Part 1

Background

Chapter 1
INTRODUCTION

1.1 Clinical Judgement Versus Statistical Decision Making

The comparative merits of clinical judgement versus statistical decision-making were first formally examined by Meehl [1954,1957]. He defined "clinical judgement" as the practice of making clinical decisions on the basis of one's own judgement and "statistical decision making" as the practice of making clinical decisions mechanically using a formula or rule that was derived from an empirical relationship. An example of the use of clinical judgement would be a psychiatrist deciding that a patient has schizophrenia on the basis of interview, presentation and background information. An example of statistical decision-making would be a psychiatrist deciding to treat a patient with antidepressant medication on the basis that the patient's score on the CORE scale described by Parker and Hadzi-Pavlovic [1993] was above the cutoff value of 27.

Meehl [1954,1957] reviewed the empirical literature at the time and concluded that clinicians would make more accurate predictions about their patients if they used statistically based formulae rather than their own clinical judgement. This has been the consistent finding of most researchers and reviewers who have visited the issue since then. The empirical base of the reviews has gradually burgeoned over the years. For instance Sawyer [1966] reviewed 45 studies, Wiggins [1973] reviewed 51 studies. Dawes et al [1989] reviewed 100 studies and Grove et al [2000] reviewed 136 studies.

The method of the earlier studies was to identify one particular narrowly defined clinical prediction problem, give the same set of input information to a formula and to a group of clinicians, and have both the clinicians and the formula make a prediction about some sort of criterion. The clinicians and the formula are then directly compared as to their accuracy in predicting this criterion. A representative study is that of Goldberg [1968]. He used MMPI (Minnesota Multiphasic Personality Inventory) profiles from 861 male psychiatric patients collected from seven different clinical settings and who had received a primary diagnosis of either "psychosis" or "neurosis". The MMPI profiles were used as input information and the criterion predicted was the patient's final discharge diagnosis. The clinicians in this study were 13 PhD clinical psychologists (experienced clinicians) and 16 pre-doctoral trainees (novice clinicians). The formula used, an unweighted linear composite of scores on MMPI subscales, was derived from a multiple regression analysis of another similar dataset of 402 cases. The findings were that experienced clinicians, on average, had diagnostic agreement with the criterion diagnosis in 66% of cases, novice clinicians also agreed with the criterion diagnosis in 66% of cases and the formula had a diagnostic agreement in 74% of cases. Contrary to common expectations, experienced clinicians were no better than novice clinicians. This is not so surprising since a number of other studies have also found experience to be unrelated to accuracy of clinical judgements [Wiggins, 1973].

Critics of this kind of research, and of the generalisation that one can conclude from it that statistical decision making is superior to clinical decision making, have pointed out that in many of the studies the prediction task used for comparison was an artificial one, that bore little resemblance to real world clinical decision making. Holt

[1986] contends that the contrived nature of the decision making task used in many studies handicapped the clinician and was a priori biased in favour of success by the formula. He pointed out that clinicians gain large amounts of information and formulate impressions from *face to face* contact with patients. A realistic comparison would have the clinician making their predictions in such a real world setting unhandicapped, with the statistical formula operating on whatever subset of the information available to the clinician could be practically made available to it. If the statistical approach could be shown to predict better than the clinician in such a real clinical environments, then could it be advocated that the statistical approach was better. Holt also objected to the false dichotomy created by this literature, suggesting that it was often framed as a one "winner", one "loser" contest. He reasoned that clinical decision making was a complex phenomenon and that it would not be surprising to find that statistical prediction and other mechanical decision tools could be incorporated into parts of the clinical decision making so as to improve clinical judgement, rather than supercede it.

Taking into account the criticism that the clinical prediction tasks studied were artificial, some researchers have sought to make comparisons in real world settings and in cases where the clinical decisions were real ones. For example Leli and Filskov [1984] studied neuropsychologists' diagnoses of progressive brain dysfunction based on psychometric testing and observation of the patient. They showed that experienced clinicians correctly identified 58% of new cases, but that a formula based decision rule using only the test scores correctly identified 83% of cases. A number of studies in a variety of clinical settings have also found statistical decision making to be

superior to clinical judgement [Dawes et al 1989, Faust & Ziskin 1988, Kleinmutz 1991].

Despite these findings, many clinicians find it incredible that their judgement can be outperformed by a formula. A major source of disbelief may be the clinicians' view that because they can directly observe patients they will always have superiority over statistical techniques. Holt [1986] posited that for the purposes of analysis it might be useful to subdivide clinical decision making into 2 parts: clinical information gathering and clinical judgement. He contended that it would be too difficult to incorporate the richness and variety of clinical observations into formulas. Thus he concluded that across the broad range of clinical decision making clinicians have the advantage and it is only in relatively few areas (where only mechanical information gathering is required), that statistical approaches have any validity.

Mechanical approaches to clinical information gathering (eg. checklists, psychometric tests, structured interviews, computerised data collection etc) have gained in popularity over the course of the debate. This has allowed researchers to study the relative efficacy of clinical and mechanical information gathering combined with different types of decision-making. Wiggins [1973] has reviewed these studies and concluded that the least predictive approach is the "pure clinical" approach, where information is gathered clinically and clinical judgement is exercised in decision-making. If the clinician incorporates mechanically collected data, but retains clinical judgement for decision-making, this improves predictive validity. If information is mechanically gathered and then coupled with statistical decision-making, then this improves predictive validity even further. Having both clinical and mechanical

information gathering, and then using statistical decision making, does not improve predictive validity beyond that of using mechanical information gathering only, coupled with statistical decision making. Thus, there is no direct evidence from studies that adding clinical to mechanical information gathering is helpful. However there is a common sense view that clinical information gathering may allow clinicians to identify exceptions that are inappropriate for the statistical method [Schwartz & Griffin 1986].

It is possible that clinical judgement is superior to statistical decision making in some instances that have yet to be studied, but as Kleinmuntz [1991] has pointed out there is not even one study that has found any instance of this. The empirical data to date clearly suggest the superiority of the statistical approach to decision making, especially when coupled with mechanical information gathering. According to this body of research, when confronted with a clinical decision problem such as: *Is this patient suffering from schizophrenia?*; *Does this patient have brain damage?*; or *Should this child presenting with symptoms of ADHD be trialled on stimulant medication?*, then those who use structured information gathering and then use a statistically derived formula to make that decision, are more likely to have drawn accurate conclusions about their patient, than clinicians who prefer to use unstructured information gathering and/or clinical judgement.

There are no surveys of clinical practices that can tell us how popular the different approaches are, but it would not be unreasonable to conclude that a large number of clinicians rely solely on clinical judgement, that an increasing number are adopting structured and automated decision making (e.g. DSM-IV) in diagnostic decisions, and

that only a small number are using statistical decision making. The non-adoption by clinicians of statistical decision making, even in those areas where its superiority has been demonstrated, presents a problem. If it really is better than clinical decision making in many instances, then why hasn't it become more popular? Probably, it is not so much that clinicians as a group have examined what statistical prediction has to offer, and then found it unsatisfactory, but that many clinicians have not yet been acquainted with statistical decision making techniques. Schwartz & Griffin [1986] postulate that clinical training emphasises beliefs and attitudes that run contrary to the adoption of statistical decision-making. As well, it may also be the case that clinicians in general do not have the mathematical knowledge and skills that are a prerequisite for implementing statistical decision-making. Clinical training by and large focuses upon the learning of clinical observation skills and places little emphasis on mathematics. However, this is changing. Some professions, such as clinical psychology, have recognised the importance of statistical methods in clinical assessment and are incorporating the acquisition of prerequisite mathematical skills in clinical training (see for example Ley [1972])

Despite the reluctance from clinicians to adopt statistical decision making, the controversy and all the efforts of the protagonists on both sides have had some positive effects. As Kleinmuntz [1991] has pointed out, the long running debates in the literature between Meehl and Holt have had the desirable side effect of stimulating research by cognitive psychologists into the human clinical decision making processes. This research is discussed in the next section.

1.2 Biases and Heuristics in Clinical Judgements

Why does statistical decision making appear to be superior? Kahneman and Tversky (e.g. Kahneman, Slovic and Tversky [1982]) have examined the biases and heuristics inherent in human judgements made under conditions of uncertainty (i.e. where the outcome is not known). Clinical judgement is one example of this. Summaries of this research and its implications for clinical decision making are contained in Dawes et al [1989], Achenbach [1985], Schwartz & Griffin [1986] and Arkes [1991].

Achenbach [1985] has summarised this research and describes biases "that affect most human reasoning but are especially crucial in clinical assessment". These are *illusory correlation*, whereby people assume correlations where there is no such correlation. An example of this bias is when clinicians assume that any disturbed behaviour exhibited by a patient (eg. verbal aggression) is attributable to their psychopathology (e.g. schizophrenia) rather than other causes (e.g. frustration with being confined to a psychiatric admission ward). *Inability to assess covariation* occurs when clinicians determine associations between clinical phenomena intuitively from a series of cases they have seen. It is actually very difficult to do so correctly, especially where there are differences of degree, for example, differences in severity. The *representative heuristic* is viewing a limited sample as representative of a larger group. This results from three errors of judgement, namely *insensitivity to base rates*, *insensitivity to sample size*, and *insensitivity to predictability*. Achenbach provides the following example of insensitivity to base rates. He found that 85% of parents of 6-year old boys described their sons as "hyperactive, restless". Even if the true prevalence of minimal brain dysfunction were as high as 10%, and even if hyperactivity were truly a

sign of minimal brain dysfunction, then we would still be more often wrong than right if we concluded from a parent's report of hyperactivity that the child had minimal brain dysfunction. Insensitivity to sample size is the failure to recognise that high correlations with small samples may well be due to chance. Insensitivity to predicability is a variation of illusory correlation. A well-known example is that despite the consistently poor correlation between interview performance and later job performance, appointments are often still strongly influenced by performance in the interview. The *availability heuristic* is a bias resulting from previous cases, which are mentally available to the clinician at the time of assessment. Such cases may influence us on account of their "vividness, recency, intensity of involvement, similarities of mannerism", etc. rather than their similarity in ways which would actually allow similar predictions of outcome. The *confirmatory bias* is the tendency to weigh up clinical data in a way that confirms our beliefs and to ignore data that disconfirms our beliefs. The *idiographic fallacy* is the unvalidated extrapolation from the individual to the population; and the *nomothetic fallacy* is the undue extrapolation from the population to the individual.

These biases which all conspire to reduce the overall performance of clinical decision making are, according to Kleinmuntz [1991], a successful adaptation of human beings in dealing with their environment. His theory is that these biases and heuristics were successful in dealing with most everyday problems with a minimum of cognitive effort. Achenbach [1985] shares this view and refers to all these biases and heuristics under the title of "Cognitive Economics of Clinical Thinking" (p. 11). If we approached all decision making with absolute rationality, it would require excessively large cognitive effort (e.g. calculating probabilities, performing observational

experiments that rule out confounding variables). By using cognitive shortcuts (judgement heuristics and biases), we successfully solve many problems with greatly reduced demands for cognitive processing. This has come to be known as the concept of "bounded rationality" [Klienmetz 1991]. While bounded rationality may have been adaptive for hunter-gatherers in their decision-making, it is not for clinicians. They deal with problems and decisions that are best solved without such cognitive short cuts.

1.3 Expert Systems & Structured Decision making

Traditional Artificial Intelligence research in medicine has focused on the development of Expert Systems, computer applications that provide expert advice in a very narrow area of expertise. The most well known of these is the MYCIN system [Shortliffe 1976]. It is a clinical expert system designed for the specific task of making diagnoses and recommending treatment for bacterial infections. The basic design strategy in developing such a system is to get clinicians to "think aloud" so that the designer (formally referred to as a knowledge engineer) can determine what are the decision rules used by the clinical expert(s) and then to encode those rules into a computer program, so that the Expert System behaves like the expert [Dayhoff 1990].

The development strategy for expert systems seems straightforward, but there have been problems with this approach. A major stumbling block has been that the extraction of decision rules from expert clinicians has proved to be extremely difficult. It is not that clinical experts have been unco-operative or unwilling to participate, but that they have often been simply unable to state explicitly the logic

they have employed to make their clinical decisions. An explanation of this [Schwartz & Griffin 1986] is that highly experienced clinicians have through experience learned to condense logically constructed chains of inference and reasoning into "compiled" knowledge. Experienced clinicians seem to just know how to interpret particular patterns of sign and symptoms. They are able to instantly recognise clinical patterns. An expert clinician might be able to recognise and discriminate hundreds or thousands of these patterns almost instantly and with little or no conscious reasoning and often in circumstances where only partial information is available to them. Thus knowledge engineers have found it difficult to extract the kind of logical reasoning rules required to program a Clinical Expert System because logical reasoning is not the main process used by expert clinicians to make their diagnoses and other clinical decisions. Clinical expertise involves, at least in some part, pattern recognition.

It is interesting to note that in terms of the clinical judgement versus statistical decision making debate, these expert systems, though computerised, are best characterised as clinical judgement, not statistical decision making. That is, they are machine emulations of a human clinician making a clinical decision based upon judgement. If all the attributes of the human expert are copied, including their biases and heuristics, then these computer programs will perform the same as the clinician does. The important defining characteristic of statistical decision-making is not that a formula is used or that computer performs the task, but that the decision is based on a statistically derived relationship [Dawes et al 1989].

Nurcombe & Gallagher [1986] have attempted to decompose the clinicial decision-making process in psychiatry into an explicit set of steps. They observe that “medical

experts find it difficult to explicitly describe how they reach a diagnosis or make other clinical decisions" (p xv), and that when pressed they ascribe it to art or intuition. Their proposed solution is to use principles of deductive reasoning to teach clinicians "how to think". However within this framework, deduction applies to decision-making, but not necessarily to information gathering, which can involve not only mechanical methods (such as checklists or laboratory tests) but also automated or statistical processing of information which gives a putative decision to the clinician which they can use as an input to deductive reasoning.

1.4 Conclusions

Clinical judgement is probably the most popular clinical decision making practice amongst mental health professionals at this time. The use of statistical decision making practices is on the increase and may at sometime in the future overtake the popularity of clinical judgement. There is a large and growing body of empirical evidence, which has accumulated over a period of 50 years, which supports this migration of practices. And there is also evidence that humans (including clinicians) are not cognitively optimised for making the sorts of decisions involved in clinical decision-making.

It could be expected, that given 50 years of consistently favourable empirical results, there would be much effort to develop statistical decision making practices and that the uptake of these practices by clinicians would be rapid. However, the slow development and uptake of statistical decision making practices by mental health

professionals indicates that there is a significant resistance amongst clinicians to the adoption of statistical decision making practices.