (81%, 21/26)
23 264 reported that the PHQ-9 was interpreted blind to the reference test compared to 60% (3/5) of the allegiant studies. All (5/5) allegiant studies were
24 265 rated as pre-specifying the threshold on the PHQ-9 compared to 73% (19/26) of the non-allegiant studies. The two sets of studies were broadly
25 266 comparable in terms of two items from the reference test domain (correctly classify target condition, reference test interpreted blind). Only one
26 267 allegiant study used a translated version of the index test or reference test, so it is not possible to comment on differences between the two sets of
27 268 studies in terms of these items from the index or reference test domains. For the flow and timing domain, the two groups of studies were broadly
28 269 comparable for two of the criteria (interval of two weeks or less, all participants receive same reference test). However, fewer than half of the
29 270 non-allegiant studies met the criterion for 'all participants included in the analysis' (42%, 11/26); whereas all allegiant studies met this criterion.

30
31
32
33
34
35
36 271

37
38
39 272 **Discussion**

1
2
3
4
5 273 This is to our knowledge the first systematic examination of a possible ‘allegiance’ or authorship effect in the validation of screening or case
6
7 274 finding psychological instrument for a common mental health disorder. We reviewed diagnostic validation studies of the PHQ-9, a widely used
8
9 275 depression screening-instrument. We found that allegiant studies reported higher sensitivity paired with similar specificity compared to non-
10
11 276 allegiant studies. When entered as a covariate in meta-regression analyses, allegiance status was predictive of variation in the DOR for both the
12
13 277 algorithm scoring method and the summed-item scoring method at a cut-off point of 10 or above.
14
15 278

16
17 279 Previous research has proposed several possible explanations for the allegiance effect [9]–[11]. One possibility is the advertent bias that may
18
19 280 serve to inflate the performance of a test when evaluated by those who have developed it. However, before concluding that the differences are
20
21 281 due to this, it is important to explore and rule out alternative explanations. First, it is possible that any observed differences are a result of
22
23 282 differences in study characteristics of the two sets of studies (e.g., setting, clinical population). Secondly, differences in the methodological
24
25 283 quality of the studies may also account for any differences. These possibilities are examined below.
26
27 284

28 285 Difference in study characteristics as potential alternative explanations

29
30
31 286 The two sets of studies were broadly comparable in terms of gender and the prevalence of depression, so these variables are unlikely to offer an
32
33 287 explanation for the differences. While there were some indications from both sets of comparisons that the PHQ-9 may have been researcher-
34
35 288 administered more often in the independent studies, it is not immediately clear how this would lead to lowered diagnostic performance.
36
37 289

1
2
3
4
5 290 The diagnostic meta-analyses of the PHQ-9 [13], [14] have shown that the sensitivity and DOR of the PHQ-9 tends to be lower in hospital
6
7 291 settings for both algorithm and summed-item scoring methods. Whilst the fact that proportionally more non-allegiant algorithm studies were
8
9 292 conducted in secondary care could explain the lower sensitivity and DOR values in the algorithm studies, in the studies that reported the cut-off
10
11 293 point of or above this would not be the case as proportionally more allegiant studies were conducted in hospital settings.

12
13 294 Similarly, differences in the proportions of studies using translated versions of the PHQ-9 are also unlikely to offer an obvious explanation of the
14
15 295 difference in diagnostic performance, because in the algorithm set of studies more of the allegiant studies used a translated version of the test,
16
17 296 but the proportions were in the opposite direction for the studies using a cut off of 10 or above. We tested this by carrying out a sensitivity
18
19 297 analysis restricting the sample to English studies and studies with adequate translation. The allegiance effect was still predictive of DOR
20
21 298 variation between allegiance and non-allegiance studies variation in both algorithm ($p = 0.00$) and summed item scoring at cut-off point of 10
22
23 299 meta-analyses ($p = 0.02$).

24
25 300 A similar conclusion is also likely to apply to the age of the samples. There were more older adults studies in the non-allegiant than allegiant
26
27 301 studies in the algorithm comparison. Depression could be more difficult to identify in older adults due to physical co-morbidities that may
28
29 302 present with similar symptomatology to depression and could account for the lower diagnostic performance in the non-allegiant studies.
30
31 303 However, the non-allegiant samples in the studies that reported the psychometric properties at cut-off point 10 or above had younger samples
32
33 304 than the allegiant studies, so this would not support this interpretation.
34

35
36 305
37 306 The SCID was used as the gold standard in nearly all allegiant studies. The fact that some non-allegiant studies used other gold standards could
38
39 307 potentially explain the poorer psychometric properties of the PHQ-9 in these studies. The SCID is often regarded as the most valid of the
40
41 308 available semi-structured interviews used in depression diagnostic validity studies as the reference standard. If we assume that this is the case
42
43 309 and, furthermore, that the PHQ-9 is an accurate method of screening for depression, then the PHQ-9 may be more likely to agree with the SCID
44
45

1
2
3
4
5 310 than other reference standards. However, when we carried out a sensitivity analysis restricting the sample to SCID only studies the allegiance
6
7 311 effect was still predictive of DOR variation between allegiance and non-allegiance studies variation in both algorithm ($p = 0.01$) and summed
8
9 312 item scoring at cut-off point of 10 reviews ($p = 0.02$).

10 313

11 314

12 13 14 15 16 315 Differences in methodological quality as potential alternative explanations

17
18 316 The quality of the studies was evaluated using the QUADAS-2. Although there were several potential methodological differences between the
19
20 317 two groups of studies from the algorithm papers, not all of these offer obvious explanations of the observed differences and some are unlikely as
21
22 318 explanations. For example, more allegiant studies ensured that the reference test was interpreted blind to the index test. This is unlikely to
23
24 319 account for the observed differences, because a lack of blinding is typically associated with artificially increased diagnostic performance, which
25
26 320 is in the opposite direction to the pattern of results observed here. The impact of some other differences is less clear-cut. For example, a higher
27
28 321 number of the non-allegiant studies met the criterion for consecutive referrals. For this to provide an explanation of the of the observed
29
30 322 differences, the non-consecutive nature of the referrals in the studies by those who had developed the PHQ-9 would need to have led to the over-
31
32 323 inclusion of true positives or under-inclusion of false negatives given that these studies tended to report higher sensitivity relative to the non-
33
34 324 allegiant studies (and vice versa for the independent studies). It is not immediately obvious how this would occur. The allegiant studies were
35
36 325 more likely to have met the criterion of 'included all participants in the analysis'. It is possible that the greater loss of participants from the non-
37
38 326 allegiant studies may have artificially reduced the observed diagnostic accuracy, though, again, it is not immediately obvious how this would
39
40 327 have affected the true positive and false negative rates. Although there is not an obvious explanation of how these differences in methodological
41
42 328 quality could account for the observed differences in diagnostic performance, it is important to recognise that they cannot on that basis be ruled
43
44 329 out.

1
2
3
4
5 330

6
7
8 331 There are, however, two differences in methodological quality among the algorithm studies that are clearer potential alternative explanations.
9 332 The higher rate of appropriate translations among the allegiant studies is potentially important, because lower diagnostic estimates may be
10 333 expected from studies that have poorly translated versions of the index test. In the flow and timing domain, more allegiant studies ensured that
11 334 there was a less than two-week interval between the index and reference test. This is consistent with lower diagnostic performance in the non-
12 335 allegiant studies: as the interval increases it is likely that depression status may change and this would lead to lower levels of agreement between
13 336 the index test and the reference test.

14
15
16
17
18 337

19
20
21 338 There were also differences on some quality assessment items between the two sets of studies in the summed item scoring method comparison.
22 339 The threshold was reported as pre-specified in all allegiant studies in contrast to approximately three quarters of the non-allegiant studies. On the
23 340 face of it, this is unlikely to explain the observed differences, because the use of a pre-specified cut-off point is likely to be associated with lower
24 341 not higher diagnostic test performance. One possibility, however, is that studies that performed poorly at this cut-off point were less likely to be
25 342 reported by those who had developed the measure. As discussed in more detail in the limitations section, we were unable to explore this
26 343 possibility through the use of formal tests for publication bias.

27
28
29
30
31
32 344

33
34 345 All allegiant studies avoided inappropriate exclusions compared to approximately half of the non-allegiant studies. While this is a potential
35 346 alternative explanation of the differences it is not immediately obvious how this would explain the differences in diagnostic performance
36 347 between the two sets of studies. Fewer than half of the non-allegiant studies met the criterion for 'all participants included in the analysis', in
37 348 contrast to all of the allegiant studies met this criterion, but again this difference should usually work against the inclusive studies, not those

1
2
3
4
5 349 excluding cases. More of the non-allegiant studies reported that the PHQ-9 was interpreted blind to the reference test. This does offer a potential
6
7 350 explanation, because the absence of blinding may artificially inflate diagnostic accuracy.
8

9 351

10
11
12 352 **Limitations**

13
14 353 The results of this review need to be viewed in the light of the limitations of the primary studies that contributed to the review and the review
15
16 354 itself. An important consideration is to establish whether any observed differences between the diagnostic performance of the non-allegiant and
17
18 355 allegiant studies are better accounted for by study characteristic or methodological differences. Caution, however, is needed in interpreting any
19
20 356 differences, because of the small number of allegiant studies in both the algorithm and cut-off 10 or above comparisons. The small number of
21
22 357 allegiant studies also meant that we were also unable to explore the potential role of publication bias in the non-allegiant and allegiant studies. At
23
24 358 least 10 studies are required to use standard methods of examining publication bias, but the number of allegiant studies in both the algorithm and
25
26 359 cut-off 10 or above comparisons were fewer than this. Papers published from August 2013 onwards are not covered in the literature search used
27
28 360 and so it potentially misses some more recent studies that would be eligible for inclusion although it is unlikely that many, if any, new allegiant
29
30 361 studies have been published since.
31

32 362

33 363

34 364

35
36
37 365 **Conclusions and implications for further research.**
38
39
40
41
42
43
44
45

1
2
3
4
5 366 The aims of the review was to investigate whether an allegiance effect is found that leads to an increased diagnostic performance in diagnostic
6
7 367 validation studies that were conducted by teams connected to the original developers of the PHQ-9. Our analyses showed that diagnostic studies
8
9 368 conducted by independent/non-allegiant researchers had lower sensitivity paired with similar specificity compared to studies that were classified
10 369 as allegiant. This conclusion held for both the algorithm and cut-off 10 or above studies. We explored a range of possible alternative
11 370 explanations for the observed allegiance effect including both differences in study characteristics and study quality. A number of potential
12 371 differences were found, though for some of these it is not clear how they would necessarily account for the observed differences. However, there
13
14 372 were a number of differences that offered potential alternative explanations unconnected to allegiance effects. In the algorithm studies, the
15 373 studies rated as allegiant were also more likely to use an appropriate translation of the PHQ-9 and were also more likely to ensure that the index
16
17 374 and reference test were conducted within two weeks of each other, both of which may be associated with an improvement in observed diagnostic
18
19 375 performance of an instrument. The majority of studies in both meta-analyses did not provide clear statements about potential conflict of interest
20
21 376 and/or funding, however the newer studies were more likely to provide such statements, which may reflect increasing transparency in this area of
22
23 377 research.
24
25
26 378

27
28 379 We cannot, therefore, conclude that allegiance effects are present in studies examining the diagnostic performance of the PHQ-9; but nor can we
29
30 380 rule them out. Conflicts of interest are an important area of investigation in medical and behavioural research, particularly due to concerns about
31
32 381 trial results being influenced by industry sponsorship. Future diagnostic validity in this area should as a matter of routine present clear statements
33
34 382 about potential conflicts of interest and funding, particularly relating to the development of the instrument under evaluation. Future meta-
35
36 383 analyses of diagnostic validation studies of psychological measures should routinely evaluate the impact of researcher allegiance in the primary
37
38 384 studies examined in the meta-analysis.
39
40 385

1
2
3
4
5 386
6
7

8 387 **Contributors** LM led on all stages of the review and is the guarantor. We used an established database of diagnostic validation studies of the
9 388 PHQ-9 [13], [14] SG provided expert advice on methodology and approaches to assessment of the evidence base. AM carried out the literature
10 389 searches, screened the studies, extracted data and assessed the quality of the included studies for one of the systematic reviews (Moriarty et al.,
11 390 2015) . LM carried out the literature searches, screened the studies, extracted data and assessed the quality of the included studies for the other
12 391 systematic review (Manea et al., 2015), analysed the data for both systematic reviews and drafted the report. JB was involved in the development
13 392 of the study, wrote the introduction section of the review and contributed to the production of the final report. DM supervised the quality
14 393 assessment, methodology and approaches to evidence synthesis, provided senior advice and support throughout and contributed to the
15 394 production of the final report. All parties were involved in drafting and/or commenting on the report.
16
17
18
19
20

21
22 395
23

24 396 **Competing interests** None declared.
25

26
27 397 **Funding** LM was an NIHR Clinical Lecturer when this research was carried out. The NIHR had no role in the study design, methods, data
28 398 collection, analysis or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication.
29
30

31 399
32

33 400 **Provenance and peer review** Not commissioned; externally peer reviewed.
34
35

36 401
37

38 402 **Data sharing statement** No additional data are available.
39
40
41
42
43
44
45

403 REFERENCES

- 404 [1] L. Luborsky, L. Diguer, D. A. Seligman, R. Rosenthal, E. D. Krause, S. Johnson, G. Halperin, M. Bishop, J. S. Berman, and E.
405 Schweizer, "The Researcher's Own Therapy Allegiances: A 'Wild Card' in Comparisons of Treatment Efficacy," *Clin. Psychol. Sci.*
406 *Pract.*, vol. 6, no. 1, pp. 95–106, May 2006.
- 407 [2] E. Dragioti, I. Dimoliatis, and E. Evangelou, "Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of
408 psychotherapy: a systematic appraisal," *BMJ Open*, vol. 5, no. 6, pp. e007206–e007206, Jun. 2015.
- 409 [3] T. Munder, O. Brüttsch, R. Leonhart, H. Gerger, and J. Barth, "Researcher allegiance in psychotherapy outcome research: An overview of
410 reviews," *Clin. Psychol. Rev.*, vol. 33, no. 4, pp. 501–511, Jun. 2013.
- 411 [4] D. A. Winter, "Editorial." Routledge, 07-May-2010.
- 412 [5] J. McLeod, "Taking allegiance seriously—implications for research policy and practice," *Eur. J. Psychother. Couns.*, May 2010.
- 413 [6] G. L. Staines and C. M. Cleland, "Bias in meta-analytic estimates of the absolute efficacy of psychotherapy.," *Rev. Gen. Psychol.*, vol. 11,
414 no. 4, pp. 329–347, 2007.
- 415 [7] K. D. Markman and E. R. Hirt, "Social Prediction and the 'Allegiance Bias,'" *Soc. Cogn.*, vol. 20, no. 1, pp. 58–86, Feb. 2002.
- 416 [8] G. D. Walters, "The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening version as incrementally
417 valid predictors of recidivism.," *Law Hum. Behav.*, vol. 33, no. 6, pp. 497–505, 2009.
- 418 [9] J. P. Singh, M. Grann, and S. Fazel, "Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis," *PLoS*
419 *One*, vol. 8, no. 9, p. e72484, Sep. 2013.

- 1
2
3
4
5 420 [10] P. R. Blair, D. K. Marcus, and M. T. Boccaccini, "Is There an Allegiance Effect for Assessment Instruments? Actuarial Risk Assessment
6 as an Exemplar," *Clin. Psychol. Sci. Pract.*, vol. 15, no. 4, pp. 346–360, Oct. 2008.
7 421
8
9 422 [11] S. O. Lilienfeld and M. K. Jones, "Allegiance Effects in Assessment: Unresolved Questions, Potential Explanations, and Constructive
10 Remedies," *Clin. Psychol. Sci. Pract.*, vol. 15, no. 4, pp. 361–365, Oct. 2008.
11 423
12
13 424 [12] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: validity of a brief depression severity measure.," *J. Gen. Intern. Med.*, vol.
14 16, no. 9, pp. 606–13, Sep. 2001.
15 425
16
17 426 [13] L. Manea, S. Gilbody, and D. McMillan, "A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring
18 method as a screen for depression," *Gen. Hosp. Psychiatry*, vol. 37, no. 1, pp. 67–75, Jan. 2015.
19 427
20
21 428 [14] A. S. Moriarty, S. Gilbody, D. McMillan, and L. Manea, "Screening and case finding for major depressive disorder using the Patient
22 Health Questionnaire (PHQ-9): a meta-analysis," *Gen. Hosp. Psychiatry*, vol. 37, no. 6, pp. 567–576, Nov. 2015.
23 429
24
25 430 [15] P. F. Whiting, A. W. S. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. G. Leeflang, J. A. C. Sterne, P. M. M.
26 Bossuyt, and QUADAS-2 Group, "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.," *Ann. Intern.
27 Med.*, vol. 155, no. 8, pp. 529–36, Oct. 2011.
28 431
29 432
30
31 433 [16] R. Mann, C. E. Hewitt, and S. M. Gilbody, "Assessing the quality of diagnostic studies using psychometric instruments: applying
32 QUADAS," *Soc. Psychiatry Psychiatr. Epidemiol.*, vol. 44, no. 4, pp. 300–307, Apr. 2009.
33 434
34
35 435 [17] J. B. Reitsma, A. S. Glas, A. W. S. Rutjes, R. J. P. M. Scholten, P. M. Bossuyt, and A. H. Zwinderman, "Bivariate analysis of sensitivity
36 and specificity produces informative summary measures in diagnostic reviews," *J. Clin. Epidemiol.*, vol. 58, no. 10, pp. 982–990, Oct.
37 436
38 437
39
40
41
42
43
44
45
46
47
48
49

- 1
2
3
4
5 438 [18] University of York. NHS Centre for Reviews and Dissemination., *Systematic reviews : CRD's guidance for undertaking reviews in health*
6 439 *care*. CRD, University of York, 2009.
- 7
8
9 440 [19] S. D. Walter, "Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data," *Stat. Med.*, vol. 21, no.
10 441 9, pp. 1237–1256, May 2002.
- 11
12
13 442 [20] J. G. Lijmer, P. M. M. Bossuyt, and S. H. Heisterkamp, "Exploring sources of heterogeneity in systematic reviews of diagnostic tests,"
14 443 *Stat. Med.*, vol. 21, no. 11, pp. 1525–1537, Jun. 2002.
- 15
16
17 444 [21] S. G. Thompson and J. P. T. Higgins, "How should meta-regression analyses be undertaken and interpreted?," *Stat. Med.*, vol. 21, no. 11,
18 445 pp. 1559–1573, Jun. 2002.
- 19
20
21 446 [22] C. Diez-Quevedo, T. Rangil, L. Sanchez-Planell, K. Kroenke, and R. L. Spitzer, "Validation and utility of the patient health questionnaire
22 447 in diagnosing mental disorders in 1003 general hospital Spanish inpatients.," *Psychosom. Med.*, vol. 63, no. 4, pp. 679–86.
- 23
24
25 448 [23] K. Gräfe, S. Zipfel, W. Herzog, and B. Löwe, "Screening psychischer Störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-
26 449 D)," *Diagnostica*, vol. 50, no. 4, pp. 171–181, Oct. 2004.
- 27
28
29 450 [24] B. Löwe, R. L. Spitzer, K. Gräfe, K. Kroenke, A. Quenter, S. Zipfel, C. Buchholz, S. Witte, and W. Herzog, "Comparative validity of
30 451 three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses.," *J. Affect. Disord.*, vol. 78, no. 2, pp. 131–40,
31 452 Feb. 2004.
- 32
33
34 453 [25] R. L. Spitzer, K. Kroenke, and J. B. Williams, "Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study.
35 454 Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire.," *JAMA*, vol. 282, no. 18, pp. 1737–44, Nov. 1999.
- 36
37
38 455 [26] P. Thekkumpurath, J. Walker, I. Butcher, L. Hodges, A. Kleiboer, M. O'Connor, L. Wall, G. Murray, K. Kroenke, and M. Sharpe,
39
40
41
42
43
44
45
46
47
48
49

- 1
2
3
4
5 456 “Screening for major depression in cancer outpatients: the diagnostic accuracy of the 9-item patient health questionnaire.,” *Cancer*, vol.
6 457 117, no. 1, pp. 218–27, Jan. 2011.
- 8
9 458 [27] K. Muramatsu, H. Miyaoka, K. Kamijima, Y. Muramatsu, M. Yoshida, T. Otsubo, and F. Gejyo, “The patient health questionnaire,
10 459 Japanese version: validity according to the mini-international neuropsychiatric interview-plus.,” *Psychol. Rep.*, vol. 101, no. 3 Pt 1, pp.
11 460 952–60, Dec. 2007.
- 14
15 461 [28] R. Navinés, P. Castellví, J. Moreno-España, D. Gimenez, M. Udina, S. Cañizares, C. Diez-Quevedo, M. Valdés, R. Solà, and R. Martín-
16 462 Santos, “Depressive and anxiety disorders in chronic hepatitis C patients: Reliability and validity of the Patient Health Questionnaire,” *J.*
17 463 *Affect. Disord.*, vol. 138, no. 3, pp. 343–351, May 2012.
- 20
21 464 [29] S. M. Eack, C. G. Greeno, and B.-J. Lee, “Limitations of the Patient Health Questionnaire in Identifying Anxiety and Depression: Many
22 465 Cases Are Undetected.,” *Res. Soc. Work Pract.*, vol. 16, no. 6, pp. 625–631, Nov. 2006.
- 24
25 466 [30] J. R. Fann, C. H. Bombardier, S. Dikmen, P. Esselman, C. A. Warmes, E. Pelzer, H. Rau, and N. Temkin, “Validity of the Patient Health
26 467 Questionnaire-9 in assessing depression following traumatic brain injury.,” *J. Head Trauma Rehabil.*, vol. 20, no. 6, pp. 501–11.
- 28
29 468 [31] B. Gelaye, M. A. Williams, S. Lemma, N. Deyessa, Y. Bahretibeb, T. Shibre, D. Wondimagegn, A. Lemenhe, J. R. Fann, A. Vander
30 469 Stoep, and X.-H. H. Andrew Zhou, “Validity of the patient health questionnaire-9 for depression screening and diagnosis in East Africa,”
31 470 *Psychiatry Res.*, vol. 210, no. 2, pp. 653–661, Dec. 2013.
- 34
35 471 [32] T. Hyphantis, K. Kotsis, P. V. Voulgari, N. Tsifetaki, F. Creed, and A. A. Drosos, “Diagnostic accuracy, internal consistency, and
36 472 convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders,”
37 473 *Arthritis Care Res. (Hoboken)*, vol. 63, no. 9, pp. 1313–1321, Sep. 2011.

- 1
2
3
4
5 474 [33] M. Inagaki, T. Ohtsuki, N. Yonemoto, Y. Kawashima, A. Saitoh, Y. Oikawa, M. Kurosawa, K. Muramatsu, T. A. Furukawa, and M.
6 Yamada, "Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural
7 475 hospital: a cross-sectional study.," *Gen. Hosp. Psychiatry*, vol. 35, no. 6, pp. 592–7, Jan. 2013.
8 476
- 9
10
11 477 [34] M. E. Khamseh, H. R. Baradaran, A. Javanbakht, M. Mirghorbani, Z. Yadollahi, and M. Malek, "Comparison of the CES-D and PHQ-9
12 478 depression scales in people with type 2 diabetes in Tehran, Iran," *BMC Psychiatry*, vol. 11, no. 1, p. 61, Dec. 2011.
- 13
14
15 479 [35] P. Persoons, K. Luyckx, C. Desloovere, J. Vandenberghe, and B. Fischler, "Anxiety and mood disorders in otorhinolaryngology
16 480 outpatients presenting with dizziness: validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology.,"
17 481 *Gen. Hosp. Psychiatry*, vol. 25, no. 5, pp. 316–23.
- 18
19
20
21 482 [36] A. Picardi, D. A. Adler, D. Abeni, H. Chang, P. Pasquini, W. H. Rogers, and K. M. Bungay, "Screening for depressive disorders in
22 483 patients with skin diseases: a comparison of three screeners.," *Acta Derm. Venereol.*, vol. 85, no. 5, pp. 414–9, 2005.
- 23
24
25 484 [37] L. Stafford, M. Berk, and H. J. Jackson, "Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to
26 485 screen for depression in patients with coronary artery disease," *Gen. Hosp. Psychiatry*, vol. 29, no. 5, pp. 417–424, Sep. 2007.
- 27
28
29 486 [38] B. D. Thombs, R. C. Ziegelstein, and M. A. Whooley, "Optimizing detection of major depression among patients with coronary artery
30 487 disease using the patient health questionnaire: data from the heart and soul study.," *J. Gen. Intern. Med.*, vol. 23, no. 12, pp. 2014–7, Dec.
31 488 2008.
- 32
33
34
35 489 [39] A. W. Thompson, H. Liu, R. D. Hays, W. J. Katon, R. Rausch, N. Diaz, E. L. Jacob, S. D. Vassar, and B. G. Vickrey, "Diagnostic
36 490 accuracy and agreement across three depression assessment measures for Parkinson's disease," *Parkinsonism Relat. Disord.*, vol. 17, no.
37 491 1, pp. 40–45, Jan. 2011.

- 1
2
3
4
5 492 [40] A. Turner, J. Hambridge, J. White, G. Carter, K. Clover, L. Nelson, and M. Hackett, "Depression screening in stroke: a comparison of
6 alternative measures with the structured diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth edition
7 493 (major depressive episode) as criterion standard.," *Stroke.*, vol. 43, no. 4, pp. 1000–5, Apr. 2012.
8 494
- 9
10
11 495 [41] K. M. van Steenbergen-Weijenburg, L. de Vroege, R. R. Ploeger, J. W. Brals, M. G. Vloedveld, T. F. Veneman, L. Hakkaart-van Roijen,
12 496 F. F. Rutten, A. T. Beekman, and C. M. van der Feltz-Cornelis, "Validation of the PHQ-9 as a screening instrument for depression in
13 497 diabetes patients in specialized outpatient clinics," *BMC Health Serv. Res.*, vol. 10, no. 1, p. 235, Dec. 2010.
- 14
15
16
17 498 [42] B. Arroll, F. Goodyear-Smith, S. Crengle, J. Gunn, N. Kerse, T. Fishman, K. Falloon, and S. Hatcher, "Validation of PHQ-2 and PHQ-9
18 499 to Screen for Major Depression in the Primary Care Population," *Ann. Fam. Med.*, vol. 8, no. 4, pp. 348–353, Jul. 2010.
- 19
20
21 500 [43] L. Ayalon, M. Goldfracht, and P. Bech, "'Do you think you suffer from depression?' Reevaluating the use of a single item question for
22 501 the screening of depression in older primary care patients," *Int. J. Geriatr. Psychiatry*, vol. 25, no. 5, pp. 497–502, May 2010.
- 23
24
25 502 [44] V. Henkel, R. Mergl, R. Kohnen, A.-K. Allgaier, H.-J. Möller, and U. Hegerl, "Use of brief depression screening tools in primary care:
26 503 consideration of heterogeneity in performance in different patient groups," *Gen. Hosp. Psychiatry*, vol. 26, no. 3, pp. 190–198, May 2004.
- 27
28
29 504 [45] F. Lamers, C. C. M. Jonkers, H. Bosma, B. W. J. H. Penninx, J. A. Knottnerus, and J. T. M. van Eijk, "Summed score of the Patient
30 505 Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients," *J. Clin. Epidemiol.*,
31 506 vol. 61, no. 7, pp. 679–687, Jul. 2008.
- 32
33
34
35 507 [46] M. Lotrakul, S. Sumrithe, and R. Saipanish, "Reliability and validity of the Thai version of the PHQ-9," *BMC Psychiatry*, vol. 8, no. 1, p.
36 508 46, Dec. 2008.
- 37
38
39 509 [47] N. P. Zuithoff, Y. Vergouwe, M. King, I. Nazareth, M. J. van Wezep, K. G. Moons, and M. I. Geerlings, "The Patient Health
40
41
42
43
44
45

- 1
2
3
4
5 510 Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study,”
6
7 511 *BMC Fam. Pract.*, vol. 11, no. 1, p. 98, Dec. 2010.
- 8
9 512 [48] D. Gjerdingen, S. Crow, P. McGovern, M. Miner, and B. Center, “Postpartum depression screening at well-child visits: validity of a 2-
10 513 question screen and the PHQ-9,” *Ann. Fam. Med.*, vol. 7, no. 1, pp. 63–70, 2009.
- 11
12 514 [49] L. S. Williams, E. J. Brizendine, L. Plue, T. Bakas, W. Tu, H. Hendrie, and K. Kroenke, “Performance of the PHQ-9 as a screening tool
13 515 for depression after stroke,” *Stroke.*, vol. 36, no. 3, pp. 635–8, Mar. 2005.
- 14
15 516 [50] M. H. N. Chagas, V. Tumas, G. R. Rodrigues, J. P. Machado-de-Sousa, A. S. Filho, J. E. C. Hallak, and J. A. S. Crippa, “Validation and
16 517 internal consistency of Patient Health Questionnaire-9 for major depression in Parkinson’s disease,” *Age Ageing*, vol. 42, no. 5, pp. 645–
17 518 649, Sep. 2013.
- 18
19 519 [51] L. Elderon, K. G. Smolderen, B. Na, and M. A. Whooley, “Accuracy and prognostic value of American Heart Association: recommended
20 520 depression screening in patients with coronary heart disease: data from the Heart and Soul Study,” *Circ. Cardiovasc. Qual. Outcomes*,
21 521 vol. 4, no. 5, pp. 533–40, Sep. 2011.
- 22
23 522 [52] A. G. Rooney, S. McNamara, M. Mackinnon, M. Fraser, R. Rampling, A. Carson, and R. Grant, “Screening for major depressive disorder
24 523 in adults with cerebral glioma: an initial validation of 3 self-report instruments,” *Neuro. Oncol.*, vol. 15, no. 1, pp. 122–129, Jan. 2013.
- 25
26 524 [53] S. Watnick, P.-L. Wang, T. Demadura, and L. Ganzini, “Validation of 2 depression screening tools in dialysis patients,” *Am. J. Kidney*
27 525 *Dis.*, vol. 46, no. 5, pp. 919–24, Nov. 2005.
- 28
29 526 [54] Y. Zhang, R. Ting, M. Lam, J. Lam, H. Nan, R. Yeung, W. Yang, L. Ji, J. Weng, Y.-K. Wing, N. Sartorius, and J. C. N. Chan,
30 527 “Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes,” *J.*

- 1
2
3
4
5 528 *Affect. Disord.*, vol. 151, no. 2, pp. 660–666, Nov. 2013.
- 6
7
8 529 [55] A. O. Adewuya, B. A. Ola, and O. O. Afolabi, “Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression
9 530 amongst Nigerian university students,” *J. Affect. Disord.*, vol. 96, no. 1–2, pp. 89–93, Nov. 2006.
- 11 531 [56] T. H. Fine, A. A. Contractor, M. Tamburrino, J. D. Elhai, M. R. Prescott, G. H. Cohen, E. Shirley, P. K. Chan, T. Goto, R. Slembariski, I.
12 532 Liberzon, S. Galea, and J. R. Calabrese, “Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I
13 533 major depression module,” *J. Affect. Disord.*, vol. 150, no. 3, pp. 1001–1007, Sep. 2013.
- 15 534 [57] S. Gilbody, D. Richards, S. Brealey, and C. Hewitt, “Screening for depression in medical settings with the Patient Health Questionnaire
16 535 (PHQ): A diagnostic meta-analysis,” *J. Gen. Intern. Med.*, vol. 22, no. 11, pp. 1596–1602, Oct. 2007.
- 18 536 [58] E. Phelan, B. Williams, K. Meeker, K. Bonn, J. Frederick, J. LoGerfo, and M. Snowden, “A study of the diagnostic accuracy of the PHQ-
19 537 9 in primary care elderly,” *BMC Fam. Pract.*, vol. 11, no. 1, p. 63, Dec. 2010.
- 21 538 [59] A. C. Sidebottom, P. A. Harrison, A. Godecker, and H. Kim, “Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal
22 539 depression screening,” *Arch. Womens. Ment. Health*, vol. 15, no. 5, pp. 367–374, Oct. 2012.
- 24 540 [60] F. de Lima Osório, A. Vilela Mendes, J. A. Crippa, and S. R. Loureiro, “Study of the Discriminative Validity of the PHQ-9 and PHQ-2 in
25 541 a Sample of Brazilian Women in the Context of Primary Health Care,” *Perspect. Psychiatr. Care*, vol. 45, no. 3, pp. 216–227, Jul. 2009.
- 27 542 [61] V. Patel, R. Araya, N. Chowdhary, M. King, B. Kirkwood, S. Nayak, G. Simon, and H. A. Weiss, “Detecting common mental disorders in
28 543 primary care in India: a comparison of five screening questionnaires,” *Psychol. Med.*, vol. 38, no. 2, Feb. 2008.
- 30 544 [62] M. N. N. Azah, M. E. M. Shah, S. Juwita, I. S. Bahri, W. M. W. M. Rushidi, and Y. M. Jamil, “Validation of the Malay Version Brief
31 545 Patient Health Questionnaire (PHQ-9) among Adult Attending Family Medicine Clinics.,” *Int. Med. J.*, 2005.

- 1
2
3
4
5 546 [63] S.-I. Liu, Z.-T. Yeh, H.-C. Huang, F.-J. Sun, J.-J. Tjung, L.-C. Hwang, Y.-H. Shih, and A. W.-C. Yeh, "Validation of Patient Health
6 Questionnaire for depression screening among primary care patients in Taiwan," *Compr. Psychiatry*, vol. 52, no. 1, pp. 96–101, Jan. 2011.
7 547
8
9 548 [64] K. Wittkamp, H. van Ravesteijn, K. Baas, H. van de Hoogen, A. Schene, P. Bindels, P. Lucassen, E. van de Lisdonk, and H. van Weert,
10 549 "The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary
11 550 care," *Gen. Hosp. Psychiatry*, vol. 31, no. 5, pp. 451–459, Sep. 2009.
12
13
14
15 551
16
17 552
18
19
20 553
21
22 554
23
24
25 555
26
27 556
28
29
30 557
31
32 558
33
34
35 559
36
37 560
38
39
40 561
41
42
43
44
45
46
47
48
49

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

562

Table 1: Descriptive characteristics of algorithm studies (Manea et al., 2014)					a) Conflict of interest (COI) declaration b) Funding c) Relationship with original developers
Study	Sample characteristics (Country, setting, age, sex)	Sample size and % depressed	PHQ-9 characteristics	Diagnostic standard	
Diez-Quevedo et al. (2001)	Country: Spain Setting: Medical and surgical tertiary hospitals Age (yrs): M=43 (SD=14.2) Female: 45.6%	N = 1003 Depressed: 8.2%	Administration: Self-report Language: Spanish	DSM-III-R SCID	a) No COI declaration b) Funding acknowledged (academic institutions) c) Not acknowledged
Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Age (yrs): M = 41.9 (SD = 13.8) Female: 67.8%	N = 528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Language: German Administration: self-report	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) Not acknowledged

Lowe et al. (2004)	Country: Germany Setting: Outpatient clinics and family practices Age (yrs): M = 41.7 (SD = 13.8) Female: 67.1%	N = 501 Depressed: 13.2%	Administration: Self-report Language: German	DSM-IV SCID	a) COI declaration 'This study was supported by unrestricted restricted grants from Pfizer Germany and from the medical faculty of the University of Heidelberg Germany, and there are no COI.' b) Acknowledged funding from Pfizer and academic institution c) Not acknowledged
Muramatsu et al. (2007)	Country: Japan Setting: Primary care and general hospital Age (yrs): M = 43.3 (SD = 16.4) Female: 59.5%	N = 131 Depressed: 28.2%	Administration: Self-report Language: Japanese	DSM-IV MINI	a) No COI declaration b) Acknowledged funding from Pfizer c) Acknowledged one of the developers of the PHQ-9: 'The authors acknowledge Dr R L Spitzer'
Navinés et al. (2012)	Country: Spain Setting: General hospital (patients with chronic HCV) Age (yrs): M = 43.4 (SD = 10.2) Female: 28.6%	N = 500 Depressed: 6.4%	Administration: Self-report Language: Spanish	DSM-IV SCID	a) All authors declared that they had no COI. b) Role of funding source declared c) Not acknowledged

1 2 3 4 5 6 7 8 9 10 11 12 13 14	Spitzer et al. (1999)	Country: US Setting: Primary care Age (yrs): M = 46 (SD = 17.2) Female: 66%	N = 3000 (585 received SCID) Depressed: 10%	Administration: Self-report Language: English	DSM-III-R SCID	a) No COI declaration b) Acknowledged funding from Pfizer. 'Drs Spitzer and Williams receive honoraria and consulting money from Pfizer Inc, which has supported this work.' c) N/A
15 16 17 18 19 20 21 22 23 24	Thekkumpurath et al. (2010)	Country: UK Setting: Hospital (cancer patients) Age (yrs): M = 61 Female: 63%	N = 782 Depressed: 6.3% (of the whole sample)	Administration: Not stated Language: English	DSM-IV SCID	a) COI declaration: 'Supported by Cancer Research UK' b) As in a) c) Not acknowledged
25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49	Ayalon et al. (2010)	Country: Israel Age (yrs): M = 75 (SD = 8.1) Female: 40.5 %	N = 153 Depressed: 3.9 %	Administration: Researcher administered Language: Hebrew	DSM-IV SCID	a) COI declaration: 'The project was funded by an Investigator's Initiated Research Grant from Lundbeck International given to Dr Liat Ayalon. Lundbeck International had no other involvement in the project concept of design or in this paper. Per Bech has occasionally over the past 3 years until August 2008 received funding from and has been speaker or member of advisory boards for pharmaceutical companies with an interest in the drug treatment of affective disorders (Astra-Zeneca, Lilly, H. Lundbeck A/S, Lundbeck Foundation and Organon).' b) Acknowledged funding from Lundbeck International

Eack et al. (2006)	Country: US Setting: Community mental health centers for children Age (yrs): M = 39.20 (SD 9.63) Female: 100%	N = 50 Depressed: 28%	Administration: Self-report Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Fann et al. (2005)	Country: US Setting: Trauma hospital (inpatients with traumatic brain injury) Age (yrs): M = 42 (SD=17.9) Female: 29.1%	N = 135 Depressed: 16.3%	Administration: Telephone-administered Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

<p>Gelaye et al. (2011)</p>	<p>Country: Ethiopia</p> <p>Setting: General hospital</p> <p>Age (yrs): 34.9 (SD=11.6)</p> <p>Female: 63.1 %</p>	<p>N = 363</p> <p>Depressed: 12.6%</p>	<p>Administration: Researcher-administered</p> <p>Language: Amharic</p>	<p>DSM-IV SCAN</p>	<p>a) No COI declaration b) Funding acknowledged (academic /health research institutions)</p>
<p>Gjerdingen et al. (2009)</p>	<p>Country: US</p> <p>Setting: Community</p> <p>Age (yrs): M = 29.3</p> <p>Female: 100%</p>	<p>N = 438</p> <p>Depressed: 4.6%</p>	<p>Administration: Telephone or self-report</p> <p>Language: English</p>	<p>DSM-IV SCID</p>	<p>a) No COI declaration b) Funding acknowledged (academic /health research institutions)</p>
<p>Henkel et al. (2004)</p>	<p>Country: Germany</p> <p>Setting: primary care</p> <p>Age (yrs): not reported</p> <p>Female: 74%</p>	<p>N = 448</p> <p>Depressed: 10%</p>	<p>Administration: self-report</p> <p>Language: German</p>	<p>DSM-IV CIDI</p>	<p>a) No COI declaration b) Funding acknowledged (academic /health research institutions)</p>

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	Hyphantis et al. (2011)	Country: Greece Setting: Hospital – rheumatology patients Age (yrs): M = 54.2 (SD = 13.5) Female: 74%	N = 213 Depressed: 32.4%	Administration: Researcher administered Language: Greek	DSM-IV MINI	a) No COI declaration b) No funding acknowledgement
16 17 18 19 20 21 22 23 24 25 26 27	Inagaki et al. (2013)	Country: Japan Setting: General hospital Age whole sample (yrs): M = 73.5 (SD = 12.3) Female: 59.3%	N = 104 out of 511 received MINI Depressed: 7.4%	Administration: Researcher administered Language: Japanese	DSM-IV MINI	a) COI declaration: ‘The authors declare that they have no competing interests.’ b) Funding acknowledged (academic /health research institutions)
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49	Khamseh et al. (2011)	Country: Iran Setting: Diabetes clinic Age (yrs): M = 56.17 (SD = 9.60) Female: 51.9%	N = 185 Depressed: 43.2%	Administration: Self report Language: Persian	DSM-IV SCID	a) COI declaration: The authors declared no competing interests b) Funding acknowledged (academic /health research institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Lamers et al. (2008)	Country: Netherlands Setting: Primary care (elderly) Age (yrs): M = 71.4 (SD = 6.90) Female: 48.2%	N = 713 Depressed: 10.7%	Administration: Self report Language: Dutch	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Lotrakul et al. (2008)	Country: Thailand Setting: Primary care Age (yrs): M = 45.0 (SD = 14.30) Female: 73.7%	N = 279 Depressed: 6.8%	Administration: Self report Language: Thai	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
Persoons et al. (2003)	Country: Belgium Setting: Hospital (otolaryngology patients) Age (yrs): M = 48.2 (SD = 12.9) Female: 65.6%	N = 268 (97 received MINI) Depressed: 16.5%	Administration: Self-report Language: Dutch	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic /health research institutions) and Pfizer Belgium

Picardi et al. (2005)	Country: Italy Setting: Hospital (dermatology inpatients) Age (yrs): M = 37.5 Female: 56%	N = 141 Depressed: 8.5%	Administration: Self-report Language: Italian	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions). Acknowledged Pfizer Italia SRL for providing the Italian version of the PHQ-9 and for permission to use it.
Stafford et al. (2007)	Country: Australia Setting: Hospital (cardiology patients) Age (yrs): M = 64.1 (SD = 10.3) Female: 66%	N = 193 Depressed: 18%	Administration: Self-report Language: English	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
Thombs et al. (2008)	Country: US Setting: Hospital (outpatients with coronary heart disease) Age (yrs): M = 67 (SD = 11) Female: 18%	N = 1024 Depressed: 22%	Administration: Not stated Language: English	DSM C-DIS	a) COI declaration "None disclosed" b) Funding acknowledged (academic/health research institutions)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Thompson et al. (2010)	Country: US Setting: Patients with Parkinson Disease Age (yrs): 72.5 (SD = 9.6) Female: 42%	N = 214 Depressed: 14%	Administration: Self administered Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
Turner et al. (2012)	Country: Australia Setting: Stroke patients Age (yrs): 66.7 (SD = 13.1) Female: 47.2%	N = 72 Depressed: 18%	Administration: Self administered Language: English	DSM-IV SCID	a) COI declaration: Disclosures 'None'. b) Funding acknowledged (academic/health research institutions)

van Steenberg-Weijnenburg (2010)	Country: Netherlands Setting: Diabetes patients Age (yrs): M = 61.8 (SD = 13.6) Female: 48.7%	N = 197 Depressed: 18.8%	Administration: Self administered Language: Dutch	DSM-IV SCID	a) COI declaration: 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions) - 'this had no influence on the content of this article'.
Zuithoff et al. (2010)	Country: Netherlands Setting: Primary care Age (yrs): M = 51 (sd = 16.7) Female: 63%	N = 1338 Depressed: 13%	Administration: Self-report Language: Dutch	DSM-IV CIDI	a) COI declaration 'The authors declare that they have no competing interests.' b) Funding acknowledged (academic/health research institutions).

Table 2: Descriptive characteristics of the summed items scoring method studies cut-off point 10 (Moriarty et al, 2015)

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	Conflict of interest (COI) declaration Funding c) Relationship with original developers
13. Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Mean age: 41.9 (SD = 13.8) Female: 67.8%	N = 528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Administration: self-report Language: German Cut-offs: 10 to 14	DSM-IV SCID	No COI declaration Acknowledged funding from Pfizer Not acknowledged
16. Kroenke et al. (2001)	Country: USA Setting: Primary care Mean age: 46 (SD=17) Female: 66%	N = 580 7.1% MDD	Administration: Self-report Language: English Cut-offs: 9 to 15	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) N/A
22. Navinés et al. (2012)	Country: Spain Setting: General hospital (patients with chronic HCV) Mean age: 43.4 (SD = 10.2) Female: 28.6%	N = 500 6.4% MDD	Administration: Self-report Language: Spanish Cut-offs: 10	DSM-IV SCID	a) All authors declared that they had no COI. b) Role of funding source declared c) Not acknowledged
29. Thekkumpurath et al. (2010)	Country: UK Setting: Hospital (cancer patients) Mean age: 61 Female: 63%	N = 782 6.3% MDD (of the whole sample)	Administration: Not stated Language: English Cut-offs: 5 to 10	DSM-IV SCID	c) COI declaration: 'Supported by Cancer Research UK' d) As in a) e) Not acknowledged
33. Williams et al. (2005)	Country: USA	N = 316	Administration: Unclear	DSM-IV SCID	a) No COI declaration b) Funding

563

	Setting: Secondary care (Post-stroke) Mean age: Unclear Female: Unclear	33.5% MDD	Language: English Cut-offs: 10		acknowledged (academic institutions) c) Not acknowledged
1. Adewuya et al. (2006)	Country: Nigeria Setting: community (students) Mean age: 24.8 (15-40) Female: 41.2%	N = 512 2.5% MDD	Administration: Self-report Language: English Cut-offs: 8 to 12	DSM-IV MINI	a) No COI declaration b) No funding declaration
2. Arroll et al. (2010)	Country: New Zealand Setting: Primary care Mean age: 49 (17-99) Female: 61%	N = 2642 6.2% MDD	Administration: Not stated Language: English Cut-offs: 8,10,12,15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic /health research institutions)
3. Azah et al. (2005)	Country: Malaysia Setting: Primary care Mean age: 38.7 (18-79) Female: 61.7%	N = 180 16.6% MDD	Administration: Self-report Language: Malay Cut-offs: 5 to 12	DSM-IV CIDI	b) No COI declaration c) Funding acknowledged (academic /health research institutions)
4. Chagas et al. (2013)	Country: Brazil	N = 84	Administration: self-report	DSM-IV SCID	a) COI declaration "None declared"

	Setting: Secondary care Mean age: Not stated Female: 52.7%	25.5% MDD	Language: Brazilian Cut-offs: 7 to 10		b) Funding acknowledged (academic/health research institutions)
6. de Lima Osorio et al. (2009)	Country: Brazil Setting: Primary care Mean age: Unclear Female: 100%	N = 177 34% MDD	Administration: research assistants Language: Brazilian Portuguese Cut-offs: 10 to 15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)
7. Elderon et al. (2011)	Country: USA Setting: Secondary care Mean age: Unclear Female: 18%	N = 1022 18.3% MDD	Administration: self-report Language: English Cut-offs: 10	C-DIS	a) COI declaration – ‘No disclosures’ b) Funding acknowledged (academic institutions and industry – AHA Pharmaceuticals Roundtable) – ‘The funding organisations had no role in the design or conduct of the study, collection, management, analysis or interpretation of data; or preparation, review or approval of the manuscript.’
8. Fann et al. (2005)	Country: US Setting: Trauma hospital (inpatients with traumatic	N = 135 16.3% MDD	Administration: Telephone-administered Language: English	DSM-IV SCID	b) No COI declaration c) Funding acknowledged (academic

	brain injury) Mean age: 42 (SD=17.9) Female: 29.1%		Cut-offs: 10		institutions)
9. Fine et al. (2013)	Country: USA Setting: Primary care (Ohio Army National Guard) Mean age: 31 (17-60) Female: 12%	N = 498 21.5% MDD	Administration: Telephone-administered Language: English Cut-offs: 10,15	DSM-IV SCID-I	a) COI – last author disclosed financial and consulting interests (Pfizer not one of them). All other authors declared that they have no COI. b) Funding acknowledged – DoD Medical Research. ‘The sponsor had no role in study design, data collection, analysis, interpretation of results, report writing or manuscript submission.
10. Gelaye et al. (2013)	Country: Ethiopia Setting: General hospital Mean age: 34.9 (SD=11.6) Female: 63.1 %	N = 363 12.6% MDD	Administration: Researcher-administered Language: Amharic Cut-offs: 9 to 11	DSM-IV SCAN	c) No COI declaration d) Funding acknowledged (academic /health research institutions)
11. Gilbody et al.	Country: UK	N = 96	Administration: Not	DSM-IV	a) COI declaration –

(2007)	Setting: Primary care Mean age: 42.5 (SD 13.6) Female: 77%	37.5 MDD	stated Language: English Cut-offs: 9 to 13	SCID	last author involved in the development of one of the instruments (CORE-OM), 'but does not gain financially from its use. b) Funding acknowledged (academic /health research institutions)
12. Gjerdingen et al. (2009)	Country: USA Setting: Community Mean age: 29.3 Female: 100%	N = 438 4.6% MDD	Administration: Telephone or self-report Language: English Cut-offs: 10	DSM-IV SCID	c) No COI declaration d) Funding acknowledged (academic /health research institutions)
14. Hyphantis et al. (2011)	Country: Greece Setting: Hospital – rheumatology patients Mean age: 54.2 (SD = 13.5) Female: 74%	N = 213 32.4% MDD	Administration: Researcher administered Language: Greek Cut-offs: 4 to 16	DSM-IV MINI	c) No COI declaration d) No funding acknowledgement
15. Khamseh et al. (2011)	Country: Iran Setting: Outpatient diabetic clinic Mean age: 56.1 (SD=9.6)	N = 185 43.2% MDD	Administration: Self-report Language: Persian Cut-offs: 10,13	DSM-IV SCID	c) COI declaration: The authors declared no competing interests d) Funding acknowledged (academic /health

					research institutions)
19. Liu et al. (2011)	<p>Female: 51.8%</p> <p>Country: Taiwan</p> <p>Setting: Primary care</p> <p>Mean age: Not specified</p> <p>Female: 60.9%</p>	<p>N = 1532</p> <p>3.3% MDD</p>	<p>Administration: Self-report</p> <p>Language: Chinese version</p> <p>Cut-offs: 9 to 11</p>	SCAN	<p>a) a) No COI declaration</p> <p>b) Funding acknowledged (academic /health research institutions)</p>
20. Lotrakul et al. (2008)	<p>Country: Thailand</p> <p>Setting: Primary care</p> <p>Mean age: 45.0 (SD = 14.30)</p> <p>Female: 73.7%</p>	<p>N = 279</p> <p>6.8% MDD</p>	<p>Administration: Self report</p> <p>Language: Thai</p> <p>Cut-offs: 7 to 15</p>	DSM-IV MINI	<p>c) No COI declaration</p> <p>d) Funding acknowledged (academic /health research institutions)</p>
23. Patel et al. (2008)	<p>Country: India</p> <p>Setting: Primary care</p> <p>Mean age: 37.5 (18-83)</p> <p>Female: 56.4%</p>	<p>N = 299</p> <p>4.3% MDD</p>	<p>Administration: Face-to-face interview</p> <p>Language: Not specified</p> <p>Cut-offs: 7 to 15</p>	CIS-R	<p>a) COI declaration – No Declaration of Interest</p> <p>b) Funding acknowledged (academic /health research institutions)</p>
24. Phelan et al. (2010)	<p>Country: USA</p> <p>Setting: Primary care (elderly)</p> <p>Mean age: 78 (SD=7)</p> <p>Female: 62%</p>	<p>N = 71</p> <p>12% MDD</p>	<p>Administration: Research assistant</p> <p>Language: English</p> <p>Cut-offs: 8 to 12</p>	DSM-IV SCID	<p>a) COI declaration – No competing interests</p> <p>b) Funding acknowledged (academic /health research institutions) . ‘The funder had no role in the study design, methods,</p>

					data collection, analysis or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication.
25. Rooney et al. (2013)	Country: UK Setting: Secondary care (glioma) Mean age: 54.2 (SD=12.3) Female: 42.6%	N = 129 13.5% MDD	Administration: Self-report Language: English Cut-offs: 8 to 11	DSM-IV SCID	a) COI declaration "The authors declare that they have no COI" b) Funding acknowledged (academic/health research institutions)
26. Sherina et al. (2012)	Country: Malaysia Setting: Primary care Mean age: 30.9 (18-81) Female: 100%	N= 146 21.2% MDD	Administration: Self-report Language: Malay Cut-offs: 10	CIDI	a) COI declaration "The authors declare that they have no competing interests" b) Funding acknowledged (academic/health research institutions)
27. Sidebottom et al. (2012)	Country: USA Setting: Community (prenatal) Mean age: 23 (SD=5.5) Female: 100%	N = 745 3.6% MDD	Administration: Interview Language: English Cut-offs: 10	DSM-IV SCID	b) COI declaration "The authors declare that they have no financial COI" b) Funding acknowledged (academic/health research institutions)
28. Stafford et al. (2007)	Country: Australia Setting: Secondary care (cardiac procedures)	N = 193 18.1% MDD	Administration: Self-report Language: English	DSM-IV MINI	b) No COI declaration c) Funding acknowledged (academic/health

	Mean age: 64.14 (38-91) Female: 19.2%		Cut-offs: 10		research institutions)
30. Thombs et al. (2008)	Country: US Setting: Hospital (outpatients with coronary heart disease) Mean age: 67 (SD = 11) Female: 18%	N = 1024 22% MDD	Administration: Not stated Language: English Cut-offs: 7 to 10	DSM C-DIS	b) COI declaration "None disclosed" b) Funding acknowledged (academic/health research institutions)
32. Watnick et al. (2005)	Country: USA Setting: Secondary care (dialysis) Mean age: 63 (SD=15) Female: 32.3%	N = 62 19% MDD	Administration: Self-report Language: English Cut-offs: 10	DSM-IV SCID	b) No COI declaration c) Funding acknowledged (academic/health research institutions)
34. Wittkamp et al. (2009)	Country: Netherlands Setting: Primary care Mean age: 49.8 Female: 66.7%	N = 664 12.3% MDD	Administration: Self-report Language: Not specified Cut-offs: 10 and 15	DSM-IV SCIDI	No COI declaration b) Funding acknowledged (academic/health research institutions)
35. Zhang et al. (2013)	Country: Hong Kong Setting: Secondary care (diabetic outpatients) Mean age: 55.1 (SD=9.5)	N = 99 23.2% MDD	Administration: Self-report Language: Chinese version Cut-offs: 15	DSM-IV MINI	COI declaration – last author acknowledged financial COI. The other authors declare that they have no competing interests.) Funding acknowledged

	Female: 40.8%				(academic/health research institutions)
36. Zuithoff et al. (2010)	Country: Netherlands Setting: Primary care Age (yrs): M = 51 (sd = 16.7) Female: 63%	N = 1338 Depressed: 13%	Administration: Self-report Language: Dutch	DSM-IV CIDI	b) COI declaration "The authors declare that they have no competing interests." b) Funding acknowledged (academic/health research institutions)

564
565

Table 3: Quality assessment of included studies in the algorithm meta-analysis (Manea et al., 2014)

Study	Patient selection:	Patient selection:	Patient selection:	Patient selection:	Index test:	Index test:	Index test:	Index test:
	Consecutive or random sample	Avoid case-control / avoid artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias	PHQ-9 interpreted blind to reference test	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias
Allegiant studies								
Diez-Quevedo et al. (2001)	✗	✓	✗	High	?	✓	✓	Unclear
Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	✓	Unclear

1									
2									
3									
4									
5	Lowe et al.	✘	✓	✓	High	✓	✓	✓	Low
6	(2004)								
7									
8	Muramatsu et	?	✓	?	Unclear	✓	✓	?	Unclear
9	al. (2007)								
10									
11	Navines et al.	✓	✓	✓	Low	✓	✓	?	Unclear
12	(2012)								
13									
14	Spitzer et al.	✘	✓	✓	High	✓	n/a	n/a	Low
15	(1999)								
16									
17	Thekkumpurath	✘	✘	✓	High	✓	n/a	n/a	Low
18	et al. (2010)								
19									
20									
21	Non-allegiant studies								
22									
23	Arroll et al.	✓	✓	✓	Low	✓	n/a	n/a	Low
24	(2010)								
25									
26	Ayalon et al. (?	✓	✓	Unclear	?	✓	?	Unclear
27	2010)								
28									
29	Eack et al.	?	✓	?	Unclear	?	n/a	n/a	Unclear
30	(2006)								
31									
32	Fann et al.	✓	✘	✘	High	✓	n/a	n/a	Low
33	(2005)								
34									
35	Gelaye et al.	?	✘	?	High	✓	✓	?	Unclear
36	(2013)								
37									
38	Gjerdingen et	✓	✓	✓	Low	?	n/a	n/a	Unclear
39									
40									
41									
42									
43									
44									
45									
46									
47									
48									
49									

1									
2									
3									
4									
5	al. (2009)								
6									
7	Henkel et al.	✓	✓	✓	Low	?	n/a	n/a	Unclear
8	(2004)								
9									
10	Hyphantis et al.	✓	✓	✗	High	✓	?	?	Unclear
11	(2011)								
12									
13	Inagaki et al.	✓	✗	✓	High	✓	?	?	Unclear
14	(2013)								
15									
16	Khamseh et al.	✓	✓	?	Unclear	✓	✓	?	Unclear
17	(2011)								
18									
19	Lamers et al	✓	✗	✗	High	✓	?	?	Unclear
20	(2008)								
21									
22	Lotrakul et al.	✗	✓	?	High	✓	✓	?	Unclear
23	(2008)								
24									
25	Persoons et al.	✓	✓	✓	Low	✓	✓	n/a	Low
26	(2003)								
27									
28	Picardi et al.	✓	✓	✓	Low	✓	?	?	Unclear
29	(2005)								
30									
31	Stafford et al.	✓	✓	✓	Low	✓	n/a	n/a	Low
32	(2007)								
33									
34	Thombs et al.	✗	✓	?	Unclear	?	n/a	n/a	Unclear
35	(2008)								
36									
37	Thomson et	?	✓	✓	Unclear	?	n/a	n/a	Unclear
38									
39									
40									
41									
42									
43									
44									
45									
46									
47									
48									
49									

al. (2011)									
Turner et al. (2012)	✓	✓	✓	Low	✓	n/a	n/a	Low	
Van Steenberg-Wijnenburg (2010)	?	✓	✓	Unclear	?	?	?	Unclear	
Zuithoff et al. (2010)	✓	✓	✓	Low	✓	✓	?	Unclear	

✓ = criterion met; ✗ = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

566

Table 3: Quality assessment of included studies in the algorithm meta-analysis (Manea et al., 2014) (continued)

Study	Reference test: Reference test correctly classifies target condition	Reference test: Reference test interpreted blind to PHQ-9	Reference test: If translated, appropriate translation	Reference test: If translated, psychometric properties reported	Reference test: Overall risk of bias	Flow / timing: Interval of two weeks or less	Flow / timing: All participants receive same reference test	Flow / timing: All participants included in analysis?	Flow / timing: Overall risk of bias
Allegiant studies									
Diez-Quevedo et al. (2001)	✓	✓	✓	?	Unclear	✓	✓	✓	Low

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
Lowe et al. (2004)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Muramatsu et al. (2007)	✓	✓	✓	✓	Low	✓	✓	?	Unclear
Navines et al. (2012)	✓	✓	?	?	Unclear	✓	✓	✓	Low
Spitzer et al. (1999)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✗	High
Non-allegiant studies									
Arroll et al. (2010)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Ayalon et al. (2010)	✓	?	✓	?	Unclear	?	✓	✓	Unclear
Eack et al. (2006)	✓	?	n/a	n/a	Unclear	?	✓	?	Unclear
Fann et al. (2005)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Gelaye et al.	✓	✓	✓	✓	Low	✓	✓	✗	High

For peer review only

(2013)

Gjerdingen et al. (2009)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Henkel et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Hyphantis et al. (2011)	✓	✓	?	?	Unclear	✓	✓	✗	High
Inagaki et al. (2013)	✓	✓	✓	?	Unclear	✓	✓	✗	High
Khamseh et al. (2011)	✓	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. (2008)	✓	✓	?	?	Unclear	?	✓	✗	High
Lotrakul et al. (2008)	✓	✓	✓	✓	Low	?	✓	✗	High
Persoons et al. (2003)	✓	✓	?	?	Unclear	✓	✓	✓	Low
Picardi et al. (2005)	✓	✓	✓	?	Unclear	✓	✓	✗	High
Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Thombs et al.	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Table 4: Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis (Moriarty et al., 2015)

(2008)

Thompson et al. (2011)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Turner et al. (2012)	✓	?	n/a	n/a	Unclear	?	✓	✗	High
Van Steenberg-Wijenburg (2010)	✓	✗	?	?	High	✓	✓	✗	High
Zuithoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

✓ = criterion met; ✗ = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

For peer review only

567
568
569
570
571
572
573
574

Study	Patient selection:	Patient selection:	Patient selection:	Patient selection:	Index test:	Index test:	Index test:	Index test:	Index test:
	Consecutive or random sample	Avoid case-control / avoid artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias	PHQ-9 interpreted blind to reference test	Was a threshold pre-specified?	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias
Allegiant studies									
13. Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	✓	✓	Unclear
16. Kroenke et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
22. Navinés et al. (2012)	✓	✓	✓	Low	✓	✓	✓	?	Unclear
29. Thekkumpurath et al. (2010)	✗	✗	✓	High	✓	✓	n/a	n/a	Low
33. Williams et al. (2005)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
Non-allegiant studies									
1. Adewuya et	✓	✓	✗	Unclear	✓	✓	n/a	n/a	Low

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

al. (2006)										
2. Arroll et al. (2010)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	
3. Azah et al. (2005)	✓	✗	?	High	✓	✓	✓	✓	Low	
4. Chagas et al. (2013)	✓	✓	✓	Low	✓	✓	✓	✓	Low	
6. de Lima Osorio et al. (2009)	✓	✗	✓	High	?	✗	n/a	n/a	High	
7. Elderon et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low	
8. Fann et al. (2005)	✓	✗	✗	High	✓	✓	n/a	n/a	Low	
9. Fine et al. (2013)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear	
10. Gelaye et al. (2013)	?	✗	?	High	✓	✗	✓	?	High	
11. Gilbody et al.	?	✓	?	Unclear	✓	✓	n/a	n/a	Low	

1
2
3
4
5 (2007)
6
7
8 12. Gjerdingen et
9 al. (2009) ✓ ✓ ✓ Low ? ✓ n/a n/a Unclear
10
11 14. Hyphantis et
12 al. (2011) ✓ ✗ ✓ High ✓ ✓ ? ? Unclear
13
14 15. Khamseh et
15 al. (2011) ✓ ✓ ? Unclear ✓ ✓ ✓ ? Unclear
16
17
18 19. Liu et al.
19 (2011) ✓ ✓ ? Unclear ✓ ✗ ✓ ? High
20
21
22 20. Lotrakul et
23 al. (2008) ✗ ✓ ? Unclear ✓ ✓ ✓ ? Unclear
24
25
26 23. Patel et al.
27 (2008) ✓ ✓ ✓ Low ✓ ✓ ? ? Unclear
28
29
30 24. Phelan et al.
31 (2010) ✗ ✓ ✓ High ✓ ✗ n/a n/a High
32
33
34 25. Rooney et al.
35 (2013) ✓ ✓ ✓ Low ? ✗ n/a n/a High
36
37
38 26. Sherina et al. ✓ ✓ ✗ High ✓ ✓ ✓ ✓ Low
39
40
41
42
43
44
45

1											
2											
3											
4											
5		(2012)									
6											
7											
8	27.	Sidebottom	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
9		et al. (2012)									
10											
11	28.	Stafford et al.	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
12		(2007)									
13											
14	30.	Thombs et	✗	✓	?	High	✓	?	n/a	n/a	Unclear
15		al. (2008)									
16											
17	32.	Watnick et	?	✗	✓	High	✓	✓	n/a	n/a	Low
18		al. (2005)									
19											
20	34.	Wittkamp	✓	✓	✓	Low	✓	?	n/a	n/a	Unclear
21		et al. (2009)									
22											
23	35.	Zhang et al.	✓	✓	?	Unclear	?	✓	?	?	Unclear
24		(2013)									
25											
26	36.	Zuithoff et	✓	✓	✓	Low	✓	✓	✓	?	Unclear
27		al. (2010)									
28											
29											
30	575										
31											

Table 4: Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis (Moriarty et al., 2015) (continued)

Study	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:
	Reference test	Reference test	If translated,	If translated, psychometric	Overall risk of	Interval of two	All participants	All participants	Overall risk of

	correctly classifies target condition	interpreted blind to PHQ-9	appropriate translation	properties reported	bias	weeks or less	receive same reference test	included in analysis?	bias
Allegiant studies									
13. Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
16. Kroenke et al. (2011)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
22. Navinés et al. (2012)	✓	✓	?	?	Unclear	✓	✓	✓	Low
29. Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
33. Williams et al. (2005)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
Non-allegiant studies									
1. Adewuya et al. (2006)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
2. Arroll et al.	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

(2010)

3. Azah et al.
(2005)

✓ ✓ ✓ ✓ Low ✓ ✓ ✗ High

4. Chagas et al.
(2013)

✓ ✓ ? ? Unclear ✓ ✓ ✗ High

6. de Lima
Osorio et al.
(2009)

✓ ? n/a n/a Unclear ? ✓ ✓ Unclear

7. Elderon et al.
(2011)

✓ ✓ n/a n/a Low ✓ ✓ ✓ Low

8. Fann et al.
(2005)

✓ ? n/a n/a Unclear ✓ ✓ ✗ High

9. Fine et al.
(2013)

✓ ? n/a n/a Unclear ? ✓ ✓ Unclear

10. Gelaye et
al. (2013)

✓ ✓ ✓ ✓ Low ✓ ✓ ✗ High

11. Gilbody et
al. (2007)

✓ ✓ n/a n/a Low ? ✓ ✓ Unclear

12. Gjerdingen

✓ ? n/a n/a Unclear ✓ ✓ ✗ High

For peer review only

1
2
3
4
5 et al. (2009)
6

7
8 14. Hyphantis
9 et al. (2011)

✓ ✓ ? ? Unclear ✓ ✓ ✗ High

10
11 15. Khamseh et
12 al. (2011)

✓ ✓ ✓ ? Unclear ✓ ✓ ? Unclear

13
14
15 19. Liu et al.
16 (2011)

✓ ✓ ✓ ✓ Low ✓ ✓ ? Unclear

17
18
19 20. Lotrakul et
20 al. (2008)

✓ ✓ ✓ ✓ Low ? ✓ ✗ High

21
22
23 23. Patel et al.
24 (2008)

✓ ✓ ✓ ? Unclear ? ✓ ✗ High

25
26
27 24. Phelan et
28 al. (2010)

✓ ✓ n/a n/a Low ✓ ✓ ✓ Low

29
30
31 25. Rooney et
32 al. (2013)

✓ ? n/a n/a Unclear ? ✓ ✗ High

33
34
35 26. Sherina et
36 al. (2012)

✓ ✓ ✓ ✓ Low ✓ ✓ ✓ Low

37
38 27. Sidebottom
39

✓ ✓ n/a n/a Low ✓ ✓ ✗ High

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

et al. (2012)									
28. Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
30. Thombs et al. (2008)	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
32. Watnick et al. (2005)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
34. Wittkampf et al. (2009)	✓	✓	n/a	n/a	Low	?	✓	✗	High
35. Zhang et al. (2013)	✓	?	✓	✓	Unclear	✗	✓	✗	High
36. Zuithoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

576

577 **Table 5. Pooled estimates of diagnostic properties of the PHQ-9 at cut-off point 10 and using algorithm scoring method in the non-independent vs**
578 **independent studies groups**

579

Settings	No of studies	No of patients	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive likelihood ratio (95% CI)	Pooled negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)	Heterogeneity: I ²
----------	---------------	----------------	----------------------	----------------------	---	---	--------------------------------	-------------------------------

Manea et al, 2014 SR – RA group	7	4,065	0.77 (0.70 – 0.84)	0.94 (0.90 – 0.97)	14.97 (8.39 – 26.71)	0.23 (0.17 - 0.31)	64.40 (34.15 – 121.43)	78.9%
Manea et al, 2014 SR Independent studies	21	9,900	0.48 (0.41 – 0.91)	0.94 (0.91 – 0.95)	8.26 (6.15 – 11.09)	0.54 (0.48 – 0.62)	15.05 (11.03 – 20.52)	68.1%
Moriarty et al., 2015 SR – RA group	5	6,188	0.87 (0.77 – 0.93)	0.87 (0.76 – 0.94)	7.24 (3.74 – 14.03)	0.14 (0.08 - 0.25)	49.31 (25.74 – 94.48)	55.1%
Moriarty et al., 2015 SR Independent studies	26	13,164	0.76 (0.67 – 0.83)	0.88 (0.85 – 0.91)	6.72 (5.06 – 8.92)	0.26 (0.19 - 0.37)	24.96 (14.81 – 42.08)	81.5%

580

581 **Figure legends**

582 **Figure 1.** PHQ-9 algorithm scoring method summary ROC plot for the diagnosis of major depressive disorder in allegiant studies and non-
 583 allegiant studies. Pooled sensitivity and specificity estimates using a bi-variate meta-analysis (*HSROC* hierarchical receiver-operating
 584 characteristic).

1
2
3
4
5 585 Figure 2. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot for diagnosis of major depressive disorder in allegiant
6 586 studies and non-allegiant studies. Pooled sensitivity and specificity estimates using a bi-variate meta-analysis (*HSROC* hierarchical receiver-
7 587 operating characteristic).
8
9

10
11 588

12
13 589
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

For peer review only

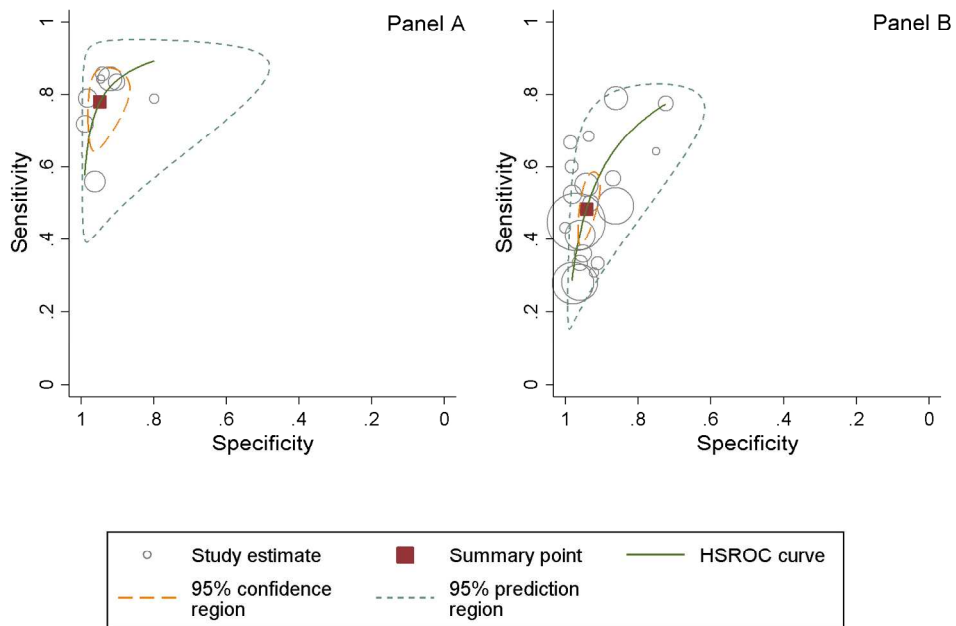
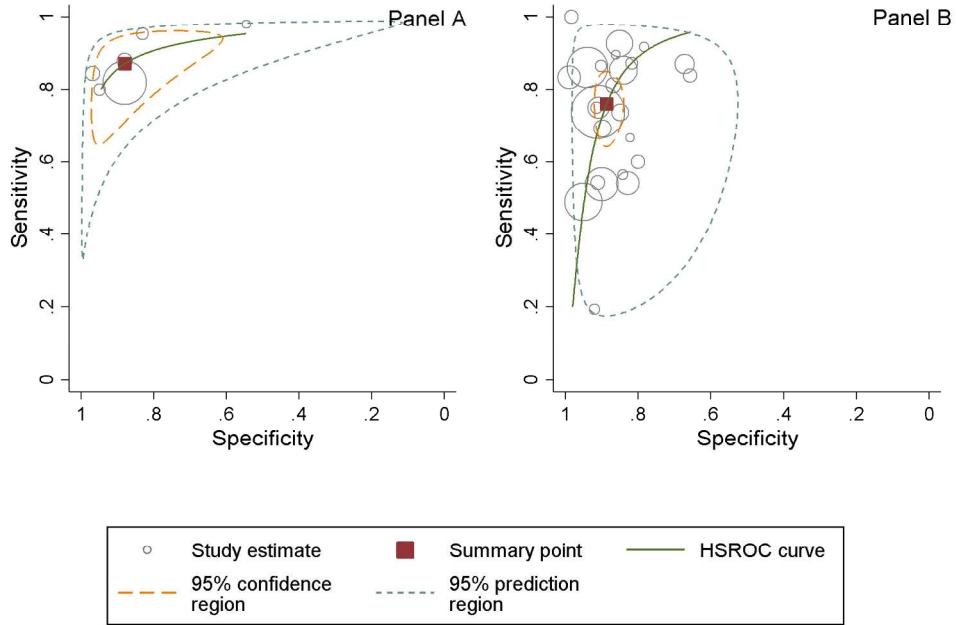


Figure 1. PHQ-9 algorithm scoring method summary ROC plot for the diagnosis of major depressive disorder in allegiant studies (Panel A) and non-allegiant studies (Panel B). Pooled sensitivity and specificity estimates using a bi-variate meta-analysis (HSROC hierarchical receiver-operating characteristic).

169x123mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



Caption : Figure 2. PHQ-9 summed items scoring method at cut-off point 10 summary ROC plot for diagnosis of major depressive disorder in allegiant studies (panel A) and non-allegiant studies (panel B). Pooled sensitivity and specificity using a bi-variate meta-analysis (HSROC hierarchical receiver-operating characteristic).

169x123mm (300 x 300 DPI)

1
2
3 **Appendices to:** Manea L, Boehnke JR, Gilbody S, Moriarty AS, McMillan D, Are there
4 researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic
5 review and meta-analysis. Manuscript submitted for publication at BMJOpen.
6
7
8
9

10
11 **Appendix 1: Search terms used in Embase, MEDLINE and PsycINFO**
12
13

14
15
16 (phq adj5 "9").ti,ab.
17

18
19 (phq adj5 item\$).ti,ab.
20

21 (patient health questionnaire adj5 "9").ti,ab.
22

23
24 (patient health questionnaire adj5 item\$).ti,ab.
25

26 (prime md adj5 "9").ti,ab.
27

28
29 (prime md adj5 item\$).ti,ab.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Appendix 2

Figure 1: PRISMA flowchart - search and selection of included diagnostic accuracy studies for the systematic review of studies reporting diagnostic accuracy of the PHQ-9 at using the summed items scoring method (Manea et al, 2014)

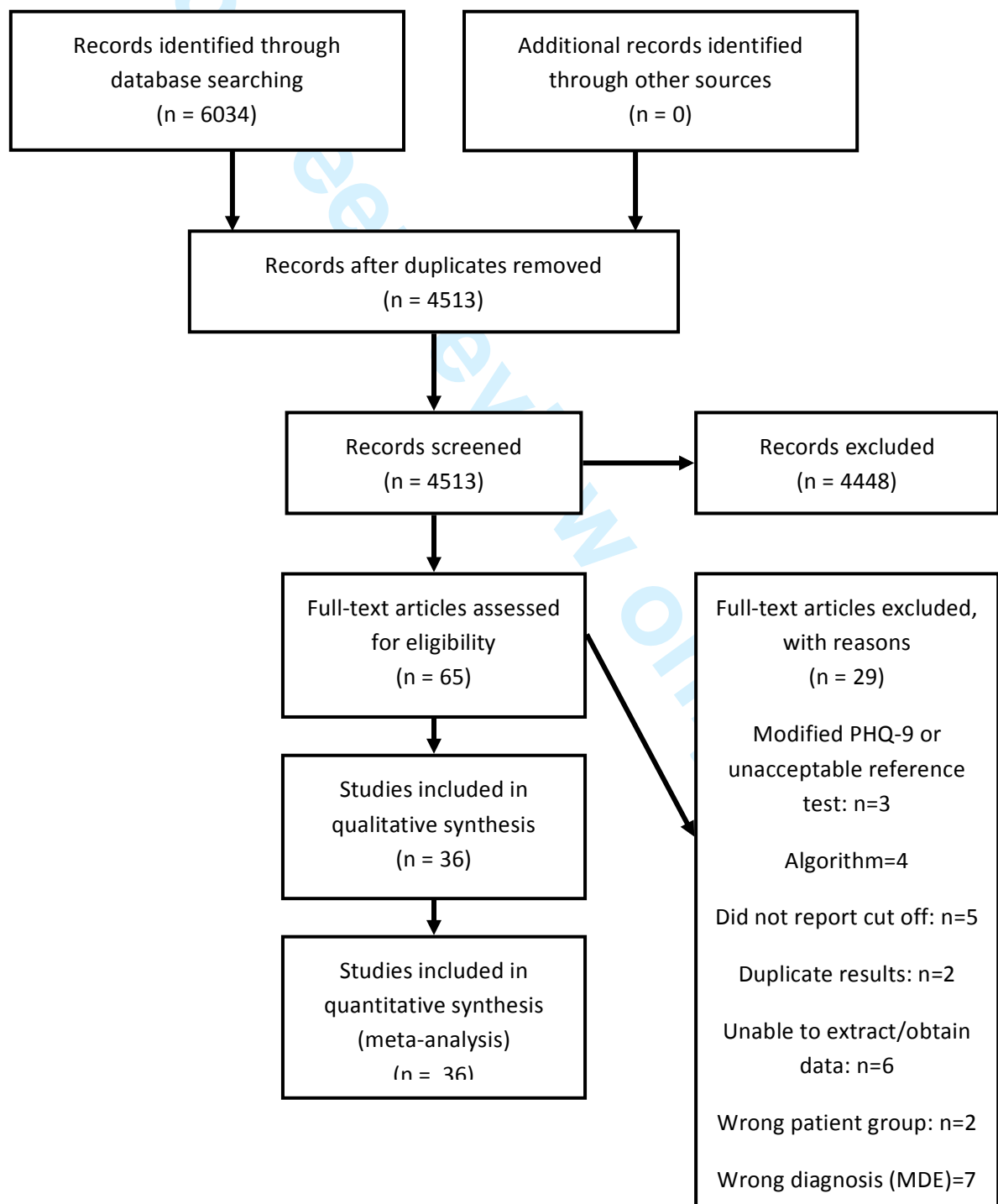
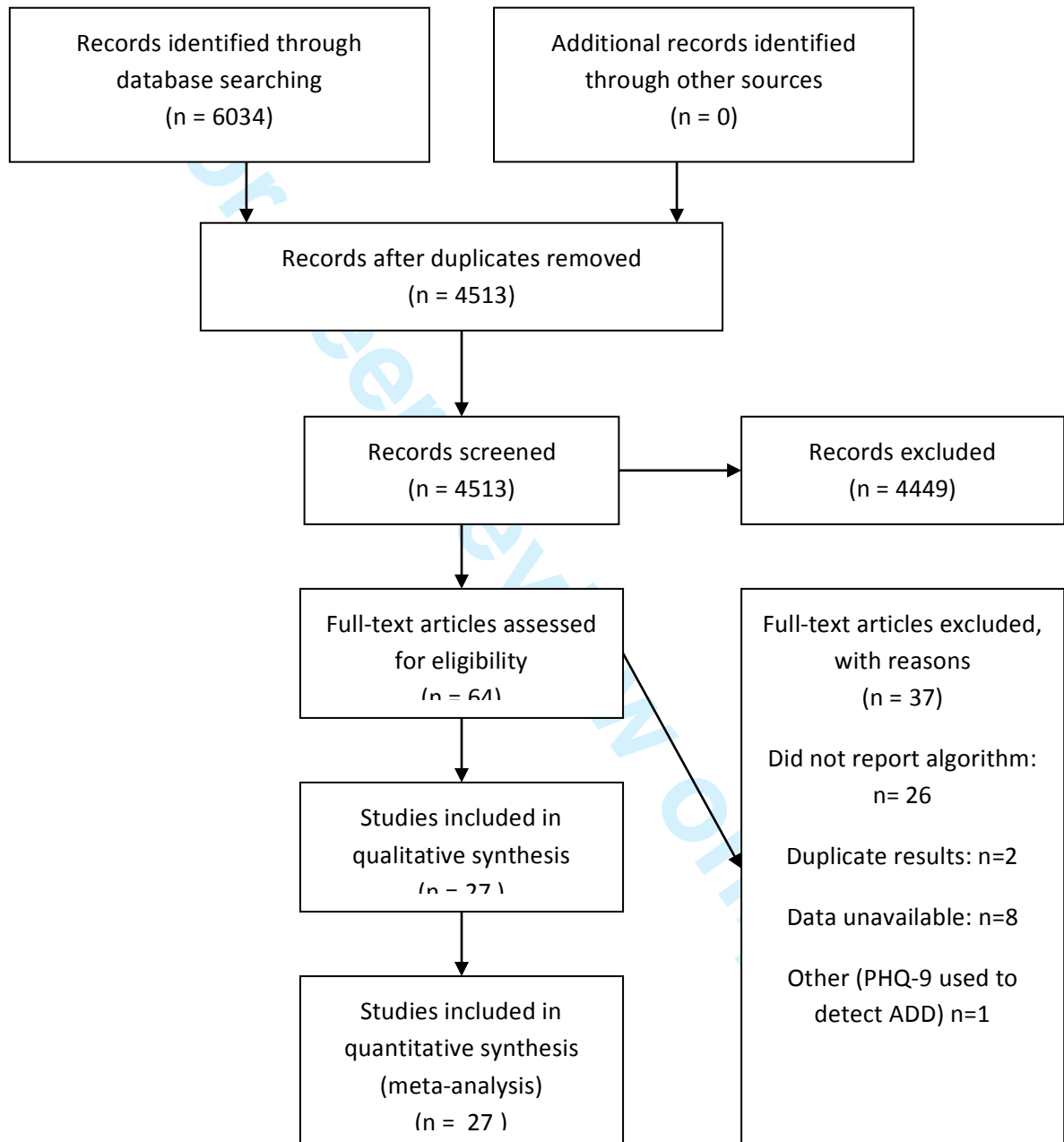


Figure 2: PRISMA flowchart - search and selection of included diagnostic accuracy studies for the systematic review of studies reporting diagnostic accuracy of the PHQ-9 at using the algorithm scoring method (Moriarty et al., 2015)





PRISMA 2009 Checklist

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	No
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	5
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Available online (see Manea et al., 2015; Moriarty et al., 2015)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	5
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	5-6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	6



PRISMA 2009 Checklist

Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	6
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	6

Page 1 of 2

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	6, 21
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	6
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Appendix
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Tables 1 and 2
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Tables 3 and 4
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Table 5
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Tables 3 and 4
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	11 and 17
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	17-21
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	21
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	21-22
FUNDING			



PRISMA 2009 Checklist

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	23
---------	----	--	----

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Page 2 of 2

For peer review only