



eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

SECOND DRAFT GUIDELINE

Meta-analysis of Diagnostic Test Accuracy Studies

Draft version – May 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

The primary objective of EUnetHTA JA2 WP 7 methodology guidelines is to focus on methodological challenges that are encountered by HTA assessors while performing relative effectiveness assessments of pharmaceuticals or non-pharmaceutical health technologies.

As such the guideline represents a consolidated view of non-binding recommendations of EUnetHTA network members and in no case an official opinion of the participating institutions or individuals.

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

This guideline on “Meta-analysis of Diagnostic Test Accuracy Studies” has been developed by HIQA – IRELAND,

with assistance from draft group members from IQWiG – GERMANY.

The guideline was also reviewed and validated by a group of dedicated reviewers from GYEMSZI – HUNGARY, HAS – FRANCE and SBU-SWEDEN.

50	Table of contents	
51		
52		
53		
54	1.1. Definitions of central terms and concepts.....	10
55	1.1.1. Diagnostic test, gold- and reference standards	10
56	1.1.2. Sensitivity and specificity.....	10
57	1.1.3. Likelihood ratios	12
58	1.1.4. Diagnostic odds ratio.....	13
59	1.1.5. Receiver Operating Characteristic (ROC) curves.....	13
60	1.1.6. Predictive values	15
61	1.1.7. Diagnostic accuracy	15
62	1.2. Problem statement	15
63	1.3. Objective(s) and scope of the guideline	17
64	1.4. Related EUnetHTA documents	17
65	2.1. Methods for meta-analysis of diagnostic accuracy studies.....	19
66	2.1.1. Separate random-effects meta-analyses of sensitivity and specificity.....	19
67	2.1.2. Separate meta-analyses of positive and negative likelihood ratios	19
68	2.1.3. Moses-Littenberg summary receiver operating characteristic (SROC) curve	19
69	2.1.4. Hierarchical summary ROC (HSROC) model.....	20
70	2.1.5. Bivariate random-effects meta-analysis for sensitivity and specificity	21
71	2.1.6. Comparison of methods	21
72	2.2. Presentation of results from a meta-analysis of a single diagnostic test	25
73	2.2.1. Tables	25
74	2.2.2. Forest plots for sensitivity and specificity	25
75	2.2.3. Confidence and prediction regions for the summary estimate of sensitivity and	
76	specificity.....	26
77	2.2.4. Summary ROC curve	26
78	2.2.5. Sensitivity analysis	27
79	2.3. Comparison of two diagnostic tests with respect to diagnostic accuracy	
80	(incorporate non-comparative studies in discussion of heterogeneity)	28
81	2.4. Sources of bias.....	28

82	2.4.1. Data gathering and publication bias	28
83	2.4.2. Heterogeneity in meta-analyses of sensitivity and specificity	29
84	2.4.3. Spectrum bias	30
85	2.4.4. Verification/work-up bias and variable gold standard	30
86	2.4.5. Bias resulting from choice of cut-off points.....	30
87	2.4.6. Disease prevalence.....	30
88	2.4.7. Potential for dependence in combined tests.....	31
89	2.4.8. Missing data/non-evaluable results	31
90	2.4.9. Individual patient data analysis	31
91	2.5. Meta-analysis of the prognostic utility of a diagnostic test.....	32
92	2.6. Assessing the quality of studies and meta-analysis	32
93	2.6.1. STARD	32
94	2.6.2. QUADAS	33
95	2.6.3. PRISMA	33
96	2.6.4. GRADE	33
97	2.7. Software	33

98 **Acronyms - Abbreviations**

99	AUC	area under the curve
100	DOR	diagnostic odds ratio
101	FN	false negative
102	FP	false positive
103	FPR	false positive rate
104	GRADE	Grading of Recommendations Applicability, Development and Evaluation
105	HSROC	hierarchical summary receiver operator characteristic
106	IPD	individual patient data
107	LR-	negative likelihood ratio
108	LR+	positive likelihood ratio
109	MESH	medical subject headings
110	NPV	negative predictive value
111	PPV	positive predictive value
112	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
113	QUADAS	Quality Assessment of Diagnostic Accuracy Studies
114	RCT	randomized controlled trial
115	ROC	receiver operator characteristic
116	Sn	sensitivity
117	Sp	specificity
118	SROC	summary receiver operator characteristic
119	STARD	Standards for Reporting of Diagnostic Accuracy
120	TN	true negative
121	TP	true positive
122	TPR	true positive rate

123 **Summary and table with main recommendations**

124 **Definitions**

125 Diagnostic tests are used for a variety of purposes including: to determine whether or not
126 an individual has a particular target condition; to provide information on a physiological or
127 pathological state, congenital abnormality, or on a predisposition to a medical condition or
128 disease; to predict treatment response or reactions; and to define or monitor therapeutic
129 measures. Ideally an evaluation should be undertaken to assess the clinical utility of a test.
130 Such an assessment is generally not supported by appropriately designed studies or by
131 long term outcome data. In the absence of clinical utility data, diagnostic tests are
132 evaluated on the basis of test accuracy: the ability of test to correctly determine the
133 disease status of an individual. A number of metrics are available to describe the
134 characteristics of a diagnostic test, such as the sensitivity, specificity, diagnostic odds
135 ratio, predictive values, likelihood ratios, and the receiver operator characteristic (ROC)
136 curve. Diagnostic tests may also be subject to a threshold effect, whereby the translation
137 of a test result into a dichotomous positive/negative result is not uniform across studies.

138 **Problem statement**

139 Diagnostic test accuracy may be evaluated across a number of studies and, to improve the
140 precision of the estimate, it may be desirable to combine data from a number of studies in
141 a meta-analysis. This guideline reviews available methods for the meta-analysis of
142 diagnostic test accuracy studies that report a dichotomous outcome, and discusses types
143 of bias that are encountered in such meta-analyses.

144 **Methods for meta-analysis of diagnostic test accuracy studies**

145 The hierarchical summary receiver operator characteristic (HSROC) and bivariate random-
146 effects techniques are considered the most appropriate methods for pooling sensitivity and
147 specificity from multiple diagnostic test accuracy studies. Both approaches take into
148 account any correlation that may exist between sensitivity and specificity. The two
149 methods offer equivalent results under certain conditions, such as when no covariates are
150 included. These two methods are considered to be more statistically rigorous than the
151 alternative Moses-Littenberg approach.

152 The most appropriate choice of meta-analytical approach is context specific and also
153 depends on the observed heterogeneity across studies, and the quantity of evidence
154 available. The type of summary data that should be reported depends on whether or not
155 there is a threshold effect. If a threshold effect is present and that the effect explains most
156 of the observed heterogeneity, then a summary ROC curve can be presented.
157 Alternatively, a summary point of sensitivity and specificity with corresponding confidence
158 region should be reported.

159 **Sources of bias**

160 Numerous sources of bias can affect the summary estimate of diagnostic test accuracy:
161 publication bias; heterogeneity; spectrum bias; verification bias; choice of cut-off points for
162 dichotomising a test result. The accuracy reported in studies can also be influenced by
163 underlying disease prevalence, dependence between combined tests, and missing data.

164 When conducting a meta-analysis, potential sources of bias should be identified and
 165 investigated in terms of how they influence the summary estimates of diagnostic test
 166 accuracy. Studies included in a meta-analysis should be appraised in terms of study
 167 quality and whether or not they are sufficiently equivalent to justify a meta-analysis.

168

Recommendations	The recommendation is based on arguments presented in the following publications and / or parts of the guideline text
1. Pooling studies of diagnostic test accuracy should only be undertaken when there are sufficient studies available. When only two studies are available, it is not recommended to undertake a meta-analysis, and reporting should be restricted to a narrative description of the available evidence.	Section 2.1.6
2. The quality of studies being pooled should be assessed using a recognised and validated quality assessment tool.	Section 2.6.2
3. Pooled studies should be equivalent in terms of the index test, the reference standard, the patient population and the indication.	Section 2.1
4. Where important differences are identified across studies in terms of disease spectrum, study setting, or disease prevalence, these should be accounted for by including covariates.	Section 2.4
5. Where potential study differences occur but cannot be readily accounted for, such as verification bias, these should be clearly identified and the potential impacts should be determined.	Section 2.4
6. The appropriate methods of meta-analysis are the hierarchical SROC and bivariate random effects techniques, unless there is an absence of heterogeneity in either the false positive rate or the true positive rate, in which case two separate univariate meta-analyses may be more appropriate.	Section 2.1.6
7. The appropriate approach to meta-analysis is defined with respect to the quantity of data, between-study	Section 2.1.6

heterogeneity, threshold effects, and the correlation between the true positive rate and the false positive rate.	
8. The reporting of meta-analysis should include all the information that justifies the choice of analytical approach and supports the exclusion of alternative approaches.	Section 2.2

169 **1. Introduction**

170 **1.1. Definitions of central terms and concepts**

171 This section describes the main concepts in diagnostic testing in terms of the test itself,
172 and the measures used to describe the accuracy of a test. Test measures may be global
173 (overall test performance) or specific (single aspect of accuracy), and they may be
174 conditional (dependent on prevalence) or unconditional (independent of prevalence).

175 **1.1.1. Diagnostic test, gold- and reference standards**

176 Diagnostic test accuracy studies estimate the ability of a diagnostic test to correctly
177 discriminate between patients with and without a particular target condition. To evaluate
178 the accuracy of a diagnostic test (also called the index test), it must be compared with a
179 reference standard test, sometimes referred to as the gold standard test.¹ A gold standard
180 with perfect discriminatory power between positive and negative status rarely exists.
181 Hence the gold standard is typically replaced by a reference standard that approximates
182 the gold standard as closely as possible.¹ In some cases, there may not be an appropriate
183 reference standard. When analysing test accuracy, the same reference standard should
184 be applied to the whole study population.

185 Test accuracy for a single study population that have been subject to a diagnostic test for
186 a given target condition is generally presented in a 2x2 table indicating the test result (as
187 positive or negative) and the true status with respect of the reference status of those
188 tested (as positive or negative) (see Figure 1).

189
190 Figure 1. The 2x2 table

		True status	
		Positive	Negative
Test result	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

191
192

193 **1.1.2. Sensitivity and specificity**

194 Sensitivity and specificity are the most commonly used measures of diagnostic test
195 performance.²

- 196 • Sensitivity (Sn) – the percentage of people with the target condition that are
197 identified as having the condition by the diagnostic index test

198
199
$$Sn = TP \times 100 / (TP + FN) \tag{1}$$

200

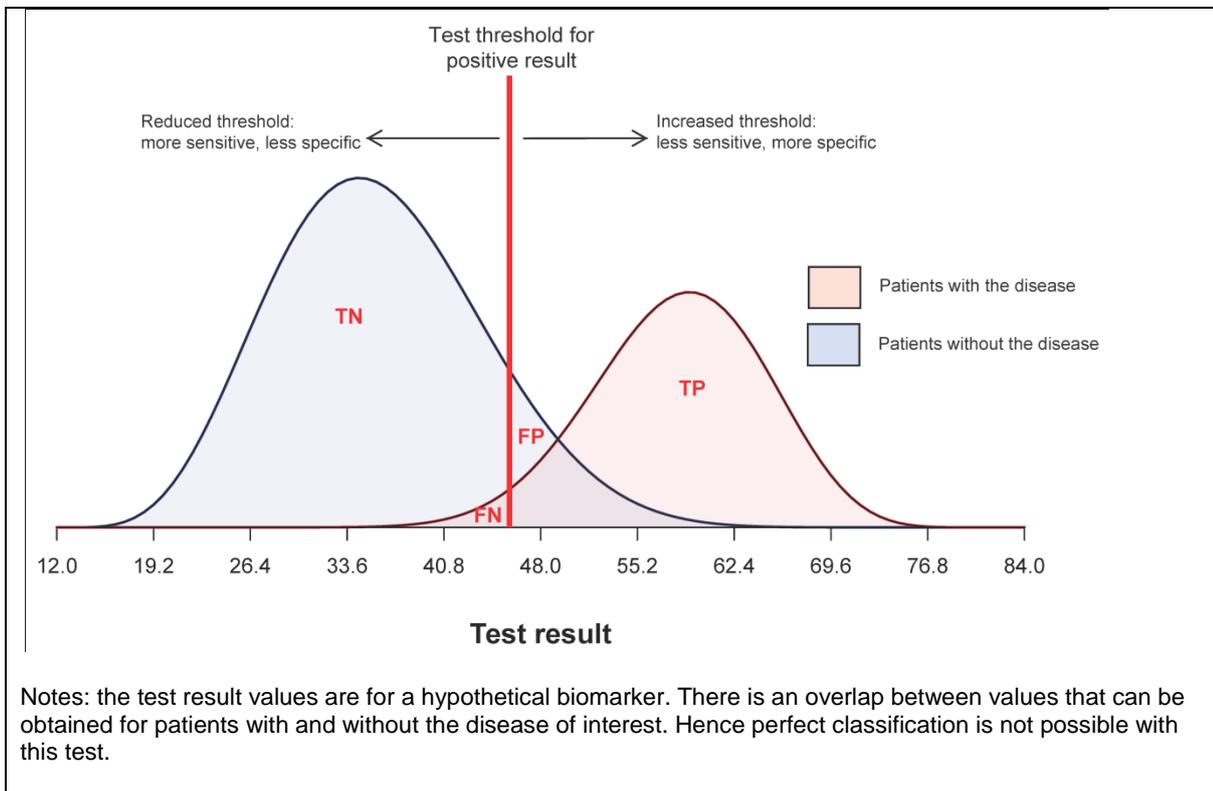
- 201 • Specificity (Sp) – the percentage of people that do not have the target condition
202 that are identified as not having the condition by the diagnostic index test

203
204
205

$$Sp = \frac{TN \times 100}{(TN + FP)} \quad (2)$$

206 A perfect test would have 100% sensitivity and specificity. However, in reality the two
207 measures are almost always negatively correlated, such that increased sensitivity is
208 associated with decreased specificity (see Figure 2). The negative correlation is often a
209 function of the threshold beyond which a test result is considered a positive. For example,
210 an increase threshold will result in fewer false positives (increased specificity) but
211 increased false negatives (reduced sensitivity). Different studies evaluating test accuracy
212 may apply the same test but apply a different threshold for defining a positive test result.
213 Decreasing the threshold decreases specificity but increases sensitivity, while increasing
214 the threshold decreases sensitivity but increases specificity. By varying the threshold for a
215 positive test, a correlation between sensitivity and specificity is observed which is known
216 as a threshold effect.

217
218 Figure 2. Test threshold and impact on diagnostic accuracy



219

220 Sensitivity and specificity are generally assumed to be independent of disease prevalence,
221 although this is not strictly the case (see section 2.4.6). The measures have no clinical
222 meaning and they do not apply to test results that are reported as levels rather than a
223 dichotomous outcome. The sensitivity is also referred to as the true positive rate (TPR),
224 while 1-Sp is referred to as the false positive rate (FPR).

225 Sensitivity and specificity are perhaps the most commonly reported measures of
226 diagnostic test accuracy. As a concept they are relatively simple to understand. They are
227 considered to be specific and unconditional measures of accuracy. However, sensitivity
228 and specificity are summary test characteristics, and do not provide information about a
229 specific patient. In other words, they provide an 'on average' accuracy for a given test.
230 Sensitivity and specificity may be reported as percentages or proportions.

231 It is generally the case that for a test to be useful at ruling out a disease it must have high
232 sensitivity, and for it to be useful at confirming a disease it must have high specificity.³ The
233 Sn-N-Out (high sensitivity, negative, rules out) and Sp-P-In (high specificity, positive, rules
234 in) mnemonics are sometimes used to make quick diagnostic decisions, although these
235 rules are serious simplifications and should be used with caution.⁴ A high sensitivity
236 implies very few false negatives, therefore nearly all patients labelled as negative are
237 correctly assigned. Similarly a high specificity implies very few false positives, meaning
238 that nearly all patients labelled positive are genuinely positive. However, a high specificity
239 combined with a poor sensitivity may not be an informative test as many genuine positives
240 will test negative. Ordinarily a high specificity can be used to rule in positives, but when
241 coupled with a poor sensitivity, few genuine positives will be captured. Hence, care must
242 be taken when interpreting sensitivity and specificity values, and both measures should be
243 reported together to consider the accuracy of a test.

244 Sensitivity and specificity are sometimes presented simultaneously as Youden's Index (J_c),
245 which is intended as a means for optimising the cut-off point (c) for a test:

$$246 \quad J_c = Sn_c + Sp_c - 1 \quad (3)$$

247 The optimal cut-off point, c^* , is the cut-off point at which J_c is maximised.⁵ However, this
248 optimisation is based on the assumption that false-positives and false-negatives are
249 equally undesirable. In reality, the incorrect classifications of healthy and diseased persons
250 may not be considered equally undesirable.⁵ For example, for a life-threatening disease
251 where early detection may significantly improve outcomes, there may be a preference to
252 minimise false-negatives.

253 The interpretation of sensitivity and specificity can be problematic when evaluating tests
254 that are applied repeatedly, such as for continuous monitoring of a patient's status.⁶

255

256 **1.1.3. Likelihood ratios**

257 The likelihood ratio (LR) associated with a positive test result is the probability of a positive
258 finding in patients with the target condition divided by the probability of a positive test result
259 in patients who do not have the target condition. Multiplying the LR by the pre-test odds of
260 having the target condition gives the post-test odds of having the condition. The LR can be
261 expressed for positive and negative test results:

$$262 \quad \text{Likelihood ratio for positive results (LR+)} = Sn / (100 - Sp) \quad (4)$$

$$263 \quad \text{Likelihood ratio for negative results (LR-)} = (100 - Sn) / Sp \quad (5)$$

265 As the likelihood ratios are a function of sensitivity and specificity, it is generally assumed
266 that they do not vary with disease prevalence. Likelihood ratios can be calculated for

267 multiple levels of test result, which can be useful in diagnostic tests for which results are
268 presented on a continuous scale.² Like sensitivity and specificity, these measures are
269 considered to be specific and unconditional measures of accuracy

270

271 By applying the Bayes' theorem, the pre-test probability of disease (e.g., the prevalence of
272 disease) can be converted into a post-test probability using the likelihood ratios in
273 conjunction with the test result. As a rule of thumb, a likelihood ratio of between 0.2 and 5
274 gives no more than weak evidence to rule the disease in or out. A likelihood ratio of
275 between 5 and 10, and between 0.1 and 0.2 gives moderate evidence to rule the disease
276 in or out, respectively. A likelihood ratio of greater than 10 or less than 0.1 gives strong
277 evidence to rule the disease in or out.² These ranges are only intended to provide an
278 approximate rule of thumb and consideration must be given to the context of the results. It
279 should also be noted that quite different combinations of sensitivity and specificity can
280 produce the same likelihood ratio values.

281

282 **1.1.4. Diagnostic odds ratio**

283 The diagnostic odds ratio (DOR) provides a single measure of test performance that is
284 assumed to be independent of the prevalence of the target condition.

$$285 \quad \text{DOR} = (\text{TP} / \text{FN}) / (\text{FP} / \text{TN}) \quad (6)$$

286 The diagnostic odds ratio describes the odds of a positive test results in participants with
287 the disease compared to the odds of a positive test results in those without the disease. A
288 single diagnostic odds ratio corresponds to a set of sensitivities and specificities depicted
289 by a symmetrical receiver operating characteristic curve (see section 2.2).³ The DOR is
290 not useful in clinical practice. As it is a single measure independent of prevalence, the
291 DOR is referred to as a global and unconditional measure.

292

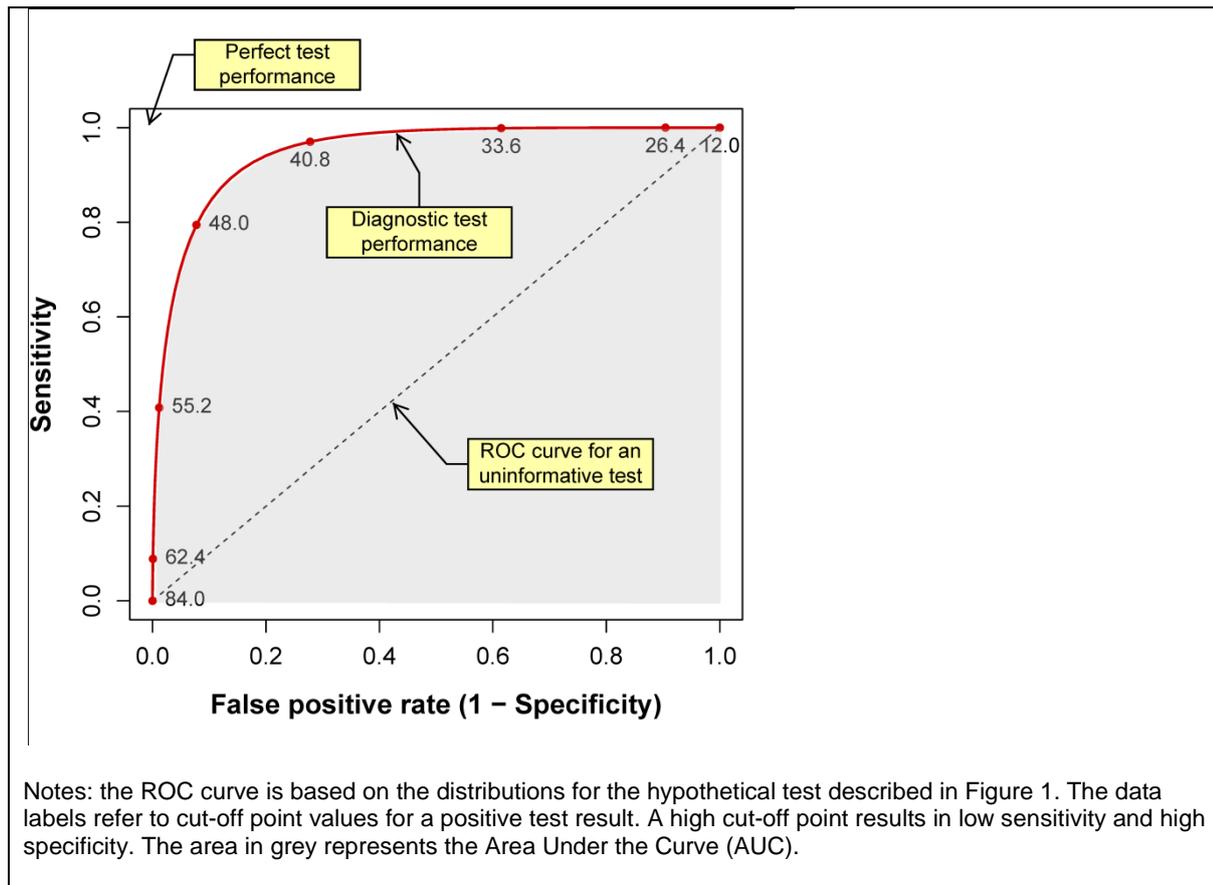
293 **1.1.5. Receiver Operating Characteristic (ROC) curves**

294 A diagnostic test may return values on a continuous scale, but this must then be converted
295 into a dichotomous positive/negative diagnosis based on a cut-off point. The choice of cut-
296 off point on the scale will impact on the test's accuracy. A threshold at one extreme will
297 result in few positives, while a threshold at the other extreme will result in many positives
298 (see Figure 2). A ROC curve is a graphical plot used to represent the performance of a
299 test over a range of threshold settings (see Figure 3).³ That is, the curve shows the impact
300 on sensitivity and specificity of varying the threshold for which a test result is labelled as a
301 positive rather than a negative. A ROC curve plots the test sensitivity as a function of the
302 false-positive rate (100 – Sp). As with the DOR, the ROC curve is a single measure
303 independent of prevalence, and hence is referred to as a global and unconditional
304 measure.

305

306

307 Figure 3. Example of a Receiver Operating Characteristic (ROC) curve
308



309

310 The diagonal line in Figure 3 represents a test that is essentially uninformative, as the
311 ability to detect genuine cases is no better than chance allocation to positive and negative.
312 The upper left-hand corner represents a test with sensitivity and specificity of 100%, in
313 other words a perfect dichotomous test. Clearly a desirable test is as close as possible to
314 the upper left-hand corner and as far from the diagonal as possible. The cut-off point that
315 yields the most upper-left point may be appropriate for clinical practice, presuming it is
316 feasible and has been validated in (preferably multiple) independent samples.

317 An area under the curve (AUC) of 1 represents a perfect test, while an AUC of 0.5
318 represents an uninformative test. The AUC is sometimes reported as a single summary
319 measure of diagnostic accuracy and gives an indication of how close to perfect, or
320 uninformative, a test is. Two tests, one with high sensitivity and the other with high
321 specificity, may have the same AUC. The AUC does not provide any information about
322 how the patients are misclassified (i.e., false positive or false negative) and should
323 therefore be reported alongside another measure of test performance, such as likelihood
324 ratios or predictive values.² The AUC is not useful in clinical practice as it summarises
325 performance over a range of possible thresholds, whereas in practice a single pre-
326 specified threshold applies. It should be noted that ROC curves of different shapes can
327 have the same AUC value, so an AUC value does not represent a set of unique
328 combinations of sensitivity and specificity.⁷

329

330 **1.1.6. Predictive values**

331 The positive predictive value (PPV) is the proportion of patients with a positive test who
332 actually have the disease, and the negative predictive value (NPV) is the proportion of
333 patients with a negative test result who are actually free of the disease.²

334
$$PPV = \frac{Sn.P}{Sn.P+(1-Sn).(1-P)} \quad (7)$$

335
336
$$NPV = \frac{Sp.(1-P)}{(1-Sn).P+Sp.(1-P)} \quad (8)$$

337 Where P is the estimated prevalence of disease, also known as the pre-test or prior
338 probability of disease.⁸ A patient belonging to a population with a higher pre-test
339 probability will have a higher PPV than a patient from a lower risk population. Predictive
340 values have a strong clinical utility. However, they vary with disease prevalence and are
341 not useful in situations where test results are reported on multiple levels rather than a
342 dichotomous outcome. The predictive values are referred to as specific unconditional
343 measures of test accuracy.

344

345 **1.1.7. Diagnostic accuracy**

346 A single overall measure of test accuracy is also used which is expressed as the
347 proportion of correctly classified cases:⁹

348
$$\text{Diagnostic accuracy} = (TP + TN) / (TP + FP + FN + TN) \quad (9)$$

349 A global, conditional measure of accuracy, it is not often used and is not pooled across
350 studies.

351

352 **1.2. Problem statement**

353

354 Diagnostic tests are used for a variety of purposes including: to determine whether or not
355 an individual has a particular target condition; to provide information on a physiological or
356 pathological state, congenital abnormality, or on a predisposition to a medical condition or
357 disease; to predict treatment response or reactions; and to define or monitor therapeutic
358 measures. As such, the test is not a treatment, but influences a clinician when deciding on
359 the appropriate course of action for a particular patient. Timely or correct detection of
360 disease does not necessarily lead to timely or correct treatment of disease, hence
361 improved diagnostic test accuracy is not synonymous with improved patient outcomes.
362 Diagnostic tests can change patient outcomes by changing diagnostic and treatment
363 decisions, impacting on timely treatment, modifying patient perceptions and behaviour, or
364 putting patients at risk of direct harm.¹⁰

365

366 To study the association between the accuracy of a diagnostic test with regard to
367 outcomes a follow-up is required, but this may be at significant risk of confounding¹¹ unless
368 studied in RCTs. In the area of health technology assessment, diagnostic test accuracy is

Copyright © EUnetHTA 2014. All Rights Reserved. No part of this document may be reproduced without an explicit acknowledgement of the source and EUnetHTA's expressed consent.

369 sometimes used as a surrogate for patient-relevant outcomes, although some agencies
 370 require long-term outcome data on patient outcomes.¹² In practice, diagnosis may also
 371 depend on factors other than just the results of a single diagnostic test, such as clinical
 372 history and additional testing, and hence other factors will impact on diagnosis, treatment
 373 and outcomes.¹³ A linked evidence approach, whereby patient outcomes can be
 374 associated with the diagnostic test, may be a pragmatic solution although in practice there
 375 are often insufficient data to enable this approach.¹⁴ It should also be noted that diagnostic
 376 tests may be relatively invasive (e.g., sentinel lymph node biopsy) or be harmful to patients
 377 (e.g., exposure to ionising radiation), and that this information is not captured in the
 378 assessment of test accuracy.

379

380 When the sensitivity of a new diagnostic test is compared with an existing test, the
 381 detected cases may be different to those detected by the existing test. Results from
 382 treatment trials based on patients detected by the old test may not be generalisable to the
 383 cases detected by the new test. Unless clinicians can be satisfied that the new test detects
 384 the same spectrum and subtype of disease as the old test or that treatment response is
 385 similar across the spectrum of disease, it is possible that the new test will result in different
 386 outcomes.¹⁵

387

388 The impact of a diagnostic test can be viewed according to a number of domains (see
 389 Table 1).¹⁶⁻¹⁸ The tiered model has been tailored to radiological testing, for which the
 390 resolution and sharpness of test images are relevant. For other types of tests, the
 391 resolution and sharpness may be analogous to the precision of the test. The stages or
 392 tiers of efficacy answer a variety of questions about a diagnostic test, from whether or not
 393 it can work to whether or not it is worth using. This guideline is restricted to methodologies
 394 for summarising diagnostic accuracy.

395 Table 1. Tiered model of diagnostic efficacy¹⁶⁻¹⁸

Stage of efficacy	Definition
Technical capacity	Resolution, sharpness, reliability
Diagnostic accuracy	Sensitivity, specificity, predictive values, ROC curves
Diagnostic impact	Ability of a diagnostic test to affect the diagnostic workup
Therapeutic impact	Ability of a diagnostic test to affect therapeutic choices
Patient outcomes	Ability of a diagnostic test to increase the length or quality of life
Societal outcomes	Cost-effectiveness and cost-utility

396

397

398 **1.3. Objective(s) and scope of the guideline**

399

400 This guideline presents a review of the available methods for the meta-analysis of
401 diagnostic test accuracy studies. The aim of the guideline is to highlight the circumstances
402 in which it is appropriate to use each of the approaches. The guideline will also elaborate
403 on:

- 404 • thresholds for positive tests
- 405 • fixed and random effects approaches
- 406 • heterogeneity across studies
- 407 • sample sizes
- 408 • the quality and quantity of evidence required for a meta-analysis
- 409 • the case where multiple diagnostic tests may be evaluated and compared
- 410 • issues regarding study selection
- 411 • the types of bias that might arise when reviewing diagnostic test accuracy
- 412 data will be reviewed.

413 The guidance is restricted to methods for pooling results from diagnostic tests that report
414 dichotomous results (i.e., the test result if either positive or negative), as opposed to tests
415 that report results on a continuous scale or as a number of discrete levels.

416 The guideline will not address issues relating to systematic reviews and meta-analysis that
417 are not restricted or unique to diagnostic test accuracy studies. These issues include:
418 bibliographic searching and study types. These issues are common to any meta-analysis
419 and are comprehensively described elsewhere.¹⁹ It is assumed that the meta-analysis is
420 undertaken using comparable studies derived from a systematic review conducted
421 according to best practice.

422

423

424 **1.4. Related EUnetHTA documents**

425

426 The following EUnetHTA methodological guidelines are relevant to the present guideline:

- 427 • Applicability of evidence in the context of a relative effectiveness assessment of
- 428 pharmaceuticals (February 2013)
- 429 • Direct and indirect comparisons (February 2013)

430 It should be noted that the EUnetHTA guidelines were developed for the relative
431 effectiveness assessment of pharmaceuticals. However, the principles contained in the
432 guidelines are also relevant to the meta-analysis of diagnostic test accuracy studies.

433 Also relevant are the assessment elements from the HTA Core Model Application for
434 Diagnostic Technologies:

- 435
- 436
- Assessment element tables for HTA Core Model Application for Diagnostic Technologies (2.0), mekathtl.fi/htacore/model/AE-tables-diagnostics-2.0.pdf

437 2. Analysis and discussion of the methodological issue

438 2.1. Methods for meta-analysis of diagnostic accuracy studies

439 A variety of methods are available for pooling data from multiple studies of diagnostic test
440 accuracy. The relevance of each method is influenced by the type of study data available
441 (e.g., individual patient data, 2x2 tables, summary measures such as sensitivity and
442 specificity). Certain data may not be available for all studies, which will also influence the
443 approach to pooling data. The most straightforward approach is a simple pooling where
444 the 2x2 tables from all of the studies are combined with no weighting.⁷ This method
445 assumes no correlation between sensitivity and specificity, no between-study
446 heterogeneity, and no variability in the diagnostic threshold. As such, simple pooling can
447 be described as a naive approach and will not be considered in these guidelines.

448 It is assumed that a meta-analysis is only undertaken when the available studies are
449 considered equivalent in terms of the index test, reference standard, the patient
450 population, and the indication. Where the studies are not equivalent it is not recommended
451 that a meta-analysis is undertaken. A lack of study equivalence gives rise to various types
452 of bias which are discussed in Section 2.4.

453

454 2.1.1. Separate random-effects meta-analyses of sensitivity and specificity

455 Sensitivity and specificity can be individually summarised across studies based on their
456 logit transforms.²⁰ This approach is a random-effects method that allows for between-study
457 heterogeneity in the two measures, but ignores the potential correlation between the two.
458 The logit transforms are used in the analysis on the basis that an assumption of a normal
459 distribution between studies is more reasonable on the logit scale, with an inverse logit
460 transformation applied to the results to return them to a [0,1] interval. In addition to the
461 point estimates of sensitivity and specificity, this approach also allows for the estimation of
462 a ROC curve using the ratio of estimated between-study variances. This approach may be
463 appropriate when there is evidence of no correlation between sensitivity and specificity
464 across studies.²¹

465 2.1.2. Separate meta-analyses of positive and negative likelihood ratios

466 As likelihood ratios are ratios of probabilities, positive and negative likelihood ratios can be
467 pooled separately by meta-analysis using the same mathematical methods as risk ratios.²⁰
468 Approaches can be based on either fixed-effect or random-effects models. These methods
469 ignore the possible correlation between positive and negative likelihood ratios, and thus
470 pooled estimates may produce values that are not possible in reality (e.g., both ratios
471 above or below 1.0).²² Should they be required, pooled estimates of the likelihood ratios
472 can be computed from summary estimates of sensitivity and specificity derived using any
473 of the other methods described in this section. Meta-analysis of predictive values is
474 possible, although it is usually discouraged because of the influence of disease
475 prevalence.

476 2.1.3. Moses-Littenberg summary receiver operating characteristic (SROC) curve

477 The Moses–Littenberg fixed-effects method is historically the most commonly used
478 method for meta-analysis of diagnostic tests. A straight line is fitted to the logits of the false
479 positive rate (FPR) and true positive rate (TPR) of each study, and its slope and intercept
480 give the parameters of the SROC curve.²¹ The SROC curve summarises pairs of

Copyright © EUnetHTA 2014. All Rights Reserved. No part of this document may be reproduced without an explicit acknowledgement of the source and EUnetHTA's expressed consent.

481 sensitivity and specificity from multiple studies. The least squares linear fit may be
482 unweighted or weighted, although in the latter case there is uncertainty as to which
483 weighting method to use.²⁰ The linear fit is then back-transformed to be plotted as the
484 SROC curve. The Moses-Littenberg may be appropriate if all the observed heterogeneity
485 is due to a threshold effect. That is, where all of the observed heterogeneity is due to the
486 use of different thresholds across the included studies.

487 This method allows for the correlation between sensitivity and specificity, but is not
488 statistically rigorous, as the assumptions of linear regression (constant variance, covariate
489 measured without error) do not hold.²⁰ Furthermore, as it is based on an analysis of the
490 DOR, summary measures of sensitivity and specificity are not directly available. By
491 selecting a value for sensitivity, it is possible to compute the corresponding specificity. It is
492 common to report the sensitivity and specificity at the Q-point, which is where sensitivity
493 equals specificity (where the SROC curve intersects the diagonal that runs from the top left
494 to bottom right of the ROC plot).²³ However, the values at the Q-point may bear little
495 relation to the values observed in the original studies used in the meta-analysis.

496 **2.1.4. Hierarchical summary ROC (HSROC) model**

497 The HSROC model for combining estimated pairs of sensitivity and specificity from
498 multiple studies is an extension of the Moses-Littenberg fixed-effects summary ROC
499 (SROC) model.²⁴ The HSROC model more appropriately incorporates both within- and
500 between-study variability, and allows greater flexibility in the estimation of summary
501 statistics. The HSROC model describes within-study variability using a binomial
502 distribution for the number of positive tests in diseased and not diseased patients.

503 The model is specified on two levels: the within study model and the between study model.
504 The within study model takes the following form:²⁵

$$505 \quad \text{logit}(\pi_{ij}) = (\theta_i + \alpha_i X_{ij}) \exp(-\beta X_{ij}) \quad (10)$$

506 The variable π_{ij} is the probability that a patient in study i with disease status j will return a
507 positive test result. By defining $j=0$ for a patient without the disease and $j=1$ for a patient
508 with the disease, it follows that for study i , π_{i0} is the false positive rate and π_{i1} is the true
509 positive rate. The parameter X_{ij} is a dummy variable for the true disease status of a
510 patient in study i with disease status j . The parameters θ_i and α_i are the cut-off point and
511 accuracy parameters, respectively, and are allowed to vary between studies. Finally, β is a
512 scale parameter for modelling the possible asymmetry in the ROC curve.

513 The between-study model allows the parameters θ_i and α_i to vary between studies. The
514 following parameter definitions include a common covariate Z which affects both
515 parameters, although they can be modelled without covariates or with multiple covariates:

$$516 \quad \theta_i \sim N(\Theta + \gamma Z_i, \sigma_\theta^2) \quad (11)$$

$$517 \quad \alpha_i \sim N(\Lambda + \lambda Z_i, \sigma_\alpha^2) \quad (12)$$

518 The model was originally formulated in a Bayesian framework, and hence also included
519 specification of priors.²⁴ The model produces an SROC curve by allowing the cut-off point
520 parameter to vary while holding the accuracy parameter at its mean value.

521 2.1.5. Bivariate random-effects meta-analysis for sensitivity and specificity

522 As with the HSROC method, the bivariate approach preserves the two-dimensional nature
523 of the original data, with pairs of sensitivity and specificity jointly analysed.²³ Like the
524 HSROC approach, this method also incorporates any correlation that might exist between
525 these two measures using a random-effects approach. Evaluation of the bivariate model
526 requires specification of an appropriate transformation (e.g., a generalised linear mixed
527 model using the logit-transformation).²⁶ Explanatory variables can be added to the
528 bivariate model and lead to separate effects on sensitivity and specificity, rather than a net
529 effect on the odds ratio scale as in the SROC approach.²³

530 The bivariate model is specified as follows:²⁵

$$531 \begin{pmatrix} \mu_{Ai} \\ \mu_{Bi} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \Sigma_{ab} \right) \quad (13)$$

$$532 \Sigma_{AB} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \quad (14)$$

533 The variables μ_{Ai} and μ_{Bi} are the logit transformed sensitivity and specificity, respectively,
534 for study i . Covariates affecting sensitivity and specificity can be included by replacing the
535 means μ_A and μ_B with linear predictors in the covariates.²⁵ The covariates can be applied
536 to one or both measures, and can have common or distinct effects.

537 2.1.6. Comparison of methods

538 The appropriate choice of methodology for the meta-analysis of diagnostic test accuracy
539 studies will depend on numerous factors. The Moses-Littenberg model is considered as
540 approximate, as the assumptions of simple linear regression are not met and because of
541 the uncertainty around the appropriate weighting.²⁵ As the Moses-Littenberg model is
542 essentially a fixed-effect model it does not provide estimates of the between study
543 heterogeneity.²⁷ This method can also lead to improper SROC curves.²⁷

544 The HSROC and bivariate methods are equivalent under certain parameterisations, such
545 as in the absence of covariates or when the same covariates affect both sensitivity and
546 specificity (in the bivariate model) and both the accuracy and cut-off point parameters (in
547 the HSROC model).²⁵ Therefore in situations where there are no covariates, the two
548 models will return equivalent estimates of the expected sensitivity and specificity (and also
549 any measures derived from those two measures).

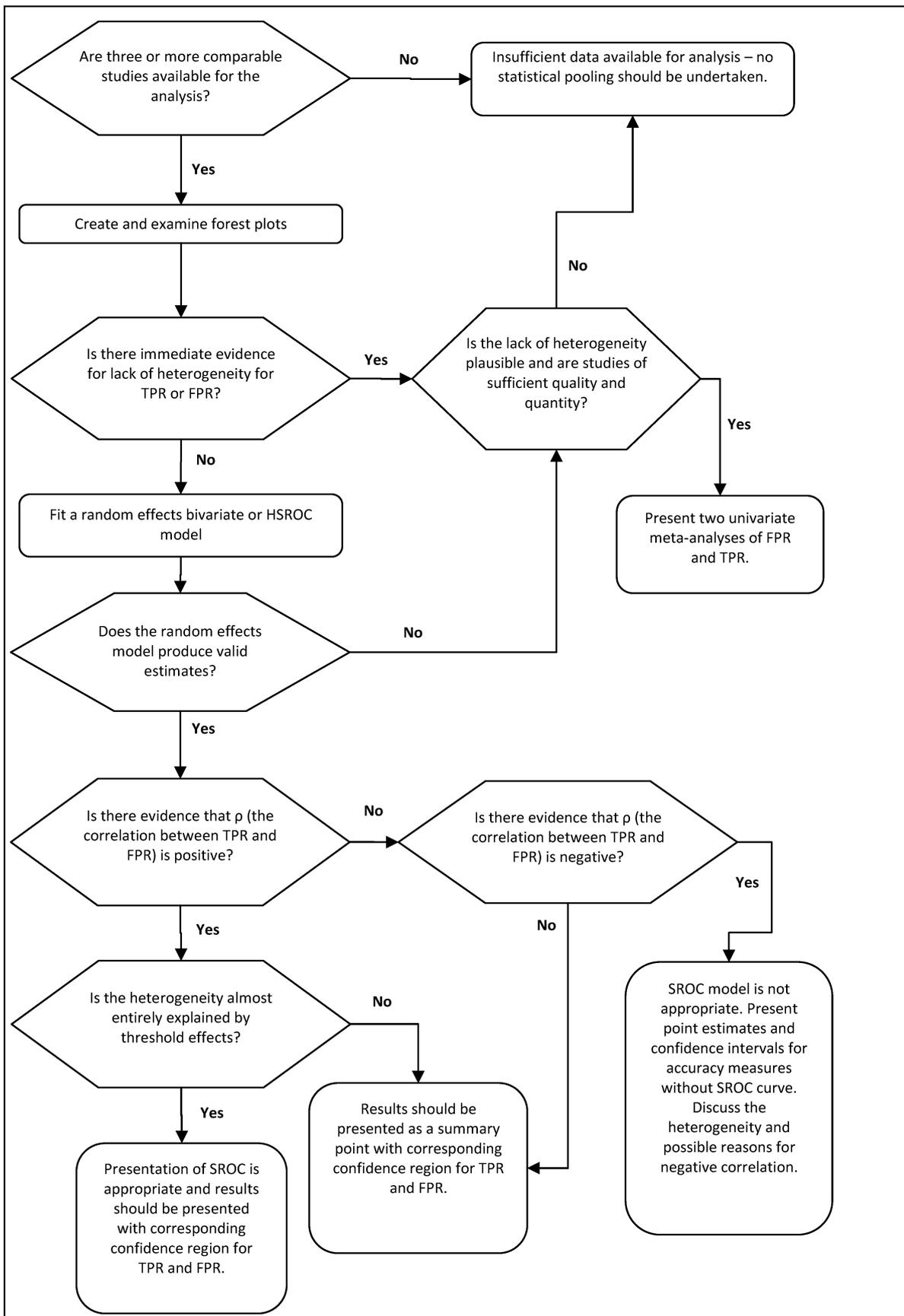
550 The HSROC and bivariate approaches are considered to be more statistically rigorous
551 than the Moses-Littenberg approach,^{20;27} although it has been questioned whether this
552 necessarily translates into improved estimates of diagnostic test accuracy in all
553 situations.²⁸ There is an increasing consensus that the HSROC and bivariate approaches
554 offer the best methodologies for pooling diagnostic test accuracy studies, but there are
555 differences between the two approaches and the nature of the underlying data may dictate
556 which approach is more appropriate.

557 A first step is to separately examine the distributions of sensitivity and specificity from the
558 included studies.²¹ If either measure shows a lack of heterogeneity, then it is more
559 appropriate to analyse the data using separate univariate meta-analyses to derive point
560 estimates and confidence bounds for sensitivity and specificity. However, a full description

561 of the included studies can provide contextual information that may justify a full analysis in
562 these situations. If only one study is available, then clearly there is no basis for meta-
563 analysis. If only two studies are available, then there is insufficient information available to
564 reliably estimate all of the parameters in the HSROC and bivariate models. Therefore, in
565 the case of two studies, it is not recommended to undertake a meta-analysis and a
566 narrative description of the studies should be presented.

567 The correlation between sensitivity and specificity is important and is estimated by the
568 HSROC and bivariate methods. Ordinarily a positive correlation is expected between TPR
569 and FPR. However, data from studies are often noisy and no correlation or a negative
570 correlation may be estimated. The confidence bounds of the correlation estimate should
571 be assessed. If there is a significant negative correlation, this implies that sensitivity
572 improves with increasing specificity, which is unlikely to occur in practice due to the
573 relationship between disease status and the test cut-off point (see Figure 2). In the event
574 of a negative correlation, the plausibility of this finding should be discussed in relation to
575 the nature of the test and the quantity of evidence.

Figure 3. Algorithm for the meta-analysis of diagnostic test data (adapted from Chappell et al.²¹)



579 A key consideration is then whether or not a threshold effect is present, which is usually
580 evidenced by a positive correlation between the false positive rate and sensitivity. When a
581 threshold effect is present, then an SROC approach is appropriate, which can be achieved
582 using either the HSROC or bivariate approaches. Estimates of test accuracy can be
583 plotted in ROC space. In the absence of a threshold effect, the SROC approach is not
584 appropriate. There will be situations where a threshold effect may or may not be plausible,
585 depending on the nature of the test and the indication. For example, some tests explicitly
586 depend on converting a measure on a continuous scale into a dichotomous measure of
587 disease status (e.g., Prostate-Specific Antigen (PSA) test). These tests are likely to give
588 rise to threshold effects. Some tests, on the other hand, may rely on a simple
589 presence/absence measure of a biomarker which is directly interpreted as a measure of
590 disease status (e.g., rapid strep test), or may employ an unequivocal cut-off point that is
591 universally adopted (e.g., Ottawa Ankle rules). Due to differences in how test results are
592 interpreted, a threshold effect may arise even when there is a universally employed cut-off
593 point.

594 If a threshold effect is plausible, and heterogeneity is observed, then it must be evaluated
595 whether the heterogeneity can be attributed to a threshold effect. Determining whether
596 observed heterogeneity is due to a threshold effect is generally based on a visual
597 inspection of the distribution of study points in relation to the SROC curve. If study points
598 are in close proximity to the SROC, then there will be reasonable confidence that the
599 threshold effect is responsible for the heterogeneity. An inspection of the shape of the
600 confidence region is also helpful, particularly to check whether the region largely
601 encompasses and follows the shape of the SROC curve. If, on the other hand, the
602 prediction region bears little relation to the SROC curve, or the study points are not close
603 to the SROC curve, then it is reasonable to conclude that factors other than just a
604 threshold effect are responsible for the observed heterogeneity.

605 The choice of method used must be justified by the context (e.g., the studies, inclusion of
606 covariates, correlation between sensitivity and specificity), and the assumptions must be
607 clearly understood and how they may impact on the interpretation of the results.

608

609

610

611

612

613

614

615

616

617

618 **2.2. Presentation of results from a meta-analysis of a single diagnostic test**

619 Reports of meta-analyses of diagnostic test accuracy must contain the requisite
620 information for a reader to know how the analysis was undertaken, what data were used,
621 what results were found, and whether or not the findings are reliable. To achieve this, a
622 number of presentational features should be included, depending on the type of analysis
623 undertaken.

624 **2.2.1. Tables**

625 Reports should include a table of all the included studies, and the relevant data from the
626 2x2 tables for each of the studies. Such tables can also include the estimated sensitivity
627 and specificity and associated confidence bounds for the two measures for each included
628 study.

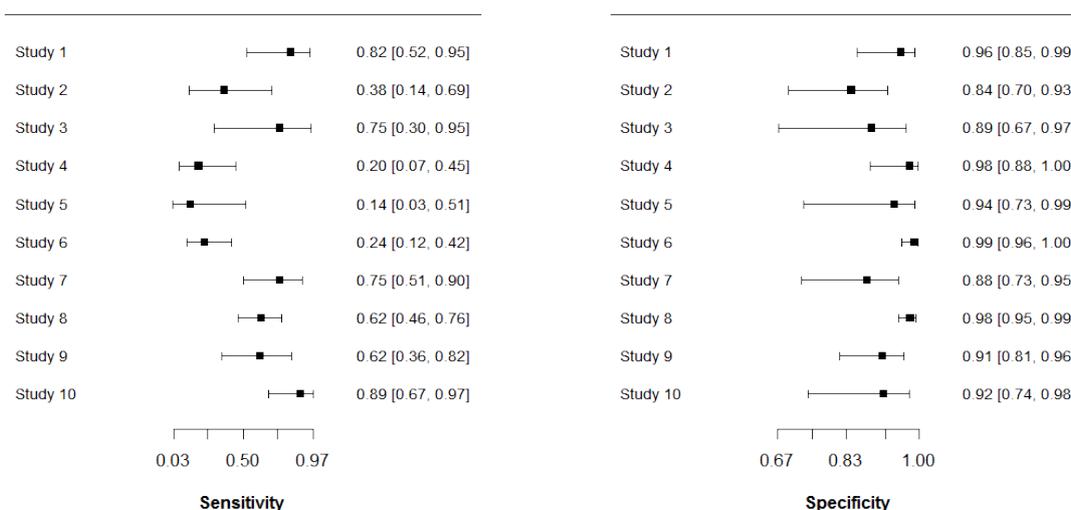
629 For the results of the meta-analysis, summary estimates of accuracy and their associated
630 confidence bounds should be reported. The main result of the bivariate and the HSROC
631 models is the pooled estimate of the summary paired sensitivity and specificity. At a
632 minimum, the results for sensitivity and specificity should be reported, although the DOR
633 and likelihood ratios may also be useful. Any other useful outputs from the analysis (e.g.,
634 the estimated correlation between sensitivity and specificity) should also be reported as
635 they may aid interpretation of the data and results.

636 The results of any subgroup analyses should also be tabulated. Results of sensitivity
637 analyses can also be included in tables, as summary points and confidence bounds
638 cannot always be easily read from graphical displays.

639 **2.2.2. Forest plots for sensitivity and specificity**

640 Forest plots (also called blobbograms) of sensitivity and specificity are useful for showing
641 heterogeneity across studies. These plots give the point estimates and confidence bounds
642 for sensitivity and specificity for the individual studies included in the analysis (see Figure
643 4). Studies may be ordered by sensitivity or specificity, which can aid interpretation or
644 make it more apparent if there is a correlation between the two measures.

645 Figure 4. Forest plots of sensitivity and specificity for a sample meta-analysis



646

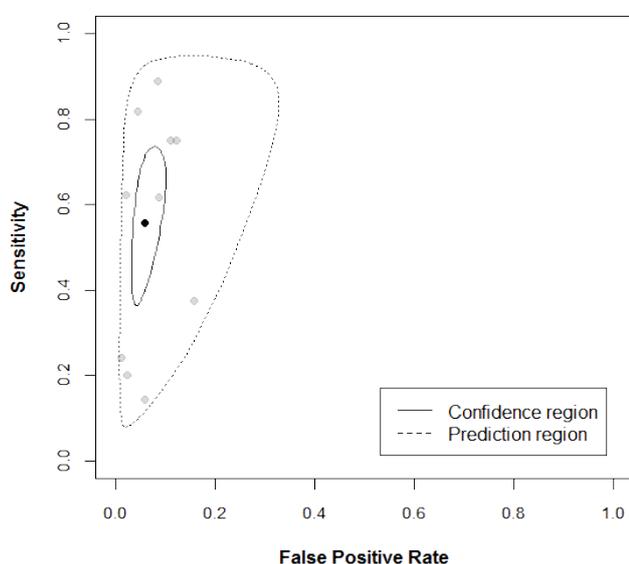
647

648 **2.2.3. Confidence and prediction regions for the summary estimate of sensitivity**
649 **and specificity**

650 From the meta-analysis of diagnostic test accuracy, it is possible to generate 95%
651 confidence and prediction regions for sensitivity and specificity. The confidence region
652 relates to the summary point estimate based on the included studies whereas the
653 prediction region refers to potential values of sensitivity and specificity that might be
654 observed in a future study. If the summary values for sensitivity and specificity are to be
655 using in a subsequent relative effectiveness assessment simulation model, the prediction
656 region may form a more realistic basis for defining parameter uncertainty than the more
657 narrowly defined confidence region. Furthermore, prediction regions can also be used for
658 the purpose of identifying studies that may be statistical outliers.

659 Both the HSROC and bivariate models facilitate the computation of confidence and
660 prediction regions around the summary point for sensitivity and specificity, usually in the
661 form of a joint confidence ellipse for sensitivity and specificity.

662 Figure 5. An example of a summary sensitivity and false positive rate point



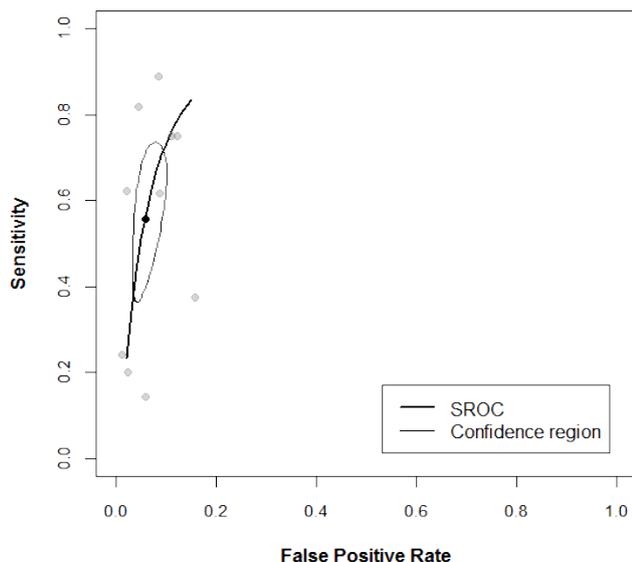
663

664 **2.2.4. Summary ROC curve**

665 The choice to display an SROC curve depends on whether or not the included studies had
666 a common positivity threshold and the subsequent analytical approach. In instances where
667 the threshold varies across studies, a summary estimate of sensitivity and specificity is of
668 limited use as it represents an average across thresholds. Where the threshold varies, it is
669 appropriate to report an SROC curve.

670 It must be noted that the SROC curve as specified for the HSROC model is constrained to
671 always be positive.²⁴ An SROC curve can also be generated from the results of the
672 bivariate model, although a variety of formulations are possible which can lead to quite
673 different curves, including those with a negative slope.²⁹ Another drawback of the SROC
674 curve is that uncertainty in the SROC curve is not generally calculated as a single common
675 SROC curve is assumed. Using a Bayesian approach, it is possible to generate numerous
676 SROC curves based on posterior densities which can be used to derive a graphical
677 indication of uncertainty in the curve within specified quantiles.²¹

678 Figure 6. An example of a summary receiver operating characteristic (SROC) curve



679

680 It is suggested that the SROC curve should be restricted to the observed range of
681 specificities in the included studies and that the analyst should not extrapolate beyond the
682 observed data.²⁴ For example, if the highest upper bound for the false positive rates
683 observed in the included studies is 0.60, then the SROC curve should not be extended
684 beyond a FPR of 0.60 in the plot.

685 2.2.5. Sensitivity analysis

686 It is typical in relative effectiveness assessment to consider the influence on results of
687 various factors. This is generally achieved through sensitivity analysis – an evaluation of
688 how much the conclusions change if the included evidence is changed. In this context,
689 sensitivity analysis refers to the quantification of uncertainty rather than an analysis of the
690 measure of diagnostic accuracy. Sensitivity analysis may be targeted (e.g., re-analysing
691 the data with data at risk of bias excluded) or systematic (for example, univariate
692 sensitivity analysis with all uncertain parameters varied one at a time). The same
693 principles apply to meta-analysis.

694 In any meta-analysis there is likely to be heterogeneity across the included studies. There
695 can be many reasons for that heterogeneity, including systematic differences between the
696 studies in terms of the patients, how the test was applied, and the choice of reference
697 standard. Study quality can also be variable and the application of a formal risk of bias
698 measure can be used to identify specific studies at high risk of bias. A targeted sensitivity
699 analysis may involve excluding studies that are considered outliers in a statistical sense, or
700 that have been evaluated as being at high risk of bias. Alternatively, the meta-analysis
701 may be restricted to a sub-group of studies with a common characteristic, although this
702 can also be achieved by a meta-regression approach, which is possible with both the
703 HSROC and bivariate methods. By extension, a systematic sensitivity analysis may be
704 based around repeating the meta-analysis with each of the studies excluded in turn. Study
705 influence can be measured using metrics such as Cook's distance, while statistical outliers
706 may be identified using standardised study-level residuals.³⁰ In both cases, these metrics
707 can be applied to sensitivity and specificity simultaneously.

708 If the results remain relatively unchanged then there can be confidence that the summary
709 estimates are accurate. If, however, the results are sensitive to the included data, then
710 greater attention needs to be paid to the included studies and what characteristics are
711 impacting on differences. Evidence for a relative effectiveness assessment must be
712 relevant to the target population and conditions under which the diagnostic test will be
713 used in practice.

714

715 **2.3. Comparison of two diagnostic tests with respect to diagnostic accuracy** 716 **(incorporate non-comparative studies in discussion of heterogeneity)**

717 Estimating diagnostic test accuracy is often for the purpose of comparing two or more tests
718 for the same indication. In this situation, the diagnostic accuracy of all tests has to be
719 compared.³¹ However, the evidence derived from comparative and non-comparative
720 studies often differs. Ideally, for the purposes of comparing two diagnostic tests, robustly
721 designed studies in which all patients receive all tests or are randomly assigned to receive
722 one or other of the tests are preferred as evidence to guide test selection.³² Irrespective of
723 the comparison, the same reference standard test should be used for all patients. The use
724 of data from non-comparative studies increases the chances of differences in the patient
725 populations, different reference standard tests, and different interpretation of test results.

726 **2.4. Sources of bias**

727 Evidence has shown that diagnostic studies with methodological shortcomings may
728 overestimate the accuracy of a diagnostic test, particularly those including non-
729 representative patients or applying different reference standards.³³ As with the meta-
730 analysis of interventions, the pooling of data across diagnostic test accuracy studies may
731 be subject to numerous sources of bias, although some forms of bias are specific to
732 diagnostic test studies.²² In this section we outline some of the main sources of bias that
733 can occur. In many cases, there is little that can be done to correct for bias beyond a
734 forensic examination of the included studies, careful documentation of potential bias, and a
735 full sensitivity analysis to examine the potential impact on results of including studies at
736 risk of bias.

737 **2.4.1. Data gathering and publication bias**

738 As with the meta-analysis of any clinical intervention, meta-analysis of diagnostic test
739 accuracy studies should be undertaken as part of a systematic review. Methods for
740 systematic review are well described elsewhere,¹⁹ with specific guidance available for
741 diagnostic test accuracy studies.³⁴ The identification of diagnostic test accuracy studies
742 can pose particular difficulties, due in part to the lack of consistent terminology or use of
743 MESH terms, indeed, in some cases methodological filters can reduce the ability to find
744 relevant studies.^{35;36} Best practice is to search on the basis of the index test and target
745 condition.³⁶ Difficulties can also arise where a single study publishes multiple articles using
746 the same or overlapping cohorts; care must be taken not to include data on the same
747 patients from several articles, which can be referred to as double data reporting bias.

748 Publication bias is believed to arise due to studies with poor test performance results not
749 getting published, leading to exaggerated estimates of test performance in a systematic
750 review.³⁷ As with meta-analyses of clinical interventions, asymmetry in the funnel plot
751 (constructed using the DOR) is often taken as an indication that there may be publication
752 bias, although there may be many other factors causing asymmetry (e.g., variations in test

753 procedures, patients, or reference standards).³⁸ It is possible that publication bias may be
754 more prevalent in studies of test accuracy than in studies of clinical effectiveness.³⁸ There
755 are a number of approaches available for estimating funnel plot asymmetry, each of which
756 may give different results in a given context. Furthermore, where the number of included
757 studies is small, the statistical methods available may be underpowered to detect
758 asymmetry. As such, funnel plot asymmetry should be used but interpreted with caution.³⁷

759 **2.4.2. Heterogeneity in meta-analyses of sensitivity and specificity**

760 Between-study heterogeneity refers to differences to variability in the results of studies.
761 Clinical, methodological, and statistical heterogeneity are distinct concepts. Clinical
762 heterogeneity refers to variability across studies in terms of participants, the intervention,
763 and outcomes. These are legitimate differences that arise because the studies are not
764 comparing like with like. Methodological heterogeneity is a function of variability in study
765 design and risk of bias. Differences in methodology may include differences in the
766 technical specifications of the test, such as the protocols for how the test is applied. This
767 may also be referred to as technical heterogeneity. Statistical heterogeneity arises when
768 there is greater variability in outcomes than would be expected by chance, and usually
769 invokes a violation of underlying assumptions. For the purposes of this guideline, the
770 outcome measure of interest is diagnostic test accuracy. Clinical and methodological
771 heterogeneity will often but not necessarily give rise to statistical heterogeneity.

772 The obvious source of statistical heterogeneity in sensitivity and specificity is related to
773 threshold differences for test positivity.³⁹ If the observed between-study heterogeneity is
774 entirely due to variation in the diagnostic threshold, estimates of summary sensitivity and
775 specificity will underestimate diagnostic performance.³ In these situations the appropriate
776 meta-analytical summary is the receiver operating characteristic curve rather than a single
777 summary point. However, it must be clear that there are no other substantial sources of
778 heterogeneity. Where there are a variety of sources of heterogeneity, including threshold
779 effects, the HSROC or bivariate method should be used with random effects. Presentation
780 of an SROC may not be informative unless some attempt to measure uncertainty in the
781 curve is included.

782 Another potentially important source of heterogeneity is due to observer variability. Within-
783 study observer variability can be of the same order of magnitude as variability across
784 studies.⁴⁰ By including studies from a wide time horizon, it is also possible that changes in
785 how a diagnostic test is used in practice may have occurred, giving rise to heterogeneity.⁴⁰

786 Although measures of heterogeneity exist for univariate meta-analyses (e.g., I^2 , τ^2), there
787 is no analogue for bivariate meta-analyses. The amount of observed heterogeneity is
788 quantified by the random effects terms in the models, but these are not easily
789 interpreted.²⁷ The distribution of study points on a plot of true versus false positive rates
790 relative to the estimated SROC can give an indication of whether there is heterogeneity
791 due to variation in the test threshold. The distribution of points relative to the prediction
792 ellipse can also provide an indication of whether or not there is heterogeneity.²⁷

793 A common approach to exploring heterogeneity is to use meta-regression whereby study-
794 level covariates are included when estimating summary statistics. Both the HSROC and
795 bivariate models facilitate the use of study-level covariates as either categorical (e.g.,
796 study design) or continuous (e.g., average patient characteristics).²⁷ In the bivariate model,
797 covariates can be incorporated to affect summary sensitivity or summary specificity, or
798 both measures. The HSROC model, on the other hand, allows covariates to be added to

Copyright © EUnetHTA 2014. All Rights Reserved. No part of this document may be reproduced without an explicit acknowledgement of the source and EUnetHTA's expressed consent.

799 affect the test positivity, position of the curve, and shape of the curve. A covariate may be
800 associated with some, but not all three model parameters.²⁷

801

802 **2.4.3. Spectrum bias**

803 As test performance often varies across population subgroups, diagnostic tests should be
804 evaluated in a clinically relevant population. The performance of the test may vary
805 depending on the mix of patients, most particularly due to differences in disease severity.
806 Inappropriate use of patient populations can occur, introducing a form of heterogeneity
807 referred to as spectrum bias.⁴¹ When there is spectrum bias the diagnostic test
808 performance varies across patient subgroups and a study of that test's performance does
809 not adequately represent all subgroups. The impact of spectrum bias on the estimated test
810 accuracy will depend on the difference between included patients and the actual target
811 population.

812

813 **2.4.4. Verification/work-up bias and variable gold standard**

814 Verification bias (also called selection or workup bias), occurs when not all recipients of
815 the index test also receive the reference or gold-standard test.⁴² This will often occur
816 where a primary study uses a two stage design, where all patients receive the index test in
817 the first stage, but only a subsample receives the reference test in the second stage. The
818 reference test is required to verify if the tested individuals did or did not have the target
819 indication. When selection of subjects for the reference standard is not completely random,
820 verification bias will occur.⁴² When verification bias is present, it will often lead to an
821 overestimate of the sensitivity of the index test.²² To prevent misleading comparisons,
822 estimates from a trial with a series or multi-stage design must always be described in
823 context of the trial design and study population.⁴³

824 **2.4.5. Bias resulting from choice of cut-off points**

825 A data-driven approach to the selection of the optimal cut-off values can result in overly
826 optimistic estimates of sensitivity and specificity, particularly in small studies.⁴⁴ Using
827 simulation it has been shown that data-driven cut-off points frequently exaggerate test
828 performance, and this bias probably affects many published diagnostic validity studies.⁴⁵
829 Bias can be reduced by optimising cut-off points using a training dataset and then applying
830 those cut-off points to second test set of data. However, such an approach is reliant on
831 sufficient data availability, which is frequently problematic when considering diagnostic test
832 accuracy studies. Pre-specified cut-off points improve the validity of diagnostic test
833 research, and this is particularly the case for studies with small samples. Alternative
834 methods can be used to reduce this bias, but finding robust estimates for cut-off values
835 and accuracy requires considerable sample sizes.⁴⁴

836

837 **2.4.6. Disease prevalence**

838 Although contrary to typical assumptions, the sensitivity and specificity of a diagnostic test
839 can vary with disease prevalence.⁴⁶ This effect is likely to be the result of a number of
840 mechanisms, including patient spectrum, which affect prevalence, sensitivity and
841 specificity. Trivariate generalised linear mixed models have been applied to jointly model

842 prevalence, sensitivity and specificity, enabling the assessment of correlations between
843 the three parameters.^{47;48}

844 **2.4.7. Potential for dependence in combined tests**

845 For the purpose of this guideline, it is presumed that combined tests are not repeated
846 applications of the same test, as might happen in a screening programme, but rather the
847 use of a variety of tests with the aim of increasing the overall diagnostic accuracy.

848 When investigating combined tests, or tests carried out in sequence, the correlation
849 between test results is important. Two perfectly correlated tests will return the same
850 results, and hence the second test does not add any information from a diagnostic point of
851 view. This is important for the clinician: if two correlated tests are treated as independent
852 then the post-test probability of disease will be over-estimated by two positive tests.⁴⁹
853 From a meta-analytic point of view, combined tests can give rise to a number of problems,
854 not least a multiplication of the issues for single diagnostic tests.

855 Where multiple tests are used for diagnosis, it is highly likely that the tests will not perform
856 independently.⁵⁰ That is, in the case of two tests, the performance of the second test may
857 depend on the results of the first test. When the assumption of dependence between tests
858 is ignored, this can lead to erroneous disease probability estimates.⁵⁰

859 A further issue is that patients testing positive may be removed from the tested population
860 to receive treatment. This change to the population may affect the disease prevalence and
861 may also introduce spectrum bias.

862 **2.4.8. Missing data/non-evaluable results**

863 Reports of diagnostic test accuracy studies will sometimes refer to missing data or non-
864 evaluable results. This may be done explicitly in the text or it may be apparent from the
865 2x2 tables where the numbers of tests are inconsistent. A potential for bias exists if the
866 number of patients enrolled differs from the number of patients included in the 2x2 table of
867 results, as patients lost to follow-up differ systematically from the remaining patients.⁵¹
868 Missing data can occur for a variety of legitimate reasons. For example, if a patient is to
869 receive two different tests and is clearly positive after the first, it may be unethical to
870 subject them to the second test if that may cause delays in treatment. Non-evaluable
871 results can occur where the results of what is intended to be a dichotomous measure
872 cannot be unequivocally classified. The exclusion of non evaluable results leads to the
873 overestimation of diagnostic accuracy.⁵² One potential solution is to adopt an intention to
874 diagnose approach, which can be formulated as a 3x2 table in which non-evaluable results
875 are included.⁵² Such an analysis can significantly decrease the estimate of diagnostic
876 performance.

877 **2.4.9. Individual patient data analysis**

878 Individual patient data meta-analysis enables the evaluation of diagnostic test accuracy in
879 relation to other relevant information.⁵³ This approach could increase the efficiency of the
880 diagnostic work-up by, for example, reducing the need for invasive confirmatory
881 tests.^{13;53;54} The addition of clinical information when interpreting the results of diagnostic
882 tests can also improve accuracy;⁵⁶ if this has been applied to some, but not all patients in a
883 study, it could be recorded as a covariate and used in an individual patient analysis.
884 Allowing for individual patient characteristics can also allow for proper accounting of

885 differences in the patient spectrum, and to enable test results to be interpreted based on
886 additional patient information.²⁵

887 **2.5. Meta-analysis of the prognostic utility of a diagnostic test**

888 A prognostic factor is typically a biomarker that is used to predict future events, such as
889 disease progression or mortality. Studies of prognostic factors aim to estimate the
890 relationship between the prognostic factor and an outcome. In some instances, prognosis
891 is based on a measure such as a relative risk or hazard ratio, in which case the meta-
892 analytic approach would be a univariate analysis. Where studies present prognostic
893 information as a 2x2 table then methods used for diagnostic test accuracy studies may be
894 appropriate.

895 As they are similar to diagnostic tests in a number of regards, the meta-analysis of
896 prognostic factors face similar issues to those of diagnostic tests. Systematic reviews of
897 prognostic factors are often affected by the difficulty in comprehensively identifying
898 relevant studies. More so than diagnostic test accuracy studies, there is a strong risk of
899 publication bias.⁵⁷ There is also a likelihood that many prognostic factors may be
900 evaluated in a single study, but only those that show a high predictive value are reported.
901 The selective reporting can mean that although the same prognostic factor may have been
902 evaluated in numerous studies, it may be selectively reported giving a biased impression
903 of its predictive power. Equally, although the relevant biomarker may be consistently used,
904 the method of measurement may vary substantially.

905 Data extraction from identified studies can also be problematic because different methods
906 of presentation may have been used. The prognostic measure, and often the outcome, are
907 frequently measured on a continuous scale (e.g., tumour size) and may be recorded as
908 longitudinal data. The manner in which these data are handled and presented can vary
909 substantially. Results may be adjusted for relevant covariates, including other prognostic
910 variables. Different studies will vary because of different choices of covariates (if any) and
911 different methods of adjustment.

912 One solution is to utilise individual patient data (IPD), as this facilitates incorporation of
913 quite detailed data. Grouped data can lose the associations between different prognostic
914 measures that may be very important.

915 **2.6. Assessing the quality of studies and meta-analysis**

916 An important component of any systematic review or meta-analysis is a formal
917 assessment of study quality, and the detailed reporting of methodology and findings. A
918 number of initiatives have taken place with a view to improving the quality of published
919 studies for both diagnostic accuracy studies and subsequent meta-analyses.

920 **2.6.1. STARD**

921 The Standards for Reporting of Diagnostic Accuracy (STARD) initiative was started with a
922 view to improving the accuracy and completeness of reporting of studies of diagnostic
923 accuracy.⁵⁸ In doing so, it was hoped that readers would be able to assess the potential for
924 bias in a study, and to evaluate a study's generalisability. The STARD checklist was
925 published in a number of journals and adopted by some as a requirement for submitting
926 diagnostic test accuracy studies. However, the impact of the initiative on the quality of
927 reporting has been questioned.^{59:60} While the STARD initiative applies to the reporting of
928 primary research, poor reporting can be indicative of poor study quality.

929 **2.6.2. QUADAS**

930 The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool was originally
931 developed in 2003 and subsequently refined and updated in 2011 as QUADAS-2.⁵¹ The
932 tool assesses study quality in four domains: patient selection, index test, reference
933 standard, and flow and timing. Each domain is assessed in terms of risk of bias, and
934 concerns regarding applicability (for the first three domains). Signalling questions are used
935 to assist judgement regarding risk of bias. Application of the tool results in a judgement of
936 risk of bias for each study categorised as low, high, or unclear. These judgements can be
937 used to exclude studies from the primary analysis or to guide sensitivity analyses.
938 Although it is the only validated tool for assessing the quality of diagnostic test accuracy
939 studies, it should be noted that QUADAS-2 does not include specific criteria for assessing
940 comparative studies although it is possible to adapt the tool for this purpose.⁶¹

941 **2.6.3. PRISMA**

942 Having identified relevant studies for a meta-analysis, assessed their risk of bias and
943 undertaken evidence synthesis through meta-analysis, the results must then be reported.
944 The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)
945 statement outlines an evidence-based minimum set of items for reporting in systematic
946 reviews and meta-analyses.⁶² The PRISMA checklist was designed for systematic reviews
947 and meta-analyses in general, and the authors acknowledged that the checklist may need
948 to be modified when the research question related to diagnostic or prognostic
949 interventions. One of the key principles underpinning the PRISMA statement is that
950 authors ensure that their methods are reported with sufficient clarity and transparency so
951 that readers can critically judge the presented evidence and replicate the research. An
952 analysis of the impact of PRISMA on the reporting of meta-analyses in diagnostic research
953 has suggested that there are still issues in quality of reporting such studies.⁶³ However, a
954 modified version of PRISMA for the reporting of diagnostic test accuracy meta-analyses
955 provides the best prospect of achieving good quality reporting.

956 **2.6.4. GRADE**

957 The Grading of Recommendations Applicability, Development and Evaluation (GRADE)
958 approach provides framework for considering the quality of evidence regarding
959 interventions, and can be applied to diagnostic tests in terms of their impact on patient-
960 relevant outcomes.⁶⁴ The use of GRADE is often in the context of developing clinical
961 guidelines and recommendations regarding the appropriate use of a technology or health
962 intervention.

963 **2.7. Software**

964 A variety of software packages have been used in the literature for carrying out the meta-
965 analysis techniques described in these guidelines. In terms of the bivariate and HSROC
966 approaches, implementations have been documented for proprietary programs SAS and
967 Stata. Coded implementations in SAS have been published in a number of studies, while
968 the metandi module for Stata computes results for both methods (without covariates).
969 MLwiN is a package created by the University of Bristol for fitting multilevel models and
970 can be applied to both techniques. The techniques can also be applied through free
971 software packages R and WinBUGS. The latter program is for analyses in a Bayesian
972 framework and code for both methods has been published. Functions for the bivariate and
973 HSROC methods in R are provided through a number of freely available packages. It
974 should be noted that a variety of implementations are available in R with different default

975 parameterisations, so users should pay careful attention to what methodology is coded
976 into each function.

977 In all cases, it is critical that the user understands how the method has been implemented
978 and what parameters can be set and what outputs are provided. Correct interpretation of
979 the output is contingent on understanding how the computations have been carried out
980 and whether the underlying assumptions are correct. Other than reporting convergence,
981 most packages will give limited information on whether the pooled estimates and
982 parameter values are valid.

983

984 Table 1. Software implementations of methods for the meta-analysis of diagnostic test
985 accuracy studies

Software	Meta-analysis method		
	Moses-Littenburg	Hierarchical SROC	Bivariate random effects
RevMan*	✓	✗	✗
Meta-DiSc*	✓	✗	✗
SPSS	✓	✗	✗
SAS	✓	✓	✓
Stata	✓	✓	✓
MLwiN ⁺	✓	✓	✓
R*	✓	✓	✓
WinBUGS/OpenBUGS*	✓	✓	✓

Notes: * Free software; ⁺ free to UK academics.

986

987 Of the software packages listed in Table 1, some (RevMan, metaDisc, R, Stata) contain
988 specific commands with implemented versions of meta-analysis methods, while some
989 (SAS, SPSS, Stata, R) allow for the computation of the corresponding algorithms. The
990 distinction being that the latter may allow for greater flexibility in how the algorithms are
991 applied.

992

993

994

995 3. Conclusion and main recommendations

996

997 The meta-analysis of diagnostic test accuracy studies can be used to generate a more
998 precise estimate by pooling data from a number of studies. Diagnostic test accuracy is not
999 a measure of clinical effectiveness and improved accuracy does not necessarily imply
1000 improved patient outcomes. There are a variety of metrics available for describing
1001 diagnostic test accuracy, although the measures most commonly summarised in a meta-
1002 analysis are sensitivity and specificity (or the corresponding true positive rate and false
1003 positive rate). Due to the likelihood of a negative correlation between sensitivity and
1004 specificity, a meta-analysis of the two measures should take this relationship into account.
1005 While a number of methodological approaches are available for the meta-analysis of
1006 diagnostic test accuracy studies, the HSROC and bivariate methods are the most
1007 appropriate. These techniques have been implemented in a variety of software
1008 environments. There are numerous forms of bias that can affect estimates of diagnostic
1009 test accuracy in individual studies. All studies included in a meta-analysis should be
1010 carefully scrutinised to ensure they are equivalent and suitable for meta-analysis.
1011 Sensitivity analysis is a useful approach for testing the influence of studies with a high risk
1012 of bias.

1013 Based on the preceding sections, a number of recommendations are proposed:

- 1014 1. Pooling studies of diagnostic test accuracy should only be undertaken when there
1015 are sufficient studies available. When only two studies are available, it is not
1016 recommended to undertake a meta-analysis, and reporting should be restricted to a
1017 narrative description of the available evidence.
- 1018 2. The quality of studies being pooled should be assessed using a recognised and
1019 validated quality assessment tool.
- 1020 3. Pooled studies should be equivalent in terms of the index test, the reference
1021 standard, the patient population and the indication.
- 1022 4. Where important differences are identified across studies in terms of disease
1023 spectrum, study setting, and disease prevalence, these should be accounted for by
1024 including covariates.
- 1025 5. Where potential study differences occur but cannot be readily accounted for, such
1026 as verification bias, these should be clearly identified and the potential impacts
1027 should be determined.
- 1028 6. The appropriate methods of meta-analysis are the hierarchical SROC and bivariate
1029 random effects techniques, unless there is an absence of heterogeneity in either
1030 FPR or TPR, in which case two separate univariate meta-analyses may be more
1031 appropriate.
- 1032 7. The appropriate approach to meta-analysis is defined with respect to the quantity of
1033 data, between-study heterogeneity, threshold effects, and the correlation between
1034 TPR and FPR.
- 1035 8. The reporting of meta-analysis should include all the information that justifies the
1036 choice of analytical approach and supports the exclusion of alternative approaches.

1037

1038

1039 ***Recommendations for those undertaking meta-analyses***

1040 For researchers undertaking a meta-analysis of diagnostic test accuracy studies, a
1041 minimum set of information must be reported:

- 1042 1. The included studies must be described in detail in terms of both similarities and
1043 differences in key components (e.g., index test, reference test, population,
1044 indication, test threshold, prevalence).
- 1045 2. Report the quality assessment of the included studies.
- 1046 3. The decision process that leads to the selection of the appropriate methodology
1047 must be clearly described.
- 1048 4. All of the estimated parameter values should be clearly reported along with their
1049 corresponding confidence or credibility intervals.
- 1050 5. Appropriate graphical outputs should be provided including forest plots, SROC (if
1051 computed), and prediction regions.
- 1052 6. Report the possible impact of different forms of bias on the results.

1053

1054 ***Recommendations for those reading meta-analyses***

1055 For those reading a meta-analysis of diagnostic test accuracy studies, certain key
1056 information must be included in order to appraise the findings:

- 1057 1. Were the included studies comparable in terms of the key features (e.g., index test,
1058 reference test, population, indication, test threshold, prevalence)?
- 1059 2. Were the included studies of acceptable quality?
- 1060 3. Was the methodology used appropriate given the nature of the included evidence?
- 1061 4. Were all of the estimated parameter values clearly reported and their values
1062 interpreted?
- 1063 5. Were the relevant graphical outputs provided including forest plots, SROC (if
1064 computed), and prediction regions?
- 1065 6. Were the possible effects of different forms of bias on the results clearly reported
1066 and supported with relevant sensitivity analyses?
- 1067 7. Were the conclusions drawn consistent with the evidence analysed?

1068

1069

Annexe 1. Bibliography

1070

- 1071 (1) Knottnerus JA, van Weel C. General introduction: evaluation of diagnostic
1072 procedures. In: Knottnerus JA, editor. The evidence base of clinical diagnosis.
1073 London: BMJ Books; 2002. 81-94.
- 1074 (2) Ebell MH. Evidence-based diagnosis. New York: Springer-Verlag; 2001.
- 1075 (3) Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of
1076 diagnostic and screening tests. *BMJ* 2001; 323(7305):157-162.
- 1077 (4) Pewsner D, Battaglia M, Minder C, Marx A, Bucher, einer C. et al. Ruling a
1078 diagnosis in or out with "SpIn" and "SnNOut": a note of caution. *BMJ* 2004; 329.
- 1079 (5) Smits N. A note on Youden's J and its cost ratio. *BMC Med Res Methodol* 2010;
1080 10(1):89.
- 1081 (6) Chen L, Reisner AT, Chen X, Gribok A, Reifman J. Are standard diagnostic test
1082 characteristics sufficient for the assessment of continual patient monitoring? *Med*
1083 *Decis Making* 2013; 33(2):225-234.
- 1084 (7) Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In:
1085 Egger M, Davey Smith G, Altman DG, editors. *Systematic Reviews in Health Care*.
1086 London: BMJ Books; 2001. 248-284.
- 1087 (8) Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ* 1994;
1088 309(6947):102.
- 1089 (9) Eusebi P. Diagnostic accuracy measures. *Cerebrovasc Dis* 2013; 36:267-272.
- 1090 (10) Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing
1091 the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ*
1092 2012; 344:e686.
- 1093 (11) Knottnerus JA, Dinant G-J, van Schayck OP. The diagnostic before-after study to
1094 assess clinical impact. In: Knottnerus JA, editor. The evidence base of clinical
1095 diagnosis. London: BMJ Books; 2002. 81-94.
- 1096 (12) Staub LP, Dyer S, Lord SJ, Simes RJ. Linking the evidence: intermediate outcomes
1097 in medical test assessments. *Int J Technol Assess Health Care* 2012; 28(1):52-58.
- 1098 (13) Broeze KA, Opmeer BC, van d, V, Bossuyt PM, Bhattacharya S, Mol BW. Individual
1099 patient data meta-analysis: a promising approach for evidence synthesis in
1100 reproductive medicine. *Hum Reprod Update* 2010; 16(6):561-567.
- 1101 (14) Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess
1102 medical tests: a critical analysis. *Int J Technol Assess Health Care* 2013; 29(3):343-
1103 350.
- 1104 (15) Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient
1105 to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern*
1106 *Med* 2006; 144(11):850-855.
- 1107 (16) Jarvik JG. Fundamentals of Clinical Research for Radiologists: The Research
1108 Framework. *Am J Roentgenol* 2001; 176(4):873-878.
- 1109 (17) Krupinski EA, Jiang Y. Anniversary paper: evaluation of medical imaging systems.
1110 *Med Phys* 2008; 35(2):645-659.
- 1111 (18) Thornbury JR. Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging:
1112 love it or leave it. *AJR Am J Roentgenol* 1994; 162(1):1-8.
- 1113 (19) *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 ed. The
1114 Cochrane Collaboration; 2011.
- 1115 (20) Harbord RM, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A et al. An
1116 empirical comparison of methods for meta-analysis of diagnostic accuracy showed

- 1117 hierarchical models are necessary. *Journal of Clinical Epidemiology* 2008;
1118 61(11):1095-1103.
- 1119 (21) Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate
1120 for diagnostic meta-analyses? *Stat Med* 2009; 28(21):2653-2668.
- 1121 (22) Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic
1122 test accuracy. *Ann Intern Med* 2008; 149(12):889-897.
- 1123 (23) Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH.
1124 Bivariate analysis of sensitivity and specificity produces informative summary
1125 measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; 58(10):982-
1126 990.
- 1127 (24) Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of
1128 diagnostic test accuracy evaluations. *Stat Med* 2001; 20(19):2865-2884.
- 1129 (25) Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JAC. A unification of models for
1130 meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; 8(2):239-251.
- 1131 (26) Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with
1132 SAS PROC GLIMMIX. *Methods Inf Med* 2010; 49(1):54-64.
- 1133 (27) Macaskill P, Gatsonis CA, Deeks JJ, Harbord R, Takwoingi Y. Chapter 10:
1134 Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors.
1135 *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version*
1136 *1.0. The Cochrane Collaboration; 2010.*
- 1137 (28) Begg CB. Meta-analysis methods for diagnostic accuracy. *Journal of Clinical*
1138 *Epidemiology* 2008; 61(11):1081-1082.
- 1139 (29) Dahabreh IJ, Trikalinos TA, Lau J, Schmid C. An Empirical Assessment of Bivariate
1140 Methods for Meta-Analysis of Test Accuracy. No. 12(13)-EHC136-EF. 2012.
1141 Rockville, Maryland, Agency for Healthcare Research and Quality.
- 1142 (30) Harbord RM, Whiting P. metandi: Meta-analysis of diagnostic accuracy using
1143 hierarchical logistic regression. In: Sterne JAC, Newton HJ, Cox NJ, editors. *Meta-*
1144 *Analysis in Stata. Texas: Stata Press; 2009. 181-199.*
- 1145 (31) Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new
1146 tests against existing diagnostic pathways. *BMJ* 2006; 332(7549):1089-1092.
- 1147 (32) Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of
1148 comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013; 158(7):544-
1149 554.
- 1150 (33) Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH et al.
1151 Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;
1152 282(11):1061-1066.
- 1153 (34) de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Chapter 7:
1154 Searching for Studies. *Cochrane Handbook for Systematic Reviews of Diagnostic*
1155 *Test Accuracy. Version 0.4 ed. The Cochrane Collaboration; 2008.*
- 1156 (35) Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic
1157 reviews of diagnostic tests was difficult due to the poor sensitivity and precision of
1158 methodologic filters and the lack of information in the abstract. *J Clin Epidemiol*
1159 2005; 58(5):444-449.
- 1160 (36) Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of
1161 methodological filters in searches for diagnostic test accuracy studies misses
1162 relevant studies. *J Clin Epidemiol* 2011; 64(6):602-607.
- 1163 (37) Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C et al.
1164 Challenges in systematic reviews of diagnostic technologies. *Ann Intern Med* 2005;
1165 142(12 Pt 2):1048-1055.

- 1166 (38) Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication
1167 bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002; 31(1):88-95.
- 1168 (39) Leeflang MMG, Deeks JJ, Rutjes AWS, Reitsma JB, Bossuyt PMM. Bivariate meta-
1169 analysis of predictive values of diagnostic tests can be an alternative to bivariate
1170 meta-analysis of sensitivity and specificity. *Journal of Clinical Epidemiology* 2012;
1171 65(10):1088-1097.
- 1172 (40) Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy
1173 evaluations: Methodologic primer. *Am J Roentgenol* 2006; 187(2):271-281.
- 1174 (41) Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in
1175 diagnostic test evaluation. *Ann Intern Med* 2002; 137(7):598-602.
- 1176 (42) de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L et al.
1177 Adjusting for partial verification or workup bias in meta-analyses of diagnostic
1178 accuracy studies. *Am J Epidemiol* 2012; 175(8):847-853.
- 1179 (43) Ringham BM, Alonzo TA, Grunwald GK, Glueck DH. Estimates of sensitivity and
1180 specificity can be biased when reporting the results of the second test in a
1181 screening trial conducted in series. *BMC Med Res Methodol* 2010; 10:3.
- 1182 (44) Leeflang MMG, Moons KGM, Reitsma JB, Zwinderman AH. Bias in Sensitivity and
1183 Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms,
1184 Magnitude, and Solutions. *Clinical Chemistry* 2008; 54(4):729-737.
- 1185 (45) Ewald B. Post hoc choice of cut points introduced bias to diagnostic research.
1186 *Journal of Clinical Epidemiology* 2006; 59(8):798-801.
- 1187 (46) Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's
1188 sensitivity and specificity with disease prevalence. *CMAJ* 2013.
- 1189 (47) Kuss O, Hoyer A, Solms A. Meta-analysis for diagnostic accuracy studies: a new
1190 statistical model using beta-binomial distributions and bivariate copulas. *Stat Med*
1191 2014; 33(1):17-30.
- 1192 (48) Ma X, Nie L, Cole SR, Chu H. Statistical methods for multivariate meta-analysis of
1193 diagnostic tests: An overview and tutorial. *Stat Methods Med Res* 2013.
- 1194 (49) van Walraven C, Austin PC, Jennings A, Forster AJ. Correlation between serial
1195 tests made disease probability estimates erroneous. *Journal of Clinical
1196 Epidemiology* 2009; 62(12):1301-1305.
- 1197 (50) Novielli N, Cooper NJ, Sutton AJ. Evaluating the Cost-Effectiveness of Diagnostic
1198 Tests in Combination: Is It Important to Allow for Performance Dependency? *Value
1199 in Health* 2013; 16(4):536-541.
- 1200 (51) Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al.
1201 QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy
1202 studies. *Ann Intern Med* 2011; 155(8):529-536.
- 1203 (52) Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to
1204 diagnose approach to assess clinical performance of diagnostic tests: meta-
1205 analytical evaluation of coronary CT angiography studies. *BMJ* 2012; 345:e6717.
- 1206 (53) Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data
1207 meta-analysis to evaluate diagnostic tests. *European Journal of Obstetrics &
1208 Gynecology and Reproductive Biology* 2003; 108(2):121-125.
- 1209 (54) Broeze KA, Opmeer BC, Coppus SFPJ, Van Geloven N, Alves MFC, Å...nestad G
1210 et al. Chlamydia antibody testing and diagnosing tubal pathology in subfertile
1211 women: an individual patient data meta-analysis. *Hum Reprod Update* 2011;
1212 17(3):301-310.
- 1213 (55) Broeze KA, Opmeer BC, Coppus SF, Van Geloven N, Den Hartog JE, Land JA et
1214 al. Integration of patient characteristics and the results of Chlamydia antibody

- 1215 testing and hysterosalpingography in the diagnosis of tubal pathology: an individual
1216 patient data meta-analysis. *Human Reproduction* 2012; 27(10):2979-2990.
- 1217 (56) Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical
1218 information: a systematic review. *JAMA* 2004; 292(13):1602-1609.
- 1219 (57) Altman DG. Systematic reviews of evaluations of prognostic variables. In: Egger M,
1220 Davey Smith G, Altman DG, editors. *Systematic Reviews in Health Care*. London:
1221 BMJ Books; 2001. 228-247.
- 1222 (58) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al.
1223 Towards complete and accurate reporting of studies of diagnostic accuracy: the
1224 STARD initiative. *BMJ* 2003; 326(7379):41-44.
- 1225 (59) Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB et al.
1226 The quality of diagnostic accuracy studies since the STARD statement: has it
1227 improved? *Neurology* 2006; 67(5):792-797.
- 1228 (60) Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since
1229 STARD statement publication--before-and-after study. *Radiology* 2008; 248(3):817-
1230 823.
- 1231 (61) Wade R, Corbett M, Eastwood A. Quality assessment of comparative diagnostic
1232 accuracy studies: our experience using a modified version of the QUADAS-2 tool.
1233 *Res Syn Meth* 2013; 4(3):280-286.
- 1234 (62) Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP et al. The
1235 PRISMA statement for reporting systematic reviews and meta-analyses of studies
1236 that evaluate health care interventions: explanation and elaboration. *Ann Intern Med*
1237 2009; 151(4):W65-W94.
- 1238 (63) Willis BH, Quigley M. The assessment of the quality of reporting of meta-analyses
1239 in diagnostic research: a systematic review. *BMC Med Res Methodol* 2011; 11:163.
- 1240 (64) Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE et al.
1241 Grading quality of evidence and strength of recommendations for diagnostic tests
1242 and strategies. *BMJ* 2008; 336(7653):1106-1110.
- 1243
1244
1245
1246
1247
1248

1249 **Annexe 2. Documentation of literature search**

1250

1251 **Keywords**

1252 Five keywords were defined that would enable identification of relevant literature:

- 1253 • diagnostic
- 1254 • test
- 1255 • accuracy
- 1256 • meta-analysis
- 1257 • systematic

1258

1259 **Search engines and sources of information**

1260 A variety of sources of information were identified to find published literature and
1261 information pertinent to the development of these guidelines.

1262 *Literature search*

- 1263 • EMBASE
- 1264 • MEDLINE
- 1265 • DARE
- 1266 • Cochrane Database of Systematic Reviews
- 1267 • CADTH/CEDAC
- 1268 • EBSCOhost

1269 *Internet search*

- 1270 • Google and Google Scholar
- 1271 • ScienceDirect
- 1272 • Wiley-Interscience
- 1273 • Hand searching of references cited in relevant documents
- 1274 • The Cochrane Collaboration
- 1275 • National Guideline Clearinghouse
- 1276 • National Institute for Health and Clinical Excellence
- 1277 • ISPOR
- 1278 • Pharmaceutical Benefits Advisory Committee (PBAC)
- 1279 • Centre for Reviews and Dissemination, University of York
- 1280 • University of Bristol

1281 *Guidelines search*

1282 The websites of EUnetHTA member agencies and those of major international
1283 agencies were searched for relevant guidelines.

1284 *Other specifically identified sources of information*

- 1285 • Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new
1286 tests against existing diagnostic pathways. *BMJ*. 2006; 332: 1089-1092.
- 1287 • Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The
1288 STARD statement for reporting studies of diagnostic accuracy: explanation and
1289 elaboration. *Ann Intern Med*. 2003; 138(1):W1-12.
- 1290 • Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate
1291 for diagnostic meta-analyses? *Statistics in Medicine*. 2009; 28(21): 2653-2668.
- 1292 • Harbord RM, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A, et al. An
1293 empirical comparison of methods for meta-analysis of diagnostic accuracy showed
1294 hierarchical models are necessary. *Journal of Clinical Epidemiology*. 2008; 61:
1295 1095-1103.
- 1296 • Moher D, Liberati A, Tetzlaff J, Altman DG, the PRISMA Group. Preferred Reporting
1297 Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann
1298 Intern Med*. 2009; 151(4):264-269.
- 1299 • Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic
1300 test into a summary roc curve: Data-analytic approaches and some additional
1301 considerations. *Statistics in Medicine*. 1993; 12: 1293-1316.
- 1302 • Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH.
1303 Bivariate analysis of sensitivity and specificity produces informative summary
1304 measures in diagnostic reviews. *Journal of Clinical Epidemiology*. 2005; 58: 982-
1305 990.
- 1306 • Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of
1307 diagnostic test accuracy evaluations. *Statistics in Medicine*. 2001; 20(19): 2865-
1308 2884.
- 1309 • Simel DL, Bossuyt PMM. Differences between univariate and bivariate models for
1310 summarizing diagnostic accuracy may not be large. *Journal of Clinical
1311 Epidemiology*. 2009; 62(12): 1292-1300.
- 1312 • Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in
1313 Medicine*. 2008; 27: 625-650.
- 1314 • Tawoingi Y, Leeflang MMG, Deeks JJ. Empirical Evidence of the Importance of
1315 Comparative Studies of Diagnostic Test Accuracy. *Annals of Internal Medicine*.
1316 2013; 158: 544-554.
- 1317 • Verde P. Meta-analysis of diagnostic test data: A bivariate Bayesian modeling
1318 approach. *Statistics in Medicine*. 2010; 29: 3088-3102
- 1319 • Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In:
1320 Egger M, Davey Smith G, Altman DG (eds). *Systematic Reviews in Health Care*.
1321 BMJ Books. London, 2001.
- 1322 • Diagnostic Test Accuracy Working Group. Handbook for DTA Reviews Version
1323 1.0.1. Cochrane Collaboration, 2009.
- 1324 • Harbord RM, Whiting P. Metandi: Meta-analysis of diagnostic accuracy using
1325 hierarchical logistic regression. In: Sterne JAC (ed). *Meta-Analysis in Stata – An
1326 Updated Collection from the Stata Journal*. Stata Press. Texas, 2009.
- 1327 • Health Information and Quality Authority. Guidelines for Evaluating the Clinical
1328 Effectiveness of Health Technologies in Ireland. HIQA. Dublin, 2011.

1329

1330 **Strategies of research**

1331 Reports, papers and other guidance documents were assessed on the basis of whether
1332 they described, applied or assessed methods of meta-analysis for diagnostic test accuracy
1333 studies. Documents that only mentioned methods but did not describe, apply or assess
1334 them were disregarded after being checked for useful references. Documents that applied
1335 methods were used to determine the scope of application, utility and possible limitations of
1336 those methods. Finally, documents that assessed methods were used to compare
1337 methods directly and to elicit recommendations. Where relevant, the quality of studies was
1338 assessed using the STARD (Standards for Reporting for Reporting of Diagnostic
1339 Accuracy) or PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-
1340 Analyses) statements.

1341 For PubMed, the search was limited to the period 1990 to date (end June 2013). In
1342 EBSCO the search was limited to 1990 to 2013 (inclusive). In both cases the search was
1343 limited to English language publications and human subjects. Database searches used the
1344 following search strategy:

1345 (diagnostic[Title/Abstract]) AND test[Title/Abstract]) AND accuracy[Title/Abstract]) AND
1346 (meta-analysis[Title/Abstract]) OR systematic[Title/Abstract])

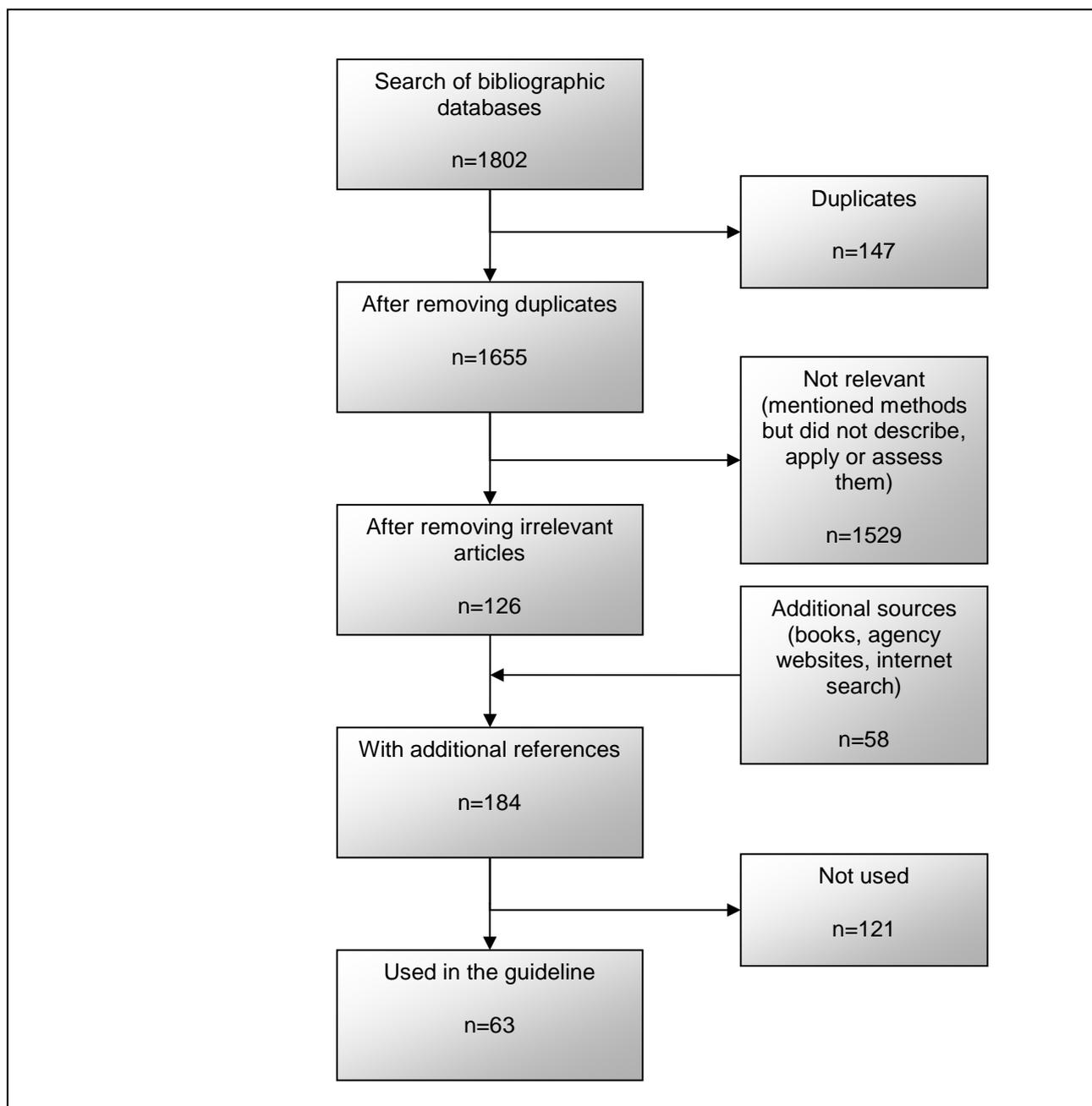
1347

1348 **Findings of literature search**

1349 The initial search returned 1802 articles that were potentially useful. After scanning titles,
1350 abstracts and, in some cases, full text, 126 articles were retained. Of these, 63 were
1351 ultimately used and referenced in the guidelines.

1352

1353 Figure 7. Flowchart of literature search.



1354

1355 **Annexe 3. Other sources of information**

1356

1357 No other sources of information were used.