e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-975

Enhancing informatics competency under uncertainty at the point of decision: a knowing about knowing vision

Mette Kjer KALTOFT[a,1] Jesper Bo NIELSEN [a] Glenn SALKELD [b] and Jack DOWIE [c]
[a] *University of Southern Denmark*
[b] *University of Sydney School of Public Health*
[c] *London School of Hygiene & Tropical Medicine*

Abstract. Most informatics activity is aimed at reducing unnecessary errors, mistakes and misjudgements at the point of decision, insofar as these arise from inappropriate accessing and processing of data and information. Healthcare professionals use the results of scientific research, when available, and 'big data', when rigorously analysed, as inputs into the probability judgements that need to be made in decision making under uncertainty. But these judgements are needed irrespective of the state of 'the evidence' and personalised evidence on person/patient-important criteria is very often poor or lacking. This final stage in 'translation to the bedside' has received relatively little attention in the medical, nursing, or health informatics literature, until the recent appearance of 'cognitive informatics'. Positive experience and feed-back from several thousand students who have experienced exercises in assigning probabilities informs our future vision in which better decisions result from healthcare professionals – indeed all of us – having accepted that probability assignment is a skill, with the internal coherence and external correspondence of the probabilities assigned as twin evaluative criteria. As a route to improved correspondence – in the absence of the systematic recording and monitoring of real world judgments that would be the normal pathway to quality improvement - a 'Prober' is a set of statements to which the respondent supplies their personal probabilities that a statement is true. They receive the proper Brier score and its decomposition as analytical feedback, along with graphic representations of their discrimination and calibration, the two key components of good correspondence. Provided with estimates of their sensitivity (mean probability true for true statements) and specificity (1 minus mean probability true for false statements) they can visualise themselves as a 'test' when making diagnostic and prognostic judgements , thereby being given the cognitive foundation for such reflection in their clinical practice, including 'reflection in action'. They acknowledge that an appropriate balance of intuition and analysis is required, as in Hammond's Cognitive Continuum, and are made aware of the cognitive and motivated biases that can prevent us knowing 'how much we know about how much we know', with its deleterious effect on decision quality. Probability exercises, such as 'Probers', are proposed as an enhancement of professional courses and virtual learning environments, such as the TIGER initiative in nursing, through which the competency portfolio of all those seeking to deliver high quality person/patient-centred care can be expanded.

[1] Corresponding Author: Mette Kjer Kaltoft, Health Visitor RN MPH, Research Unit for General Practice, Institute of Public Health, University of Southern Denmark, Odense 5000, Denmark E-mail: mkaltoft@health.sdu.dk

Keywords. Probability, judgement, coherence, correspondence, calibration, discrimination

Introduction

Our vision is of the better decisions that will characterise the coming era of person/patient-centred care as a result of healthcare professionals – indeed all of us - accepting that, in decision making, we are necessarily Bayesians.[1] We accept that the assessments of the future chances which permeate decisions are ontologically personal and subjective, whatever the extent to which they are epistemologically-based on robust frequencies and however widely they are inter-subjectively agreed. All parties have rejected the temptations of right-wrong thinking, reflected in testing by non-probabilistic Multiple Choice Questions, along with the unwarranted confidence, trust and denial it often generates. Healthcare professionals treat the results of scientific research, when available, and 'big data', when rigorously analysed, as relevant inputs into the probability judgements that need to be made irrespective of the state of 'the evidence'. It is accepted that competence in making probability judgements is the key to improved handling of uncertainty at the point of decision so it is part of the training and education of clinicians.

Most informatics activity is ultimately aimed at reducing unnecessary errors, mistakes and misjudgements at the point of decision, insofar as these arise from inappropriate accessing and processing of data and information. For some criteria and some conditions high-quality 'evidence- based' probabilities can be acquired directly or through a nomogram or 'risk calculator' (preferably a *probability* calculator). [2] But in many cases the clinician will need to use their personal belief probability judgements to remedy the absence of, or to better personalise, the available estimates.

This frequently necessary final stage in 'bench to bedside translation' has received relatively little attention in the medical, nursing or health informatics literature. The widespread assumption has been that this is an intuitive competence that can, and can only, be acquired intuitively, through experience. However, this ignores a significant literature on how the quality of probability judgements can be assessed, on the empirical evidence on clinician performance in this respect, [3,4] on the possible sources of limited performance, and on possible routes to improved quality. Since it will take time to overcome the institutional-professional barriers to systematic judgemental recording and monitoring in practice – the normal route to competence improvement - our vision is pessimistic in this respect. However, as part of the increasing interest in 'cognitive informatics', clinicians can be provided with the cognitive basis for reflecting continuously on their judgemental practice and performance, both 'in action' and outside it, [5,6] accepting that an appropriate balance of intuition and analysis is required, as in Hammond's Cognitive Continuum, [7–9] as well as an awareness of the likely cognitive as well as motivated biases that may hinder them knowing 'how much they know about how much they know'. [10] Probability exercises (such as 'Probers') are therefore an integral part of our vision, enhancing professional courses and virtual learning environments, such as The TIGER Initiative in nursing.[11]

In relation to the evaluation of probability assessments - and assessor - Kenneth Hammond and others have emphasised that two distinct criteria are relevant, and drawn attention to the fact that, for a variety of reasons, including different meta-theoretic

paradigms, the two have attracted different sets of adherents. [12] There are those who wish to judge probabilities primarily by their *internal coherence* and those who wish to judge them primarily by their *external correspondence*. The vast majority of those who emphasise coherence are pessimistic about judgemental competence, because clinicians typically perform poorly on coherence tests, such as calculating the predictive value of a test result, given the sensitivity of the test and the prevalence of the target condition. Most optimists emphasise external correspondence, arguing that abstract tests of coherence are not 'ecologically valid', [13] since the items are not representative of those that actually arise. But there are also pessimists among those who favour the correspondence criterion, doubting whether experience will be productive in the absence of quick and unbiased feedback. [14,15] The 'clinical versus actuarial' controversy, associated primarily with the name of Paul Meehl, [16] rumbles on.

1. Methods

In the case of the coherence criterion, teaching of the way in which probabilities should be combined is required. Correspondence can only be taught through probabilistic exercises with relevant feedback. A Prober is a set of statements to which the respondent supplies their personal probability that a statement such as 'The true positive rate indicates the sensitivity of a test' is true. The set used currently consists of 50 statements relating mainly to research methods. A variety of probability response sets are available for use in the software. A compromise between response granularity and item set size is necessary to achieve a reasonable number of observations for an individual at each probability. We currently use seven discrete probabilities: 0, 10, 30, 50, 70, 90 and 100%. Respondents are advised that they should enter their honest probabilities and in order to avoid 'motivated biasing', they will receive full marks for completion of the exercise. In any case, the accompanying teaching makes clear that the assessments are scored by a proper scoring rule (Brier's) which ensures that respondent's expected score will always be maximised by reporting honest beliefs [17].

After completion the respondent can learn whether each statement was actually true or false, along with short elaborations, mainly in the case of false items. The main, analytical feedback comes in the form of the Brier score and its decomposition, [18] (Figure 1a) One key measure is that of *discrimination*, the difference between the average probabilities assigned to true and false items, plotted on the right and left axes respectively. (These represent the sensitivity and 1 minus the specificity of the judge interpreted as a 'test'.) Graphically discrimination is represented by the slope of the line joining them. This can be compared with the 45 degree slope of the diagonal which indicates perfect discrimination. An associated diagram (Figure 1b) provides information relevant to the other key competence, *calibration*. Calibration is measured by the degree to which the 'frequency correct' matches 'probability assigned'. For example, if a respondent assigned 70% to 10 statements, then perfect calibration exists if 7 of these are actually true. Deviations from 7 in *either* direction indicate poorer calibration. Accompanying teaching stresses that calibration should not be improved at the expense of using whatever discrimination ability is possessed

2. Results

The latest in 35 years of Probers use has been in the Translational Health Masters course at the Sydney School of Public Health. In 2012 and 2013, 63 students responded. (Completion rates were high as the exercises were a compulsory assignment). Their Brier scores ranged from .1 to .55 (where 0 is perfect and 1 is worst possible.) The mean score of .25 (SD .08) is actually that which would be achieved by assigning .5 probability to all 50 statements, so that on average the population did no better than chance. The average sensitivity (mean probability true assigned to true statements) was 75% and average specificity (1 minus probability true assigned to false statements was 64%. Only one of the 63 had a specificity exceeding sensitivity and hence a discrimination line with a negative slope.

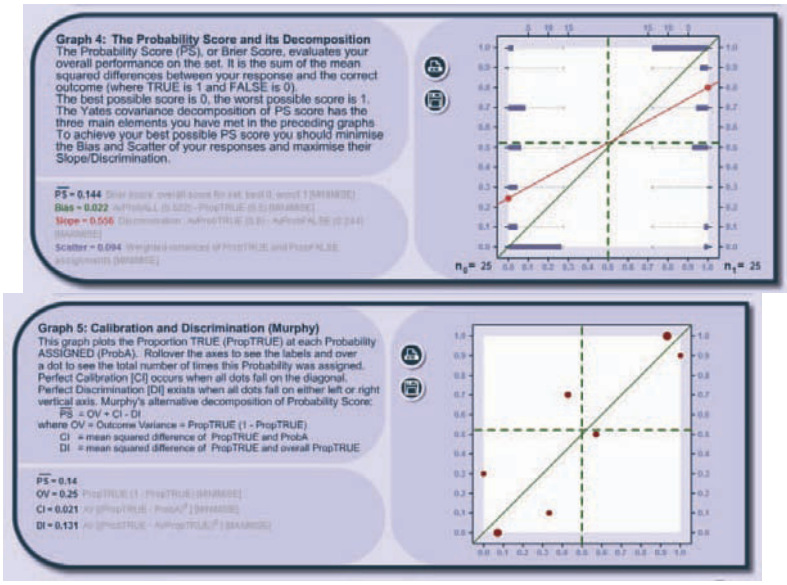

Figure 1 Student showing good discrimination (top) and calibration botton

As in previous settings there was no indication of respondent difficulty in completing the task at a practical level. Feedback comments have been solely about the unfamiliar nature of the task, without questioning of its relevance, and mainly doubts about whether using numerical probabilities and regarding it as a skill would be acceptable 'where I work' because it would be disruptive of organisational routines and/or professional hierarchies

3. Discussion

Having arrived at a numerical estimate of, say, 30%, the Prober-aware health professional will recall that if all their 30%s were monitored and collated the frequency correct should be 30%. They will be able to reflect 'in action' on their *calibration.* In relation to their whole set of judgements and outside any specific case, they can ask themselves whether they assigned a (much) higher average probability to the occasions

when the target outcome occurred, than the average assigned when it did not occur. They will be able to reflect, outside of action, on their sensitivity and specificity and overall *discrimination* competence.

Where is the 'evaluation' of Probers? Real world evaluation requires the systematic recording and monitoring of judgements that seems almost impossible in larger clinical settings. In our vision the 'anatomy of judgment' is taught alongside the anatomy of the human body in clinical curricula. Probers are part of the new cognitive informatics.

References

[1] J. Dowie The Bayesian approach to decision making, in A. Killoran, C. Swann, M. Kelly, Editors. *Public Health Evidence: Tackling Health Inequalities*, Oxford, Oxford University Press; 2006. p. 309–21

[2] J Dowie, Against risk, *Risk Decision and Policy* 4 (1999) 57–73

[3] J.G. Dolan, D.R. Bordley, A.I. Mushlin, An evaluation of clinicians' subjective prior probability estimates. *Medical Decision Making* 6 (1986) 216–23

[4] T.G.Tape, J. Kripal, R.S.Wigton, Comparing methods of learning clinical prediction from case simulations. *Medical Decision Making,* 12 (1992) 213–221

[5] D Schön *The Reflective Practitioner. How professionals think in action*, London, Temple Smith, 1983

[6] P Benner *From Novice to Expert: Excellence and Power in Clinical Nursing Practice*, London, Addison-Wesley, 1984

[7] K.R.Hammond *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice,* New York, Oxford University Press; 1996

[8] J. Dowie, M.K.Kaltoft Deciding how to decide – and how to support decisions. *Nuffield Trust Webinar*. http://www.slideshare.net/NuffieldTrust/jack-dowie-211111. 2011

[9] J. Dowie, JUDEMAKIA : a personal map of the world of judgement and decision making. https://www.dropbox.com/s/ph800ycdah5no92/Judemakia pdf.pdf.2006

[10] A. Tversky, D Kahneman, Judgment under uncertainty: heuristics and biases, *Science* 185 (1974) 1124–31

[11] M.K.Kaltoft Nursing Informatics AND Nursing Ethics : Addressing their disconnect through an enhanced TIGER-vision, *Studies in Health Technology and Informatics*, *192* (2013), 879–883

[12] K.R. Hammond How convergence of research paradigms can improve research on diagnostic judgment. *Medical Decision Making* 1996 16 (1996) 281–287

[13] Hammond KR. Ecological Validity: Then and Now.1998 http://www.brunswik.org/notes/essay2.html

[14] B Brehmer, In one word: not from experience. *Acta Psychologica* 45 (1980) 223–41

[15] I. Fischer, D.V. Budescu, When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organisational. Behavior and Human Processes* 98 (2005) 39–53

[16] P.E. Meehl, Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50 (1986) 370–5

[17] G.W. Fischer, Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting *Organisational. Behavior and Human Performance* 29 (1982) 352–69

[18] J.F.Yates, External Correspondence: Decompositions of the Mean Probability Score, *Organisational. Behavior and Human Performance* 30 (1982) 132–56.

[19] A.S. Elstein, Beyond Multiple-choice Questions and essays: the need for a new way to assess clinical competence, *Academic Medicine*. 68 (1993) 244–9

Access To obtain access to the Prober set, email jack.dowie@sydney.edu.au **Conflicts of Interest** Prober is ©Maldaba Ltd; Contact: info@maldaba.co.uk . Jack Dowie has a financial interest in Prober, but has not benefited from its use in any of the specified settings. **Authorship** JD and MKK jointly planned the paper's structure and content. JD's first draft was extensively revised with MKK. GS and JBN also provided comments. All authors approved the final version.