Judgment Pitfalls in Estimating Premorbid Intellectual Function

David A. Kareken

Departments of Neurology and Psychiatry, Indiana University School of Medicine and the Richard L. Roudebush VA Medical Center

Neuropsychological diagnosis first requires estimation of patients' premorbid cognitive ability. Although a number of methods have been proposed for this purpose, in practice there is little standardization. Often, clinical judgment and expert synthesis of multiple sources of information are prescribed to derive a premorbid estimate. There are, however, a number of systematic biases in human cognition that may lead clinicians to produce inaccurate estimates of premorbid function. This paper reviews the cognitive biases that are most likely to interfere with clinical judgment in this context. Given findings in the human judgment literature, actuarial strategies are recommended to avoid the contamination of cognitive bias. © 1997 National Academy of Neuropsychology. Published by Elsevier Science Ltd

Determination of premorbid ability is critical to neuropsychological diagnosis, as the lack of a valid estimate precludes the accurate diagnosis of decline. Despite a number of methods and theories (Franzen, Burgess & Smith-Seemiller, this issue), there is little standardization or general agreement about which method of premorbid IQ calculation is best. Further, quantitative methods of estimating the premorbid status of more specific cognitive domains, such as memory or visual-spatial perception are virtually nonexistent (Franzen et al., this issue; Williams, this issue).

In the absence of a generally accepted standard, clinicians must rely on their judgment. Lezak (1983) proposes that clinicians render an estimate synthesized from interview, anecdotal information, demographic characteristics, and patients' patterns of test performance. Similarly, Vanderploeg (1994) recommends integrating estimates from a number of formulas with past academic performance, occupational success, and consistency in performance throughout the neuropsychological battery. For example, a patient with a successful career in medicine suggests a Full Scale IQ higher than an obtained value of 98. A history of academic awards, peaks in test performance, and the patient's use of sophisticated words or formal grammar during interview may also suggest superior intellectual endowment.

Thanks to James Townsend, Frederick Unverzagt, David McAllister, and David Pfenniger for comments and criticisms on earlier versions of this manuscript.

Address correspondence to: David A. Kareken, PhD, Department of Neurology (RI 3751), Indiana University School of Medicine, Indianapolis, IN 46202. E-mail: dkareken@indyvax.iupui.edu.
This complex synthesis of multiple information sources leaves the clinician susceptible to biases in human decision making that transcend expertise or experience (Dawes, 1989; Dawes, Faust, & Meehl, 1989; Garb, 1989). This article will describe limitations in human judgment, and explore potential ways in which they may distort estimates of premorbid neuropsychological ability derived from clinical judgment (also see Wedding & Faust, 1989).

### STATISTICAL VERSUS CLINICAL PREDICTION

Since Sarbin's (1942) seminal study and Meehl's (1954) subsequent review, a number of experiments have compared the decision accuracy of actuarial formulas to human judges. In nearly all studies of this sort, an actuarial or statistical formula either equals or exceeds the decision accuracy of human diagnosticians (Dawes, Faust, & Meehl, 1989, 1993; Sawyer, 1966). Holt (1958) suggested that clinicians might surpass the accuracy of an actuarial formula by considering the formula's decision, and supplementing it with additional data. This approach is clearly suggested by Vanderploeg (1994), who recommends integrating actuarial premorbid estimates with other data, such as academic history, job performance, and pattern of test performance. However, even when experts are allowed to integrate an actuarial formula's decision with data not available to the formula, their accuracy still does not exceed the formula's (Sawyer, 1966).

These results are not peculiar to the behavioral sciences. Einhorn (1972) found that actuarial formulas using histological signs identified by three expert pathologists were more accurate at predicting the survival time of Hodgkin's disease patients than the same pathologists' clinical judgments. In fact, the pathologists' judgments were uncorrelated with actual survival time. This particular experiment demonstrated that clinical judgment is indeed important. Notably, the histological signs provided by the pathologists were quite useful in predicting survival time. The pathologists were, however, unreliable in combining the information they believed important. The robust consistency of the actuarial method's greater accuracy has been summarized in a recent meta-analysis by Grove and his colleagues (Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, submitted, cited in Grove & Meehl, 1996). They found that of 136 studies that satisfied criteria for inclusion in the study, only 8 showed an advantage for the clinician. This was not peculiar to psychologists. Medically trained judges were no different in their accuracy, and amount of training, experience in the field, and task-relevant experience did not contribute to decision accuracy.

To date, no direct comparison of clinical versus actuarial estimation of premorbid IQ has been reported in the literature. The paradigm is, nonetheless, quite similar to the large number of studies previously reported: a human expert is required to estimate an outcome (in this case, IQ) by considering a number of predictor variables. The expert's prediction can then be compared to the prediction of a statistical formula or actuarial table that uses the same or a subset of the same variables. Although no objective IQ data were available, Kareken and Williams (1994) found that neuropsychologists' premorbid IQ estimates were, as a group, very similar to the Barona estimates of the same individuals. Given that the psychologists' and the formula's estimates were roughly equal, the potential for experts to exceed the Barona formula appears limited, at least using these preditors.

### BIASES IN HUMAN JUDGMENT

The mere existence of the disparity between human and actuarial decision accuracy is less interesting, and less informative, than why it exists (Wiggins, 1981). Clearly, the accuracy of statistical prediction does not surpass that of clinical judgments because clinicians lack
intelligence or knowledge about the phenomena that they try to predict. As exemplified by Einhorn’s (1972) study, experts provide a critical role in determining the variables that enter such formulas. Rather, flaws and biases inherent in cognition prevent judges from combining and integrating information accurately (Tversky & Kahnemann, 1974). The first and most encompassing of those biases is the reliability and consistency with which humans think.

**Reliability**

The human brain has many advantages over the computer, such as the ability to formulate hypotheses (Meehl, 1954). The computer’s strength lies in its reliability and consistency in combining information. This aspect alone will give the computer a superior degree of decision accuracy (Kleinmuntz, 1975). Phrased psychometrically, reliability is a necessary condition for validity, and poor reliability necessarily attenuates validity. Inconsistency in a clinician’s day-to-day judgment due to fatigue, distraction, time pressure, and other extraneous “noise” compromises the validity of diagnosis.

This can be examined by contrasting the accuracy of a statistical model of a clinician’s decisions to actual decisions made by the same clinician (the “paramorphic” model of Hoffman, 1960). For example, one could take the premorbid IQ estimates that a clinician gives to a number of subjects, and correlate those estimates with the predictor variables that the clinician used to create the estimates. The resulting regression equation is a statistical model describing the expert’s behavior. Pitting the clinician against his or her own model on a new set of cases, one can determine which is more accurate: the clinician or the model derived from that clinician’s behavior. Goldberg (1970) undertook such a study, and examined the differential diagnoses of psychosis from neurosis using the Minnesota Multiphasic Personality Inventory (MMPI). When the actual clinicians and their statistical models were given new cases to diagnose, 97% of the statistical models were equally or more accurate than the clinicians on whom the models were based.

Goldberg (1970) derived the conditions in which the “model of man” will better the man. Applied to premorbid IQ estimation, the more linearly predictable a clinician’s estimate of IQ is from a set of predictors (i.e., as \( r \) approaches unity), the more difficult it will be for the clinician to surpass the accuracy of his or her own statistical model (see Goldberg, 1970 for the proof). In fact, expert reasoning can often be described well by linear terms, even when the experts believe that their judgment process is highly nonlinear (e.g., Hoffman, Slovic, & Rorer, 1968; Oskamp, 1967; Summers, Taliaferro, & Fletcher, 1970). IQ estimation appears not to be an exception, as clinicians’ IQ estimates are highly linearly related to the demographic cues on which they are based (median multiple \( R = .93 \) for Verbal IQ, median \( R = .86 \) for PIQ; Kareken & Williams, 1994).

Therefore, if on a sample series of cases, a clinician’s IQ estimates were sufficiently valid, as well as strongly linearly related to the predictor variables, the clinician would be better advised to use his or her model instead of clinical judgment. Unreliability in a clinician’s decision process will work against the clinician, as the clinician’s model is 100% reliable.

**Specific Biases in Judgment**

Unreliability may be due to a number of reasons. Fatigue and distraction are two mentioned above. There are, however, more systematic and nonrandom sources that contribute to unreliability in judgment (Nisbett & Ross, 1980; Tversky & Kahnemann, 1974).

*The representative heuristic.* Under the effects of representativeness, people make predictions about an event by relying too heavily on the similarity (representativeness) of evidence to that
event. This reliance on feature similarity is at the exclusion of base rates, or information about how likely the event is to occur (Kahnemann & Tversky, 1973). To illustrate, Kahnemann and Tversky (1973) gave subjects the description of a highly intelligent, orderly, and obsessive man, who had little interest in others. Asked to guess the man’s field of study, the subjects believed that he was more likely to be studying for a degree in computer science (an infrequently encountered field at the time the of the research) than more common fields, such as humanities or social science. The subjects ignored the low base rate of the computer science major, and instead attended to how well the man’s description matched the stereotypic picture of an intelligent, introverted computer scientist. People appear to lack the intuitive sense of what Bayes law shows to be true: Even if an indicator variable is valid (the probability of being introverted is high for computer scientists, perhaps itself a dubious hypothesis), its utility is severely limited by a low base rate (the probability of being a computer scientist given an introverted personality is low).

Representativeness also influences subjects when they are asked to estimate the values of a continuous variable, as in premorbid IQ estimation. Kahnemann and Tversky (1973) asked subjects to predict grade point average (GPA) from a test of mental concentration. Although the subjects knew that mental concentration was an imperfect predictor of IQ, their predictions were near translations of concentration into GPA. A second group of subjects displayed the same tendency when using a presumably less valid predictor of GPA: sense of humor. The subjects failed to regress their estimates of GPA according to the validity of the predictor variables. Their estimates instead reflected the degree to which one trait was similar to the variable being estimated. This was apparently true even when the validity of the relationship was dubious (i.e., humorous people perform well in school).

Kareken and Williams (1994) found a similar tendency in professional neuropsychologists who were asked to estimate IQ for a set of hypothetical individuals using the demographic variables of the Barona formula (Barona, Reynolds, & Chastain, 1984). The high correlations between education and the clinicians’ Verbal IQ (VIQ) estimates ($r = .85$) suggests that education was so representative of VIQ that the clinicians merely translated education into VIQ. This behavior was at least consistent with the judges’ average (incorrect) belief that VIQ correlated with education at $r = .83$ (the actual relationship is $r = .56$; Barona et al., 1984). In the case of Performance IQ (PIQ), the judges believed that education correlated with PIQ at $r = .47$, and that occupation was a slightly better predictor ($r = .56$). However, the correlation between the clinicians’ PIQ estimates and education was .78. The correlation between the PIQ estimates and occupation was .66. The clinicians behaved as if education accounted for nearly three times more variance in PIQ than they believed it should. Moreover, although they believed that occupation was a better predictor than education, they actually relied more on education. Finally, the expert judges surrounded their IQ estimates with 95% confidence intervals that were nearly half the size of Barona equation (Barona et al., 1984). That is, the clinicians were far more certain about the accuracy of their estimates than the true empirical relationships dictate.

Representativeness prevented these experts from acting in accordance with their own beliefs about the validity of demographic predictors. The psychologists made IQ estimates that were near translations of education and they surrounded their estimated values with very small margins of error. In practice, these narrow confidence intervals may form a sensitive trigger to the diagnosis of decline, leading clinicians to interpret small differences between predicted and obtained scores as significant.

The availability heuristic and illusory correlation. Judgments based upon the ease with which information is retrieved from memory capitalize on “availability” (Tversky & Kahnemann, 1973). The memory of a particularly striking or vivid case might, for instance, covertly
Judgment Pitfalls and Premorbid IQ Estimation

influence the diagnosis of the next patient. Certainly, neuropsychologists know how consciously forgotten information primes subsequent recognition, even in profoundly amnesic subjects (e.g., Cohen & Eichenbaum, 1993; Squire, 1987). Poses and Anthony (1991) showed that diagnoses of bacteremia were one third more frequent when physicians reported thinking of recent patients with the condition, or factors associated with it. A clinician’s attribution of premorbid function might therefore differ depending on recent experience, such as the vivid memory of a similar patient or a recent seminar attended.

Availability of information can also engender a related distortion in clinical judgment: Illusory Correlation, or the appearance of a relationship between variables where none exists (Chapman & Chapman, 1967, 1969). People attend readily to conspicuous details, such as events in temporal or spatial contiguity. Less salient events fit better into the background (Einhorn, 1988; Nisbett & Ross, 1980). A diagnostic sign paired with a disease is a combination likely to attract a clinician’s attention. The disease without the marker, and the marker without the disease, are less vivid pairings, and not likely to draw attention. The result is the inference of a correlation based, not on actual degree of association, but on the conspicuousness of the information (Jenkins & Ward, 1965; Smedslund, 1963; Ward & Jenkins, 1965). Thus automotive mechanics with superior Block Design scores are more likely to constitute an enduring memory (and mental normative standard) than mechanics with average or low-average scores (which may be more common). Even when explicitly told that a case example is atypical of the larger group, subjects continue to make unwarranted inferences about the population if the example is sufficiently graphic (Hamill, DeCamp Wilson, & Nisbett, 1980).

In daily practice, clinicians are deprived the luxury of complete information, and presented only with the vivdness of a diagnostic sign accompanied by a particular disorder (Dawes, 1989). People without cognitive impairment, but who also have the diagnostic sign, do not come to a neuropsychologist’s office. The information to which the clinician is exposed is the “available” information leading to an association of symptom and disorder. The remaining data needed to form a true empirical correlation are rarely encountered in clinical settings. Similarly, individuals with excellent mechanical ability, average to low average Block Design scores, and no brain injury, do not present themselves for examination. Premorbid test data are almost always nonexistent, and no amount of clinical experience can help a clinician make accurate premorbid estimations when the dependent variable is itself missing (Dawes, 1989).1

Finally, apart from the vividness and completeness of information presented to the expert, the limited working memory of the human brain is alone sufficient to prevent the accurate mental calculation of covariance between a number of characteristics and premorbid ability. As the amount of information to be processed increases, so does illusory correlation (Leuger & Petzel, 1979). This may be one reason why illusory correlation has proven to be a robust phenomenon that is difficult to change, even when judges are trained to be aware of its presence (Kurtz & Garfield, 1978).

Labeling. Faust and colleagues found a tendency for more experienced neuropsychologists to overdiagnose impairment (Faust, Guilmette, Hart, Arkes, Fishburne, & Davey, 1988). The physicians in Poses and Anthony’s (1991) study also significantly overestimated the presence of bacteremia. A bias to find pathology might stem from a covert belief that, by virtue of being a “patient” or “referral,” an individual possesses a pathological condition until proven

1There is little empirical evidence to support the common belief that experience is related to accuracy in clinical judgment (see Garb, 1989, for a review).
normal. For example, Temerlin (1968) found that a label of psychosis from a prestigious psychiatrist convinced a group of clinicians that an individual, displaying no pathology whatsoever, was indeed psychotic. A graphic example of this effect is found in Temerlin and Trousdale (1969). After hearing the tape of a patient and being told that a well-known psychologist believed the person to be psychotic, a clinician offered the following diagnosis:

... "manic-depressive psychosis, paranoid-schizophrenia, ambulatory schizophrenia, and undifferentiated schizophrenia." During the debriefing he was asked: "Did the patient really look like four different kinds of psychosis?" He answered: "Hell no, he looked normal, so I figured his personality must reflect some weird combination of psychotic patterns." When asked, "If he looked normal why didn't you say so on the data sheet?" he replied: "Because normal people do not come into a clinic." (p. 25)

Other striking examples are found in Rosenhan's (1973) classic study in which clinicians interpreted routine behaviors of normal individuals in a psychiatric setting as pathological. Such effects may result from the fact that abnormality is what clinicians are trained to find (see Kayne & Alloy, 1988 for a discussion of expectation-based distortions). The referral of a patient for evaluation following a minor head injury might therefore covertly induce clinicians to look for reasons why obtained IQ is lower than the estimated score. The same estimated-obtained discrepancy might go unremarked if the individual were instead volunteering as a healthy control in a research project.

Anchoring. Tversky and Kahnemann (1974) noted that judges make estimates subsequent to an initial value that is "anchored" to their original estimate. In other words, judges do not sway far from their first impressions. Meehl (1960) showed that clinicians form enduring clinical impressions in psychotherapy in 2 to 4 hours, and remain invariant in their opinions for at least 16 hours of contact. Anchoring may affect judgments across patients as well. An initial patient can form a background (anchor) to which subsequent patients are compared. For example, Bieri, Orcutt, and Leaman (1963) found that clinical judges conceptualized identical, moderately pathological cases differently, according to whether an anchor case of high or low pathology preceded them. Friedlander and Stockman (1983) also found that, after initial exposure to a patient presenting as asymptomatic, clinicians disregarded pathognomonic signs in the same patient that were introduced at a later time.

Anchoring might influence judgments of premorbid ability by restraining clinicians to their initial impressions. Once an estimate of premorbid ability has been created, perhaps only implicitly (e.g., a neuropsychologist is told that a well-educated executive with recent trauma is scheduled to be tested), the clinician might not use subsequently obtained information to revise optimally an original estimate. Anchoring does not necessarily imply that new information will be ignored or discounted entirely, although this can be the case. Even though a clinician believes that he or she is appropriately taking account of new information, new estimates still might not be revised in accordance with they weight that they deserve.

Anchoring can also be exacerbated by the human tendency to seek evidence that confirms a hypothesis or first impression (e.g., Wason, 1960). Setting out to disprove a theory (and failing to do so) is a far more stringent test of a theory than gathering supporting evidence (Meehl, 1978). It holds that (a) if a theory is true, then the supporting data will be observed; and (b) if the data are not observed, then the theory cannot be true. It is logically invalid to assert that since the data are observed, the theory must be true, an error known as affirming the consequent (Rorer, 1991). Therefore, having formed an initial premorbid estimate, a clinician is more likely to remain anchored to that estimate by searching only for validating
evidence (affirming the consequent). A more accurate revision will come from looking for data inconsistent with the estimate.

**AVOIDING JUDGMENT BIAS**

Einhorn (1986, 1988) made a compelling argument about the nature of human reasoners, and in particular, the assumptions that clinicians hold about probability and random error. He argued that clinicians believe, at least implicitly, that (a) all variance in behavior can be explained; and (b) only lack of information prevents perfect explanation. Einhorn suggested that clinicians harbor this belief because the vast majority of clinical activity is from a backward-looking vantage point. There is little true forecasting in routine clinical diagnosis, and most diagnostic opinions are formed from backward reasoning in which past events are arranged to create a *post hoc* causal model of present symptoms (Einhom, 1986). This backward reconstruction is often accomplished with an exhaustive search of prior history for additional information that completes the model, and satisfies the clinician's personal threshold of diagnostic confidence. *Post hoc* explanation is always much easier to accomplish than point prediction of future events, and most often does not generalize.

In the Kareken and Williams (1994) study of premorbid IQ estimation, the clinicians did not have access to information beyond age, race, gender, education, occupation, school grades, and quality of schooling. When asked if more information or a diagnostic interview would have made them more accurate or more confident, the overwhelming majority of clinicians answered affirmatively, and in the context of confidence intervals that were already smaller than the actual empirical relationships dictate. This behavioral response lends strong support to Einhom's supposition that clinicians assume that near perfect estimations are attainable, and that the more information sought, the better the estimate (Einhom, 1986).

Einhom suggests that, in contrast to clinical reasoners, the assumptions in statistical estimation are that (a) all knowledge is at best probabilistic; (b) perfect prediction is never possible; and (c) a certain amount of error is random and cannot be reduced with more information (Einhom, 1986). In fact, more data may have the opposite effect, and compromise predictive power. Statistical practice teaches that "over-fitting" a regression model with too many parameters will cause the equation to generalize poorly due to shrinkage. Trying to account for all of the variance in a data set capitalizes on random noise, increases redundancy between variables, and decreases generalizability to future clinical cases (Einhom, 1989).

Kleinmuntz (1990) wrote that a first step toward minimizing human judgment bias in decision making is deciding which decisions are best made by an actuarial process, and which require human intervention. Meehl (1986) suggested that most individuals would be uncomfortable with a grocery clerk who visually scans all of the merchandise and avers a quick opinion about their cost. Pending research that demonstrates otherwise, I would suggest the following: Given the similarity of premorbid IQ estimation to other situations examined in the Statistical versus Clinical Judgment debate, the likely intrusion of systematic judgment bias, and the linear manner in which clinicians apparently estimate premorbid function, estimation of premorbid intellect is exactly the kind of task that is best left to a well-developed actuarial formula.

Implementing this assertion requires accepting Einhom's advice that one must "[accept] error to make less error" (Einhom, 1986, p. 387). That is, clinicians must explicitly acknowledge that x/100 decisions will be incorrect by choosing a particular cutoff point, and be sufficiently comfortable with this so as to avoid engaging in the endless *post hoc* search for numerous other variables. Whatever the error rate of a particular formula, it will almost certainly be exceeded by overfitting it with extraneous variables. Unfortunately, there is a low
ceiling of variance that can be accounted for in behavioral data, and after a certain number of variables have been taken into account, further incremental validity is highly unlikely, no matter how much data are acquired. Whatever additional data can be found are likely to be redundant with information already known, and may unjustifiably increase diagnostic confidence without making the estimate more accurate. Faulty decisions of decline are therefore more likely.

REFERENCES


