Psychological Bulletin
1970, Vol. 73, No. 6, 422–432

MAN VERSUS MODEL OF MAN:

A RATIONALE, PLUS SOME EVIDENCE, FOR A METHOD OF IMPROVING ON CLINICAL INFERENCES [1]

LEWIS R. GOLDBERG [2]

University of Oregon and Oregon Research Institute

Clinical psychologists, physicians, and other professionals are typically called upon to combine cues to arrive at some diagnostic or prognostic decision. Mathematical representations of such clinical judges can often be constructed to capture critical aspects of their judgmental strategies. An analysis of the characteristics of such models permits a specification of the conditions under which the model itself will be a more valid predictor than will the man from whom it was derived. To ascertain whether such conditions are met in natural clinical decision making, data were reanalyzed from a study of the judgments of 29 clinical psychologists attempting to differentiate psychotic from neurotic patients on the basis of their MMPI profiles. The results of these analyses indicate that for this diagnostic task models of the men are generally more valid than the men themselves. Moreover, the finding occurred even when the models were constructed on a small set of cases, and then man and model competed on a completely new set.

A psychologist who works for a municipal Suicide Prevention Center uses the cues from telephone interviews to assess the probability that each of his callers will try to kill himself. A pathologist who works for a large University hospital uses medical laboratory findings to assess the likelihood that each of his patients has a particular disease. A stock-broker who works in a metropolitan brokerage house uses daily financial indexes to assess the chances of gain or loss for each of the stocks in his clients' portfolios. The admissions officer of a private college uses aptitude-test scores and ratings by high school teachers to estimate the odds that each of the applicants to his institution will perform creditably there. All four of these professionals can be viewed as relying on their clinical wisdom to combine fallible cues (predictors) so as to arrive at some overall prognostic or diagnostic decision (Goldberg, 1968).

If questioned, all four decision makers would admit that their judgments are not perfect: Some of the callers whom the psychologist feels are bluffing actually commit suicide; some of the pathologist's diagnoses are way off the mark; sometimes a stock, which the broker forecasts will gain, simply drops out of sight; and sometimes a student, whom the admissions officer admits to his college, flunks right out of it. What can be done to help such professionals make more accurate decisions?

Clinical versus Actuarial Prediction

If information can be obtained about the empirical relationships between the cues (predictors) and the criterion outcomes, then

[1] The ideas presented here are hardly novel. They rest on the clinical judgment research of Paul J. Hoffman, Kenneth R. Hammond, and James C. Naylor, each of whom has contributed substantially to the methodology utilized in this project. The present rationale for improving clinical inference was stimulated by a study designed by Leonard G. Rorer, and by the reactions to his data from two other Oregon Research Institute investigators: Robyn M. Dawes and Paul Slovic. Without the ideas and encouragement of Slovic, Dawes, and Rorer, this report could not have been written.

The data analyzed in the present study were originally collected by Paul E. Meehl, and all data analyses were programmed and carried out by William Chaplin. The author is deeply indebted to both Meehl and Chaplin. Data analyses were carried out at the Health Sciences Computing Facility, Los Angeles, California. The study was supported by Grant MH12972 and Grant MH10822 from the National Institute of Mental Health, United States Public Health Service.

[2] Requests for reprints should be sent to Lewis R. Goldberg, Oregon Research Institute, P.O. Box 3196, Eugene, Oregon 97403.

the decision maker can substitute for his own clinical judgments the empirically derived "best-fitting" actuarial model of the phenomena he seeks to predict, and thereby increase the accuracy of the resulting predictions. Such an enterprise, originally viewed with considerable disdain by clinical psychologists, has recently weathered a period of intense controversy (Gough, 1962; Meehl, 1954; Sawyer, 1966), and may soon become a reasonably well accepted procedure in psychology—if not in medicine, stock forecasting, and other professional endeavors. Consequently, it now seems safe to assert rather dogmatically that when acceptable criterion information is available, the proper role of the human in the decision-making process is that of a scientist: (*a*) discovering or identifying new cues which will improve predictive accuracy, and (*b*) constructing new sorts of systematic procedures for combining predictors in increasingly more optimal ways. The resulting success of the clinician-as-scientist, when he is relegated to such an hypothesis-generating role, obviously must be evaluated empirically.

Improving on Clinical Inferences

Where criterion information is available, research is the obvious solution—hopefully providing the answers to many of the applied problems we face today. But, can anything be done to improve the accuracy of our forecasts in those situations where criterion information is *not* available?

Given a clinical judge making important decisions in a situation where (*a*) the environmental structure is as yet unknown, and (*b*) criterion information is at best available only on an irregular and unreliable basis, and consequently (*c*) criterion class base-rates are also unknown—thus obviating the usefulness of a procedure like mixed-group analysis (Dawes, 1967; Dawes & Meehl, 1966)—is there any procedure which might generally improve on the judge's forecasts? The psychologist at the Suicide Prevention Center is in roughly this situation. As an experienced clinician who has worked in his setting for the past several years, he has received some aperiodic indications of the sort of patient who does and does not commit suicide. While he knows that he is far from infallible, he feels that he has learned something from his experience. He is able to code reliably the various cues available from the telephone interview (e.g., client's sex, education, number and severity of past suicide attempts, availability of a weapon, etc.). While we do not know the actual validity of his predictions from these cues, let us pretend that the correlation between his prognostic ratings and the true state of affairs hovers around .40. Is there any way—short of collecting some criterion data—that we could boost that figure a bit, say up to .50? Seemingly, the answer is no.

Actually, however, the answer may be yes. For the clinician is not a machine. While he possesses his full share of human learning and hypothesis-generating skills, he lacks the machine's reliability. He "has his days": Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configuration are not identical. He is subject to all those human frailties which lower the reliability of his judgments below unity. And, if the judge's reliability is less than unity, there must be error in his judgments—error which can serve no other purpose than to attenuate his accuracy. If we could remove some of this human unreliability by eliminating the random error in his judgments, we should thereby increase the validity of the resulting predictions. The problem, then, may be reformulated: Can the clinician's judgmental unreliability be separated from his—hopefully, somewhat valid—judgmental strategy?

Capturing the Policies of Clinical Judges

Ten years of research on the clinical judgment process have demonstrated that for many types of common clinical decisions and for many sorts of clinical judges, a simple linear regression equation can be constructed which will predict the responses of a judge at approximately the level of his own reliability. For documentation of this assertion and for details of the methodology, see Hoffman (1960), Hammond, Hursch, and Todd (1964), Naylor and Wherry (1965), and Goldberg (1968). While such regression models have

been utilized (probably somewhat inappropriately) to *explain* the manner in which clinicians combine cues in making their diagnostic and prognostic decisions (see Green, 1968; Hoffman, 1968), there is little controversy about their power as *predictors* of the clinical judgments. In addition, of course, such "paramorphic" (Hoffman, 1960) models of human decision making all possess at least one asset which humans typically lack: perfect reliability.

These judgmental representations are constructed in the following fashion: A clinician (henceforth referred to as a "clinical judge" or simply a "judge") is asked to make his diagnostic or prognostic judgments from a previously quantified set of K cues (predictors) for each of M target individuals (e.g., patients, applicants, etc.). These M judgments, when quantified, are then used as the dependent variable in a standard linear regression analysis. The independent variables in this analysis are the values of the K predictors. The results of such an analysis are a set of K regression weights, one for each predictor, and these sets of regression weights are referred to as the judge's "model" (Hoffman, 1960) or his "policy" (Naylor & Wherry, 1965).

Now, how would such models fare as predictors themselves? That is, if the set of K regression weights (generated from an analysis of one clinical judge) were used to obtain a "predicted score" for each target individual, would these scores be more valid, or less valid, than the original clinical judgments from which the regression weights were derived? To the extent that the model fails to capture valid nonlinear variance in the judge's decision processes, it should perform less creditably than the judge. To the extent that it eliminates the random error component in human judgments, it should perform more validly than the judge. Which of these counteracting factors is more important in typical clinical decision making? Can we construct a mathematical representation of a clinical judge—without any recourse to criterion information—which is more valid as a decision maker than the human we have used as a model?

Man versus Model of Man: The Mathematics of their Comparison

If we can specify the conditions under which models of men will surpass their human counterparts, then we should be better able to decide on the sorts of clinical problems, if any, where models might profitably be substituted for men. Fortunately, the pioneering formulations of Hammond, Hursch, and Todd (1964) permit these conditions to be specified rather precisely.

Tucker (1964) has reformulated the fundamental "lens model" equation of Hammond et al. (1964), as follows:

$$r_a = GR_eR_s + C\sqrt{1 - R_e^2}\sqrt{1 - R_s^2} \quad [1]$$

where

r_a = the validity coefficient of the judge: the correlation between the judge's predictions and the actual criterion values ($r_{Y_s \cdot Y_e}$).

G = the linear component of judgmental accuracy: the correlation between the predicted scores from the linear model of the judge and those from the linear model of the criterion ($r_{\hat{Y}_s \cdot \hat{Y}_e}$).

R_e = the linear predictability of the criterion: the multiple correlation between the cues and the criterion values ($r_{Y_e \cdot \hat{Y}_e}$).

R_s = the linear predictability of the judge: the multiple correlation between the cues and the judge's predictions ($r_{Y_s \cdot \hat{Y}_s}$).

C = the nonlinear component of judgmental accuracy: the correlation between the residual values of the criterion and the residual values of the judge's predictions after the linear components in both the criterion and the judge have been removed.

$$C = \frac{r_a - GR_eR_s}{\sqrt{1 - R_e^2}\sqrt{1 - R_s^2}} \quad [2]$$

and where, preserving the original notation, the subscript e refers to the environment (e.g., criterion), and s to the subject (e.g., judge);

Y_e refers to the criterion values, and Y_s to the clinical judgments; and $\hat{Y}_e$ and $\hat{Y}_s$ are the predicted components of Y_e and Y_s, respectively, using the predictor (cue) variables in a standard linear regression analysis.

Now, let us define two more terms:

r_m = the validity coefficient of the linear model of the judge: the correlation between the predicted scores from the judge's model and the actual criterion values ($r_{\hat{Y}_s \cdot Y_e}$).

Δ = the differential validity of model over man: the difference in validity coefficients between the model (r_m) and the man (r_a).

$$\Delta = r_m - r_a \qquad [3]$$

If r_m is substituted for r_a in Equation 1, then

$$R_s = r_{\hat{Y}_s \cdot \hat{Y}_s} = 1$$

and, therefore:

$$r_m = GR_e(1) + C\sqrt{1 - R_e^2}(0)$$
$$r_m = GR_e \qquad [4]$$

Substituting Equations 1 and 4 in Equation 3, we obtain:

$$\Delta = GR_e - (GR_eR_s + C\sqrt{1 - R_e^2}\sqrt{1 - R_s^2})$$
$$\Delta = GR_e(1 - R_s) - C\sqrt{1 - R_e^2}\sqrt{1 - R_s^2} \qquad [5]$$

Equation 5 specifies the conditions under which models will outperform men (i.e., when $GR_e(1 - R_s) > C\sqrt{1 - R_e^2}\sqrt{1 - R_s^2}$). Note that when R_e is unity (i.e., when the criterion is perfectly predictable from a linear model) and man has *any* positive validity, the model will always surpass the man. And, as R_s approaches unity, Δ approaches zero—indicating that as man becomes more linearly predictable, his judgments become less distinguishable from those of his model, and thus the differences between man and model disappear.

Since we know that in natural settings neither man nor environment is perfectly predictable, it becomes an empirical question whether the conditions favoring model over man *generally* hold. The following study was designed to provide a preliminary answer to this important question.

Man versus Model of Man: An Illustrative Example

While the present method is being proposed only for those situations where criterion information is *not* available, an experimental test of the procedure demands the utilization of a decision-making task where criterion data has already been collected. If human judges do not turn out to make more valid predictions than do their models in situations where criterion information is available, then it is reasonable to assume that this same result would be obtained in settings where such information is unknown.

The Judgment Task

The judgmental problem used for this study was that of differentiating psychotic from neurotic patients on the basis of their MMPI profiles. Paul Meehl (1959) originally focused research attention in this arena on the grounds that

> the differences between psychotic and neurotic profiles are considered in MMPI lore to be highly configural in character, so that an atomistic treatment by combining single scales linearly should theoretically be a very poor substitute for a configural approach [p. 104].

While subsequent research has failed to confirm this conjecture (Goldberg, 1965, 1969), the fact that the clinicians utilized in this study believed it to be true may allow some generalization of the present findings to other clinical situations where rather complicated interactions are presumed.

The MMPI profiles from 861 psychiatric patients who had been diagnosed as being clearly either psychotic or neurotic were used as the predictors (cues) in this study; roughly half of these patients had been diagnosed psychotic. These materials, collected by Meehl from seven hospitals and clinics throughout the United States, have been described in a number of previous publications (e.g., Goldberg, 1965, 1968, 1969; Meehl, 1959; Meehl & Dahlstrom, 1960).

The Judges

Meehl engaged 29 clinical psychologists of varying experience and training to make a diagnostic decision on the basis of each

MMPI profile. The 29 judges rated each profile on an 11-step–forced-normal distribution from least to most psychotic, within each of the seven samples. If the judge's validity coefficient—the point-biserial correlation between his clinical judgments and the actual (dichotomous) criterion diagnoses (r_a)—is used as an index of diagnostic accuracy, then any differences between judges in their presumptions regarding the base-rates of psychosis will not affect the present findings.

The Models of the Judges

For each clinical judge, least-squares regression weights were computed using the clinician's own judgmental responses as the dependent variable and the 11 MMPI scale scores as the independent variables; these 11 regression weights constitute the judge's model. A "composite judge" was created by averaging the judgments of all 29 clinicians to each MMPI profile. A model of this hybrid clinician, which yields a validity coefficient identical to that of the average of the predictions from the 29 individual judgmental models, was also obtained.

The Comparative Validity of Man, Model, and Composite

Since criterion information was available for this task, it was possible to compare the validity coefficients of each judge's model (r_m) with that achieved by the judge himself (r_a), and also to ascertain the actual values of G, R_e, R_s, and C in Equation 5. Moreover, since each clinician judged MMPI profiles from seven clinical settings, it was possible to compare these values across the samples in order to check on the generality of the findings over different sorts of clinical populations. Table 1 presents the results of these analyses. The terms in Equation 5 for predicting the differential validity of model and man are used as column headings in Table 1, and their values are presented for each sample, as well as for the total sample. Descriptions

TABLE 1

THE VALIDITY OF MAN VERSUS MODEL OF MAN FOR THE TYPICAL AND THE COMPOSITE CLINICIAN WITHIN EACH OF SEVEN SAMPLES OF MMPI PROFILES

Sample	N	$\Delta = r_m - r_a = G$	R_e	$1 - R_s - C$	$\sqrt{1 - R_e^2}$	$\sqrt{1 - R_a^2}$
		Typical judge				
A	92	−.004 .42 .43 .77	.55	.17 .17	.84	.55
B*	77	.019 .18 .16 .37	.49	.16 .03	.87	.54
C	103	.027 .29 .26 .58	.50	.15 .03	.87	.52
D*	42	.104 .25 .15 .38	.66	.16 −.16	.76	.54
E*	181	−.067 .13 .20 .24	.55	.22 .19	.83	.62
F	166	.043 .25 .20 .50	.49	.21 .02	.87	.61
G	200	.053 .48 .43 .74	.65	.16 .05	.76	.54
Total	861	.028 .31 .28 .68	.46	.23 .08	.89	.64
		Composite judge				
A	92	−.058 .46 .52 .83	.55	.07 .29	.84	.37
B*	77	−.001 .20 .20 .41	.49	.06 .04	.87	.34
C	103	−.006 .31 .32 .63	.50	.04 .08	.87	.30
D*	42	.108 .30 .19 .45	.66	.08 −.28	.76	.40
E*	181	−.107 .15 .26 .27	.55	.10 .33	.83	.45
F	166	.016 .27 .26 .56	.49	.11 .04	.87	.45
G	200	.003 .51 .51 .79	.65	.06 .10	.76	.34
Total	861	−.017 .33 .35 .72	.46	.11 .13	.89	.45

Note:—For an explanation of the column headings, see text.
* Uncontaminated samples (see Goldberg, 1965).

TABLE 2

A COMPARISON OF THE VALIDITY COEFFICIENTS FOR THE MOST ACCURATE, COMPOSITE, TYPICAL, AND LEAST ACCURATE CLINICIANS (r_a) AND MODELS (r_m) IN EACH SAMPLE

Condition	Sample							Total
	A	B	C	D	E	F	G	
Judge								
Most Accurate	.56	.32	.45	.43	.34	.33	.55	.39
Composite	.52	.20	.32	.19	.26	.26	.51	.35
Typical (Mean)	.43	.16	.26	.15	.20	.20	.43	.28
Least Accurate	.18	−.10	.04	−.12	.01	.09	.23	.14
Model								
Most Accurate	.52	.35	.43	.55	.32	.39	.60	.43
Composite	.46	.20	.31	.30	.15	.27	.51	.33
Typical (Mean)	.42	.18	.29	.25	.13	.25	.48	.31
Least Accurate	.29	−.06	.09	−.01	−.11	.04	.25	.16
N	92	77	103	42	181	166	200	861
% Psychotic	64	52	47	52	50	43	37	47
Actuarial Formula ($L+Pa+Sc-Hy-Pt$)	.39	.45	.42	.46	.27	.42	.56	.44
Linear Predictability (R_c)	.55	.49	.50	.66	.55	.49	.65	.46

of each sample can be found in Goldberg (1965).

The upper half of Table 1 presents the findings for the typical judge (the mean values of each index across the 29 clinicians). In five out of the seven samples (and in the total sample) the typical judge produced a positive Δ value (a positive Δ indicates that the model was more valid than the judge). Only in one sample was there any sizable discrepancy favoring the judge over his model. The values of G (the correlation between the linear model of the judge and the linear model of the criterion) ranged across the seven samples from .24 to .77. While the values of C (the nonlinear analogue of G) ranged from −.16 to .19, in all but the smallest sample the average value of C was positive, indicating at least some slight nonlinear contribution to the accuracy of the clinicians' judgments (see Wiggins & Hoffman, 1968). This latter finding makes the major finding (that judges are not more valid than linear models) all the more remarkable.

The lower half of Table 1 presents the findings for the composite judge (the average judgment of all 29 clinicians to each MMPI profile). In contrast to the findings for the typical individual judge, modeling did not serve to increase the diagnostic accuracy of the judgmental composite. In four of the seven samples (and in the total sample) the composite judge was at least as valid as its model. Moreover, the values of all the terms in Equations 3 and 5, other than R_c, were higher for the composite than for the typical judge. The largest of these differences occurred for R_s, indicating that the unreliability which attenuated the validity of the individual clinicians has been substantially removed from the composite judge.

Table 2 presents a comparison of the validity coefficients for the most accurate, composite, typical, and least accurate clinicians (r_a) and models (r_m) in each of the seven samples, and in the total sample. The most accurate and least accurate judges differed from sample to sample (Goldberg, 1965), and the most accurate judge did not necessarily produce the most accurate model. Also included in Table 2 are the multiple correlation coefficients based upon all 11 MMPI scale scores (R_c), as well as the validity coefficients for one actuarial formula ($L + Pa + Sc - Hy - Pt$). For a discussion

of the procedures used to derive the actuarial formula, see Goldberg (1965).

A ranking of the validity coefficients based upon the total sample ($N = 861$) generates the following order:

Linear Predictability (R_e)	.46
Actuarial Formula	.44
Most Accurate Model	.43
Most Accurate Judge	.39
Composite Judge	.35
Model of Composite Judge	.33
Typical Model	.31
Typical Judge	.28
Least Accurate Model	.16
Least Accurate Judge	.14

Note that for this task a relatively simple actuarial formula ($r = .44$) and the most accurate judgmental model ($r = .43$) were almost as valid as the optimal linear regression function ($R = .46$). Moreover, the composite judge ($r = .35$) and its model ($r = .33$) were somewhat more valid than the typical model ($r = .31$) and the typical judge ($r = .28$).

Consequently, in settings where the incremental accuracy of composite over typical judge does not justify the inefficiency of having all cases diagnosed by all judges, a model of the composite judge might profitably be substituted for the individual judgments. But, if these findings are to be generalized to other clinical settings, it becomes especially important to discover the extent to which models based upon a sample of cases will fare as predictors when cross-validated on new cases. The present data can also provide a preliminary answer to this critical question.

Comparing Man and Model on New Cases

Since 861 MMPI profiles were judged by each clinician, it was possible to build his model using successively smaller samples of cases, cross-validate the model on the remaining cases, and then compare the cross-validity of the model with that of the judge himself when both were compared on a completely new set of cases.

To carry out these analyses, the validity of each judge and his model was compared:

(*a*) using *all* the cases in the sample;

(*b*) splitting the sample into two equal parts (odd half versus even half), constructing the judges' models using the cases in each half in turn, calculating the validity of the judges and their models on the other half, and then averaging these "cross-validities" across both half samples;

(*c*) splitting the samples into three equal parts (Profiles 1, 4, 7, . . . versus 2, 5, 8, . . . versus 3, 6, 9, . . .), constructing the models using only a third of the cases, calculating "cross-validities" on the remaining two-thirds of the cases, and then averaging these three indexes of "cross-validity";

(*d*) repeating the procedure for derivation samples of one-fourth, one-fifth, one-sixth, one-seventh, one-eighth, one-ninth, and finally one-tenth the original sample size. While only selected derivation sample sizes are summarized in the tables to follow, the others are available from the author.

Table 3 presents these comparisons of each clinician with his model, based upon the total set of 861 profiles. The 29 judges have been ordered from the top to the bottom of Table 3 on their validity coefficients for these MMPI profiles—from a high of .39 to a low of .14. When models of each clinician were constructed on the basis of the 861 cases, 86% of these models turned out to be more accurate predictors of the actual criterion diagnoses than were the clinicians from whom they were derived ($p < .001$; Sign Test). Moreover, if one accepts as "essentially identical" those cases where the differences between man and model were zero when rounded to two decimal places, then for 28 of the 29 judges (97%), the models were at least as valid as were the judges themselves. In no case was there a sizable superiority of man over his model.

As Table 3 indicates, when models were constructed on one-half of the cases, in 86% of the comparisons the model was more accurate—on the remaining half of the cases—than was the clinician. When the model was derived on only one-seventh of the cases, it was still superior to its human counterpart 79% of the time. Even when the sample was reduced to one-tenth of its original size, and both man and model were compared on the remaining nine-tenths of the cases, models fared better than judges in 72% of these comparisons.

TABLE 3

DIFFERENCES IN VALIDITY COEFFICIENTS BETWEEN THE MODEL OF EACH CLINICAL JUDGE AND HIS OWN JUDGMENTS

Judge		Validity	Δ				
Derivation Sample			861	430	215	123	86
Cross-Validation Sample			0	431	646	738	775
(No. of Samples Averaged)				(2)	(4)	(7)	(10)
Most Accurate	1	(.39)	.03	.03	.02	.02	.02
	2	(.38)	.01	.01	.00	.00	−.00
	3	(.36)	.07	.07	.06	.06	.05
	4	(.36)	.02	.02	.01	.01	.01
	5	(.35)	.01	.00	−.01	−.00	−.01
	6	(.33)	.03	.02	.02	.02	.01
	7	(.33)	−.00	−.00	−.00	−.01	−.01
	8	(.33)	.01	.00	.00	.01	−.00
	9	(.32)	.05	.05	.05	.04	.04
	10	(.31)	.07	.08	.07	.06	.05
	11	(.31)	.01	.01	.01	.01	.01
	12	(.31)	.07	.07	.06	.06	.06
	13	(.30)	.06	.06	.06	.06	.04
	14	(.30)	.04	.04	.04	.04	.03
Middle Judge	15	(.30)	.05	.05	.05	.05	.05
	16	(.29)	.04	.04	.03	.03	.03
	17	(.29)	.02	.02	.02	.02	.01
	18	(.28)	.02	.02	.02	.02	.01
	19	(.26)	.06	.06	.06	.06	.06
	20	(.26)	.05	.05	.04	.04	.03
	21	(.25)	−.00	−.01	−.01	−.00	−.01
	22	(.24)	.01	.00	.01	.00	.00
	23	(.23)	.01	.01	.01	.01	.00
	24	(.22)	−.00	−.00	−.01	−.00	−.00
	25	(.22)	.00	.00	−.01	−.01	−.01
	26	(.21)	.03	.03	.03	.03	.02
	27	(.17)	−.01	−.01	−.01	−.01	−.02
	28	(.15)	.04	.03	.02	.03	.01
Least Accurate	29	(.14)	.02	.02	.02	.02	.02
Average		(.28)	.03	.03	.02	.02	.02
Composite Judge		(.35)	−.02	−.02	−.02	−.02	−.02
% of Cases Model > Man			86	86	79	79	72

Note:—Total sample (N = 861). The values in the first column of this table are point-biserial correlations between the clinicians' judgments (scaled 1–9) and the actual criterion diagnoses (coded 0 or 1), based upon all 861 cases (r_{pb}). (The corresponding validity coefficients for all of the other columns in the table are virtually identical.) The values presented in the other columns are the arithmetic differences between the validity coefficient of the judge's model and that of the judge, himself (Δ). Positive differences favor the model; negative differences favor the judge. For differences which round to zero, the sign of the (four place) computed value has been retained.

One final analysis was focused on the 200 MMPI profiles from one setting (Sample G) —the sample for which the clinicians' judgments were most accurate. Table 4 presents the comparisons of each clinician with his model for these 200 cases, the 29 judges again having been ordered on their overall diagnostic accuracy for profiles from this sample. When models of each clinician were constructed on the basis of these 200 cases, every one of the 29 models turned out to be a more accurate predictor of the actual criterion diagnoses than was the clinician from whom it was derived. When models were constructed on only one-third of the cases, in 28 out of 29 comparisons (97%) the model was more accurate on the remaining two-thirds of the cases than was the clinician. When the model was derived on only one-quarter of the cases, it was still superior to its human counterpart 86% of the time. Even when the sample was reduced to one-sixth of its original size (11 predictors and

TABLE 4

DIFFERENCES IN VALIDITY COEFFICIENTS BETWEEN THE MODEL OF EACH CLINICAL JUDGE AND HIS OWN JUDGMENTS

Judge		Validity	Δ			
Derivation Sample			200	66	50	33
Cross-Validation Sample			0	134	150	167
(No. of Samples Averaged)				(3)	(4)	(6)
Most Accurate	1	(.55)	.02	.01	.00	−.03
	2	(.54)	.06	.04	.04	.00
	3	(.53)	.03	.02	.01	.00
	4	(.53)	.05	.03	.02	−.00
	5	(.53)	.04	.02	.01	−.01
	6	(.49)	.02	−.01	−.04	−.05
	7	(.48)	.04	.03	.01	.01
	8	(.47)	.05	.03	.02	.00
	9	(.47)	.09	.09	.07	.07
	10	(.47)	.07	.07	.05	.04
	11	(.46)	.03	.02	.01	.01
	12	(.46)	.04	.02	.01	.01
	13	(.45)	.06	.05	.06	.04
	14	(.44)	.11	.08	.08	.02
Middle Judge	15	(.44)	.07	.05	.04	.03
	16	(.44)	.02	.00	−.02	−.01
	17	(.42)	.05	.04	.02	−.00
	18	(.40)	.02	.01	−.01	−.02
	19	(.40)	.05	.06	.02	.04
	20	(.40)	.08	.07	.06	.03
	21	(.39)	.03	.02	.00	−.01
	22	(.38)	.02	.01	.01	−.01
	23	(.38)	.03	.01	−.00	−.01
	24	(.37)	.14	.12	.09	.06
	25	(.36)	.04	.02	.01	−.01
	26	(.36)	.06	.04	.02	.02
	27	(.31)	.06	.07	.06	.05
	28	(.26)	.11	.10	.06	−.04
Least Accurate	29	(.23)	.03	.01	.01	.01
Average		(.43)	.05	.04	.02	.01
Composite Judge		(.51)	.00	.00	−.01	−.01
% of Cases Model > Man			100	97	86	62

Note:—See Note to Table 3. Sample G (N = 200).

only 33 cases) models fared better than judges in 62% of the cross-validated comparisons.

Discussion

The findings presented in Tables 1–4 suggest the following conclusions, each of which will be discussed in turn:

1. Within the generality restrictions obviously imposed by any one study, it has been demonstrated that linear regression models of clinical judges can be more accurate diagnostic predictors than are the humans who are modeled. Moreover, even when the model of the judge was constructed on a relatively small subset of cases, and then the clinician and his model competed on the remaining (larger) set of additional cases, this same effect occurred.

2. The composite judgment of all 29 clinicians, which was more accurate than that of the typical individual judge, was *not* improved by the modeling procedure.

Man versus Models of Man

The present findings demonstrate that, at least in one clinical situation, clinical judgments can be improved upon without recourse to any criterion data, whatsoever. Are these results surprising? Whatever the reader's own reactions, it should be useful to realize the extent to which these findings have already been anticipated. For example, Yntema and Torgerson (1961) had this to say about man and his (computerized) models: "Men and computers could cooperate more efficiently . . . if a man could tell the computers how he wanted decisions made, and then let the machine make the decisions for him [p. 20]." They then went on to describe the following "intriguing possibility": "Artificial, precomputed judgments may in some cases be better than those the man could make himself if he dealt with each situation as it arose [p. 24]."

Yntema and Torgerson reported a study in which subjects were asked to learn a rather complicated configural pattern involving three cues (size, shape, and color), two of which had six values and one of which had five. The 180 possible combinations of these cues were presented to the subjects repeatedly, under conditions of outcome feedback. By the end of the training period, one subject had achieved a validity coefficient of .84; a linear model of his judgmental responses yielded a validity coefficient of .89. This evidence favoring model over man is extremely compelling, since we know a priori that the true state of affairs, far from being linear, was set up by the experimenters to be configural.

A related study by Ward and Davis (1965) has been described as follows:

> An alternative mentioned by Ward and Davis but not specifically applied to classification is to feed the computer the data describing persons, together with the assignments actually made by the institutional decision maker. The computer develops regression weights for predicting the decision that would be made regarding any subsequent case. This "predicted decision" for a new person then becomes his assignment. Such a method substitutes "machine judgment" for human judgment by reproducing the human's implicit strategy within the computer. Computer judgments will be more consistent and therefore somewhat more valid than human judgments from the same information; in effect, the validity is raised by correcting for "attenuation" arising from fluctuations in the judge's standards and attentiveness, and, if several judges are used in establishing weights, for attenuation due to idiosyncracies of the judges. But if the judges have common misconceptions about the validity of any of the information and so weight it improperly, the simulation method will also use a less-than-optimal strategy [Cronbach & Gleser, 1965; p. 164].

A similar rationale has been applied to management decision making by Bowman (1963) and Kunreuther (1969).

In another type of application, Naylor and his collaborators have proposed that if the correlation between the regression models of raters are substituted for the usual correlation between the judges' ratings in an inverse factor analysis of the raters, then more interpretable rater groupings emerge:

> let us now define the policy of a judge as his least-squares regression equation. In other words, his policy is defined as our best prediction of what he will do (given our model), rather than what he actually does [Naylor & Schenck, 1966; p. 818].

> It can be argued that such "policy agreement" correlations based upon predicted values are better [than the typical correlations as] indices of the similarity or "matching" of two raters' policies . . . [Naylor, Dudycha, & Schenck, 1967; p. 7].

In an important theoretical and empirical attack on this problem, Dudycha and Naylor (1966) suggested that

> humans tend to generate "correct" strategies but then, in turn, fail to use their own strategy with any great consistency. . . . One is left with the conclusion that humans may be used to generate inference strategies but that once the strategy is obtained the human should be removed from the system and replaced by his own strategy! [Dudycha & Naylor, 1966; p. 127].

The findings from the present study certainly corroborate this conclusion.

Modeling the Composite Judge

Given enough clinical judges, all possessing at least some validity, the most accurate predictions (in situations where criterion information is lacking) may well come from the composite judgments of the total group. Moreover, when the set of judges is as large as in the present study ($N = 29$), this composite judgment may not be improvable by modeling techniques. On the other hand, when only a few judges are available, it is possible that their composite might be improved by modeling. Cooke (1967a, 1967b) investigated this possibility, using the same clinical problem studied earlier by Kleinmuntz (1963): predicting the adjustment of college students from their MMPI profiles. Cooke employed six MMPI experts, obtained a composite judgment, and then fitted a linear regression equation to the latter. As classical test theory would predict, Cooke found that the test-retest reliability coefficient of the composite was higher than that of any single judge. Unfortunately, Cooke's analyses do not permit any confident conclusions regarding the validity of the judges versus that of the composite and its model.

On the other hand, the fact that composite judgments are predictively superior to the average of individual judgments has already received considerable support. For example, Winkler (1967) has carried out an unusually thorough analysis of the judgments of experts estimating the point-spread between winning and losing football teams over a season of weekly encounters. In this context, Winkler found that a composite of the individual judgments "in all cases . . . performed better than the average of the individual subjects' scores [p. 393]."

Moreover, in the present study the composite judge was more valid than the average of the *modeled* individual judgments (see Tables 1 and 2). This finding suggests that pooling the responses of a set of judges may generally be a more powerful technique than modeling as a means of improving the accuracy of clinical judgments. Unfortunately, the use of a composite judge is an extremely inefficient procedure, since each clinician must judge all of the cases. Consequently, it would be desirable to find some method of locating a single judge whose model outperforms the composite, thus eventually eliminating the clinical judgments altogether. However, in situations where criterion information is unavailable, it is not apparent how the best judge can be spotted. Analyses of the present data indicate that neither R_s nor the correlation of the clinician with the judgmental composite will unfailingly locate such a judge. For example, the similarity of the clinician to the composite bore a curvilinear relationship to the model's validity: Models of judges who were highly similar to the composite performed approximately as well as it did, while models of judges who were less similar to the composite achieved either quite superior or quite inferior validities. And, R_s bore no relationship to judgmental accuracy.

Consequently, in situations where clinicians' time is at a premium, it might be well to model either the judge with the highest correlation with the composite or the composite itself—thereby insuring no substantial loss over the composite judgments, and some gain over the validity of the typical clinician's judgments. On the other hand, in situations where clinicians' time is of no great concern, or where a small increase in validity is extremely important, then a composite judgment might be used.

Proponents of actuarial prediction have repeatedly emphasized the lower cost of such procedures when compared to the cost of human experts. Analogously, one should realize that the use of judgmental models is inherently less costly than the use of human judges, for after an expert has been used to derive his model he is free to perform other

activities. Consequently, if cost is a factor in deciding between the use of men or their models, then in many situations judges would have to be substantially more valid than their models before the overall utilities would favor the continued use of expensive professional time. However, in the present study there were no instances of any substantial incremental validity of man over his model. Consequently, if these findings can be generalized to other sorts of judgmental problems, it would appear that only rarely—if at all—will the utilities favor the continued employment of a man over a model of man.

REFERENCES

BOWMAN, E. H. Consistency and optimality in managerial decision making. *Management Science,* 1963, **9**, 310–321.

COOKE, J. K. Clinicians' decisions as a basis for deriving actuarial formulae. *Journal of Clinical Psychology,* 1967, **23**, 232–233. (a)

COOKE, J. K. MMPI in actuarial diagnosis of psychological disturbance among college males. *Journal of Counseling Psychology,* 1967, **14**, 474–477. (b)

CRONBACH, L. J., & GLESER, G. C. *Psychological tests and personnel decisions.* Urbana, Ill.: University of Illinois, 1965.

DAWES, R. M. How clinical probability judgment may be used to validate diagnostic signs. *Journal of Clinical Psychology,* 1967, **23**, 403–410.

DAWES, R. M., & MEEHL, P. E. Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin,* 1966, **66**, 63–67.

DUDYCHA, L. W., & NAYLOR, J. C. Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance,* 1966, **1**, 110–128.

GOLDBERG, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs,* 1965, **79**, (9, Whole No. 602).

GOLDBERG, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist,* 1968, **23**, 483–496.

GOLDBERG, L. R. The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavioral Research,* 1969, **4**, 523–536.

GOUGH, H. G. Clinical vs. statistical prediction in psychology. In L. Postman (Ed.), *Psychology in the making.* New York: Knopf, 1962.

GREEN, B. F., JR. Descriptions and explanations: A comment on papers by Hoffman and Edwards. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

HAMMOND, K. R., HURSCH, C. J., & TODD, F. J. Analyzing the components of clinical inference. *Psychological Review,* 1964, **71**, 438–456.

HOFFMAN, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin,* 1960, **57**, 116–131.

HOFFMAN, P. J. Cue-consistency and configurality in human judgment. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

KLEINMUNTZ, B. MMPI decision rules for the identification of college maladjustment: A digital computer approach. *Psychological Monographs,* 1963, 77 (14, Whole No. 577).

KUNREUTHER, H. Extensions of Bowman's theory on managerial decision-making. *Management Science,* 1969, **15**, 415–439.

MEEHL, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press, 1954.

MEEHL, P. E. A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology,* 1959, **6**, 102–109.

MEEHL, P. E., & DAHLSTROM, W. G. Objective configural rules for discriminating psychotic from neurotic MMPI profiles. *Journal of Consulting Psychology,* 1960, **24**, 375–387.

NAYLOR, J. C., DUDYCHA, A. L., & SCHENCK, E. A. An empirical comparison of ρ_a and ρ_m as indices of rater policy agreement. *Educational and Psychological Measurement,* 1967, **27**, 7–20.

NAYLOR, J. C., & SCHENCK, E. A. ρ_m as an "error-free" index of rater agreement. *Educational and Psychological Measurement,* 1966, **26**, 815–824.

NAYLOR, J. C., & WHERRY, R. J., SR. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement,* 1965, **25**, 969–986.

SAWYER, J. Measurement *and* prediction, clinical *and* statistical. *Psychological Bulletin,* 1966, **66**, 178–200.

TUCKER, L. R. A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review,* 1964, **71**, 528–530.

WARD, J. H., JR., & DAVIS, K. Teaching a digital computer to assist in making decisions. In L. J. Cronbach, & G. C. Gleser, *Psychological tests and personnel decisions.* Urbana, Ill.: University of Illinois, 1965.

WIGGINS, N., & HOFFMAN, P. J. Three models of clinical judgment. *Journal of Abnormal Psychology,* 1968, **73**, 70–77.

WINKLER, R. L. The quantification of judgment: Some experimental results. *Proceedings of the American Statistical Association,* 1967, **62**, 386–395.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics,* 1961, Vol. HFE-2, No. 1, 20–26.

(Received May 29, 1969)