Introduction to diagnostic test accuracy studies

Sitch, A.J.; Dekkers, O.M.; Scholefield, B.R.; Takwoingi, Y.

Introduction to diagnostic test accuracy studies

Alice J Sitch[1,2], **Olaf M Dekkers**[3,4], **Barnaby R Scholefield**[5,6] and **Yemisi Takwoingi**[1,2]

[1]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK, [2]Test Evaluation Research Group (TERG), Institute of Applied Health Research, University of Birmingham, Birmingham, UK, [3]Department of Clinical Epidemiology, [4]Department of Endocrinology, Leiden University Medical Center, Leiden, the Netherlands, [5]Birmingham Acute Care Research Group (BACR), Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK, and [6]Paediatric Intensive Care Unit, Birmingham Women & Children's Hospital NHS Foundation Trust, Birmingham, UK

Correspondence should be addressed to A J Sitch
Email
a.j.sitch@bham.ac.uk

Abstract

Diagnostic accuracy studies are fundamental for the assessment of diagnostic tests. Researchers need to understand the implications of their chosen design, opting for comparative designs where possible. Researchers should analyse test accuracy studies using the appropriate methods, acknowledging the uncertainty of results and avoiding overstating conclusions and ignoring the clinical situation which should inform the trade-off between sensitivity and specificity. Test accuracy studies should be reported with transparency using the STAndards for the Reporting of Diagnostic accuracy studies (STARD) checklist.

European Journal of Endocrinology
(2021) **184**, E5–E9

Introduction

Diagnosing diseases is crucial in medicine, and for this purpose, many diagnostic tests and procedures are applied. For the diagnosis of a suspected adrenal carcinoma a CT scan is performed, and an insulin tolerance test (ITT) for adrenal insufficiency. The performance of these tests can be investigated in diagnostic accuracy studies.

Medical diagnostic tests are evaluated in different ways, depending on the stage of evaluation and the purpose of the test. A fundamental aspect of the evaluation of diagnostic tests is test accuracy, that is, the ability of a test to differentiate between those who have and those who do not have the condition or disease of interest. In this article, we define key terminology (Fig. 1) used in the context of test accuracy, and describe basic aspects of study design and analysis.

Guidelines for reporting of test accuracy studies, the STAndards for the Reporting of Diagnostic accuracy studies (STARD) checklist (1), have been published and we recommend their use to increase quality and transparency of reporting.

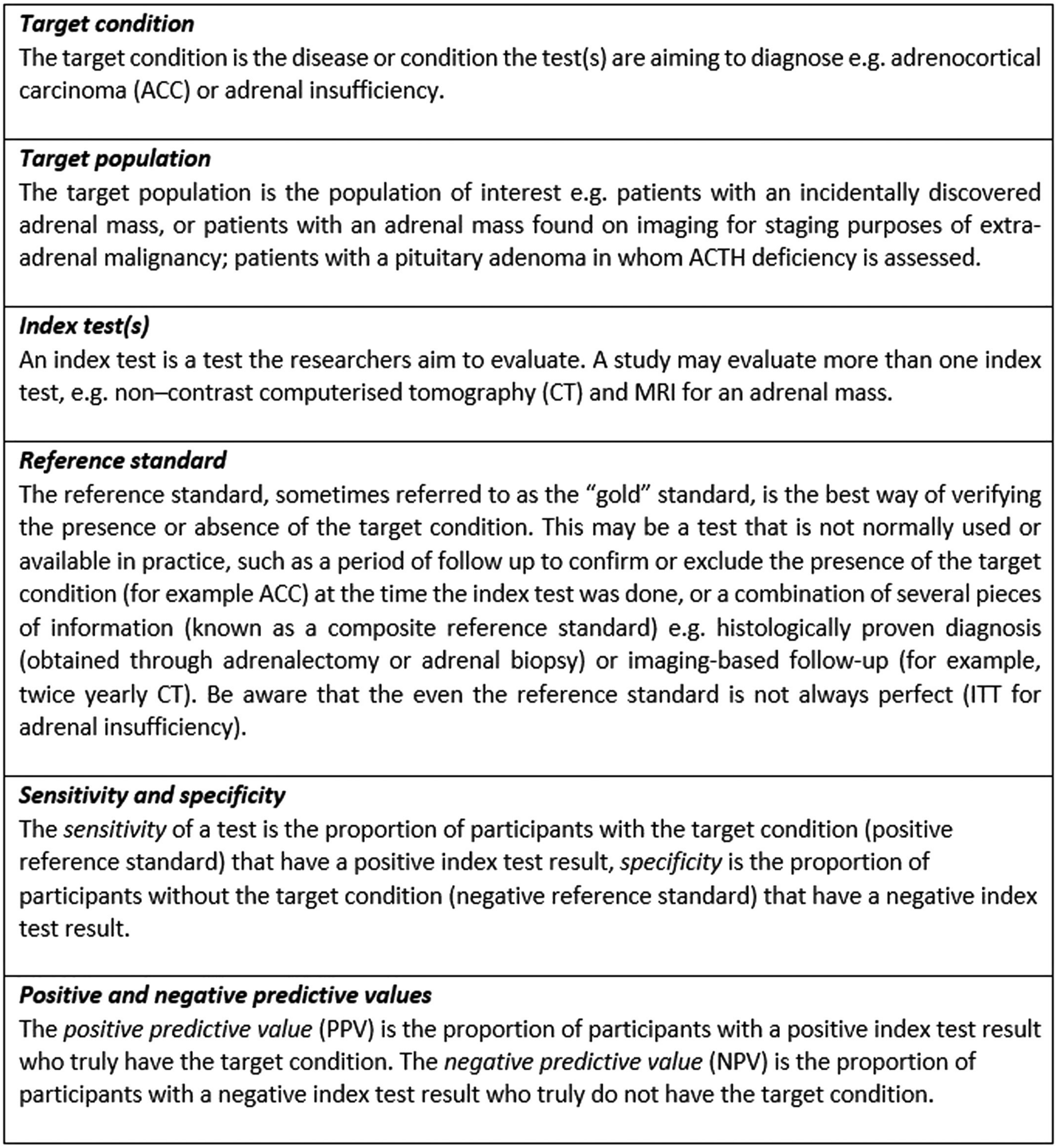
Measures of test accuracy

Test accuracy is determined by cross classifying the results (positive and negative) of an index test against those of the reference standard. This produces a two-by-two table giving the number of true positives, false positives, false negatives and true negatives (Fig. 2). Standard methods for estimating test accuracy require binary classification of the results of the index test and the reference standard. As such when test results are non-binary, criteria (referred to as thresholds, cut-offs or cut-points) are needed to define test negatives and test positives. For example, when assessing the test accuracy for the CRH-test for adrenal insufficiency, a cut-off needs to be defined.

Measures of test accuracy should always be accompanied by a 95% CI, which is a measure of uncertainty for the point estimate. In example given in Fig. 2, the 95% CI for the sensitivity ranges from 0.96 to 0.99; the 95% CI for specificity is wider, ranging from 0.67 to 0.78.

https://eje.bioscientifica.com
https://doi.org/10.1530/EJE-20-1239

Published by Bioscientifica Ltd.



Target condition

The target condition is the disease or condition the test(s) are aiming to diagnose e.g. adrenocortical carcinoma (ACC) or adrenal insufficiency.

Target population

The target population is the population of interest e.g. patients with an incidentally discovered adrenal mass, or patients with an adrenal mass found on imaging for staging purposes of extra-adrenal malignancy; patients with a pituitary adenoma in whom ACTH deficiency is assessed.

Index test(s)

An index test is a test the researchers aim to evaluate. A study may evaluate more than one index test, e.g. non–contrast computerised tomography (CT) and MRI for an adrenal mass.

Reference standard

The reference standard, sometimes referred to as the “gold” standard, is the best way of verifying the presence or absence of the target condition. This may be a test that is not normally used or available in practice, such as a period of follow up to confirm or exclude the presence of the target condition (for example ACC) at the time the index test was done, or a combination of several pieces of information (known as a composite reference standard) e.g. histologically proven diagnosis (obtained through adrenalectomy or adrenal biopsy) or imaging-based follow-up (for example, twice yearly CT). Be aware that the even the reference standard is not always perfect (ITT for adrenal insufficiency).

Sensitivity and specificity

The *sensitivity* of a test is the proportion of participants with the target condition (positive reference standard) that have a positive index test result, *specificity* is the proportion of participants without the target condition (negative reference standard) that have a negative index test result.

Positive and negative predictive values

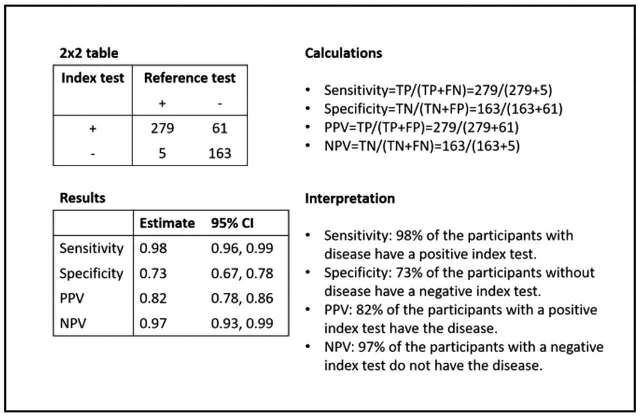
The *positive predictive value* (PPV) is the proportion of participants with a positive index test result who truly have the target condition. The *negative predictive value* (NPV) is the proportion of participants with a negative index test result who truly do not have the target condition.

Figure 1

Key terminology for test accuracy studies.

Study population and design

There are different phases in the evaluation of a diagnostic test. First, test performance is determined in a population of clearly established cases and non-cases (2, 3), this is referred to as proof-of-concept or exploratory study. Secondly, assessment in a representative population in an appropriate clinical setting (prospective consecutive recruitment of suspected cases) can be performed (4). The spectrum of



2x2 table

Index test	Reference test	
	+	-
+	279	61
-	5	163

Calculations

- Sensitivity=TP/(TP+FN)=279/(279+5)
- Specificity=TN/(TN+FP)=163/(163+61)
- PPV=TP/(TP+FP)=279/(279+61)
- NPV=TN/(TN+FN)=163/(163+5)

Results

	Estimate	95% CI
Sensitivity	0.98	0.96, 0.99
Specificity	0.73	0.67, 0.78
PPV	0.82	0.78, 0.86
NPV	0.97	0.93, 0.99

Interpretation

- Sensitivity: 98% of the participants with disease have a positive index test.
- Specificity: 73% of the participants without disease have a negative index test.
- PPV: 82% of the participants with a positive index test have the disease.
- NPV: 97% of the participants with a negative index test do not have the disease.

Figure 2

Example calculations, results and interpretation. TP, the number of true positive results; TN, the number of true positive results; FP, the number of false positive results; FN, the number of false negative results; PPV, the positive predictive value; NPV, the negative predictive value.

the disease will vary between these designs; researchers should be aware of this difference when planning studies and generalising results of studies to clinical settings (5, 6). When researchers perform an exploratory study involving known cases and non-cases (referred to as a diagnostic case–control or two-gate design (2)), (positive and negative) predictive values should not be directly calculated using two-by-two data from such studies. This is because predictive values are directly related to prevalence and the proportion of participants with the target condition in case–control studies is artificial, that is, determined by the study investigators. For example, doubling the number of cases would directly affect the calculated NPV and PPV. This is not the case when a representative population is sampled, for example, all pituitary adenoma patients with suspected ACTH deficiency.

Test accuracy studies often evaluate a single index test but where alternative tests exist that can be used at the same point in the diagnostic pathway (providing the tests do not interfere with each other and the patient burden is not too great), these tests can be evaluated in one study population (7) (Fig. 3). The ideal comparative study design is to perform all tests and the reference standard on all participants (paired or within-subject design) or to randomise participants to receive one of the index tests (8). The randomised design is preferred when it is not possible to perform multiple index tests on each individual for ethical or logistical reasons. Additionally, the role of the test in the diagnostic pathway – replacement, triage or add-on – should be considered when designing a study (8).

Sample size

Sample size calculations for test accuracy studies should be determined prior to recruitment; see (9, 10) for details. When evaluating a single test, a common approach is based on the precision around an estimate of sensitivity and/or specificity (i.e. the width of the CIs). The precision of the sensitivity estimate will increase with the number of participants with the target condition (reference standard positive) and the precision of the specificity estimate will increase with the number of participants without the target condition (reference standard negative). Hence, it is vital to have an estimate of the prevalence of the target condition in the study population to plan the sample size.

Statistical analysis

Measures of test accuracy (Fig. 1) can be calculated along with 95% CIs (11, 12). For test accuracy studies comparing two tests, additionally, the difference in sensitivity and specificity between the index tests can be computed. With the paired comparative design, McNemar's test can be used to test differences in sensitivity and specificity. Alternatively, regression modelling taking into account the paired nature of the data can be performed. The effect of important clinical characteristics on test accuracy can also be explored using such models; for example, it can be assessed whether age determines differences of two index tests. For the randomised comparative design, a test of independent proportions can be used to compare sensitivity and specificity between groups.

For tests with non-binary results, receiver operating characteristic (ROC) curve analysis is typically performed. There is a negative relationship between sensitivity and specificity as the cut-point changes (threshold effect); if we, for example, lower the cortisol threshold for the diagnosis of adrenal insufficiency, this will increase sensitivity (less false negatives), as a consequence, however, the specificity will be lower (more false positives). An ROC curve displays this trade-

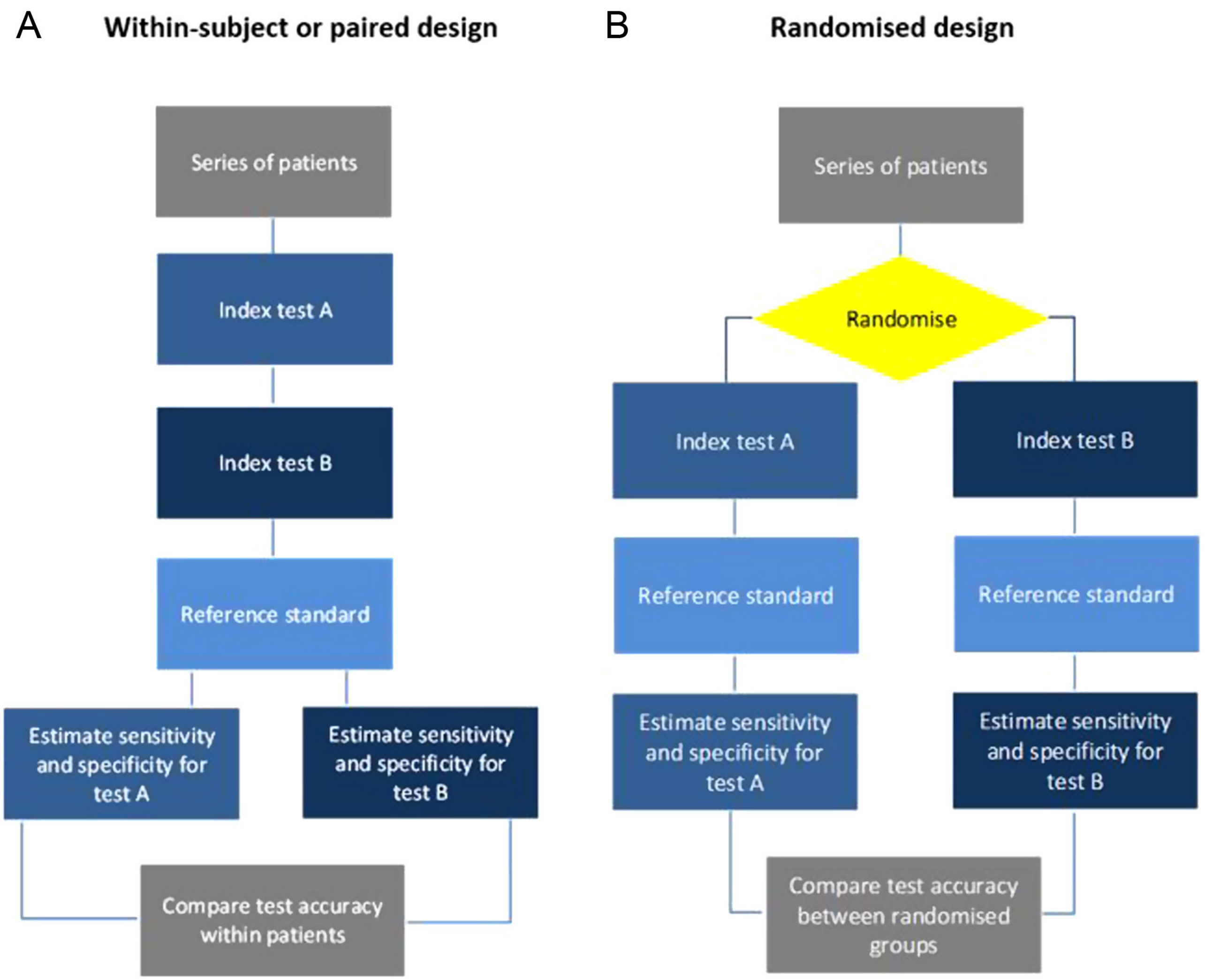

A
Within-subject or paired design
Series of patients
Index test A
Index test B
Reference standard
Estimate sensitivity and specificity for test A
Estimate sensitivity and specificity for test B
Compare test accuracy within patients
B
Randomised design
Series of patients
Randomise
Index test A
Index test B
Reference standard
Reference standard
Estimate sensitivity and specificity for test A
Estimate sensitivity and specificity for test B
Compare test accuracy between randomised groups

Figure 3
Robust study designs for comparing test accuracy (13). In (A) all patients undergo all index tests while in (B) patients are randomly assigned to only one of the index tests. In both (A) and (B), all patients receive the reference standard. Both designs are valid, although the paired design requires a smaller study sample.

off between sensitivity and specificity at different cut-points for a test (Fig. 4), and curves for different tests in a comparative study can be compared on a single ROC plot. A simplistic cut-point would maximise sensitivity and specificity. However, an appropriate cut-point for use in clinical practice should be driven by the consequences for false positive and false negative results. If a study is used to derive a cut-point for a test the performance, external validation is required, as a single study will likely overestimate the test's performance. This is especially the case for small studies.

Concluding remarks

Diagnostic test accuracy studies are required to understand the potential for new diagnostic technologies. It is vital that researchers understand the implications of the design of their studies and the impact on the study conclusions. Researchers need to understand the clinical situation and weigh the consequences of misidentifying positive and negative participants. There is a need for clear and transparent reporting allowing the limitations of studies to be identified. We encourage researchers to seek specialist support when embarking on these studies.

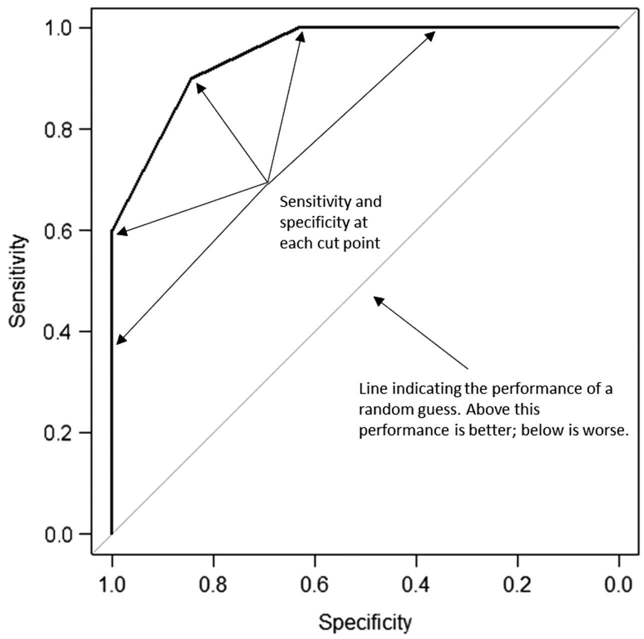


Figure 4
ROC curve example.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this editorial.

Funding

B S is funded by a National Institute for Health Research (NIHR) Clinician Scientist Fellowship (NIHR-CS-2015-15-016). Y T is funded by an NIHR Postdoctoral Fellowship. Y T is a co-convenor of the Cochrane Screening and Diagnostic Tests Methods Group. This work is supported by the NIHR Statistics Group (https://statistics-group.nihr.ac.uk/). A S and Y T are supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham.

Acknowledgements

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. A J S is an Advisory Editor and O M D is a Deputy Editor for *European Journal of Endocrinology*. Neither were involved in the review or editorial process for this paper, on which they are listed as authors.

References

1 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW *et al*. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015 **351** h5527. (https://doi.org/10.1136/bmj.h5527)

2 Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS & Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry* 2005 **51** 1335–1341. (https://doi.org/10.1373/clinchem.2005.048595)

3 Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M & Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 2001 **93** 1054–1061. (https://doi.org/10.1093/jnci/93.14.1054)

4 Sackett DL & Haynes RB. The architecture of diagnostic research. *BMJ* 2002 **324** 539–541. (https://doi.org/10.1136/bmj.324.7336.539)

5 Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM & Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of Internal Medicine* 2004 **140** 189–202. (https://doi.org/10.7326/0003-4819-140-3-200402030-00010)

6 Irwig L, Bossuyt P, Glasziou P, Gatsonis C & Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002 **324** 669–671. (https://doi.org/10.1136/bmj.324.7338.669)

7 Takwoingi Y, Leeflang MM & Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of Internal Medicine* 2013 **158** 544–554. (https://doi.org/10.7326/0003-4819-158-7-201304020-00006)

8 Bossuyt PM, Irwig L, Craig J & Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006 **21** 1089–1092. (https://doi.org/10.1136/bmj.332.7549.1089)

9 Obuchowski NA. Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research* 1998 **7** 371–392. (https://doi.org/10.1177/096228029800700405)

10 Pepe MS & Pepe PBMS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

11 Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927 **22** 209–212. (https://doi.org/10.1080/01621459.1927.10502953)

12 Brown LD, Cai TT & DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001 **16** 101–133. (https://doi.org/10.1214/ss/1009213286)

13 Takwoingi Y. Meta-analytic approaches for summarising and comparing the accuracy of medical tests [PHD thesis]. Birmingham, University of Birmingham; 2016. Accessed at https://etheses.bham.ac.uk/id/eprint/6759/ on 1st June 2020.

Received 27 October 2020
Accepted 3 November 2020