When Doctors Meet Numbers

DONALD M. BERWICK, M.D.
HARVEY V. FINEBERG, M.D., Ph.D.
MILTON C. WEINSTEIN, Ph.D.
Boston, Massachusetts

A Statistical Skills Self-Assessment Questionnaire (SAQ) was developed using hypothetical clinical questions to explore respondents' mastery of vocabulary and rules of inference that seem relevant to the use of quantitative information. The SAQ was administered to 281 subjects, including 36 medical students, 45 interns and residents, 49 physicians engaged in research and 151 physicians in full-time practice.

All groups of subjects showed frequent lack of consensus on the meaning of terms in common use (e.g., "false-positive rate" and "p value") and unfamiliarity with some important principles in quantitative inference (e.g., the Central Limit Theorem and Regression to the Mean). Subjects often seemed willing to draw conclusions unsupported by available data. Performance on the SAQ was inversely correlated with length of time since graduation from medical school, and practicing physicians tended to err more frequently than the other three groups.

The practice of modern medicine is inseparable from the use of numbers. Physicians must consume prodigious quantities of data in their daily work; laboratory tests alone yield over 20,000 individual bits of information per practicing physician per year [1]. As the volume of quantitative information in medicine has increased, so has the vocabulary for the description of that information become richer and more widely used. Furthermore, modern concepts of the design, reporting and interpretation of clinical experiments depend upon statistical principles.

We cannot take for granted the ability of physicians to understand and interpret quantitative information and to use it to the best advantage of the patient. Both theoretic work in cognitive psychology [2] and a few reports of direct assessment of the abilities of physicians to use quantitative information [3–5] suggest that physicians and other decision-makers often fail to make full use of the data available to them. Human beings faced with the task of drawing conclusions from complicated or voluminous information fall victim to hazards that degrade that information. We may reasonably expect that the same hazards bedevil doctors when they meet numbers.

The experiment reported here is an attempt to describe the performance of physicians in coping with quantitative information. We devised a 36-item multiple-choice questionnaire to test physicians' knowledge and skill in drawing inferences from quantitative clinical information. Our objective was to assess the strengths and weaknesses of physicians-to-be in coping with numeric information, and to compare these skills at various levels of professional training and practice.

From the Center for the Analysis of Health Practices, Harvard School of Public Health, and the Department of Pediatrics, Harvard Medical School, Boston, Massachusetts. This work was supported in part by grants from the Robert Wood Johnson Foundation and the Merrill Trust through the Center for the Analysis of Health Practices. Reprint requests should be addressed to Dr. Donald M. Berwick, Center for the Analysis of Health Practices, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115. Manuscript accepted July 6, 1981.

METHODS AND MATERIALS

We developed a Self-Assessment Questionnaire (SAQ) to measure a clinician's ability to understand and use quantitative clinical information. All questions employed a multiple-choice or true/false format. In order to avoid variations in performance attributable to differences in substantive knowledge of clinical medicine, all of the questions in the SAQ dealt with fictitious diseases and hypothetic clinical circumstances that were designed to simulate actual clinical problem-solving circumstances. Biostatisticians and physicians with specialized training in quantitative sciences reviewed candidate questions for appropriateness and clarity, and we gradually refined the SAQ to the point that six physicians highly skilled in the rules of statistical inference were able to complete the test missing no more than a single question each on their first attempt. The SAQ as distributed contained 41 questions. Two questions were excluded from analysis because of concern about ambiguity in their wording, and four questions were combined into a single scorable item dealing with the selection of an appropriate definition of "positive" on a hypothetic clinical test. Thus, the final instrument as reported here consisted of 36 scorable items.

All subjects who completed the SAQ received a pamphlet that discussed in detail correct and incorrect responses to each question. The authors presented an analysis of test performance and a more general interpretation of quantitative clinical information as part of continuing medical education at three hospitals where physicians had previously completed the questionnaire.

We grouped items in the SAQ into five separate categories according to the particular skill being explored. The selection of these categories and of individual test items was based on the authors' beliefs about the particular statistical skills that are pertinent to the practice of clinical medicine and to the interpretation of reports on clinical research. The five subscore categories are as follows.

(1) Definitions. The ability to define and use terms closely associated with the statistical properties of tests and experiments (10 items). Illustrative questions:

(I) *A study of the effectiveness of a new drug indicates that the difference in outcomes between the treatment and placebo groups was "significant with $p < 0.05$." The most accurate interpretation of this result is:*

a. *The probability that the drug is better than the placebo is at least 95 percent.*

b. *The probability of observing this large a difference would be less than 5 percent if the drug were no better than the placebo.*

c. *The drug is better than the placebo—unless the advantage of the drug over the placebo is actually less than 5 percent, in which case the study is inconclusive.*

d. *The placebo is no more than 5 percent more effective than the drug.*

e. *In a given case, the probability that the placebo will outperform the drug is at most 5 percent.*

Correct answer: **b.**

(II) *In terms of the following table, where a, b, c and d represent the number of patients in each cell of the table, drawn from a population whose total number is $a + b + c + d$, what is the "false-positive rate"?*

		Disease	
		Present	*Absent*
Test Result	*Positive*	*a*	*b*
	Negative	*c*	*d*

False-Positive Rate =

a. $\frac{b}{a+b}$

b. $\frac{b}{a+b+c+d}$

c. $\frac{c}{a+c}$

d. $\frac{b+c}{a+d}$

e. $\frac{b}{b+d}$

Correct answer: **e.***

Random guesses on all 10 items in the Definitions category would achieve an average score of 31 percent correct. (The number of choices offered per question varied from two to five.)

(2) Behavior of Statistical Data. Knowledge of basic properties of collections of quantitative data, and of the rules for making simple statistical calculations, e.g., Bayes' Formula (six items). Illustrative questions:

(III) *In a city of 1 million people, there are 1,000 people who have contracted Disease K. A test for Disease K is positive in 95 percent of people with the disease and is negative in 95 percent of people without the disease. The test is given to all of the people in the city. In this city, what is the probability that a person with a positive test has Disease K?*

a. *1 to 3 percent*
b. *10 to 20 percent*
c. *50 to 60 percent*
d. *80 to 94 percent*
e. *95 percent*

Correct answer: **a.**

(IV) *Limits for "normal" for each of 12 independent tests done on an autoanalyzer in a certain laboratory are set to include all but the upper and lower 2.5 percent of a popula-*

* Some authorities define "False-Positive Rate" as in answer "a." See Comments.

tion. John Doe, who is in reality entirely well, has a 12-test screen done on his blood. What is the percentage chance that John Doe will have at least one "abnormal" test value?

a. *1 to 3 percent*
b. *7 to 10 percent*
c. *20 to 30 percent*
d. *40 to 50 percent*

Correct answer: **d.**

Random answers in Category 2 would achieve an average score of 24 percent correct.

(3) Going Beyond the Data. The ability to limit inferences to those actually supported by available information; the ability to avoid a conclusion when it is not proved (10 items). Illustrative question:

(V) *The town of Blueburg has two districts, each with its own elementary school. In District A, the average per capita income is $5,000; in District B, the average per capita income is $12,000. A study is done on the percentage of children in the two schools who are on medication for hyperactivity. The study reveals the following: in District A School, 3 percent of children are on medication; in District B School, 1 percent of children are on medication. Appropriate tests of statistical significance show this difference to be significant with p less than 0.005. From this information, we can conclude that in Blueburg a child who lives in a poor family (income under $5,000 per person) has a higher risk of being on medication for hyperactivity than a child who lives in a rich family (income over $12,000 per person).*

True or **False?**

Correct answer: **False.**

Random guesses in Category 3 would achieve an average score of 41 percent correct.

(4) Stopping Short of the Data. The ability to glean useful information from data; the ability to recognize when a conclusion is justified (five items). Illustrative question:

(VI) *All people with Disease Q have both Factor I and Factor II in their blood. Nobody without Disease Q has Factor I; 20 percent of people without Disease Q have Factor II. If a person has Factor II, then measuring his Factor I will help to determine with more certainty if he has Disease Q.*

True or **False?**

Correct answer: **True.**

Random guesses in Category 4 would achieve an average score of 44 percent.

(5) Expected Value Calculations. The ability to combine utilities (i.e., the values attached to outcomes) with probabilistic information according to the rules of decision theory so as to maximize expected utility (five items). Clinical decisions frequently involve the combination of values with information on probabilities. For example, a clinician may reasonably take strong action with respect to a highly unlikely event if the consequences of that event would be dire and can be avoided. Illustrative questions:

(VII) *An operation with 10 percent mortality for anyone always cures AcuteTarkism, a rapidly progressive disease with 20 percent mortality. No other treatment is available. (In our simplified world, all we care about is achieving a minimum mortality rate.) There is available a test, the "AT Test," which identifies people with Acute Tarkism. The problem with the AT Test is that it is sometimes wrong; that is, it sometimes says that a person has Acute Tarkism when, in reality, he is perfectly well. Would you recommend an operation for a person with a positive AT Test if a positive test indicates the presence of the disease in:*

10 percent of the cases in which the test is positive?
40 percent of the cases in which the test is positive?
60 percent of the cases in which the test is positive?
80 percent of the cases in which the test is positive?

Correct answers: **no, no, yes, yes.**

(VIII) *Bewefi's disease is an uncommon and serious infection which has a substantial mortality rate if left untreated. A new chemotherapy agent called HOFRA-B can reduce the mortality, but it is highly cytotoxic and would cause some deaths in normal people. Bewefi's disease is difficult to diagnose. The only diagnostic test available is serum rubar, but it is far from perfect. The mean level of serum rubar is higher in patients with Bewefi's disease than in patients without the disease, but the distributions of patients with and without the disease also overlap on the rubar scale. At present, the expert consensus is to treat patients who have a serum rubar ≧ 100 mmole/liter with HOFRA-B. If mortality among the untreated goes up, but treatment can lower mortality to the same level as previously, what should happen to the level of rubar at which physicians treat for Bewefi's disease?*

a. *it should go up*
b. *it should go down*
c. *it should remain unchanged*
d. *it is impossible to say*

Correct answer: **b.**

Random guesses in Category 5 would achieve an average score of 26 percent.

A more extensive description of the SAQ and the grouping of items into subscore categories is available from the authors.

Each SAQ examination yielded 42 scores per subject: one overall score (percentage correct), five subscores (percentage correct in each category), and 36 item scores (correct or incorrect on each item).

Statistical analyses of these scores, as described later, were performed at the Harvard University Computing facility using SPSS as the language of analysis.

Subjects. The SAQ was administered to a total of 281 subjects drawn primarily from Boston area medical training and practice settings. The selection of these settings was based on the affiliations of the investigators and the willing-

ness of the relevant department to cooperate in this work. In most cases, the test was taken in a group setting, with a time limit of 1 hour. Some subjects received the test as a mailed questionnaire and could answer without limit of time. The subjects came from the following groups:

Medical students: 36 second-year medical students entering a required quantitative methods course as a component of their introduction to clinical medicine;

House staff: 45 interns and residents, including 30 pediatric house staff and 15 family practice residents;

Practicing physicians: 151 physicians in full-time or extensive part-time practice. Some of these doctors had limited teaching responsibilities, but none was primarily engaged in academic research. This group consisted of doctors from three sites: 16 from Community Hospital "A," 42 from Community Hospital "B," and 93 from a medical school Continuing Education Course in general internal medicine.

Academic physicians: 49 doctors whose primary career interest was in medical teaching, medical research, or both. This group included members from four sites: 12 members of the faculty of a pediatric teaching hospital, eight members of the department of family medicine at a second teaching hospital, 12 members of the department of family medicine at a third teaching hospital, and 17 post-doctoral fellows in a preventive medicine training program. A subgroup of 25 of these academic physicians had published an average of 8.7 articles each in the medical literature in the five years prior to our experiment.

Although the selection of subjects was nonrandom, this four-way grouping of subjects was intended to permit preliminary exploration of the hypothesis that stage of training or career path, or both, may correlate with skills related to the interpretation and use of quantitative information.

Our method of comparison among the four major groups was to perform one-way analysis of variance for each score, and to examine differences between individual pairs of groups only if the analysis of variance showed intergroup variation to be significantly ($p \leq 0.05$) higher than intragroup variation. In addition, a simple regression analysis was performed on scores as a function of the number of years since graduation from medical school.

RESULTS

The performance of each group of subjects on all questions and in each of the five categories of questions is summarized in **Table I.**

The 281 subjects correctly answered an average of 63 percent of the test items. This is significantly ($p < 0.0001$) better than the score (approximately 34 percent correct) to be expected from random answers to this multiple-choice test.

The overall performances of medical students, house officers, and academic physicians were virtually identical (average score, 72 percent correct), but these groups differed significantly from the 151 practicing physicians (average score, 55 percent correct) ($p < 0.001$).

In three categories (Going Beyond the Data, Stopping Short of the Data, and Expected Value Calculations) the scores of medical students, house officers and academic physicians were not significantly different from one another. House officers scored significantly ($p < 0.05$) lower than medical students in the Definitions category and significantly ($p < 0.05$) lower than academic physicians in the Behavior of Data category. Practicing physicians consistently scored lower than the other groups in all categories of questions.

Following are salient results based on responses to individual questions in each of the five categories.

Definitions (Category 1). On individual definition questions, academic physicians, house officers and medical students consistently performed better than practicing physicians. The terms "p value" (illustrative question I), "prevalence" and "incidence" were each correctly defined by 45 percent of practicing physicians and by between 82 and 90 percent of each of the other groups ($p < 0.001$). The differences in scores were smaller on items requiring the definition or use of the terms "sensitivity" and "specificity," but the average score of practicing physicians was still lower than that for the rest of the subjects.

Behavior of Statistical Data (Category 2). Performance on illustrative question III confirmed the previously reported [3] lack of knowledge of physicians about the relation between the prevalence of a condition and the predictive value of a positive test. This question was answered correctly by 32 percent of all subjects. Among practicing physicians, 21 percent answered correctly; among research physicians, 65 percent answered correctly. **Table II** shows the response rates of the four groups divided into answers that are either correct or close to correct (**a** or **b**) and those that are far from correct (**c** or **d** or **e**).

Illustrative question IV deals with the interpretation of abnormal results in a battery of clinical chemistry tests. Fifty-five percent of academic physicians and 18 percent of practicing physicians gave the correct answer, **d.** Again grouping answers as "close to correct" (**c** or **d**) or "far from correct" (**a** or **b**), we observed the performance shown in **Table III.**

Two additional questions within this category required qualitative understanding of statistical principles that are important in interpreting clinical research: the Central Limit Theorem, which posits that the means of large samples tend to be less variable than the means of small samples taken from the same population, and the phenomenon of Regression to the Mean, which holds that repeated measurements on a selected sample will be best predicted not by the sample's measured mean, but by a weighted average of the sample's measured mean and the mean of the popu-

TABLE I Performance of Groups Answering the Self-Assessment Questionnaire (Score = Percent Correct)

Score	All Subjects (N = 281)	Medical Students (N = 36)	House Officers (N = 45)	Practicing Physicians (N = 151)	Academic Physicians (N = 49)	Expected Score from Random Guessing
Overall	63%	73%	70%	55%*	74%	33.7%
Definitions	63	78	70‡	54*	73	31
Behavior of Data	35	44	37‡	26*§	52	24
Going Beyond the Data	71	81	83	62*	83	41
Stopping Short of the Data	85	86	91	82†	90	44
Expected Value Calculations	56	65	64	49*	66	26

* Significantly lower than three other groups, $p \leq 0.005$.
† Significantly lower than two highest groups, $p < 0.01$.
‡ Significantly lower than the highest group, $p < 0.05$.
§ Not significantly different from score expected with random guessing, $p > 0.05$.

TABLE II Distribution of Answers to a Question on Prevalence and Predictive Value (Illustrative Question III)

Answer	Medical Students	House Officers	Practicing Physicians	Academic Physicians
Correct or Close (**a** or **b**)	33	33	24	73
Incorrect (**c, d** or **e**)	67	67	74	26
Totals	100%	100%	98%*	99%*

* Sums other than 100% are due to nonrespondents.

TABLE III Distribution of Answers to a Question on the Meaning of Abnormal Results in a Battery of Tests (Illustrative Question IV)

Answer	Medical Students	House Officers	Practicing Physicians	Academic Physicians
Correct or Close (**c** or **d**)	75	56	39	65
Incorrect (**a** or **b**)	25	40	60	33
Totals	100%	96%*	99%*	98%*

* Sums other than 100% are due to nonrespondents.

lation from which the sample was drawn. Fifty percent of subjects gave the correct answer to the question involving the Central Limit Theorem (which had three options); the practicing physicians scored lower (38 percent correct) than the other three groups (60 to 70 percent correct) ($p < 0.005$). There were no significant intergroup differences in the accuracy of response to the single question involving Regression to the Mean. The correct answer was given 32 percent of the time overall, little better than the yield of random responses to this four-item multiple-choice question.

Going Beyond the Data (Category 3). To illustrative question V, 70 percent of practicing physicians incorrectly answered **True,** compared with 50 percent of medical students, 53 percent of house officers and 57 percent of academic physicians. On other items in this category, which required questioning of the (unsupported) assumption of statistical independence in order to be answered correctly, about half of the practicing physicians and one-third to one-fourth of the others made errors.

Another test item in Category 3 involved the estimation of the degree of correlation between test results and the presence or absence of disease. Subjects were presented with a list of results of two tests in 20 hypothetic patients, each of whom was said either to have or not to have a fictitious disease. The table of "clinical data" was so constructed that one test correlated not

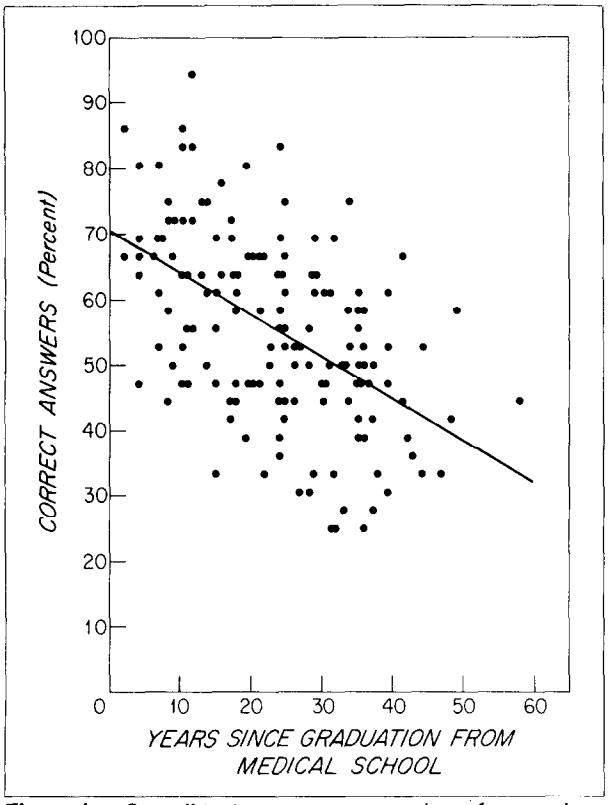

Figure 1. *Overall test score versus number of years since graduation from medical school. Linear regression, performed on results for the 151 practicing physicians, shows a highly significant inverse correlation ($r = -0.53$; $p < 0.0001$).*

at all ($\chi^2 = 0$) with the disease state or with the other test. Nevertheless, 27 percent of subjects perceived that a correlation (either positive or negative) was present. Practicing physicians were more prone to the error of illusory correlation (63 percent correct) than were the other three groups (83 percent, 86 percent and 87 percent correct for medical students, house officers and academic physicians, respectively).

Stopping Short of the Data (Category 4). All groups performed best in this category, and intergroup differences were less striking than in other areas. Because of a higher proportion of "true/false" questions, the expected score from random guessing was also highest in this category.

Expected Value Calculations (Category 5). Several questions in this category required judgment about the decision to treat based on test results (illustrative question VII) and changes in the severity of disease (illustrative question VIII). In question VII, the theoretically correct cutoff level is at a predictive value positive of 50 percent. Forty-one percent of subjects set their cutoff level higher than this; that is, they showed a tendency to avoid surgery even when surgery would have produced the better outcome under the terms of the question. Only 12 percent of subjects showed too great a readiness to use surgery in this hypothetic case. According to the precepts of decision analysis, **b** is the theoretically correct answer to question VIII, an answer given by 60 percent of all subjects.

Age effects: A linear regression model was fitted, with test performance as the dependent variable and the number of years since graduation from medical school as the independent variable, for the practicing physician group. Regression analysis revealed a highly significant inverse correlation ($r = -0.53$; $p < 0.0001$), with poorer scores being achieved by those longer out of training (**Figure 1**). The same effect was observed, with slightly lower correlations, for all of the subscores.

Similar analysis for the academic physicians failed to reveal a significant relation between performance and the number of years out of medical school, for either the overall score or the five subscores.

COMMENTS

Our results suggest that important problems exist in the consumption of quantitative information by medically trained individuals, and that there are significant differences among selected groups of physicians. Although the selection of participants in this study was opportunistic and nonrandomized, our study does include a broad cross-section of physicians and physicians-in-training.

We believe that our results are more useful in comparing groups and exploring very specific skills than in providing meaningful absolute ratings of performance on statistical tasks. Tests such as ours may be made as easy or difficult as the inventor wishes, even on the simplest topic, and representing a subject's level of understanding through a simple score like "percentage correct" is a hazardous business at best.

The measured performance on this test raises concerns in several areas. Questions involving definitions revealed lack of consensus on two terms in common use ("false-positive rate" and "false-negative rate") and surprisingly high levels of error in selecting the correct definitions of such terms as "p value," "sensitivity" and "specificity." More subjects seemed unaware, as well, of several statistical principles that relate to clinical inference, including the connection between the prevalence of a disease and the predictive value of a test for that disease [7], the proper estimation of conjoint and disjoint probability and knowledge of the phenomena related to the Central Limit Theorem and Regression to the Mean. Subjects in all groups tended to draw conclusions that could not be supported by available data. Finally, many subjects did not properly combine probabilistic data with information on utilities.

At least some of the "errors" found in our Definitions

TABLE IV **Distribution of Answers to a Question on the Definition of False-Positive Rate (Illustrative Question II)**

Answer	Medical Students (N = 36)	House Officers (N = 45)	Practicing Physicians (N = 151)	Academic Physicians (N = 49)
a	0%*	60%	26%	49%
b	0	24	41	16
c	0	0	9	6
d	31	0	5	0
e	69	16	15	29
Totals	100%	100%	96%†	100%

* Table entries are percent of subjects answering as indicated.
† Sums other than 100% due to nonrespondents.

subscore may be due to lack of consensus instead of lack of information. For example, experts disagree on the meaning of "false-positive rate" (illustrative question II). We believe most writers who discuss test properties use definition **e** in question II, but some authorities favor definition **a** [8]. **Table IV** summarizes the response to question II by the various groups. Our results raise the interesting possibility that when the 49 percent of academic doctors who prefer answer **a** write a paper reporting the "false-positive rate" of a test, two-thirds of medical students may confuse their meaning with the test property indicated by answer **e**, whereas 40 percent of the practicing physicians may think that the property in question is that indicated by answer **b**.

Some of our results might have been predicted from extant studies of human cognition. Tversky and Kahneman [2], for example, described psychologic tendencies to detect correlations where none exist, to overestimate the informational value of samples of small size, to cling too tenaciously to estimates based upon poor information and to confuse easily noticed events with highly probable events. The consequences of these tendencies include many of the errors that we have noted among our subjects. However, general psychologic characteristics do not explain the striking differences we found between the groups of subjects. Of the 26 individual items on which intergroup variation exceeded intragroup variation, the practicing physicians received the lowest score on 25.

The highly significant inverse correlation of practicing physicians' test performance with years out of medical training has several plausible explanations. The result may represent a pure "learning effect," reflecting the loss of a taught and learned skill over time. Alternatively, it may be evidence of increasing stringency in entrance requirements for medical training, suggesting thereby that our experimental tool correlates with some of those requirements, such as mathematical ability. Finally, the correlation could be attributable to the atrophy of skills that are not very relevant to good clinical practice; experienced clinicians may learn that statistical skills do not affect their efficacy as clinicians and therefore do not expend effort in maintaining skills in that sphere. Elstein et al. [9] have reported that substantive knowledge of clinical medicine is a far more reliable predictor of performance on simulated clinical problems than are any of a large number of traits relating to personality or problem-solving strategy.

We do not claim that performance on this Self-Assessment Questionnaire correlates with clinical acumen or necessarily determines or reflects the quality of medical care. It seems reasonable, however, that a proper understanding of the meaning of terms used in reports of medical experiments is necessary to the accurate consumption of those reports. In view of our results, it may be advisable for authors to define explicitly the statistical terms they use in order to be certain that their meaning is clear to readers.

It also seems reasonable that physicians who must utilize large amounts of quantitative data can better act in the patient's interests if they understand some of the theoretic principles of statistical sampling, and can avoid drawing conclusions that are not warranted by the information at hand. For example, a physician who is unaware of the phenomenon of Regression to the Mean may wrongly credit treratment with observed improvements in a clinical condition such as blood pressure when that apparent effect is a statistical artifact. A physician who consistently underestimates disjoint probabilities may place too much emphasis on unanticipated abnormal results that occur as statistical flukes in a large body of clinical data.

Our subjects frequently gave answers that would not attain the greatest expected value for the patient, in terms of the combination of expressed utility values and probabilities. If this result also holds in clinical practice, then the use of formal decision theory in clinical work may have a role in helping doctors and patients to make certain that their decisions reflect the value structure that they wish to be using.

Further research on the statistical skills of physicians

may move profitably in several directions. Links between statistical skills and clinical performance remain to be forged, although the research challenges in doing so are formidable. For which (if any) of the quantitative analytic skills does the doctor who knows more make better decisions as a result? For which clinical tasks is mastery of statistical concepts most useful?

For those skills shown to be related to the quality of clinical performance, it is important to determine whether, where, to whom and how they can be taught. We are optimistic about finding ways to help clinicians perform their work more efficiently, but not all human tendencies to err in using data can be corrected simply with better education. If errors derive from natural short-cuts in human cognition, then doctors can no more learn to avoid them without mechanical assistance than human travelers can learn to fly without machines. Once we know where our ignorance lies, and how much better we *could* do, we must ask what strategies for avoiding error are best: making human problem-solvers better at their work, or enlisting forms of mechanical aid in situations in which the structural capacities and natural tendencies of the human mind are the limiting factors.

ACKNOWLEDGMENTS

We are indebted to Ms. Penny Kefalas for preparing the test materials, to Ms. Susan Driehaus-Greenberg for supervising data collection, and to Dr. Stephen Schoenbaum, Professor Frederick Mosteller, Dr. Marc Lieberman and the members of the Center for the Analysis of Health Practices at the Harvard School of Public Health for assisting in design, review, and correction of the self-assessment questionnaire, to Ms. Barbara Nash for conducting the computer analyses, and to our subjects, physicians and medical students, for giving generously of their time.

REFERENCES

1. Fineberg HV: Clinical chemistries: the high cost of low-cost diagnostic tests. In: Altman SH, Blendon R, eds. Medical technology: the culprit behind health care costs. U.S. Department of Health, Education, and Welfare Publication Number (PHS) 79-3216. Washington, DC: U.S. Government Printing Office, 1979.
2. Tversky A, Kahneman D: Judgment under uncertainty: heuristics and biases. Science 1974; 185: 1124–1131.
3. Casscells W, Schoenberger A, Grayboys TB: Interpretation by physicians of clinical laboratory results. N Engl J Med 1978; 299: 999–1001.
4. Detmer DE, Fryback DG, Gassner K: Heuristics and biases in medical decision-making. J Med Educ 1978; 53: 682–683.
5. Koran LM: The reliability of clinical methods, data, and judgments. N Engl J Med 1975; 293: 642–646, 695–701.
6. McNeil BJ, Keeler E, Adelstein SJ: Primer on certain elements of medical decision making. N Engl J Med 1975; 293: 211–215.
7. Vecchio TJ: Predictive value of a single diagnostic test in unselected populations. N Engl J Med 1966; 271: 1171.
8. Feinstein AR: Clinical biostatistics. St. Louis: CV Mosby, 1977; 216.
9. Elstein AS, Shulman LS, Sprafka SA: Medical problem solving: an analysis of clinical reasoning. Cambridge, Massachusetts: Harvard University Press, 1978.