How well do health professionals interpret diagnostic information?  
A systematic review

Penny F Whiting,1,2 Clare Davenport,3 Catherine Jameson,1 Margaret Burke,1 Jonathan A C Sterne,1 Chris Hyde,4 Yoav Ben-Shlomo1

ABSTRACT
Objective: To evaluate whether clinicians differ in how they evaluate and interpret diagnostic test information.
Design: Systematic review.
Data sources: MEDLINE, EMBASE and PsycINFO from inception to September 2013; bibliographies of retrieved studies, experts and citation search of key included studies.
Eligibility criteria for selecting studies: Primary studies that provided information on the accuracy of any diagnostic test (eg, sensitivity, specificity, likelihood ratios) to health professionals and that reported outcomes relating to their understanding of information on or implications of test accuracy.
Results: We included 24 studies. 6 assessed ability to define accuracy metrics: health professionals were less likely to identify the correct definition of likelihood ratios than of sensitivity and specificity. 25 studies assessed Bayesian reasoning. Most assessed the influence of a positive test result on the probability of disease: they generally found health professionals’ estimation of post-test probability to be poor, with a tendency to overestimation. 3 studies found that approaches based on likelihood ratios resulted in more accurate estimates of post-test probability than approaches based on estimates of sensitivity and specificity alone, while 3 found less accurate estimates. 5 studies found that presenting natural frequencies rather than probabilities improved post-test probability estimation and speed of calculations.
Conclusions: Commonly used measures of test accuracy are poorly understood by health professionals. Reporting test accuracy using natural frequencies and visual aids may facilitate improved understanding and better estimation of the post-test probability of disease.

INTRODUCTION
Making a correct diagnosis is a prerequisite for appropriate management.1 Probabilistic reasoning is suggested to be a prominent feature of diagnostic decision-making,2–3 but to the extent to which this is based on quantitative revision of health professionals’ estimated pretest probabilities, rather than intuitive judgements, is not known.

Test accuracy can be summarised using a range of measures derived from a 2×2 contingency table (table 1). Measures that distinguish between the implications of a positive test result (positive predictive value (PPV), positive likelihood ratio (LR), specificity) and a negative test result (negative predictive value, negative LR, sensitivity) are more useful for decision-making than global test accuracy measures such as diagnostic ORs and the area under the curve (AUC).4–6 Predictive values and LRs, which are applied based on the test result, are believed to be more clinically intuitive than sensitivity and specificity, which are applied based on disease status.7–8 The promotion of evidence-based testing, including the use of LRs,5–10 is based on the premise that formal probabilistic reasoning is necessary for informed diagnostic decision-making.11 12 Such reasoning requires use of Bayes’ theorem to revise the pretest odds of disease, based on the test result, to give the post-test odds of disease.13

There is a widespread belief that health professionals and decision-makers have difficulty understanding and applying test
Difficulties are thought to arise from the need to interpret conditional probabilities, and the complex nature of probability revision. However, to date there has been no systematic review of the literature pertaining to clinician’s understanding of test accuracy evidence. Here, we aimed to evaluate whether clinicians differ in how they evaluate and interpret different diagnostic test information. The findings will be used to provide recommendations about how the results of test accuracy research should be presented in order to promote evidence-based testing.

**METHODS**

We followed standard systematic review methods and established a protocol for the review (available from the authors on request).

**Data sources**

We searched MEDLINE, EMBASE and PsycINFO from inception to September 2013. We combined terms for measures of accuracy AND terms for communicating and interpreting AND terms for health professionals (see web appendix 1). Additional studies were identified by screening the bibliographies of retrieved studies, contacting experts and through a citation search of four key included studies that is, identifying studies that had cited these papers. Contacting experts involved presenting results at a national conference and obtaining literature passively through discussions with experts at national and international conferences and meetings concerned with test evaluation. No language or publication restrictions were applied.

**Inclusion criteria**

Primary studies of any design that provided information on the accuracy of any diagnostic test (eg, sensitivity, specificity, LRs, predictive values, and receiver operator characteristic (ROC) plots/curves) to health professionals (eg, doctors, nurses, physiotherapists, midwives), or student health professionals, from any specialty and that reported outcomes relating to their understanding of test accuracy were eligible for inclusion. Studies were screened for relevance independently by two reviewers;
disagreements were resolved through consensus. Full-text articles of studies considered potentially relevant were assessed for inclusion by one reviewer and checked by a second.

Data extraction
Data extraction was carried out by one reviewer and checked by a second using a standardised form. Study quality was not formally assessed due to a lack of any agreed tools for studies of this type.

Synthesis
We combined results using a narrative synthesis due to heterogeneity between studies in terms of design, type of health professionals and measures of accuracy investigated, making a quantitative summary (meta-analysis) inappropriate. We grouped studies according to their objective: (1) accuracy definition (ability to define measures of accuracy); (2) self-reported understanding (doctors self-rating of their understanding or use of accuracy measures); (3) assess Bayesian reasoning (combining data on the pretest probability of disease with accuracy measures to obtain information on the post-test probability of disease) and (4) presentation format (impact of presenting accuracy data as frequencies rather than probabilities). Groupings were defined based on the data.

RESULTS
The searches identified 4808 records of which 24 studies reported in 28 publications were included in the review (figure 1). Table 2 presents a summary of the included studies, grouped according to objective; further details are provided in web appendix 2. The majority of studies investigated health professionals understanding of sensitivity and specificity (or false-positive rate), six studies assessed LRs and two studies assessed other measures such as graphical displays. Only one study assessed a global measure of accuracy, the ROC curve, this was a study of doctors’ self-reported understanding. Box 1 provides examples of some of the types of scenario used in the included studies.

Self-reported understanding: How do doctors self-rate their understanding or use of accuracy measures?
Two studies assessed doctors self-report of their understanding or use of diagnostic information. One study, which also contributed information on doctors’ ability to define measures of accuracy, found that 13/50 general practitioners (GPs) self-reported understanding of the definitions of sensitivity, specificity and PPV. However, when interviewed only one could define any measures of accuracy, suggesting that GPs self-rating of understanding overestimates their ability. A second study found that although 82% of doctors interviewed reported using sensitivity and specificity only 58% actually used information on sensitivity and specificity when interpreting test results and <1% reported being familiar with and using ROC curves or LRs.

Accuracy definition: “Can health professionals define measures of accuracy?”
Six single-group studies assessed health professionals’ understanding of the definition of measures of accuracy. Four studies asked doctors to identify correct definitions of sensitivity and specificity, three using multiple choice questionnaires and one based on information provided in a research study. The proportion of doctors who correctly identified sensitivity...
ranged from 76% to 88%, the proportion who correctly identified specificity ranged from 80% to 88%.20 23 24 30

LRs and predictive values were generally less well understood. One study comparing sensitivity, specificity and LRs found only 17% of healthcare professionals could define LR+ compared with 76% sensitivity and 80% specificity.30 One study found that PPV was less well understood compared with sensitivity (sensitivity 76%, PPV 61%).20 As a study that interviewed GPs to elicit their definitions of various accuracy parameters found that only 1/13 could define PPV, 1/13 could define some aspects of sensitivity and 0/13 could define specificity.45 One study compared health professionals’ ability to define sensitivity, specificity, predictive values and LRs. Health professionals were less able to define predictive values and LRs compared with sensitivity and specificity.21 A final study, that involved asking participants to identify definitions based on a 2×2 table, reported that practicing physicians were less able to select correct definitions of sensitivity and specificity compared with medical students and research doctors but exact values were not reported.24

Bayesian reasoning: “How well can health professionals combine data on pre-test probability and test accuracy to obtain information on the post-test probability of disease?”

Twenty-two studies assessed whether health professionals could combine information on prevalence with data on sensitivity and specificity (or false-positive rate) to calculate the post-test probability of disease.17 19 20 22–32 36–42 44 Nine studies used the terms ‘sensitivity’, ‘specificity’, or

---

**Table 2** Summary of included studies

<table>
<thead>
<tr>
<th>Study design</th>
<th>Number of studies</th>
<th>Total</th>
<th>Self-rating of understanding</th>
<th>Accuracy definition</th>
<th>Bayesian reasoning</th>
<th>Presentation format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single group</td>
<td>17</td>
<td>2</td>
<td>6</td>
<td>20</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>RCT</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Multiple groups, unclear allocation</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical students</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Mixed physicians</td>
<td>17</td>
<td>2</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Single specialty</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>How was the diagnostic information presented?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vignette/case study</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Population scenario</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Simulated patient</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2×2 table</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Research study extract</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>No information/unclear</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>How was understanding assessed?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questionnaire (multiple choice)</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Questionnaire (open ended)</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Unclear</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Type of scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fictitious</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Real life</td>
<td>16</td>
<td>0</td>
<td>2</td>
<td>15</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Unclear</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Measure of test accuracy assessed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>22</td>
<td>2</td>
<td>6</td>
<td>20</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Specificity/FPR</td>
<td>24</td>
<td>2</td>
<td>5</td>
<td>22</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>LR+</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>LR−</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>LR categories</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Graphical display</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>PPV</td>
<td>21</td>
<td>1</td>
<td>3</td>
<td>19</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>NPV</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ROC</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

FPR, false positive rate; LR−, negative likelihood ratio; LR+, positive likelihood ratio; NPV, negative predictive value; PPV, positive predictive value; RCT, randomised controlled trial; ROC, receiver operating characteristic.
### Box 1  Example of population based scenarios and clinical vignettes

**Self-rating of understanding:**

1. Some authorities recommend that diagnostic decisions be made first by obtaining a test’s sensitivity and specificity, estimating the prevalence of disease (in the patient under evaluation), then calculating a positive or negative predictive value. Do you perform these calculations when you make diagnostic decisions? If no, can you tell me why you do not do them?

2. Many authorities recommend that we use receiver operator characteristic (ROC) curves to set test thresholds before making diagnostic decisions. Do you use ROC curves? If no, why not?

3. Another recommendation is to use test likelihood ratios for certain diagnostic calculations. Do you use likelihood ratios before ordering tests or when interpreting test results? If no, why not?

4. Do you use test sensitivity and specificity values when you order tests or interpret test results? (For positive responses) Can you tell me in what way you use them?

5. When you use sensitivity and specificity, where do you get your values from?

6. Do you prefer to use published values for sensitivity and specificity, or values based on your clinical experience with the test?

7. Do you use positive and negative predictive accuracies when you interpret test results?

8. Do you use any other methods to help you determine the effectiveness, or accuracy of the tests you use in practice?

9. During your medical training either in medical school, residency, or perhaps fellowship training, did you participate in any formal educational activities to teach you how to use test sensitivity, specificity, or likelihood ratios?

10. Since finishing your medical training have you participated in any formal educational activities such as seminars, workshops, or CME courses designed to teach you how to use test sensitivity and specificity or likelihood ratios?

---

**Accuracy definition:**

The sensitivity of a test is: **Please check the correct answer**

- the percentage of false positive test results
- the percentage of false negative test results
- the percentage of persons with disease having a positive test result
- the percentage of persons without the disease having a negative test result

**Population based scenario: Bayesian reasoning and presentation format**

**Probability format**

The probability that one of these women has breast cancer is 1%. If a woman has breast cancer, the probability is 80% that she will have a positive mammography test. If a woman does not have breast cancer, the probability is 10% that she will still have a positive mammography test.

**Frequency format**

Ten out of every 1,000 women have breast cancer. Of these 10 women with breast cancer, 8 will have a positive mammography test. Out of the remaining 990 women without breast cancer, 99 will still have a positive mammography test.

**Bayesian reasoning: vignette/case study**

Typical angina chest pain: A 55 year old man presented to your office with a 4 week history of sub-ternal pressure-like chest pain. The chest pain is induced by exertion, such as climbing stairs, and relieved by 3–5 minutes of rest. It sometimes radiated to the throat, left shoulder, down the arm.

1. Do you understand about the idea of sensitivity, specificity, pre-test probability, post-test probability (Yes/No)
2. What is the sensitivity of the exercise stress test?
3. What is the specificity of the exercise stress test?
4. What is the probability that this patient has significant coronary artery disease?
5. What is the probability that this patient has significant coronary artery disease if the exercise stress test is positive?
6. What is the probability that this patient has significant coronary artery disease if the exercise stress test is negative?

---

Comparison of effects of positive and negative test results on Bayesian reasoning

Fourteen studies provided test accuracy information to help with interpretation of a positive test result, one study provided information for a negative test result, and five provided information for both a positive and a negative test result. In one study it was unclear whether the test result provided should be interpreted as positive or negative, and in one study participants were questioned on how they interpreted test results in general. Most participants overestimated the post-test probability of disease given a positive test result; where reported (4 studies) overestimates ranged between 46% and 73%. Two studies found that post-test probabilities were poorly estimated.
estimated for positive and negative test results. One study found that correct reasoning was applied for positive test results but that post-test probability was poorly estimated for negative test results. One study found that although the post-test probability was consistently overestimated for a positive test result, estimates were correct for negative test results. The study that assessed interpretation of a negative test result only found that 56% of participants estimated post-test probability of disease as higher than pretest probability (ie, estimate moved in the wrong direction).

Comparison of summary metrics for Bayesian reasoning
Six studies assessed the effects of providing test accuracy information using LRs (LRs), only two of these studies provided information on the positive LR (LR+) and the negative LR (LR−). Three studies provided a text description rather than using the term ‘likelihood ratio,’ and in one study a categorical approach based on the LR was used (‘quite useless’, ‘weak’, ‘good’, ‘strong’, or ‘very strong’). Two studies included an additional scenario in which the LR information was provided graphically—one provided the information as a probability modifying plot, the other as a graphic featuring five circles in a row in which an increasing number of circles were coloured black to correspond with increasing positive LRs or decreasing negative LRs.

Two studies demonstrated less correct responses for post-test probability estimation with LRs (described in words in one and numerical in the other) compared with sensitivity and specificity presented numerically. One study demonstrated similarly poor post-test probability estimation for LRs (described in words) compared with sensitivity and specificity (presented numerically). Two studies demonstrated more correct responses for post-test probability estimation with LRs (described in words or using the categorical approach) compared with sensitivity and specificity presented numerically. Two studies found that graphical presentation of LRs improved post-test probability estimation compared with LRs described in words or sensitivity and specificity presented numerically.

The effect of clinical experience, profession and academic training on Bayesian reasoning
Two studies found no effect of experience (medical students vs qualified doctors) on Bayesian reasoning, and a further study found no influence of age. One study found that a greater proportion of newly qualified doctors were more accurate in their estimation of post-test probability (29%) compared with more experienced doctors with or without an academic affiliation (15%). Two studies demonstrated that research experience improved doctors’ ability to correctly estimate post-test probability. One study found that midwives were less likely than obstetricians to correctly estimate post-test probability of disease.

Presentation format: “Does presenting accuracy data as frequencies and using graphic aids improve understanding compared to presenting results as probabilities?”
Five studies (3 randomised controlled trials (RCTs), 1 two-group study, and 1 single-group study) found that post-test probability estimation was more accurate when accuracy data were presented as natural frequencies than as probabilities (see box 1 for example). Natural frequencies are joint frequencies of two events, for example the number of women who test positive and who have breast cancer. The same information presented as a probability would just present the probability that a woman with breast cancer has a positive test result (sensitivity), usually expressed as a percentage.

Two studies also found that health professionals spent an average of 25% more time assessing the scenarios based on a probability format compared with a natural frequency format. One RCT demonstrated that presenting test accuracy information as natural frequencies with graphical aids resulted in the highest proportion of correct post-test probability estimates (73%) compared with probabilities with graphical aids (68%), natural frequencies alone (48%) or probabilities alone (23%).

DISCUSSION
Statement of principal findings
This review suggests that summary test accuracy measures, including sensitivity and specificity are not well understood. Although health professionals are able to select the correct definitions of sensitivity and specificity and to a lesser extent predictive values when presented with a series of options, they are less able to verbalise the definitions themselves. LRs are least well understood, although this may reflect a lack of familiarity with these measures rather than suggesting that they are less comprehensible. Few studies found evidence of successful application of Bayesian reasoning: most studies suggested that post-test probability estimation is poor with wide variability and a tendency to overestimation for both positive and negative test results. There was some evidence that post-test probability estimation is poorer for negative than positive test results, although few studies assessed the impact of negative test results. The impact of LRs on estimation of post-test probability is unclear. Presenting data as natural frequencies rather than as probabilities improved post-test probability estimation and also the speed of calculations. The use of visual aids to present information (both on probabilities and natural frequencies) was found to further improve post-test probability estimation, although this was based on a single study. No study investigated understanding of other test accuracy metrics such as ROC curves, AUC and forest plots.

Explanation of findings
Difficulty in interpreting summary test accuracy measures is likely to be related to their complexity. Summary test accuracy statistics used to describe test
performance (e.g., sensitivity and specificity and positive and negative predictive values) are conditional probabilities and misinterpretation as evidenced in this review is proposed to be a function of confusion over the subgroup of study participants the measures refer to. For example, the subgroup may be those with or without disease (sensitivity and specificity), or those with positive or with negative test results (positive and negative predictive values).

Our finding that presenting probabilities as frequencies may facilitate probability revision by healthcare professionals mirrors the findings of research carried out in the psychological literature. Research in the psychological literature has also shown that individuals are often conservative when asked to estimate probability revisions based on Bayes’ theorem. However, this has been shown only to be the case for information having reasonably high diagnostic value. For information with the least diagnostic value, participants are generally more extreme than would be expected based on Bayes’ theorem. This is consistent with our findings where most examples presented combinations of low pretest probabilities of disease or values of sensitivity and specificity that were not sufficiently high for ruling in or ruling out disease. The findings of this review are important for those attempting to facilitate the integration of test accuracy evidence into diagnostic decision-making. Indeed qualitative research conducted recently suggests that interpretation of findings of systematic reviews of test accuracy by decision-makers is poor.

Strengths and weaknesses

To the best of our knowledge, this is the first systematic review of health professionals’ understanding of diagnostic information. We conducted extensive literature searches in an attempt to maximise retrieval of relevant studies. However, a potential limitation of our review is that the search was conducted in September 2013 and so any recently published articles will not have been captured. The possibility of publication bias remains a potential problem for all systematic reviews. Publication bias was not formally assessed in this review because there is no reliable method of assessing publication bias when studies report a variety of outcomes in different formats. However, the potential impact of publication bias is likely to be less for these types of studies where there is no clear ‘positive’ finding than for RCTs of treatment effects which may be more likely to be published if a positive association between the treatment and outcomes is demonstrated. Study quality assessment is an important component of a systematic review. For this review we did not perform a formal risk of bias assessment as study designs included in the review varied and, although we included some RCTs, most were single-group studies that examined how well doctors could perform certain calculations or understand pieces of diagnostic information. There is no accepted tool for assessing the risk of bias in these types of study and so we were unable to provide a formal assessment of risk of bias in these studies.

Conclusions and implications for practice, policy and future research

Perhaps the more important finding of this review is the lack of understanding of test accuracy measures by health professionals. This review suggests that presenting probabilities as frequencies may improve understanding of test accuracy information and this has been embraced by both the Cochrane Collaboration and GRADE. Further research is needed to capture the needs of healthcare professionals, policymakers and guideline developers with respect to presentation of test accuracy evidence for diagnostic decision-making and how this may actually influence disease management especially as regards initiating or withholding treatment.

Author affiliations

1School of Social and Community Medicine, University of Bristol, Bristol, UK
2The National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West at University Hospitals Bristol NHS Foundation Trust
3Unit of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, UK
4Peninsula Technology Assessment Group, Peninsula College of Medicine & Dentistry, Exeter, UK

Contributors PFW and CD contributed to the conception and design of the study, analysis and interpretation of data, and drafting of the manuscript. JACS, CH and YB-S contributed to the conception and design of the review. CJ acted as second reviewer performing inclusion assessment and data extraction. MB conducted the literature searches. All authors commented on drafts of the manuscript and gave final approval of the version to be published. PFW is the guarantor.

Funding This work was partially funded by the UK Medical Research Council (Grant Code G0801405).

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for any purpose, subject to the terms of the licence. This is an Open Access article distributed in accordance with the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for any purpose, subject to the terms of the licence. This work is c\textcopyright\ 2015 the authors, some rights reserved; exclusive licence held by BMJ Publishing Group Limited. You may use, download and print the article for personal use and non-commercial research and teaching. You may not alter the article in any way.

REFERENCES

**Web Appendix 1: MEDLINE Search Strategy (1950 to present)**

1. `exp "Sensitivity and Specificity"/` (325577)
2. `likelihood functions/` (13059)
3. `diagnostic accuracy.tw.` (16207)
4. `(sensitivity and specificity).tw.` (99257)
5. `likelihood ratio$.tw.` (5874)
6. `predictive value$.tw.` (50099)
7. `receiver operating curve$.tw.` (720)
8. `roc.tw.` (12868)
9. `or/1-8` (406204)
10. `statistics as topic/` (74395)
11. `exp Diagnosis/` (5292818)
12. `di.f.s.` (1692215)
13. `diagnos$.af.` (2740205)
14. `or/11-13` (6392103)
15. `10 and 14` (26625)
16. `9 or 15` (430307)
17. `exp Decision Making/` (95388)
18. `Clinical Competence/` (52085)
19. `(communicat$ adj5 statistic$).tw.` (182)
20. `(communicat$ adj5 risk$).tw.` (2209)
21. `(skill$ adj5 evidence).tw.` (632)
22. `(probabilistic$ adj5 reason$).tw.` (127)
23. `(understand$ adj5 statistic$).tw.` (450)
24 (understand$ adj5 risk$).tw. (4439)
25 (interpret$ adj5 statistic$).tw. (1478)
26 (interpret$ adj5 test$).tw. (6619)
27 (diagnos$ adj5 probabilit$).tw. (1435)
28 (clinical adj5 reason$).tw. (4191)
29 bayes theorem/ (12418)
30 bayesian.tw. (12435)
31 or/17-30 (180963)
32 16 and 31 (9839)
33 exp Health Personnel/ (319758)
34 doctor$.tw. (70235)
35 physician$.tw. (218341)
36 nurse$.tw. (159624)
37 practitioner$.tw. (76174)
38 clinician$.tw. (88378)
39 Family Practice/ (57311)
40 Physician's Practice Patterns/ (31957)
41 Nurse's Practice Patterns/ (214)
42 or/33-41 (797624)
43 32 and 42 (1772)
### Web Appendix 2: Individual Study details

#### a. Self-rating of understanding

<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reid (1998) 41</td>
<td>Single group</td>
<td>300 practicing doctors</td>
<td>Sensitivity Specificity LR+ LR- ROC curves</td>
<td>None</td>
<td>Questioned regarding use and understanding of various measures</td>
<td>Telephone interview</td>
<td>None</td>
<td>8 (3%) used the recommended formal Bayesian calculations, 3 used ROC curves, and 2 used likelihood ratios. The main reasons cited for non-use included impracticality of the Bayesian method (74%), and non-familiarity with ROC curves and likelihood ratios (97%). 246 (82%) used sensitivity and specificity but only 174 (58%) physicians used them when interpreting test results.</td>
</tr>
<tr>
<td>USA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young (2002) 45</td>
<td>Single group</td>
<td>50 GPs</td>
<td>Sensitivity Specificity PPV</td>
<td>No information</td>
<td>Asked to self-rate understanding of diagnostic terms.</td>
<td>Telephone interview</td>
<td>None</td>
<td>13 of 50 indicated that “I understand this and could explain to others’ the above answer” for the 3 diagnostic terms. Participants self ratings of their understanding differed from an objective, criterion based assessment.</td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### b. Accuracy Definition

<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argimon-Pallas (2011)</td>
<td>Single group</td>
<td>152 family medicine residents in their second year of the Family Medicine training programme</td>
<td>Sensitivity, Specificity, PPV, NPV, LR+</td>
<td>Population based scenario</td>
<td>Information provided on total number of patients with target condition and number with and without condition testing positive</td>
<td>Questionnaire asked to calculate accuracy measures from raw data in scenario</td>
<td>Unclear</td>
<td>Before task number of doctors correctly calculating figures were: Sensitivity: 42% Specificity: 34% PPV: 33% NPV: 26% LR+: 8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Administered before and after educational intervention (intensive and interactive four half-day sessions)</td>
<td></td>
<td></td>
<td>After intervention numbers more than doubled for all accuracy measures. Sensitivity: 82% Specificity: 79% PPV: 82% NPV: 80% LR+: 48%</td>
</tr>
<tr>
<td>Bergus (2004)</td>
<td>Single group</td>
<td>43 medical students and residents (psychiatry and Internal Medicine)</td>
<td>Sensitivity, Specificity</td>
<td>Extract from research study</td>
<td>Asked to identify sensitivity and specificity from report</td>
<td>Questionnaire (open ended)</td>
<td>Real life (major depression and panic disorder, congestive heart failure)</td>
<td>88% correctly identified the specificity and sensitivity of the test from the paper.</td>
</tr>
<tr>
<td>Berwick (1981)</td>
<td>Single group</td>
<td>36 medical students, 45 interns and residents, 49 research doctors, 151</td>
<td>Sensitivity, Specificity</td>
<td>2x2 table</td>
<td>Asked to identify definitions based on 2x2 table (a, b, c, d used rather than numbers)</td>
<td>Questionnaire (MC)</td>
<td>Hypothetical (Disease K)</td>
<td>Practicing physicians were less able to correctly define sensitivity and specificity than medical students and research doctors. Exact values not reported</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Country</td>
<td>Sample Size</td>
<td>Study Methods</td>
<td>Study Setting</td>
<td>Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>---------</td>
<td>-------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estellat (2006)</td>
<td>Single group</td>
<td>Senior doctors research and full time practice</td>
<td>Sensitivity Specificity LR+</td>
<td>2x2 table and short extract from study report.</td>
<td>Questionnaire. (multiple choice, Postal or given directly by one investigator)</td>
<td>Real life (CT for Pulmonary Embolism)</td>
<td>85% selected correct definition for sensitivity, 80% for specificity and 17% for LR+. High rate of ‘do not know’ for LR’s (72%)</td>
<td></td>
</tr>
<tr>
<td>Steurer (2002)</td>
<td>Single group</td>
<td>263 GPs</td>
<td>Sensitivity PPV</td>
<td>No information</td>
<td>Questionnaire (multiple choice)</td>
<td>Real life (Transvaginal ultrasound for endometrial cancer)</td>
<td>76% (95% CI 70-81%) correctly identified the definition of sensitivity, 61% (95% CI 45-67%) correctly identified the definition of PPV</td>
<td></td>
</tr>
<tr>
<td>Related publication: Bachmann (2003)</td>
<td>Single group</td>
<td>Switzerland</td>
<td>Sensitivity PPV</td>
<td>No information</td>
<td>Interview</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Young (2002)</td>
<td>Single group</td>
<td>13 GPs</td>
<td>Sensitivity Specificity PPV</td>
<td>No information</td>
<td>Interview</td>
<td>None</td>
<td>Sensitivity: In interview, 1 met some of the criteria to show that they knew the correct meaning of the term, 7 met none of the criteria and 5 could not or refused to answer or participate. Specificity: In interview, 6 met none of the criteria and 7 could not answer or refused to participate. PPV: In interview, 1 met all the criteria, 1 met none of the criteria and 11 could not answer or refused to participate.</td>
<td></td>
</tr>
</tbody>
</table>
## c. Bayesian Reasoning

<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agoritsas(2011) 22</td>
<td>RCT</td>
<td>1361 physicians of all clinical specialties</td>
<td>Sensitivity, Specificity</td>
<td>Population-based scenario</td>
<td>Sensitivity and specificity described in words and numerical frequencies (terms not used) for very accurate test (sensitivity and specificity 99%)</td>
<td>Multiple choice Questionnaire: Different categories of post-test probability offered: &lt;60%, 60-79%, 80-94%, 95-99.9%, &gt;99.9%</td>
<td>Screening test for viral disease in primary school</td>
<td>Test result evaluated (positive or negative): Positive Post-test probability proportion correct: 22% Most respondents (66.7% to 80.3%) selected a post-test probability of 95–99.9%, regardless of the prevalence of disease and even when no information on prevalence was provided. We estimated that 9.1% (95% CI 6.0–14.0) of respondents knew how to assess correctly the post-test probability. This proportion did not vary with clinical experience or practice setting.</td>
</tr>
</tbody>
</table>

Doctors randomised to receive information on different prevalence (1%, 2%, 10%, 25%, 95%) and no information
<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bergus (2004)</td>
<td>Single group</td>
<td>43 medical students and incoming residents (psychiatry and Internal Medicine)</td>
<td>Sensitivity, Specificity</td>
<td>Extract from research study and simulated patient</td>
<td>Asked to identify sensitivity and specificity from report and asked to apply these to a patient with a specified pre-test probability</td>
<td>Questionnaire (open ended)</td>
<td>Real life (major depression and panic disorder, congestive heart failure)</td>
<td>Test result evaluated: Unclear PPV/NPV proportion correct: 1/28 Med students, 0/15 residents PPV proportion over/under: NR</td>
</tr>
<tr>
<td>USA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berwick (1981)</td>
<td>Single group</td>
<td>36 medical students, 45 interns and residents, 49 research doctors, 151 full time doctors</td>
<td>Sensitivity, Specificity</td>
<td>Population based scenario</td>
<td>Sensitivity and specificity described in words (terms not used)</td>
<td>Questionnaire (MC)</td>
<td>Hypothetical (Disease K)</td>
<td>Test result evaluated: Positive PPV proportion correct: 32% PPV proportion over: 68% PPV proportion under: 0 Effect of research: 65% research vs 21% practicing correct</td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>---------------------------------</td>
<td>-----------------</td>
<td>--------------------------------------------</td>
</tr>
</tbody>
</table>
| Borak (1982)  | Single group | 42 practising physicians based in a non-teaching hospital, 43 'statistically sophisticated' community medicine physicians, 43 nurses | Sensitivity, Specificity      | 2 population based and 1 simulated patient scenario | Sensitivity and specificity described in words (terms not used) to a population or a patient with a specified pre-test probability also described in words | Questionnaire (open ended) | Real life (streptococcal sore throat, bowel cancer) | Non-medical scenarios also included but not presented here | Test result evaluated: Positive  
PPV proportion correct: 34% statistically sophisticated doctors, <2% of nurses and other doctors  
PPV proportion over/under: NR |
| Bramwell (2006) | RCT          | 42 midwives, 41 obstetricians                                                 | Sensitivity, FPR              | Population based scenario                     | Sensitivity and FPR described in words; terms not used. Group 1 received information in % format, group 2 in natural frequencies | Questionnaire (open ended) | Real life (Down's screening)                | Test result evaluated: Positive  
PPV proportion correct: 0 midwives, 5% obstetricians  
PPV proportion over: 46% midwives, 76% obstetricians  
PPV proportion under: 55% midwives, 19% obstetricians |
<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casscells (1978) 17</td>
<td>Single group</td>
<td>40 doctors</td>
<td>FPR</td>
<td>Population based scenario</td>
<td>Single scenario</td>
<td>Interview (1 on 1 corridor discussion)</td>
<td>Hypothetical</td>
<td>Test result evaluated: Positive</td>
</tr>
<tr>
<td>USA</td>
<td></td>
<td>20 medical students</td>
<td></td>
<td>including prevalence and FPR</td>
<td></td>
<td></td>
<td></td>
<td>PPV proportion correct: 11/60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV proportion over: not stated; 27/60 said 95% and mean was 56% - correct value was 2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPV proportion under: NR</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Effect of experience: No effect</td>
</tr>
<tr>
<td>Chernushkin (2012) 27</td>
<td>Single group</td>
<td>94 Pharmacists; 55 completed diagnostics knowledge and skills section (extracted here)</td>
<td>Sensitivity Specificity LR+ (numerical)</td>
<td>Population based scenario</td>
<td>Various different knowledge and skills questions related to application of accuracy measures</td>
<td>Online questionnaire</td>
<td>Real life</td>
<td>Test result evaluated (positive or negative): Positive and negative Post-test probability proportion correct: When information on sensitivity was provided 61% were correct, when information on specificity was provided 48% were correct, when information on LR+ was provided 39% were correct. The mean proportion of “don’t know” answers was 13% for sensitivity, 9% for specificity and 49% for LR+.</td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>---------------------------------</td>
<td>------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Curley 1990   | Unclear allocation to 1/8 scenarios | 36 fellowship physicians, 29 chief medical residents, 18 medical students. 208 undergraduates (non-medical) also included but results not presented here | Sensitivity Specificity | Vignette/Case-study | In 6/8 scenarios sensitivity, specificity and prevalence in words (terms not provided). In 2/8 scenarios specificity was purposefully not provided | Questionnaire (open ended) | Real life (Coronary heart disease) | **Test result evaluated:** Positive  
**PPV proportion correct:** Most participants revised probability in correct direction but reasonable proportion did not. Between 0% and 69% of participants correctly estimated the magnitude and direction of change in post-test probability following a positive test result (PPV) (on a visual scale from 0-100%).  
**Values of sens/spec:** Values of sens/spec did not influence proportion correct  
**Effect of experience:** No significant difference in correct responses between medical students, physicians and undergraduates. |
| Eddy (1982)   | Single group | 100 doctors | FPR | Population based scenario including prevalence and FPR | Unclear | Real life (mammography breast cancer) | **Test result evaluated:** Positive  
**PPV proportion correct:** 95/100 estimated answer as 75% rather than 7.5% |
<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estellat (2006)³⁰</td>
<td>Single group</td>
<td>130 Senior doctors research and full time practice</td>
<td>Sensitivity, Specificity LR+</td>
<td>Population scenario (different scenarios for sens/spec and LR+)</td>
<td>Sensitivity, specificity, LR+ (in words) and prevalence given</td>
<td>Questionnaire. (multiple choice for sens/spec and open for LR+)</td>
<td>Hypothetical</td>
<td>Test result evaluated: Positive PPV proportion correct: 32% correct, 42% incorrect, 25% do not know based on sens and spec. PPV proportion over/under: NR LR Effect: 9% correct PPV with LR+, 58% incorrect, 25% did not know</td>
</tr>
<tr>
<td>Garcia-Retamero (2013)³¹</td>
<td>RCT</td>
<td>81 GPs with a minimum of 1 year of practice and 81 patients; data only extracted for GPs</td>
<td>Sensitivity FPR</td>
<td>Population based scenario</td>
<td>Information on sensitivity FPR and prevalence reported in words (terms not used) or as natural frequencies. Half participants received this information depicted with visual aids</td>
<td>Paper questionnaire</td>
<td>Real life (Breast cancer, colon cancer, diabetes)</td>
<td>Test result evaluated (positive or negative): Positive Post-test probability proportion correct: Probabilities alone: 23% Natural frequencies alone: 48% Probabilities with visual aid: 68% Natural frequencies with visual aid: 73%</td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------</td>
<td>-------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>-------------------------------</td>
<td>------------------</td>
<td>--------------------------------------------</td>
</tr>
</tbody>
</table>
| Hoffrage (1998)
**Related publications:**
Gigerenzer (1996)
Gigerenzer (2003) | Two groups   | 48 Doctors, mixture of full time and research | Sensitivity FPR | Vignette/Case study | Information on sensitivity and specificity reported in words (terms not used) or as natural frequencies | Questionnaire (multiple choice) & interview about reasoning strategies | Real life (Breast cancer, colorectal cancer, Phenylketonuria and Ankylosing Spondylitis.) | Test result evaluated: Positive
**PPV proportion correct:** 10% as probabilities, 46% as natural frequencies
**PPV proportion over:** 17/24 for prob, 8/24 for nat freq
**PPV proportion under:** 5/25 for prob, 5/24 for nat freq |
| Germany               |              |                                      |                               |                                               |                      |                               |                  |                                            |
| Hoffrage (2000)
**Related publication:**
Hoffrage (2004)        | Single group | 87 medical students, 9 first year interns | Sensitivity FPR | Population based scenario | 4 different scenarios 2 presented as probabilities (terms defined in words), and two as natural frequencies. Short and long formats used. | Questionnaire        | Real life (colorectal cancer, breast cancer, phenylketonuria, ankylosing spondylitis) | Test result evaluated: Positive
**PPV proportion correct:** Long prob 18%, long nat 57%, short prob 50%, short nat 68% |
<p>| Germany               |              |                                      |                               |                                               |                      |                               |                  |                                            |</p>
<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lyman (1993)</td>
<td>Single group</td>
<td>29 doctors; 21 nurses and pharmacists</td>
<td>Sensitivity</td>
<td>Vignette/Case study</td>
<td>Questionnaire</td>
<td>Real life (mammography for breast cancer)</td>
<td>Test result evaluated: Positive and negative PPV: Consistently overestimated NPV: Estimates correct</td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td></td>
<td></td>
<td>Specificity</td>
<td></td>
<td>(open ended)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td></td>
<td></td>
<td>Specificity</td>
<td></td>
<td>(open ended)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>--------------------------------</td>
<td>-----------------</td>
<td>---------</td>
</tr>
<tr>
<td>Moreira (2008)</td>
<td>Single group</td>
<td>50 Doctors attending course on tropical medicine</td>
<td>Sensitivity, Specificity Categorical grouping based on LR</td>
<td>Unclear</td>
<td>Sensitivity and specificity values and LRs categorised as: ‘quite useless’, ‘weak’, ‘good’, ‘strong’, ‘very strong’.</td>
<td>Questionnaire (multiple choice and open ended)</td>
<td>Mixed (4 real diseases and 2 dummy diseases)</td>
<td>Test result evaluated: Positive PPV proportion over: Overestimated for real and dummy diseases. PPV not estimate: 40% could not calculate PPV with sensitivity and specificity data LR Effect: More accurate results with categorical description of LR compared to numerical presentation of sens and spec</td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------------------------</td>
<td>-----------------------------------------------</td>
<td>---------------------------------------</td>
<td>---------------------------------</td>
<td>----------------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>Noguchi (2002) (^{39})</td>
<td>Single group</td>
<td>224 medical students</td>
<td>Sensitivity Specificity</td>
<td>Vignette/Case-study</td>
<td>Participants provided with 1/3 descriptions of a patients’ history representing low, intermediate or high pre-test probability and a diagnostic test result (+ve or – ve) and asked to estimate pre-test probability and PPV and NPV</td>
<td>Questionnaire (open ended)</td>
<td>Coronary Heart Disease and Exercise Stress Test</td>
<td><strong>Test result evaluated:</strong> Positive and negative PPV: Correct reasoning NPV: Poorly estimated</td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------</td>
<td>---------------------------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>---------------------------------</td>
<td>-----------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Puhan (2005) a)</td>
<td>RCT</td>
<td>183 Senior family and internal medicine doctors</td>
<td>Sensitivity, Specificity</td>
<td>Vignette/Case study</td>
<td>Group 1: Sensitivity and specificity</td>
<td>Questionnaire</td>
<td>Pulmonary Embolus, Myocardial Infarction, COPD, Temporal arteritis, flu, heart failure.</td>
<td>Test result evaluated: Positive and negative Post-test probability proportion correct: Deviations from correct estimates were similar for all modes of presentation, for some scenarios the graphic produced the closest estimates Post-test probability proportion over: Overall post-test probability in wrong direction in 9% of sens/spec group, 4% in LR group, and 4% in LR graphic group</td>
</tr>
<tr>
<td>Switzerland</td>
<td></td>
<td></td>
<td>LR+, LR-</td>
<td></td>
<td>Group 2: Positive or negative likelihood ratio defined in words Group 3: simple graphic of 5 circles based on LR.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reid (1998) b)</td>
<td>Single group</td>
<td>300 practicing doctors</td>
<td>Sensitivity, Specificity</td>
<td>Questioned regarding use and understanding of various measures</td>
<td>Telephone interview</td>
<td>None</td>
<td>None</td>
<td>Test result evaluated: No test result defined PPV proportion correct: Of the 174 physicians who said they used sensitivity and specificity, 165 (95%) did not do so in the recommended formal manner.</td>
</tr>
<tr>
<td>USA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>---------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>----------------------</td>
<td>-------------------------------</td>
<td>----------------</td>
<td>---------</td>
</tr>
<tr>
<td>Sox (2009)</td>
<td>RCT</td>
<td>653 paediatricians</td>
<td>Sensitivity Specificity</td>
<td>Vignette/Case study</td>
<td>Group 1: No test accuracy info&lt;br&gt;Group 2: Sensitivity and specificity (%)&lt;br&gt;Group 3: Sensitivity and specificity (natural frequencies)</td>
<td>Questionnaire (open ended postal)</td>
<td>Real life (DFA for pertussis)</td>
<td><strong>Test result evaluated:</strong> Negative&lt;br&gt;<strong>Post-test probability proportion correct:</strong> 1% (n=5) (all from group 3) estimated correct value. Proportion nearly correct was 13% (group 1), 20% (group 2) and 19% (group 3)&lt;br&gt;<strong>Post-test probability proportion over:</strong> 56% estimated post test prob higher than pre-test prob, 11% estimated post test probability same as pre-test probability. 32% estimated post-test prob as 50% (same as sensitivity)&lt;br&gt;<strong>Effect of experience:</strong> Greater proportion of residents estimated a nearly correct probability (29%) compared to paediatricians with (15%) or without (15%) an academic affiliation.</td>
</tr>
<tr>
<td>Study</td>
<td>Study design</td>
<td>Participants</td>
<td>Measures of accuracy assessed</td>
<td>How was the diagnostic information presented?</td>
<td>Information provided</td>
<td>How was understanding assessed?</td>
<td>Type of scenario</td>
<td>Results</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------</td>
<td>----------------------------------</td>
<td>------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Steurer (2002) \(^\text{10}\) | RCT          | 263 GPs      | Sensitivity Specificity LR+ (described in words) | Vignette/Case study                             | Generic question based on sensitivity and specificity for population based scenario. | Questionnaire (multiple choice and open ended) | Real life (Transvaginal ultrasound for endometrial cancer) | Test result evaluated: Positive  
PPV proportion correct: 22%.  
PPV proportion over: 56% selected value close to 100%.  
PPV overestimated: no test accuracy info > sensitivity & specificity (%) > LR in plain language. |
| Related publication:     |              |              |                               |                                               |                       |                                  |                  |                                                                          |
| Bachmann (2003) \(^\text{43}\) |              |              |                               |                                               |                       |                                  |                  |                                                                          |
| Switzerland              |              |              |                               |                                               |                       |                                  |                  |                                                                          |
| Vermeesch (2010) \(^\text{44}\) | Single group | 117 GPs and 55 specialists in internal medicine | Sensitivity Specificity LR+ Probability modifying plot | Population based scenario | Three questions with different info:  
Q 1: Sensitivity, specificity and prevalence  
Q 2: Prevalence & LR+ described in words (terms not used)  
Q 3: Prevalence and probability modifying plot | Questionnaire (multiple choice, conference) | Hypothetical | Test result evaluated: Positive  
PPV proportion correct: Q1: 7%, Q2: 27%, Q3: 50%.  
PPV “Don’t know”: Q1 15%, Q2 22%, Q3 33%  
PPV proportion over: Q1: 73%, Q2: 43%, Q2: 7%  
PPV proportion under: Q1: 6%, Q2: 8%, Q3: 2%  
Effect of experience: Results similar according to age |
<table>
<thead>
<tr>
<th>Study</th>
<th>Study design</th>
<th>Participants</th>
<th>Measures of accuracy assessed</th>
<th>How was the diagnostic information presented?</th>
<th>Information provided</th>
<th>How was understanding assessed?</th>
<th>Type of scenario</th>
<th>Results</th>
</tr>
</thead>
</table>
| Bramwell (2006)²⁶  | RCT          | 42 midwives, 41 obstetricians | Sensitivity (1-specificity) FPR | Population based scenario                      | Information on sensitivity and 1-specificity (as FPR) reported in words (terms not used) or as natural frequencies | Questionnaire (open ended) | Real life (Down’s screening) | **Probability format (sensitivity and FPR as words):**  
  - None of the midwives and 1 (5%) of the obstetricians gave the correct answer.  
  - 46% of midwives and 76% of obstetricians overestimated the PPV  
  - 55% of midwives and 19% of obstetricians underestimated the PPV.  
  **Natural frequency format:**  
  - None of the midwives and 13 (65%) of the obstetricians gave the correct answer.  
  - 35% of midwives and 15% of obstetricians overestimated the PPV  
  - 65% of midwives and 20% of obstetricians underestimated the PPV. |
<table>
<thead>
<tr>
<th>Study (??)</th>
<th>Design</th>
<th>Population</th>
<th>Presentation</th>
<th>Information</th>
<th>Delivery</th>
<th>Example</th>
<th>Results</th>
</tr>
</thead>
</table>
| Garcia-Retamero (2013) | RCT | 81 GPs with a minimum of 1 year of practice and 81 patients; data only extracted for GPs | Sensitivity FPR | Population based scenario | Information on sensitivity FPR and prevalence reported in words (terms not used) or as natural frequencies. Half participants received this information depicted with visual aids | Paper questionnaire | Real life (Breast cancer, colon cancer, diabetes) | Test result evaluated (positive or negative): Positive  
Post-test probability proportion correct:  
Probabilities alone: 23%  
Natural frequencies alone: 48%  
Probabilities with visual aid: 68%  
Natural frequencies with visual aid: 73% |
| Hoffrage (1998) | Two groups | 48 Doctors, mixture of full time and research | Sensitivity Specificity | Vignette/Case study | Information on sensitivity and specificity reported in words (terms not used) or as natural frequencies | Questionnaire (multiple choice) & interview | Real life (Breast cancer, colorectal cancer, Phenylketonuria and Ankylosing Spondylitis) | Probability format: Clinicians correct post-test probability only 10%  
Natural frequency format: Clinicians correct post-test probability increased to 46%.  
Doctors spent an average of 25% more time on probability formats than natural frequency formats |
| Hoffrage (2000) | Single group | 87 medical students, 9 first year interns | Sensitivity FPR | Population based scenario | Information on sensitivity and specificity reported in words (terms not used) or as natural frequencies. Four scenarios two for each presentation format using short and long versions | Questionnaire (open ended) | Real life (colorectal cancer, breast cancer, Phenylketonuria, ankylosing spondylitis) | LONG FORMAT:  
Probability format: Clinicians correct post-test probability only 10% correct  
Natural frequency format: Clinicians correct post-test probability increased to 57%.  
SHORT FORMAT:  
Probability format: Clinicians correct post-test probability only 50% correct  
Natural frequency format: Clinicians correct post-test probability increased to 68%. |
<table>
<thead>
<tr>
<th>Study</th>
<th>Design</th>
<th>Participants</th>
<th>Intervention</th>
<th>Outcome Measures</th>
<th>Summary</th>
</tr>
</thead>
</table>
| Sox (2009)^2 | RCT | 635 paediatricians | Sensitivity and Specificity | Vignette/Case study | Group 1: No test accuracy info  
Group 2: Sensitivity and specificity  
Group 3: Sensitivity and specificity (natural frequencies)  
Questionnaire (open ended postal) | Real life (DFA for pertussis) | 18% correctly estimated post-test probability.  
There was no difference (p=0.16) in the mean post-test probability between groups 1 and 2 (38% and 41%). Group 3 (45%) had a significantly higher mean post-test probability than group 1 (p=0.007).  
Even though test result was negative 56% of participants gave a higher post-test probability than the pre-test probability and 11% estimated a post-test probability of 30% (same as pre-test probability). Five participants (all in group 3) correctly estimated the post-test probability. There was no significant difference in the proportion of doctors who nearly estimated the correct post-test probability (defined as within range 13% to 23%) - 13% in group 1, 20% in group 2, and 19% in group 3 - p=0.06 comparing groups 1 and 2, p=0.08 and comparing groups 3 and 1 |
References (same as main document)
47. Gigerenzer G. What are natural frequencies?, 2011.
53. GRADE working group [Internet]. Secondary GRADE working group [Internet] 2014 [accessed 27.3.2014].