Psychological Assessment
2000, Vol. 12, No. 1, 19–30

Copyright 2000 by the American Psychological Association, Inc.
1040-3590/00/$5.00 DOI: 10.1037//1040-3590.12.1.19

Clinical Versus Mechanical Prediction: A Meta-Analysis

William M. Grove, David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson
University of Minnesota, Twin Cities Campus

The process of making judgments and decisions requires a method for combining data. To compare the accuracy of clinical and mechanical (formal, statistical) data-combination techniques, we performed a meta-analysis on studies of human health and behavior. On average, mechanical-prediction techniques were about 10% more accurate than clinical predictions. Depending on the specific analysis, mechanical prediction substantially outperformed clinical prediction in 33%–47% of studies examined. Although clinical predictions were often as accurate as mechanical predictions, in only a few studies (6%–16%) were they substantially more accurate. Superiority for mechanical-prediction techniques was consistent, regardless of the judgment task, type of judges, judges' amounts of experience, or the types of data being combined. Clinical predictions performed relatively less well when predictors included clinical interview data. These data indicate that mechanical predictions of human behaviors are equal or superior to clinical prediction methods for a wide range of circumstances.

Two general classes of data combination procedures have been extensively studied in the psychological and medical literatures: clinical judgment and mechanical prediction. *Clinical judgment* refers to the typical procedure long used by applied psychologists and physicians, in which the judge puts data together using informal, subjective methods. Clinicians differ in how they do this: The very nature of the process tends to preclude precise specification.

Mechanical prediction, including statistical prediction (using explicit equations), actuarial prediction (as with insurance companies' actuarial tables), and what we may call algorithmic prediction (e.g., a computer program emulating expert judges), is by contrast well specified. Once developed, application of mechanical prediction requires no expert judgment. Also, mechanical predictions are 100% reproducible.

Clinical and mechanical predictions sometimes disagree (Meehl, 1986). One cannot both admit and not admit a college applicant, or parole and imprison a convict. To fail to grapple with the question—Which prediction should one follow?—is to ignore the obligation of beneficence. Hence, the comparison of clinical judgment and mechanical prediction takes on great importance.

Meehl's (1954) famous review of clinical versus statistical prediction inspired many psychologists to conduct studies on this topic. Sawyer's (1966) later review included 40 studies addressing mechanical versus clinical data combination. He concluded that mechanical prediction often outshines clinical prediction; that is, when it is not superior, it performs as well as clinical prediction. Since then, several reviews and polemics have appeared (Dawes, Faust, & Meehl, 1989; Garb, 1994; Holt, 1970; Marchese, 1992; Sines, 1971; Wiggins, 1981).

We report the results of the first completed meta-analysis to be conducted on studies comparing clinical and mechanical prediction. Except for Holt (1970), the previous reviews of this area have reached largely similar conclusions, favoring statistical prediction. So why bother to resurvey the field? Because, as Holt pointed out, the older reviews were flawed by problems like low statistical power, subjective decisions about superiority of one method versus another, and failure to examine study-design variables when summarizing the literature. The present study avoided many of these problems by analyzing far more studies than have been included in previous reviews (thus improving statistical power) and by applying quantitative meta-analytic techniques (Cooper & Hedges, 1994; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Smith, Glass, & Miller, 1980) as opposed to a subjective literature review.

As we apply it, meta-analysis yields two encoded sets of measurements for each comparison of clinical and mechanical prediction. First, study-design characteristics that might influence the study's outcome are encoded. Second, there is an objectively defined effect size (ES). Each ES quantifies the degree to which mechanical prediction out- or underperforms clinical prediction. Thus, we are able to provide a quantitative analysis of the relative superiority of clinical and statistical predictions, and to determine the extent to which these differences are influenced by study-design characteristics.

Method

Finding Studies

Our search for studies began by reviewing reference lists of previous reviews and an additional list of studies provided by Paul E. Meehl. We

William M. Grove, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson, Department of Psychology, University of Minnesota, Twin Cities Campus; David H. Zald, Department of Psychiatry, University of Minnesota, Twin Cities Campus.

We thank Roger Blashfield for providing a citation search of references to Meehl's (1954) book; Robyn Dawes and David Faust for many useful ideas; and Kelly Hughes and Jennifer Monson for assistance with library research. We extend particular thanks to Paul E. Meehl and Leslie J. Yonce, without whose encouragement and help this study would not have been done. We also thank Howard Garb for many helpful comments on a previous draft of this manuscript.

Correspondence concerning this article should be addressed to William M. Grove, Department of Psychology, University of Minnesota, Twin Cities Campus, N218 Elliott Hall, 75 East River Road, Minneapolis, Minnesota 55455-0344. Electronic mail may be sent to william.m.grove-1@tc.umn.edu.

then ran keyword searches using PsycLit and MedLine. These searches covered the years 1966–1988 for the key words *actuarial, algorithm, artificial intelligence, Bayesian, clinical, clinician, computer-aided diagnosis, computer-assisted diagnosis, inference, judgment, physician, practitioner, prediction, psychologist, psychotherapist, psychiatrist, psychodiagnosis,* and *statistical.* We also ran computer searches for citations of articles by Holt (1958, 1970), Sarbin (1942), Meehl (1954), Sawyer (1966), Sines (1971), and any publication by De Dombal. These are prominent summaries of the literature (i.e., authors Meehl, Sarbin, Sawyer, and Sines) or articles by leading investigators (i.e., authors De Dombal and Holt).

These searches were supplemented by a hand search for articles published during 1945–1994 in many journals that had published relevant articles (list available on request). We also contacted prominent researchers in the field, asking for articles. Finally, we checked references in studies found by any of these means. Our search procedure is comprehensive, and it appears unlikely that we missed many studies.

Criteria for Including Studies

To be included, clinicians and mechanical procedures had to predict human behavior, make psychological or medical diagnoses or prognoses, or assess states and traits (including abnormal behavior and normal personality). Only studies within the realm of psychology and medicine were included. Thus, we excluded attempts to predict nonhuman outcomes (e.g., horse races, weather, stock market prices).

Studies also had to meet several additional inclusion criteria. First, studies had to compare the performance of at least one human judge to at least one mechanical-prediction scheme. The latter included explicit mathematical formulas, actuarial tables, and other algorithmic procedures (e.g., a computer model of a human judge). Second, the clinician and the mechanical procedure had to have access to the same (or almost the same) predictor variables and predict a common criterion. Operationally, we excluded studies where the predictor variables used by one method were not a subset of those used by the other.

Studies were excluded if they operated on different sets of participants, unless the assignment of participants to prediction conditions was random. This was to rule out the possibility that an ostensibly superior prediction method was favored by being applied to participants for whom predictions were easier to make. Using these selection criteria, 163 studies were identified. Of the 163 studies, 136 qualified for inclusion once closely examined for coding.

Coding Studies

A 25-page manual (available on request) was used to code each study for publication variables and study-design characteristics. Publication variables included year of publication and publication source (e.g., journal article, dissertation, etc.). In regard to design characteristics, the nature of the criterion variable was coded and subcategorized where appropriate. Predictors were coded into major categories (e.g., personality test data), then subcategorized (e.g., objective or projective). Clinician characteristics included professional identification (e.g., physician or psychologist), amount of education, extent of general post-education experience, and amount of experience with the specific prediction task. The form of prediction task was coded (e.g., were judges asked to assign patients to categories, or were they asked to make rank ratings?). We coded the type of mechanical scheme (e.g., regression analysis, actuarial table), and cross-validation used (e.g., none, split sample, bootstrapping, or jackknifing; Efron, 1982). We coded whether clinicians, the mechanical method, or neither had more information on which to base predictions.

The 25-page manual was also used to code each study for the accuracy of clinical and mechanical predictions. We coded variables directly related to ES computation and weighting: number of subjects, number of judges, and number of judgments per judge per subject; and the number of subjects given mechanical predictions. Finally, we coded the accuracy of clinical and mechanical predictions and identified the type of accuracy statistic (e.g., hit rate, correlation).

In many studies, more than one clinical or mechanical prediction was reported. For example, predictions might be made for more than one criterion (e.g., college GPA and college leadership ratings). In such a case, as a first step the different predictands were separately coded. Also, in many studies, several clinicians or several mechanical-prediction algorithms made judgments. For example, accuracy for experienced versus novice judges might be reported, or the efficiency of unweighted sums of predictor variables might be compared with a linear regression equation. All possible paired clinical–mechanical prediction comparisons were calculated.

One study (Goldberg, 1965) was voluminous, yielding 124 ESs. To ensure that Goldberg's results did not overwhelm those from other studies, and to reduce the multiplicity of results from other studies, we conducted a study effect-size analysis (explained below).

Reliability

After initial training, a series of 14 studies (31 ESs) were coded by a minimum of two coders each to check reliability. These studies had not been used in training. Reliability of ES coding was satisfactory (intraclass $R = .97$). The rest of the studies were coded by only one rater. Coding problems were resolved by consensus.

Study Effect Sizes

In our primary data analyses, we reduced the data to one representative ES per study. This approach prevents potential biases arising from the inclusion of studies with multiple, nonindependent ESs. To select the most representative ES, we proceeded as follows: We took as representative of the clinicians' performance for the study the performance of that judge (or group of judges) that would be expected to have the highest accuracy (or no lower accuracy than any other judge). Thus, we preferred clinicians with the most experience with the particular prediction task at hand; failing this, we preferred clinicians with the most general clinical experience; as a fallback, we took clinicians with the most training. If the study did not describe the experience or training of the judges well enough to make such distinctions, then we took the median of all judges. Among the various mechanical-prediction schemes used in a study, we preferred cross-validated rules to ones that were not; among cross-validated rules, we took the one with the highest accuracy.[1] If there were no cross-validated rules, then we took the median performance of all actuarial prediction rules. In other words, the clinicians we would expect a priori to perform best (or as well as any others) on a new sample were compared with the mechanical-prediction scheme we would expect to perform best (or as well as any other) on a new sample. We acknowledged that the procedures utilized to select the most representative ES might bias the results. We, therefore, also analyzed the data in several ways: utilizing the median ES for each study, examining every ES from every study, and also looking at the data using different criteria for selecting the most representative ES.

Transformation of ESs

Studies assessed prediction accuracy in incommensurable ways, therefore we converted accuracy measures to a common metric. First, we found a suitable transformation of the reported accuracy statistic (e.g., hit rate, correlation, etc.) with a known asymptotic variance and an approximately normal distribution (Rao, 1973). We then computed the transformed ES as

[1] We treated $N \geq 1000$ studies as if they were cross-validated, because there is little validity shrinkage with such large samples.

the difference between mechanical and clinical predictions' transformed accuracy statistics. Positive ESs indicate superiority for mechanical prediction.

Weighted least squares were used to compute descriptive and inferential statistics (Draper & Smith, 1981). The weight for a transformed ES was chosen to be inversely proportional to the square root of the asymptotic variance for that transformed ES. For example, suppose a study reported hit rates as its accuracy statistic. Because a suitable variance-stabilizing transform for a hit rate (proportion) is the arcsin transformation (Rao, 1973), the transformed ES would be ES = $(\sin^{-1}\sqrt{P_m} - \sin^{-1}\sqrt{P_c})$, where P_m and P_c are proportions of accurate predictions for mechanical and clinical prediction, respectively. The asymptotic variance of a single transformed proportion is $1/(4N)$, therefore our ES, being the difference between two independent, transformed proportions, has asymptotic variance, $[1/(4N) + 1/(4N)] = 1/(2N)$. The weight, given such an ES when aggregating ESs to produce descriptive or inferential statistics, is then chosen to be proportional to $\sqrt{2N}$. In parallel fashion, correlations were transformed by Fisher's z transform; differences between two independent rs have asymptotic variance, $2/N$ (Rao, 1973). Other accuracy statistics reported in some studies (e.g., area under a receiver operating characteristic [ROC] curve) could all be transformed into either a hit rate or a correlation, which was then transformed and weighted as described.

Nonindependence of Statistics

Accuracy statistics for clinical and mechanical prediction from a single study are usually nonindependent, because they are generally evaluated on the same participants. This creates a statistical problem, compounded by the fact that a study ES, which is the median of several highly correlated individual ESs, has a smaller asymptotic variance than the median of several uncorrelated ESs. Lacking necessary information to compensate for nonindependence of statistics, we had to ignore it.[2] This results in overestimation of the sampling error of the ES, which is most notable with studies having close agreement between clinical and mechanical predictions. It does not introduce a predictable, systematic bias to the comparisons.

Data Analysis

We report the entire ES distribution and provide histograms and summary statistics for the distribution. Also, we report the number of studies that exceed 0.1 in the transformed ES scale. This is a way to describe how many ESs show pragmatically useful superiority for clinical or mechanical prediction. A 0.1 ES difference corresponds with about a 9%–10% difference in hit rates, for predictions of intermediate (50%–75%) accuracy.

Results

Table 1 provides a brief description of the studies that met inclusion criteria. The Predictand column describes the type of judgment. Most of the accuracy statistics are reported as hit rates (HR) in percent or as correlation coefficients (corr), most typically Pearson product moment correlations. The accuracy statistics reported in Table 1 represent the raw (untransformed) accuracy statistics.

Figure 1 gives a stem-and-leaf diagram for transformed study ESs. Each line contains an N, followed by a stem to the left of the colon and zero or more leaves to the right of it; together, the stem and a single leaf (with the decimal point adjusted as indicated) give one observation's value. For example, in the first line, 1 −3 : 0, N = 1 observations' values lay between −0.30 inclusive and −0.35 exclusive; the lone observation's value was about −0.30. In the next line, 2 −2 : 65, says N = 2 observations fell between

```
 1  -3 : 0
 2  -2 : 65
 0  -2 :
 3  -1 : 875
 4  -1 : 3200
 7  -0 : 9988875
18  -0 : 443332222222111111
19   0 : zzz1111122223333334
18   0 : 555667777788889999
17   1 : 00011122222233344
10   1 : 5555778899
13   2 : 0000111222444
 6   2 : 567788
 5   3 : 00224
 1   3 : 5
 1   4 : 3
 3   4 : 699
 3   5 : 224
 1   5 : 8
 2   6 : 34
 0   6 :
 1   7 : 3
```

Figure 1. Stem-and-leaf diagram for differences (mechanical–clinical) of transformed effect sizes. Decimal point is one place to the left of the colon. N = 136; *Mdn* = 0.0876968; Quartiles = −0.00802525 and 0.203482.

−0.25 inclusive and −0.30 exclusive. One was about −0.25, the other about −0.26. The figure's shape forms a histogram, whereas the numerals give the actual data to two decimal places.

Transformed ESs ranged from −0.30 (*clinical prediction superior*) to 0.74 (*marked superiority for mechanical prediction*). Although wide variability exists between studies, the typical study modestly favors mechanical prediction. Half of the transformed ESs lie between about 0 and 0.2. A simple rubric for summarizing the data treats ESs < −0.1 as substantially favoring the clinician, those falling from −0.1 to 0.1 as being relatively equal, and those > 0.1 as substantially favoring the mechanical method. Using this categorization scheme, we found about half of the studies (N = 63; 47%) notably favor mechanical prediction, with as many (N = 65) yielding equal performance. In contrast, only eight studies (6%) notably favor clinical prediction.

We tested whether the ESs are homogeneous by Q_T statistic (Hedges & Olkin, 1985, p. 153). They are not, Q_T = 1635.2, distributed as $\chi^2(135, N = 136)$, $p < .0001$. Weighted summary statistics for the ESs are M = 0.086, SD = 0.12, Quartile 1 = −0.0080, *Mdn* = 0.080, Quartile 3 = 0.20.

Relationship of ES to Study Design

We examined factors that might influence study outcomes: year published, sample size, type of criterion, predictors used, background of clinical judges (e.g., medical versus psychological), judges' level of experience, relative amount of data available to the clinicians versus mechanical formulas, and whether the mechanical algorithm was cross-validated. We used weighted least squares

[2] Hence, most significance tests comparing clinical- and mechanical-prediction accuracies, as given in the original reports, are likewise in error. Almost no studies use appropriate matched-samples statistics.

Table 1
Studies Included in Meta-Analysis

Citation	Predictand	Accuracy statistic	Accuracy	
			Clinical	Mechanical
Alexakos (1966)	college academic performance	HR	39	56
Armitage & Pearl (1957)	psychiatric diagnosis	HR	30	31
Ashton (1984)	*magazine advertising sales*	*corr*	*0.63*	*0.88*
Barron (1953)	psychotherapy outcome	HR	62	73
Blattberg & Hoch (1988)	catalog sales; coupon redemption	corr	0.52	0.66
Blenkner (1954)	case work outcome	corr	0.00	0.62
Bobbitt & Newman (1944)	success in military training	regression coefficient	0.93	0.87
Bolton et al. (1968)	vocational rehabilitation outcome	corr	0.30	0.40
Boom (1986)	diagnosis of jaundice	HR	85	90
Boom et al. (1988)	diagnosis of jaundice	HR	88	96
Boyle et al. (1966)	diagnosis of thyroid disorder	HR	77	85
Brodman et al. (1959)	general medical diagnosis	HR	43	48
Brown et al. (1989)	diagnosis of lateralized cerebral dysfunction	corr	0.43	0.64
Buss et al. (1955)	prediction of anxiety	corr	0.60	0.64
Caceres & Hochberg (1970)	diagnosis of heart disease	HR	74	84
Campbell et al. (1962)	*job performance*	*corr*	*0.15*	*0.29*
Cannon & Gardner (1980)	general medical diagnoses, optimality of treatment recommendations	HR	63	64
Cebul & Poses (1986)	presence of throat infection	HR	69	99
Clarke (1985)	surgery recommendation	HR	59	69
Cooke (1967)	psychological disturbance	HR	77	76
Cornelius & Lyness (1980)	job analysis	corr	0.73	0.76
Danet (1965)	future psychiatric illness	HR	65	70
Dannenberg et al. (1979)	prognosis of medical illness	accuracy coefficient	0.22	0.21
Dawes (1971)	success in graduate school	corr	0.10	0.51
De Dombal et al. (1974)	diagnosis of gastrointestinal disorders	HR	71	92
De Dombal et al. (1975)	diagnosis of gastrointestinal disorders	HR	83	85
De Dombal, Horrocks, et al. (1972)	diagnosis of gastrointestinal disorders	HR	50	97
De Dombal, Leaper, et al. (1972)	diagnosis of appendicitis	HR	83	92
Devries & Shneidman (1967)	course of psychiatric symptoms	HR	75	100
Dicken & Black (1965)	supervisory potential	corr	0.09	0.30
Dickerson (1958)	client compliance with counseling plan	HR	57	52
Dickson et al. (1985)	diagnosis of abdominal pain	HR	55	73
Dunham & Meltzer (1946)	length of psychiatric hospitalization	HR	34	70
Dunnette et al. (1960)	job turnover	HR	53	73
Durbridge (1984)	diagnosis of hepatic or biliary disorder	HR	62	74
Edwards & Berry (1974)	psychiatric diagnosis	HR	63	74
Enenkel & Spiel (1976)	diagnosis of myocardial infarction	HR	78	57
Evenson et al. (1973)	medication prescribed	HR	77	75
Evenson et al. (1975)	length of hospitalization	HR	76	71
Geddes et al. (1978)	degree of pulmonary obstruction	HR	96	95
Glaser & Hangren (1958)	probation success	HR	83	84
Glaser (1955)	criminal recidivism	mean cost rating	0.14	0.35
S. C. Goldberg & Mattsson (1967)	improvement of schizophrenia	significance test	8.15	10.78
L. R. Goldberg (1965)	psychiatric diagnosis	corr	0.28	0.38
L. R. Goldberg (1969)	psychiatric diagnosis	HR	62	69
L. R. Goldberg (1976)	business failure	corr	0.51	0.56
Goldman et al. (1981)	cardiac disease survival or remission	corr	−0.12	−0.11
Goldman et al. (1982)	diagnosis of acute chest pain	HR	79	73
Goldman et al. (1988)	prediction of myocardial infarction	HR	73	76
Goldstein et al. (1973)	cerebral impairment	HR	95	75
Gottesman (1963)	personality description	HR	62	53
Grebstein (1963)	prediction of IQ	corr	0.59	0.56
Gustafson et al. (1973)	diagnosis of thyroid disorder	HR	88	87
Gustafson et al. (1977)	suicide attempt	HR	63	81
Halbower (1955)	personality description	corr	0.42	0.64
Hall (1988)	criminal behavior	HR	54	83
Hall et al. (1971)	diagnosis of rheumatic heart disease	HR	62	73
Harris (1963)	game outcomes and point spread	HR	60	69
Hess & Brown (1977)	academic performance	HR	68	83
Holland et al. (1983)	criminal recidivism	corr	0.32	0.34
Hopkins et al. (1980)	surgical outcomes	HR	84	91
Hovey & Stauffacher (1953)	personality characteristics	HR	74	63

Table 1 (*continued*)

Citation	Predictand	Accuracy statistic	Accuracy	
			Clinical	Mechanical
Ikonen et al. (1983)	diagnosis of abdominal pain	HR	67	59
Janzen & Coe (1973)	"diagnosis" of female homosexuality	HR	57	85
Jeans & Morris (1976)	diagnosis of small bowel disease	HR	83	83
Johnston & McNeal (1967)	length of psychiatric hospitalization	HR	72	75
Joswig et al. (1985)	diagnosis of recurrent chest pain	HR	69	86
Kahn et al. (1988)	detection of malingering	HR	21	25
Kaplan (1962)	psychotherapy outcome	HR	66	70
Kelly & Fiske (1950)	success on psychology internship	corr	0.32	0.41
Khan (1986)	business startup success	corr	−0.09	0.13
Klehr (1949)	psychiatric diagnosis	HR	67	64
Klein et al. (1973)	psychopharmacologic treatment outcome	corr	0.12	0.90
Kleinmuntz (1963)	maladjustment	HR	70	72
Kleinmuntz (1967)	maladjustment	HR	68	75
Klinger & Roth (1965)	diagnosis of schizophrenia	HR	77	43
Kunce & Cope (1971)	job success	HR	67	77
Lee et al. (1986)	death and myocardial infarction	corr	0.58	0.64
Leli & Filskov (1981)	presence, chronicity and lateralization of cerebral impairment	HR	79	79
Leli & Filskov (1984)	diagnosis of intellectual deterioration	HR	75	73
Lemerond (1977)	suicide	HR	50	50
Lewis & MacKinney (1961)	career satisfaction	corr	0.09	0.56
Libby (1976)	business failure	HR	74	72
Lindzey (1965)	"diagnosis" of homosexuality	HR	70	57
Lindzey et al. (1958)	"diagnosis" of homosexuality	HR	95	85
Lyle & Quast (1976)	diagnosis of Huntington disease	HR	61	68
Martin et al. (1960)	diagnosis of jaundice	HR	87	79
Mathew et al. (1988)	diagnosis of low back pain	HR	74	87
McClish & Powell (1989)	intensive care unit mortality	ROC	0.89	0.83
Miller et al. (1982)	general medical diagnosis	HR	53	40
Mitchell (1975)	managerial success	corr	0.19	0.46
Oddie et al. (1974)	diagnosis of thyroid disorder	HR	97	99
Orient et al. (1985)	diagnosis of abdominal pain	HR	64	63
Oskamp (1962)	presence of psychiatric symptoms	HR	70	71
Peck & Parsons (1956)	work productivity	corr	0.71	0.61
Pierson (1958)	college success	HR	43	49
Pipberger et al. (1975)	diagnosis of cardiac disease	HR	72	91
Plag & Weybreun (1968)	fitness for military service	corr	0.19	0.30
Popovics (1983)	cerebral dysfunction	corr	0.17	0.16
Poretsky et al. (1985)	diagnosis of myocardial infarction	HR	80	67
Reale et al. (1968)	diagnosis of congenital heart disease	HR	73	82
Reich et al. (1977)	diagnosis of hematologic disorders	HR	68	71
Reitan et al. (1964)	diagnosis of cerebral lesions	HR	75	73
Rosen & Van Horn (1961)	academic performance	HR	55	57
Royce & Weiss (1975)	marital satisfaction	corr	0.40	0.58
Sacks (1977)	criminal recidivism	HR	72	78
Sarbin (1942)	academic performance	corr	0.35	0.45
Schiedt (1936)	parole success or failure	HR	68	76
Schofield & Garrard (1975)	performance in medical school	HR	76	78
Schofield (1970)	performance in medical school	deviation score	0.07	−0.06
Schreck et al. (1986)	diagnosis of acid-base disorders	HR	55	100
Schwartz et al. (1976)	diagnosis of metabolic illnesses	HR	92	85
Shapiro (1977)	outcome of rheumatic illness	Q	0.20	0.15
Silverman & Silverman (1962)	diagnosis of schizophrenia	HR	55	64
Smith & Lanyon (1968)	juvenile criminal recidivism	HR	52	54
Speigelhalter & Knill-Jones (1984)	diagnosis of dyspepsia	ROC	0.85	0.83
Stephens (1970)	schizophrenia prognosis and course	corr	0.51	0.29
Storment & Finney (1953)	assaultive behavior	corr	0.00	0.57
Sutton (1989)	diagnosis of abdominal pain	HR	65	57
Szucko & Kleinmuntz (1981)	lie detection	corr	0.23	0.42
Taulbee & Sisson (1957)	psychiatric diagnosis	HR	63	63
Thompson (1952)	juvenile delinquency	HR	64	91
Truesdell & Bath (1957)	academic dropouts	HR	71	75
Ullman (1958)	course of group home placement	HR	59	78
Walters et al. (1988)	malingering	HR	56	93

Table 1 (*continued*)

Citation	Predictand	Accuracy statistic	Accuracy	
			Clinical	Mechanical
Warner (1964)	diagnosis of congenital heart disease	HR	66	66
Watley & Vance (1963)	college achievement and leadership	HR	59	72
Webb et al. (1975)	occupational choice	HR	35	55
Wedding (1983)	diagnosis of cerebral impairment	corr	0.74	0.84
Weinberg (1957)	personality characteristics	corr	0.41	0.65
Werner et al. (1984)	assault by psychiatric inpatients	corr	0.14	0.56
Wexler et al. (1975)	medical diagnosis	HR	65	85
Wiggins & Kohen (1971)	graduate school success	corr	0.33	0.58
Wilkinson & Markus (1989)	minor psychiatric morbidity	ROC	0.74	0.89
Wittman & Steinberg (1944)	psychiatric prognosis	HR	41	68
Wormith & Goldstone (1984)	criminal recidivism	corr	0.21	0.39
Yu et al. (1979)	optimality of treatment for meningitis	HR	30	65

Note. For Accuracy Statistic, HR = hit rate (nearest %), corr = correlation coefficient (generally Pearson), ROC = area under Receiver Operating Characteristic curve.

regression to estimate influences of study design on the ES (Hedges & Olkin, 1985).

Publication date did not influence ESs, $t(134) = 0.76$, *ns*. There is no significant relationship between sample size and ES, $t(134) = -0.15$, *ns*, nor between source of publication (journal versus other) and ES, $t(134) = 0.12$, $p < .91$. Table 2 breaks down ESs by type of criterion. There was a trend toward greater advantage for mechanical prediction in medical and forensic settings, $F(5, 130) = 2.11$, $p < .07$. We ran a Ryan–Einot–Gabriel–Welsch multiple F follow-up test, which failed to show any significant differences, between categories. We concluded that the superiority of mechanical prediction holds across many prediction domains.

When results of an interview are used as predictive data, the ES favors the mechanical prediction more than when no interview is available [with interview, weighted $M \pm SD = 0.224 \pm 5.06$; without interview, 0.070 ± 2.29, $t(134) = 5.02$, $p < .0001$]. Use of medical data (physical examination, laboratory tests) as predictors is associated with smaller differences [with medical data, weighted $M \pm SD = 0.083 \pm 3.00$; without medical data, 0.16 ± 3.61, $t(134) = 2.66$, $p < .009$]. On the other hand, use versus nonuse of psychological tests, trait ratings, behavioral observations, and the examinee's track record (e.g., record of previous crimes used to predict recidivism) as predictors does not significantly influence the difference in accuracy between mechanical and clinical prediction.

Table 2
Mean Difference of Transformed Effect Sizes by Type of Criterion

Criterion type	*N*	*M*	*SD*
Educational	18	0.09	0.96
Financial	5	0.20	1.53
Forensic	10	0.89	2.16
Medical	51	0.82	3.05
Clinical–Personality	41	0.19	4.83
Other	11	0.14	1.34

Note. All statistics are computed on weighted observations, with weights as explained in the text. $F(5, 130) = 2.11$, $p < .07$.

Judges' training does not substantially influence the results. Medically trained judges ($N = 27$ studies) did not differ from psychologists ($N = 56$ studies) in how inferior they were to mechanical prediction: The weighted $M \pm SD$ for medical judges was 0.11 ± 4.93, close to the figure for psychologist judges, 0.065 ± 1.53, $t(81) = 0.66$, *ns*. Similarly, training and experience (amount of training, general experience in the field, specific task-relevant experience) do not significantly predict the degree of superiority of mechanical over clinical prediction (all t tests were nonsignificant, $p > .10$).

Even the comparative amount of data available to clinicians versus mechanical prediction did not significantly influence the study ES [$M \pm SD$ when clinician had more data 0.10 ± 3.88; 0.11 ± 2.19 when they had the same data, $t(130) = 0.25$, *ns*]. The mechanical-prediction formula was never based on more data than the clinician had.[3] Furthermore, whether the judges had more data or equal amounts of data relative to the mechanical formula made little difference in the relative superiority of mechanical predictions ($M = .10 \pm 3.88$ when the clinician had more data and $M = .11 \pm 2.19$ when the clinician had the same amount of data).

There is a potential bias in favor of mechanical prediction in studies in which mechanical-prediction formulas were not cross-validated. However, this was not observed. The ES was not significantly larger when the mechanical prediction was cross-validated (0.12 ± 4.56) than when it was not, 0.10 ± 2.28, $t(134) = -0.62$, *ns*.

Effect of Varying Definitions of Study ES

Our definition of the study ES is somewhat elaborate and could have been done otherwise. Therefore, we reanalyzed the data in three ways. These results are given in Table 3. The first row lists the results of our preferred ES definition. In the second row, the study ES was defined as the median of all individual ESs for a given study. The last two rows are variations on an idea: comparing the best-performing clinician, or group of clinicians, with the

[3] Four studies did not report the relative amount of data available to clinicians versus mechanical prediction.

Table 3
Sensitivity of Results to Definition of Effect Size

ES definition	*M*	*SD*	Q1	*Mdn*	Q3	*N* of studies with ES < −0.1	*N* of studies with ES >0.1
Main definition	0.12	0.19	−0.0080	0.088	0.20	8	64
Mdn, all ESs	0.11	0.18	−0.0064	0.095	0.20	8	67
Best clinician	0.09	0.20	−0.023	0.068	0.18	14	59
Best mechanical	0.11	0.20	−0.012	0.078	0.20	10	62

Note. All statistics are computed on weighted observations, with weights as explained in the text. Main definition is the preferred ES definition given in the text. Best clinician means comparing clinical versus mechanical-prediction performance on criterion yielding most accurate clinician predictions. Best mechanical means comparing performance on criterion yielding most accurate mechanical predictions. Q1 and Q3 are first and third quartiles, respectively.

best-performing mechanical-prediction rule. Although this approach fails to consider capitalization on chance and cross-validation, it does have the strength of showing each method to maximum advantage. For this approach, the ES was computed as follows. First, the data were reduced to one ES per dependent variable per study, by choosing that ES for each dependent variable on which the clinician predicted best, or on which the mechanical-prediction scheme was most accurate. Then the comparison was computed in two ways. For the Best clinician row (Row 3), the dependent variable was chosen on which the clinician performed best for that study. For the Best mechanical row (Row 4), the dependent variable consisted of the mechanical prediction that performed best for that study. For either row, the other term in the comparison represents the best available performance by the other prediction method, for the dependent variable in question. Inspection of Table 3 shows that it makes little difference how the effect size is defined, even in the tails of the ES distribution. We conclude that the analysis is relatively insensitive to this aspect of our method.

Conclusions and Discussion

This study confirms and greatly extends previous reports that mechanical prediction is typically as accurate or more accurate than clinical prediction. However, our results qualify overbroad statements in the literature opining that such superiority is completely uniform; it is not. In half of the studies we analyzed, the clinical method is approximately as good as mechanical prediction, and in a few scattered instances, the clinical method was notably more accurate.

Even though outlier studies can be found, we identified no systematic exceptions to the general superiority (or at least material equivalence) of mechanical prediction. It holds in general medicine, in mental health, in personality, and in education and training settings. It holds for medically trained judges and for psychologists. It holds for inexperienced and seasoned judges.

Unfortunately, we were unable to identify many study design variables that robustly predict an advantage for clinical prediction. Only one consistent feature emerged in the eight studies in which clinical judgment outperformed mechanical prediction. In seven of the eight studies, the clinicians received more data than the mechanical prediction. (However, in the entire set of 136 studies, whether the clinician had more data did not significantly influence the relative superiority of mechanical prediction.) The only design variable that substantially influenced the relative efficacy of the mechanical- and clinical-prediction methods was whether the clinicians had access to a clinical interview. Alas, clinical predictions were outperformed by a substantially greater margin when such data was available to the clinician.

Why do we obtain these results? Humans are susceptible to many errors in clinical judgment (Garb, 1998; Kahneman, Slovic, & Tversky, 1982). These include ignoring base rates, assigning nonoptimal weights to cues, failure to take into account regression toward the mean, and failure to properly assess covariation. Heuristics such as representativeness (which leads to belief in the law of small numbers) or availability (leading to over-weighting vivid data) can similarly reduce clinicians' accuracy. Also, clinicians often do not receive adequate feedback on the accuracy of their judgments (Einhorn & Hogarth, 1978), which gives them scant opportunity to change maladaptive judgment habits. In this regard, it is notable that experienced psychologists frequently show little improvement in the accuracy of their clinical judgments relative to the clinical judgments of psychology graduate students (Garb, 1989, 1998).

Answers to Potential Objections

It might be objected that many of the studies are methodologically unsound. Indeed, many were not even designed specifically to test clinical against mechanical data combination. The classic criticism of meta-analyses as "garbage in, garbage out" might be brought to bear. In principle, one could answer such a criticism by showing that better studies do not produce results much different from poorer ones, or even that the better studies allow stronger conclusions in the same direction as all the studies put together. Unfortunately, *better studies* is too ambiguous a phrase to be objectively applied. Therefore, we instead examined specific study design factors that are rationally related to quality (e.g., peer-reviewed journal versus chapter or dissertation, sample size, level of training and experience for judges, cross-validated versus non-cross-validated statistical formulae). Essentially all of these study-design factors failed to significantly influence study effect sizes; no such factor produced a sizable influence on study outcomes.

This being the case, our data give us no reason to suppose that methodologically weaker studies yield different results from better studies. Arguments to the contrary would need to be substantiated by new studies with methods inarguably better than those we have reviewed. The minimal effect of most of the study design factors provides no support for Holt's (1970) contention that the superiority of the mechanical-prediction method is reflective of study-design factors that may bias the results in favor of mechanical predictions. For instance, mechanical predictions outperformed both expert and novice judges. The effect is also clearly not limited the prediction of purely quantitative data. In fact, inclusion of nonquantitative data such as clinical interviews accentuates the superiority of mechanical predictions. Similar conclusions arise if the criterion for considering one method superior to the other are made more lenient (i.e., smaller ES) or more stringent (i.e., larger ES). The trend in our data is so strong that we conjecture the following: There is no selection of studies, based on anything except study outcome itself, that will yield a conclusion directly contrary to ours.

Does our more novel finding, that clinical prediction is often as good as mechanical prediction, matter for any practical purpose? Historically, many advocates of mechanical prediction have assumed that the (near-) equality of mechanical and clinical predictions' performances could be counted as a *win* for mechanical prediction. The rationale for this position has not always been made explicit, but one reason put forth is cost: Mechanical prediction is supposedly much cheaper. No expensive team meetings of high paid professionals are held; a clerk, or a computer program, can make the prediction for a pittance. Hence, our near-equality results should arguably be counted as a real superiority for mechanical predictions.

One potential problem with this position lies in its failure to distinguish various costs. Costs in decision theory include the cost of acquiring information needed to make a decision (whether via clinical or mechanical data combination), the cost of making the decision (here, combining the data), and the costs of the various errors that might be made. Either the data-gathering cost, or the cost of errors, could wash out the advantage of cheaper data combination.

However, data-gathering costs also tend to favor mechanical prediction. In most studies we found, the mechanical-prediction rule operated on fewer data. For example, using a two-variable regression equation (grade point average and college board exam score), Sarbin (1942) could out-predict guidance counselors, who had to obtain and look at personal statements, interest inventories, and letters of recommendation. Moreover, it has not been demonstrated that the costs associated with mechanical-prediction errors are greater than those for clinical prediction. A full analysis of this problem would require thorough examination of relative costs (e.g., false positive determinations of likely criminal recidivism versus false negatives). Studies like this have apparently not been conducted. Future research might profitably address such issues. Perhaps the accuracy of mechanical prediction, appropriately weighted by the costs of its errors, really is less than that of clinical prediction.

Although we cannot rule this possibility out, it is fair to say that the ball is in the clinicians' court. Given the overall deficit in clinicians' accuracy relative to mechanical prediction, the burden falls on advocates of clinical prediction to show that clinicians' predictions are more beneficial to clients in terms of cost-weighted errors, overall costs of decision making, or both.

On the other hand, the assumption that mechanical prediction is always cheaper is simply false. A few years ago WMG watched a highly seasoned MMPI interpreter clinically interpret a number of MMPIs. He wrote his interpretations out much faster than WMG could look up the actuarial data in Marks and Seeman's (1963) actuarial MMPI interpretation guide. In fact, given his billing rates, he was less expensive as an MMPI interpreter than the Minnesota Report, a mechanical interpretation scheme implemented as a computerized expert system.

Although this example may not be typical, it does demonstrate that one cannot simply assume that mechanical prediction is always cheaper. We would, however, be quite surprised if the general run of costs did not favor mechanical prediction. Because many clinical decisions are made not by one clinician but by teams or committees, their costs are often considerably larger than any reasonable estimate of the cost of mechanical prediction. Moreover, unlike the Minnesota Report, most actuarial or statistical data-combination procedures are not proprietary and do not have an appreciable cost per case, after initial setup (on the other hand, freely distributed statistical formulas do not generate revenue that can be put back into research designed to improve the prediction scheme).

On balance, the basic assumption that mechanical prediction is cheapest, and hence to be preferred when it and clinical prediction perform about equally well, seems sound even though there are counterexamples to be found. There seem, then, to be no barriers to a general preference for mechanical prediction where an appropriate mechanical algorithm is available.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Alexakos, C. E. (1966). Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. *Journal of Educational Psychology, 57,* 207–306.

*Armitage, S. G., & Pearl, D. (1957). Unsuccessful differential diagnosis from the Rorschach. *Journal of Consulting Psychology, 21,* 479–484.

*Ashton, A. H. (1984). A field test of implications of laboratory studies of decision making. *Accounting Review, 59,* 361–375.

*Barron, F. (1953). Some test correlates of response to psychotherapy. *Journal of Consulting Psychology, 17,* 235–241.

*Blattberg, R. C., & Hoch, S. J. (1988). *Database models and managerial intuition: 50% model + 50% manager.* Unpublished manuscript.

*Blenkner, M. (1954). Predictive factors in the initial interview in family casework. *Social Service Review, 28,* 65–73.

*Bobbitt, J. M., & Newman, S. H. (1944). Psychological activities at the United States Coast Guard Academy. *Psychological Bulletin, 41,* 568–579.

*Bolton, B. F., Butler, A. J., & Wright, G. N. (1968). *Clinical versus statistical prediction of client feasibility.* Madison: University of Wisconsin Regional Rehabilitation Research Institute.

*Boom, R. (1986). Looking for "indicants" in the differential diagnosis of jaundice. *Medical Decision Making, 6,* 36–41.

*Boom, R., Chavez-Oest, J., Gonzalez, C., Cantu, M. A., Rivero, F., Reyes, A., Aguilar, E., & Santamaria, J. (1988). Physicians' diagnoses compared with algorithmic differentiation of causes of jaundice. *Medical Decision Making, 8,* 177–181.

*Boyle, J. A., Grieg, W. R., Franklin, D. A., Harden, R. M., Buchanan,

W. W., & McGirr, E. M. (1966). Construction of a model for computer-assisted diagnosis: Application to the problem of non-toxic goitre. *Quarterly Journal of Medicine, 35,* 565–588.

*Brodman, K., Van Woerkom, A. J., Erdmann, A. J., & Goldstein, L. S. (1959). Interpretation of symptoms with a data-processing machine. *Archives of Internal Medicine, 103,* 776–782.

*Brown, G. G., Spicer, K. B., Robertson, W. M., Baird, A. D., & Malik, G. (1989). Neuropsychological signs of lateralized arteriovenous malformations: Comparison with ischemic stroke. *Clinical Neuropsychologist, 3,* 340–352.

*Buss, A. H., Wiener, M., Durkee, A., & Baer, M. (1955). The measurement of anxiety in clinical situations. *Journal of Consulting Psychology, 19,* 125–129.

*Caceres, C. A., & Hochberg, H. H. (1970). Performance of the computer and physician in the analysis of the electrocardiogram. *American Heart Journal, 79,* 439–443.

*Campbell, J. T., Prien, E. P., & Brailey, L. G. (1962). Predicting performance evaluations. *Personnel Psychology, 15,* 63–74.

*Cannon, S. R., & Gardner, R. M. (1980). Experience with a computerized interactive protocol system using HELP. *Computers in Biomedical Research, 13,* 399–409.

*Cebul, R. D., & Poses, R. M. (1986). The comparative cost-effectiveness of statistical decision rules and experienced physicians in pharyngitis management. *Journal of the American Medical Association, 256,* 3353–3357.

*Clarke, J. R. (1985). A comparison of decision analysis and second opinions for surgical decisions. *Archives of Surgery, 120,* 844–847.

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

Cooper, H., & Hedges, V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

*Cornelius, E. T., & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. *Journal of Applied Psychology, 65,* 155–163.

*Danet, B. N. (1965). Prediction of mental illness in college students on the basis of "nonpsychiatric" MMPI profiles. *Journal of Consulting Psychology, 29,* 577–580.

*Dannenberg, A. L., Shapiro, A. R., & Fries, J. F. (1979). Enhancement of clinical predictive ability by computer consultation. *Methods of Information in Medicine, 18,* 10–14.

*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., Faust, D., & Meehl, P. E. (1989, March 31). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

*De Dombal, F. T., Clamp, S. E., Leaper, D. J., Staniland, J. R., & Horrocks, J. C. (1975). Computer-aided diagnosis of lower gastrointestinal tract disorders. *Gastroenterology, 68,* 252–260.

*De Dombal, F. T., Horrocks, J. C., Staniland, J. R., & Guillou, P. J. (1972). Pattern recognition: A comparison of the performance of clinicians and non-clinicians with a note on the performance of a computer-based system. *Methods of Information in Medicine, 11,* 32–37.

*De Dombal, F. T., Leaper, D. J., Staniland, J. R., & McCann, A. P. (1974). Human and computer-aided diagnosis of abdominal pain: Further report with emphasis on performance of clinicians. *British Medical Journal, 1,* 376–380.

*De Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., & Horrocks, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal, 2,* 9–13.

*Devries, A. G., & Shneidman, E. S. (1967). Multiple MMPI profiles of suicidal persons. *Psychological Reports, 21,* 401–405.

*Dicken, C. F., & Black, J. D. (1965). Predictive validity of psychometric evaluations. *Journal of Applied Psychology, 49,* 34–47.

*Dickerson, J. H. (1958). *The biographical inventory compared with clinical prediction of post counseling behavior of VA hospital counselors.* Unpublished doctoral dissertation, University of Minnesota, Twin Cities Campus, Minneapolis.

*Dickson, J. A. S., Edwards, N., & Jones, A. P. (1985, June 15). Computer-assisted diagnosis of acute abdominal pain in childhood. *Lancet, 2,* 1389–1390.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

*Dunham, H. W., & Meltzer, B. N. (1946). Predicting length of hospitalization of mental patients. *American Journal of Sociology, 52,* 123–131.

*Dunnette, M. D., Kirchner, W. K., & Erickson, J. R. (1960). Predicting turnover of female office employees. *Personnel Administration, 23,* 45–50.

*Durbridge, T. C. (1984). Applying descriptive discriminant analysis as a visual aid for physicians interpreting biochemical test results. *Clinical Biochemistry, 17,* 321–326.

*Edwards, D., & Berry, N. H. (1974). Psychiatric decisions: An actuarial study. *Journal of Clinical Psychology, 30,* 153–159.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans* (Society for Industrial and Applied Mathematics–National Science Foundation Monograph No. 38). Washington, DC: Society for Industrial and Applied Mathematics.

Einhorn, J. H., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

*Enenkel, W., & Spiel, R. (1976). Comparison of the performance of a computer compared to the effectiveness of a physician's analysis of infarction ECGs. *Advances in Cardiology, 16,* 240–245.

*Evenson, R. C., Altman, H., Sletten, I. W., & Cho, D. W. (1973). Clinical judgment vs. multivariate formulae in assignment of psychotropic drugs. *Journal of Clinical Psychology, 29,* 332–337.

*Evenson, R. C., Altman, H., Sletten, I. W., & Cho, D. W. (1975). Accuracy of actuarial and clinical predictions for length of stay and unauthorized absence. *Diseases of the Nervous System, 36,* 250–252.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105,* 387–396.

Garb, H. N. (1994). Toward a second generation of statistical prediction rules in psychodiagnosis and personality assessment. *Computers in Human Behavior, 11,* 313–324.

Garb, H. N. (1998). *Studying the clinician.* Washington, DC: American Psychological Association.

*Geddes, D. M., Green, M., & Emerson, P. A. (1978). Comparison of reports on lung function tests made by chest physicians with those made by a simple computer program. *Thorax, 33,* 257–260.

*Glaser, D. (1955). The efficacy of alternative approaches to parole prediction. *American Sociological Review, 20,* 283–287.

*Glaser, D., & Hangren, R. F. (1958). Predicting the adjustment of federal probationers. *National Probation and Parole Association Journal, 4,* 258–267.

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

Goldberg, L. R. (1968). Seer over sign: The first "good" example? *Journal of Experimental Research in Personality, 3,* 168–171.

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

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

*Goldberg, L. R. (1976). Man versus model of man: Just how conflicting is that evidence? *Organizational Behavior and Human Performance, 16,* 13–22.

*Goldberg, S. C., & Mattsson, N. (1967). Symptom changes associated

with improvement in schizophrenia. *Journal of Consulting Psychology, 31,* 175–180.

*Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., Acampora, D., Stasiulewicz, V., 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. *New England Journal of Medicine, 318,* 797–803.

*Goldman, L., Waternaux, C., Garfield, F., Cohn, P. F., Strong, R., Barry, W. H., Cook, E. F., Rosati, R., & Sherman, H. (1981). Impact of a cardiology data bank on physicians' prognostic estimates. *Archives of Internal Medicine, 141,* 1631–1634.

*Goldman, L., Weinberg, M., Weisberg, M., Olshen, R., Cook, E. F., Sargent, R. K., Lamas, G. A., Dennis, C., Wilson, C., Deckelbaum, L., Fineberg, H., & Stiratelli, R. (1982). A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. *New England Journal of Medicine, 307,* 588–596.

*Goldstein, S. G., Deysach, R. E., & Kleinknecht, R. A. (1973). Effect of experience and amount of information on identification of cerebral impairment. *Journal of Consulting and Clinical Psychology, 41,* 30–34.

*Gordon, L. V. (1967). Clinical, psychometric, and work sample approaches in the prediction of success in Peace Corps training. *Journal of Applied Psychology, 51,* 111–119.

*Gottesman, I. I. (1963). Heritability of personality: A demonstration. *Psychological Monographs, 77* (9, Whole No. 572).

*Grebstein, L. C. (1963). Relative accuracy of actuarial prediction, experienced clinicians, and graduate students in a clinical judgment task. *Journal of Consulting Psychology, 27,* 127–132.

*Gustafson, D. H., Greist, J. H., Stauss, F. F., Erdman, H., & Laughren, T. (1977). A probabilistic system for identifying suicide attempters. *Computers and Biomedical Research, 10,* 1–7.

*Gustafson, D. H., Kestly, J. J., Ludke, R. L., & Larson, F. (1973). Probabilistic information processing: Implementation and evaluation of a semi-PIP diagnostic system. *Computers and Biomedical Research, 6,* 355–370.

*Halbower, C. C. (1955). *A comparison of actuarial versus clinical prediction to classes discriminated by the Minnesota Multiphasic Personality Inventory.* Unpublished doctoral dissertation, University of Minnesota, Twin Cities Campus, Minneapolis.

*Hall, D. L., Lodwick, S., Kruger, R. P., Dwyer, S. J., & Townes, J. R. (1971). Direct computer diagnosis of rheumatic heart disease. *Radiology, 101,* 497–509.

*Hall, G. C. N. (1988). Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. *Journal of Consulting and Clinical Psychology, 56,* 773–775.

*Harris, J. G. (1963). Judgmental versus mathematical prediction: An investigation by analogy of the clinical versus statistical controversy. *Behavioral Science, 8,* 324–335.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego: Academic Press.

*Hess, T. G., & Brown, D. R. (1977). Actuarial prediction of performance in a six year A.B.-M.D. program. *Journal of Medical Education, 52,* 68–69.

*Holland, T. R., Holt, N., Levi, M., & Beckett, G. E. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology, 68,* 203–211.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56,* 1–12.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25,* 337–349.

*Hopkins, J. A., Shoemaker, W. C., Greenfield, S., Chang, P. C., McAuliffe, T., & Sproat, R. W. (1980). Treatment of surgical emergencies with and without an algorithm. *Archives of Surgery, 115,* 745–750.

*Hovey, H. B., & Stauffacher, J. C. (1953). Intuitive versus objective prediction from a test. *Journal of Clinical Psychology, 9,* 341–351.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

*Ikonen, J. K., Rokkanen, P. U., Gronroos, P., Kataja, J. M., Nykanen, P., de Dombal, F. T., & Softley, A. (1983). Presentation and diagnosis of acute abdominal pain in Finland: A computer aided study. *Annales de Chirurgiae et Gynaecologiae, 72,* 332–336.

*Janzen, W. B., & Coe, W. C. (1973). Clinical and sign prediction: The Draw-a-Person and female homosexuality. *Journal of Clinical Psychology, 31,* 757–765.

*Jeans, W. D., & Morris, A. F. (1976). The accuracy of radiological and computer diagnoses in small bowel examinations in children. *British Journal of Radiology, 49,* 665–669.

*Johnston, R., & McNeal, B. F. (1967). Statistical versus clinical prediction: Length of neuropsychiatric hospital stay. *Journal of Abnormal Psychology, 72,* 335–340.

*Joswig, B. C., Glover, M. U., Nelson, D. P., Handler, J. B., & Henderson, J. (1985). Analysis of historical variables, risk factors and the resting electrocardiogram as an aid in the clinical diagnosis of recurrent chest pain. *Computers in Biology and Medicine, 15,* 71–80.

*Kahn, M. W., Fox, H., & Rhode, R. (1988). Detecting faking on the Rorschach: Computer versus expert clinical judgment. *Journal of Personality Assessment, 52,* 516–523.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

*Kaplan, R. L. (1962). *A comparison of actuarial and clinical predictions of improvement in psychotherapy.* Unpublished doctoral dissertation, University of California, Los Angeles.

*Kelly, E. L., & Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist, 5,* 395–406.

Kendell, R. E. (1973). Psychiatric diagnoses: A study of how they are made. *British Journal of Psychiatry, 122,* 437–445.

*Khan, A. M. (1986). Entrepreneur characteristics and the prediction of new venture success. *International Journal of Management Science, 14,* 365–372.

*Klehr, R. (1949). Clinical intuition and test scores as a basis for diagnosis. *Journal of Consulting Psychology, 13,* 34–38.

*Klein, D. F., Honigfeld, G., & Feldman, S. (1973). Prediction of drug effect in personality disorders. *Journal of Nervous and Mental Disease, 156,* 183–198.

*Kleinmuntz, B. (1963). MMPI decision rules for the identification of college maladjustment: A digital computer approach. *Psychological Monographs: General and Applied, 77,* 1–22.

*Kleinmuntz, B. (1967). Sign and seer: Another example. *Journal of Abnormal Psychology, 72,* 163–165.

*Klinger, E., & Roth, I. (1965). Diagnosis of schizophrenics by Rorschach patterns. *Journal of Projective Techniques and Personality Assessment, 29,* 323–335.

*Kunce, J. T., & Cope, C. S. (1971). *Studies in vocational rehabilitation* (Monograph). Regional Rehabilitation Research Institute, University of Missouri—Columbia.

*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: Statistical models versus expert clinicians. *American Journal of Medicine, 80,* 553–560.

*Leli, D. A., & Filskov, S. B. (1981). Clinical-actuarial detection and description of brain impairment with the W-B Form I. *Journal of Clinical Psychology, 37,* 623–629.

*Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual

deterioration associated with brain damage. *Journal of Clinical Psychology, 40,* 435–441.

*Lemerond, J. N. (1977). *Suicide prediction for psychiatric patients: A comparison of the MMPI and clinical judgments.* Unpublished doctoral dissertation, Marquette University, Marquette, MI.

*Lewis, E. C., & MacKinney, A. C. (1961). Counselor vs. statistical predictions of job satisfaction in engineering. *Journal of Counseling Psychology, 8,* 224–230.

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

*Lindzey, G. R. (1965). Seer vs. sign. *Journal of Experimental Research in Personality, 1,* 17–26.

*Lindzey, G. R., Tejessy, C., & Zamansky, H. S. (1958). Thematic Apperception Test: An empirical examination of some indices of homosexuality. *Journal of Abnormal and Social Psychology, 57,* 67–75.

*Lyle, O., & Quast, W. (1976). The Bender Gestalt: Use of clinical judgment versus recall scores in prediction of Huntington's disease. *Journal of Consulting and Clinical Psychology, 44,* 229–232.

Marchese, M. C. (1992). Clinical versus actuarial prediction: A review of the literature. *Perceptual and Motor Skills, 75,* 583–594.

Marks, P. A., & Seeman, W. (1963). *Actuarial description of abnormal personality: An atlas for use with the MMPI.* Baltimore: Williams & Wilkins.

*Martin, W. B., Apostolakos, P. C., & Roazen, H. (1960). Clinical versus actuarial prediction in the differential diagnosis of jaundice. *American Journal of the Medical Sciences, 240,* 73–80.

*Mathew, B., Norris, D., Hendry, D., & Waddell, G. (1988). Artificial intelligence in the diagnosis of low-back pain and sciatica. *Spine, 13,* 168–172.

*McClish, D. K., & Powell, S. H. (1989). How well can physicians estimate mortality in a medical intensive care unit? *Medical Decision Making, 9,* 125–132.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

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

Meehl, P. E. (1965). Seer over sign: The first good example. *Journal of Experimental Research in Personality, 1,* 27–32.

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

*Miller, R. A., Pople, H. E., & Myers, J. D. (1982). Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine, 307,* 468–476.

*Mitchell, J. O. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60,* 573–579.

*Oddie, T. H., Hales, I. B., Steil, J. N., Reeve, T. S., Hooper, M., Boyd, C. M., & Fisher, D. A. (1974). Prospective trial of computer program for diagnosis of thyroid disease. *Journal of Clinical Endocrinology and Metabolism, 38,* 876–882.

*Orient, J. M., Kettek, L. J., & Lim, J. (1985). A test of a linear discriminant for identifying low-risk abdominal pain. *Medical Decision Making, 5,* 77–87.

*Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied, 76,* 1–27.

*Peck, R. F., & Parsons, J. W. (1956). Personality factors in work output: Four studies of factory workers. *Personnel Psychology, 9,* 49–79.

*Pierson, L. R. (1958). High school teacher prediction of college success. *Personnel and Guidance Journal, 37,* 142–145.

*Pipberger, H. V., McCaughan, D., Littmann, D., Pipberger, H. A., Cornfield, J., Dunn, R. A., Batchlor, C. D., & Berson, A. S. (1975). Clinical application of a second generation electrocardiographic computer program. *American Journal of Cardiology, 35,* 597–608.

*Plag, J. A., & Weybreun, B. (1968). Naval recruit selection program at the neuropsychiatric research unit. In J. A. Plag & B. Weybreun (Eds.), *Personnel selection in the U.S. Navy* (pp. 15–19). Washington, DC: Government Printing Office.

*Popovics, A. J. (1983). Predictive validities of clinical and actuarial scores for the Gesell Incomplete Man Test. *Perceptual and Motor Skills, 56,* 864–866.

*Poretsky, L., Leibowitz, I. H., & Friedman, S. A. (1985). The diagnosis of myocardial infarction by computer-derived protocol in a municipal hospital. *Journal of Vascular Diseases, 16,* 165–170.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

*Reale, A., Maccacaro, G. A., Rocca, E., D'Intine, S., Gioffre, P. A., Vestri, A., & Motolese, M. (1968). Computer diagnosis of congenital heart disease. *Computers and Biomedical Research, 1,* 533–549.

*Reich, P. R., Geer, D. E., & Bleich, H. L. (1977). A computer program for the diagnosis of hematologic disorders. *American Journal of Hematology, 3,* 127–135.

*Reitan, R. M., Warren, J. M., & Akert, K. (1964). Psychological deficits resulting from cerebral lesions in man. In J. M. Warren & K. Akert (Eds.), *The frontal granular cortex and behavior* (Vol. 14, pp. 295–312). New York: McGraw-Hill.

*Rosen, N. A., & Van Horn, J. W. (1961). Selection of college scholarship students: Statistical vs. clinical methods. *Personnel and Guidance Journal, 40,* 150–154.

*Royce, W. S., & Weiss, R. L. (1975). Behavioral dues in the judgment of marital satisfaction: A linear regression analysis. *Journal of Consulting and Clinical Psychology, 43,* 816–824.

*Sacks, H. R. (1977). Promises, performance, and principles: An empirical study of parole decision making in Connecticut. *Connecticut Law Review, 9,* 349–422.

*Sarbin, T. L. (1942). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48,* 593–602.

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

*Schiedt, R. (1936). *Ein Beitrag zum Problem der Rueckfallsproznose.* Munich, Germany: Muenchner-Zeitungs-Verlag.

*Schofield, W. (1970). A modified actuarial method in the selection of medical students. *Journal of Medical Education, 45,* 740–744.

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

*Schreck, D. M., Zacharias, D., & Grunau, C. F. V. (1986). Diagnosis of complex acid base disorders: Physician performance versus the microcomputer. *Annals of Emergency Medicine, 15,* 164–170.

*Schwartz, S. L., Vertinsky, I., Ziemba, W. T., & Bernstein, M. (1975). Some behavioural aspects of information use in decision making: A study of clinical judgements. In H. Thiriez & S. Zionts (Eds.), *Multiple criteria decision making* (Lecture Notes in Economics and Mathematical Systems No. 130. (pp. 136–146). Berlin: Springer-Verlag.

*Shapiro, A. R. (1977). The evaluation of clinical predictions: A method and initial application. *New England Journal of Medicine, 296,* 1509–1514.

*Silverman, L. H., & Silverman, D. K. (1962). Ego impairment in schizophrenia as reflected in the Object Sorting Test. *Journal of Abnormal and Social Psychology, 64,* 381–385.

Sines, J. O. (1971). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry, 116,* 129–144.

Singer, M., & Wynne, L. C. (1965). Thought disorder and family relations of schizophrenics: IV. Results and implications. *Archives of General Psychiatry, 12,* 201–212.

*Smith, J., & Lanyon, R. I. (1968). Prediction of juvenile probation violators. *Journal of Consulting and Clinical Psychology, 32,* 54–58.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins Press.

*Speigelhalter, D. J., & Knill-Jones, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society Series A, 147,* 35–37.

*Stephens, J. H. (1970). Long-term course and prognosis in schizophrenia. *Seminars in Psychiatry, 2,* 464–485.

*Storment, C. T., & Finney, B. C. (1953). Projection and behavior: A Rorschach study of assaultive mental hospital patients. *Journal of Projective Techniques, 17,* 349–360.

*Sutton, G. (1989, October 14). How accurate is computer-aided diagnosis? *Lancet, 2,* 905–908.

*Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36,* 488–496.

*Taulbee, E. S., & Sisson, B. D. (1957). Configural analysis of MMPI profiles of psychiatric groups. *Journal of Consulting Psychology, 21,* 413–417.

*Thompson, R. E. (1952). A validation of the Glueck Social Prediction Scale for proneness to delinquency. *Journal of Criminal Law: Criminology and Police Science, 43,* 451–470.

*Truesdell, A. B., & Bath, J. A. (1957). Clinical and actuarial predictions of academic survival and attrition. *Journal of Counseling Psychology, 4,* 50–53.

*Ullman, L. P. (1958). Psychological reports related to behavior and benefit of placement in home care. *Journal of Clinical Psychology, 14,* 254–259.

*Walters, G. D., White, T. W., & Greene, R. L. (1988). Use of the MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Consulting and Clinical Psychology, 56,* 111–117.

*Warner, H. R. (1964). Experience with Bayes' Theorem for computer diagnosis of congenital heart disease. In S. Gollub & A. W. Ulin (Eds.), *Annals of the New York Academy of Sciences: Vol. 115. Computers in Medicine and Biology* (pp. 558–567). New York: New York Academy of Sciences.

*Watley, D. J., & Vance, F. L. (1963). *Clinical versus actuarial prediction of college achievement and leadership activity* (Cooperative Research Project No. 2202). Minneapolis: University of Minnesota, Student Counseling Bureau.

*Webb, S. C., Hultgen, D. D., & Craddick, R. A. (1975). Predicting occupational choice by clinical and statistical methods. *Journal of Counseling Psychology, 24,* 98–110.

*Wedding, D. (1983). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology, 5,* 49–55.

*Weinberg, G. H. (1957). *Clinical versus statistical prediction with a method of evaluating a clinical tool.* Unpublished doctoral dissertation, Columbia University, New York, NY.

*Werner, P. D., Rose, T. L., Yesavage, J. A., & Seeman, K. (1984). Psychiatrists' judgment of dangerousness on an acute care unit. *American Journal of Psychiatry, 141,* 263–266.

*Wexler, J. R., Swender, P. T., Tunnessen, W. W., & Oski, F. A. (1975). Impact of a system of computer-assisted diagnosis. *American Journal of Diseases of Children, 129,* 203–205.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychology Review, 1,* 3–18.

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

*Wilkinson, G., & Markus, A. C. (1989). PROQSY: A computerised technique for psychiatric case identification in general practice. *British Journal of Psychiatry, 154,* 378–382.

*Wittman, M. P., & Steinberg, L. (1944). Follow-up of an objective evaluation of prognosis in dementia praecox and manic-depressive psychoses. *Elgin Papers, 5,* 216–227.

*Wormith, J. S., & Goldstone, C. S. (1984). The clinical and statistical prediction of recidivism. *Criminal Justice and Behavior, 11,* 3–34.

*Yu, V. L., Fagan, L. M., Wraith, S. M., Clancey, W. J., Scott, A. C., Hannigan, J., Blum, R. L., Buchanan, B. G., & Cohen, S. N. (1979). Antimicrobial selection by a computer. *Journal of the American Medical Association, 242,* 1279–1282.

Received August 3, 1998
Revision received January 5, 1999
Accepted February 4, 1999 ■

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.