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 overstimated **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.