Bayesian clinical reasoning: does intuitive estimation of likelihood ratios on an ordinal scale outperform estimation of sensitivities and specificities?

Juan Moreira MD,[1,2] Zeno Bisoffi MD,[3] Alberto Narváez MD PhD[4] and Jef Van den Ende PhD[5,6]

[1]Researcher, Centro de Epidemiología Comunitaria y Medicina Tropical (CECOMET), Esmeraldas, Ecuador
[2]PhD Student, Department of Clinical Sciences, Institute of Tropical Medicine (ITM), Antwerpen, Belgium
[3]Head, Centro per le malattie tropicale-Ospedale Sacrocuore, Negrar, Italy
[4]Professor, Instituto de Investigaciones-Facultad de Ciencias Medicas-Universidad Central del Ecuador, Quito, Ecuador
[5]Professor, Department of Clinical Sciences, Institute of Tropical Medicine (ITM), Antwerpen, Belgium
[6]Head, Department of Tropical Medicine, University Hospital, Antwerpen, Belgium

Keywords
Bayes' theorem, clinical reasoning, educational tools

Correspondence
Juan Moreira
Institute of Tropical Medicine, Antwerpen
Belgium
Department of Clinical Sciences
Nationalestraat 155
2000 Antwerpen
Belgium
E-mail: jmoreira@itg.be

Accepted for publication: 13 December 2007

doi:10.1111/j.1365-2753.2008.01003.x

Abstract

Rationale: Bedside use of Bayes' theorem for estimating probabilities of diseases is cumbersome. An alternative approach based on five categories of powers of tests from 'useless' to 'very strong' has been proposed. The performance of clinicians using it was assessed.

Methods Fifty clinicians attending a course of tropical medicine estimated powers of tests and post-test probabilities using the classical vs. the categorical Bayesian approach. The estimation of post-test probability was assessed for real and dummy diseases in order to avoid the bias of previous knowledge. Accuracy of answers was measured by the difference with reference values obtained from an expert system (Kabisa).

Results Clinicians estimated positive likelihood ratios (LRs) a median of -1.07 $\log_{10}$ lower than Kabisa [interquartile range (IQR): -1.47; -0.80] when derived classically and -0.17 (IQR: -0.42; $+0.04$) when estimated categorically ($P < 0.001$). For negative LRs the median was $+0.39$ $\log_{10}$ higher (IQR: $+0.71$; $+0.08$) when derived classically and -0.18 $\log_{10}$ lower (IQR: $+0.03$; -0.36) when estimated categorically ($P < 0.001$).

Twenty (40%) disclosed not being able to calculate post-test probabilities using sensitivities and specificities. Regardless the approach post-test probabilities were overestimated both for real and dummy diseases [respectively $+1.23$ $\log_{10}$ (IQR: +0,67; +2.08) and $+2.03$ $\log_{10}$ (IQR: +0.49; +2.42)] ($P = 0277$), but the range was wider for the latter ($P = 0.001$).

Conclusions Participants were more accurate in estimating powers with a categorical approach than with sensitivities and specificities. Post-test probabilities were overestimated with both approaches. Knowledge of the disease did not influence the estimation of post-test probabilities. A categorical approach might be an interesting instructional tool, but the effect of training with this approach needs assessment.

Introduction

After exhausting all available information, the clinical reasoning process should yield a final probability of having a disease. This is also known as post-test probability, which is the probability of a disease conditional on the results of previously obtained data. The probability of any given disease at the beginning of a clinical encounter is known as pre-test probability. After asking the first question (or test) the new information converts the pre-test probability in a new (post-test) probability. The sequential processing yields the final post-test probability. The engine of this process is Bayes' theorem, which has been extensively advocated during the last years as an essential tool for clinical logic [1,2]. However, its use requires several complex and time-consuming mathematics that most clinicians are not inclined to do [3–7]. To overcome these constraints, a number of educational tools have been promoted, but none have been extensively applied in clinical practice [5,8,9]. It has also been argued that, although clinicians do not understand the underlying rules of conditional probabilities, they could nonetheless be experts in applying it intuitively [10].

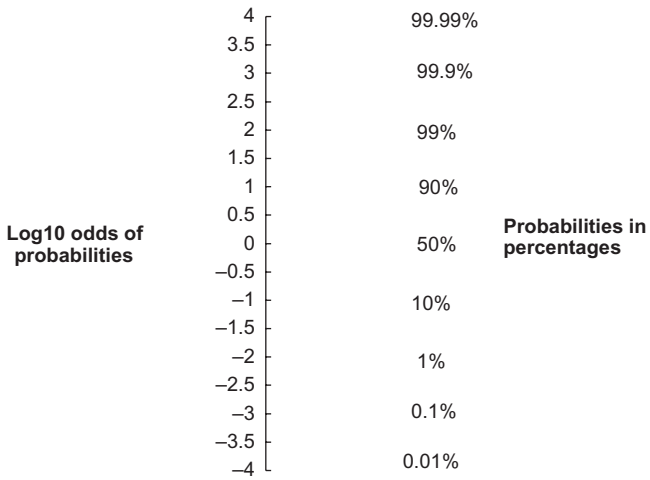

Figure 1 Correspondence of a scale in Log_{10} odds of probabilities and a scale of probabilities in percentages. In the left side the scale of Log_{10} odds of probabilities is depicted. In the right side the correspondent percentages. A probability of 50% corresponds to an even odds (1:1), its Log_{10} being 0. A probability of 90% corresponds to an odds of 9:1, which log_{10} is 0.95 (rounded 1). Inversely a probability of 10% corresponds to an odds of 1:9, which Log_{10} is –0.95 (rounded –1). In the same sense a probability of 99% corresponds to a rounded Log_{10} odds of 2 and a probability of 1% corresponds to a rounded Log_{10} odds of –2.

Seasoned clinicians commonly bypass mathematics just relying on knowledge, experience and common sense. However, mistakes could occur, especially when inappropriate guess estimations are done for the frequency of a disease or for the intrinsic power [likelihood ratio(LR)] of a test [11]. Nowadays, this becomes an important concern as new diagnostic tools are being made increasingly available. A feeling of overconfidence on the test result vs. clinical reasoning could easily arise. This is a major problem in low income countries where a considerable effort should be made for an efficient use of resources while in the meantime defending the people's rights for an optimal standard of care. To this purpose, user friendly tools to help clinicians to correctly estimate the disease probability without excessive testing would be most welcome.

Just 400 years ago, Napier developed logarithms to make complex calculations easier, and later Oughtred developed the first slide rule [12]. Fagan proposed a nomogram to find post-test probabilities by drawing a line passing from a pre-test probability through the LR of the applied test [8]. Although Fagan's nomogram was an interesting contribution to the bedside application of conditional probabilities in medicine, its use is still limited because (i) it is necessary to know how to calculate a positive and negative LR; (ii) only the effect of one test can be drawn at the same time, requiring repeated steps to see the entire evolution of probabilities; and (iii) interpreting negative LR represented as decimal numbers below one is very difficult.

Gathering Napier's original ideas behind transforming scales in logarithms and the mathematician Alan Turing's proposal of representing LR as base-10 logarithmic values [13], we proposed a logarithmic scale of odds, symmetrical around a 50% probability with values going up to 'certain' on one side and to 'impossible' on the other (Fig. 1). This scale [14]allows a summation of LR of test results as well as an understandable visualization of gains in probabilities, which is closer to clinical intuition. Indeed, we should keep in mind that LR are not linear – as they are ratios, having a skewed distribution – therefore they should be interpreted on a logarithmic scale and not on a linear one [15]. Transforming values of LR to base-10 logarithms and rounding to half the unit allows a representation in categories from 'a quite useless test', 'a weak test', 'a good test', 'a strong test' and 'very strong test'. A 'good confirmer' would be situated around a Positive Likelihood Ratio (LR+) of 10 allowing a gain of one step upwards, a 'good excluder' around a Negative Likelihood Ratio (LR–) of 0.1 bringing the pre-test probability one step downwards [14,16] (Table 1 and Fig. 1). Suppose for a given disease one has got a baseline suspicion of 1 in 1000, corresponding to a Log_{10} odds of probability of –3. If a 'good confirming test' turned out to be positive, it allows a gain of one step resulting in a log_{10} odds of post-test probability of -2 ($-3 + 1 = -2$) which means 1 in 100. In other words, a 'good confirming test' applied to a baseline suspicion of 1 in 1000 brings the probability one step upwards to 1 in 100, not more! In the same way a test which turned out to be negative can be subtracted from the pre-test probability according to its excluding power, namely the negative LR. We proposed as well to use the wording 'confirming power' and 'excluding power' instead of positive and negative LR, because these terms are a lot closer to physicians' intuition and daily language [4]. Actually this is an alternative and easier way to perform Bayesian calculations.

Table 1 Correspondence between likelihood ratios and categorical classes of power

Likelihood ratio	Log_{10} Likelihood ratio	Categorical class of power
100	2	very strong confirmer
33	1.5	strong confirmer
10	1	good confirmer
3	0.5	weak confirmer
1	0	useless
0.3	–0.5	weak excluder
0.1	–1	good excluder
0.03	–1.5	strong excluder
0.01	–2	very strong excluder

The usefulness of this ordinal log_{10} odds scale of LR should be proved. To this purpose, we intended to test how accurate physicians are in estimating the power of a test, when data about sensitivities and specificities are asked, compared with directly ranking a guess confirming and excluding power based on this scale. Additionally we wanted to assess how accurate they are in obtaining a post-test probability for a common disease both applying formal Bayes' theorem and guessing a post-test probability from given confirming and excluding powers. As previous knowledge could influence the results in probability estimation, we designed an experiment to confront the accuracy when dealing with real against dummy diseases.

Methods

The population

The study was carried out in a teaching setting at the Institute of Tropical Medicine of Antwerp in Belgium. Participants were all

physicians attending postgraduate courses of tropical medicine during the academic year 2005–2006. During the weeks before the study all received a formal training in epidemiology including test accuracy (sensitivity, specificity) and classical Bayes' theorem (positive and negative predictive value). They also had the opportunity to exercise with Kabisa, an interactive software for training in diagnosis in tropical contexts [17]; however, they were not yet taught the principles and use of LR or the ordinal $\log_{10}$ odds scale.

Reference data

For real diseases, values of test sensitivity and specificity were obtained from Kabisa software which was set for an African context (free access at http://www.kabisa.be). Kabisa has a logical engine to perform Bayesian computations based on data of around 250 diseases commonly found in a district hospital setting [17]. Each disease is linked with a list of clinical characteristics (including anamnesis, physical examination, basic laboratory and imaging) through their respective positive and negative LR. For LR calculation, data of the 'waiting room' prevalence are supplied in a database for every disease, along with sensitivity of all clinical characteristics related with the disease. These values are obtained from literature, when available, or from a consensus of experts. In Kabisa specificity values are derived from the complement of the false positive rate, which depends on the prevalence of other diseases in the same clinical setting and on the probability of finding the concerned characteristic in these diseases.

For the sake of offering a friendly and useful environment to the final user, negative LRs in Kabisa are displayed with values above 1. Instead of the ratio between false negatives and true negatives, the inverse is calculated (the 'unlikelihood ratio'). This trick allows a better comprehension of the power without changing the interpretation. However, for the purpose of comparison between approaches in this study, negative LRs in its original form, with the corresponding $\log_{10}$ transformation, were used.

The questionnaire

A questionnaire combining open ended and single choice questions was anonymously filled in by all participants. They were informed of the purpose of the questionnaire and gave their consent to participate; however, the hypothesis was not disclosed in order not to bias the answers. The content of the questionnaire was divided in three sections coinciding with the main study questions: (i) the estimation of confirming and excluding powers of four clinical findings related with two common diseases: pulmonary tuberculosis and appendicitis; (ii) the estimation of post-test probabilities for two suspected diagnosis (pulmonary embolism and ectopic pregnancy), taking into account a baseline pre-test probability and the combination of three clinical findings; and (iii) the estimation of post-test probabilities for two 'dummy' diseases given a baseline pre-test probability and the combination of three clinical findings.

In each section we evaluated two different approaches: a formal Bayesian approach with values of sensitivities and specificities; and the ordinal $\log_{10}$ odds (categorical) approach with classes of confirming and excluding powers: 'quite useless', 'weak', 'good', 'strong', 'very strong' (Table 1). In the first section we asked to estimate the values, while in the second and third section we asked to solve the problem using supplied values.

Two fictitious (dummy) diseases, both with a pre-test probability of 0.01% were 'concocted' by the researchers. For each disease three tests were related, one with a 'good', the other with a 'strong' and the third with a 'very strong' confirming power, but all with a 'weak' excluding power. A coined name and a short description of the disease were supplied. In the second and third section an option for disclosing not being able to perform Bayesian calculations was provided.

In order to avoid bias related with better general knowledge of one of the real diseases, the entire group was randomly split in two subgroups using a crossover methodology. For the third section this was not required – participants did not know neither of both dummy diseases.

Analysis

The outcome for the first section was the median of the differences in the $\log_{10}$ odds scale between a standard value derived from Kabisa and those estimated by participants: the Kabisa standard value was subtracted from the estimation of participants. Computing the difference for answers using the categorical approach was straightforward. For answers with the classical approach, LRs were calculated by researchers from data of sensitivities and specificities brought by participants. A further transformation of LRs in $\log_{10}$ was done to allow subtracting reference values. For the difference of excluding powers positive values must be interpreted as an underestimation and negative values as an overestimation, as negative LRs are usually below 1, hence their $\log_{10}$ transformation yields a negative number. This outcome was tested for every finding in both diseases (pulmonary tuberculosis and appendicitis).

For the second and third section the outcome was the median of the differences between the actual post-test probability and the estimations of participants. The actual post-test probability calculated by researchers was subtracted from the one estimated by participants. Once again these estimates were asked both through a classical and a categorical approach.

For post-test probabilities obtained through a classical calculation, exact values were asked. These values were further transformed in $\log_{10}$ odds to allow comparisons. For the categorical approach an ordinal scale was supplied. These categories were: 'between 1% and 9%', 'between 10% and 49%', 'between 50% and 89%', 'between 90% and 99%' and 'more than 99%' which matched the following $\log_{10}$ odds: −1.5; −0.5; 0.5; 1.5; 2.5 respectively (Fig. 1).

A detailed explanation of formulas used to calculate the median of differences for the power as well as the English version of one branch of the questionnaire is available in: http://www.itg.be/itg/uploads/clinicalsciences/MDMitg.zip.

The median of differences with reference values between approaches (classical or categorical) were compared. For power estimation and post-test probabilities of real diseases the Mann–Whitney *U*-test, its respective *P*-value and the effect size R was used. As for post-test probabilities of dummy diseases a crossover methodology was not required, comparison was done as two measures in the same group, therefore the Wilcoxon Signed-rank test with its respective *P*-value and the effect size R was used. Pitman's

t statistic was used to test the null hypothesis that samples were drawn from populations with identical variances [18].

Analysis was performed both with SPSS V. 10 (SPSS Inc, Chicago, IL, USA) and Microsoft Excel (Microsoft, Redmond, WA, USA).

Results

Population

Fifty medical doctors participated in the study. Mean age was 28.6 years (Std. Dev. 4.4). The female to male ratio was 1:5. Eighty-four per cent came from a high income country. They had a mean of 2.5 years (Std. Dev. 3.6) of professional experience. Seventy per cent were mainly concerned with clinical practice, 20% with public health and 8% with both.

Estimating powers

Concerning the confirming power, for appendicitis the median of differences for all findings considered together was –0.93 [interquartile range (IQR): –1.12; –0.80] when derived classically compared with –0.08 (IQR: –0.46; +0.17) when estimated categorically (U = 19; $P < 0.0001$; R = –0.80). For tuberculosis the median of differences were –1.20 (IQR: –1.54; –0.74) vs. –0.29 (IQR: –0.42; –0.07) (U = 33; $P < 0.0001$; R = –0.76). When every finding was assessed separately the same trend was observed (Table 2).

When analysing the excluding powers, the median of differences for appendicitis for all findings considered together was +0.49 (IQR: +0.85; +0.12) when estimated classically, while when estimated categorically it was –0.24 (IQR: +0.01; –0.39) (U = 56; $P < 0.0001$; R = –0.70). For tuberculosis the medians of differences were + 0.34 (IQR: +0.57; –0.01) vs. – 0.18 (IQR: +0.07; –0.30) (U = 125; $P = 0.0008$; R = –0.48). When findings were considered separately different trends were observed for appendicitis and tuberculosis. In the former significant differences were observed in all characteristics, while in the latter assessment of Ziehl staining was not different among groups, both overestimating its excluding power: –0.15 vs. –0.11 (Table 3).

When all findings for both diseases were considered together, the median difference for confirming powers was –1.07 (IQR: –1.47; –0.80) when derived classically and –0.17 (IQR: –0.42; +0.04) when estimated categorically (z-score = –6.010; *P*-value <0001; R = –0.6071). For excluding powers the median of differences was +0.39 (IQR: +0.71; +0.08) when derived classically and –0.18 (IQR: +0.03; –0.36) when estimated categorically (z-score = –4.831; *P*-value <0001; R = 0.4880) (Fig. 2).

Estimating post-test probabilities

Twenty of 50 (40%) participants disclosed not being able to calculate post-test probabilities through a formal Bayesian approach; nevertheless, some of them tried it anyhow.

Regardless the approach (classical or categorical), participants overestimated post-test probabilities both for real and dummy diseases. For the former the median of differences with actual post-test probabilities was 1.23 (IQR: 0.67; 2.08) while for the latter it was 2.03 (IQR: 0.49; 2.42). The difference between real and dummy diseases was not statistically significant (z-score = –1.086; *P*-value = 0.277) but the range was remarkably wider for the latter: variances were different ($t = -3.416$; d.f. = 48; $P = 0.001$) according with Pitman's *t*-test (Fig. 3).

Concerning real diseases, the median of differences with actual post-test probabilities for pulmonary embolism was 0.80 (IQR: 0.09; 1.81) with the classical approach and 1.23 (IQR: 0.98; 2.23) with the categorical. For ectopic pregnancy these values were 1.64 (IQR: 0.41; 2.66) and 1.14 (IQR: 1.14; 2.14)

Table 2 Median of differences in confirming powers between estimates by participants and Kabisa either with a classical (se & sp.) or with an alternative categorical (powers) approach

		Median of differences (IQR)	Mann–Whitney U	Asymp. 2-tailed *P*-value	effect size(R)
Appendicitis					
pain	powers	–0.20 (–0.20; 0.30)	22.50	<0.0001	–0.8042
	se & sp.	–0.98 (–1.19; 0.36)			
fever	powers	–0.40 (–0.40; 0.10)	58.50	<0.0001	–0.7028
	se & sp.	–0.94 (–1.38; –0.76)			
guarding	powers	–0.17 (–0.67; –0.17)	35.00	<0.0001	–0.7661
	se & sp.	–1.37 (–1.53; –1.08)			
leukocytes	powers	0.44 (–0.06; 0.94)	59.00	<0.0001	–0.6992
	se & sp.	–0.52 (–0.65; –0.21)			
Pulmonary tuberculosis					
cough	powers	0.27 (–0.23; 0.77)	40.00	<0.0001	–0.7428
	se & sp.	–0.70 (–0.87; –0.36)			
haemoptysis	powers	–0.14 (–0.89; 0.11)	40.50	<0.0001	–0.7412
	se & sp.	–1.39 (–1.66; –0.95)			
ziehl	powers	0.23 (–0.27; 0.23)	109.00	0.0001	–0.5564
	se & sp.	–0.77 (–1.20; 0.12)			
cavities	powers	–1.03 (–1.28; –0.78)	20.00	<0.0001	–0.8057
	se & sp.	–2.39 (–2.71; –1.92)			

IQR, interquartile range.

Table 3 Median of differences in excluding powers between estimates by participants and kabisa

		Median of differences (IQR)	Mann–Whitney U	Asymp. 2-tailed *P*-value	effect size(R)
Appendicitis					
pain	powers	–0.03 (0.47; –0.53)	89.50	<0.0001	–0.6166
	se & sp.	0.82 (0.95; 0.46)			
fever	powers	0.15 (0.15; –0.35)	89.50	<0.0001	–0.6189
	se & sp.	0.80 (1.35; 0.35)			
guarding	powers	–0.31 (0.19; –0.31)	139.50	0.0007	–0.4783
	se & sp.	0.39 (0.58; –0.02)			
leukocytes	powers	–0.76 (–0.26; –0.89)	96.00	0.0000	–0.5976
	se & sp.	0.21 (0.47; –0.25)			
Pulmonary tuberculosis					
cough	powers	0.22 (0.72; –0.28)	84.00	<0.0001	–0.6103
	se & sp.	1.08 (1.52; 0.68)			
haemoptisis	powers	–0.35 (0.15; –0.35)	171.00	0.0136	–0.3562
	se & sp.	0.15 (0.35; –0.17)			
ziehl	powers	–0.11 (–0.11; –0.98)	219.00	0.1015	–0.2340
	se & sp.	–0.15 (0.08; –0.58)			
cavities	powers	0.02 (0.02; –0.35)	147.00	0.0030	–0.4281
	se & sp.	0.20 (0.50; –0.05)			

IQR, interquartile range.

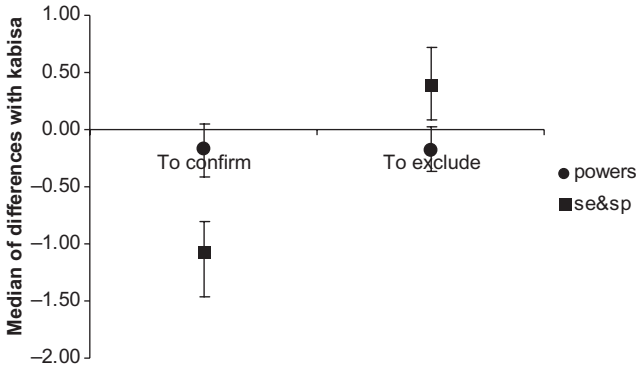

Figure 2 Median of differences between reference values and participant's estimates for all findings considered together. Y axis represents the differences between Kabisa and estimates by participants. The line crossing at 0.00 represents the point where no difference was observed. Error bars at left represent the median and interquartile range of differences of all findings to confirm considered together for both diseases (appendicitis and tuberculosis). Error bars at right represent the median and interquartile range of differences of all findings to exclude considered together. Rounded points represent the categorical approach while squared points represent the approach based on sensitivities and specificities. For excluding powers a negative sign means overestimation while a positive sign means underestimation.

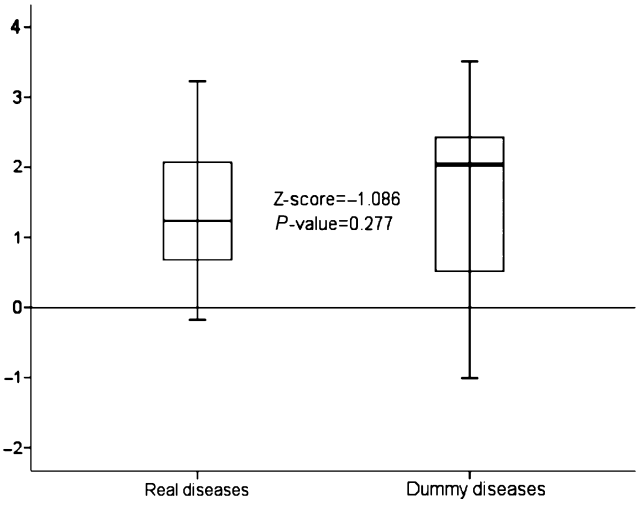

Figure 3 Differences between actual post-test probabilities and estimations of participants for real and dummy diseases. Y axis represents the difference between actual post-test probabilities and estimations of participants. The line crossing at 0 represents the point where no difference was observed. The box represents the interquartile range; the thick line represents the median. The 'T' bar includes distant values.

respectively. For neither disease was the difference statistically significant among those who made the calculation with the classical approach compared with those who used the categorical one (Table 4).

For dummy diseases, a classical approach yielded a median of differences of 1.23 (IQR: –0.40; 2.20) and a categorical 2.43 (IQR: 0.93; 2.43), which was significantly less accurate (z = –3.199, *P*-value = 0.001, R = –0.35) (Fig. 4).

Discussion

Main results

Participants were more accurate – closer to reference values – when weighing power of tests with an ordinal scale than when asked to estimate sensitivities and specificities; however, when estimating post-test probabilities inaccurate results were observed with both a classical and a categorical approach.

Table 4 Median differences with Kabisa in post-test probability estimation for two common diseases

		Median of differences (IQR)	Mann–Whitney U	Asymp. Sig (2-tailed)	effect size (R)
Pulmonary embolism	Formal Bayes	0.80 (0.09; 1.81)	168	0.1115	–0.2340
	Categorical approach	1.23 (0,98; 2.23)			
Ectopic pregnancy	Formal Bayes	1.64 (0.41; 2.66)	186	0,6304	–0.0751
	Categorical approach	1.14 (1.14; 2.14)			

IQR, interquartile range.

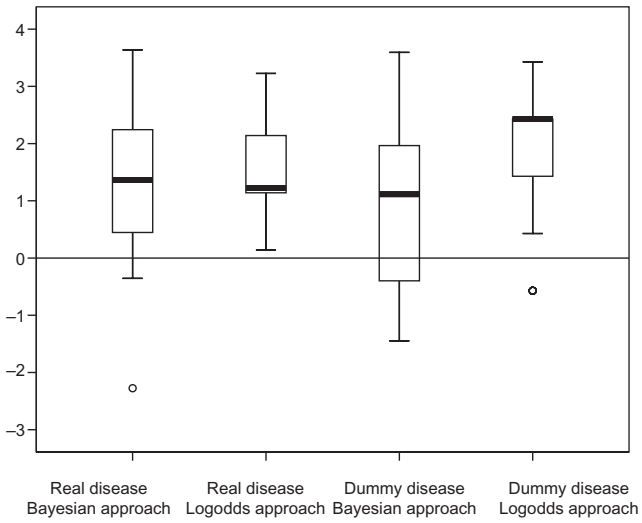

Figure 4 Differences between actual post-test probabilities and estimations of participants grouped by approaches for real and dummy diseases. Y axis represents the difference between actual post-test probabilities and estimations of participants. The line at crossing at 0 represents the point where no difference was found. The box represents the interquartile range; the thick line represents the median. The 'T' bar includes distant values and rounded points represent extreme values. Both approaches (Bayesian and ordinal scale) are compared for real and dummy diseases.

The population

A homogeneous group of physicians – regarding their age, experience and training – was chosen to test our hypothesis. This group is certainly not representative of the entire population of physicians, because of sample size and their characteristics. On the other hand, as they had training in epidemiology, one might expect worse results in other groups.

The fact that they had the opportunity to exercise with Kabisa could be a bias influencing their accuracy. However, we should consider on one side that they were not yet taught the principles and applications of the logodds scale and, on the other side, that in Kabisa there are roughly 6000 relations (predictor-disease), something impossible to memorize in a few weeks.

The method

The four discussed diseases were supposed to be well known by any general practitioner. Pulmonary tuberculosis is not frequent in western European countries, but participants were all following courses of tropical medicine. On the other hand, pulmonary tuberculosis is an ideal model to test medical decision issues, as the decision to start a treatment should be correctly taken with a low disease probability: a frequently observed error is a delayed treatment because of negative microscopy results [19].

The rationale for including two 'dummy' diseases was based on the hypothesis that previous knowledge could be of influence. If clinicians rely on previous clinical knowledge rather than on computing, one might suppose that for a real disease results would be more accurate. Surprisingly we found no differences.

The use of a logarithmic scale – instead of a linear one – for test power is justified by the fact that LRs are intrinsically exponential as they are ratios [6,15]. As clinicians often work at the extremes (low or high suspicion), the use of a logarithmic scale was also preferred for post-test probabilities in order to visualize these lower and upper values [4].

It could be argued that Kabisa is not a reliable reference standard for comparing estimations of participants as part of the Kabisa values is based on expert opinion rather than on evidence. As students are trained for a virtual district hospital in a tropical country, sensitivities found in the literature, 'waiting room' prevalence estimated by experts, and specificities computed with these values are probably the best estimates of real figures. Moreover, finding reliable evidence for setting specific sensitivities is almost impossible: many diseases, especially in a tropical context, do not have a reliable gold standard. Specificities, on the other hand, depend on the mix of other conditions which is highly setting dependent as explained above.

The results

Concerning the estimation of powers, when findings of both diseases were mixed a significant difference between different ways of estimation was still observed; however, the difference was smaller for excluding powers (Fig. 2). This is surprising as we supposed that excluding powers were more counter intuitive, hence worse results were expected, but this could be due to the fact that sensitivities (which have the most important influence on excluding powers), are more readily available in the literature and easier to estimate.

As expected, a previous training in Bayes' theorem does not improve the ability of clinicians, either to estimate the discriminative value of a test or to estimate pre-test and post-test probabilities. Our results confirm previous research in the topic [20]; nonetheless, many authors still argue that clinical training should emphasize the teaching of Bayes' theorem [3,10,21]. It is undeniable that the principles of conditional probabilities should be

applied in medicine; however, a simplified, more practice-based system, with a sound common language, should be used [4,14].

The inferiority of sensitivities and specificities for estimation of post-test probabilities compared to crude LRs explained in plain language or an inexact numerical graphical representation of LRs was already proven by others [22,23]. Interestingly Puhan split LRs in the inexact numerical graphical representation in four categories under 10, and only one over 10, concealing the difference between a strong and very strong test (the former contributing for an increment of 1.5 $\log_{10}$odds and the latter contributing for an increment of 2 $\log_{10}$odds) [14].

Contrarily to what we were expecting physicians do not estimate better post-test probabilities with the categorical approach. It is not surprising that they were not able to perform a complex Bayesian calculation; however, we thought that offering an ordinal, user-friendly scale would result in better estimates of post-test probabilities. The explanation could be that even a simplified model needs some training. Interestingly, this seems to be confirmed by the fact that they performed significantly worse with the categorical approach, when a dummy disease was concerned.

The interpretation of the results should pay attention to the fact that median differences were obtained in a base-10 logarithmic scale; hence the differences between the expected result and the estimations of participants would appear quite larger if applied on a linear scale. Indeed, the meaning of an overestimation of two categories (one $\log_{10}$) for a characteristic with a LR of 10 is that the participant assigned a LR that was around 100. This could mean a shift from a post-test probabilities of 10% to 90%!

Future research

Although participants did not perform better with the categorical scale, it would be worthwhile to see what the effect of training with this approach is. A further step should be a formal assessment of the impact of training in the simplified (categorical) didactic model, compared with the classical training on Bayes.

Acknowledgements

The authors wish to thank the postgraduate students attending the course in Tropical Medicine and International Health at the Institute of Tropical Medicine of Antwerp during the academic year 2005–2006, who agreed to participate in this study, as well as the group corresponding to the academic year 2004–2005, who participated in the pilot survey. We are also deeply grateful with Joris Menten for his input as consultant statistician.

References

1. Sackett, D., Haynes, R., Guyatt, G. H. & Tugwell, P. (1991) Clinical Epidemiology: A Basic Science for Clinical Medicine, 2nd, edn. Boston: Little, Brown and Company.
2. Jaeschke, R., Guyatt, G. H. & Sackett, D. L. (1994) Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *Journal of the American Medical Association*, 271 (9), 703–707.
3. Grimes, D. A. & Schulz, K. F. (2005) Refining clinical diagnosis with likelihood ratios. *Lancet*, 365 (9469), 1500–1505.
4. Van den Ende, J., Moreira, J., Basinga, P. & Bisoffi, Z. (2005) The trouble with likelihood ratios. *Lancet*, 366 (9485), 548.
5. McGee, S. (2002) Simplifying likelihood ratios. *Journal of General Internal Medicine*, 17 (8), 646–649.
6. Dujardin, B., Van den Ende, J., Van Gompel, A., Unger, J. P. & Van der Stuyft, P. (1994) Likelihood ratios: a real improvement for clinical decision making? *European Journal of Epidemiology*, 10 (1), 29–36.
7. Eddy, D. M. (1982) Probabilistic reasoning in clinical medicine: problems and opportunities. In Judgement Under Uncertainty: Heuristics and Biases (eds Kahneman, D., Slovic, P., Tversky, A.), pp. 249–267. Cambridge: Cambridge University Press.
8. Fagan, T. J. (1975) Nomogram for Bayes theorem. *New England Journal of Medicine*, 293 (5), 257.
9. Glasziou, P. (2001) Which methods for bedside Bayes? *ACP Journal of Club*, 135 (3), A11–A12.
10. Gill, C. J., Sabin, L. & Schmid, C. H. (2005) Why clinicians are natural bayesians. *British Medical Journal*, 330 (7499), 1080–1083.
11. Tversky, A. & Kahneman, D. (1974) Judgement under uncertainty: heuristics and biases. *Science*, 185, 1124–1131.
12. Stoll, C. (2006) When slide rules ruled. *Scientific American*, 294, 69–75.
13. Good, I. J. (1979) Studies in the history of probability and statistics XXXVIII. AM Turing's statistical work in world war II. *Biometrika*, 66, 393–396.
14. Van den Ende, J., Bisoffi, Z., Van Puymbroek, H., *et al.* (2007) Bridging the gap between clinical practice and diagnostic clinical epidemiology: pilot experiences with a didactic model based on a logarithmic scale. *Journal of Evaluation in Clinical Practice*, 13 (3), 374–380.
15. Zhou, X. H., Obuchowski, N. A. & McClish, D. K. (2002) Statistical Methods in Diagnostic Medicine, 1nd, edn. New York: Wiley.
16. Van den Ende, J., Van Gompel, A., Van den Enden, E., Van Damme, W. & Janssen, P. A. J. (1994) Bridging the gap between clinicians and clinical epidemiologists: bayes theorem on an ordinal scale. *Theoretical Surgery*, 9, 195.
17. Van den Ende, J., Blot, K., Kestens, L., Van Gompel, A., Van den Enden, E. (1997) Kabisa: an interactive computer-assisted training program for tropical diseases. *Medical Educ*, 31 (3), 202–209.
18. Howell, D. (1997) Statistical Methods for Psychology, 4nd, edn. Belmont, CA: Duxbury.
19. Basinga, P., Moreira, J., Bisoffi, Z., Bisig, B. & Van den Ende, J. (2007) Why are clinicians reluctant to treat smear-negative tuberculosis? An inquiry about treatment thresholds in Rwanda. *Medical Decision Making*, 27 (1), 53–60.
20. Noguchi, Y., Matsui, K., Imura, H., Kiyota, M. & Fukui, T. (2004) A traditionally administered short course failed to improve medical students' diagnostic performance. A quantitative evaluation of diagnostic thinking. *Journal of General Internal Medicine*, 19 (5 Part 1), 427–432.
21. Peirce, J. C. & Gerkin, R. D. (2003) The likelihood ratio as one of – if not the most important – operating characteristic of a diagnostic test. *Journal of General Internal Medicine*, 18 (1), 75.
22. Steurer, J., Fischer, J. E., Bachmann, L. M., Koller, M. & ter, R. G. (2002) Communicating accuracy of tests to general practitioners: a controlled study. *British Medical Journal*, 324 (7341), 824–826.
23. Puhan, M. A., Steurer, J., Bachmann, L. M. & ter, R. G. (2005) A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. *Annals of Internal Medicine*, 143 (3), 184–189.