Dawes, R. M., Faust, D., & Meehl, P. E. (1993) Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 351-367). Hillsdale, NJ: Lawrence Erlbaum.



# Statistical Prediction versus Clinical Prediction: Improving What Works

Robyn M. Dawes
*Carnegie Mellon University*

David Faust
*University of Rhode Island*

Paul E. Meehl
*University of Minnesota*

In Pennsylvania, offenders sentenced to maximum prison terms of 2 years or longer are considered for parole under the authority of the Pennsylvania Board of Probation and Parole after they have completed half of their maximum sentence. The decision to grant or withhold parole is based on a four-step procedure beginning with a summary recommendation from the correctional staff, proceeding through a "parole case analyst" and then to a "parole interviewer." This interviewer is either a board member or a specialized hearing examiner who has access to the previous reports of the staff and the analyst, and who makes a final recommendation to the parole board, which has the ultimate responsibility for the decision. One thousand thirty-five prison inmates were interviewed for parole between October 1977 and May 1978, yielding 743 cases in which the parole board made final decisions, of which 84.7% were to grant parole. In all but one of these cases, the decision of the parole board was identical to the final recommendation of the interviewer, who also made four- or five-point ratings on: (a) prognosis for supervision, (b) risk of future crime, (c) risk of future dangerous crime, and (d) assaultive potential. On the basis of a 1-year follow-up study, J. Carroll, Winer, Coates, Galegher, and Alibrio (1982) were able to compare the prediction of the parolee's behavior based on the interviewer ratings with its prediction based on simple background factors, such as number of previous convictions. (These factors also were available to the interviewers and were shown to be correlated with their clinical judgments.)

Approximately 25% of the parolees were considered "failures" by the board within a 1-year period—for reasons such as being recommitted to prison, absconding, committing a criminal act, being apprehended on a criminal charge, or committing a technical violation of parole. None of the interviewers' ratings predicted any of the outcomes, the largest correlation being .06. In contrast, a three-variable model based on offense type, number of convictions, and number of (noncriminal) violations of prison rules during the last year of prison did have (very) modest predictability, $R = .22$, a result consistent with earlier findings that actuarial predictions based primarily on prior record predict parole violation with a multiple $R$ of approximately .30 (Gottfredson, Wilkins, & Hoffman, 1978). When parolees were convicted of new offenses, the seriousness of such crimes was

correlated .27 with the interviewers' ratings of assaultive potential, but a simple dichotomous evaluation of past heroin use correlated .45. Parole revocation and violence are very difficult to predict, partly because offenses of record are a small minority of those committed, but these outcomes are better predicted on a statistical than a judgmental basis, as has been found in other studies examining criminal recidivism (Glaser, 1964).

J. Carroll et al.'s results illustrate the outcome of research comparing statistical to clinical prediction, where these two types of prediction refer to these two ways of combining data (not to its source). The purpose of this chapter is (a) to present a brief synopsis of this research; (b) to present a (possibly new) framework for interpreting this evidence; (c) to discuss the characteristics of the predictive problem that may be primarily responsible for the superiority of "formula over head"; (d) to discuss some of the objections to the research, and (e) to propose a way of implementing statistical models that overcomes a major objection to their use. The first topic has been discussed at length in previous books and papers (e.g., Dawes, 1988; Dawes, Faust, & Meehl, 1989; Meehl, 1954), as have the third (e.g., B. Carroll, 1987; Dawes, 1979), and the fourth (e.g., Faust, Meehl, & Dawes, 1990; Meehl, 1986). This chapter, therefore, focuses on the second and fifth.

## THE RESEARCH

Here, we list 10 diverse areas in which studies have shown the superiority of statistical prediction. There are other areas as well, but we list these 10 so that the reader can have access to a representative set of studies on which we base our conclusions of superiority of statistical prediction. (We are omitting those covered in Meehl's 1954 book or Sawyer's 1966 review, except for areas in which we know no subsequent studies.) These areas are those that predict:

1. Academic success (Dawes, 1971; Schofield & Garrard, 1975; Wiggins & Cohen, 1971)
2. Business bankruptcy (Beaver's, 1966, and Deacon's, 1972, models compared to Libby's, 1976, experts)
3. Longevity (Einhorn, 1972)
4. Military training success (Bloom & Brundage, 1947)
5. Myocardial infarction (Goldman et al., 1988; Lee et al., 1986)
6. Neuropsychological diagnosis (Leli & Filskov, 1984; Wedding, 1983)
7. Parole violation (J. Carroll et al., 1982; Gottfredson, Wilkins, & Hoffman, 1978)
8. Police termination (Inwald, 1988)
9. Psychiatric diagnosis (Goldberg, 1965)
10. Violence (Miller & Morris, 1988; Werner, Rose, & Yesavage, 1983)

Some of the studies in some of these 10 areas can be summarized briefly.

Dawes (1971) compared clinical and statistical prediction of success in graduate school. Statistical methods, even those based on single variables (e.g., grade-point average), outperformed the clinical predictions of an admissions committee that had access to more extensive information. Using the statistical method, it would

also have been possible to eliminate 55% of applicants that the admissions committee considered and later rejected, without eliminating any applicants the committee considered and later accepted.

Dawes and his colleagues at Oregon decided that this finding concerning automatic elimination was of sufficient importance that it be formally implemented—both to save the psychology faculty there from the meaningless work of evaluating applicants who had no chance of admission, and to save these applicants themselves the work, expense, and heartache of applying to a program into which they had no chance of being admitted. Before doing so, however, Dawes (1979) cross-validated this elimination procedure on the subsequent year, even though the ratio of observations to variables involved in the elimination were so large that there was little question of the statistical stability of the earlier results. The elimination procedure was, after all, "radical." What he did was to inform the other members of the Oregon admissions committee that the procedure was being implemented and they would be asked to examine only those applicants who passed the screening, but then deceptively pass on to them for evaluation any applicant who appeared to him to have any particular strengths not reflected in grade-point average or test scores. None of his colleagues noticed this deception. Nor were any of those applicants who would not have passed the screening given a rating by the other committee members high enough to have any chance of being admitted. The reason was that for every applicant below the cut-score who had a particular strength, there was one above this score who had a comparable strength, and there was no reasonable or ethical reason for admitting the former applicant rather than the latter. Subsequently, Goldberg (1977) informed potential applicants of a revised formula being used for screening: Average Graduate Record Exam score plus grade point average multiplied by 100. Not only were the applicants informed that potential applicants with a score less than 9.50 should not waste their time and energies applying, but the probability of being admitted with particular scores above that level was shared as well. Although this procedure was met with some cries of "dehumanization," the psychology professors at Oregon felt they were being perfectly open and honest with the applicants by providing them with as much information as possible about the chances of being admitted and therefore highly ethical. (Of course, the number of $25 admissions fees to the university went down.)

Einhorn (1972) studied the prediction of survival time following the diagnosis of Hodgkin's disease, a previously untreatable, and hence fatal form of cancer. Pathologists rated patients' biopsy slides along nine dimensions they deemed relevant in appraising disease severity and also formulated an overall rating of severity. Statistical formulae were first developed and then validated by examining relations between the pathologists' ratings and actual survival time. Although the pathologists' overall ratings of severity showed minimal relations to survival time, a statistical method achieved modest but significant relations. Of particular interest, the study shows that the pathologists' ratings did contain information of potential predictive value, but only the statistical combination method captured this potential.

Finally, Libby (1976) had loan officers from either relatively small or large banks predict which 30 of 60 firms about which financial information was avail-

able would go bankrupt within 3 years after issuing financial statements. Overall, the loan officers achieved 74% predictive accuracy, in comparison to the 82% accuracy achieved by the use of a statistical method (Beaver, 1966; Deacon, 1972).

There are some exceptions—particularly in the medical domain (e.g., Brannen, Godfrey, & Goetter, 1989; Sutton, 1989)—but the framework proposed in the next section of this chapter may yield some insight into why they occur. Overall, we reiterate Meehl's (1986) conclusion:

> There is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one [the relative validity of statistical versus clinical prediction]. When you are pushing 90 investigations [now closer to 140], predicting everything from the outcome of football games to the diagnosis of liver disease and when you can hardly come up with a half dozen studies showing even a weak tendency in favor of the clinician, it is time to draw a practical conclusion. (pp. 372-373)

## THE FRAMEWORK

In his 1954 book, Meehl proposed certain ground rules for the comparison of clinical versus statistical prediction. The most important of these were that the prediction should be based on exactly the same data, and the statistical prediction should avoid capitalization on chance due to overfitting a sample of data. The latter rule was followed either (a) by using crossvalidation (which might better be termed *validation,* with data on which the statistical rule is derived termed the *development sample*); or (b) by using unit weights (such an a priori weighting system being equivalent to a single predictor—as in a system using a validation sample), or (c) by using a sufficiently large sample that the stability of the statistical model was not in question. Since that time, both jackknife procedures (e.g., Drehmer & Morris, 1981; Gollob, 1967) and the use of the Wherry-Lord "prophecy" formula in multiple regression contexts have become common ways of dealing with the overfitting problem. The Wherry-Lord formula occurs in many different algebraic forms; in the simplest its numerator is equal to the actual squared multiple regression coefficient minus that expected on a chance basis (= $k/(n-1)$, where $k$ is the number of predictors and $n$ is the sample size), while its numerator is equal to 1 minus this chance expectation, a form that happens to be isomorphic to almost all "correction" formulas. Simulations have shown this formula to be quite accurate (e.g., Schmitt, Coyle, & Rauschenberger, 1977); in fact, a formula identical to that of Wherry-Lord has been proposed as a method for determining how many variables to enter into a regression equation in order to maximize expected predictability of a new sample (Breiman & Freedman, 1983). Aware of the problems of overfitting, most researchers in this area have either followed Meehl's second ground rule, or have used one of the subsequent procedures.

The first ground rule, in contrast, has often been violated in studies purporting to investigate the clinical versus statistical prediction problem. In the studies assessing interviews, for example, the clinician often has access to more information than is used in the statistical prediction model. A larger information set is also available in the medical studies of which we are aware that show greater accuracy for clinical prediction; for example, in the study by Brannen, Gottfred, and Goetter

(1989), the predictions of an acute physiological and chronic health scoring system (APACHE-II) were found to be inferior to those of "the critical care fellow," who was board certified in internal medicine and who "had seen the patient, obtained a history, and conducted a physical examination, as well as reviewed the pertinent laboratory and roentgenogran and data available" (p. 1083);[1] in the study by Sutton (1989), Bayes' formula was found to be inferior to the diagnosis of doctors who actually saw the patients. In both of these studies there is a possibility that valid predictors were noted in the live examination, predictors not available to the statistical systems (but—as is pointed out later—which might be integrated into such systems, to create predictions of equal or greater accuracy). A clear statement of the information problem in a comparison that violates Meehl's first rule can be found in a business context examined by Blattberg and Hoch (1990), in which they concluded that predictions should be made by "50% model and 50% manager." "Experts also had inside information, not the Machiavellian variety available to corporate officers, but they clearly had more information available to them than did the models. We have elected to label this inside information as intuition and see it as a valuable decision input."

Basically, we distinguish four types of relations between the information available to the statistical model and to the clinician. Based on Gergonne's (1817) naive set theory (which does not involve the paradoxes of material implication), these relations are illustrated in Fig. 13.1. The information sets are either exclusive (rarely the basis for a comparison thus far), identical (following Meehl's ground rules), inclusive—in which the information available to one method is a subset of that available to the other—or disjunctive. The vast majority of the studies in the "clinical versus statistical" field have been based on information sets that are either identical or inclusive, with the information on which the model is based being a subset of the information available to the clinician, not vice versa. In the studies mentioned earlier, for example, most involved a comparison of predictions based on the same data; in contrast, the studies on personnel selection by Bloom and Brundage (1947) and by Dawes (1971), on longevity by Einhorn (1972), on parole violation by J. Carroll et al., and on medical diagnosis by Lee et al. (1986) indicated superiority for statistical prediction even when the information on which it was based was a subset of the information on which the clinical prediction was

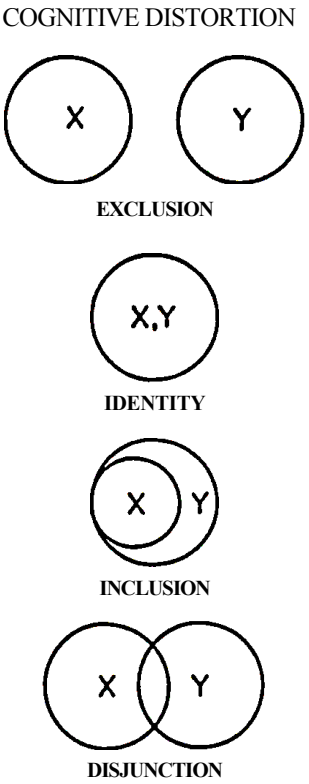


FIG. 13.1. The Gergonne relations.

[1] Following the completion of the present article, a new prediction system, APACHE-III, has been demonstrated to be superior to such judgment—specifically about survival before hospital discharge in an intensive care unit; see Knaus, Wagner, & Lynn (1991).

based. For example, in the Lee et al. (1986) study "all baseline characteristics used in computing model predictions were among the descriptors listed on the one-page summary given to the doctors," (p. 555). The study by Goldman et al. (1988) predicting whether chest pain was due to myocardial infarction is unusual in that the data bases available to the computer simulation model and the clinicians (apparently) had a disjunctive relationship. We have found no studies in which the data available to the clinician is the subset of that available to the model, seemingly because one of the assumed benefits of the clinician is the ability to gather data that might not be considered relevant prior to the evaluation process (although that might be balanced by the ability to be distracted by irrelevant data; see Dawes, 1988, chaps. 5 and 6).

When the data bases are identical, the findings have been uniform in showing that statistical *combination* of data is superior to clinical combination. There are, as noted, a few exceptions when the information on which the model is based is included in the information available to the clinician; the problem with this structure is, however, that the conclusion can be unambiguous only if the statistical prediction is superior; in the event that the clinical prediction is superior, there is always the possibility that had the additional data been incorporated into the statistical model—so that ~~if~~ an identity relationship following Meehl's ground rules had been formed—the model might have turned out to be superior. Conversely, we have been able to discover no studies in which the data available to the clinician is a proper subset of that available to the model; given that structure, an outcome in favor of the clinician would be a particularly strong one in the context studied, but to the best of our knowledge no such outcomes have been observed. Finally, a study using the disjunctive structure is particularly difficult to interpret.

We recommend that if the researcher is interested in studying statistical versus clinical predictions as *general methods for combining data,* then any relationships between the data on which the model is based and on which the clinical or judgment is based should—as far as possible—be transformed to an identity relation. There are, of course, many important problems involved in such a transformation, particularly when the basis of clinical judgment is unclear—for example, some characteristics of a biopsy that are not included in coded ratings (as in the Einhorn study, in which the statistical model was still superior). Moreover, there are other ambiguities; should, for example, a "gestalt characteristic" be included as an input in a statistical model, or is it really an act of clinical integration? The latter position may appear obvious, until it is noted that virtually all inputs of statistical models involve some form of human information processing and coding (see Dawes & Corrigan, 1974), and it is not always clear where to distinguish coding from integration. Nevertheless, the "practical conclusion" of all these studies is quite clear.

Finally, there is one method of combining statistical information with clinical judgment that has been tested at least in a few contexts—and found wanting. The method is to inform the clinician of the output of a statistical model and then to allow the clinician to "improve" on it (Arkes, Dawes, & Christensen, 1986; Goldberg, 1968; Sawyer, 1966). In the few contexts studied, that combination does worse than the statistical models alone. For example, in the Arkes, Dawes, and Christensen study, people with and without expertise in baseball (as evaluated by a

test dealing with rules and terminology) were (correctly) informed that among nonpitching candidates for "Most Valuable Player Award" in the National League between 1940 and 1961, the player on the team ending higher in the standings was chosen 70% of the time. The nonexpert group of subjects did better than the expert group, apparently because they employed the rule more often (according to self-report)—even though these subjects in the expert group may have been more likely to recall the award-winning player directly than those in the nonexpert group. (The average number of correct judgments in both groups fell short of the 14 correct judgments that would have been made were the rule applied "blindly!") Although these particular judgments of experts defined according to knowledge of rules and procedures of baseball may be of passing interest, the study was constructed specifically to be parallel to that of Goldberg (1968), in which experts were provided with a simple rule for differentiating psychiatric diagnoses of neurosis versus psychosis from Minnesota Multiphasic Personality Inventory (MMPI) profiles and proceeded to make judgments that were less valid than the rule itself; in fact, Arkes, Dawes, and Christensen chose a rule of 70% accuracy because the Goldberg (1965) rule has that accuracy. Unfortunately, the number of contexts studied is small.

## CHARACTERISTICS OF THE PROBLEM

As Dawes (1979) and B. Carroll (1987) emphasized, these prediction problems involve many factors that are not assessed by either the model or the clinician, that often cannot be assessed because they are unknown, or cannot be assessed because their influence does not occur until after the prediction is made. Some of these (e.g., possible genetic dispositions tied to particular genes) may be capable of evaluation at some later point in time, whereas others (e.g., the nature of the old friends a parolee happens to meet the first few days out of jail) may be totally unpredictable. Such factors (either considered singly or in interactions with others), may be of a most ephemeral nature—but nevertheless of great importance in influencing outcomes. See, for example, Malmquist and Meehl (1978, p. 155) for a discussion of the role of luck in psychopathology. For the purposes of predicting future outcomes from *stable* predictors, these factors, however important, must be nevertheless considered as "noise," although not always of a random variety. How well do people perform in a context involving a great deal of such unpredictable noise?

Not well. Performance can be conceptualized by employing what is termed the *lens model analysis* of components of clinical inference (Hammond, Hursch, & Todd, 1964; Hammond & Summers, 1965; Tucker, 1964). When such inference is made on the basis of codable multivariable input, the resulting judgments can be broken down into three additive components: (a) a random component due to unreliability of judgment, (b) a component that can be predicted by a linear combination of the input variables, and (c) a reliable "residual" component (which presumably reflects "configurality," "intuition," and so on). The latter two components can then be correlated separately with both the criterion values and with the predicted criterion values based on a linear model of the input variables (the "ecological" linear model). The consistent finding is that only the second com-

ponent of judgment, the linearly predictable one, is correlated with either the criterion values or the ecological model, more highly with the model values than with the actual ones. The finding that the residual component of judgment is unrelated to accuracy indicates why the clinical modification of the statistical model does not improve it when the statistical prediction is made in a linear manner. Moreover, the high correlation between the two models (of judgment and of reality) follows from the fact that any two linear models with weights of the same sign will correlate highly (Castellan, 1973); thus, it may simply reflect the fact that judges are weighting the variables in the appropriate directions.

In contrast to people, many statistical models are specifically designed to work in a context of unpredictable noise, which is most often captured by an "uncorrelated error" term in these models. The linear regression model is designed to achieve the best *relative* weighting of variables in order to maximize predictability from the resulting composite. There are many variants of this model (ordinary least squares, ridge regression, etc.), but they all are based on the same principle of attempting to maximize predictability—most usually for subsequent samples in the same population—in a context with an explicitly defined error component.

Linear models—again, those most commonly used in the contexts studied—also have the advantage that their predictions are often insensitive to differences in weights, provided all variables are weighted in the appropriate direction (Bloch & Moses, 1988; Dawes & Corrigan, 1974; Tukey, 1948; Wainer, 1976; Wilks, 1938). The output of linear models is particularly insensitive when the predicted variables form a positive manifold, as often occurs in the type of situation we reviewed; because many of the predictors (e.g., aptitude test score and grades, past arrest record, and number of prison violations) are related to the same sets of unobserved variables (e.g., intellectual competence and motivation, lack of impulse control, and sufficient cleverness not to get caught), they are often related positively to each other. It is, of course, always possible to *hypothesize* "suppressor relationships," in which the simple correlations with the dependent variable are in one direction while the weight in a regression equation is in another, as a result of the covariance structure between predictors. Such variables are typically not found in the context we have reviewed.

Finally, although incapable of forming some "gestalt" judgments, these models nevertheless are immune to many of the cognitive heuristics to which people are subject—for example, those stemming from availability, representative thinking, and framing. (For a review, see Dawes, 1988.) Research results indicate that even if the clinician, in contrast, may be capable of perceiving such gestalts, they are often not particularly important in predicting human outcomes in the "booming buzzing, confusion" of human adult life.

## OBJECTIONS TO THE RESULTS

To state that the research results reported in the first section of this chapter do not arouse universal enthusiasm is an extreme understatement.

Some object that there is no well-specified population of prediction of human outcomes. Therefore, any sampling of particular ones in which the accuracies of clinical and actuarial predictions are compared cannot result in a general conclu-

sion. Moreover, there could always be other contexts in which clinical prediction is superior (the reviewer's own often being the example). There is a problem with this objection because "could" is not equivalent to "are," and if such contexts do exist, they have not yet been discovered—and not for want of trying. As Meehl (1986) pointed out, when you discover such a uniform result across such diverse domains, you should reach, at minimum, a practical conclusion.

Individual studies can also be criticized. Whereas the results across contexts may be interpreted as supporting the superiority of actuarial prediction, each separate context, when considered in isolation, admits to an alternative interpretation or two. For example, in one context in which violence on a psychiatric ward was predicted (Werner, Rose, & Yesavage, 1983), the clinical predictions might have been superior, except that they served as a basis for taking precautions aimed at keeping patients from having any chance to become violent. (The additional finding that the trained clinicians made the same judgments as untrained high school students [Lierer, Warner, Rose, & Yesavage, 1985] could simply indicate that the students, too, would have achieved accuracy, were it not for the "self-negating" nature of the predictions.) Or in the J. Carroll et al. (1982) study of parole boards, actuarial predictions of who would succeed (based on variables such as number of past convictions and violations of prison rules) proved superior to predictions of parole interviewers (who had access to information about these variables). But perhaps the interviewers were nevertheless better at making the simple decision of who should or should not be let out on parole. So, if the interviewers' recommendations had not been followed, their assessments might have been revealed as superior—even though these assessments were shown to be inferior when they were in fact followed.

Perhaps. There is a problem with this kind of analysis, though, because it becomes necessary to concoct a new alternative explanation to "explain away" the findings in each separate context, rather than simply accept each replicable finding. And, as was true for the first objection, the alternatives are simply hypothesized, rather than supported by any data.

In contrast to these structural objections, Meehl (1986) speculated about some of the personal factors that lead to frequent, and often vehement, rejection of the research conclusion. These factors include simple ignorance, narcissistic belief about the validity of one's own judgment, and threat to professional status. There is an additional motivational factor that statistical prediction for human outcomes appears "dehumanized," and there is even a widespread belief that "statistics do not apply to the individual." (If so, there would be no point in conducting randomized trials on groups of people—such as the Salk vaccine experiment in 1954—in order to determine the validity or relative validity of various medical techniques, which must, of course, be applied to an individual person.)

Furthermore, there is the aversion to the lack of predictability of human life that is demonstrated in many of these studies; for example, the best statistical models in most appear to have maximal predictabilities expressed by correlation coefficients of .3 or .4, which can be threatening in such contexts as longevity, academic success, and parole violation. An unpredictable world cannot be a just one, and although predictability is not a sufficient condition for "justice," it is a necessary one. Thus, people who wish to believe that the world itself will give

them certain "entitlements" (Lerner, 1980, 1987) are faced with the implication from these studies that "the race is not to the swift, … but time and chance," (Ecclesiastes, chap. 9, Verse 11).

There are, in addition, objections based on cognitive factors. First, the lack of predictability found in the studies appears to violate our belief that we really do "understand" the course of human life. As Dawes (1993) suggested elsewhere, this belief is based on the ability when given an outcome to identify important antecedent factors acting either singly or in monotonic interactions—for example, the success of "the right person at the right time." These factors are then identified as predictive. The problem, however, is they must predict as main effects combined with others when they are entered as variables of outcome studies; first, the effects of a single influence (variable) are vitiated by the effects of others; second, monotone "combinations of ingredients" are rare statistically; third, prediction must be made *to a specific point in time*—whereas retrospective examination of factors affecting outcome allows the analyst freedom in scanning across time for the existence of the important factors. Finally, and most importantly, even though the observer can create a monotone ("many-one") relationship when observing a consequence and searching for antecedents, the focus on consequences creates an "availability bias" whereby the fact that the same set of antecedents often leads to a different outcome (a "many-many" structure in reality) is unobserved. Knowing why something occurred therefore involves different cognitive processes than does predicting its occurrence, but the knowledge of "knowing why" is easily identified with knowledge that allows prediction. (See Dawes, 1993, for further discussion.)

The last cognitive objection we discuss concerns the "rigidity" of statistical models. It is a clearly correct intuition that situations change. A model, however, is (or generally should be) based on a large sample of data and—the argument runs —once in place produces a determinant predicted value. As the situation changes, however, it is quite reasonable to postulate that the predictors will as well. The argument then runs that even though the statistical model may do better than clinical judgment on the data studied, it is not as capable as such judgment in "altering when it alteration finds." The extreme form of this objection is that no model should be proposed as a substitute for clinical intuition unless it was developed at one point in time and validated at a subsequent point.

This objection is based on a simple error, which is the belief that models cannot be modified from feedback. In fact, model parameters can be systematically altered by feedback, whereas clinical judges often do not have the feedback presented in a systematic manner, or at all. How many clinical psychologists, for example, carefully check the accuracy of their diagnoses and prognoses by evaluating their clients' lives years later? And if they do check, do they keep careful records of these judgments so they are not subject to hindsight biases (that "I knew it all along") or benign memories about these diagnoses and prognoses—or to the reconstructive nature of memory that leaves the past compatible with the present? (Pearson, Ross, & Dawes, 1992.) "Every reminiscence is colored by today's being what it is, and therefore by a deceptive point of view" (Albert Einstein, referenced in Schilpp, 1949, p. 3).

## IMPLEMENTATION

The statistical method has been demonstrated to be superior to the clinical method. In response, many people (not referenced here) have proposed that this finding mandates the improvement of clinical methods, perhaps by improving their reliability. But given that the statistical methods of prediction are not difficult to construct, are based on empirically observable relationships, and are far less expensive than clinicians making the same predictions, a reasonable alternative is to improve the models.

Those variables that clinicians believe to be important that are not captured in a particular model can be incorporated in it. Such variables may even include "gestalt judgments" based on clinical information-gathering techniques. (Recall that the basic question is one of how to combine data, given human involvement is always necessary in order to collect it; even an "objective" test depends on the efforts of a people to create it.) Once entered, the validity of these variables can be discovered rather than simply hypothesized.

Finally, statistical predictive models are easily modified as time or context demands. Although such models are derived and checked (e.g., "cross-validated") in particular contexts at certain points in time, it is possible—in fact desirable—to use feedback as the models are implemented to examine the validity of both their variables and the weighting of these variables. Even the most "subjective" variable thought to be important, such as overall rating of liking or disliking an interviewee, can be examined to determine its predictive validity. And we urge doing so—especially in contrast to postulating in the absence of evidence that such a variable may be predictive (and then often not even bothering to assess it in some reasonably reliable manner).

Many procedures exist for the type of "updating" we propose of models, particularly linear ones (e.g., Duncan & Horn, 1972; Kalman, 1960, 1963; Rumelhart, Hinton, & Williams, 1986). Perhaps the most relevant to predicting human outcomes (and the most easily understood) is *empirical Bayesian* updating. A prototypical example may be found in Rubin's (1980) analysis of optimal weighting for combining undergraduate grade-point averages and Law School Aptitude Test scores to predict academic performance across 82 law schools over 3 years. In effect, the weights computed for a particular school in a given year were regressed toward a common mean; this regression, however, was not total; as a consequence, particularistic information was retained. The empirical result was that with new data the regressed estimates generally outperformed both the weighting systems obtained for the same school the previous year and the overall mean estimates. (It should be pointed out again that this procedure is in no way limited to evaluating "objective" variables—although the degree to which a grade-point average is as objective a variable as it is often asserted to be consists simply of the fact that it is an average of multiple subjective judgments.)

Another method that may be used is modification of the Chow (1965) test. Suppose an investigator believes that at a certain point in time the best prediction equation has changed—through the introduction of new variables, new weightings of the same variables, or even interactions between variables; then, a dummy variable may be entered into the regression that indicates whether the prediction is made before or after this point. The *interaction* of this dummy variable with the

predictor then indicates changes in the best prediction equation. Here, standard procedures of regression analysis can be employed—in particular, the incremental magnitude and significance of having the *set* of predictors involving interactions with this dummy variable.

In fact, the Bayesian and dummy variable interaction methods can be combined by having the analyst develop a prior distribution of belief about when and how the prediction equation might be best modified and then evaluating this belief in light of the data obtained. We present no simple, "canned," method of doing so. In fact, there is—and should be—an element of judgment in such modification. Are we then back to intuitive prediction? No, because the judgment is of the type that considers the prediction problem as a whole, and the individual predictions as elements of the set of predictions to be made; that is, the judgment adopts an "external" view of the prediction problem rather than an "internal" one focusing simply on the problem at hand; it views this problem as one of a similar set to which a solution is required (see Kahneman & Lovallo, 1993 ~~1990~~).

One goal that cannot be accomplished by such updating methods is that of allowing a clinician to make a judgment about a particular individual at a particular setting at a certain time in a manner that is unrelated to judgments about other people in other settings or at other times. Such judgments can be variously characterized as the exercise of expertise, or as "shooting from the hip" (Russo & Schoemaker, 1989, p. 143). In light of the research-covered in this chapter and elsewhere, we see little reason for endorsing such judgments. (Even the expert claiming to make a clinical judgment on a "purely intuitive" basis attempts to substantiate its validity on the basis of experience with *people* [plural] similar to the person about whom the judgment is made.) Thus, for example, the conservative stance of Rubin's work in using regressed weighting coefficients is in our view a virtue, not a problem. The major point that statistical models are every bit as amenable to modification as is intuitive judgment is not vitiated by our recommendation to proceed cautiously in such modifications.

We suggest that there is no reason *not* to use available statistical techniques to maximize the predictability of models, when in fact models have been shown to be superior to clinical judgment. The research has been focussed primarily on demonstrating, or often questioning, this superiority, without a simultaneous consideration of how best to take advantage of it. Although others (e.g., Kleinmuntz, 1990) have urged that people combine their heads with their models in prediction, we urge that people use their heads to improve their models.

## REFERENCES

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes, 37,* 93-110. (Reprinted in J. Dowie and A. Elstein, Eds., *Professional judgment,* Cambridge, England: Cambridge University Press, 1987.)

Beaver, W. H. (1966). *Empirical research and accounting: Selective studies.* Chicago, IL: University of Chicago, Graduate School of Business, Institute of Professional Accounting.

Blattberg, R. C., & Hoch, S. J. (1990 ~~in press~~). Database model and managerial intuition: 50% model + 50% manager. *Management Science*. *36*, 887-899.

Bloch, D. A., & Moses, L. E. (1988). Non-optimally weighted least squares. *The American Statistician, 42*, 50-53.

Bloom, R. F., & Brundage, E. G. (1947). Predictions of success in elementary school for enlisted personnel. In D. B. Stuit (Ed.), *Personnel research and test development in the Naval Bureau of Personnel* (pp. 233-261). Princeton, NJ: Princeton University Press.

Brannen, A. L., Godfrey, L. J., & Goetter, W. E. (1989). Prediction of outcome from critical illness: A comparison of clinical judgment with a prediction rule. *Archives of Internal Medicine, 149*, 1083-1086.

Breiman, L., & Freedman, D. (1983). How many variables should be entered into a regression equation? *Journal of the American Statistical Association, 78*, 131-136.

Carroll, B. (1987). Artificial intelligence expert systems for clinical diagnosis: Are they worth the effort? *Behavioral Science, 32*, 274-292.

Carroll, J. S., Winer, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review, 17*, 199-228.

Castellan, N. J., Jr. (1973). Comments on the "lens model" equation and the analysis of multiple-cue judgment tasks. *Psychometrika, 38*, 87-100.

Chow, W. M. (1965). Adaptive control of the exponential smoothing constant. *Journal of Industrial Engineering, 16*, 314-317.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*, 180-188.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist, 34*, 571-582.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

Dawes, R. M. (1993 ~~in press~~). Prediction of the future versus an understanding of the past: A basic asymetry. *American Journal of Psychology*, *106*, 1-24.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Deacon, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research, 10*, 167.

Drehmer, D. E., & Morris, G. W. (1981). Cross-validation with small samples: An algorithm for computing Gollob's estimator. *Educational and Psychological Measurement, 41*, 195-200.

Duncan, D. B., & Horn, S. D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association, 67*, 815-821.

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

Faust, D., Meehl, P E., & Dawes, R. M. (1990). Clinical and actuarial judgment: Response. *Science, 247*, 146-147.

Gergonne, J. D. (1817). Essai de dialectique rationelle. *Annales des mathematiques pures et appliques, 7*.

Glaser, D. (1964). *The effectiveness of a prison and parole system.* Indianapolis, IN: Bobbs-Merrill.

Goldberg, L. R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. *Psychological Monographs: General and Applied, 79* (No. 9).

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23,* 483-496.

Goldberg, L. R. (1977). Admission to the Ph.D. Program in the Department of Psychology at the University of Oregon. *American Psychologist, 32*, 663-668.

Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C, Acampora, D. A., Stasiulewicz, C, Walshon, J., Terranova, G., Gottlieb, L., Kobernick, M., Goldstein-Wayne, B., Copen, D., Daley, K., Brandt, A. A., Jones, D., Mellors, J., & Jakubowski, R. (1988). A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *The New England Journal of Medicine, 318* (13), 797-802.

Gollob, H. F. (1967, September). *Cross-validation using samples of size one.* Paper presented at the American Psychological Association meetings, Washington, DC.

Gottfredson, D., Wilkins, L. T., & Hoffman, T. B. (1978). *Guidelines for parole and sentencing.* Lexington, MA: Lexington Books.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*, 438-456.

Hammond, K. R., & Summers, D. A. (1965). Cognitive dependence on linear and nonlinear cues. *Psychological Review, 72*, 215-224.

Inwald, R. E. (1988). Five year follow-up study of departmental terminations as predicted by 16 pre-employment psychological indicators. *Journal of Applied Psychology, 73*, 703-710.

Kahneman, D., & Lovallo, D. (1993 ~~in press~~). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*, 17-31.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions ASME Journal of Basic Engineering, 82*, 35-45.

Kalman, R. E. (1963). New methods in Wiener filtering theory. In J. L. Bodanoff & F. Kozin (Eds.), *Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability* (pp. 270-388). New York: Wiley.

Kleinmuntz, B. (1990). Clinical and actuarial judgment. *Science, 247.*

Knaus, W. A., Wagner, D. P., & Lynn, J. (1991). Short-term mortality predictions for critically ill hospitalized adults: Science and ethics. *Science, 254*, 389-395.

Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, R. M., Behar, V. S., Floyd, W. L., Morris, J. J., Waugh, R. A., Whalen, R. E., & Rosati, R. A. (1986). Predicting outcome in coronary disease. *The American Journal of Medicine, 80*, 553-560.

Leirer, V O., Warner, P. D., Rose, T. L., & Yesavage, J. A. (1985, August). *Predictions of violence by high school students and clinicians.* Presented at the American Psychological Association Convention, Los Angeles.

Leli, D. A., & Filskov, S. B. (1984). Clinical deterioration associated with brain damage. *Journal of Clinical Psychology, 40*, 1435-1441.

Lerner, M. J. (1980). *The belief in a just world: A fundamental delusion.* New York: Plenum.

Lemer, M. J. (1987). Integrating societal and psychological rules of entitlement: The basic task of each social actor and fundamental problem for the social sciences. *Social Justice Research, 1*, 107-125.

Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance, 16*, 1-12.

Malmquist, C. P., & Meehl, P. E. (1978). Barabbas: A study in guilt-ridden homicide. *The International Review of Psycho-Analysis, 5*, 149-179.

Meehl, P. E. (1954). *Clinical versus statistical prediction.* Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370-375.

Miller, M., & Morris, N. (1988). *Violence and victims, 3,* 263-328.

Pearson, R. W., Ross, M., & Dawes, R. M. (1992 ~~in press~~). Personal recall and the limits of retrospective questions in surveys. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys.* New York: Sage.

Rubin, D. E. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association, 75*, 801-816.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

Russo, J. E., & Schoemaker, P. J. H. (1989). *Decision traps.* New York: Doubleday.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178-200.

Schilpp, P. A. (1949). *Albert Einstein: Philosopher-scientist.* Evanston, IL: Library of Living Philosophers.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. *Psychological Bulletin, 84*, 751-758.

Schofield, W., & Garrard, J. (1975). Longitudinal study of medical students selected for admissions to medical school by actuarial and committee methods. *British Journal of Medical Education, 9*, 86-90.

Sutton, G. C. (1989). How accurate is computer-aided diagnosis? *Lancet,* October 14, 905-908.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review, 71*, 528-530.

Tukey, J. W. (1948). Approximate weights. *Annals of Mathematical Statistics, 19,* 91-92.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 312-317.

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin, 85*, 267-273.

Wedding, D. (1983). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology, V*, 49-54.

Werner, P. D., Rose, T. L., & Yesavage, J. A. (1983). Reliability, accuracy, and decision making strategy in clinical predictions of imminent dangerousness. *Journal of Consulting and Clinical Psychology, 51*, 815-825. (Companion piece: Liere, V. O., Werner, P D., Rose, T. L., & Yesavage, J. A. *Predictions of violence by high school students and clinicians.* Paper presented at the 1984 Oregon Psychological Association Spring Convention, Newport, Oregon.)

Wiggins, N., & Cohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*, 100-106.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 8*, 23-30.

pdf by LJY, March 2016