INTRODUCTION

PART I: THE ART AND SCIENCE OF UNCERTAINTY

Clinicians have traditionally held clinical judgment and decision making in high regard and been suspicious of attempts to explore them systematically with a view to making explicit their precise character. This is true even when such efforts are undertaken for the presumably laudable purpose of helping novices to acquire the ability to make good clinical judgments and decisions with minimal distress to themselves, their teachers and – not least – their patients. But it is especially true in relation to those who seem to wish to demystify for the sake of demystification, or hope thereby to facilitate the achievement of some such (apparently) non-clinical goal as economic efficiency or the redistribution of welfare in the community.

The suspicions remain widespread despite, or perhaps because of, the increased attention focussed on clinical judgment and decision making in recent years, from both inside and outside the medical profession. The broad reasons for this growing interest are fairly well appreciated. Internally, the growth of biomedical knowledge on the one hand and technological developments on the other have vastly expanded the range of investigative and therapeutic possibilities confronting the practitioner and those on the way to becoming practitioners: the clinical and educational interests reinforce one another because of the central place of clinical training in the profession. Externally, rising public expectations concerning both the length and quality of life have, along with the general growth of "consumerism", changed the attitudinal basis of professional-client relations. At the same time they have brought major changes in the institutional and legal contexts within which these relations occur. The profession has undoubtedly contributed in some measure to these "external" developments, because the growth of its technical knowledge and technological capability has been used as the basis for claims to increased funding and other resources. Much of the increased interest from outside therefore represents calls for greater accountability for professional judgments and decisions.

The conceptually separable questions of *how* clinicians make judgments and decisions and of *how well* they make them have both become of greater interest to a wide range of parties, the evaluative question being the dominant concern of those who see themselves as bearing the costs, monetary or personal, of inferior judgments and decisions. The wide variation in clinical practices uncovered by virtually all studies of clinical behaviour – whether the comparison is between colleagues, communities or countries – has been one empirical focus for those attempting to assess professional performance. The significant percentage of discrepancies between clinical diagnoses and pathological findings which emerges in most autopsy studies is another (Goldman *et al.* 1983 and references cited therein). While both types of study are always questionable on a variety of methodological grounds the idea that clinical judgment and decision making are in "perfect health" and that, at most, "fine tuning" is required, has become increasingly hard to sustain to independent observers. This criticism of everyday clinical judgment is complemented by a growing number of techniques and systems that claim to provide clinicians with the means to improve their judgments and decisions, accompanied by evidence that (under at least some circumstances) these claims are warranted.

It is not difficult to accept that attitudinal and behavioural changes in society at large have brought new uncertainties to the clinicians in their dealings with individual patients. It is more paradoxical, though only until the matter is given second thought, that increased biomedical knowledge and technological capability have increased rather than reduced the complexity and difficulty of the clinician's task. Eddy [1] sees contemporary medical practice as being saturated with uncertainties:

> Whether a physician is defining a disease, making a diagnosis, selecting a procedure, observing outcomes, assessing probabilities, assigning preferences, or putting it all together, he is walking on very slippery terrain. It is difficult for nonphysicians, and for many physicians, to appreciate how complex these tasks are, how poorly we understand them, and how easy it is for honest people to come to different conclusions. (p. 45)

After exploring the nature of the uncertainties present in each of the components into which he divides the clinical task, Eddy concludes that the management decision for a single patient requires, in principle, such a complex synthesis of imperfect information that it "would be an extremely hard task for a research team; there is no hope that it could occur with any precision in the head of a busy clinician" (p. 54). For him the appropriate clinical response is not the denial or avoidance of the uncertainties, reactions which far too often characterize practice and produce, among other things, conflicting policies defended with great

confidence by different clinicians. The uncertainties should be *explicitly* – indeed "scientifically" – confronted, using the relevant techniques and languages that have been developing alongside the biomedical knowledge that clinicians so admire and use.

> Over the past few hundred years languages have been developed for collecting and interpreting evidence (statistics), dealing with uncertainty (probability theory), synthesizing evidence and estimating outcomes (mathematics), and making decisions (economics and decision theory). These languages are not currently learned by most clinical policymakers; they should be. (p. 58)

On the contrary, Schön sees [2] clinical wisdom and expertise residing precisely in the ability to accomplish the tremendously difficult synthesizing task described by Eddy *without* resort to formal analysis. Expert clinicians move efficiently to the appropriate *resolution* of problems, where "resolution" encompasses the "framing" or "setting" of the problem as well as its "solving" once it has been framed. The framing of the client's problem is not amenable to the systematic, formal, analytical modelling Eddy recommends, even if the solving of already framed problems is. Eddy's call for greater openness and the public evaluation and justification of clinical policies (i.e. "what I do for my patients"), using the language of the decision sciences, comes up against Schön's belief that a substantial part of clinical expertise, and probably the most important part, comes in the form of "knowledge-*in*-action". In contrast to "knowledge-*for*-action" (which can be acquired from books or colleagues) this sort of knowledge is only revealed *in action* – in the doing of the task, whether it be tightrope walking or clinical diagnosis. There is no reason to expect the performer to be able to articulate it, or even for outside observers to be able to capture and make it into "knowledge-for-action" by others. Based on the protracted and unique personal experience of the person concerned, it only exists in that individual professional's actions.

It would be tempting and not altogether misleading to characterize Eddy as a pessimist about the quality of professional judgment and decision making and Schön as an optimist. Eddy certainly seems broadly unimpressed with the quality of clinical practice at *all* levels, including that of the highest experts: even the best clinicians are performing below levels of achievements which are feasible. If they were to face up fully to the difficulties of the clinical task and address the uncertainties that characterize all aspects of it much more explicitly and analytically, the overall quality of practice would be raised. For Schön, on the other hand, the best experts are *by definition* performing at the highest possible level (given the resource and other contexts of their practice), so it is only a question of bringing those who are not currently exhibiting "best prac-

tice" judgment and decision making up to this level. This can be done by increasing their "knowledge-in-action" and by encouraging and facilitating the other hallmark of the best clinicians, which is their ability to "reflect-in-action". This is not reflection *on* action *after* the action is over, but reflection on action *in* action, that is during the action.

> The practitioner allows himself to experience surprise, puzzlement, or confusion in a situation which he finds uncertain or unique. He reflects on the phenomena before him, and on the prior understandings which have been implicit in his behavior. He carries out an experiment which serves to generate both a new understanding of the phenomena and a change in the situation.
>
> When someone reflects-in-action, he becomes a researcher in the practice context. He is not dependent on the categories of established theory and technique, but constructs a new theory of the unique case . . . He does not keep means and ends separate, but defines them interactively as he frames a problematic situation. He does not separate thinking from doing, ratiocinating his way to a decision which he must later convert to action . . . Thus reflection-in-action can proceed, even in situations of uncertainty and uniqueness, because it is not bound by the dichotomies of Technical Rationality. (pp. 75–6)

It is difficult to pin down the precise nature of Schön's "reflection-in-action" (and hence its distinction from reflection *on* action), because its character is said to vary with the length of the "action present", "the zone of time in which action can still make a difference to the situation". This zone may vary from seconds to years depending on the type and duration of action (acute surgery, community geriatric care). It will also vary for one practitioner with different patients and for the same patient through time. Equally in need of clarification is the idea of "a theory of the unique case". But what is very clear is that for Schön the activity is essentially *craftlike* in character. He greatly respects what has traditionally been referred to as the artistic, or intuitive, aspect of clinical practice.

Are the manifold uncertainties of contemporary clinical practice to be seen as the legitimate basis for regarding clinical judgment and decision making as significantly – even "essentially" – artistic in character? Albeit, of course, an *art* that uses, and gains credibility from, the knowledge produced by the medical and other *sciences*. Or are claims to artistry just the way the profession dresses up its refusal to apply the same scientific approach to its own cognitive processes and behaviour that it insists upon in relation to processes and knowledge at the levels of organ and cell? Those, like Eddy, who take the latter view would insist that even if clinical practice *can* be usefully regarded as an art or craft this does not, and should not, prevent the results of that activity being evaluated scientifically or at least systematically. Those, like Schön, who are deeply impressed by the dangers of applying to human activity the positivism of the natural sciences, question the possibility of evaluating craft activities

"scientifically". The "variations in practice" detected by investigators are precisely what one would expect if individual clients' problems are being framed appropriately – and differently – even if by some "scientific" criterion the frames are equivalent.

The long-standing debate as to whether clinical practice is essentially artistic or scientific, to which Eddy and Schön may be regarded as recent contributors, is, fortunately, giving way to more productive and research-based discussion about the precise mixture of activities in which clinicians engage. From this it has become clear (Sober 1979) that if clinical judgment is an art it is not because it is a fundamentally non-logical, qualitative activity which is emotionally concerned with the unique individual. Clinical judgment can only be successful to the extent that it is also – implicitly if not explicitly – logical, quantitative, detached and statistical. Hamm [3] introduces a theoretical framework in which the dichotomy between "intuitive art" and "analytical science" gives way to a continuum of modes of practice, in which the nature of clinical inference and decision making varies with the structure and context of the task and the way of thinking adopted by the clinician in response to that structure and context.

Cognitive Continuum Theory, developed by Kenneth Hammond, identifies *intuitive judgment* and *scientific experiments* (of the sort done in physics and chemistry) as the poles of a continuum of modes of inquiry (or practice) which we, as humans, have more or less available to us. We select a mode for practice on the basis of the structural characteristics of the task facing us and the time and other resources available in the context. Broadly speaking the less well structured a task is – or is perceived to be – the more we will be induced to adopt the way of thinking (mode of cognition) in which *intuition* is dominant (i.e. intuitive judgment). The better structured the task is – or is perceived to be – the more we will be induced to draw on the mode in which analysis is dominant (i.e. scientific experiment).

While both task structure and cognitive mode are continuous variables, six characteristic combinations of them – six main modes of practice – are identified in Cognitive Continuum Theory. The intermediate modes are variously labelled, depending on whether the focus is on inference or decision, but may be generally referred to as "peer-aided judgment", "system-aided judgment", "quasi-experiment" (e.g. epidemiological studies) and "controlled trial". As we move through these modes from "intuitive judgment" to "scientific experiment" the possibility of manipulation by the clinician/researcher increases: at the intuitive judgment end one is engaged in developing a "theory of the single case" with little possibility of holding other things equal and observing the effect of changing one variable (e.g. *this* patient's age), precisely what a scientific

experiment offers. Increasing in parallel is the openness of his/her activity to examination and replication but also, of course, is the time (and usually other resourcing) necessary to carry out the manipulation. For the *clinician* the most relevant modes are therefore the three more intuitive ones and many of the papers in this collection may be interpreted as contributions to the debate about the "proper" or "best" mode of practice for clinical activity. In most cases they advance the case for moving towards more explicit peer-aided judgments (Mode 5) and to system-aided judgments (Mode 4), where "system" embraces such things as data-based aids, knowledge-based decision support systems and particular types of decision analysis.

It is important to note that Cognitive Continuum Theory is fundamentally descriptive, designed for empirical testing to see whether judges' modes of practice *are* substantially determined by the structure of the particular task they are undertaking – rather than by, for example, some personality-based preference for a particular blend of intuition and analysis irrespective of the task. But while Cognitive Continuum Theory identifies a number of attributes on which to "objectively" index task structure (among others, the number and redundancy of the cues available to the clinician) it is clearly possible that the *perceived* structure of a given situation will vary from person to person. Others would go further and argue that structures are never given (and perceived) but always *generated*, this being the view of those like Schön who hold problem-framing rather than problem-solving to be the heart of clinical judgment. Arguments in the real world about the appropriate mode of practice will then be arguments about *what* sort of task is being undertaken and how much time and how many resources *are* available to tackle it, rather than about the mode of cognition appropriate for an agreed type of task. Alternatively these arguments may be completely *normative* ones (e.g. about the cues and time and resources which *should* be available to clinicians), in which case the *descriptive* validity of the theory is irrelevant.

There is no suggestion in Cognitive Continuum Theory that clinicians' ability to draw correct inferences or arrive at optimal decisions for their patients is *guaranteed* by the adoption of the mode of cognition (and hence practice) appropriate to the task. While practitioners at all levels of expertise are seen as having each of the modes of cognition available in their repertoire, they may be more or less expert within any of them, be it more intuitive or more analytical. This is not so in the alternative model introduced by Hamm, that of Hubert and Stuart Dreyfus. They postulate only two practice modes, retaining a fundamental dichotomy between *intuition* and *analysis*, and, in contrast to the determination-by-task thrust of Cognitive Continuum Theory, argue that it is the clinician's level of expertise that determines whether an intuitive or analytical approach is

taken to the various components of the clinical task. Put at its simplest, a novice is one who tackles each of the clinical sub-processes (which they identify as "perception", "action", "orientation" and "decision") analytically and the expert is one who tackles all of them intuitively. Instead of intermediate modes of *practice* we have, in the Dreyfus schema, intermediate levels of *expertise*, through which clinicians graduate by increasing the number of sub-processes which they approach intuitively. (They can only do this having established a sound analytical ability at each first.)

There are clear echoes of Schön in the Dreyfus' assertion that "better thinking is done intuitively, because experts, who think better, think intuitively" (p. 99). Unfortunately, it is difficult to see any way of empirically testing such a proposition and the question of whether a valid *independent and objective* evaluation of expert performance in the clinical professions is possible underlies much of the tension between the artistic-intuitive and rational-analytic views of clinical reasoning. If an evaluative study shows experts performing less well than non-expert humans or some impersonal system such as an equation, is it, by conclusion, methodologically defective? This question reverberates through the papers in the next section.

PART 2: MODELLING THE CLINICIAN AND THE CLINICAL TASK

In this section we are concerned primarily with two major approaches to the description of *how* clinicians make judgments and decisions, although the question of *how good* these judgments and decisions are arises frequently, and so, by implication, does that of the appropriate normative standard for evaluation.

The policy-capturing or *clinical judgment analysis* approach to the modelling of clinician behaviour in a task seeks to capture, in the form of a mathematical equation, the relationship between the inputs (cues or indicants) available to the clinician and his/her outputs (judgments or decisions). It consciously makes no claim to model what goes on inside the clinician's head, which is officially treated as a "black box", though the temptation to make inferences about what is going on in there is not always easy to resist. The *process-tracing approach* has, in its pure form, precisely the reverse aim: to model only what goes on inside the clinician's head. Since, however, such modelling is of little *practical* use in either educational or decision support development if it is unable to predict the clinician's behaviour reasonably well, the pure forms of process tracing have tended to give way (e.g. in computer simulations) to modes which incorporate intelligent reasoning of various sorts, whether or not it is reported by the clinician. Elstein and Bordage [4] provide introductory summaries of the earlier literature in these two traditions.

Statistical approaches

The judgment tradition, adopting the theoretical framework provided by Egon Brunswik's lens model, has been remarkably successful in showing that the intuitive judgments of clinicians can be successfully reproduced by simple (linear, additive) equations. In the absence of gold standard verdicts this sort of analysis is restricted to comparing the cue weights captured from a clinician's behaviour with the weights they *say* they *think* they give the cues, or comparing them with a colleague's captured weights. Significant differences between captured and expressed weights and between colleagues' captured weights (even when they report the same expressed weights) are typically found.

Wigton, Hoellerich and Patil [5] found little variation in the captured group average weights of faculty members, house officers and students in diagnosing simulated cases of pulmonary embolism, but great individual variation within each group. Most surprising to them, they found no convergence towards a consensus weighting as experience increased – surprising because they had deliberately selected a task where definitive laboratory tests were lacking and clinical experience would therefore be the main basis of clinical judgment. The paper canvasses the possible reasons for its findings, assuming that they are not experimental artifacts, but also notes (in relation to that assumption) the significant methodological problems presented by all studies based on case vignettes ("paper patients"), and the possibility that the variations in diagnostic policy had little or no therapeutic significance.

It is the two latter responses – "I don't believe the results would hold in real clinical life" and "So what, it doesn't matter for management" – that have dominated clinicians' reactions to the results coming from policy-capturing studies, particularly those which have brought in gold standard verdicts for each case and thereby introduced evaluations as well as comparisons. The performance of a judge can, in these circumstances, be compared with that of the equation captured from his behaviour. If the equation outperforms the judge (has a higher number of correct diagnoses, roughly speaking) then the judge can "bootstrap" himself – achieve a better performance – by replacing himself with this equation. The equation probably does better because it eliminates the "noise" in his judgments that arises from very "human" variations in mood and attention, while retaining his consistent "signal" (i.e. the judge's policy, in the terminology of this literature).

But if a gold standard verdict exists it is possible to model the relationship between it and the cues directly, leaving out the judge entirely. If the conclusion that judges could usually be bootstrapped was controversial, vastly more so was the oft-repeated finding that a purely

statistical modelling of clinical tasks would outperform the judge (and even his captured, "bootstrapping", equation). The demonstrated superiority of statistical over clinical judgment – of Mode 3 over Mode 6 in Cognitive Continuum terms – became the centre of intense methodological debate. The clinical fraternity and their defenders claimed that the competition was being set up in a way that ensured "the computer" would win. Cues were presented in clusters rather than sequentially, and as pre-coded information rather than as part of the texture of the real world. The judgments asked of the clinician had to be supplied in quantitative form; and so on. Whatever the validity of these objections a deeper and somewhat more appealing explanation eventually emerged, ironically from those responsible for many of the original findings.

This deeper explanation is discussed by Dawes [6] in a paper which expands on his claim that

> in a wide variety of psychological contexts, systematic decisions based on a few explicable and defensible principles are superior to intuitive decisions – because they work better, because they are not subject to conscious or unconscious biases on the part of the decision maker, because they can be explicated and debated, and because their basis can be understood by those most affected by them. (p. 151)

(The empirical illustration in this case is graduate school admission procedures, but the arguments transfer to many other clinical situations.)

Why do regression equations "work better" than clinical judgments? Dawes suggests that it is because most of the clinical tasks we undertake are ones in which, as far as current knowledge permits us to know, the relationship between the cues we know to have some relevance and the thing we are predicting is "conditionally monotonic": a higher value on each of these cues is associated with a higher value on the variable being predicted, irrespective of the values of the other cues. This is precisely the statistical situation in which regression equations are very good predictors, so that the clinicians were right, though in a very different way from the one they thought, when they felt the competition favoured "the computer". It is the difficult tasks we are left to tackle clinically, rather than in a routine way based on secure causal knowledge, that are the ones in which "the computer", grinding out its regression equations, is more or less bound to win. In other words the clinician is the person left with the tasks that we know relatively little about – relative to their immense complexity – and her attempts to deal with these in a sophisticated way, on the basis of highly developed knowledge and skills, often lead to inferior performance and wasted mental energy. The consolation to the clinician is that it is only she who can establish the relevant set of non-redundant cues for the computer to work on.

What about the procedural and ethical arguments against system-aided

(or system-*made*) judgments, which are invariably raised when the output performance case has had to be conceded? Dawes points out that the defence that clinical judgment avoids "dehumanizing judgment by numbers" is one that needs to be examined carefully in the light of the tradeoffs between process and outcome, as well as wider considerations of justice. In this examination is it really the clinician who should be deciding whether a "more human process" is worth the increased number of inferior decisions or unfairnesses it can produce compared with the regression equation?

Clinicians have great difficulty accepting that their expansive and complex technical knowledge does not necessarily guarantee that they can do better, in most clinical cases that come to them, than something as simple as "add up how many cues are in favour of each possible judgment and go with the highest score". Arkes, Dawes and Christensen [7] raise the alarming possibility that resistance to the use of simple decision rules, which cannot be outperformed because existing knowledge does not allow it, increases with expertise and the size of the stake. It seems to be precisely when highly qualified experts are engaged in tasks where the consequences of error are rated extremely serious that we will find most reluctance to use a simple mechanical rule – and end up with a higher than necessary number of errors as a result.

Einhorn [8] looks for a wider framework within which to interpret the persistence of the clinical versus statistical controversy, despite the one-sided results. Arkes *et al.* set up their experimental task in such a way that there was no way of improving on the performance of their simple decision rule, which in fact had a 70 percent success rate. Implicitly many of their respondents, and particularly the more expert, refused to believe that they could not do better than this, an attitude which reflects a wider belief that even if we cannot predict perfectly now, perfect prediction should be our goal. It is unacceptable to "settle for" a given number of errors (in this case 30 percent) when it is *possible* that 100 percent success may be achieved. Insofar as such an attitude motivates search for greater knowledge it may indeed further the movement towards the goal of perfect prediction *in the long run*, but will an increased number of current errors result from pursuing that possibility? For the "statisticians" the empirical evidence is that there is *at least* a short-run cost of this sort to be paid – and a long-run one too if the world is *not* really determinate or determinable. Einhorn suggests that the clinical versus statistical controversy is therefore between people who hold fundamentally different beliefs about the degree of predictability of that part of the world about which they are making judgments and/or have very different valuations of the two sorts of error which one may make: treating a situation which is "truly" random as being systematic and predictable and treating a