Clinical Judgement Analysis: A one day international meeting

17th April, 2002
Centre for Decision Research,
Leeds University Business School (LUBS), UK.

Supported financially by the European Association for Decision Making (EADM)
and Leeds University Business School (LUBS).

Clinical Judgment analysis is the application of Judgement Analysis to medical situations. In fact, Judgement Analysis has been used in a clinical context for almost 50 years (see Hammond, 1955 for an early publication). Indeed since it first emerged from the Brunswikian principle of probabilistic functionalism, the main application of judgement analysis has been to clinical judgement (see e.g. Wigton, 1988; Wigton, 1996; Engel, Wigton, LaDuca and Blacklow, 1990 for overviews). Several key findings have resulted from this considerable body of applied and experimental research. For example, Tape, Heckerling, Ornato and Wigton, (1991) demonstrated the importance of examining the predictability of the environment and the validity of cues in their study of physicians' judgements about pneumonia in three different states in the USA. Kirwan, Chaput de Saintonge, Joyce and Currey (1983 a,b) examined the usefulness of paper cases as a research technique. Dhami and Harries (2001) used a fast and frugal model to describe clinical judgements. However, despite the large amount of research in this area, most review or summary papers moot multiple linear regression as the main analytical technique, and do not discuss alternatives. In an approach reminiscent of behaviourism most methodological techniques collect data on judgement as an outcome, rather than as a process. Since they are based in applied research, most cja studies use these modal methodological and analytical techniques as tools to investigate particular behaviours or examine particular topic-related hypotheses. The implications of use of particular scientific techniques for the shape of results should not be overlooked (see Eddington, 1949). The specific aims of this international one-day meeting were to identify current and past methodological and analytical techniques of clinical judgment analysis, and to identify the purpose, pros and cons of <u>potential</u> methods and analyses. The overarching aim of the meeting was to create opportunities for discussion between researchers who are new to this approach and more experienced researchers. Were the aims of the meeting met?

The specific aims of the meeting were met. The 13 papers at the meeting were grouped into those with a methodological slant, papers with an analytical slant, papers with a prescriptive slant and papers that gave an overview – of clinical judgment analysis and of issues in its application. Researchers presented applications of judgement analysis to assessments, diagnoses and treatment decisions such as, for example, General Practitioners' (GP) diagnosis of heart failure, rheumatologists' assessment of arthritis, community occupational therapists' referral prioritisation decisions in a mental health team, psychiatric nurses' risk assessments and decisions to observe, GPs', Cardiologists' and care of the elderly physicians' management of patients presenting with chest pain, a variety of practitioners' decisions to admit asthmatic children to hospital, prescription decisions for treatment of asthma and UTI, and diagnosis and treatment in youth and family care. Experimental research exploring methodological issues was also presented. Many issues of application and development of clinical judgment analysis were raised. Here I shall focus on three of these: representativeness, representation of knowledge and the applied context of clinical judgement.

<u>Representativeness.</u>
At the start of the day, Tom Tape (University of Nebraska) described the historical development and emergence of clinical judgement analysis. He outlined three principles that distinguished the work of Egon Brunswik: his framing of human behaviour within the context of an unpredictable world

(probabilistic functionalism), his focus on each individual's decision making (an idiographic approach), and his emphasis on the familiarity of the participant or subject with the task or the stimulus and behaviour being examined (representativeness). The purpose of representative design is essentially to allow generalisations or conclusions to be drawn about an individual's behaviour. Performance with lab-based stimuli allows us to generalise about lab-based stimuli, but not necessarily beyond that. In his examination of perception, Brunswik had a student followed around campus. She was stopped at various times to make judgments of size of an object in her visual field, its distance and its projected size onto the equivalent of the retina and these actual sizes and distances were measured (Brunswik, 1944). Few studies using judgement analysis have attempted the equivalent (see Dhami, Hertwig and Hoffrage, 2000). In practice, where naturally occurring judgments have been studied, analysis is often nomothetic rather than idiographic. For example, Tape's studies of judgments of pneumonia grouped together the judgements of different physicians in a region. Where analysis is idiographic, each individual's judgements are more often than not made on 'paper' cases. With such cases, researchers have emphasised maximisation of the statistical representativeness (in particular the correlation between cues) rather than the fidelity of the study stimuli (*cf* Elstein, Shulman and Sprafka, 1978).

In the first talk illustrating an application of clinical judgment analysis, Ylva Skånér (Dept of Family Medicine, Stockholm Karolinska Institute) emphasised the difficulties of identifying a representative set of cases that capture the "conditions and constraints" of the judgment task. Although, Skånér had used data from real patients transformed into case vignettes, these patients were representative of a particular condition or health service experience (patients identified with heart failure in health centres in study one and patients referred from general practice to a cardiology outpatient clinic suspected of having heart failure in study two), and therefore of a particular age distribution. Of course these patients were not necessarily representative of the broader class of patients. But they were also not necessarily representative of the patients of this type seen by any one individual doctor nor those of this type seen by physicians of different specialities. As Tom Stewart (State University of New York at Albany) pointed out in his talk, even a sample of naturally occurring judgments is not necessarily (statistically) representative of an individual's experience.

Representativeness of a sample of cases is one focus; a second is whether people's judgments on 'paper' cases are similar to their judgments on real cases. Use of 'paper' cases allows for measurement of consistency (test-retest reliability), for comparison of different physicians' behaviour over the same set of cases, and, by reducing the statistical representativeness of the intercorrelations between cues in the sample (and hence losing the naturally occurring patterns of information), the independent effects of information on a person's judgments can be measured. These analyses and more were illustrated in Kirwan's (University of Bristol) romp through his still influential MD thesis of 20 years ago. Kirwan's work has had a major impact on the development and uptake of universal criteria for assessment of rheumatism, however his methodological work examining the validity of paper cases is also widely cited. Rheumatologists judged real life cases and paper presentations of the same real life cases, with a high degree of consistency. However, as he emphasised, the validity of any set of paper cases has to be established for itself and, as Tom Stewart phrased it, environmental validity is the property of a study, not of a methodology.

Issues of both statistical representativeness and measurement of validity of paper cases are recurrent ones. Forrest, Harries, Harvey, Bowling and Hemingway (University College London) increased the fidelity of the experience compared to a vignette, by forcing each physician to collect information on computer presented cases, just as they would in a real life consultation. Their non-representative set of hypothetical cases consisted of possible (likely even) patients with chest discomfort. One aim of their study was to examine age-related effects on decision making – hence age ranges from middle to old age were more evenly spread than the usual high proportion of elderly patients, and were independent of other effects. But they would be hard-pushed to find a set of cases that would be equally representative of patients with chest pain seen by their General Practitioner, by Cardiologist and Care

of the Elderly physicians. Dowding, Cassells and Brodie (University of Stirling) face related problem of statistical representativeness. In their aim to examine nurses' judgments of risk of suicide and decisions to observe psychiatric inpatients, initial data regarding the predictiveness (ecological validity) of, and relationship between cues was based on a thorough examination of the literature, which draws conclusions across a different and broader population than is seen on inpatient wards, and in terms of risk of suicide in the long rather than the short term. In fact, both research teams had applied for grants that included analysis of judgements on real cases. In both cases, only the hypothetical patient halves of the grants were funded. Clearly issues of representativeness need to be raised with grant funders as well as researchers.

Representation of knowledge

The representation of knowledge was addressed in several talks. Papers were presented that explored simple process tracing models and information search (Kee and Forrest et al), that explored nonlinear and configural aspects of cue use (Harries, Koele), that contrasted simple or weighted linear additive rules with exemplar based decision making (Juslin), and that formulated models of judgment and decision making in terms of a set of arguments (Fox).

Frank Kee (Queens University, Belfast) presented a lens model analysis of fast and frugal decision making models. Although Dhami and Harries (2000) and Smith and Gilhooly (2001) have explored idiographic models of General Practitioners' decision making, and Tape and colleagues (2001) have explored a nomothetic two-sided lens model of physicians' judgements, Kee developed two-sided idiographic lens model analyses of management of asthma in an emergency department, using agreement by more than 75% of consultants as the gold standard on the environmental side. Using area under the ROC curve as a proxy for the C index (non-linear matching) Kee compared the ability of fast and frugal models (a type of Take the Best rule) of each practitioner's judgments and logistic models of practitioner's judgments, Dawes' rule based on their judgments, and their actual judgements in terms of their ability to predict this gold standard. Kee expected the less-is-more principle to imply that those who were less expert would be better at exploiting simple models. Although across the 50 practitioners, the take the best style model and the Dawes' rule model did worse than the logistic regression model (areas under the curve were significantly smaller), the logistic regression model (and the Dawes' rule model) of consultants' decisions performed better than that based on other participants, but consultants and other participants did not differ in performance of their take the best models. The take the best model of paediatricians was better than that of non-paediatricians, but they did not differ in terms of the performance of the other two models.

Although fast and frugal models are quintessentially simple, as Pieter Koele (University of Amsterdam, The Netherlands) pointed out in his talk, often experts explicitly describe a judgment or decision making policy that is superficially more complex (involving non-linear, and configural relationships between pieces of information) than that captured by a weighted linear additive model. Of course, as Einhorn (1971) made clear, mathematical and cognitive simplicity do not necessarily equate. Einhorn (1972) demonstrated that a disjunctive model was the best descriptor of the judgment making of three physician participants. Indeed in his overview of good practice in applying judgment analysis, Tom Stewart emphasised the importance of analysing the C index of the lens model, as a partial antidote to the linear additive assumptions of regression models. In an effort to capture these more non-linear relationships more precisely, Priscilla Harries (Brunel University, UK) elicited explicit judgements about the effect of each level of a cue, as well as an overall rating of its impact on judgment and decision making. In his study, Pieter Koele used the C index of the lens model (non-linear matching) as a measure of people's ability to learn relatively simple configural relationships. After ninety trials presenting outcome feedback, only one of the ten participants' judgments matched the configural aspects of this environment.

Like multiple cue probability learning (MCPL), judgment analysis assumes a need to discover rules underlying people's behaviour. In his exploration of the differing exemplar based and rule based

explanations of behaviour found in categorization research and MCPL research respectively, Peter Juslin (University of Umeå, Sweden) demonstrated the importance of the role of outcome feedback on learning in representation of knowledge and formation of judgements. People's ability to extract and apply rules to distinguish between dangerous and harmless bugs, can be seen in the way they make judgements on exemplars outside the range seen during training. Juslin found that participants tended to integrate information rather than to rely on simple lexicographic rules, and that participants who learnt both on analogue exemplars (the equivalent of patients) and those who learnt on conceptually described exemplars (the equivalent of paper patients) showed behaviour more reminiscent of exemplar based judgment than rule based judgement when feedback was binary (as it usually is in categorisation tasks) but those who had outcome feedback on a continuous variable were relatively more inclined to use rule based reasoning. When feedback was probabilistic (as it is for most physicians) rather than deterministic, participants again tended to extrapolate rules, rather than rely on exemplars. The more similar training was to test phases, the more people used exemplars. In fact of course, most doctors are unlikely to have access to complete outcome feedback. When they do have feedback, with the exception of news of death, the feedback is unlikely to arrive in a neatly dichotomised form. The information is likely to be probabilistic. With more experience, a physician's "test phases" (the current case) will resemble their "learning phase" (everything they have experienced up till now). As physicians gain experience, their potential for use of exemplar based, intuitive judgment increases. Capturing this behaviour in terms of rules (fast and frugal, weighted linear additive, or configural) may of course be missing something.

Although his PROforma model examines the broader context of clinical processing, John Fox (Cancer Research, UK) placed the Lens model in the cognitive strategy part of this model. Fox's argument-based representation of behaviour is yet another means of representing physicians' knowledge and judgment and decision processes. Usually clinical judgement analysis has captured the probabilistic nature of peoples' behaviour and the inherent uncertainty of the environment in terms of predictiveness or descriptive fit of static and deterministic models, in which the relative weight of each component a model reflects its ecological validity or utilisation. In contrast, Fox's argument-based approach encapsulates probabilism or uncertainty within the components of a model based on relative aggregated net support. Each argument contains a numerical or carefully defined qualitative qualifier that is often operationalised into supporting, conclusively supporting, opposing and conclusively opposing signs. These arguments, and their qualifiers, apply to general beliefs ((medical) knowledge), preferences, and situational (case specific) beliefs. In a standard judgment analysis of course medical knowledge, and preferences are captured in a different way to situational factors. The former are contained within regression model weights and constant, and the equivalent of situational arguments are the application of the regression model to any one case. Each argument-based representation of the reasoning process is specific to an exemplar, the equivalent of the particular increasing and decreasing sum of weighted information as a regression model is applied to a particular case.

<u>The applied context of clinical judgement</u>

The applied nature of work using clinical judgment analysis cannot be escaped. For example, risk of suicide assessments and decisions to observe patients are studied in order ultimately to reduce suicide rates; age-related inequalities are studied in order to maximise the clinical appropriateness and equity of medical decision making. Priscilla Harries (Brunel University, UK) used cluster analysis to identify a gold standard for Occupational Therapists' decisions of referral prioritisation. This could then be used to train occupational therapy students about the most appropriate referrals to accept - essential knowledge for a profession where demand exceeds service availability. Fox's PROforma model is also a prescriptive decision aid and his argument-based approach is compelling in terms of comprehendibility to the practitioner.

Despite their lack of immediate interpretability, feedback based on linear models has been shown to be extremely effective in improving clinical judgment analysis (see Wigton, 1996). Some conditions for

the positive effects of this feedback were explored by Petra Denig (University of Groningen, The Netherlands). Denig described studies using cognitive and outcome feedback or cognitive feedback and task information or all three to change the routine prescribing of doctors in different countries. Prescribing practice for Asthma was improved in the Netherlands studies, (in which all three types of information were given as 'feedback') but there were relatively negative effects in the Norwegian and Swedish studies (in which task information was not included). In contrast outcome and cognitive feedback were sufficient to improve prescribing practice for UTI in Norway and Sweden, and all three types of feedback improved UTI prescribing practice in The Netherlands. Although there may be cultural differences in training behaviour, Denig reported little impact of feedback where policies were simple, but that graphs representing policies were easier to understand when comparisons were made in an interactive context. Denig also discussed the role of critical cases in maximising the effect of outcome and cognitive feedback. Earlier, Juslin had highlighted the role of critical exemplars on people's ability to extract rules. Interestingly, the role of critical cases for an experimenter's ability to describe a decision making model has also been emphasised by Gary McLelland (1999) in his essay on representative and efficient study designs.

The relative role of feedback in changing practice in particular domains was echoed in Kirwan's review of the impact his work has had on assessment of rheumatoid arthritis: cognitive feedback had a large impact on identifying disagreements between individuals, and on reducing these, but the main impact of his work was to stimulate development of internationally agreed standards of RA assessment that dramatically changed the practice, and teaching of rheumatology.

The day finished with Huub Pijnenburg's (Praktikon, Nijmegen, The Netherlands) presentation on clinical decision making, which firmly put decision making into its applied and socio-political context. Pijnenburg described the dramatic legal and logistical restructuring that youth care is undertaking in the Netherlands. In the old care structure separate organisations within health, legal and welfare systems sequentially assessed a client's needs in relation to the services that they provided. This probably exacerbated the identified problems with clinical decision making whereby there was a lack of attention to clients' desires, limited and confirmatory diagnoses, and a lack of relationship between clinical needs and treatment regime. Within such a structure, handheld Bayesian based decision aids, and diagnostic expert systems had limited impact on improving clinical decision making. Presumably judgment analysis and feedback of task information would not have faired better. In the new system, each client's diagnostic and treatment assessments will occur at a central point and be based on standardised protocols and their treatment regime will be provided by an allocated care centre. Pijnenburg's emphasis was on prediction as to how clinical decision making will be able to improve and how it will still be limited. Self-interest of care centres will be avoided, and diagnostic decision making should be improved, avoiding hypothesis confirmation etc, but, unless specific measures are put into place, outcome feedback from this two tier system will be just as just as elusive as it was in the old system. The main recommendation from the audience on this point was the need, not for analysis of individuals' models of decision making but for task analysis based on multiple methods including focus groups. In his talk Tom Stewart had emphasised the specificity of good judgment analyses: in which experts' behaviour is studied on natural samples, cross validation samples and reliability testing samples of cases in relation to a clearly identified gold standard which is itself based on a thorough 'front end' study of the environment. In the context of youth care in the Netherlands, as in many clinical contexts, the change of practice and role necessitated by changing circumstances means that there will now be no expert practitioners. As in many other clinical contexts, analysis of the task, and of the multi-faceted environment is of primary importance.

During the day, numerous applications of clinical judgment analysis were discussed, the potential for different analyses and different representations were seen, and limits to practical implications were explored. The 27 participants were academics in psychology, medicine, policy research, occupational therapy, pharmacology, nursing, and social care, and been working on clinical judgment for 30 years,

20 years, 10 years, one or two years or no years. They met and exchanged ideas, and discussed potential developments. Yes, the overarching aims of the meeting were also met.

Dr Clare Harries
University Research Fellow
Centre for Decision Research
Leeds University Business School
Maurice Keyworth Building
University of Leeds
Leeds LS2 9JT
U.K.
Tel: + 44 113 343 2634 (NB Change)
Fax: + 44 113 343 4465 (mark FAO Clare Harries)
e-mail: ch@lubs.leeds.ac.uk
Website of the Centre for Decision Research: http://www.leeds.ac.uk/decision-research

References

Brunswik E. (1944) Distal Focussing of Perception: Size Constancy in a Representative Sample of Situations. *Psychological Monographs, 56*:1-49. Reprinted in K.R.Hammond and T.R.Stewart (Eds) The Essential Brunswik: Beginnings, Explications, Applications. New York: Oxford University Press, 2001.

Dhami, M.K. and Harries, C. (2001) Fast and frugal versus regression models of human judgement. *Thinking and Reasoning, 7*, 5-28.

Dhami, M.K., Hertwig, R. and Hoffrage, U. (2000) A review of the use of representative design in social judgment theory research. Presented at the *Millenium Meeting of the Brunswik Society*, Berlin, Germany, July, 2000.

Eddington, A. (1949) *The Philosophy of Science.* Cambridge: Cambridge University Press.

Einhorn, H.J. (1971) Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance, 6*, 1-27.

Einhorn, H.J. (1972) Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*, 86-106

Elstein, A.S., Shulman, L.S. and Sprafka, S.A. (1978) *Medical Problem Solving: An analysis of clinical reasoning.* Harvard University Press.

Engel, J.D., Wigton, R.S., LaDuca, A. and Blacklow, R.S. (1990) A social judgment theory perspective on clinical problem solving. *Evaluation and the Health Professions 13*, 63-78.

Hammond, K.R. (1955) Probabilistic functioning and the clinical method. *Psychological Review, 62*, 255-262.

Kirwan, J.R., Chaput de Saintonge, D.M. Joyce, C.R.B. and Currey, H. (1983a) a Clinical judgment in rheumatoid arthritis I. Rheumatologists' opinions and the development of 'paper patients. *Annals of the Rheumatic Diseases, 42*, 644-647.

Kirwan, J.R., Chaput de Saintonge, D.M. Joyce, C.R.B. and Currey, H. (1983b) Clinical judgment in rheumatoid arthritis II. Judging 'current disease activity' in clinical practice. *Annals of the Rheumatic Diseases, 42*, 648-651.

McClelland, G. (1999) Representative and efficient designs. *Brunswik Society web-site*: http://www.brunswik.org/notes/essay5/essay5.html, September, 1999.

Smith, M.E. and Gilhooly, K (2001) Medical Decision Making: Linear and Fast and frugal models. *Presentation in the University of Leeds Centre for Decision Research Seminar Series. March, 2001*.

Tape, T., Heckerling, P., Ornato, J. and Wigton, R.S. (1991) Use of clinical judgment analysis to explain regional variations in physicians' accuracies in diagnosing pneumonia. *Medical Decision Making 11*, 189-197.

Tape T.G., Konigsberg S., Jacobson M.S., Bessmer J.R., O'Dell, D.V. (2001) How Physicians Decide Whether to Admit "Low Risk" Chest Pain Patients. (Presented at SMDM October 2001). *Medical Decision Making 21*, 541.

Wigton, R.S. (1988) Applications of judgment analysis and cognitive feedback to medicine. In B. Brehmer and CRB Joyce (Eds.) *Human Judgment: The SJT View.* North-Holland: Elsevier Science Publishers.

Wigton, R.S. (1996) Social Judgement Theory and Medical Judgement. *Thinking and Reasoning 2,* 175-190.