Working Papers on The Nature of Evidence:
How Well Do ‘Facts’ Travel?
No. 16/07

Battle in the Planning Office: Biased Experts versus Normative Statisticians

Marcel Boumans

Department of Economics
University of Amsterdam

May 2007

how
well do
'facts'
travel?

“The Nature of Evidence: How Well Do ‘Facts’ Travel?” is funded by The Leverhulme Trust and the ESRC at the Department of Economic History, London School of Economics.

For further details about this project and additional copies of this, and other papers in the series, go to:

http://www.lse.ac.uk/collection/economichistory/

Series Editor:

Dr. Jonathan Adams
Department of Economic History
London School of Economics
Houghton Street
London, WC2A 2AE

Tel: +44 (0) 20 7955 6727
Fax: +44 (0) 20 7955 7730

Battle in the Planning Office: Biased experts versus normative statisticians[1]

Marcel Boumans

Abstract

For the purposes of calculation, context is irrelevant: one is expected to strip away the "contingent" details, slot bare numbers into the equations, and perform the relevant maths. Medical doctors must know this. So why, asked just such a question about the likelihood of a diagnostic test being accurate, do two thirds of respondents get the answer wrong? These results are usually used to demonstrate the medics' woeful comprehension of probability theory. This paper, however, argues that the results can be understood as a reminder of the importance of context to the constitution of "rationality." Reinterpreting the results in light of "ecological rationality" – which takes account of context – reveals that the problem may not be with the respondents, but with the conception of rationality as necessarily context independent. "Facts" are statements about the world for which there is consensus, and consensus will be achieved when a statement can be accepted on rational arguments. But what kind of arguments can be considered as rational?[2]

1. Introduction

Most planners consider maximising free choice to be consistent with economic efficiency and, thus, the most effective means of promoting or enhancing social welfare. This link between the augmentation of choices and increase of social welfare is based on the assumption that decisions are

[1] This paper is presented at the 'Biased experts versus plain facts' session of the Annual Meeting of the Society for Social Studies of Science (November 2006), Vancouver, Canada and at the International congress 'The Social Sciences and Democracy: A philosophy of science perspective (September 2006), Ghent University, Belgium. I am grateful for the comments and encouragements received from participants at both sessions.

[2] This abstract has been composed by the series editor, and is not the work of the paper's author.

made rationally. However, new behavioural economists and psychologists, notably 2002 Nobel prize laureate Daniel Kahneman, have shown over the past three decades that people, including experts like physicians, do not exhibit rational expectations, fail to make judgements that are consistent with Bayes' rule, use heuristics that lead them to make systematic blunders, exhibit preference reversals, make different choices depending on the wording of the problem, and suffer from problems of self-control. As a result of these findings, these economists and psychologists recommend what they call 'libertarian paternalism', that is an approach that preserves freedom of choice but that authorizes both private and public institutions to steer people in directions that will promote their welfare (Thaler and Sunstein 2003). For this steering a new type of expert is called for, namely the 'normative statistician', the expert in rational reasoning with uncertainty.

As a result, we face a battle for the position of the planner's counsellor between two kinds of expert: 1) The field expert with skilled knowledge of a specific field, inclusive knowledge about usage and application of the appropriate instruments. 2) The 'normative statistician' with skilled knowledge of statistical reasoning. This battle is in fact a confrontation between two kinds of rationality. To explore this encounter, the paper will compare both kinds of expertise within the context of decision making in medical practice. The starting point is a classic example of a so-called 'base rate fallacy': the Harvard Medical School Test (presented below, in section 2). It appeared that, when a laboratory test result is given, physicians do not take account of the base rate, or pre-test probability, to reach a clinical decision.

A base rate fallacy is considered to be a bias, in the sense of a violation of the axioms of probability and/or a misperception of probabilities. Biases (discussed in section 3) are errors that anyone would want to correct if the matter were brought to his/her attention. A lot of experiments have shown that ‘reasoning with uncertainty’ is tough – even to experts – and that training can be worthwhile. This ‘normative statistical’ perspective on scientific reasoning is compared with another expert perspective on rational decision-making, the so-called Evidence-Based Medicine approach. In this approach (which will be extensively discussed in section 4), a test is only meaningful when the evidence is not clear yet, and is recommended not to apply in extreme cases, as was actually the case in the above Harvard Medical School Test. From this perspective, clinical judgments are unbiased when tests are used appropriately.

Facts are statements about the world for which there is scientific consensus. Consensus will be achieved when a statement can be accepted on rational arguments. The problem being studied in this paper is what kind of arguments can be considered as rational. Rationality is here roughly defined as correctly applying the rules of logic and those of the probabilistic calculus. It will appear that decision processes in both kinds of expertises are rational and so neither is biased in that sense. If, however, one would compare both approaches with a criterion of biasedness as defined in classical statistics (which is what is done here, in section 5), it appears that *both* approaches are biased. The distinguishing criterion between both expertises is not their rationality but the way they take the environment into account, or in other words, how they have modelled the context in which the decisions have to be taken. In section 6, two different positions will be compared. One is that a decision, inference, or conclusion is rational when arrived at by correct reasoning insusceptible for any context. The other position, ‘ecological rationality’, is characterized by correct reasoning where one is highly susceptible for the environment in which one takes a decision. Both positions will be discussed for the case of ‘rational clinical decision making', where one has to decide to ask for a test and subsequently has to interpret its possible outcomes.

2. Interpretations by physicians of clinical laboratory results

The so-called Harvard Medical School Test, carried out by Casscells, Schoenberger and Graboys (1978), was a small survey to obtain some idea of how physicians interpret a laboratory result.

> We asked 20 house officers, 20 fourth-year medical students and 20 attending physicians, selected in 67 consecutive hallway encounters at four Harvard Medical School teaching hospitals, the following question: "If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs?" (Casscells, Schoenberger and Graboys 1978: 999)

Using Bayes' theorem, the 'correct' answer should be: 2 percent.[3] The result of this test was that only 11 of 60 participants gave this answer. The most common answer, given by 27, was 95 percent. The average of all answers was 55.9 percent, 'a 30-fold overestimation of disease likelihood' (p. 1000).

Discussing these results, Casscells, Schoenberger and Graboys observe that, despite probabilistic reasoning has been presented in prominent clinical journals for a decade, 'in this group of students and physicians, formal decision analysis was almost entirely unknown and even commonsense reasoning about the interpretation of laboratory data was

[3] $\Pr(P\,|\,+) = \dfrac{\Pr(+\,|\,P)\Pr(P)}{\Pr(+\,|\,P)\Pr(P) + \Pr(+\,|\,A)\Pr(A)} \approx \dfrac{0.001}{0.001 + 0.05 \cdot 0.999} = 0.02$, where *P*: disease is present, *A*: disease is absent, and +: positive test result.

uncommon’ (p. 1000). This problem, however, was considered to be remediable by practical instruction in the theory of test interpretation.

Four years later a similar result was published by David Eddy (1982). He discusses a more specific case of deciding whether to perform a biopsy on a woman who has a breast mass that might be malignant. Specifically, he studied how physicians process information about the results of a mammogram, an X-ray test used to diagnose breast cancer.

The prior probability, $\Pr(ca)$, ‘the physician’s subjective probability’, that the breast mass is malignant is assumed to be 1 percent. To decide whether to perform a biopsy or not, the physician orders a mammogram and receives a report that in the radiologist’s opinion the lesion is malignant. This is new information and the actions taken will depend on the physician’s new estimate of the probability that the patient has cancer. This estimate also depends on what the physician will find about the accuracy of mammography. This accuracy is expressed by two figures: sensitivity, or true-positive rate $\Pr(+ \mid ca)$, and specificity, or true-negative rate $\Pr(- \mid benign)$. They are respectively 79.2% and 90.4%. Applying Bayes’ theorem leads to the following estimate of the posterior probability:

$$\Pr(ca \mid +) = \frac{\Pr(+ \mid ca)\Pr(ca)}{\Pr(+ \mid ca)\Pr(ca) + \Pr(+ \mid benign)\Pr(benign)} = \frac{0.792 \cdot 0.01}{0.792 \cdot 0.01 + 0.096 \cdot 0.99} = 7.7\%$$

In an informal sample taken by Eddy, most physicians (approximately 95 out of 100) estimated the posterior probability to be about 75%.

When Eddy asked the ‘erring’ physicians about this, they answered that they assumed that the probability of cancer given that the patient has a positive X-ray, $\Pr(ca \mid +)$, was approximately equal to the probability of a positive X-ray in a patient with cancer, $\Pr(+ \mid ca)$.

> The latter probability is the one measured in clinical research programs and is very familiar, but it is the former probability that is needed for clinical decision making. It seems that many if not most physicians confuse the two. (Eddy 1982: 254)

According to Eddy, it is not only the physicians who are erring, but a review of the medical literature on mammography reveals a ‘strong tendency’ to equate both probabilities, that is, to equate $\Pr(ca \mid +) = \Pr(+ \mid ca)$. Generally, erroneous probabilistic reasoning is widespread among practitioners, and according to Eddy, focusing on improving this kind of reasoning will have an important impact on the quality of medical care:

> The probabilistic tools discussed in this chapter have been available for centuries. In the last two decades they have been applied increasingly to medical problems […], and the use of systematic methods for managing uncertainty has been growing in medical school curricula, journal articles, and postgraduate education programs. At present, however, the application of these techniques has been sporadic and has not yet filtered down to affect the thinking of most practitioners. As illustrated in this case study, medical problems are complex, and the power of formal probabilistic reasoning provides great opportunities for improving the quality and effectiveness of medical care. (Eddy 1982: 267)

3. Heuristics and biases

Eddy’s article was published in *Judgment under Uncertainty: Heuristics and Biases* (1982), edited by Daniel Kahneman, Paul Slovic and Amos Tversky. The first chapter of this volume is a reprint of an article by Tversky and Kahneman (1974), with the same title as the book, published in *Science*, eight years earlier.

Tversky and Kahneman (1974, 1982) explain that to assess the probability of an uncertain event, people rely on a limited number of heuristic principles that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations. However, though these 'heuristics' are useful, they sometimes lead to 'severe and systematic errors', also called 'biases'. They describe three heuristics with accompanying biases: 'representativeness', 'availability', and 'adjustment and anchoring'. Only the first heuristics is of relevance here, because this is the one that leads to the bias discussed above: 'insensitivity to prior probability of outcomes'.

One type of probabilistic questions are questions like: 'What is the probability that event *B* will generate event *A*?' In answering this question, Tversky and Kahenman (1974, 1982) claim that people rely on the representativeness heuristics, in which probabilities are evaluated by the degree to which *A* is representative of *B*, that is, by the degree to which *A* resembles *B*. One of the biases that go along with this heuristic is the 'base-rate fallacy': the neglect of prior probabilities.

When discussing the different heuristics and biases, Tversky and Kahneman (1974, 1982) emphasize that reliance on heuristics and prevalence of biases are not restricted to laymen and 'naive subjects', but when they think 'intuitively', experienced researchers are prone to the same biases. 'Statistical principles are not learned from everyday experience because the relevant instances are not coded appropriately' (1974: 1130, 1982: 18). This lack of an appropriate code also explains why people do not detect the biases in their judgments of probability, when no one brings this to their attention.

4. Evidence-Based Medicine

An increasingly influential movement to rationalize clinical examination is the Evidence-Based Medicine (EBM) approach, which appeared in the early 1990s. This approach was developed by the so-called Evidence-Based Medicine Working Group[4], chaired by Gordon Guyatt (EBM 1992), and made public by an editorial of the *ACP Journal Club*, a year earlier (Guyatt 1991). The primary purpose of *ACP* (*American College of Physicians*) *Journal Club*, which originally appeared as a supplement to the *Annals of Internal Medicine*, was 'to help make evidence-based medicine more feasible' (Guyatt 1991: A-16). EBM was presented as a 'new paradigm for medical practice':

> Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. Evidence-based medicine requires new skills of the physician, including efficient literature searching and the application of formal rules of evidence evaluating the clinical literature. (EBM 1992: 2420)

This approach resulted in a pocketbook (Sackett et al. 2000, first published in 1997) with a CD and coloured cards in the cover pocket and a book's website http://hiru.mcmaster.ca/ebm.htm, in which EBM is defined as 'the integration of best research evidence with clinical expertise and patient

[4] This group consisted of the following members: P. Brill-Edwards, J. Cairns, D. Churchill, D. Cook, A. Detsky, M. Enkin, P. Frid, M. Gerrity, H. Gerstein, J. Gibson, B. Haynes. J. Hirsch. J. Irvine, R. Jaeschke, A. Kerigan, A. Laupacis, V. Lawrence, Mark Levine, Mitchell Levine, J. Menard, V. Moyer, C. Mulrow, P. Links, A. Neville, J. Nishikawa, A. Oxman, A. Panju, D. Sackett, J. Sinclair, and P. Tugwell.

values' (p. 1).[5] The book describes the practice of EBM in five steps (Sackett et al. 2000: 3-4):

- Step 1 – converting the need for information (about prevention, diagnosis, prognosis, therapy, causation, etc.) into an answerable question.
- Step 2 – tracking down the best evidence with which to answer that question.
- Step 3 – critically appraising that evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in our clinical practice).
- Step 4 – integrating the critical appraisal with our clinical expertise and with out patient's unique biology, values and circumstances.
- Step 5 – evaluating our effectiveness and efficiency in executing steps 1-4 and seeking ways to improve them both for next time.

The use of test results are part of Step 3, which is dealt with in chapters 3-7, and of which chapter 3 'Diagnosis and Screening' is of direct relevance here. The main part of this chapter has been written to help one answering three questions about diagnostic testing, which can also be found on the yellow-ochre card 2A:

[5] A third edition by S.E. Straus, W.S. Richardson, P. Glasziou, and R.B. Haynes was published in 2005. The 2nd edition is however used for this paper.

1. Is this evidence about the accuracy of a diagnostic test valid?
2. Does this (valid) evidence demonstrate an important ability of this test to accurately distinguish patients who do and don't have a specific disorder?
3. Can I apply this valid, important diagnostic test to a specific patient?

The third question is subsequently split up in another set of questions:

I. Is the diagnostic test available, affordable, accurate, and precise in our setting?

II. Can we generate a clinically sensible estimate of our patient's pre-test probability?
   - Are the study patients similar to our own?
   - Is it unlikely that the disease possibilities or probabilities have changed since this evidence was gathered?

III. Will the resulting post-test probabilities affect our management and help our patient?
   - Could it move us across a test-treatment threshold?
   - Would our patient be a willing partner in carrying it out?

In clinical practice, physicians are faced with three choices: to withhold therapy, to order a diagnostic test, or to treat without testing. Therefore they must take into account the reliability, value and risks of both testing and treatment to maximize both diagnostic accuracy and cost effectiveness (Scherokman 1997).

An ideal test should distinguish absolutely between patients who do and who do not have disease. The clinical usefulness of a test is determined by how much it deviates from this ideal. Data on test characteristics are derived from studying the test against a ‘golden standard test’, the test that definitively determines the presence or absence of disease. An example of a ‘golden standard test’ would be biopsy. Patients whom biopsy has shown to have the disease and patients shown not to have the disease are given the diagnostic test in question. To review the accuracy of the test, the results of biopsy and diagnostic test are presented in a two-by-two table (see table 1).

		Target disorder	
		present *P*	absent *A*
Diagnostic test result	positive +	*a*	*b*
	negative –	*c*	*d*

Table 1: Systematic review of a diagnostic test

Two characteristics define the accuracy of a test:

- ‘Sensitivity’ describes the ability of a test to correctly detect disease, $\Pr(+ \mid P) = a/(a + c)$.
- ‘Specificity’ describes the ability of a test to correctly identify absence of disease, $\Pr(- \mid A) = d/(b + d)$.

Sensitivity and specificity are considered to be stable properties of a test. They do not vary with pre-test probability of disease, also called base rate or prevalence, $\Pr(P)$. In contrast with these test characteristics, the predictive value is not a stable property and varies with the pre-test probability:

- ‘Positive predictive value’: $\Pr(P\mid +)=\dfrac{\Pr(+\mid P)}{\Pr(+)}\Pr(P)=\dfrac{a}{a+b}$

- ‘Negative predictive value’: $\Pr(A\mid -)=\dfrac{\Pr(-\mid A)}{\Pr(-)}\Pr(A)=\dfrac{d}{c+d}$

Instead of using these ‘old-fashioned’ concepts of sensitivity and specificity, EBM recommends to use the ‘new-fangled and more powerful’ concepts of likelihood ratios to represent the accuracy of a test (Sackett et al.: 72). When dealing with more then one test results, it is easier to use for calculating the post-test probabilities.

- ‘Likelihood ratio for positive test result’: $LR(+) = \Pr(+ \mid P)/\Pr(+ \mid A)$
- ‘Likelihood ratio for negative test result’: $LR(–) = \Pr(– \mid P)/\Pr(– \mid A)$

In general, the likelihood ratio is: $LR(X) = \Pr(X \mid P)/\Pr(X \mid A)$, where X is the random variable indicating a test result and taking values + or –. Then the interpretation of diagnostic test runs as follows:

- Pre-test odds $= \Pr(P)/\Pr(A)$
- Post-test odds = likelihood ratio × pre-test odds $= LR(X) \times \Pr(P)/\Pr(A)$

Information about post-test probabilities (if required) can easily be inferred from information about post-test odds: post-test probability = post-test odds / (post-test odds + 1).

Tests can be painful and/or risky, so a clinician only asks for a test after a well-considered evaluation of reliability, value and risk. The model for making this rational decision in (Sackett et al. 2000) is based on Pauker and

Kassirer (1980). This article describes a model that uses two thresholds to aid physicians in making clinical decisions:

1) a ‘no treatment/test’ threshold, T_t, which is the disease probability at which the expected utility of withholding treatment is the same as that of performing a test;
2) a ‘test/treatment’ threshold, T_{trx}, which is the disease probability at which the expected utility of performing is the same as that of administering treatment.

The decision not to treat, to test, or to treat is determined by pre-test disease probability and both thresholds, see figure 1. The best clinical decision for probabilities below the ‘no treatment/test’ threshold T_t is to refrain from treatment; for probabilities above the ‘test/treatment’ threshold T_{trx}, the best decision is to administer treatment. When the pre-test disease probability lies between the thresholds, the test result could change the probability of the disease enough to alter the decision, so the best decision would be to administer a test.

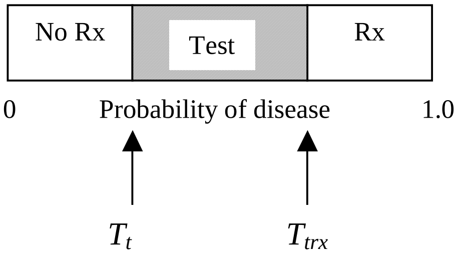

Figure 1 Test-treatment thresholds

Adapted from Pauker and Kassirer 1980: 1111

So, for clinical decision-making, estimates of disease prevalences are crucial. It is noteworthy to see that Pauker and Kassirer, though referring to (Tversky and Kahneman 1974), make, unlike Tversky and Kahneman (1974), a distinction between expert opinions and of those outside the medical domain: ‘Studies in non-medical domains show that people have biases and often make inaccurate estimates and that training improves the reliability of such estimates’ (Pauker and Kassirer 1980: 1112). The sort of studies Tversky and Kahneman are referring to rely, according to Pauker and Kassirer, on ‘simple tests in which an actual probability is known (for example, the number of various coloured balls in an urn)’, whereas in medicine a prevalence ‘represents a belief or opinion for which no actual or true value exists’ (p. 1112). Moreover, it appears that physicians make probability estimates with ‘reasonable reliability’ (p. 1112). When published data on probabilities are not specific enough, the ‘opinions of experts’ are needed and used.

In the EBM approach one will find the same position towards expert opinions, that is, needed where data on probabilities are not available, but with caution:

> Clinical experience and the development of clinical instincts (particularly with respect to diagnosis) are a crucial and necessary part of becoming a competent physician. Many aspects of clinical practice cannot, or will not, ever be adequately tested. Clinical experience and its lessons are particularly important in these situations. At the same time, systematic attempts to record observations in a reproducible and unbiased fashion markedly increase the confidence one can have in knowledge about patient prognosis, the value of diagnostic tests, and the efficacy of treatment. In the absence of systematic observations one must be cautious in the interpretation of information derived from clinical experience and intuition, for it may at times be misleading. (EBM 1992: 2421)

From the above-described threshold model test criteria can be inferred. First, according to Scherokman (1997), tests that do not change the probability of disease enough to cross the threshold probability T_{trx} are not useful and should not be ordered. This means that when the pre-test disease probability lies between the thresholds and we have a positive test result, the post-test disease probability should lie above the test/treatment probability:

$$\Pr(P \mid +) > T_{trx}.$$

This is in fact a weak criterion, because it implies that the disease should (causally) influence the test result:

$$\Pr(+ \mid P) > \Pr(+).^{6} \qquad (1)$$

In statistics, an event A is independent of event B if $\Pr(A \mid B) = \Pr(A)$. In probabilistic accounts of causality, it is crudely stated that B causes A if $\Pr(A \mid B) > \Pr(A)$. So, the above requirement (1) obviously excludes tests like flipping a coin.

A stronger test requirement is that it should be 'most informative'. A test is most informative when its predictive values, $\Pr(P \mid +)$ and $\Pr(A \mid -)$, are optimal. As is said above, these values depend on the pre-test probabilities. It can be shown that both predictive values are optimal when:[7]

[6] If $\Pr(P) < T_{trx}$, and $\frac{\Pr(+ \mid P)}{\Pr(+)}\Pr(P) > T_{trx} \Rightarrow \frac{\Pr(+ \mid P)}{\Pr(+)} > \frac{T_{trx}}{\Pr(P)} > 1$.

[7] This can be seen by maximizing $\Pr(P \mid +)\cdot\Pr(A \mid -)$ for $\Pr(P)$.

$$\Pr(P) = \frac{1}{\sqrt{LR(+)LR(-)}+1}.$$

Usually the test characteristics sensitivity and specificity are about equal, which means that the optimal pre-test probability is about 50%.[8] Generally, it expected that a test is most informative when the pre-test probability of disease is between 40% and 60% (Scherokman 1997).

These demands on tests with respect to accuracy and applicability give new light on the interpretation by physicians of clinical laboratory results. First, assume that condition for using the test is optimal: $\Pr(P) \approx 0.5$, so $\Pr(A) = 1 - \Pr(P) \approx 0.5$. When sensitivity and specificity are about equal, then

$$\Pr(+) = \Pr(+ \mid P)\Pr(P) + \Pr(+ \mid A)\Pr(A) \approx \Pr(+ \mid P)0.5 + \Pr(- \mid P)0.5 = 0.5$$

So, if physicians assume that a test is used for optimal conditions, there is no question of base rate fallacy, because:

$$\Pr(P \mid +) = \frac{\Pr(+ \mid P)}{\Pr(+)}\Pr(P) \approx \Pr(+ \mid P)$$

Secondly, let us take Eddy's figures: $\Pr(+ \mid P) = 79.2\%$ and $\Pr(- \mid A) = 90.4\%$, and assume that $40\% < \Pr(P) < 60\%$, then $37.44\% < \Pr(+) < 51.36\%$, and so

$$84.6\% < \Pr(P \mid +) < 92.5\%.$$

Most physicians estimated the post-test probability to be about 75%.

[8] When $\Pr(+ \mid P) \approx \Pr(- \mid A)$, then also $\Pr(- \mid P) \approx \Pr(+ \mid A)$, and thus $LR(+)LR(-) \approx 1$.

And finally, the Harvard Medical School Test figure, $\Pr(+ \mid A) = 5\%$, leads even to higher post-test probabilities, when the prevalence is between 40% and 60%:

$$93\% < \Pr(P \mid +) < 95\%.$$

Recall that most common answer, given by 27 of 60, was 95 per cent.

Physicians are trained not to ask for diagnostic tests when prevalences are too small (or too large). Faced with test results they might have assumed automatically that the test was performed for the right conditions. So, they might have developed a heuristic to read the sensitivity and specificity as predictive values. Seen from this perspective, the physician's high estimates of the post-test probabilities in the case of the Harvard Medical School Test and in Eddy's test are not biased, but show 'ecological rationality'. (This type of rationality takes account of the environment, and will be discussed in section 6.)

5. Statistical bias

In the literature discussed above, it is assumed that Bayesian reasoning is an unbiased heuristic. In mathematical statistics, however, unbiasedness has a very specific meaning: An estimator, $\hat{\theta}$, is unbiased if and only if it's expected value is equal to the parametric value, θ, it is intended to estimate: $E[\hat{\theta}] = \theta$. A consequence of this specific definition is that an estimator based on Bayesian reasoning is not automatically unbiased. In a widely used standard textbook on statistics *Introduction to the*

Theory of Statistics[9], one will find the following remarkable observation: 'in general a posterior Bayes estimator is not unbiased' (p. 343). A 'posterior Bayes estimator' is defined as $E[Y \mid X]$, where X is a random variable with probability $\Pr(X \mid Y = y)$, and Y a random variable with probability $\Pr(Y)$. A posterior Bayes estimator is an 'unbiased' estimator of *y* when $E[E[Y \mid X] \mid y] = y$. It is shown that a posterior Bayes estimator is unbiased only when this estimator correctly estimates *y* with probability one. In all other cases the estimator is not unbiased. So, in an early training in statistics, one is already warned that Bayesian tools and unbiasedness might be incompatible.

Being warned, let us check whether the post-test probability, that is the probability taking account of test results, $\Pr(P \mid X)$, is an unbiased estimator of the pre-test probability, $\Pr(P)$. Let X be the random variable indicating the test result, taking value + or –.

$$E[\Pr(P \mid X)] = E\left[\frac{\Pr(X \mid P)}{\Pr(X)}\right]\Pr(P) = \left[\frac{\Pr(+ \mid P)}{\Pr(+)}\Pr(+) + \frac{\Pr(- \mid P)}{\Pr(-)}\Pr(-)\right]\Pr(P) = \Pr(P)$$

So it seems that our worry was unnecessary. Unfortunately, this is not the case. Generally, in rational decision-making (including in EBM), it is highly recommended to use likelihood ratios to estimate the disease odds. When discussing the use of likelihood ratios, Roger Cooke (1991) gives an expression how one can 'learn' from observations (adapted from his theorem 6.3, p. 97):

$$E[LR(X) \mid P] \geq 1, \text{ and equality holds if and only if } \Pr(LR(X) = 1 \mid P) = 1$$

[9] First edition by Alexander Mood was published in 1950. The 2nd edition coauthored by Franklin Graybill appeared in 1963, and the 3rd edition with Duane Boes as third author was published in 1974.

The equality condition can hold only if $\Pr(X \mid P) = \Pr(X \mid A) = \Pr(X)$. A test that would have this latter characteristic is not informative because it is then independent of disease, and should therefore be excluded, see equation (1).

However, this theorem shows only that one can learn from a test in case the disease is present. It surprisingly happens to be that in case of an absent disease, a test will not ‘learn’ us about the absence of this disease:

$$E[LR(X) \mid A] = \frac{\Pr(+ \mid P)}{\Pr(+ \mid A)} \Pr(+ \mid A) + \frac{\Pr(- \mid P)}{\Pr(- \mid A)} \Pr(- \mid A) = 1$$

This result makes the test biased

$$E[LR(X)] = E[LR(X \mid P] \cdot \Pr(P) + E[LR(X) \mid A] \cdot \Pr(A) > 1$$

So, it appears to be the case that post-test odds are not unbiased estimators for the pre-test odds:

$$E\left[\frac{\Pr(P \mid X)}{\Pr(A \mid X)}\right] = E[LR(X)] \frac{\Pr(P)}{\Pr(A)} > \frac{\Pr(P)}{\Pr(A)}$$

The undesired result of this bias is that each time a test result is being taken account of (whatever the result is, positive or negative) the expected disease odds will increase.

This may have the curious effect of going to a hospital because you feel some vague disorder, the physicist tries to find out what is the case with you by asking for one or more tests, and the result will be that both s/he and you have higher expectations of having a disease, and so you will feel more miserable when leaving the hospital.

6. Ecological rationality

An important critic of Kahneman and Tversky's normative statistical approach is Gerd Gigerenzer:

> If you open a book on judgment and decision making, chances are that you will stumble over the following moral: Good reasoning must adhere to the laws of logic, the calculus of probability, or the maximization of expected utility; if not, there must be a cognitive or motivational flaw. Don't be taken in by this fable. (Gigerenzer 2004: 62)

Gigerenzer describes Kahneman and Tversky's approach as a study of cognitive illusions: its primary aim seems to be is to demonstrate that people's judgments do not actually follow the laws of probability or the maximization of expected utility. 'The result is a list of deviations from norms, which are interpreted as cognitive fallacies, emphasizing irrationality rather than rationality' (p. 65).

In Gigerenzer's account of heuristics, the rationality of heuristics is not logical, but ecological. Ecological rationality implies that a heuristic is not good or bad, rational or irrational per se, only relative to an environment. Gigerenzer (2004) mentions twelve examples of phenomena that were interpreted as 'cognitive illusions' but which he re-evaluated as 'reasonable judgments given the environmental structure' (p. 66).

In fact, the interpretation of a test result by physicians can be seen as another example of ecological rationality. The (fast and frugal) heuristic is to read the sensitivity of a test as the predictive value when the test result is positive. This is reasonable in an environment of Evidence Based Medicine practice where test results are only asked for when prevalence's are not decisive yet, and tests are most informative.

Not taking the environment into account can lead to all kind of so-called ‘paradoxes’ in statistics. These paradoxes are used to show that when making judgments regarding the likelihood of uncertain events, even mathematically sophisticated people do not follow the principles of probability theory. According to Kahneman et al. (1982), ‘this conclusion is hardly surprising because many of the laws of chance are neither intuitively apparent, nor easy to apply’ (p. 32). A famous example is the Monty-Hall problem.[10] Discussing this problem in the *American Statistician*, Morgan et al. (1991) ended their conclusions with the following question: ‘“How do you expect me to solve a problem that stumped scores of Ph.D.’s [sic] and confused the world’s most intelligent person?”!’ (p. 287). In his Comment, Seymann (1991) separated this question into two (in his view) distinct issues. The first is concerned with clarity of problem definition, and the second is concerned with why ‘sensible and mathematically well-trained people, given that they agree on what the problem is, still get the wrong solution’ (p. 287). To address the latter issue, Seymann gives a few examples (Bertrand’s Box Problem, Birthday Problem) well known in statistics, but he refers also to Kahneman, Slovik, and Tversky’s edited volume *Judgement Under Uncertainty* (1982), in particular to the Harvard Medical School Test. Interestingly, it is this example of the Harvard Medical School Test which provoked others to respond to Seymann’s commentary. The first comment is by John P. Wendell, and is worth quoting at length:

[10] This problem raised a good deal of commotion, even among mathematicians, when discussed by vos Savant (1990). She phrased the problem as follows: ‘Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, “Do you want to pick door #2?” Is it to your advantage to switch your choice of doors?’ (p. 13).

> This answer of 2% apparently assumes that everyone in the population, whether they have the disease or not, has an equally likely chance of receiving the test and that the false negative rate is zero […]. Neither of these assumptions is stated or clearly implied in the problem. Stating “you know nothing about the person’s symptoms or signs” is not the same as stating that the test has an equal chance of being administered to people in the population, even if that was the intent of the phrase. The medical students and staff that were given this question would know full well that patients having a disease are almost more likely to have a test for their disease administered to them than the general public […]. The majority response of 95% is consistent with the assumption that persons having the test applied to them have a 50% chance of actually having the disease […]. Certainly these assumptions are more reasonable than those needed to support the 2% answer. Perhaps this illustration shows not that medically trained people don’t understand probability but that some statisticians don’t understand medicine. (Wendell 1992, 242)

In a subsequent response, Seymann (1992) stated that to ‘know nothing about the person’s symptoms or signs’ is not an instruction to assume random testing, but a ‘a clear instruction to disregard all other information, biases, or prejudices we might have’ (p. 242):

> one must ask where a 50% prior, though perhaps understandable in other circumstances, here results in the fabrication of a new prior and the dismissal of a vital piece of explicit information. (Seymann 1992, 242)

The problem is here of course is that what is a ‘vital piece of explicit information’ for a statistician is not necessarily the same as for a physician.

To illustrate this, let us first have a look at one of the fables of Ackoff:

> In a conversation with one of my colleagues I was asked how I would go about determining the probability that the next ball drawn from an urn would be black if I knew the proportion and

> number of black balls that had previously been drawn. He told me that the urn contained only black and white balls. I replied that I would first find out how the urn had been filled. “No”, he said, “that is not permissible”. “Why?” I asked, “Certainly you have such information”. “No, I don’t”, he replied. “Then how do you know the urn contains only black and white balls?” I asked. “I have it on good authority”, he answered. “Then let me talk to that authority”, I countered. In disgust he told me to forget about the whole thing because I clearly missed the point. I certainly did. (Ackoff 1974, 89)

The moral of this fable is that the ability to solve a textbook exercise is not equivalent to the ability to solve a real-world problem. Textbook exercises are usually formulated so as to have only one correct answer and one way of reaching it. Real-world problems have neither of these properties. An essential part of problem solving, according to Ackoff, lies in determining what information is relevant and in collecting it.

By discussing six problems in reasoning with probabilities, so-called ‘teasers’, Bar-Hillel and Falk (1982) show that the way we model a problem is strongly dependent on the way the information was obtained.

> The kind of problem in which the conditioning event does turn out to be identical to what is perceived as ‘the information obtained’ can only be found in textbooks. Consider a problem which asks for ‘the probability of A *given* B’. This non-epistemic phrasing sidesteps the question of how the event B came to be known, since the term ‘give’ supplies the conditional event, by definition. […] Outside the never-never land of textbooks, however, conditioning events are not handed out on silver platters. They have to be inferred, determined, extracted. In other words, real-life problems (or textbook problems purporting to describe real life) need to be *modelled* before they can be solved formally. And for the selection of an appropriate model (i.e., probability space), the way in which information is obtained (i.e. the statistical experiment) is crucial. (Bar-Hillel and Falk 1982: 120-121)

Bar-Hillel and Falk emphasize that a probability space for modelling verbal problems should allow for the representation of the given outcome and the statistical experiment which yields it. They illustrate how different scenarios for obtaining some information yield different solutions. In other words, the way one model a problem is strongly dependent on how the information is obtained. Different ways of obtaining the selfsame information can significantly alter the revision of probability contingent upon it. Real-life problems need to be modelled before they can be solved formally. And for the selection of an appropriate model (e.g., probability space), the way in which information is obtained (i.e. the statistical experiment) is crucial.

In the case of The Harvard School Test (Casscells et al 1978) and in the later test by Eddy (1982), it was simply assumed that both questioner and respondent had the same model in mind. However, both were trained differently and therefore had modelled the problem differently.

7. Conclusions

Generally, rational decision-making is conceived as arriving at a decision by a correct application of the rules of logic and statistics. If not, the conclusions are called biased. After an impressive series of experiments and tests carried out the last few decades, the view arose that rationality is tough for all, skilled field experts not excluded. A new type of planner’s counsellor is called for: the normative statistician, the expert in uncertainty par excellence.

To unravel this view, the paper has explored a specific practice of clinical decision-making, namely Evidence-Based Medicine. This practice is chosen, because it is a very explicit about how to rationalize practice.

One of the key examples of biased expertise is the Harvard Medical School Test, which shows that physicians often commit a base rate fallacy: they confuse the accuracy of a diagnostic test for its predictive value. However, it is shown that for the base rate given in the Harvard Medical School Test it is not rational to ask for a diagnostic test. Moreover, it is shown that for base rates between the test-treatment thresholds, it is an unbiased heuristic to take a test's accuracy as its predictive value.

Most practices of rational decision making prefer the ratios of likelihoods to simple likelihoods because they are easier and more practical to update when new evidence (e.g. a test result) comes in. The term bias has a specific meaning in mathematical statistics. Using this specific interpretation of biasedness, it is shown that paradoxically a rational application of likelihood ratios leads to biased results.

It has also been shown that whether a decision-making process is rational cannot be assessed without taking into account the environment in which the decisions have to be taken. To be more specific, the decision to call for new evidence should be rational too. This decision and the way in which this evidence is obtained are crucial to validate the base rates. Rationality should be model-based, which means that not only the isolated decision-making process should take a Bayesian updating process as its norm, but should also model the acquisition of evidence (priors and tests results) as a rational process. The use of thresholds is an option for that.

References

Ackoff, Russell L. 1974. *Redesigning the future: A systems approach to societal problems*. New York: Wiley.

Bar-Hillel, Maya and Ruma Falk. 1982. Some teasers concerning conditional probabilities. *Cognition* 11: 109-122.

Casscells, Ward, Arno Schoenberger and Thomas Grayboys. 1978. Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine* 299 (18): 999-1000.

Cooke, Roger M. 1991. *Experts in uncertainty: Opinion and subjective probability in science*. New York and Oxford: Oxford University Press.

Eddy, David M. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, Slovic and Tversky (1982).

EBM, Evidence-Based Medicine Working Group. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *The Journal of the American Medical Association* 268 (17): 2420-2425.

Gigerenzer, Gerd. 2004. Fast and frugal heuristics: The tools of bounded rationality. In *Blackwell handbook of judgment and decision making*, edited by D. Koehler and N. Harvey. Oxford: Blackwell.

Guyatt, Gordon H. 1991. Evidence-based medicine. *Annals of Internal Medicine* 114 (*ACP Journal Club* supplement): A-16.

Kahneman, Daniel, Paul Slovic and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge et al.: Cambridge University Press.

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes. 1974. *Introduction to the theory of statistics,* 3rd edition. Tokyo: McGraw-Hill.

Morgan, J.P., N.R. Chaganty, R.C. Dahiya, and M.J. Doviak. 1991. Let’s make a deal: The player’s dilemma. *The American Statistician* 45 (4): 284-287.

Pauker, Stephen G. and Jerome P. Kassirer. 1980. The threshold approach to clinical decision making. *The New England Journal of Medicine* 302 (20): 1109-1117.

Sackett, D.L., S.E. Straus, W.S. Richardson, W. Rosenberg and R.B. Haynes. 2000. *Evidence-based medicine: How to practice and teach EBM*, 2nd edition. Edinburgh et al.: Churchill Livingstone.

Scherokman, Barbara. 1997. Selecting and interpreting diagnostic tests, *The Permanente Journal*, http://xnet.kp.org/permanentejournal/fall97pj/tests.html.

Seymann, Richard G. 1991. Comment, *The American Statistician* 45 (4): 287-288.

Seymann, R.G. 1992. Response, *The American Statistician* 46 (3): 242-243.

Thaler, Richard H. and Cass R. Sunstein. 2003. Libertarian paternalism. *The American Economic Review* 93 (2): 175-179.

Tversky, Amos and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124-1131. Reprinted as chapter 1 in Kahneman, Slovic and Tversky (1982).

vos Savant, M. 1990. ‘Ask Marilyn’, *Parade Magazine*, September 9: 13.

Wendell, John P. 1992. Comment, *The American Statistician* 46 (3): 242.

LONDON SCHOOL OF ECONOMICS
DEPARTMENT OF ECONOMIC HISTORY

WORKING PAPERS IN: 'THE NATURE OF EVIDENCE: HOW WELL DO "FACTS" TRAVEL?'

For further copies of this, and to see other titles in the department's group of working paper series, visit our website at:
http://www.lse.ac.uk/collections/economichistory/

2005

01/05: Transferring Technical Knowledge and innovating in Europe, c.1200-c.1800
Stephan R. Epstein

02/05: A Dreadful Heritage: Interpreting Epidemic Disease at Eyam, 1666-2000
Patrick Wallis

03/05: Experimental Farming and Ricardo's Political Arithmetic of Distribution
Mary S. Morgan

04/05: Moral Facts and Scientific Fiction: 19th Century Theological Reactions to Darwinism in Germany
Bernhard Kleeberg

05/05: Interdisciplinarity "In the Making": Modelling Infectious Diseases
Erika Mattila

06/05: Market Disciplines in Victorian Britain
Paul Johnson

2006

07/06: Wormy Logic: Model Organisms as Case-Based Reasoning
Rachel A. Ankeny

08/06: How The Mind Worked: Some Obstacles And Developments In The Popularisation Of Psychology
Jon Adams

09/06: Mapping Poverty in Agar Town: Economic Conditions Prior to the Development of St. Pancras Station in 1866
Steven P. Swensen

10/06 "A Thing Ridiculous"? Chemical Medicines and the Prolongation of Human Life in Seventeenth-Century England
David Boyd Haycock

11/06 Institutional Facts and Standardisation: The Case of Measurements in the London Coal Trade
Aashish Velkar

12/06 Confronting the Stigma of Perfection: Genetic Demography, Diversity and the Quest for a Democratic Eugenics in the Post-war United States
Edmund Ramsden

13/06 Measuring Instruments in Economics and the Velocity of Money
Mary S. Morgan

14/06 The Roofs of Wren and Jones: A Seventeenth Century Migration of Technical Knowledge from Italy to England
Simona Valeriani

15/06 Rodney Hilton, Marxism, and the Transition from Feudalism to Capitalism
*Stephan R. Epstein (*also published as 94/06 in the Department of Economic History Working Paper series)

2007

16/07 Battle in the Planning Office: Biased Experts versus Normative Statisticians
Marcel Boumans