BMJ Open

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-Shlomo[1]

To cite: Whiting PF, Davenport C, Jameson C, *et al*. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open* 2015;**5**:e008155. doi:10.1136/bmjopen-2015-008155

▶ Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-008155).

PFW and CD are joint first authors.

Received 10 March 2015
Revised 1 July 2015
Accepted 2 July 2015

For numbered affiliations see end of article.

Correspondence to
Dr Penny Whiting;
penny.whiting@bristol.ac.uk

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.

Strengths and limitations of this study

- 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.
- We did not perform a formal risk of bias assessment as study designs included in the review varied and 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.

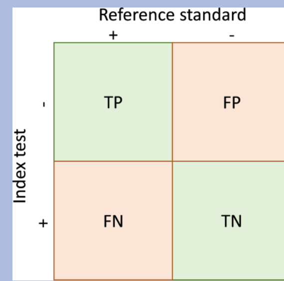
INTRODUCTION

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 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,[8–10] 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

Table 1 A 2×2 table showing the cross-classification of index test and reference standard results and overview of measures of accuracy that can be calculated from these data*



True positives	People with the target condition who have a positive test result	TP
True negatives	People without the target condition who have a negative test result	TN
False positives	People without the target condition who have a positive test result	FP
False negatives	People with the target condition who have a negative test result	FN
Sensitivity	Proportion of patients with the target condition who have a positive test result	TP/(TP+FN)
Specificity	Proportion of patients without the target condition who have a negative test result	TN/(FP+TN)
Positive predictive value (PPV)	Probability that a patient with a positive test result has the target condition	TP/(TP+FP)
Negative predictive value (NPV)	Probability that a patient with a negative test result does not have the target condition	TN/(FN+TN)
Prevalence	The proportion of patients in the whole study population who have the target condition	(TP+FN)/(TP+FP+FN+TN)
Positive likelihood ratio (LR+)	The number of times more likely a person with the target condition is to have a positive test result compared with a person without the target condition	(TP/(TP+FN))/(FP/(FP+TN)) or sensitivity/(1−specificity)
Negative likelihood ratio (LR−)	The number of times more likely a person with the target condition is to have a negative test result compared with a person without the target condition	(FN/(TP+FN))/(TN/(FP+TN)) or (1−sensitivity)/specificity

*Adapted from Whiting P, Martin RM, Ben-Shlomo Y, et al. How to apply the results of a research paper on diagnosis to your patient. JRSM Short Reports 2013;4:7.
FN, False negatives; TP, true positives.

accuracy evidence.[14 15] 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[16] 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.[17–20] 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[17 19–45] 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.[41 45] 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.[45] 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.[41]

Accuracy definition: "Can health professionals define measures of accuracy?"

Six single-group studies assessed health professionals' understanding of the definition of measures of accuracy.[20 21 23 24 30 45] 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

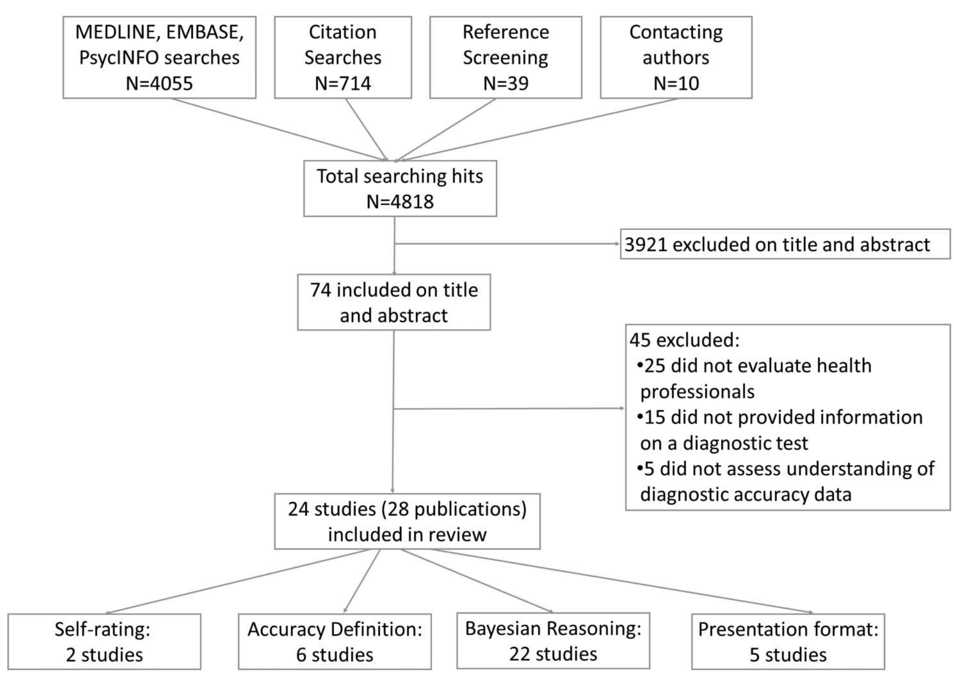


Figure 1 Flow of studies through the review process.

Table 2 Summary of included studies

	Total	Self-rating of understanding	Accuracy definition	Bayesian reasoning	Presentation format
Number of studies	24	2	6	22	5
Study design					
Single group	17	2	6	14	1
RCT	6	0	0	6	3
Multiple groups, unclear allocation	2	0	0	2	1
Participants					
Medical students	6	0	2	6	1
Mixed physicians	17	2	3	15	2
Single specialty	8	0	3	7	3
Other	4	0	0	4	1
How was the diagnostic information presented?					
Vignette/case study	6	0	0	6	2
Population scenario	13	0	1	12	3
Simulated patient	3	0	0	2	0
2×2 table	0	0	2	0	0
Research study extract	1	0	1	1	0
No information/unclear	3	2	2	2	0
How was understanding assessed?					
Questionnaire (multiple choice)	7	0	3	7	0
Questionnaire (open ended)	16	0	2	15	5
Interview	5	2	1	3	1
Unclear	1	0	0	1	0
Type of scenario					
Fictitious	7	0	2	7	0
Real life	16	0	2	15	5
Unclear	1	0	1	0	0
None	1	2	1	1	0
Measure of test accuracy assessed					
Sensitivity	22	2	6	20	4
Specificity/FPR	24	2	5	22	4
LR+	5	1	2	5	0
LR−	2	1	0	2	0
LR categories	1	0	0	1	0
Graphical display	2	0	0	2	1
PPV	21	1	3	19	3
NPV	6	0	1	6	1
ROC	1	1	0	0	0

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.

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] 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

Box 1 Example of population based scenarios and clinical vignettes

Self-rating of understanding:[41]

QUESTIONS USED IN TELEPHONE SURVEY

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:[40]

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[33]

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[39]

Typical angina chest pain: A 55year old man presented to your office with a 4 week history of sub-sternal 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?

'false-positive rate', seven provided a text description equivalent to these terms, one used both[39] and in five it was unclear whether terms or test descriptions were provided.[27 29 36–38]

Post-test estimation of probability was generally poor with a tendency to overestimation; only two studies found some evidence of successful application of Bayesian reasoning.[39 40] Thirteen studies provided data on the proportion of participants who correctly estimated the post-test probability of disease when provided with data on sensitivity and specificity (or false-positive rate) and the pretest probability of disease.[17 19 20 23–27 30 32 42 44 46] This varied from 0% to 61%, but the proportion of study participants who did not respond was between <1% and 40%.

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,[42] and five provided information for both a positive and a negative test result.[27 36 37 39 40] In one study it was unclear whether the test result provided should be interpreted as positive or negative[23] and in one study participants were questioned on how they interpreted test results in general.[41] 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 for positive and negative test results.[37 40] One study found that correct reasoning was applied for positive test results but that post-test probability was poorly estimated for negative test results.[39] One study found that although the post-test probability was consistently overestimated for a positive test result, estimates were correct for negative test results.[36] 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).[42]

Comparison of summary metrics for Bayesian reasoning

Six studies assessed the effects of providing test accuracy information using LRs (LRs),[20 27 30 38 40 44] only two of these studies provided information on the positive LR (LR+) and the negative LR (LR−).[27 40] Three studies provided a text description rather than using the term 'likelihood ratio',[30 40 44] and in one study a categorical approach based on the LR was used ('quite useless', 'weak', 'good', 'strong', or 'very strong').[38] Two studies included an additional scenario in which the LR information was provided graphically—one provided the information as a probability modifying plot,[44] 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.[40]

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.[27 30] One study demonstrated similarly poor post-test probability estimation for LRs (described in words) compared with sensitivity and specificity (presented numerically).[40] 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.[20 38 44] 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.[40 44]

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,[17 28] and a further study found no influence of age.[44] 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%).[42] Two studies demonstrated that research experience improved doctors' ability to correctly estimate post-test probability.[24 25] One study found that midwives were less likely than obstetricians to correctly estimate post-test probability of disease.[26]

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[19 26 31 32] than as probabilities (see box 1 for example).[42] 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.[47]

Two studies[19 32] 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%).[31]

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 (eg, 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.[18 48 49] 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.[50] 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.[51]

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[52] and GRADE.[53] 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
[1]School of Social and Community Medicine, University of Bristol, Bristol, UK
[2]The National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West at University Hospitals Bristol NHS Foundation Trust
[3]Unit of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, UK
[4]Peninsula 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.

REFERENCES

1. Kostopoulou O, Oudhoff J, Nath R, *et al*. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. *Med Decis Making* 2008;28:668–80.
2. Heneghan C, Glasziou P, Thompson M, *et al*. Diagnostic strategies used in primary care. *BMJ* 2009;338:b946.
3. Eddy D, Clanton C. The art of diagnosis: solving and clinicopathological exercise. In: Dowie J, Elstein A, eds. *Professional judgment: a reader in clinical decision making*. Cambridge: Cambridge University Press, 1988:200–11.
4. Falk G, Fahey T. Clinical prediction rules. *BMJ* 2009;339:b2899.
5. Knottnerus JA. Interpretation of diagnostic data: an unexplored field in general practice. *J R Coll Gen Pract* 1985;35:270–4.
6. Stengel D, Bauwens K, Sehouli J, *et al*. A likelihood ratio approach to meta-analysis of diagnostic studies. *J Med Screen* 2003;10:47–51.
7. Moons KG, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Acad Radiol* 2003;10:670–2.

8. Sackett DL, Straus S. On some clinically useful measures of the accuracy of diagnostic tests. *ACP J Club* 1998;129:A17–19.
9. Dujardin B, Van den Ende J, Van Gompel A, *et al*. Likelihood ratios: a real improvement for clinical decision making? *Eur J Epidemiol* 1994;10:29–36.
10. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet* 2005;365:1500–5.
11. Hayward RS, Wilson MC, Tunis SR, *et al*. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence-Based Medicine Working Group. *JAMA* 1995;274:570–4.
12. Wilson MC, Hayward RS, Tunis SR, *et al*. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. B. what are the recommendations and will they help you in caring for your patients? The Evidence-Based Medicine Working Group. *JAMA* 1995;274:1630–2.
13. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural Bayesians. *BMJ* 2005;330:1080–3.
14. Cochrane AJ. *Effectiveness and efficiency: random reflections on health services*. The Nuffield Provincial Hospitals Trust. London: The Royal Society of Medicine Press Ltd, 1972.
15. Knottnerus JA. *Evidence base of clinical diagnosis*. Wiley, 2002.
16. Centre for Reviews and Dissemination. *Systematic reviews: CRD's guidance for undertaking reviews in health care [Internet]*. York: University of York, 2009. (accessed 23 Mar 2011).
17. Casscells W, Schoenberger A, Graboys TB. Interpretation by physicians of clinical laboratory results. *N Engl J Med* 1978;299:999–1001.
18. Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol Rev* 1995;102:684–704.
19. Hoffrage U, Lindsey S, Hertwig R, *et al*. Medicine. Communicating statistical information. *Science* 2000;290:2261–2.
20. Steurer J, Fischer JE, Bachmann LM, *et al*. Communicating accuracy of tests to general practitioners: a controlled study. [Erratum appears in BMJ 2002 Jun 8;324(7350):1391]. *BMJ* 2002;324:824–6.
21. Argimon-Pallas JM, Flores-Mateo G, Jimenez-Villa J, *et al*. Effectiveness of a short-course in improving knowledge and skills on evidence-based practice. *BMC Fam Pract* 2011;12:64.
22. Agoritsas T, Courvoisier DS, Combescure C, *et al*. Does prevalence matter to physicians in estimating post-test probability of disease? A randomized trial. *J Gen Intern Med* 2011;26:373–8.
23. Bergus G, Vogelgesang S, Tansey J, *et al*. Appraising and applying evidence about a diagnostic test during a performance-based assessment. *BMC Med Educ* 2004;4:20.
24. Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. *Am J Med* 1981;71:991–8.
25. Borak J, Veilleux S. Errors of intuitive logic among physicians. *Soc Sci Med* 1982;16:1939–44.
26. Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ* 2006;333:284.
27. Chernushkin K, Loewen P, De Lemos J, *et al*. Diagnostic reasoning by hospital pharmacists: assessment of attitudes, knowledge, and skills. *Can J Hosp Pharm* 2012;65:258–64.
28. Curley SP, Yates JF, Young MJ. Seeking and applying diagnostic information in a health care setting. *Acta Psychol (Amst)* 1990;73:211–23.
29. Eddy DM. Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. *Judgement under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, 1982:249–67.
30. Estellat C, Faisy C, Colombet I, *et al*. French academic physicians had a poor knowledge of terms used in clinical epidemiology. *J Clin Epidemiol* 2006;59:1009–14.
31. Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Soc Sci Med* 2013;83:27–33.
32. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. *Acad Med* 1998;73:538–40.
33. Gigerenzer G. The psychology of good judgment: frequency formats and simple algorithms. *Med Decis Making* 1996;16:273–80.
34. Gigerenzer G. *Reckoning with risk: learning to live with uncertainty*. UK: Penguin, 2003.
35. Hoffrage U, Gigerenzer G. How to improve the diagnostic inferences of medical experts. In Kurz-Milcke E, Gigerenzer G, eds. *Experts in science and society*. New York: Kluwer Academic/Plenum Publishers, 2004:249–268.
36. Lyman GH, Balducci L. Overestimation of test effects in clinical judgment. *J Cancer Educ* 1993;8:297–307.
37. Lyman GH, Balducci L. The effect of changing disease risk on clinical reasoning. *J Gen Intern Med* 1994;9:488–95.
38. Moreira J, Bisoffi Z, Narvaez A, *et al*. Bayesian clinical reasoning: does intuitive estimation of likelihood ratios on an ordinal scale outperform estimation of sensitivities and specificities? *J Eval Clin Pract* 2008;14:934–40.
39. Noguchi Y, Matsui K, Imura H, *et al*. Quantitative evaluation of the diagnostic thinking process in medical students. *J Gen Intern Med* 2002;17:848–53.
40. Puhan MA, Steurer J, Bachmann LM, *et al*. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. *Ann Intern Med* 2005;143:184–9.
41. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. *Am J Med* 1998;104:374–80.
42. Sox CM, Doctor JN, Koepsell TD, *et al*. The influence of types of decision support on physicians' decision making. *Arch Dis Child* 2009;94:185–90.
43. Bachmann LM, Steurer J, ter Riet G. Simple presentation of test accuracy may lead to inflated disease probabilities. *BMJ* 2003;326:393.
44. Vermeersch P, Bossuyt X. Comparative analysis of different approaches to report diagnostic accuracy. *Arch Intern Med* 2010;170:734–5.
45. Young JM, Glasziou P, Ward JE. General practitioners' self ratings of skills in evidence based medicine: validation study. *BMJ* 2002;324:950–1.
46. Sassi F, McKee M. Do clinicians always maximize patient outcomes? A conjoint analysis of preferences for carotid artery testing. *J Health Serv Res Policy* 2008;13:61–6.
47. Gigerenzer G. *What are natural frequencies?* 2011;343:d6386.
48. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. *BMJ* 2003;327:741–4.
49. Hoffrage U, Gigerenzer G, Krauss S, *et al*. Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition* 2002;84:343–52.
50. Edwards W. 25. Conservatism in human information processing. In: Kahneman D, Slovic P, Tversky A, eds. *Judgement under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press, 1982:359–69.
51. Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision-makers when reading a Cochrane diagnostic test accuracy review. *Syst Rev* 2013;2:32.
52. Cochrane Diagnostic Test Accuracy Working Group. *Handbook for DTA reviews [Internet]*. The Cochrane Collaboration, 2013 (accessed 13 Oct 2014).
53. GRADE working group [Internet]. Secondary GRADE working group [Internet]. 2014, (accessed 27 Mar 2014). http://www.gradeworkinggroup.org/index.htm

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.fs. (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

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Reid (1998) [41] USA	Single group	300 practicing doctors	Sensitivity Specificity LR+ LR- ROC curves	None	Questioned regarding use and understanding of various measures	Telephone interview	None	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.
Young (2002) [45] Australia	Single group	50 GPs	Sensitivity Specificity PPV	No information	Asked to self-rate understanding of diagnostic terms.	Telephone interview	None	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.

b. Accuracy Definition

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Argimon-Pallas (2011) [21] Spain	Single group	152 family medicine residents in their second year of the Family Medicine training programme	Sensitivity Specificity PPV NPV LR+	Population based scenario	Information provided on total number of patients with target condition and number with and without condition testing positive	Questionnaire asked to calculate accuracy measures from raw data in scenario Administered before and after educational intervention (intensive and interactive four half-day sessions)	Unclear	Before task number of doctors correctly calculating figures were: Sensitivity: 42% Specificity: 34% PPV: 33% NPV: 26% LR+: 8% After intervention numbers more than doubled for all accuracy measures. Sensitivity: 82% Specificity: 79% PPV: 82% NPV: 80% LR+: 48%
Bergus(2004) [23] USA	Single group	43 medical students and residents (psychiatry and Internal Medicine)	Sensitivity Specificity	Extract from research study	Asked to identify sensitivity and specificity from report	Questionnaire (open ended)	Real life (major depression and panic disorder, congestive heart failure)	88% correctly identified the specificity and sensitivity of the test from the paper.
Berwick (1981)[24] USA	Single group	36 medical students, 45 interns and residents, 49 research doctors, 151	Sensitivity Specificity FPR	2x2 table	Asked to identify definitions based on 2x2 table (a, b, c, d used rather than numbers)	Questionnaire (MC)	Hypothetical (Disease K)	Practicing physicians were less able to correctly define sensitivity and specificity than medical students and research doctors. Exact values not reported

		full time doctors						
Estellat (2006)[30]	Single group	Senior doctors research and full time practice	Sensitivity Specificity LR+	2x2 table	2x2 table and short extract from study report.	Questionnaire. (multiple choice, Postal or given directly by one investigator)	Real life (CT for Pulmonary Embolism)	85% selected correct definition for sensitivity, 80% for specificity and 17% for LR+. High rate of 'do not know' for LR's (72%)
Steurer (2002)[20] *Related publication:* Bachmann (2003)[43] Switzerland	Single group	263 GPs	Sensitivity PPV	No information	Asked to select correct definition for various accuracy measures	Questionnaire (multiple choice)	Real life (Transvaginal ultrasound for endometrial cancer)	76% (95% CI 70-81%) correctly identified the definition of sensitivity, 61% (95% CI 45-67%) correctly identified the definition of PPV
Young (2002)[45] Australia	Single group	13 GPs	Sensitivity Specificity PPV	No information	Asked for verbal explanations of diagnostic terms	Interview	None	**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.

c. Bayesian Reasoning

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Agoritsas(2011) [22] Switzerland	RCT	1361 physicians of all clinical specialties	Sensitivity Specificity	Population based scenario	Sensitivity and specificity described in words and numerical frequencies (terms not used) for very accurate test (sensitivity and specificity 99%) Doctors randomised to receive information on different prevalence (1%, 2%, 10%, 25%, 95%) and no information	Multiple choice Questionnaire: Different categories of post-test probability offered: <60%, 60-79%, 80-94%, 95-99.9%, >99.9%	Screening test for viral disease in primary school	**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.

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Bergus(2004) [23] USA	Single group	43 medical students and incoming residents (psychiatry and Internal Medicine)	Sensitivity Specificity	Extract from research study and simulated patient	Asked to identify sensitivity and specificity from report and asked to apply these to a patient with a specified pre-test probability	Questionnaire (open ended)	Real life (major depression and panic disorder, congestive heart failure)	**Test result evaluated:** Unclear **PPV/NPV proportion correct:** 1/28 Med students, 0/15 residents **PPV proportion over/under:** NR
Berwick (1981) [24] USA	Single group	36 medical students, 45 interns and residents, 49 research doctors, 151 full time doctors	Sensitivity Specificity	Population based scenario	Sensitivity and specificity described in words (terms not used)	Questionnaire (MC)	Hypothetical (Disease K)	**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

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Borak(1982) [25] USA	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 sohpisticated doctors, <2% of nurses and other doctors **PPV proportion over/under:** NR
Bramwell (2006) [26]	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

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Casscells (1978) [17] USA	Single group	40 doctors 20 medical students	FPR	Population based scenario	Single scenario including prevalence and FPR	Interview (1 on 1 corridor discussion)	Hypothetical	**Test result evaluated:** Positive **PPV proportion correct:** 11/60 **PPV proportion over:** not stated; 27/60 said 95% and mean was 56% - correct value was 2% **PPV proportion under:** NR **Effect of experience:** No effect
Chernushkin (2012) [27] Canada	Single group	94 Pharmacists; 55 completed diagnostics knowledge and skills section (extracted here)	Sensitivity Specificity LR+ (numerical)	Population based scenario	Various different knowledge and skills questions related to application of accuracy measures	Online questionnaire	Real life	**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+.

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Curley 1990[28] USA	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 probabilty 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)[29] USA	Single group	100 doctors	FPR	Population based scenario	Single 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%

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Estellat (2006) [30] France	Single group	130 Senior doctors research and full time practice	Sensitivity Specificity LR+	Population scenario (different scenarios for sens/spec and LR+)	Sensitivity, specificity, LR+ (in words) and prevalence given	Questionnaire. (multiple choice for sens/spec and open for LR+)	Hypothetical	**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
Garcia-Retamero (2013)[31] Spain	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%

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Hoffrage (1998) [32] *Related publications:* Giggerenzer(1996) [33] Giggerenzer (2003) [34] Germany	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
Hoffrage (2000) [19] *Related publication:* Hoffrage (2004) [35] Germany	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, phynylketonuria, ankylosing spondylitis)	**Test result evaluated:** Positive **PPV proportion correct:** Long prob 18%, long nat 57%, short prob 50%, short nat 68%

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Lyman (1993) [36] USA	Single group	29 doctors; 21 nurses and pharmacists	Sensitivity Specificity	Vignette/Case study	Asked to estimate prevalence, sensitivity and specificity based on vignette then apply their values to get a post-test probability	Questionnaire (open ended)	Real life (mammography for breast cancer)	**Test result evaluated:** Positive and negative **PPV:** Consistently overestimated **NPV:** Estimates correct
Lyman (1994) [37] USA	Single group	39 mixed doctors, 15 nurses and pharmacists, 4 medical students	Sensitivity Specificity	Population based scenario	Various different estimates of sensitivity, specificity and prevalence	Questionnaire (open ended)	Hypothetical	**Test result evaluated:** Positive and negative **PPV:** Physicians and non-physicians overestimate post-test probabilities with increasing error associated with decreasing disease risk.

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Moreira (2008) [38] Belgium	Single group	50 Doctors attending course on tropical medicine	Sensitivity Specificity Categorical grouping based on LR	Unclear	Sensitivity and specificity values and LRs categorised as: 'quite useless', 'weak', 'good', 'strong', 'very strong'.	Questionnaire (multiple choice and open ended)	Mixed (4 real diseases and 2 dummy diseases)	**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

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Noguchi (2002) [39] Japan	Single group	224 medical students	Sensitivity Specificity	Vignette/Case-study	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	Questionnaire (open ended)	Coronary Heart Disease and Exercise Stress Test	**Test result evaluated:** Positive and negative **PPV:** Correct reasoning **NPV:** Poorly estimated

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Puhan (2005) [40] Switzerland	RCT	183 Senior family and internal medicine doctors	Sensitivity Specificity LR+ LR- Graphic based on LR	Vignette/Case study	Group 1: Sensitivity and specificity Group 2: Positive or negative likelihood ratio defined in words Group 3: simple graphic of 5 circles based on LR.	Questionnaire (open ended, conference)	Pulmonary Embolus, Myocardial Infarction, COPD, Temporal arteritis, flu, heart failure.	**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
Reid (1998)[41] USA	Single group	300 practicing doctors	Sensitivity Specificity	None	Questioned regarding use and understanding of various measures	Telephone interview	None	**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.

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Sox (2009)[42] USA	RCT	653 paediatricians	Sensitivity 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)	**Test result evaluated:** Negative **Post-test probability proportion correct: 1% (n=**5) (all from group 3) estimated correct value. Proportion nearly correct was 13% (group 1), 20% (group 2) and 19% (group 3) **Post-test probability proportion over**: 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) **Effect of experience:** Greater proportion of residents estimated a nearly correct probability (29%) compared to paediatricians with (15%) or without (15%) an academic affiliation.

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
Steurer (2002) [20] *Related publication:* Bachmann (2003) [43] Switzerland	RCT	263 GPs	Sensitivity Specificity LR+ (described in words)	Vignette/Case study	Generic question based on sensitivity and specificity for population based scenario. Group 1: Test positive, no information on accuracy Group 2: sensitivity and specificity Group 3: LR+ defined in words	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.
Vermeesch (2010)[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

d. Presentation Format

Study	Study design	Participants	Measures of accuracy assessed	How was the diagnostic information presented?	Information provided	How was understanding assessed?	Type of scenario	Results
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.

Garcia-Retamero (2013)[31] Spain	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)[32] *Related publications:* Giggerenzer(1996)[33] Giggerenzer (2003)[34] Germany	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)[19] *Related publication:* Hoffrage (2004)[35] Germany	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, phynylketonuria, 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%.

Sox (2009)[42] USA	RCT	635 paedia-tricians	Sensitivity 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)

1. Kostopoulou O, Oudhoff J, Nath R, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. Medical Decision Making 2008;**28**(5):668-80.
2. Heneghan C, Glasziou P, Thompson M, et al. Diagnostic strategies used in primary care. BMJ 2009;**338**:b946.
3. Eddy D, Clanton C. The art of diagnosis: solving and clinicopathological exercise. In: Dowie J, Elstein A, eds. Professional Judgment: A Reader in Clinical Decision Making. Cambridge: Cambridge University Press, 1988:200-11.
4. Falk G, Fahey T. Clinical prediction rules. BMJ 2009;**339**:b2899.
5. Knottnerus JA. Interpretation of diagnostic data: an unexplored field in general practice. The Journal of the Royal College of General Practitioners 1985;**35**(275):270-4.
6. Stengel D, Bauwens K, Sehouli J, et al. A likelihood ratio approach to meta-analysis of diagnostic studies. Journal of medical screening 2003;**10**(1):47-51.
7. Moons KG, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Academic radiology 2003;**10**(6):670-2.
8. Sackett DL, Straus S. On some clinically useful measures of the accuracy of diagnostic tests. ACP journal club 1998;**129**(2):A17-9.
9. Dujardin B, Van den Ende J, Van Gompel A, et al. Likelihood ratios: a real improvement for clinical decision making? European journal of epidemiology 1994;**10**(1):29-36.
10. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet 2005;**365**(9469):1500-5.
11. Hayward RS, Wilson MC, Tunis SR, et al. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence-Based Medicine Working Group. Jama 1995;**274**(7):570-4.
12. Wilson MC, Hayward RS, Tunis SR, et al. Users' guides to the Medical Literature. VIII. How to use clinical practice guidelines. B. what are the recommendations and will they help you in caring for your patients? The Evidence-Based Medicine Working Group. Jama 1995;**274**(20):1630-2.
13. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural bayesians. BMJ 2005;**330**(7499):1080-3.
14. Cochrane AJ. *Effectiveness and Efficiency: Random Reflections on Health Services*. The Nuffield Provincial Hospitals Trust. London: The Royal Society of Medicine Press Ltd., 1972.
15. Knottnerus JA. *Evidence Base of Clinical Diagnosis*: Wiley, 2002.
16. Centre for Reviews and Dissemination. Systematic Reviews: CRD’s guidance for undertaking reviews in health care [Internet]. York: University of York, 2009 [accessed 23.3.11].
17. Casscells W, Schoenberger A, Graboys TB. Interpretation by physicians of clinical laboratory results. N Engl J Med 1978;**299**(18):999-1001.
18. Gigerenzer G, Hoffrage U. How to Improve Bayesian Reasoning Without Instruction:
Frequency Formats. Psychological Review 1995;**102**(4):684-704.
19. Hoffrage U, Lindsey S, Hertwig R, et al. Medicine. Communicating statistical information. Science 2000;**290**(5500):2261-62.
20. Steurer J, Fischer JE, Bachmann LM, et al. Communicating accuracy of tests to general practitioners: a controlled study.[Erratum appears in BMJ 2002 Jun 8;324(7350):1391]. BMJ 2002;**324**(7341):824-26.

21. Argimon-Pallas JM, Flores-Mateo G, Jimenez-Villa J, et al. Effectiveness of a short-course in improving knowledge and skills on evidence-based practice. BMC Family Practice 2011;**12**:64.
22. Agoritsas T, Courvoisier DS, Combescure C, et al. Does prevalence matter to physicians in estimating post-test probability of disease? A randomized trial. Journal of General Internal Medicine 2011;**26**(4):373-8.
23. Bergus G, Vogelgesang S, Tansey J, et al. Appraising and applying evidence about a diagnostic test during a performance-based assessment. BMC Medical Education 2004;**4**:20.
24. Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. Am J Med 1981;**71**(6):991-98.
25. Borak J, Veilleux S. Errors of Intuitive Logic Among Physicians. Social Science & Medicine 1982;**16**(22):1939-44.
26. Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. British Medical Journal 2006;**333**(7562):284-86A.
27. Chernushkin K, Loewen P, De Lemos J, et al. Diagnostic reasoning by hospital pharmacists: Assessment of attitudes, knowledge, and skills. Canadian Journal of Hospital Pharmacy 2012;**65**(4):258-64.
28. Curley SP, Yates JF, Young MJ. Seeking and applying diagnostic information in a health care setting. Acta Psychol (Amst) 1990;**73**(3):211-23.
29. Eddy DM. Probabilistic reasoning in clinical medicine: problems and opportunities. In: Kahneman D, Slovic P, Tversky A, eds. Judgement under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982:249-67.
30. Estellat C, Faisy C, Colombet I, et al. French academic physicians had a poor knowledge of terms used in clinical epidemiology. Journal of Clinical Epidemiology 2006;**59**(9):1009-14.
31. Garcia-Retamero R, Hoffrage U. Visual representation of statistical information improves diagnostic inferences indoctors and their patients. Social Science & Medicine 2013;**83**:27-33.
32. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Academic Medicine 1998;**73**(5):538-40.
33. Gigerenzer G. The psychology of good judgment: Frequency formats and simple algorithms. Medical Decision Making 1996;**16**(3):273-80.
34. Gigerenzer G. *Reckoning with Risk: Learning to live with uncertainty*. UK: Penguin, 2003.
35. Hoffrage U, Gigerenzer GE-MA, Hoffrage Uhm-bmd. How to Improve the Diagnostic Inferences of Medical Experts. [References]. Kurz-Milcke, Elke [Ed]; Gigerenzer, Gerd [Ed] 2004;**:**(2004):314.
36. Lyman GH, Balducci L. Overestimation of test effects in clinical judgment. Journal of Cancer Education 1993;**8**(4):297-307.
37. Lyman GH, Balducci L. The effect of changing disease risk on clinical reasoning. Journal of General Internal Medicine 1994;**9**(9):488-95.
38. Moreira J, Bisoffi Z, Narvaez A, et al. Bayesian clinical reasoning: does intuitive estimation of likelihood ratios on an ordinal scale outperform estimation of sensitivities and specificities? Journal of Evaluation in Clinical Practice 2008;**14**(5):934-40.
39. Noguchi Y, Matsui K, Imura H, et al. Quantitative evaluation of the diagnostic thinking process in medical students. Journal of General Internal Medicine 2002;**17**(11):848-53.

40. Puhan MA, Steurer J, Bachmann LM, et al. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. Annals of Internal Medicine 2005;**143**(3):184-89.
41. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy. American Journal of Medicine 1998;**104**(4):374-80.
42. Sox CM, Doctor JN, Koepsell TD, et al. The influence of types of decision support on physicians' decision making. Archives of Disease in Childhood 2009;**94**(3):185-90.
43. Bachmann LM, Steurer J, ter RG. Simple presentation of test accuracy may lead to inflated disease probabilities. BMJ 2003;**326**(7385):393.
44. Vermeersch P, Bossuyt X. Comparative Analysis of Different Approaches to Report Diagnostic Accuracy. Archives of Internal Medicine 2010;**170**(8):734-35.
45. Young JM, Glasziou P, Ward JE. General practitioners' self ratings of skills in evidence based medicine: validation study. BMJ 2002;**324**(7343):950-51.
46. Sassi F, McKee M. Do clinicians always maximize patient outcomes? A conjoint analysis of preferences for carotid artery testing. J Health Serv Res Policy 2008;**13**(2):61-66.
47. Gigerenzer G. *What are natural frequencies?*, 2011.
48. Gigerenzer G, Edwards A. Simple tools for understanding risks: from innumeracy to insight. BMJ 2003;**327**(7417):741-44.
49. Hoffrage U, Gigerenzer G, Krauss S, et al. Representation facilitates reasoning: what natural frequencies are and what they are not. Cognition 2002;**84**(3):343-52.
50. Edwards W. 25. Conservatism in human information processing. In: Kahneman D, Slovic P, Tversky A, eds. Judgement under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982:359-69.
51. Zhelev Z, Garside R, Hyde C. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. Systematic reviews 2013;**2**:32.
52. Cochrane Diagnostic Test Accuracy Working Group. Handbook for DTA Reviews [Internet]: The Cochrane Collaboration, 2013 [accessed 13.10.14].
53. GRADE working group [Internet]. Secondary GRADE working group [Internet] 2014 [accessed 27.3.2014]. http://www.gradeworkinggroup.org/index.htm.