#13

Clinical Versus Actuarial Judgment

ROBYN M. DAWES, DAVID FAUST, PAUL E. MEEHL

Professionals are frequently consulted to diagnose and predict human behavior; optimal treatment and planning often hinge on the consultant's judgmental accuracy. The consultant may rely on one of two contrasting approaches to decision-making—the clinical and actuarial methods. Research comparing these two approaches shows the actuarial method to be superior. Factors underlying the greater accuracy of actuarial methods, sources of resistance to the scientific findings, and the benefits of increased reliance on actuarial approaches are discussed.

A PSYCHIATRIC PATIENT DISPLAYS AMBIGUOUS SYMPTOMS. Is this a condition best treated by psychotherapy alone or might it also require an antipsychotic medication with occasionally dangerous side effects? An elderly patient complains of memory loss but neurologic examination and diagnostic studies are equivocal. The neuropsychologist is asked to administer tests to help rule out progressive brain disease. A medical work-up confirms a patient's worst fears: he has terminal cancer. He asks the doctor how long he has to put his life in order.

These three brief scenarios illustrate a few of the many situations in which experts are consulted to diagnose conditions or to predict human outcomes. Optimal planning and care often hinge on the consultant's judgmental accuracy. Whether as physicians, psychiatrists, or psychologists, consultants perform two basic functions in decision-making: they collect and interpret data. Our interest here is in the interpretive function, specifically the relative merits of clinical versus actuarial methods.

Methods of Judgment and Means of Comparison

In the clinical method the decision-maker combines or processes information in his or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rest solely on empirically established relations between data and the condition or event of interest. A life insurance agent uses the clinical method if data on risk factors are combined through personal judgment. The agent uses the actuarial method if data are entered into a formula, or tables and charts that contain empirical information relating these background data to life expectancy.

Clinical judgment should not be equated with a clinical setting or a clinical practitioner. A clinician in psychiatry or medicine may use the clinical or actuarial method. Conversely, the actuarial method should not be equated with automated decision rules alone. For example, computers can automate clinical judgments. The computer can be programmed to yield the description "dependency traits," just as the clinical judge would, whenever a certain response appears on a psychological test. To be truly actuarial, interpretations must be both automatic (that is, prespecified or routinized) and based on empirically established relations.

Virtually any type of data is amenable to actuarial interpretation. For example, interview observations can be coded quantitatively (patient appears withdrawn: [1] yes, [2] no). It is thereby possible to incorporate qualitative observations and quantitative data into the predictive mix. Actuarial output statements, or conclusions, can address virtually any type of diagnosis, description, or prediction of human interest.

The combination of clinical and actuarial methods offers a third potential judgment strategy, one for which certain viable approaches have been proposed. However, most proposals for clinical-actuarial combination presume that the two judgment methods work together harmoniously and overlook the many situations that require dichotomous choices, for example, whether or not to use an antipsychotic medication, grant parole, or hospitalize. If clinical and actuarial interpretations agree, there is no need to combine them. If they disagree, one must choose one or the other. If clinical interpretation suggests brain damage but the actuarial method indicates otherwise, one does not conclude that the patient is and is not brain damaged.

Although some research appeared on clinical and actuarial judgment before the mid-fifties, Meehl (*1*) introduced the issue to a broad range of social scientists in 1954 and stimulated a flurry of studies. Meehl specified conditions for a fair comparison of the two methods.

First, both methods should base judgments on the same data. This condition does not require that clinical judge and statistical method, before comparison, use the same data to derive decision strategies or rules. The clinician's development of interpretive strategies depends on prior experience and knowledge. The development of actuarial methods requires cases with known outcome. The clinical and actuarial strategies may thus be derived from separate or overlapping data bases, and one or the other may be based on more or fewer cases or more or less outcome information. For example, the clinician may have interpreted 1000 intelligence tests for indications of brain dysfunction and may know the outcome for some of these cases based on radiologic examination. The actuarial method may have been developed on the subset of these 1000 cases for which outcome is known.

Second, one must avoid conditions that can artificially inflate the accuracy of actuarial methods. For example, the mathematical procedures (such as regression analysis or discriminant analysis) used to develop statistical actuarial decision rules may capitalize on chance (nonrepeating) relations among variables. Thus, derivation

R. M. Dawes is head of the Department of Social and Decision Sciences and professor of psychology, Carnegie Mellon University, Pittsburgh, PA 15213. D. Faust is director of psychology, Rhode Island Hospital, Providence, RI 02903, and associate professor, Department of Psychiatry and Human Behavior, Brown University Program in Medicine. P. E. Meehl is Regents' Professor of Psychology and professor of psychiatry and philosophy of science, University of Minnesota, Minneapolis, MN 55455.

typically should be followed by cross-validation, that is, application of the decision rule to new or fresh cases, or by a standard statistical estimate of the probable outcome of cross-validation. Cross-validation counters artificial inflation in accuracy rates and allows one to determine, realistically, how the method performs. Such application is essential because a procedure should be shown to work where it is needed, that is, in cases in which outcome is unknown. If the method is only intended for local use or in the setting in which it was developed, the investigator may partition a representative sample from that setting into derivation and cross-validation groups. If broader application is intended, then new cases should be representative of the potential settings and populations of interest.

Results of Comparative Studies

The three initial scenarios provide examples of comparative studies. Goldberg studied the distinction between neurosis and psychosis based on the Minnesota Multiphasic Personality Inventory (MMPI), a personality test commonly used for such purposes (*2, 3*). This differential diagnosis is of practical importance. For example, the diagnosis of psychosis may lead to needed but riskier treatments or to denial of future insurance applications. Goldberg derived various decision rules through statistical analysis of scores on 11 MMPI scales and psychiatric patients' discharge diagnoses. The single most effective rule for distinguishing the two conditions was quite simple: add scores from three scales and then subtract scores from two other scales. If the sum falls below 45, the patient is diagnosed neurotic; if it equals or exceeds 45, the patient is diagnosed psychotic. This has come to be known as the "Goldberg Rule."

Goldberg next obtained a total of 861 new MMPIs from seven different settings, including inpatient and outpatient services from either medical school, private, or Veterans Administration hospital systems in California, Minnesota, and Ohio. The accuracy of the decision rules when applied to these new cases was compared with that of 29 judges who analyzed the same material and attempted the same distinction. Some of the judges had little or no prior experience with the MMPI and others were Ph.D. psychologists with extensive MMPI experience.

Across the seven settings, the judges achieved mean validity coefficients ranging from $r = 0.15$ to 0.43, with a total figure of 0.28 for all cases, or 62% correct decisions. The single best judge achieved an overall coefficient of 0.39, or 67% correct decisions. In each of the seven settings, various decision rules exceeded the judges' mean accuracy level. The Goldberg Rule performed similarly to the judges in three of the settings and demonstrated a modest to substantial advantage in four of the settings (where the rule's validity coefficient exceeded that of the judges by 0.16 to 0.31). For the total sample, the Goldberg Rule achieved a validity coefficient of 0.45, or 70% correct decisions, thereby exceeding both the mean accuracy of the 29 judges and that of the single best judge.

Rorer and Goldberg then examined whether additional practice might alter results. Judges were given MMPI training packets consisting of 300 new MMPI profiles with the criterion diagnosis on the back, thus providing immediate and concrete feedback on judgmental accuracy. However, even after repeated sessions with these training protocols culminating in 4000 practice judgments, none of the judges equaled the Goldberg Rule's 70% accuracy rate with these test cases. Rorer and Goldberg finally tried giving a subset of judges, including all of the experts, the outcome of the Goldberg Rule for each MMPI. The judges were free to use the rule when they wished and knew its overall effectiveness. Judges generally made modest gains in performance but none could match the rule's accuracy; every judge would have done better by always following the rule.

In another study using the same 861 MMPI protocols, Goldberg constructed mathematical (linear) models of each of the 29 judges that reproduced their decisions as closely as possible (*4*). Modeling judges' decisions requires no access to outcome information. Rather, one analyzes relations between the information available to the judge and the judge's decisions. In principle, if a judge weights variables with perfect consistency or reliability (that is, the same data always lead to the same decision), the model will always reproduce that judge's decisions. In practice, human decision-makers are not perfectly reliable and thus judge and model will sometimes disagree. Goldberg found that in cases of disagreement, the models were more often correct than the very judges on whom they were based. The perfect reliability of the models likely explains their superior performance in this and related studies (*5*).

Leli and Filskov studied the diagnosis of progressive brain dysfunction based on intellectual testing (*6*). A decision rule derived from one set of cases and then applied to a new sample correctly identified 83% of the new cases. Groups of inexperienced and experienced clinicians working from the same data correctly identified 63% and 58% of the new cases, respectively. In another condition, clinicians were also given the results of the actuarial analysis. Both the inexperienced and experienced clinicians showed improvement (68% and 75% correct identifications, respectively), but neither group matched the decision rule's 83% accuracy. The clinicians' improvement appeared to depend on the extent to which they used the rule.

Einhorn (*7*) studied the prediction of survival time following the initial diagnosis of Hodgkin's disease as established by biopsy. At the time of the study, survival time was negatively correlated with disease severity (Hodgkin's is now controllable). All of the 193 patients in the study subsequently died, thus tragically providing objective outcome information.

Three pathologists, one an internationally recognized authority, rated the patients' initial biopsy slides along nine histological dimensions they identified as relevant in determining disease severity and also provided a global rating of severity. Actuarial formulas were developed by examining relations between the pathologists' ratings and actual survival time on the first 100 cases, with the remaining 93 cases used for cross-validation and comparison. The pathologists' own judgments showed virtually no relation to survival time; cross-validated actuarial formulas achieved modest but significant relations. The study revealed more than an actuarial advantage. It also showed that the pathologists' ratings produced potentially useful information but that only the actuarial method, which was based on these ratings, tapped their predictive value.

Additional research. These three studies illustrate key features of a much larger literature on clinical versus actuarial judgment. First, the studies, like many others, met the previously specified conditions for a fair comparison.

Second, the three studies are representative of research outcomes. Eliminating research that did not protect sufficiently against inflated results for actuarial methods, there remain nearly 100 comparative studies in the social sciences. In virtually every one of these studies, the actuarial method has equaled or surpassed the clinical method, sometimes slightly and sometimes substantially (*8–10*). For example, in Watley and Vance's study on the prediction of college grades the methods tied (*11*); in Carroll *et al.*'s study on the prediction of parole violation, the actuarial method showed a slight to modest advantage (*12*); and in Wittman's study on the prediction of response to electroshock therapy, the actuarial method was correct almost twice as often as the clinical method (*13*).

The earlier comparative studies were often met with doubts about

validity and generalization. It was claimed, for example, that the studies misrepresented the clinical method either by denying judges access to crucial data sources such as interviews, by using artificial tasks that failed to tap their areas of expertise, or by including clinicians of questionable experience or expertise.

The evidence that has accumulated over the years meets these challenges. First, numerous studies have examined judgments that are not artificial but common to everyday practice and for which special expertise is claimed. Examples include the three studies described above, which involved the differential between less serious and major psychiatric disorder, the detection of brain damage, and the prediction of survival time. Other studies have examined the diagnosis of medical versus psychiatric disorder (*14*); the description or characterization of personality (*15*); and the prediction of treatment outcome (*16*), length of psychiatric hospitalization (*17*), and violent behavior (*18*). These are decisions that general practitioners or specialists often address, and in a number of studies investigators did not introduce judgment tasks that clinicians then performed, but rather examined decisions already made in the course of everyday practice.

Other studies have provided clinicians or judges with access to preferred sources of information. Even in 1966, Sawyer was able to locate 17 comparisons between actuarial and clinical judgment based on the results of psychological testing and interview (*8*). Other investigators have allowed judges to collect whatever data they preferred in whatever manner they preferred. In Carroll *et al.*'s naturalistic study on the prediction of parolees' behavior after release, the parole board did not alter the data collection procedures (*12*). In Dawes's study on the prediction of graduate student performance, the admissions committee relied on the same data normally used to reach decisions (*19*). None of the 17 comparisons reviewed by Sawyer and neither the study by Carroll *et al.* nor Dawes favored clinical over actuarial judgment.

Nor has the outcome varied within or across studies involving judges at various levels of experience or expertise. In Goldberg's study novice and experienced MMPI interpreters performed similarly when using the clinical method and neither group surpassed the actuarial method, results parallel to those of Leli and Filskov in their study on the detection of brain damage (*2, 6*). Other studies on the detection and localization of brain damage have yielded similar results (*20, 21*). For example, Wedding found that neither clinicians with extensive experience interpreting the tests under study nor a nationally prominent neuropsychologist surpassed the overall accuracy of actuarial methods in determining the presence, location, and cause of brain damage (*20*).

The comparative studies often do not permit general conclusions about the superiority of one or another specific actuarial decision rule. Some studies, such as Goldberg's, do show application across settings, but much of the research has involved restricted samples. Investigators have been less interested in a specific procedure's range of application than in performing an additional test of the two methods and thereby extending the range of comparative studies.

The various studies can thus be viewed as repeated sampling from a universe of judgment tasks involving the diagnosis and prediction of human behavior. Lacking complete knowledge of the elements that constitute this universe, representativeness cannot be determined precisely. However, with a sample of about 100 studies and the same outcome obtained in almost every case, it is reasonable to conclude that the actuarial advantage is not exceptional but general and likely encompasses many of the unstudied judgment tasks. Stated differently, if one poses the query: "Would an actuarial procedure developed for a particular judgment task (say, predicting academic success at my institution) equal or exceed the clinical method?", the available research places the odds solidly in favor of an affirmative reply. "There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly . . . as this one" (*9*, p. 373).

Possible exceptions. If fair comparisons consistently favor the actuarial method, one may then reverse the impetus of inquiry and ask whether there are certain circumstances in which the clinical judge might beat the actuary. Might the clinician attain superiority if given an informational edge? For example, suppose the clinician lacks an actuarial formula for interpreting certain interview results and must choose between an impression based on both interview and test scores and a contrary actuarial interpretation based on only the test scores. The research addressing this question has yielded consistent results (*8, 10, 22*). Even when given an information edge, the clinical judge still fails to surpass the actuarial method; in fact, access to additional information often does nothing to close the gap between the two methods.

It is not difficult to hypothesize other circumstances in which the clinical judge might improve on the actuarial method: (i) judgments mediated by theories and hence difficult or impossible to duplicate by statistical frequencies alone, (ii) select reversal of actuarial conclusions based on the consideration of rare events or utility functions that are not incorporated into statistical methods, and (iii) complex configural relations between predictive variables and outcome (*23–25*).

The potential superiority of theory-mediated judgments over conclusions reached solely on the basis of empirical frequencies may seem obvious to those in the "hard" sciences. Prediction mediated by theory is successful when the scientist has access to the major causal influences, possesses accurate measuring instruments to assess them, and uses a well-corroborated theory to make the transition from theory to fact (that is, when the expert has access to a specific model). Thus, although most comparative research in medicine favors the actuarial method overall, the studies that suggest a slight clinical advantage seem to involve circumstances in which judgments rest on firm theoretical grounds (*26*).

The typical theory that underlies prediction in the social sciences, however, satisfies none of the needed conditions. Prediction of treatment response or violent behavior may rest on psychodynamic theory that permits directly contradictory conclusions and lacks formal measurement techniques. Theory-mediated judgments may eventually provide an advantage within psychology and other social sciences, but the conditions needed to realize this possibility are currently but a distant prospect or hope.

Clinicians might be able to gain an advantage by recognizing rare events that are not included in the actuarial formula (due to their infrequency) and that countervail the actuarial conclusion. This possibility represents a variation of the clinical-actuarial approach, in which one considers the outcome of both methods and decides when to supersede the actuarial conclusion. In psychology this circumstance has come to be known as the "broken leg" problem, on the basis of on an illustration in which an actuarial formula is highly successful in predicting an individual's weekly attendance at a movie but should be discarded upon discovering that the subject is in a cast with a fractured femur (*1, 25*). The clinician may beat the actuarial method if able to detect the rare fact and decide accordingly. In theory, actuarial methods can accommodate rare occurrences, but the practical obstacles are daunting. For example, the possible range of intervening events is infinite.

The broken leg possibility is easily studied by providing clinicians with both the available data and the actuarial conclusion and allowing them to use or countervail the latter at their discretion. The limited research examining this possibility, however, all shows that greater overall accuracy is achieved when clinicians rely uniformly on actuarial conclusions and avoid discretionary judgments (*3, 8*).

When operating freely, clinicians apparently identify too many exceptions," that is, the actuarial conclusions correctly modified are outnumbered by those incorrectly modified. If clinicians were more conservative in overriding actuarial conclusions they might gain an advantage, but this conjecture remains to be studied adequately.

Consideration of utilities raises a related possibility. Depending on the task, certain judgment errors may be more serious than others. For example, failure to detect a condition that usually remits spontaneously may be of less consequence than false identification of a condition for which risky treatment is prescribed. The adjustment of decision rules or cutting scores to reduce either false-negative or false-positive errors can decrease the procedure's overall accuracy but may still be justified if the consequences of these opposing forms of error are unequal. As such, if the clinician's counter-actuarial judgments, although less likely than the actuarial to be correct, were shown empirically to lower the probability of the rule's deliverances being correct (say, from 0.8 to 0.6), then in some contexts consideration of the joint probability-utility function might rationally reverse the action suggested by reliance on the formula alone. This procedure is formally equivalent to putting the clinician's judgment (as a new variable) into the actuarial equation, and more evidence on this process is needed to adequately appraise its impact. Here again, one cannot assume that the clinician's input helps. The available research suggests that formal inclusion of the clinician's input does not enhance the accuracy, nor necessarily the utility, of the actuarial formula and that informal or subjective attempts at adjustment can easily do more harm than good (*8*).

The clinician's potential capacity to capitalize on configural patterns or relations among predictive cues raises two related but separable issues that we will examine in order: the capacity to recognize configural relations and the capacity to use these observations to diagnose and predict. Certain forms of human pattern recognition still cannot be duplicated or equaled by artificial means. The recognition of visual patterns has challenged a generation of researchers in the field of artificial intelligence. Humans maintain a distinct advantage, for example, in the recognition of facial expressions. Human superiority also exists for language translation and for the invention of complex, deep-structure theories. Thus, for example, only the human observer may recognize a particular facial expression or mannerism (the float-like walk of certain schizophrenic patients) that has true predictive value. These observational abilities provide the potential for gathering useful (predictive) information that would otherwise be missed.

The possession of unique observational capacities clearly implies that human input or interaction is often needed to achieve maximal predictive accuracy (or to uncover potentially useful variables) but tempts us to draw an additional, dubious inference. A unique capacity to observe is not the same as a unique capacity to predict on the basis of integration of observations. As noted earlier, virtually any observation can be coded quantitatively and thus subjected to actuarial analysis. As Einhorn's study with pathologists and other research shows, greater accuracy may be achieved if the skilled observer performs this function and then steps aside, leaving the interpretation of observational and other data to the actuarial method (*7*).

Factors Underlying the Superiority of Actuarial Methods

Contrasts between the properties of actuarial procedures and clinical judgment help to explain their differing success (*27*). First, actuarial procedures, unlike the human judge, always lead to the same conclusion for a given data set. In one study rheumatologists's and radiologists's reappraisals of cases they themselves had evaluated previously often resulted in different opinions (*28*). Such factors as fatigue, recent experience, or seemingly minor changes in the ordering of information or in the conceptualization of the case or task can produce random fluctuations in judgment (*29*). Random fluctuation decreases judgmental reliability and hence accuracy. For example, if the same data lead to the correct decision in one case but to a different, incorrect decision in the second case, overall accuracy will obviously suffer.

Perhaps more importantly, when properly derived, the mathematical features of actuarial methods ensure that variables contribute to conclusions based on their actual predictive power and relation to the criterion of interest. For example, decision rules based on multiple regression techniques include only the predictive variables and eliminate the nonpredictive ones, and they weight variables in accordance with their independent contribution to accurate conclusions. These achievements are essentially automatic with actuarial prediction but present formidable obstacles for human judges.

Research shows that individuals have considerable difficulty distinguishing valid and invalid variables and commonly develop false beliefs in associations between variables (*30*). In psychology and psychiatry, clinicians often obtain little or no information about the accuracy of their diagnoses and predictions. Consultants asked to predict violence may never learn whether their predictions were correct. Furthermore, clinicians rarely receive immediate feedback about criterion judgments (for example, diagnoses) of comparable validity to that physicians obtain when the pathologist reports at the end of a clinicopathological conference (*31*). Lacking sufficient or clear information about judgmental accuracy, it is problematic to determine the actual validity, if any, of the variables on which one relies. The same problem may occur if actuarial methods are applied blindly to new situations or settings without any performance checks.

In other circumstances, clinical judgments produce "self-fulfilling prophecies." Prediction of an outcome often leads to decisions that influence or bias that outcome (*32*). An anecdote illustrates this problem. A psychiatrist in a murder trial predicted future dangerousness, and the defendant was sentenced to death. While on death row the defendant acted violently, which appeared to support the psychiatrist's predictive powers. However, once sentenced to death this individual had little to lose; he may have acted differently had the psychiatrist's appraisal, and in turn the sentence, been different.

Additionally, known outcomes seem more predictable than they are in advance (*33*), and past predictions are mistakenly recalled as overly consistent with actual outcomes (*34, 35*). For example, Arkes *et al.* presented the same case materials to groups of physicians and asked them to assign probabilities to alternate diagnoses. When probabilities were assigned in foresight, each diagnosis was considered about equally likely. However, when the physicians were informed that one or another diagnosis had been established previously and they were then asked to state what initial diagnosis they likely would have made, they assigned the highest probability to whatever diagnosis they were told had been established (*36*). If one's view or recall of initial judgments is inadvertently shaped to fit whatever happens to occur, outcome information will have little or no corrective value.

The clinician is also exposed to a skewed sample of humanity and, short of exposure to truly representative samples, it may be difficult, if not impossible, to determine relations among variables. For example, suppose that about half of the adolescents appraised for a history of juvenile delinquency show subtle electroencephalographic (EEG) abnormalities. Based on these co-occurrences, the clinician may come to consider EEG abnormality a sign of delinquency or may conclude that delinquency is associated with brain dysfunction.

In fact, clinicians have often postulated these relations (*37*).

One cannot determine, however, whether a relation exists unless one also knows whether the sign occurs more frequently among those with, versus those without, the condition. For example, to determine whether EEG abnormality is associated with delinquency, one must also know the frequency with which delinquents do not obtain EEG abnormalities and the frequencies with which nondelinquents do and do not obtain EEG abnormalities. Further, even should a valid relation exist, one cannot determine the sign's actual utility unless one knows: (i) how much more frequently it occurs when the condition is present than when it is absent and (ii) the frequency of the condition. For example, a sign that is slightly more common among those with the condition may be of little diagnostic utility. If the condition is infrequent, then positive identifications based on the sign's presence can even be wrong in most cases, for most individuals who display the sign will not have the condition. If 10% of brain-damaged individuals make a particular response on a psychological test and only 5% of normals, but nine of ten clinic patients are not brain-damaged, most patients who show the feature will not be brain-damaged.

In practice, the clinician is far more likely to evaluate individuals with significant problems than those without them, and this skewed exposure hinders attempts to make all of the needed comparisons. In fact, empirical study shows that EEG "abnormalities" are common among normal children and further suggests that the incidence of delinquency is no greater among those with than without neurological disorder (*37, 38*). The formation of such false beliefs is further compounded by a decided human tendency to overattend to information consistent with one's hypotheses and to underattend to contradictory information (*39*). The result is that mistaken beliefs or conclusions, once formed, resist counterevidence. Error is also fostered by a tendency to disregard frequency data and instead to form diagnostic judgments based on the perceived match between one or more of the presenting symptoms (for example, EEG abnormality) and some prototype or instance of the diagnostic category (delinquency) stored in memory (*40, 41*).

The same factors that hinder the discovery of valid relations also promote overconfidence in clinical judgment. When the clinician misinterprets contrary evidence as indicative of judgmental accuracy, confidence will obviously be inflated. Research shows that judges are typically more confident than their accuracy warrants (*42*). In one study demonstrating the upper range of misappraisal, most clinicians were quite confident in their diagnosis although not one was correct (*43*).

The difficulty in separating valid and invalid variables on the basis of clinical experience or judgment is demonstrated in many studies examining diagnostic or predictive accuracy (*44*). Research shows that clinical judgments based on interviews achieve, at best, negligible accuracy or validity (*12*). Other studies show that clinical judgments based on psychological test results may be of low absolute validity (*6, 18, 20, 21*). Although clinical interviews or psychological tests can produce useful information, the clinical judge often cannot distinguish what is useful from what is useless. In all studies cited immediately above, statistical analysis of the same data uncovered useful variables or enhanced predictive accuracy.

The optimal weighting of variables is a less important advantage of the statistical method than is commonly assumed. In fact, unit (equal) weights yield predictions that correlate highly with those derived from optimally weighted composites, the only provisos being that the direction in which each predictor is related to the criterion can be specified beforehand and the predictors not be negatively correlated with each other (*5, 45–47*). Further, optimal weights are specific to the population in which they were derived, and any advantage gained in one setting may be lost when the same method is applied in another setting. However, when optimal weighting adds meaningfully to predictive accuracy, the human judge is at a decided disadvantage. As Meehl (*9*, p. 372) has stated:

> Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. There are no strong arguments . . . from empirical studies . . . for believing that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently.

It might be objected that this analogy, offered not probatively but pedagogically, presupposes an additive model that a proponent of configural judgment will not accept. Suppose instead that the supermarket pricing rule were, "Whenever both beef and fresh vegetables are involved, multiply the logarithm of 0.78 of the meat price by the square root of twice the vegetable price"; would the clerk and customer eyeball that any better? Worse, almost certainly. When human judges perform poorly at estimating and applying the parameters of a simple or component mathematical function, they should not be expected to do better when required to weight a complex composite of these variables.

Lack of Impact and Sources of Resistance

Research on clinical versus statistical judgment has had little impact on everyday decision making, particularly within its field of origin, clinical psychology. Guilmette *et al.*'s survey showed that most psychologists specializing in brain damage assessment prefer procedures for which actuarial methods are lacking over those for which actuarial formulas are available (*48*). The interview remains the sine qua non of entrance into mental health training programs and is required in most states to obtain a license to practice (*49*). Despite the studies that show that clinical interpretation of interviews may have little or no predictive utility, actuarial interpretation of interviews is rarely if ever used, although it is of demonstrated value.

Lack of impact is sometimes due to lack of familiarity with the scientific evidence. Some clinicians are unaware of the comparative research and do not even realize an issue exists. Others still refer to earlier studies and claim that the clinician was handicapped, unaware of the subsequent research that has rendered these arguments counterfactual.

Others who know the evidence may still dismiss it based on tendentiousness or misconception. Mental health professionals' education, training, theoretical orientations and identifications, and personal values may dictate against recognition of the actuarial advantage. Some psychologists, for example, believe that the use of a predictive equation dehumanizes their clients. The position overlooks the human costs of increased error that may result.

A common anti-actuarial argument, or misconception, is that group statistics do not apply to single individuals or events. The argument abuses basic principles of probability. Although individuals and events may exhibit unique features, they typically share common features with other persons or events that permit tallied observations or generalizations to achieve predictive power. An advocate of this anti-actuarial position would have to maintain, for the sake of logical consistency, that if one is forced to play Russian roulette a single time and is allowed to select a gun with one or five bullets in the chamber, the uniqueness of the event makes the choice arbitrary.

Finally, subjective appraisal may lead to inflated confidence in the accuracy of clinical judgment and the false impression that the actuarial method is inferior. Derivation and cross-validation of an

actuarial method yields objective information on how well it does and does not perform (*50*). When the clinician reviews research that shows, for example, that the Goldberg Rule for the MMPI achieved 70% accuracy in a comparable setting and exceeded the performance of all 29 judges in the study, this may still seem to compare unfavorably to self-perceived judgmental powers. The immediacy and salience of clinical experience fosters the misappraisal. The clinician may recall dramatic instances in which his interpretations proved correct or in which he avoided error by countervailing an actuarial conclusion, failing to recognize or correctly tally counter instances.

Ultimately, then, clinicians must choose between their own observations or impressions and the scientific evidence on the relative efficacy of the clinical and actuarial methods. The factors that create difficulty in self-appraisal of judgmental accuracy are exactly those that scientific procedures, such as unbiased sampling, experimental manipulation of variables, and blind assessment of outcome, are designed to counter. Failure to accept a large and consistent body of scientific evidence over unvalidated personal observation may be described as a normal human failing or, in the case of professionals who identify themselves as scientific, plainly irrational.

Application of Actuarial Methods: Limits, Benefits, and Implications

The research reviewed in this article indicates that a properly developed and applied actuarial method is likely to help in diagnosing and predicting human behavior as well or better than the clinical method, even when the clinical judge has access to equal or greater amounts of information. Research demonstrating the general superiority of actuarial approaches, however, should be tempered by an awareness of limitations and needed quality controls.

First, although surpassing clinical methods, actuarial procedures are far from infallible, sometimes achieving only modest results. Second, even a specific procedure that proves successful in one setting should be periodically reevaluated within that setting and should not be applied to new settings mindlessly. Although theory and research suggest that the choice of predictive variables is often more important than their weighting, statistical techniques can be used to yield weights that optimize a procedure's accuracy when it is applied to new cases drawn from the same population. Moreover, accuracy can be easily monitored as predictions are made, and methods modified or improved to meet changes in settings and populations. Finally, efforts can be made to test whether new variables enhance accuracy.

When developed and used appropriately, actuarial procedures can provide various benefits. Even when actuarial methods merely equal the accuracy of clinical methods, they may save considerable time and expense. For example, each year millions of dollars and many hours of clinicians' valuable time are spent attempting to predict violent behavior. Actuarial prediction of violence is far less expensive and would free time for more productive activities, such as meeting unfulfilled therapeutic needs. When actuarial methods are not used as the sole basis for decisions, they can still serve to screen out candidates or options that would never be chosen after more prolonged consideration.

When actuarial methods prove more accurate than clinical judgment the benefits to individuals and society are apparent. Much would be gained, for example, by increased accuracy in the prediction of violent behavior and parole violation, the diagnosis of disorder, and the identification of effective treatment. Additionally, more objective determination of limits in knowledge or predictive power can prevent inadvertent harm. Should a confident but incorrect clinical diagnosis of Alzheimer's disease be replaced by a far more cautious statement, or even better by the correct conclusion, we would avoid much unnecessary human misery.

Actuarial methods are explicit, in contrast to clinical judgment, which rests on mental processes that are often difficult to specify. Explicit procedures facilitate informed criticism and are freely available to other members of the scientific community who might wish to replicate or extend research.

Finally, actuarial methods—at least within the domains discussed in this article—reveal the upper bounds in our current capacities to predict human behavior. An awareness of the modest results that are often achieved by even the best available methods can help to counter unrealistic faith in our predictive powers and our understanding of human behavior. It may well be worth exchanging inflated beliefs for an unsettling sobriety, if the result is an openness to new approaches and variables that ultimately increase our explanatory and predictive powers.

The argument that actuarial procedures are not available for many important clinical decisions does not explain failure to use existent methods and overlooks the ease with which such procedures can be developed for use in special settings. Even lacking any outcome information, it is possible to construct models of judges that will likely surpass their accuracy (*4, 5*). What is needed is the development of actuarial methods and a measurement assurance program that maintains control over both judgment strategies so that their operating characteristics in the field are known and an informed choice of procedure is possible. Dismissing the scientific evidence or lamenting the lack of available methods will prove much less productive than taking on the needed work.

REFERENCES AND NOTES

1. P. E. Meehl, *Clinical Versus Statistical Prediction* (Univ. of Minnesota Press, Minneapolis, MN, 1954).
2. L. R. Goldberg, *Psychol. Monogr.*, 79 (no. 9) (1965).
3. ———, *Am. Psychol.* 23, 483 (1968).
4. ———, *Psychol. Bull.* 73, 422 (1970).
5. R. M. Dawes and B. Corrigan, *ibid.* 81, 95 (1974).
6. D. A. Leli and S. B. Filskov, *J. Clin. Psychol.* 40, 1435 (1984).
7. H. J. Einhorn, *Organ. Behav. Human Perform.* 7, 86 (1972).
8. J. Sawyer, *Psychol. Bull.* 66, 178 (1966).
9. P. E. Meehl, *J. Personal. Assess.* 50, 370 (1986).
10. W. M. Grove, who is conducting the first formal meta-analysis of studies in the social sciences and medicine (and a few other areas) that compared clinical and actuarial judgement, has reported a preliminary analysis in simple "box score" terms (paper presented at the Annual Meeting of the Minnesota Psychological Association, Minneapolis, 8 May 1986). The clinical method showed an advantage in only 6 of 117 studies. These exceptions mainly involved the medical field. The clinical advantage was generally slight and the rarity of this outcome across so many comparisons raises the possibility that some or most of these exceptions were statistical artifacts.
11. D. J. Watley and F. L. Vance, U.S. Office of Education Cooperative Research Project No. 2022 (University of Minnesota, Minneapolis, MN 1974).
12. J. S. Carroll *et al.*, *Law Society Rev.* 17, 199 (1982).
13. M. P. Wittman, *Elgin Pap.* 4, 20 (1941).
14. S. Oskamp, *Psychol. Monogr.* 76 (no. 28) (1962).
15. C. C. Halbower, thesis, University of Minnesota, Minneapolis, MN (1955).
16. F. Barron, *J. Consult. Clin. Psychol.* 17, 233 (1953).
17. H. W. Dunham and B. M. Meltzer, *Am. J. Sociology* 52, 123 (1946).
18. P. D. Werner, T. L. Rose, J. A. Yesavage, *J. Consult. Clin. Psychol.* 51, 815 (1983).
19. R. M. Dawes, *Am. Psychol.* 26, 180 (1971).
20. D. Wedding, *Clin. Neuropsychol.* V, 49 (1983).
21. D. A. Leli and S. B. Filskov, *J. Clin. Psychol.* 37, 623 (1981).
22. J. S. Wiggins, *Clin. Psychol. Rev.* 1, 3 (1981).
23. P. E. Meehl, *J. Counseling Psychol.* 6, 102 (1959).
24. ———, *Problems in Human Assessment*, D. N. Jackson and S. Messick, Eds. (McGraw-Hill, New York, 1976), pp. 594–599.
25. ———, *J. Counseling Psychol.* 4, 268 (1957).
26. W. B. Martin, P. C. Apostolakos, H. Roazen, *Am. J. Med. Sci.* 240, 571 (1960).
27. L. R. Goldberg, in preparation.
28. J. F. Fries *et al.*, *Arthritis Rheum.* 29, 1 (1986).
29. D. Kahneman and A. Tversky, *Am. Psychol.* 39, 341 (1984); K. R. Hammond and D. A. Summers, *Psychol. Rev.* 72, 215 (1965).
30. L. J. Chapman and J. P. Chapman, *J. Abnorm. Psychol.* 72 193 (1967); *ibid.* 74, 271 (1969).
31. P. E. Meehl, *Psychodiagnosis: Selected Papers* (Univ. of Minnesota Press, Minneapolis, MN 1973).