Appraising Diagnostic Test Studies

Diagnostic test studies

These are studies which evaluate a test for diagnosing a disease. In the studies we shall be considering here, we have an established diagnostic method for the disease and we look at how well a new test predicts it. We ask questions such as 'how well does the test identify people with the disease?' and 'how well does the test exclude people without the disease?'.

The basic design is to compare test results on people with the disease with test results on people without the disease. Clearly, we need to know who has the disease.

There are two types of study design:

- ❖ Prospective or cohort design, or cross-sectional design. We take a sample of subject eligible for the test, test them all and get the true diagnosis on them all.
- ❖ Retrospective or case-control design. We take a sample with true diagnosis established as positive and another sample of controls. We may have a negative diagnosis established on controls and we may not.

We need to know who actually has the disease, the true diagnosis. We can never be absolutely sure that the 'true' diagnosis is correct, but in order to proceed we decide to accept one diagnostic method as 'true'. We call this the **gold standard** or **reference standard**. This is often more invasive than the test, e.g. biopsy and histology compared to an ultrasound image. It is always possible that the reference standard is wrong for some subjects.

Statistics of diagnostic test studies

There is no single statistic that can adequately represent the agreement between a diagnostic test and a reference standard. Many different statistics have a part to play in the analysis of such studies. These include

- 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)

To illustrate these, for our first example we shall use data from a study of diabetic eye tests (Harding *et al.*, 1995). This was a cross-sectional study in which diabetic patients being screening for eye problems were examined using direct opthalmoscopy (the test) and slit lamp stereoscopic biomicroscopy (the reference standard). A single sample of subjects all received both the diagnostic test and the reference standard test.

The following table shows the results for all eye problems combined:

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

From this table we can calculate all diagnostic test statistics other than a ROC curve:

sensitivity = 40/45 = 0.89 = 89%

specificity = 237/275 = 0.86 = 86%

LR (+ve test) = 0.89/(1 – 0.86) = 6.4

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%

We shall now look at what these mean and how they were calculated.

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

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

For eye disease in diabetics, there were 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%.

A good test will have high sensitivity and high specificity. We are looking for values exceeding 80%, preferably 90% or 95%.

Odds = number of positives divided by number of negatives.

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

For eye disease in diabetics:

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

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

As the test and the reference standard should have a strong positive relationship, we expect the odds ratio to be much greater than 1.0.

The likelihood ratio (LR) for a positive test = sensitivity/(1 – specificity).

We use this as follows. If we start with the probability that a subject has the disease, which is the prevalence of the disease, we can convert this to odds:

odds = prevalence/(1 – prevalence)

Then if we test a subject from a population with this prevalence, we can estimate the odds of having the disease if the test is positive:

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

For eye disease in diabetics:

Likelihood ratio for a positive test = 0.89/(1 – 0.86) = 6.4

Suppose the prevalence of eye problem in the local diabetic population is 10% = 0.10. The odds of eye problems is 0.10/0.90 = 0.11. If a subject has a positive test, the odds of eye disease will be increased:

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

This corresponds to a probability of eye disease = 0.41. (Probability = odds/(1 + odds)).

Similarly, the likelihood ratio for a negative test = specificity/(1 – sensitivity). As before, if we start with the probability that the subject does not have the disease = 1 – prevalence of disease and convert to odds = (1 – prevalence)/prevalence, we can look at the effect on the odds of not having the disease if the test is negative:

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

Likelihood ratio for a negative test = 0.86/(1 – 0.89) = 7.8

Suppose the prevalence of eye problem in the local diabetic population is 10% = 0.10. The odds of no eye problems is 0.90/0.10 = 9.0. If a subject has a negative test, the odds of no eye disease will be increased:

odds of disease if test negative = 9.0 × 7.8 = 70.2

This corresponds to a probability of no eye disease = 0.986.

The positive predictive value (PPV) is the proportion of test positives who are reference positive. The negative predictive value (NPV) is the proportion of test negatives who are reference negative.

For eye disease in diabetics, there were 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%.

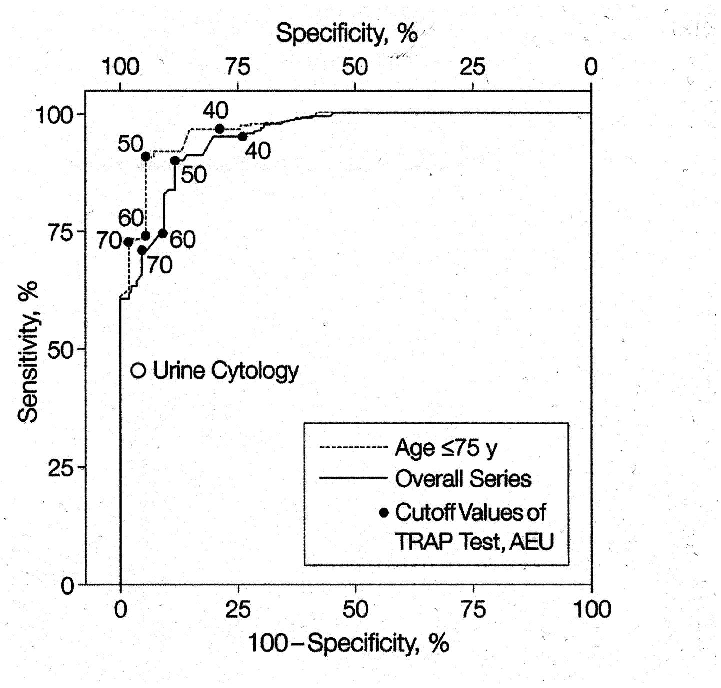
NPV = 237/242 = 98%.

Hence if a subject is positive on the test, the probability that eye disease will be found using the reference standard is 51%. If a subject negative on the test, the probability that no eye disease will be found using the reference standard is 98%.

For a receiver operating characteristic (ROC) curve, we need a different example. Sanchini *et al.* (2005) looked at the early detection of bladder cancer using a test of elevated urine telomerase, an enzyme involved in cell proliferation. The reference standard was histologically confirmed bladder cancer. This was a case-control study conducted in 218 men: 84 healthy individuals and 134 patients at first diagnosis of histologically confirmed bladder cancer.

Urine telomerase is a measurement taking a range of possible values rather the presence or absence of a sign. If we change the value of telomerase which we classify as elevated, this will change the sensitivity and specificity. We can do this and plot the sensitivity against the specificity to see how they vary together. For obscure historical reasons, it is usual to plot sensitivity against one minus specificity, also called the false positive rate. This is the ROC curve, a plot of sensitivity against 1 – specificity. (The name comes from telecommunications. As far as we are concerned, it is just a name.)

This is the ROC curve of Sanchini *et al.* (2005):



They have drawn two separate ROC curves, one for their whole sample and the other for men aged 75 years or older.

Sensitivity increases as one minus specificity increase, i.e. as specificity decreases. We make our test more sensitive at the expense of making it less specific. We are looking for a compromise cut-off which will give both high sensitivity and high specificity. Sanchini *et al.* (2005) chose 50 as being a reasonable compromise between a test which is sensitive, so finding most cases with the disease, and specific, so does not pick up a lot of people who do not have the disease.

The diagnostic tests on a ROC curve do not have to determined by a continuous measurement, though they often are. All we need to plot the curve is more than one test. Sanchini *et al.* (2005) also show a different, non-numerical test: urine cytology, not sensitive but fairly specific.

For the detection of bladder cancer using as a test that urine telomerase > 50 against a reference standard of histologically confirmed bladder cancer, the 2 by 2 table is:

	Reference standard	
Test	+ve	-ve
+ve	120	10
-ve	14	74
total	134	84

We can calculate most of the statistics as before:

sensitivity = 120/134 = 0.90 = 90%

specificity = 74/84 = 0.88 = 88%

LR (+ve test) = 0.90/(1–0.88) = 7.5

LR (-ve test) = 0.88/(1–0.90) = 8.8

OR = 120×74/(10×14) = 63.4

However, the row totals would be meaningless and they are not shown in the table. This is because we took two separate groups of subjects. The row totals will depend on what ratio of cases to controls we used. They do not tell us anything about how many people would be test positive or test negative. As a result, PPV and NPV cannot be found in this study. We cannot estimate PPV and NPV in a case-control study. Their values depend on the prevalence of the disease in the population being tested. See Bland (2004) for more information.

Here is a quotation from a newspaper article:

> "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, a positive result on the test which would not be positive for the reference standard? 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.

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. What the 'chance' or probability of this is depends on whether we are talking about the probability of a truly HIV negative subject being test positive, or of a test positive subject being truly HIV negative. The first of these is one minus the specificity, also called the false positive rate. The second is one minus the positive predictive value. Whether either the sensitivity or the PPV of a test could be 0.999999 (which is one minus one in a million) I doubt very much and we would need to test several million people repeatedly to establish such a figure.

Reproducibility in diagnostic test studies

Tests may fail because the biology is wrong or because the measurement error is too great. In most diagnostic test studies the reference standard is assumed true, therefore has no measurement error (although it must do so in practice). It is regarded as given for the study of the diagnostic test. The diagnostic test may have error, such as observer variation or variation from occasion to occasion. For example, if we use blood pressure as a test for hypertension, this varies from occasion to occasion and so we usually require successive measurements made weeks apart to be above a threshold before we diagnose hypertension.

Some diagnostic studies incorporate reproducibility studies. This might be observer variation or repeatability, or any combination.

For example, Derksen *et al.* (2005) looked at whether the Ottawa Ankle Rules (OAR) and Ottawa Foot Rules (OFR) could be applied satisfactorily by a specialized emergency nurse (SEN) rather than by a doctor. In a prospective study, all ankle sprains presented in the emergency department of a hospital from April to July 2004 were assessed by both an SEN and a junior doctor. The order in which the nurse and the doctor made their observations was randomized. In all patients, radiography was performed and acted as the reference standard.

The following table was produced:

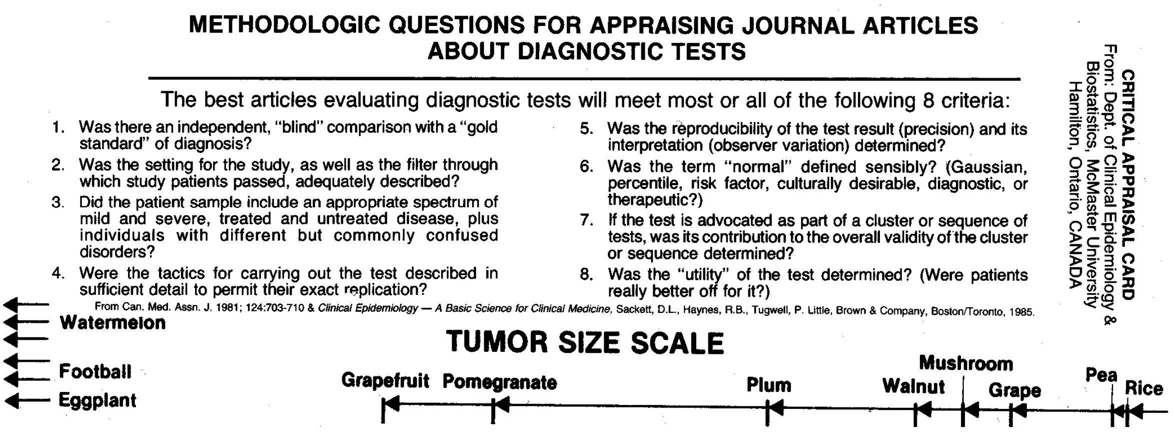
	sensitivity	specificity	PPV	NPV
nurses	0.93	0.49	0.22	0.98
doctors	0.93	0.39	0.19	0.97

This showed that the doctors and nurses produced very similar diagnostic accuracy. 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. (Kappa is an index of the agreement between different observers. Kappa = 1.0 if agreement is perfect and kappa = 0.0 if the agreement is what we would expect by chance, if observers were recording at random. kappa values above 0.8 are usually regarded as representing very good agreement, between 0.6 and 0.8 good, 0.4 to 0.6 moderate, 0.2 to 0.4 fair, and below 0.2 poor agreement. See Bland 2005.)

Critical appraisal of diagnostic test studies

Many sets of criteria or guidelines have been produced to help in the critical appraisal of diagnostic test studies. An early example was due to Sackett *et al.* (1991), in their ground-breaking book *Clinical Epidemiology: a Basic Science for Clinical Medicine*. These authors produces sets of criteria for different types of study, including diagnostic test studies. The criteria were printed on plastic cards to be carried in the pockets of clinicians, as a rapid aide memoir during practice.

This is diagnostic test study card:



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: Dept. of Clinical Epidemiology &
Biostatistics, McMaster University
Hamilton, Ontario, CANADA

TUMOR SIZE SCALE

The card has a description of the statistics of diagnostic tests on the other side.

The Sackett *et al.* (1991) criteria are

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?
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 *et al.* thought that the best articles evaluating diagnostic tests would meet most or all of the 8 criteria.

Another list of points for looking at diagnostic test studies were given by Greenhalgh (1997). The Greenhalgh guidelines are:

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 work-up bias been avoided?

5: Has expectation bias been avoided?

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?

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

Greenhalgh (1997) gives a commentary on each of these. Some of them are not very clear as expressed here. Item 4, work-up bias, means 'was the reference standard group originally identified because they were positive on the test?'. Item 5, expectation bias, means 'was the reference standard blind to the test?' If not, knowledge of the test may have influenced the judgement about the reference standard rather than the reference standard itself. Item 9, normal range, is relevant only for continuous test variables.

More recently, Whiting *et al.* (2003) developed what they call a tool, QUADAS (**Qu**ality **A**ssessment of **D**iagnostic **A**ccuracy **S**tudies). The QUADAS tool is structured as a list of 14 questions which should each be answered 'yes', 'no', or 'unclear'.

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?
5. Did the whole sample or a random selection of the sample, receive verification using a reference standard?
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?
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?
13. Were uninterpretable/intermediate test results reported?
14. Were withdrawals from the study explained?

Whiting *et al.* (2003) give a more detailed description of each item together with a guide on how to score each item. I find item 5 rather confusing. The authors say it is the same as work-up bias. They say that each item should be 'scored' as 'Yes', 'No' or 'Unclear', but having produced these scores they do not use them to produce an overall quality score for the study or paper, so their function is rather unclear. However, the list can be taken as a set of questions worth asking of a diagnostic test study.

None of these lists include what I might call "The Bland criterion", though it is certainly not original to me:

> 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 criteria for a test to be positive as are then used to estimate sensitivity and specificity, the criteria will be tailored to fit the data. Sensitivity, specificity, etc., will be too big and the test will appear better than it really is. We sometimes get this when a continuous measurement is used as a test. The cut-off for test positive is chosen so as to give the best available combination of sensitivity and specificity. The sensitivity that we get from them fits the data and to get an unbiased estimates we need a separate sample. These two samples are sometimes called the training sample and the evaluation sample. This criterion is not in any of the checklists.

Application of appraisal guidelines to a study

We shall look at the Greenhalgh guidelines and the QUADAS tool for a diagnostic test study example. The example is a cross sectional study of conventional cervical smear, monolayer cytology, and human papillomavirus DNA testing for cervical cancer screening (Coste *et al.* 2003). This was a cross-sectional study in that all women had both the three tests and a colposcopy with biopsy and histology if this was indicated during the colposcopy. There were two samples of women:

1. a sample referred for colposcopy because of a positive smear (high risk),
2. a sample arriving for screening (normal risk).

Application of any guidelines to appraisal requires both knowledge of the subject area and judgement. There are no right answers, certainly not mine. What follows are my own comments on this study in the light of the Greenhalgh and QUADAS list.

If we apply the Greenhalgh guidelines to Coste *et al.* (2003):

1: Is this test potentially relevant to my practice?

Not to mine! This is a question which is clearly relevant in practice, but not for the purposes of this exercise.

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

I think so. What we want the test to tell us is would a colposcopy lead to a biopsy with positive histology, rather than would it detect cancer.

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

Yes, they keep the two populations, normal risk and high risk, separate. Only the normal risk group is appropriate. Most of the high risk group have been detected by a smear test.

4: Has workup bias been avoided? (I.e. 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. Only the normal risk group avoids this bias.

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? We are not told, but if they did this would clearly affect their expectations of what they might find.

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. This is something which we would clearly want to know: what does the study tell us about the test.

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 rather than being derived from the study data.

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.

The study seems to hold up quite well by these guidelines, at least for the normal risk sample.

For the QUADAS tool:

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

Yes, results are given separately for colposcopy and screening samples. The normal risk sample is normal spectrum for screening subjects, the high risk sample is the normal spectrum for smears repeated because of positive findings.

2. Were selection criteria clearly described?

Not in this paper, but we are referred to an 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, and if so whether that biopsy would be positive, so that is we are trying to predict.

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, done at the same time.

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

Yes. But if this really means work-up bias, then the high risk sample have work-up bias.

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

Yes. Everybody had a colposcopy.

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

Yes. The test is not used to decide whether a biopsy is required or to do the histology.

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. They are not described in this paper and no reference is given.

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. As for 8, they are not described in this paper, but two references are given.

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

13. Were uninterpretable/intermediate test results reported?

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

14. Were withdrawals from the study explained?

This is not mentioned. I find it hard to believe that all women arriving for screening agreed to colposcopy!

The study performs quite well on QUADAS. This does not mean that it is above reproach, however, and there is a lot of correspondence on the BMJ website as Rapid Responses and published letters, including a critique from one of the authors who did not agree with some of the interpretation.

Finally, 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 and criteria. There are other available, too. None of those given here 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 which I can recommend. I suggest that you choose one that suits you for regular use if you are appraising many diagnostic test studies.

For the assignment, I would use all of them.

J. M. Bland
October 2010.

References

Bland JM. (2004) Interpretation of diagnostic tests. http://www-users.york.ac.uk/~mb55/msc/meas/senspec.pdf

Bland JM. (2005) Cohen's Kappa. http://www-users.york.ac.uk/~mb55/msc/meas/kappash2.pdf

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.

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.

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

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

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

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.

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