Appraising Diagnostic Test Studies

Martin Bland

Prof. of Health Statistics
Dept. of Health Sciences
University of York

http://www-users.york.ac.uk/~mb55/msc/

Diagnostic Test Studies

How well does a test

- identify people with the disease?
- exclude people without the disease?

Compare test results on people with the disease with test results on people without the disease.

Need to know who has the disease.

Diagnostic Test Studies

Two designs

Prospective or cohort design, or cross-sectional design:
take a sample of subject eligible for the the test, test them all and get true diagnosis on them all.

Retrospective or case-control design:
take a sample with true diagnosis established as positive and another sample of controls. We may have negative diagnosis established on controls and we may not.

Who has the disease?

True diagnosis.

We can never be absolutely sure that the 'true' diagnosis is correct.

We decide to accept one method as 'true':
call this the **gold standard** or **reference standard**.

Often more invasive than the test, e.g. histopathology compared to ultrasound image.

It is always possible that the reference standard is wrong for some subjects.

Statistics of diagnostic test studies

- Sensitivity
- Specificity
- Receiver operating characteristic curve (ROC curve)
- Likelihood ratio (LR) for positive test
- Likelihood ratio (LR) for negative test
- Odds ratio (OR)
- Positive predictive value (PPV)
- Negative predictive value (NPV)

Statistics of diagnostic test studies

Example: diabetic eye tests (cross-sectional)
test = direct opthalmoscopy
reference standard = slit lamp stereoscopic biomicroscopy

Single sample of subjects all received reference standard test.

Harding SP, Broadbent DM, Neoh C, White MC, Vora J. Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight threatening eye disease: the Liverpool diabetic eye study. *BMJ* 1995;311:1131-1135

Statistics of diagnostic test studies

Example: diabetic eye tests (cross-sectional)
test = direct opthalmoscopy
reference standard = slit lamp stereoscopic biomicroscopy

Single sample of subjects all received reference standard test.

	Reference			
Test	+ve	-ve	Total	sensitivity = 40/45 = 0.89 = 89%
+ve	40	38	78	specificity = 237/275 = 0.86 = 86%
-ve	5	237	242	LR (+ve test) = 0.89/(1–0.86) = 6.4
Total	45	275	320	LR (-ve test) = 0.86/(1–0.89) = 7.8
				OR = 40×237/(38×5) = 49.9
				PPV = 40/78 = 51%
				NPV = 237/242 = 98%

Statistics of diagnostic test studies

Sensitivity = proportion of reference positive cases who are positive on the test = proportion of true cases that the test finds.

Specificity = proportion of reference negative cases who are negative on the test = proportion of true non-cases that the test finds.

Example: eye disease in diabetics

45 reference standard positive cases of whom 40 were positive on the test, 275 reference standard negative non-cases of whom 237 were negative on the test.

Sensitivity = 40/45 = 0.89 = 89%.

Specificity = 237/275 = 0.86 = 86%.

Statistics of diagnostic test studies

Odds = number of positives/number of negative.

Odds ratio (OR) = odds in one group divided by odds in another.

Example: eye disease in diabetics

	Reference		
Test	+ve	-ve	Total
+ve	40	38	78
-ve	5	237	242
Total	45	275	320

Odds test +ve for those reference +ve = 40/5 = 8.0

OR = (40/5)/(38/237) = 40×237/(38×5) = 49.9

Statistics of diagnostic test studies

Likelihood ratio for a positive test
= sensitivity/(1 – specificity)

Start with the probability that the subject has the disease
= prevalence of disease

Convert to odds = prevalence/(1 – prevalence)

Odds of disease if test positive
= odds of disease × likelihood ratio

Example: eye disease in diabetics

Likelihood ratio for a positive test = 6.4

Suppose prevalence = 0.10 = 10%, odds = 0.10/0.90 = 0.11

Odds of disease if test positive = 0.11 × 6.4= 0.70

Probability = 0.41

Statistics of diagnostic test studies

Likelihood ratio for a negative test
= specificity/(1 – sensitivity)

Start with the probability that the subject does not have the disease = 1 – prevalence of disease

Convert to odds = (1 – prevalence)/prevalence

Odds of not disease if test negative
= odds of not disease × likelihood ratio

Example: eye disease in diabetics

Likelihood ratio for a negative test = 7.8

Suppose prevalence = 0.10 = 10%, odds = 0.90/0.10 = 9.0

Odds of no disease if test negative = 9.0 × 7.8 = 70.2

Probability = 0.986

Statistics of diagnostic test studies

Positive predictive value (PPV) = proportion of test positives who are reference positive.

Negative predictive value (NPV) = proportion of test negatives who are reference negative.

Example: eye disease in diabetics

78 test positives of whom 40 were positive on the reference standard, 242 test negatives of whom 237 were negative on the reference standard.

PPV = 40/78 = 51%.

NPV = 237/242 = 98%.

Statistics of diagnostic test studies

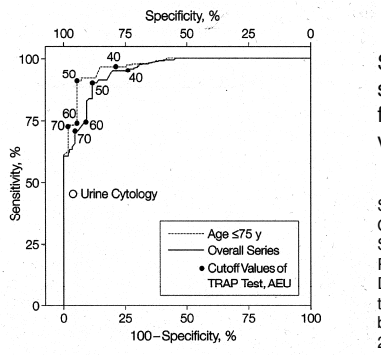
Example: early detection of bladder cancer (case-control)
test = urine telomerase > 50
reference standard =histologically confirmed bladder cancer

Case-control study conducted in 218 men: 84 healthy individuals and 134 patients at first diagnosis of histologically confirmed bladder cancer.

Sanchini MA, Gunelli R, Nanni O, Bravaccini S, Fabbri C, Sermasi A, Bercovich E, Ravaioli A, Amadori D, Calistri D. (2005) Relevance of urine telomerase in the diagnosis of bladder cancer. *JAMA* **294**, 2052-2056.

Statistics of diagnostic test studies

ROC curve: plot of sensitivity against 1 – specificity.



Sensitivity and specificity calculated for different cut-off values of test variable.

Sanchini MA, Gunelli R, Nanni O, Bravaccini S, Fabbri C, Sermasi A, Bercovich E, Ravaioli A, Amadori D, Calistri D. (2005) Relevance of urine telomerase in the diagnosis of bladder cancer. *JAMA* **294**, 2052-2056.

Statistics of diagnostic test studies

Example: early detection of bladder cancer (case-control)
test = urine telomerase > 50
reference standard =histologically confirmed bladder cancer

Case-control study.

	Reference		
Test	+ve	-ve	sensitivity = 120/134 = 0.90 = 90%
+ve	120	10	specificity = 74/84 = 0.88 = 88%
-ve	14	74	LR (+ve test) = 0.90/(1–0.88) = 7.5
total	134	84	LR (-ve test) = 0.88/(1–0.90) = 8.8
			OR = 120×74/(10×14) = 63.4
Row totals are meaningless.			PPV = ? --- cannot be found.
			NPV = ? --- cannot be found.

Cannot estimate PPV and NPV in case-control study.

Statistics of diagnostic test studies

A doctor writes . . .

"Doctors … now believe that [a man who went from HIV+ to HIV–] probably hasn't recovered from the disease but probably never had it in the first place, and that his first results were false positives. These are extremely rare because the HIV blood test – which checks for antibodies to the virus – is so sensitive. There are no official figures, but the experts estimate that the chance of getting a false positive is one in a million." – Dr. Simon Atkins, *The Guardian*, 17 November 2005.

Does a very sensitive test produce many false positives?

What is 'the chance of getting a false positive'?

Statistics of diagnostic test studies

Does a very sensitive test produce many false positives?

The more sensitive we make the test, the less specific it tends to become.

We expect that the more sensitive a test is the more false positives it will produce, not fewer, as this author seems to think.

Statistics of diagnostic test studies

What is 'the chance of getting a false positive'?

A false positive means that the test is positive but the person does not have the disease.

Probability of a truly HIV negative subject being test positive is one minus the specificity, also called the false positive rate.

Probability of a test positive subject being truly HIV negative is one minus the positive predictive value.

Could either the sensitivity or the PPV of a test be 0.999999 (one minus one in a million)?

Reproducibility in diagnostic test studies

Tests may fail because the biology is wrong or because the measurement error is too great.

The reference standard is assumed true, therefore has no measurement error (although it must do so in practice).

The test may have error, such as observer variation.

Some diagnostic studies incorporate reproducibility studies.

Reproducibility in diagnostic test studies

Example: Ottawa Ankle Rules (OAR) and Ottawa Foot Rules (OFR) applied by a specialized emergency nurse (SEN).

In a prospective study, all ankle sprains presented in the ED from April to July 2004 were assessed by both an SEN and a junior doctor, randomized for first observer. In all patients, radiography was performed (gold standard).

Derksen RJ, Bakker FC, Geervliet PC, de Lange-de Klerk ESM, Heilbron EA, Veenings B, Patka P, Haarman HJTM. (2005) Diagnostic accuracy and reproducibility in the interpretation of Ottawa ankle and foot rules by specialized emergency nurses. *American Journal of Emergency Medicine* **23**, 725-729.

Reproducibility in diagnostic test studies

Example: Ottawa Ankle Rules (OAR) and Ottawa Foot Rules (OFR) applied by a specialized emergency nurse (SEN).

	sensitivity	specificity	PPV	NPV
SENs	0.93	0.49	0.22	0.98
HOs	0.93	0.39	0.19	0.97

The interobserver agreement for the OAR and OFR subsets was kappa = 0.38 for the lateral malleous; kappa = 0.30, medial malleolus; kappa = 0.50, navicular; kappa = 0.45, metatarsal V base; and kappa = 0.43, weight-bearing. The overall interobserver agreement for the OAR was kappa = 0.41 and kappa = 0.77 for the OFR.

Appraisal of diagnostic test studies

Many sets of criteria or guidelines have been produced.

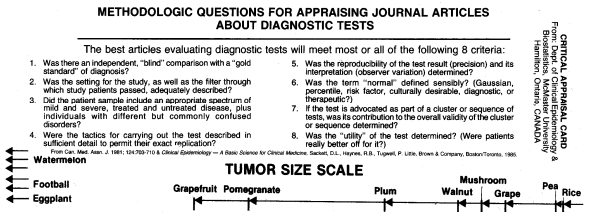
An early example: Sackett DL, Haynes RB, Guyatt GH, Tugwell P. (1991) *Clinical Epidemiology: a Basic Science for Clinical Medicine*, Little Brown, Chicago

Greenhalgh, T. (1997) How to read a paper: Papers that report diagnostic or screening tests. *BMJ* **315**, 540-543.

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* **3**, 25, http://www.biomedcentral.com/1471-2288/3/25

Sackett, Haynes, Guyatt, and Tugwell criteria

8 criteria on a plastic card to be carried in pocket.



METHODOLOGIC QUESTIONS FOR APPRAISING JOURNAL ARTICLES ABOUT DIAGNOSTIC TESTS

The best articles evaluating diagnostic tests will meet most or all of the following 8 criteria:

1. Was there an independent, "blind" comparison with a "gold standard" of diagnosis?
2. Was the setting for the study, as well as the filter through which study patients passed, adequately described?
3. Did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?
4. Were the tactics for carrying out the test described in sufficient detail to permit their exact replication?
5. Was the reproducibility of the test result (precision) and its interpretation (observer variation) determined?
6. Was the term "normal" defined sensibly? (Gaussian, percentile, risk factor, culturally desirable, diagnostic, or therapeutic?)
7. If the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined?
8. Was the "utility" of the test determined? (Were patients really better off for it?)

From Can. Med. Assn. J. 1981; 124:703-710 & *Clinical Epidemiology — A Basic Science for Clinical Medicine*. Sackett, D.L., Haynes, R.B., Tugwell, P. Little, Brown & Company, Boston/Toronto, 1985.

CRITICAL APPRAISAL CARD
From Clinical Epidemiology & Biostatistics, McMaster University Hamilton, Ontario, CANADA

TUMOR SIZE SCALE

Watermelon
Football
Eggplant
Grapefruit
Pomegranate
Plum
Walnut
Mushroom
Grape
Pea
Rice

Sackett DL, Haynes RB, Guyatt GH, Tugwell P. (1991) *Clinical Epidemiology: a Basic Science for Clinical Medicine*, Little Brown, Chicago

Sackett, Haynes, Guyatt, and Tugwell criteria

1. Was there an independent or 'blind' comparison with a 'gold standard' or diagnosis?
2. Was the setting for the study, as well as the filter through which study patients passed, adequately described?
3. Did the patient sample include an appropriate spectrum of mild and severe, treated and untreated disease, plus individuals with different but commonly confused disorders?
4. Were the tactics for carrying out the test described in sufficient detail to permit their exact replication?

Sackett DL, Haynes RB, Guyatt GH, Tugwell P. (1991) *Clinical Epidemiology: a Basic Science for Clinical Medicine*, Little Brown, Chicago

Sackett, Haynes, Guyatt, and Tugwell criteria

5. Was the reproducibility of the test result (precision) and its interpretation (observer variation) determined?
6. Was the term 'normal' defined sensibly? (Gaussian, percentile, risk factor, culturally desirable, diagnostic, or therapeutic?)
7. If the test is advocated as part of a cluster or sequence of tests, was its contribution to the overall validity of the cluster or sequence determined?
8. Was the 'utility' of the test determined? (Were the patients really better off for it?)

Sackett DL, Haynes RB, Guyatt GH, Tugwell P. (1991) *Clinical Epidemiology: a Basic Science for Clinical Medicine*, Little Brown, Chicago

Sackett, Haynes, Guyatt, and Tugwell criteria

The best articles evaluating diagnostic tests will meet most or all of the . . . 8 criteria.

Sackett DL, Haynes RB, Guyatt GH, Tugwell P. (1991) *Clinical Epidemiology: a Basic Science for Clinical Medicine*, Little Brown, Chicago

Greenhalgh guidelines

1: Is this test potentially relevant to my practice?
2: Has the test been compared with a true gold standard?
3: Did this validation study include an appropriate spectrum of subjects?
4: Has workup bias been avoided?
(Was the reference standard group originally identified because they were positive on the test?)
5: Has expectation bias been avoided?
(I.e. was the reference standard blind to the test?)

Greenhalgh, T. (1997) How to read a paper: Papers that report diagnostic or screening tests. *BMJ* **315**, 540-543.

Greenhalgh guidelines

6: Was the test shown to be reproducible?

7: What are the features of the test as derived from this validation study?

8: Were confidence intervals given?

9: Has a sensible 'normal range' been derived? (Only relevant for continuous test variables.)

10: Has this test been placed in the context of other potential tests in the diagnostic sequence?

Greenhalgh, T. (1997) How to read a paper: Papers that report diagnostic or screening tests. *BMJ* **315**, 540-543.

QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool

1. Was the spectrum of patients representative of the patients who will receive the test in practice?
2. Were selection criteria clearly described?
3. Is the reference standard likely to correctly classify the target condition?
4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* **3**, 25, http://www.biomedcentral.com/1471-2288/3/25

QUADAS tool

5. Did the whole sample or a random selection of the sample, receive verification using a reference standard? (Rather confusing, = work-up bias)
6. Did patients receive the same reference standard regardless of the index test result?
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?
8. Was the execution of the index test described in sufficient detail to permit replication of the test?

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* **3**, 25, http://www.biomedcentral.com/1471-2288/3/25

QUADAS tool

9. Was the execution of the reference standard described in sufficient detail to permit its replication?
10. Were the index test results interpreted without knowledge of the results of the reference standard?
11. Were the reference standard results interpreted without knowledge of the results of the index test?
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* **3**, 25, http://www.biomedcentral.com/1471-2288/3/25

QUADAS tool

13. Were uninterpretable/ intermediate test results reported?
14. Were withdrawals from the study explained?

Each of these questions is 'scored' as 'Yes', 'No', or 'Unclear', but no total quality score is recommended.

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* **3**, 25, http://www.biomedcentral.com/1471-2288/3/25

The Bland criterion

Were the cut-off points for the test determined using data different from those used for evaluation?

If the same data are used to decide the cut-off and to validate it, the test will appear better than it really is.

Sensitivity, specificity, etc., will be too big.

Not in any of the checklists.

An example: cervical cancer screening

Coste J, Cochand-Priollet B, de Cremoux P, Le Galès C, Cartier I, Molinié V, Labbé S, Vacher-Lavenu, M-C, Vielh P. (2003) Cross sectional study of conventional cervical smear, monolayer cytology, and human papillomavirus DNA testing for cervical cancer screening. *British Medical Journal* **326**, 733-736.

Cross-sectional study with two samples of women:

1. a sample referred for colposcopy because of a positive smear,
2. a sample arriving for screening.

All women had all tests.

Greenhalgh: cervical cancer screening

1: Is this test potentially relevant to my practice?

Not to mine!

2: Has the test been compared with a true gold standard?

I think so.

3: Did this validation study include an appropriate spectrum of subjects?

Yes, they keep the two populations, normal risk and high risk, separate.

Greenhalgh: cervical cancer screening

4: Has workup bias been avoided?
(Was the reference standard group originally identified because they were positive on the test?)

Yes, this is specifically mentioned. However, the sample referred for colposcopy have been referred because of similar tests. We avoid it by keeping the groups separate in the analysis.

5: Has expectation bias been avoided?
(I.e. was the reference standard blind to the test?)

This is not clear. I doubt it. Would the colposcopists know which sample the women belonged to?

Greenhalgh: cervical cancer screening

6: Was the test shown to be reproducible?

Yes, kappa statistics are given for each test and conventional smears were most reliable.

7: What are the features of the test as derived from this validation study?

Sensitivity and specificity are best for conventional smears. Sensitivities are fairly high in the referred for colposcopy group and specificities high in the screening group. Sensitivity is not good in the screening group.

Greenhalgh: cervical cancer screening

8: Were confidence intervals given?

Yes.

9: Has a sensible 'normal range' been derived? (Only relevant for continuous test variables.)

This is not relevant here. I think that the cut-off for HPV was pre-specified.

10: Has this test been placed in the context of other potential tests in the diagnostic sequence?

Yes, in that we have separate results given for the tests as screening and as follow-up tests.

QUADAS: cervical cancer screening

1. Was the spectrum of patients representative of the patients who will receive the test in practice?

Yes, results given separately for colposcopy and screening samples.

2. Were selection criteria clearly described?

Not in this paper, but referred to earlier paper.

3. Is the reference standard likely to correctly classify the target condition?

The reference standard is colposcopy followed by biopsy and histology if an abnormality is detected. I think that the tests are to decide whether a colposcopist would want a biopsy, so that is we are trying to predict.

QUADAS: cervical cancer screening

4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?

Yes, at the same time.

5. Did the whole sample or a random selection of the sample, receive verification using a reference standard?

Yes.

6. Did patients receive the same reference standard regardless of the index test result?

Yes.

QUADAS: cervical cancer screening

7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?

Yes.

8. Was the execution of the index test described in sufficient detail to permit replication of the test?

I think that these methods are fairly well established.

9. Was the execution of the reference standard described in sufficient detail to permit its replication?

I think that these methods are fairly well established.

QUADAS: cervical cancer screening

10. Were the index test results interpreted without knowledge of the results of the reference standard?

Yes.

11. Were the reference standard results interpreted without knowledge of the results of the index test?

Yes.

12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?

Yes.

QUADAS: cervical cancer screening

13. Were uninterpretable/ intermediate test results reported?

I think so. We have the proportion of unsatisfactory slides.

14. Were withdrawals from the study explained?

Not mentioned. Hard to believe that all women arriving for screening agreed to colposcopy!

The Bland criterion: cervical cancer screening

Were the cut-off points for the test determined using data different from those used for evaluation?

Yes. The criteria for reading the slides were established before these data were collected.

Choice of guidelines/criteria

There is considerable overlap between the various lists of guidelines/criteria, but none includes any of the other lists completely.

Each list includes good ideas for things to watch out for in diagnostic test studies.

There is no definitive list.

Choose one that suits you.

For the assignment, use all of them.