Journal of Evaluation in Clinical Practice ISSN 1356-1294

Bridging the gap between clinical practice and diagnostic clinical epidemiology: pilot experiences with a didactic model based on a logarithmic scale

Jef Van den Ende MD PhD,[1,2] Zeno Bisoffi MD,[3] Hugo Van Puymbroek MD,[4] Patrick Van der Stuyft MD PhD,[1,5] Alfons Van Gompel MD,[6] Anselm Derese MD PhD,[7] Lutgarde Lynen MD,[8] Juan Moreira MD MSc[9,10] and Paul Adriaan Jan Janssen MD[11]

[1]Professor, Department of Clinical Sciences, Institute of Tropical Medicine, Antwerp, Belgium.
[2]Head, Department of Tropical Medicine, University Hospital, Antwerp, Belgium
[3]Doctor, Centro per le Malattie Tropicali, Negrar, Verona, Italy
[4]Lecturer, Universitaire Instellingen Antwerpen, Belgium
[5]Professor, Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium
[6]Professor, Clinical Services, Institute of Tropical Medicine, Antwerp, Belgium
[7]Professor, Faculty of Medicine, Ghent University, Belgium
[8]Doctor, Department of Clinical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
[9]PhD student, Department of Clinical Sciences, Institute of Tropical Medicine, Antwerp, Belgium
[10]Centro de Epidemiologia Comunitaria y Medicina Tropical, Esmeraldas, Ecuador.
[11]Founder (deceased) Janssen Research Foundation, Beerse, Belgium

Keywords
Bayes' theorem, clinical epidemiology, clinical reasoning, didactic model, likelihood ratio

Correspondence
J. Van den Ende
Institute of Tropical Medicine
Nationalestraat 155
2000 Antwerpen
Belgium
E-mail: jvdende@itg.be

Accepted for publication: 23 January 2006

doi:10.1111/j.1365-2753.2006.00710.x

Abstract

Rationale From general practitioners to academic staff, clinicians continue to have difficulties in applying clinical epidemiology in their everyday work. They do not fully understand the logical rules behind the numbers and they do not recognize these rules in their work. We present a new model where the pre-test and the post-test probabilities are converted to $\log_{10}$ of odds, and the likelihood ratio (LR) to its own $\log_{10}$.
Methods Following Bayes' theorem, adding the $\log_{10}$LR to the $\log_{10}$ pre-test odds gives the $\log_{10}$ post-test odds, which can easily be represented on a logarithmic scale. In addition, by rounding the $\log_{10}$LR to half the unit, we create classes of discriminative power of tests, close to intuitive estimation. This model generates also a user-friendly diagram, adding considerably to the understanding of Bayes' theorem. We evaluated the effect of the rounding, the current use of the classical model and the acceptability of the new model.
Results Rounding 10 disease characteristics to half the unit gives an absolute error of less than half a unit. After six explanations of Bayes' theorem, only 6/16 medical students were capable of answering simple questions about predictive value. When asked about weight of disease characteristics, no one of the 50 clinicians mentioned sensitivity, specificity, predictive value or LR. With the new model, more than 80% of trainees found medical decision making easier to understand and recognized the theory in their practice.
Conclusions We conclude that our model of diagnostic clinical epidemiology offers a logical environment for an easy and rapid assessment of the evolution of disease probability with consecutive tests, providing a scientific format for 'qualitative' clinical estimations.

Introduction

Clinical decision making in a context of limited diagnostic facilities

Our interest in clinical epidemiology began more than a decade ago, when we reasoned about the difference in the clinical decision process between the 'rich' Western medicine and the 'poor' tropical medicine. As our working hypothesis we assumed that in a more developed setting, medical decisions for a particular disease most often require a 'diagnosis' (i.e. a probability of disease close to 100%) or the exclusion of a diagnosis (i.e. a probability of disease close to 0%), as very specific and sensitive tests are usually available for most conditions. On the contrary, in a poorer context,

many decisions are inevitably taken under conditions of uncertainty, as only a limited set of diagnostic facilities is available. Therefore, we thought that the teaching of diagnostic clinical epidemiology should be particularly helpful under those conditions. We hoped to reduce diagnostic uncertainty through a more rigorous weighing of disease characteristics (symptoms, signs and the available diagnostic facilities), and to deal with the remaining uncertainty by accurately balancing the consequences of alternative decisions, through a threshold approach [1,2]. Since 1991, we integrate systematically this 'new' discipline in our teaching, not as a separate course, but embedded in the clinical lectures. Based on a continuous feedback by trainees, we developed a simplified didactic model that we believe could be of far more general interest. The incentive to this model was given by the late Paul Janssen, famous founder of the Janssen Pharmaceutical Company, who stated that clinicians think with classes of certainty and classes of weight of evidence of disease characteristics, not with numbers. He stressed further that clinical epidemiology would never find his way to clinicians if we continued to go on with multiplications, as clinicians are 'adding' evidence, not 'multiplying'. We designed a representation of probability and likelihood ratio's (LR) estimates on a logodds scale, but we found afterwards that this idea came very close to (unpublished) ideas developed by the well-known mathematician Alan Turing during World War II [3].

Crucial messages of diagnostic clinical epidemiology

Experience in training clinical epidemiology from the level of village health workers to academic staff leaves us and others with a disappointing impression of the user-friendliness of current models for teaching diagnostic logic [4,5].

Any new model should foster the teaching of five crucial messages of diagnostic clinical epidemiology: (1) The clinician should always start from the *pre-test probability* of a disease: without this starting point, interpretation of tests is impossible. (2) Every symptom, sign or test has a *'relative power'*. No single test is absolute, and its intrinsic value is independent of the patient. The relative power of a test can be calculated directly as a LR or indirectly from sensitivity and specificity. Pre-test probability and the intrinsic value of the test together make the final evidence after test. (3) The discriminative power of a test is often *asymmetric* (different for a positive vs. a negative test result), contrary to an intuitive feeling. A test can have a spectacular LR+, but an insignificant LR– and hence be worthless when negative, or vice versa. (4) For each disease in each setting, clinicians need to reach a *decision threshold* [1]. In case of a decision concerning a costly or dangerous test, the test and the test-treatment threshold should also be estimated [2,6]. (5) In order to avoid a 'freeze' ('diagnostic tunnel', 'premature closure') starting from key findings, all conditions that are both dangerous and treatable, even if they are rare, should be explored until all relevant disease characteristics are exhausted [7].

Analysis of the present model

Classical clinical epidemiology teaches the first four messages, but many professionals have difficulties recognizing them in the hundreds of numbers they have to deal with during training, and applying them afterwards in their clinical work. The calculations are so cumbersome that few clinicians do them. Instead, they rely on expertise and intuitive assessment. Current clinical epidemiology language is foreign to clinicians.

An often-heard criticism of the Bayesian approach to clinical diagnosis is that clinicians should apply strict mathematical calculations on rather *weak data*: first they estimate probabilities with a wide range of uncertainty, and afterwards they calculate with high precision [8]. Intuitively, many clinicians find that this does not make sense.

Clinicians have difficulty assigning numbers to their probability estimates in a given patient. A large array of numbers can result from each frequency determinant (e.g. the word 'probable' is expressed in probabilities from 0.3 to 0.95 by clinicians) [9,10]. Moreover, the strict application of Bayesian calculation when multiple tests are involved may be misleading, as we often ignore whether tests are interrelated and to which extent. With all these limitations in mind, one could wonder if the Bayesian approach is useful at all, that is, if it provides any advantages over the purely 'qualitative' and intuitive reasoning.

Prevalence or pre-test probability and post-test probability are estimated by clinicians on a nominal or an ordinal scale, not on a strictly numerical one. To apply the rules of clinical epidemiology, they have to convert a nominal scale to numbers. Moreover, the representation of pre-test and post-test probability on a continuous *linear scale* from 0% to 100% is not always adequate, as the ranges in which clinicians work are often close to 0% or to 100% [11,12]. Pre-test probability often starts from very low numbers, as in all screening and 'check-ups' which are a substantial part of general practice. Treatment thresholds, specially in a western, referral hospital setting, may be close to 100%. The distinction between 1% and 0.01%, or between 99% and 99.99%, might be more decisive than a distinction within the range of 10% to 90% (The first two distinctions affect the number of unnecessarily treated or worried patients with a factor 100, the last only with a factor 9, though it looks the most impressive on a linear scale). An identical scaling problem is seen with LRs. LR values are not linear, as they are ratios; they have a skewed distribution. Intuitively, one would estimate the discriminative power of a LR+ of 100 as 10 times that of a LR+ of 10, which is false. If we compare the effect of a test with a LR+ = 10 with that of a test with LR+ = 100, for example, for a pre-test probability of 1%, the increment in positive predictive value would be 8% versus 49%. The difference is less than we intuitively expect from the difference between LR 10 and 100. The values of LRs are intrinsically exponential; therefore, they should be interpreted on a logarithmic scale, as is done in logistic regression, and for the calculation of their confidence intervals [3,13–19]. Some authors have proposed to represent LRs as logarithms because users can add them instead of multiplying them when performing more than one test (in case of independence) [20].

Current values of LR are also confusing for clinicians, in that they do not show an existing symmetry: for example, the negative equivalent of a LR+ of 80 would be an LR– of 0.0125, which is not clear at a first glance, and is time-consuming to explain. Current representation is not adequate for the explanation of symmetry and asymmetry of tests.

Pre-test odds can be multiplied by the LR of the test, resulting in the post-test odds. Clinicians thus have to convert pre-test probabilities to odds, and post-test odds again to probabilities. This is

another, and major obstacle to clinical use of current clinical epidemiology. Fagan's nomogram can be of help, but clinicians need to wander around with this nomogram for everyday clinical work, they need to know the data, and it helps for just one step [21]. The use of a frequency format, instead of probabilities, has also been proposed to overcome this problem, but it reduces calculations by just one step, and clinicians have to convert data from the literature to this format first [22]. The use of a 2 by 4 contingency table, instead of formulas, comes also close to the frequency format solution [23].

Up until now, the *marginal benefit* of a test has to be calculated as the difference between pre-test and post-test probability for every individual clinical situation. It depends largely on the level of pre-test probability: the same test will give a low gain when applied to a low pre-test probability, and a high gain when applied to moderate pre-test probability (Fig. 1).

In the case of a negative result of a test, a confusion might arise between the post-test probability and the often-quoted negative predictive value, as the latter gives the true negatives over all negatives, whereas the former gives the false negatives over all negatives.

The classical threshold formulas are appropriate to calculate formally thresholds for clinical decisions, but they are inapplicable in daily practice [2,24]. Two of us (VdEJ & BZ) misinterpreted for years the classical formulas, until we were able to recognize them in the simplified model. Yet, we are convinced that the threshold concept (the fourth 'crucial message') is a basic element of the expert clinical reasoning, and should therefore be integrated in any teaching of clinical decision making [23].

Finally, since the late Feynman, famous through his Feynman diagrams, emphasis has been put on the importance of an intelligent, intelligible and user-friendly graphical representation of concepts. Recently, this law has been called the Feynman–Tufte principle [25]. We were looking for a more convivial model than the 2 by 2 or 2 by 4 table, hitherto the cornerstone of graphical representation of clinical epidemiology [12]. Parallel to this graphic representation, we were looking for a more intelligible language than the current terms of clinical epidemiology, which are based on mathematical grounds, not on clinical concepts.

We have developed a simplified method for the teaching of clinical epidemiology applied to diagnostic reasoning [7]. Based on this model, countless teaching sessions have been carried out during the last decade in four continents (Europe, Africa, Asia and Latin America). We describe our model and show how it fits the five crucial messages of diagnostic clinical epidemiology. We will also try to demonstrate that our model, yet greatly simplified if compared with the classical one, is mathematically sound, and easily accepted by clinicians.

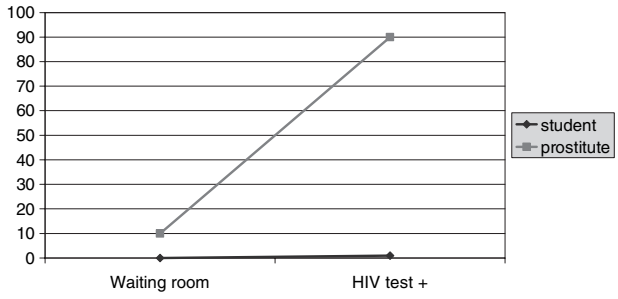

Figure 1 Effect of a positive ELISA HIV test, with different pre-test probabilities, shown on a linear scale. We order an HIV test for a worried student with a history of one unprotected sexual contact, and for a sex worker. The student starts at 0.01%, the prostitute at 10%, before questioning or physical examination (waiting room pre-test probability). The marginal benefit depends largely on the level of pre-test probability: the same test will give a low gain when applied on a low pre-test probability (0.99% in the student), and a high gain when applied on an already high pre-test probability (80% in the sex worker. ELISA, enzyme linked immuno sorbent assay.

Materials and methods

Description of the model

When we try to stay close to the intuitive clinical estimation and language, without numerical conversion, applying Bayes' theorem is impossible. We present a simplified numerical transformation, which could make sense to clinicians. 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, intuitively makes sense to clinicians (Table 1). The values clinicians are often interested in are close to 0% or 100%. Therefore, we offer clinicians values of odds with exponents of 10: 10/10, 100/10, 1000/10 and so on towards 'certain', and the inverse 10/100, 10/1000 and so on towards 'impossible'. At the lower end, the scale can be presented as probabilities because values for odds and probabilities are almost the same under 10%. This scale is so familiar that it takes only a few minutes to explain. It also comes close to an ordinal scale, representing the classes of certainty that clinicians are interested in.

Few clinicians know the exact value of the LRs of test results, but they have in mind an estimation of their value as 'quite useless', 'weak', 'good', 'strong' and the like. When rounded to half the unit of the logarithm of 10, this value could correspond to a limited number of clinical power classes (Table 2). A 'good confirmer' would be situated around a LR+ of 10, a 'good excluder' around a LR– of 0.1.

Another advantage of representing LR's in logarithms is offered by the possibility to represent LR– in the same way as the negative

Table 1 Comparison of probability, odds and $\log_{10}$ odds for different values of $\log_{10}$ odds

Probability	Odds	Log_{10} odds
0.000009	0.00001	–5
0.00009	0.0001	–4
0.0009	0.001	–3
0.009	0.01	–2
0.09	0.1	–1
0.50	**1**	**0**
0.91	10	1
0.99	100	2
0.999	1000	3
0.9999	10000	4
0.99999	100000	5

Table 2 Comparison of absolute values of LR, $\log_{10}$LR and intuitive classes of LR

LR	Log_{10}LR	Clinical power class
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

LR, likelihood ratio.

predictive value, that is as the ratio between specificity and 1-sensitivity, instead of the ratio between 1-sensitivity and specificity. In that way, LR– is represented by a whole number instead of by a fraction of 1, and is symmetric to LR+. Assigning a minus sign to the logarithm of this ratio gives the classical LR–, because log (a/b) = –log (b/a).

Suppose a test with sensitivity 90 and specificity 90:

- The LR– would be 10/90, with a LR– of 0.11, and a $\log_{10}$LR– of –0.95.
- If we reverse the ratio, we would have:
- LR–: 9, $\log_{10}$LR–: 0.95.
- The 'nine' makes much more sense to the clinician than the 0.11, and the logs are the same, except for the sign.

If we represent disease probability as $\log_{10}$ odds instead of probabilities or odds (Table 1), and if we represent LRs as $\log_{10}$ odds (Table 2), we can apply directly Bayes' theorem, while staying close to clinical intuition for pre-test probability and for test power.

pre-test odds × LR = post-test odds

$\log_{10}$ pre-test odds + $\log_{10}$LR = $\log_{10}$ post-test odds

Trainees can estimate their pre-test probability in classes (corresponding to rounded $\log_{10}$ pre-test odds). They might find the LRs in the literature, calculate them from published sensitivities and specificities or, most often, they can estimate them from their clinical practice. Adding the rounded $\log_{10}$LR or the power class of the test gives the post-test class.

clinical class pre-test + $\log_{10}$LR = clinical class post-test

clinical class pre-test + power class = clinical class post-test

A representation of this logarithmic transformation is easy to understand. Figure 2 shows the evolution of the probability of the acquired immunodeficiency syndrome (AIDS) in a patient who complains of diarrhoea of a 3-month duration and important weight loss, with estimated values of pre-test probability and LRs in a given situation. Physical examination reveals enlarged lymph nodes in the cervical and axillary regions. Inspection of the mouth is negative. An enzyme linked immuno sorbent assay (ELISA) HIV test is positive. The values of pre-test probability and LRs are drawn from the Kabisa software [26].

The marginal benefit coincides directly with the logarithm of the LR, thus with our representation of the test power itself. If we

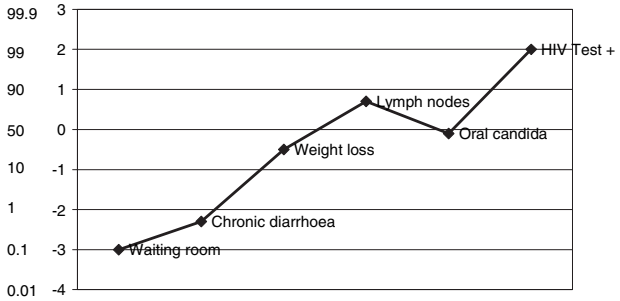

Figure 2 Effect of consecutive disease characteristics on the probability of AIDS in a given patient, shown on a scale based on the $\log_{10}$ odds. A patient complains of diarrhoea of a 3-month duration and important weight loss; physical examination reveals enlarged lymph nodes in the cervical and axillary regions. Inspection of the mouth is negative. Finally an ELISA test is positive. The Y-axis shows two scales: the first is the probability scale, giving the post-test probability; the second one is the $\log_{10}$ odds scale. ELISA, enzyme linked immuno sorbent assay.

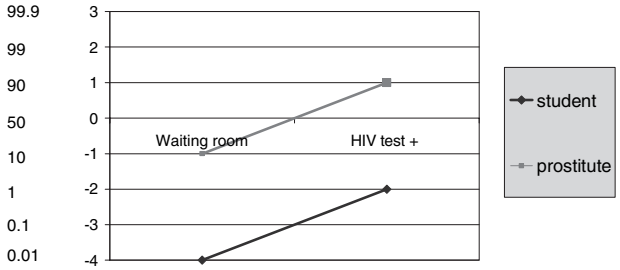

Figure 3 Effect of a positive ELISA HIV test with different pre-test probabilities, shown on a logarithmic scale. We order an HIV test for a worried student with a history of one unprotected sexual contact, and for a sex worker. The student starts at 0.01%, the sex worker at 10%. The marginal benefit is two classes in both cases, equal to the value of the LR of the test. ELISA, enzyme linked immuno sorbent assay; LR, likelihood ratio.

order an ELISA test for a worried student with a history of one unprotected sexual contact, we might start from 1/10 000 ($\log_{10}$ odds –4) (Fig. 3). After a positive result for a 'very strong confirmer' test (LR = 100, $\log_{10}$LR = 2), we would arrive at 1/100 ($\log_{10}$ odds –2). A sex worker might start at 10%, and arrive at 90%. In the case of the student, the marginal benefit of the HIV test is 0.99% in probability (from 0.0001 to 0.01), while in the case of the sex worker it is 80%. In our model, it is two classes in both cases, equal to the value of the LR of the test.

This logarithmic scale is also useful in threshold positioning. Our log model provides an extremely simple way to calculate the *maximum range* provided by the test's quality around the treatment threshold [1], as:

- log test/treatment threshold [2] = log treatment threshold PLUS log LR–
- log test threshold = log treatment threshold MINUS log LR+

Both formulas correspond mathematically to the classical ones, thus showing in a simple way the test's maximal theoretical utility, if the risk and cost are not considered.

Finally, a 'clinician-friendly' language has been tried out. Instead of LR+, we use 'confirming power'; instead of LR– 'excluding power', pre-test probability becomes 'clinical suspicion' and post-test probability becomes 'certainty level' or 'certainty'.

Evaluation of the model

Our model should be mathematically sound. The mathematical correspondence between the linear and the log model is obvious, but as we thought to be able to eliminate complex calculations through rounding, we assessed the rounding effect on the final post-test probability. We computed the effect on rounding values of logarithms of LRs to the unit or to half the unit for 2, 5 or 10 hypothetical disease characteristics of a hypothetical disease, in 50 cases with randomly computer generated LR's.

Further, our approach should be more acceptable to clinicians than the classical model. We presented a self-administered questionnaire to 50 medical doctors of different nationalities, attending clinical decision-making courses, with the following questions: (1) How do you estimate the power of arguments? (2) How do you estimate pre-test probability? (3) Is pre-test probability useful for decision making? Sixteen Belgian last-year medical students, to whom Bayes' theorem had been taught six times during their 7-year training, were asked two simple questions on predictive value. All data (sensitivity, specificity, pre-test probability) were provided.

In order to evaluate the acceptability of the new model, we subjected attendants to seminars to self-administered questionnaires. The first group consisted of 96 Belgian general practitioners, who participated in an 'appetiser' for medical decision making. The session took only 3 hours. With examples using our model, we explained the five crucial messages. Formal mathematical deduction from Bayes' theorem was stated, but not explained. We asked (1) whether this training helped in understanding medical decision making; (2) whether the model fits everyday practice; (3) whether this training is considered useful for daily work.

In three different 4-day training sessions, we subjected 45 doctors of various nationalities to the following questions: (1) Does this logarithmic representation of classes of probability, and of power of arguments seems more natural to you than the classical, Bayesian approach? (2) Did this model help you to understand the classical theory (Bayes' theorem)? (3) Did this model help you to add up a series of arguments? (4) Did this model help you to understand thresholds, or to estimate thresholds more correctly?

Results

Is this model mathematically sound?

The effect of rounding values of logarithms of LRs to the unit or to half the unit for 2, 5 or 10 disease characteristics on the final post-test probability is shown in Table 3 and in Fig. 4. As expected, the mean error (difference with the final outcome using absolute values) diminishes with increasing steps, and rounding to half the unit gives an absolute error of less than half a unit for 10 disease characteristics. Figure 5 shows the insignificant effect of rounding in the HIV example.

Table 3 Comparison of absolute and rounded values of LR for 50 cases

	Mean difference in post-test probability when using rounded value instead of absolute	Mean difference per step
Rounding to the unit		
2 steps	0.33 ± 0.06*	0.16
5 steps	0.56 ± 0.13	0.11
10 steps	0.74 ± 0.14	0.07
Rounding to half the unit		
2 steps	0.15 ± 0.03	0.07
5 steps	0.28 ± 0.05	0.05
10 steps	0.40 ± 0.06	0.04

*±0.xx indicates the 95% confidence interval. [15]
The table shows the difference in post-test probability, expressed in $\log_{10}$ odds, while rounding values of logarithms of likelihood ratios to the unit or to half the unit for 2, 5 or 10 disease characteristics compared with non-rounding of these values.
LR, likelihood ratio.

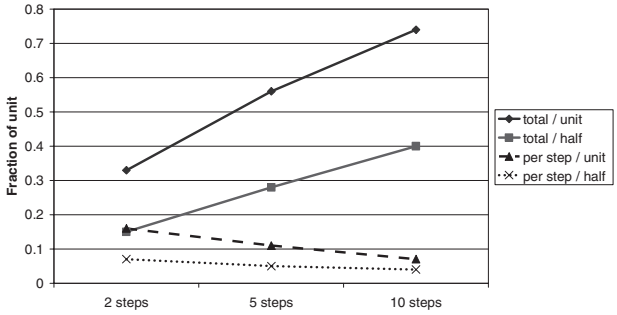

Figure 4 Effect of rounding of likelihood ratio (LR) values on post-test probability and on error per step. For 50 cases, we compute the sum of randomly generated LR. The Y-axis presents the mean error after computation with absolute versus rounded values of LR. The upper line gives the mean error of the final result when the LR is rounded to the unit, the second gives the same when rounded to half. The lower lines show the mean errors per step.

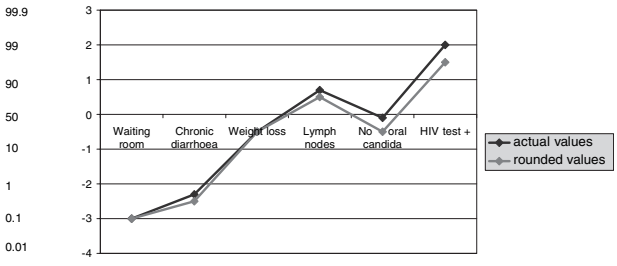

Figure 5 Comparison of actual versus rounded values of consecutive disease characteristics on the probability of AIDS in a given patient. A patient complains of diarrhoea of a 3-month duration and important weight loss; physical examination reveals enlarged lymph nodes in the cervical and axillar regions. Inspection of the mouth is negative. Finally an enzyme linked immune sorbent assay (ELISA) test is positive.

Use of classical clinical epidemiology parameters

Of 50 professionals, no one has even mentioned sensitivity, specificity, predictive value or LR when asked about the weight of disease characteristics: all clinicians worked with categories. For the estimation of prevalence, 48/50 (96%) declared to be able to estimate a prevalence: 25/48 (52%) express prevalence in nominal categories ('rare', 'extremely rare', etc.), 10/48 (20%) in percentages, and 13/48 (27%) use another numerical format (1/4, 1/1000, etc.). For 38/50 (76%), the disease prevalence is not estimated useful in decision making!

Of 16 last-year medical students, only 6/16 were capable of answering the questions on predictive value, after a total of six explanations of Bayes' theorem in the preceding years: of those answering, all had correct answers.

Subjective appreciation

In the appetiser session, 44/96 had already received some training in medical decision making. When confronted with the new model, 85/96 (88%) found medical decision making easier to understand, 94/96 (98%) recognized the theory in their practice, and 81/96 (84%) considered this training useful for daily work.

In the 4-day training sessions, the logarithmic representation seemed more natural than the classical Bayesian approach in 37/45 (82%), the model helped to understand the classical theory in 34/45 (75%), the model helped in summing a series of arguments in 42/45 (93%) and the model helped in understanding or estimating thresholds in 36/45 (80%) of participants.

Discussion

Bayesian calculations seem easy to most clinical epidemiologists, but tests with both freshly trained tropical doctors and with last-year medical students were in an astonishing contrast to that belief, and confirm earlier experiences of other authors [5,27–29]. Therefore, we present a new logical framework, based on a logarithmic scale both for disease probabilities and for LRs. We propose the use of discrete numerical classes for disease probability and LR and prove that the loss of accuracy is negligible. This logical framework leads to a diagram which allows Bayesian calculations in a user-friendly yet mathematically correct way. Finally, we propose a more clinician-friendly language for logical concepts. Pilot experiences suggest consistently that our model was easily understood by a wide variety of interviewees and considered useful in clinical problem-solving.

Model

Does representing clinical data on a *logarithmic scale* make logic more difficult? We should not forget that several phenomena in daily life are represented on a logarithmic scale: pitch, sound intensity, severity of earthquakes, power of wind, orders of magnitude from cosmology to fundamental physics etc. People can handle them, without realizing the underlying mathematics. In medicine also, many quantitative tests are logarithmic: titres of serologic tests, viral load in HIV infection. Perhaps the logarithmic scale for clinical epidemiology could be treated like any other logarithmic scale and it would be sufficient to explain the mathematics behind it only once (or even never, e.g. when teaching village health workers).

Why log_{10} and not natural logarithms? Even if natural logarithms possibly might come closer to 'real' orders of magnitude for test power, probability estimation intuitively goes by orders of 10, not by 'natural' order (1 in 10, 1 in 100, 1 in 1000). It is striking that some expert programmes rely also on a direct estimation of power in categories [30]. We realized that it is extremely difficult to formalize the intuitive feeling of the discriminative power of a test. We thought it was simple to deduce it from the effect of a LR on disease probability, but we realized that there is no agreement about the representation of a clinician's feeling of disease probability. Yet several authors suggest a logarithmic interpretation, some natural logarithms [20], others logarithms of 10 [17,18]. Turing considered both, but used the $\log_{10}$ format, moreover rounded [3].

Why *log_{10} of odds* instead of $\log_{10}$ of probabilities? Log_{10} odds have several advantages: $\log_{10}$ odds = 0 stands for the symmetry line of probability (50%) and the values which we find are close to those intuitively used by clinicians. Moreover, with $\log_{10}$ probability we would not adequately represent the very important post-test probability classes around 100%. However, the main advantage of odds over probability is the possibility to apply directly Bayes' theorem, summing pre-test odds and LRs.

Is this model mathematically sound?

It is not our aim to replace a *numerical* scale by an *ordinal* one, where 'often', 'rare', and the like, would receive a fixed value. Instead, we emphasize the relativity of actual numerical transformation, and we propose a *discrete numerical* scale, at least for didactical purposes. These classes can be sufficient for clinical diagnostic work in this didactic model. The question: 'How many times would you expect to encounter this diagnosis in a year, divided by the number of patient contacts per year?' (a proxy to prevalence) allows class estimation, without further mathematics. We discourage trainees to substitute these discrete numerical classes with 'nominal' probability classes as they are not generally standardized.

When we use discrete numerical classes for disease probability and LR, we loose very little accuracy, as is shown by the results of the computer simulation: when rounding to half the unit, the mean absolute error is still smaller than half a unit after 10 disease characteristics. A small error is obviously acceptable as clinicians work on rather weak data anyway. An error is clinically significant only if it would lead to a wrong decision. Of course, this model gives no solution for erroneous estimations of pre-test probability and LRs of tests, which would have the same deleterious effect in the classical model. It gives no solution either for the conditional interdependency of many tests. Students and clinicians should remain aware of this weakening of the power of a lot of arguments.

Utility

Few attempts to prove the clinical utility of clinical epidemiology training have been made, with contradicting results [31,32]. Moreover, there is more to diagnostic reasoning than clinical epidemiology. Kassirer describes four steps in diagnosis: hypothesis triggering, framing, information gathering and processing, and

verification [33]. The five crucial messages of diagnostic clinical epidemiology interfere only and partially with the third and the fourth step. Therefore, the evaluation of the utility should focus on the yield in the specific logical steps in the diagnostic process, as an effect on the whole probably will be difficult to demonstrate. We are convinced that clinical epidemiology (the five key messages) can be useful in clinical work, provided that this discipline is properly understood and easily applicable.

Further research

Several assumptions should be tested before this model should be applied widely. How accurately clinicians are able to estimate LRs, or power classes? Are intuitive classes of discriminative power parallel to $\log_{10}$LR? Initially, we thought it was simple to deduce the power of a characteristic from the effect of a LR on disease probability, but we realized that first we had to prove clinicians' intrinsic logarithmic perception of disease probability.

If this research leads to clear conclusions, we might test clinicians' accuracy to estimate post-test probability when given $\log_{10}$LR classes for a given disease, and analyse in real cases if, and how frequently, our simplified, discrete numerical approach to scales, tests and thresholds would lead to a wrong decision, if compared with classical, formal calculations.

Finally, the negligible effect of rounding, and the advantage of a log10 model over a natural log model is theoretical: it should be tested with clinicians on real cases.

Acknowledgements

We are indebted to Mattias Borchert, Arthur Elstein, Robert Hamm, R. D. Glass and Wim Van Damme for critical reading and/or invaluable suggestions.

References

1. Pauker, S. G. & Kassirer, J. P. (1975) Therapeutic decision making: a cost–benefit analysis. *New England Journal of Medicine*, 293, 229–234.
2. Pauker, S. G. & Kassirer, J. P. (1980) The threshold approach to clinical decision making. *New England Journal of Medicine*, 302, 1109–1117.
3. Good, I. J. (1979) AM Turing's statistical work in world war II. (Studies in the history of probability and statistics XXXVII). *Biometrika*, 66, 393–396.
4. Kassirer, J. P. (1995) Teaching problem-solving – how are we doing? [editorial; comment] [see comments]. *New England Journal of Medicine*, 332, 1507–1509.
5. Gigerenzer, G. (1996) The psychology of good judgment: frequency formats and simple algorithms [see comments]. *Medical Decision Making*, 16, 273–280.
6. Kassirer, J. P. (1989) Teaching clinical problem solving. *Acta Clinica Belgica*, 44, 291–294.
7. Van Puymbroeck, H., Remmen, R., Denekens, J., *et al.* (2003) Teaching problem solving and decision making in undergraduate medical education: an instructional strategy. *Medical Teacher*, 25, 547–550.
8. Leaper, D. J., Gill, P. W., Staniland, J. R., Horrocks, J. C. & de-Dombal, F. T. (1973) Clinical diagnostic process: an analysis. *British Medical Journal*, 3, 569–574.
9. Bryant, G. D. & Norman, G. R. (1980) Expressions of probability: words and numbers [letter]. *New England Journal of Medicine*, 302, 411.
10. Kenney, R. M. (1981) Between never and always [letter]. *New England Journal of Medicine*, 305, 1097–1098.
11. Van den Ende, J. & Van Gompel, A. (1996) Clinical problem-solving: if at first you don't succeed [letter]. *New England Journal of Medicine*, 334, 919.
12. Van den Ende, J., Moreira, J., Basinga, P. & Bisoffi, Z. (2005) The trouble with likelihood ratios. *Lancet*, 366, 548.
13. Szilard, L. (1929) Uber die Entropieverminderung eines thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitscrift für Physik*, 53, 850–856.
14. Johnson, H. A. (1991) Diminishing returns on the road to diagnostic certainty [see comments]. *Journal of the American Medical Association*, 265, 2229–2231.
15. 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, 29–36.
16. Glass, R. D. (1996) Diagnosis, A Brief Introduction. Melbourne, Australia: Oxford University Press.
17. 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, 703–707.
18. Goodman, S. N. (1999) Toward evidence-based medical statistics. 2: The Bayes factor [see comments]. *Annals of Internal Medicine*, 130, 1005–1013.
19. Zhou, X.-H., Obuchowski, N. & McClish, D. (2002) Statistical Methods in Diagnostic Medicine. New York: John Wiley & Sons.
20. Spiegelhalter, D. J. (1985) Statistical methodology for evaluating gastrointestinal symptoms. *Clinical Gastroenterology*, 14, 489–515.
21. Fagan, T. J. (1975) Letter: nomogram for Bayes theorem. *New England Journal of Medicine*, 293, 257.
22. Hoffrage, U. & Gigerenzer, G. (1998) Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, 538–540.
23. Sackett, D. L., Haynes, R. B., Guyatt, G. & Tugwell, P. (1991) Clinical Epidemiology: A Basic Science for Clinical Medicine, 2nd Edn. Boston, MA: Little, Brown and Company.
24. Kopelman, R. I., Wong, J. B. & Pauker, S. G. (1999) Clinical problem-solving. A little math helps the medicine go down. *New England Journal of Medicine*, 341, 435–439.
25. Shermer, M. (2005) The Feynman–Tufte principle. *Scientific American*, 292, 21.
26. 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 Education*, 31, 202–209.
27. Weiss, S. T. & Samet, J. M. (1980) An assessment of physician knowledge of epidemiology and biostatistics. *Journal of Medical Education*, 55, 692–697.
28. Wulff, H. R., Andersen, B., Brandenhoff, P. & Guttler, F. (1987) What do doctors know about statistics? *Statistics in Medicine*, 6, 3–10.
29. Reid, M. C., Lane, D. A. & Feinstein, A. R. (1998) Academic calculations versus clinical judgments: practicing physicians' use of quantitative measures of test accuracy [see comments]. *The American Journal of Medicine*, 104, 374–380.
30. Miller, R. A., Pople, H. E. J. & Myers, J. D. (1982) Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307, 468–476.
31. Round, A. P. (1999) Teaching clinical reasoning – a preliminary controlled study. *Medical Education*, 33, 480–483.
32. Rogers, J. C., Swee, D. E. & Ullian, J. A. (1991) Teaching medical decision making and students' clinical problem solving skills. *Medical Teacher*, 13, 157–164.
33. Kassirer, J. P. & Kopelman, R. I. (1989) Cognitive errors in diagnosis: instantiation, classification, and consequences. *The American Journal of Medicine*, 86, 433–441.