1987

The effects of test result and diagnosticity on physicians' revisions of probability of disease in medical diagnosis

Ann Elizabeth Sinclair
Portland State University

AN ABSTRACT OF THE THESIS OF Ann Elizabeth Sinclair for the Master of Science in Psychology presented November 13, 1987.

Title: The Effects of Test Result and Diagnosticity on Physicians' Revisions of Probability of Disease in Medical Diagnosis.

APPROVED BY MEMBERS OF THE THESIS COMMITTEE:

[redacted]

Barry F. Anderson, Ph.D., Chair

[redacted]

James Paulson, Ph.D.

[redacted]

Eric M. Wall, M.D., M.P.H.

This study examined the effects of sensitivity, specificity and result of diagnostic tests on the uses which physicians make of those results. These were compared with the Bayesian model of probabilty adjustment, which is generally accepted for medical diagnosis. Ninety six active members of the Oregon Academy of Family Physicians were interviewed by telephone, using a case scenario describing a patient with a newly discovered breast lump. Subjects estimated prior probability of malignancy, based on history and

physical findings, and then estimated posterior probability following results of a mammogram. Mammograms varied by result (positive or negative) and by high and low values for sensitivity and specificity. Subjects were asked to indicate their confidence in each probability estimate. About one third of the subjects were also asked for their treatment threshold -- that point at which they would change from a policy of watchful waiting to one of taking some action, which was usually biopsy of the lesion.

Subjects who received positive test results made greater probability adjustments ($F(1,88)=12.89$, $p<.001$). No other factors affected the magnitude of change. The accuracy of probability adjustment, compared to the Bayesian model, was also affected by test result ($F(1,88)=21.07$, $p<.001$), with overadjustment by subjects who received positive results and underadjustment by subjects who received negative results. Test specificity also affected accuracy ($F(1,88)=5.26$, $p<.025$), as did interaction between result and specificity ($F(1,88)=4.56$, $p<.05$). Subjects who received positive test results showed greater confidence increase in their revised estimates ($F(1,88)=13.06$, $p<.001$). The results suggest a tendency to be influenced by caution in avoidance of a false negative diagnosis. The departure from normative probability adjustment also indicates lack of understanding of the importance of test accuracy in the disease-absent condition. By comparing subjects' revised probabilities with Bayesian revised probabilities, relative to the treatment threshold, it was concluded that over 20% of the practicing physicians in this study would have more effective treatment plans if they understood better the process of probability revision.

THE EFFECTS OF TEST RESULT AND DIAGNOSTICITY ON PHYSICIANS' REVISIONS OF PROBABILITY OF DISEASE IN MEDICAL DIAGNOSIS

by

ANN ELIZABETH SINCLAIR

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
PSYCHOLOGY

Portland State University
1987

TO THE OFFICE OF GRADUATE STUDIES AND RESEARCH:

The members of the Committee approve the thesis of Ann Elizabeth Sinclair presented November 13, 1987.

[redacted]

Barry F. Anderson, Ph.D., Chair

[redacted]

James Paulson, Ph.D.

[redacted]

Eric M. Wall, M.D., M.P.H.

APPROVED:

[redacted]

Roger Jennings, Chair, Department of Psychology

[redacted]

Bernard Ross, Dean of Graduate Studies and Research

ACKNOWLEDGEMENTS

The author wishes to thank the members of her thesis committee -- Barry Anderson, Eric Wall, and Jim Paulson -- for their guidance and assistance at every step of the design and execution of this project. Thank you to members of the Oregon Health Sciences University Department of Family Medicine: to the physicians who served as pilot subjects, to the clerical staff who assisted with details, and to all those co-workers who provided constant encouragement. Thank you to Andrea Marsden and Chuck McCart, who helped with data collection. Special thanks to Jon Sinclair and Laurie Herrick, without whose support this project might never have been undertaken in the first place.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

INTRODUCTION

Physicians are expected to make decisions with a consistently high level of accuracy, as errors of judgment may lead to serious consequences for their patients and for themselves. Despite expectations of "superhuman" performance, they bring to their professional tasks the cognitive abilities and disabilities possessed by any other human judge. The purpose of this study was to examine some of the human cognitive limitations as part of the process of medical diagnosis. The diagnostic task under scrutiny was revision of the probability of disease, based on the results of a medical test.

In their article describing problems and techniques in the diagnostic process, Schechter and Sheps (1985) proposed the following first two principles of diagnostic decision analysis: (1) "In the diagnostic context, patients do not have disease, only a probability of disease. (2) Diagnostic tests are merely revisions of probabilities." Errors in reasoning--specifically in revising the probability of disease--may seriously reduce the quality of medical care. Serious consequences of overtreatment or undertreatment may arise from errors in determining on which side of the "treatment threshold" (a concept describing the cutoff point at which a physician decides whether or not a patient should be treated) a patient's probability of disease falls (Eddy, 1982). This treatment threshold is analogous to the "critical ratio" discussed by Fischoff and Beyth-

Marom (1983) as that point on the posterior probability continuum which is the "threshold for translating posterior odds into action" (p. 250).

Bayes' Theorem as a Model for Medical Decision Making

The process of revising the probability of disease can be described as hypothesis evaluation. In its most general terms, a hypothesis (H) is evaluated in the light of a new datum (D), which provides evidence for or against the hypothesis. Bayesian inference is an accepted normative model for hypothesis evaluation (Fischoff & Beyth-Marom, 1983) and human inference (Kozielecki, 1970). Cast in terms of (H), and (D), it reads

$$P(H|D) = \frac{P(D|H) \times P(H)}{[P(D|H) \times P(H)] + [P(D|\text{-}H) \times P(\text{-}H)]} \quad (1)$$

where P(H|D) is the probability that the hypothesis is true, given the datum; P(D|H) is the probability of the datum occuring, given the hypothesis is true; P(H) is the prior probability of the hypothesis being true; P(D|-H) is the probability of the datum occuring, given the hypothesis is false; and P(-H) is the prior probability that the hypothesis is false.

Ledley and Lusted (1959) were among the first to introduce the Bayesian formula for probability revision into medical literature as a model for medical diagnosis. By recasting the formula in terms of evaluating the hypothetical presence of disease (D+) and the result of a finding, in this case a positive test result (T+), the formula now

reads

$$P(D+|T+) = \frac{P(T+|D+) \times P(D+)}{[P(T+|D+) \times P(D+)] + [P(T+|D-) \times P(D-)]} \quad (2a)$$

where P(D+|T+) is the probability of disease given a positive test result, P(T+|D+) is the probability of testing positive if disease is present, P(D+) is the prior probability of disease, P(T+|D-) is the probability of testing positive if the disease is absent, and P(D-) is the prior probability of absence of the disease. The analogous formula for finding the posterior probability of disease given a negative test result (T-) is

$$P(D+|T-) = \frac{P(T-|D+) \times P(D+)}{[P(T-|D+) \times P(D+)] + [P(T-|D-) \times P(D-)]} \quad (2b)$$

The posterior probability of disease then becomes the next prior probability in assessing results of new tests, new observations, or other new data (Pascoe, 1986).

Bayesian inference continues in popularity as a normative model for diagnosis, even as methods of analyzing information become more sophisticated. One of the first articles in the medical literature to recommend decision analysis as a strategy for dealing with complex clinical decisions advocated Bayes' formula as a shortcut to be used in the diagnostic process (Schwartz, Gorry, Kassirer, & Essig, 1973). The Bayesian process of diagnostic reasoning has recently been represented in the literature as branches on a diagnostic "tree" (Schechter & Sheps, 1985).

The conditional probabilities of test result given disease state are referred to in medicine as the sensitivity and specificity of the

test. The sensitivity is the proportion of individuals with disease who test positive [P(T+|D+)]; specificity is the proportion of disease-free individuals who test negative [P(T-|D-)]. Sensitivity and specificity are evaluated by comparing test results against the presence or absence of disease as determined by a "gold standard," defined as a "set of criteria, external to the diagnostic test, by which participants are classified as diseased or not diseased" (Sheps & Schechter, 1984, p. 2419). By substituting these terms for their equivalents in formulas 2a and 2b, the formulas for updating the probability of disease given results of a diagnostic test now read

$$P(D+|T+) = \frac{(\text{Sensitivity}) \times P(D+)}{[(\text{Sensitivity}) \times P(D+)] + [(1-\text{Specificity}) \times P(D-)]} \quad (3a)$$

and

$$P(D+|T-) = \frac{(1-\text{Sensitivity}) \times P(D+)}{[(1-\text{Sensitivity}) \times P(D+)] + [(\text{Specificity}) \times P(D-)]}. \quad (3b)$$

To revise the probability of disease on the basis of a test result, it is necessary to know not only the result, but the sensitivity and specificity of the test when applied to that disease. The prior probability of the disease must be considered as well. These factors are mathematically interrelated, with practical implications in the selection and interpretation of medical tests. For example, decreased prior probability increases the effect of specificity on a positive test result (Katz, 1974; Schwartz et al., 1973). Equation 3a shows that the higher P(D-), the greater the impact of the complement of specificity. Furthermore, specificity exerts more influence on the power of a positive test result to turn a diagnosis to

"disease present" (Connell & Koepsell, 1985; Griner, Mayewski, Mushlin, & Greenland, 1981). This is because a change in specificity affects only the denominator, while a change in sensitivity affects both the numerator and the denominator. By the same reasoning, sensitivity exerts more influence on the power of a negative test result to turn a diagnosis to "disease absent".

In summary, specificity has more effect on a positive test than sensitivity, and sensitivity has more effect on a negative test than specificity. A positive test has more effect if the prior probability is low, and a negative test has more effect if the prior probability is high. Therefore, the diagnostician should seek a high-specificity test in a situation of low probability of disease, and a high-sensitivity test in a situation of high probability of disease (Griner et al, 1981; Sox, 1986).

For the Bayesian equation to be used properly with more than one observation, the observations should be conditionally independent (Fischoff & Beyth-Marom, 1983) -- i.e. they should not be correlated within either disease present or disease absent conditions, although they may be correlated overall. The assumption of independence of test results and historical/clinical observations may not be met in medical diagnosis (Doubilet & McNeil, 1985; Hammond, Kelly, Schneider, & Vancini, 1967), and indeed is seldom evaluated in clinical practice (Sox, 1986). For example, it is possible that the sensitivity of a test may vary with clinical indications which led to assessment of the prior probability of disease before the results of the test were received. Furthermore, a test may be less sensitive in the early stages of disease (Sox, 1986). Both are

examples of conditional dependence. Wrongly assuming the independence of indicators and test results may lead to errors in determining the revised estimate of probability of disease (Pascoe, 1986). For example, in cases where clinical indicators increase the sensitivity of the test, the posterior probability will be underestimated if this is not taken into account. There are more complex models of Bayesian inference which apply in situations where information may be dependent (Von Winterfeldt & Edwards, 1986). At this point, however, it is the simplest form of the Bayesian equation which is most likely to appear in the medical literature.

Despite the complexities which arise from questions of conditional independence, there are many who support the use of the Bayesian model in medical diagnosis (Griner et al., 1981), at least as a starting point (Hammond et al., 1967). It is useful as an aid in an increasingly complex medical decision making environment, brought about by advances in medical technology techniques and therapy alternatives (Balla, Iansek, & Elstein, 1985). It may also remind clinicians to use diagnostic tests--"not as infallible technologic tools providing definitive answers for all patients, but as aids with which we may revise probabilities in individual patients" (Schechter & Sheps, 1985, p. 759).

Diagnosticity and the Likelihood Ratio. To revise a hypothesis (H) on the basis of a datum (D), the decision maker must compare the two conditional probabilities [P(D|H) and P(D|-H)] in order to evaluate the evidence provided by the datum. Equation 1 can be rewritten to describe the odds of H as an effect of D (the ratio of how much D supports H to how much D supports -H) as

$$P(H|D)/P(-H|D) = [P(D|H)/P(D|-H)] \times [P(H)/P(-H)]. \quad (4)$$

"Diagnosticity" is a measure of how much the odds favoring H change as a result of D (Beyth-Marom & Fischoff, 1983). From the above equation, the diagnosticity of D is simply defined as P(D|H)/P(D|-H). It is immediately obvious that the decision maker should seek a datum with the highest diagnosticity (Beyth-Marom & Fischoff, 1983).

This diagnosticity value is also called the "likelihood ratio" of the datum. In medical testing terms, the likelihood ratio of a positive test result (LR+) is written

$$LR(T+) = (\text{Sensitivity}) / (1\text{-Specificity}). \quad (5a)$$

and the likelihood ratio of a negative test result to prove disease is

$$LR(T-) = (1\text{-Sensitivity}) / (\text{Specificity}) \quad (5b)$$

(Doubilet, 1983).

Schechter and Sheps (1985) advocate use of the likelihood ratio to determine the impact of a test result on the probability of disease, as neither sensitivity nor specificity is sufficient alone to describe a test's contribution to diagnosis. Considering the likelihood ratio helps one to understand the optimal relationship between sensitivity and specificity. An increase in the sum (Sensitivity + Specificity) leads to an increase in the "gain in certainty", which can be described as a measure of the change in probability (Connell & Koepsell, 1985). Fischoff and Beyth-Marom (1983) admonish decision makers that one should never ask for a piece of information whose likelihood ratio is equal to 1.00. Cast in medical terms, it can be seen that there is no gain of information when (Sensitivity + Specificity) = 1.00 (Connell & Koepsell, 1985).

Clinicians as Bayesians

Although sensitivity and specificity rates are given for many diagnostic tests, most clinicians do not understand how to use these values (Katz, 1974; Sheps & Schechter, 1984). In an informal "hallway encounter" survey of 60 physicians and medical students, Casscells, Schoenberger, and Grayboys (1978) presented a single hypothetical case and asked for revised probability on the basis of a prevalence (prior probability of disease in the population) of 1/1000 and a positive result on a test with a "false positive rate" of 5 percent. Although the sensitivity value was not stated, an upper limit of 100% sensitivity would revise the probability to be 2% or less. Eighteen percent of the subjects gave the correct value as their response. Forty five percent of the subjects returned the specificity value (95%) as the revised probability. The mean of all answers was 55.9%. These results show that the meaning of specificity was not clearly understood, and would suggest a danger of overtreatment due to overestimation of the probability of disease.

Billings and Bernstein (1985) replicated the above study seven years later, to see whether advances in medical training programs had made a difference. Performance seemed to be improved, in that 33% made the "correct" choice ("about 2%"), compared to the 18% found in the earlier study. Two methodological points should be noted, however. This study gave subjects six multiple choice categories ("about 80%", "about 40%", "about 20%", "about 2%", "about 0.01%", and "I don't know") from which to select the correct answer,

as opposed to the previous study which simply asked for the correct answer. Furthermore, the authors did not consider that, since the sensitivity value must be assumed, a conscientious subject could have truthfully answered "I don't know." The "I don't know" answer was, indeed, selected by 40% of the subjects. Despite the study's flaws, the mean response of 30% which subjects estimated indicates once again a tendency toward misunderstanding and suggests a danger of overtreatment due to misdiagnosis.

In a study of nurses' revisions of clinical judgments (Hammond et al., 1967), subjects were told to select one piece of clinical information at a time and to revise the probability of disease with each new datum. These revised probabilities were compared to Bayesian posterior probabilities, using sensitivity and specificity values which had been estimated by the subjects as their revisions were made. Although subjects revised their probabilities in the right direction, they revised them conservatively -- about 1/3 the distance between the prior probability and the Bayesian posterior probability. This was an early study in clinical diagnosis, and no attempt was made at separate analyses of the effects of prior probability, sensitivity or specificity.

Mathematically sophisticated physicians may not perform much better than their colleagues. Borak and Veilleux (1982) compared intuitive reasoning ability of four groups: statistically sophisticated physicians (SP), practicing physicians (PP), clinical nurses (CN), and hospital laborers (HL). When asked to revise the probability of disease on the basis of test result, sensitivity, and specificity, 37% of the SP group and 78% of the PP group erroneously

returned the sensitivity or specificity values. Percentages of the other two groups lay between these values. In a study which compared "quantitatively sophisticated" clinicians' posterior probability estimates with Bayesian probability revisions (Swets, Feehrer, Greenes, & Bynum, 1986), the mean absolute deviation from the normative value was 10%. In the first experiment, subjects set their own likelihood ratios. When they were given the same case scenarios and assigned likelihood ratios, 2/3 of the subjects had mean estimates closer to the Bayesian values (over eight case scenarios), and 1/3 adjusted their estimates farther from the Bayesian value.

Errors in medical probabilistic reasoning are probably not evenly distributed. Griner et al. (1981) showed that false positive clinical errors are more common than false negative errors when reasoning intuitively. Scheff (1971) presented a strong argument that medical diagnosis favors avoiding the possibility of a Type 1 ("rejecting a hypothesis which is true" or false negative) error, even at the expense of increased Type 2 ("accepting a hypothesis which is false" or false positive) errors. In a false negative error, a diseased patient is mistakenly classed as being disease-free; in a false positive error, a well patient is misdiagnosed as being diseased. Decision analysis and common sense may show that favoring avoidance of Type 1 errors is not always optimal. Casscells et al. (1978) remind their medical colleagues that, although the consequences of false negative errors may be obvious, the "mischief that derives from the small percentage of false positive results" may also be considerable (p. 1000). The consequences of unnecessary medical workups and

worry by the patient should not be ignored.

Casscells et al. also express concern that a misunderstanding of laboratory data may contribute to the overuse of laboratory tests. This may become a serious problem in the profession. As a recent article in the medical literature stated, "The continued excessive and often inappropriate use of diagnostic tests should be a matter of concern to the entire medical community. From a wider perspective, these practices represent an important source of waste of limited resources in a time when such waste can no longer be tolerated. More important, from the perspective of individual patient care these behaviors lead to unnecessary morbidity and mortality, arising both from the tests themselves and from the misinterpretation of their results." (Schechter & Sheps, 1985, p. 759).

Cognitive Errors in Probability Revision

Representativeness Bias. Kahneman and Tversky (1973) demonstrated the phenomenon which they named "representativeness bias" in a series of experiments in which subjects overpredicted outcomes which were most representative of (similar to) new evidence. In these tasks, they focused on specific new-case information [cf. test result] and ignored the two pieces of information necessary to make an accurate judgment: the prior, or base-rate, probability that the state exists [cf. prior probability of disease], and the validity of the specific-case information [cf. test diagnosticity]. Although subjects in one of the experiments indicated that they thought the specific-case information had an accuracy rate of only

about 23% "hits", this made no difference in their overuse of the information. A later experiment in the series made the probability values [cf. sensitivity and specificity] of the specific-case information more salient, which resulted only in a minimal reduction in overuse of the information to update probability. The authors applied these findings to "category predictions," one example of which is medical diagnosis. Indeed, Balla, Elstein, and Gates (1983) used representativeness bias to account for some of the errors they found in a hypothetical-case-scenario study of clinical decision making.

The Likelihood Ratio Fallacy. Fischoff and Beyth-Marom (1983) suggest the term "likelihood ratio fallacy" to explain the tendency of judges to ignore the likelihood ratio [P(D|H) / P(D|-H] when revising probability. As the likelihood ratio fallacy has a direct bearing on medical diagnosis, their findings, published in another article (Beyth-Marom & Fischoff, 1983) will now be discussed in terms of medical testing. In the first experiment, most subjects thought P(D|H) [cf. sensitivity] was relevant to testing whether H was true [cf. deciding whether disease is present], while only about half (53.6%) judged P(D|-H) [cf. specificity or its complement] to be relevant. This increased to 78.0% when subjects were asked to decide between H and -H [cf. deciding whether disease is present or absent]. Focusing attention on the possibility of -H [cf. disease absent] and then giving subjects their choice of whether to ask for P(D|-H) increased the percentage of subjects who asked for that information to 70.7%. When asked their reasons, however, most indicated that they thought it was a direct indication of P(-H|D) [cf. confusing the complement of specificity, or the probability that the test is positive

if disease is absent, with the posterior probability that disease is absent when the test is positive]. Only 20% thought that P(D|-H) was necessary to assess P(H|D). When they were told not to seek irrelevant information, only 34.5% of subjects asked for P(D|-H) while most subjects continued to ask for P(D|H). Finally, the experimenters varied P(D|H) and P(D|-H) [cf. sensitivity and (complement of) specificity] orthogonally, although two of the four possible combinations were nondiagnostic (likelihood ratio was 1.00). The prior probability was assigned. Subjects tended to adjust probability in the right direction (91.7%), but the experimenters do not state how far the adjustments were nor how they fell relative to the normative Bayesian value. The authors concluded that P(D|-H) will be included in probability revision if it is presented equally with P(D|H), and that subjects are better at using the relevant information than they are at seeking it out or understanding why it is important.

Covariation Assessment Errors. Judgment whether a disease is present or absent based on a positive or negative test result is an instance of covariation assessment, or judgment of the relationship between two states. Alloy and Tabachnik (1984) decomposed the covariation judgment process into five steps: (1) choose data to be considered, (2) sample, (3) classify new pieces of information as confirming or disconfirming evidence for the coviariation of the two states, (4) recall the data and estimate frequencies of confirming and discomfirming evidence, (5) make judgment on the degree of covariation. They cited numerous examples showing that humans assess covariation poorly.

Why do errors in covariation assessment occur? Anderson,

Gaffuri and Morris (1986) demonstrated a "treatment-column only" strategy in covariation judgment. Cast in Bayesian terms, this is described as attention only to the P(D|H) and P(-D|H) possibilities. This may be seen as a manifestation of the likelihood ratio fallacy, described above. In terms of medical diagnosis, it is analogous to attention only to test sensitivity in determining the diagnostic value of the result. The treatment-column strategy was proposed as an alternative to "confirmation bias" as an explanation of covariation assessment error. Confirmation bias is characterized by attention only to data supporting the hypothesis. In diagnostic terms, this would mean the number (not proportion) of patients with disease who test positive (it could also include the number of patients without disease who test negative). The dominant strategy found by these investigators, used by over half of their subjects, involved equal attention to confirming and disconfirming data and could result in disconfirmation of an hypothesis on the basis of a confirming pattern of data.

Pseudodiagnosticity. Doherty, Mynatt, Tweney and Schiavo (1979) suggested a two-part explanation of errors in covariation judgment. Confirmation bias can explain why subjects select data related to only one hypothesis (H). The tendency to revise the probability of H on the basis only of the relationship between D and H, ignoring the relationship between the D and -H, was identified and named by the authors as "pseudodiagnosticity." This is the tendency to (1) ignore relevant information necessary to revise probability, and (2) revise probability on the basis of irrelevant information.

Pseudodiagnosticity is directly applicable to diagnosis through

medical test results. For this reason, the findings of Doherty et al. (1979) will be compared to analogous terms in medical testing in the following discussion. The authors demonstrated that pseudodiagnosticity was used as a strategy in 82% of their subjects. In the first experiment, subjects were given the opportunity to seek whatever information they thought was necessary to revise the probability of a hypothesis (H) [cf. that a certain disease is present] on the basis of new data (D) [cf. a test result]. Not only did they not choose information regarding P(D|-H) [cf. (complement of) specificity], but they actively chose irrelevant information in the form of the relation between other data and H [cf. performing other tests and asking only for their sensitivity values with the results]. In two experiments in which subjects updated probability on the basis of one datum [cf. revising the probability of disease on the basis of one test result], subjects unquestioningly revised probability using only the relationship P(D|H) [cf. that test's sensitivity]. The authors suggest, as an explanation for pseudodiagnosticity, that a high relationship between the datum and one hypothesis implies a low relationship between the datum and an alternative hypothesis [high P(D|H1) implies low P(D|H2), or high P(D|H) implies low P(D|-H)]. Fischoff and Beyth-Marom (1983) also suggested that increased probability of P(D|H) results in increased reliance on that information alone, as though it implied a decrease in P(D|-H). In medical test terms, this translates to high sensitivity implying high specificity.

In discussing the tendency of judges to attend only to P(D|H) in evaluating the evidence provided by a datum, Fischoff and Beyth-Marom (1983) suggest that a misuse of "efficiency" in the gathering

of information may contribute to failure to ask questions which would disconfirm a hypothesis. In medicine, this would be indicated by ignoring the specificity value of the test. Doherty et al. (1979) enumerated three major consequences of pseudodiagnosticity: (1) it leads to premature closure in solving the problem; (2) it precludes the consideration of other hypothetical solutions; (3) it inhibits the search for other, possibly disconfirming, evidence. The parallels in medicine to each of these consequences might include treatment based on a wrong diagnosis, failure to recognize the true cause for a syndrome, and suboptimal selection of laboratory tests.

Illusory Confidence. In their experiment described above which demonstrated poor ability to revise probability on the basis of specific-case and base-rate information, Kahneman and Tversky (1973) also uncovered an alarming tendency toward high confidence in the erroneous predictions, measured by a subject's estimated probability that the answer was correct. They called this high confidence in fallible judgments, which often persists even when the fallibility is pointed out, the "illusion of validity." Perhaps one of the most interesting observations about this confidence is that it increased with perceived accuracy of the specific-case information, although a change in accuracy of that information did not lead to a change in estimated posterior probability. In medical testing, this would be analogous to increasing confidence on the basis of high sensitivity and high specificity, while not using these values normatively.

Purposes of the Study

1. The first purpose of this study was to examine the effects of the sensitivity and specificity of diagnostic tests on the use which physicians make of the test results. The work on pseudodiagnosticity has been done in a hypothetical context, although the condition describes precisely the task of revising the probability of disease on the basis of a medical test result with less than 100% sensitivity and specificity.

2. The effect of whether the test result was positive or negative was also examined. There were two reasons for this. First, test result is one of the determinants of whether sensitivity or specificity is more important. Second, an effect of test result alone would indicate whether representativeness bias affects the quality of medical diagnosis, and whether a tendency to avoid Type 1 errors interacts with use of diagnostic information.

3. A final purpose of this study was to examine confidence in a setting of medical diagnosis, to determine whether it is affected by the results of diagnostic tests or by the stated validity of those tests.

METHOD

Design

The design of this study was a 2 X 2 X 2 factorial design, using the following variables and levels within each variable:

A = Test results

1. Negative result
2. Positive result

B = Sensitivity of test

1. Low (.81)
2. High (.91)

C = Specificity of test

1. Low (.81)
2. High (.91).

Subjects

Participants in the study were ninety six members of the Oregon Academy of Family Physicians (OAFP), a state-wide organization with an active, practicing membership of about 500 physicians. The median year of graduation from medical school was 1969 (range =1947 to 1983). Subjects had been in practice an

average of 17.8 years (range =1 to 38).

Procedure

Two hundred members of the OAFP were selected by applying the results of a random number generator to the membership list. These physicians were sent a letter from the investigator (see Appendix A). The letter described the purpose of the study and informed the physicians that they might be contacted by telephone and be asked to participate in the study by way of a telephone interview (Dillman, 1978). One week after the letters were mailed, telephone interviewers selected subjects sequentially from the pool using a randomly generated list of numbers 1-200, and began to make telephone calls. Almost all telephone contacts were made initially with a receptionist or nurse. Each encounter began with an introduction, a reference to the letter, and an inquiry whether the physician was willing to be interviewed. It was usually explained that the interview would take about four minutes, and was in the form of a case scenario with an opportunity for the physician to make comments. If the potential subject was willing, it was sometimes more practical to set up a "call back" appointment than to complete the interview at that time.

Once the interviewer was speaking to the subject, the presentation began with the first part of a short case scenario (see Appendix B), describing a female patient with a newly discovered breast lump. The symptoms, history and physical findings in this scenario were chosen on the basis of a literature review (Eddy, 1982;

Griner et al., 1981; Mushlin, 1985) and conversations with family practice physicians who were not participants in the study. The independence of clinical and historical indicators of breast cancer has not been established (Mushlin, 1985), and therefore true conditional probabilities cannot be assigned to each factor. Furthermore, some important factors were deliberately omitted, such as prior history of breast cancer and movability of the lump. For these reasons, a truly "correct" prior probability of breast cancer cannot be determined from the scenario. The scenario was designed, however, as a mixture of factors to encourage the subject to set a mid-range probability of breast cancer, as opposed to benign cystic disease. A mid-range prior probability was desired because a test result has its largest effect on revising probability when the prior probability is mid-range (Sox, 1986), and the figure allowed for adjustment in both directions on the basis of a positive or negative test result. All subjects were given the same first part of the scenario. The interviewer then asked the subject to estimate the probability that the lesion was malignant. Those subjects who did not answer directly with a percentage figure were asked to recast their estimate in percent terms. Once the percent chance of malignancy was estimated, the subject was asked to describe his or her confidence in this estimate "on a scale from 1 to 10, where 1 means 'just a guess' and 10 means 'absolutely certain.'"

Once the subject had estimated a prior probability and expressed a level of confidence in that estimate, the interviewer selected the next available page from a randomized set of the eight versions of the second part of the scenario (see Appendix C for an

example). In this part, the results of a mammogram "ordered" by the subject have been returned from the radiologist as appearing malignant or appearing benign, along with the sensitivity and specificity values of mammograms in that hospital. The information in the second part differed among the eight combinations of conditions described above: negative or positive test result (expressed as "appears malignant" or "appears benign"), sensitivity of .81 or .91 and specificity of .81 or .91 (expressed as 91% or 81%). Sensitivity and specificity values were selected on the basis of congruity with the current range of mammogram validity (Mushlin, 1985), similarity to sensitivity and specificity values which were used in a previous study (Griner et al., 1981), and the maximum possible divergence to keep the ranges equal yet result in different posterior probabilities of disease. In order to reduce as a factor the possible misunderstanding of the terms "sensitivity" and "specificity," these values were first given in sentences describing their meanings without using the terms. The interviewer read this part of the scenario, and again asked the subject to estimate the probability of malignancy and to describe the confidence in that estimate by the same techniques as those in part one.

When about half the interviews had been completed, two questions were added, to be asked of subjects after the scenario was completed. The first question was "If you are forced to choose between a test with high sensitivity and one with high specificity, which do you tend to prefer?" Subjects were permitted to answer this question in whatever way they felt comfortable. The second question was "at what probability of breast cancer would your

treatment plan change from just 'watchful waiting' to actually doing something (such as biopsy, or whatever you would do next)?" At the close of the interview, all subjects were asked what year they had graduated from medical school, how many years they had been in practice, whether sensitivity and specificity values were routinely available to them in practice, and whether they would like to receive a copy of the study.

All of the interview material was scripted, and the scripts served also as data collection sheets. Interviews were conducted in randomized blocks of eight, until 12 interviews in each condition (96 subjects) had been completed.

Immediately after receipt of the letter, one physician's receptionist called to notify the investigator that the doctor did not want to be contacted. Of the remaining subject pool, 150 were selected for telephone calls before the requisite 96 interviews were completed. Of these, 100 interviews were actually completed: three were replacements for those with subjects whose responses were identified as outliers by SYSTAT, and one was with a subject who called back after the quota had been reached but wanted an interview anyway. Thirty five physicians refused to be interviewed. Twelve did not return repeated telephone calls. Telephone numbers were not available for three of the physicians who were selected from the pool as potential subjects.

RESULTS

The range of subjects' estimates of the probability of malignancy after hearing only the first part of the scenario was from 2% to 100% (see Figure 1). The mean of all estimates was 50.3%.

The magnitude of change in probability estimate (ChangeP) was calculated for each subject by using the formula

ChangeP = |(PostP - PriorP)|

where PostP and PriorP are the subject's estimates of posterior and prior probabilities of malignancy. Means and standard deviations of these values are shown in Table I. A three-way analysis of variance was performed on ChangeP, using test result, sensitivity level, and specificity level as independent variables. The ANOVA showed a significant difference due to test result ($F(1,88)=12.89$, $p<.001$). No other effects were significant.

A normative posterior probability value (NormP) was calculated for each subject, using the appropriate form of the Bayesian equation (formulas 3a and 3b above) according to whether the mammogram result was "appears benign" (T-) or "appears malignant" (T+). The values used were the stated sensitivity and specificity of the mammogram reading, and the subject's estimated prior probability of disease after hearing the first part of the scenario (PriorP). In the two cases in which prior probability was

estimated at 100%, the figure used for calculation was 99%. Figure 2 shows a scatter plot of estimated posterior probability as a function of normative posterior probability for each subject.

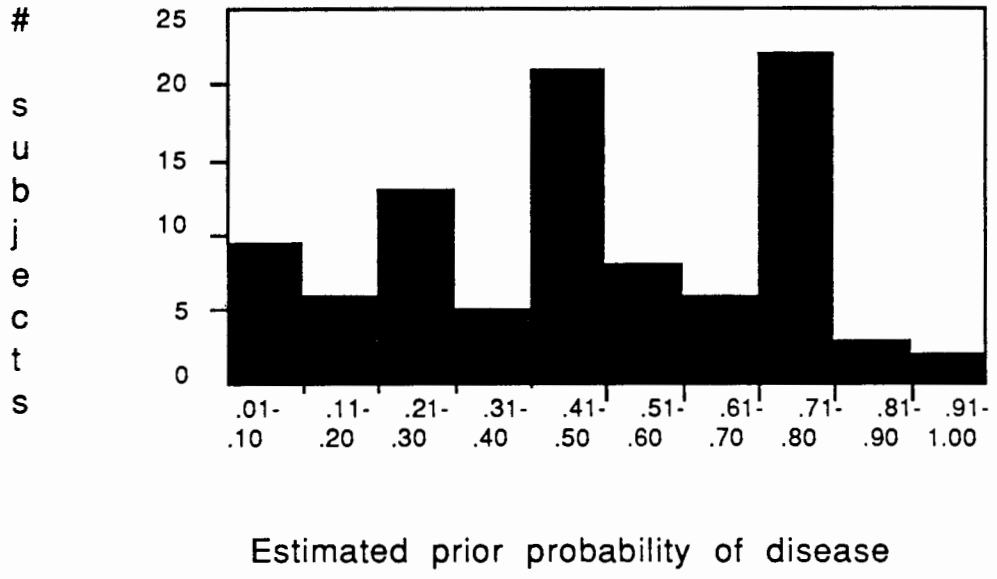


Figure 1. Prior probability of malignancy (PriorP), estimated after part one of the case scenario.

TABLE 1

MEANS (STANDARD DEVIATIONS) OF ABSOLUTE VALUE OF CHANGE FROM PRIOR PROBABILITY OF DISEASE TO POSTERIOR PROBABILITY OF DISEASE (CHANGEP)

Negative Test Result

		Sensitivity: Low	Sensitivity: High
Specificity:	Low	.17 (.12)	25 (.18)
	High	.25 (.22)	29 (.22)

Positive Test Result

		Sensitivity: Low	Sensitivity: High
Specificity:	Low	.42 (.25)	44 (.30)
	High	.33 (.18)	40 (.21)

Note: N = 12 in each cell.

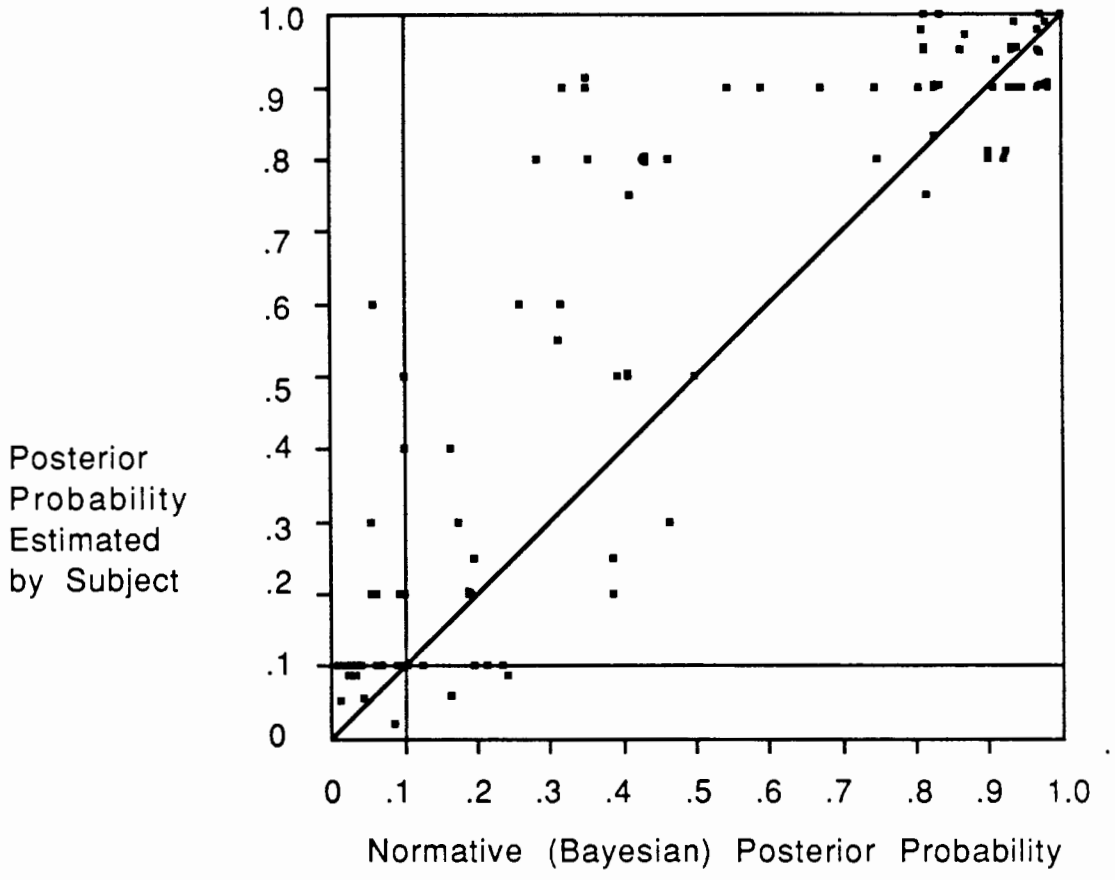
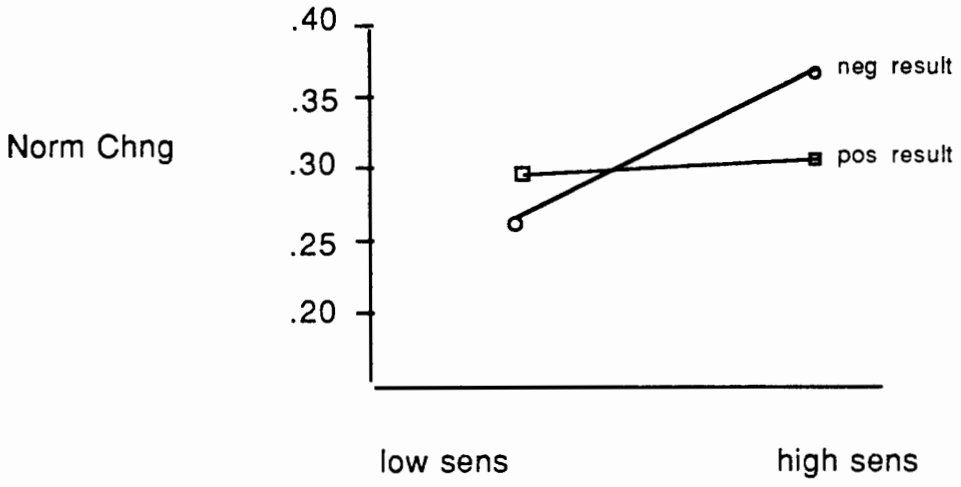


Figure 2. Posterior probability of malignancy as estimated by subjects (PostP), as a function of that calculated by the Bayesian formula (NormP).

To compare with the ChangeP results and to describe the normative change for the subjects in this study, the absolute value of the normative change in probability for each subject (NormChng) was calculated by the formula

NormChng = |(NormP - PriorP)|.

A three-way analysis of variance showed that there would be significant main effects of test sensitivity ($F(1,88)=5.33$, $p<.025$) and test specificity ($F(1,88)=4.22$, $p<.05$). There would be no significant difference attributable to test result as a main effect. As discussed earlier (see page 5), the two-way interactions between result and sensitivity ($F(1,88)=4.72$, $p<.05$) and result and specificity ($F(1,88)=4.58$, $p<.05$) would also be significant (see Figure 3). No other interactions would be significant. These results were not empirical, but indicate how normative behavior would appear in this subject population and with this design. Such an analysis was especially indicated here because of the wide range in PriorP.



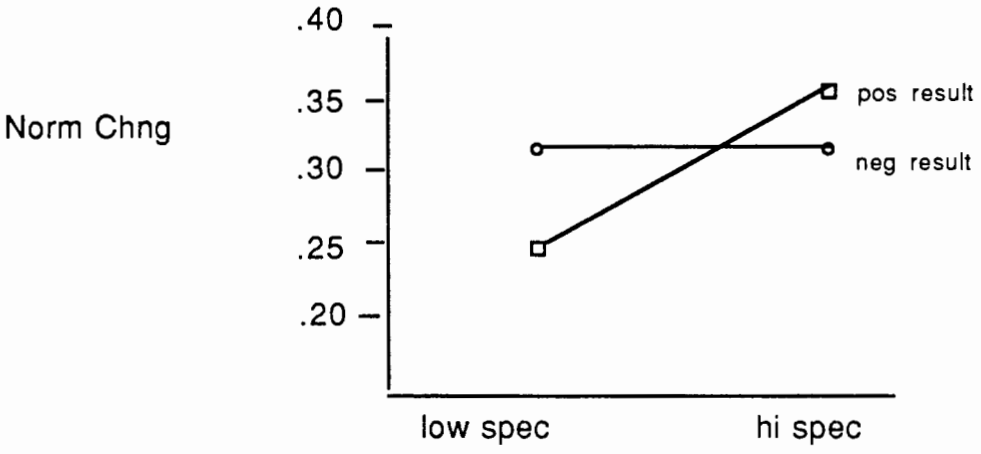


Figure 3. Interactive effects of test result and sensitivity, and test result and specificity on the normative change in probability (NormChng).

For each subject, the degree of accuracy of change from prior probability (PriorP) to posterior probability (PostP) was determined by comparing that change to the normative change in probability through the formula

AccuProb = (PostP - PriorP) / (NormP - PriorP).

This value is simply the proportion of the normative change which each subject actually adjusted, and was computed in order to remove the effects of the variability in PriorP. Algebraic differences were used instead of absolute differences, to capture any instances in which subjects changed in the wrong direction. The four subjects who performed as Bayesians had scores of 1.00. The 42 who adjusted too far had scores of greater than 1.00; the 36 who underadjusted had scores of less than 1.00. Thirteen subjects did not change their probability estimates and had scores of 0. The one subject who revised in the wrong direction had a negative score. Table II shows the means and standard deviations of AccuProb values for the 12 subjects in each experimental condition.

TABLE II

MEANS (STANDARD DEVIATIONS) OF ACCUPROB--THE RATIO OF CHANGE FROM PRIOR PROBABILITY OF DISEASE TO POSTERIOR PROBABILITY OF DISEASE, TO THE CHANGE FROM PRIOR PROBABILITY TO NORMATIVE POSTERIOR PROBABILITY OF DISEASE

Negative Test Result

		Sensitivity: Low	Sensitivity: High
Specificity:	Low	0.639 (0.397)	0.771 (0.735)
	High	0.751 (0.630)	0.617 (0.557)

Positive Test Result

		Sensitivity: Low	Sensitivity: High
Specificity:	Low	1.646 (0.821)	1.568 (1.093)
	High	0.921 (0.227)	1.105 (0.358)

Note: AccuProb = 1.000 means subject performed as a Bayesian

Note: N = 12 in each cell.

The three-way ANOVA on AccuProb showed significant main effects of test result and test specificity (see Table III). The two-way interaction between test result and specificity was also significant. Figure 4 shows a graph of the interactions. Subjects with a negative test result had a mean AccuProb of 0.705 for low specificity and 0.684 for high specificity ($t(45)=0.12$, p N.S.). Subjects who had a positive test result had a mean AccuProb of 1.607 for low specificity and 1.013 for high specificity ($t(27)=2.92, p<.01$). There were no significant effects attributable to sensitivity.

TABLE III

ANALYSIS OF VARIANCE ON ACCUPROB

Factor	SS	df	MS	F ratio	p
Result	9.093	1	9.093	21.07	<.001
Sensitivity	0.017	1	0.017	0.04	N.S.
Specificity	2.269	1	2.269	5.26	<.025
Res X Sens	0.017	1	0.017	0.04	N.S.
Res X Spec	1.970	1	1.970	4.56	<.05
Sens X Spec	0.000	1	0.000	0.00	N.S.
Res X Sens X Spec	0.416	1	0.416	0.96	N.S.
Error	37.978	88	0.432		

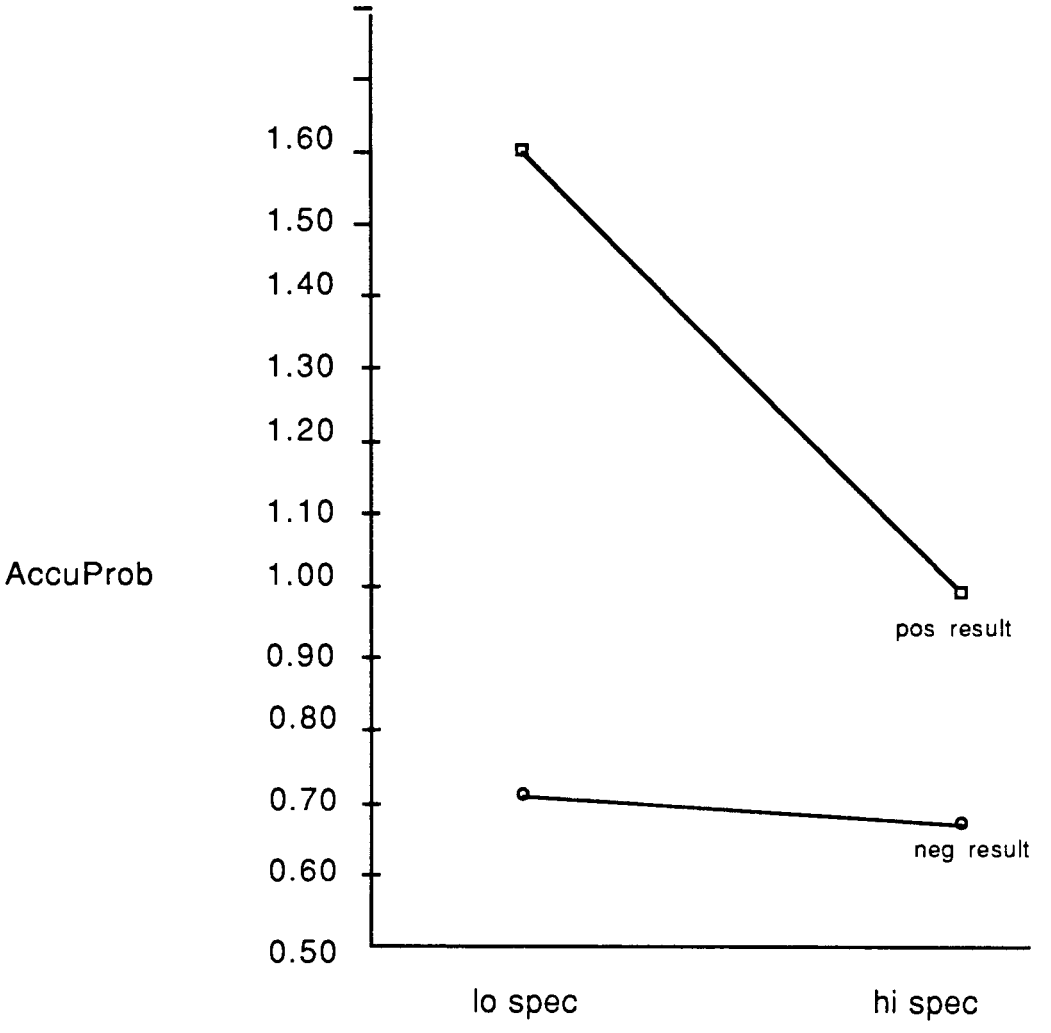


Figure 4. The interaction of test result and specificity on AccuProb.

Thirty seven subjects were asked the question, "at what probability of cancer would your treatment plan change from just 'watchful waiting' to actually doing something?" Four subjects were unable to respond quantitatively to the question. Of the 33 subjects who did respond, the median probability was 10% (range = 01% to 95%). A value of 01% was used to describe variations on the response, "at absolutely any probability." Using the median value of 0.10 as a treatment threshold (that point at which treatment changes from waiting to action), the scatterplot in Figure 2 has been divided into four quadrants. Beginning at the lower left and reading clockwise, the quadrants indicate true wait, false treat, true treat, and false wait. Seventy two subjects fall in one of the "true" quadrants. Twelve subjects estimated a posterior probability of 0.10, three subjects had NormP values of 0.10, and two subjects had both PostP and NormP equal to 0.10. Five subjects fell in the "false treat" quadrant and two in the "false wait" quadrant.

Probabilities and thresholds were also compared individually for the 33 subjects who indicated treatment thresholds. Twenty six had estimated posterior probabilities which were on the same side of their own thresholds as their Bayesian probabilities. Four subjects estimated posterior probabilites that were the same as their thresholds: of these subjects, three had normative values below the threshold and one had a normative value above the threshold. Three subjects estimated posterior probabilites on the opposite side of the threshold as their normative probabilites, all errors being of the "false treat" type.

The mean level of confidence which all subjects expressed in their estimates of prior probability (PrConf) was 5.5 (range = 1 to 10). The mean level of confidence in their estimates of posterior probability (PostConf) was 7.4 (range = 1 to 10). The change in confidence for each subject was determined by calculating

ChngConf = (PostConf - PrConf).

Means and standard deviations of these values for all conditions are shown in Table IV. A three-way analysis of variance (see Table V) showed a significant effect of test result. There was also a significant two-way interaction between test result and sensitivity. A graph of this interaction is shown in Figure 5. Subjects who had received negative results had a mean confidence change of 0.8 for low-sensitivity tests and 1.4 for high-sensitivity tests ($t(45)=-.84$, p N.S.). Subjects who had received positive test results had a mean confidence change of 3.3 for low-sensitivity tests and 2.1 for high-sensitivity tests ($t(44)=2.48$, $p<.025$). There were no effects on change in confidence attributable to specificity.

TABLE IV

MEANS (STANDARD DEVIATIONS) OF CHNGCONF--THE CHANGE IN CONFIDENCE IN ESTIMATE OF POSTERIOR PROBABILITY OF DISEASE FROM THAT EXPRESSED IN ESTIMATE OF PRIOR PROBABILITY OF DISEASE

Negative Test Result

		Sensitivity: Low	Sensitivity: High
Specificity:	Low	+1.2 (2.9)	+1.3 (3.0)
	High	+0.3 (2.2)	+1.5 (2.2)

Positive Test Result

		Sensitivity: Low	Sensitivity: High
Specificity:	Low	+3.7 (1.8)	+2.1 (1.7)
	High	+2.9 (1.8)	+2.1 (1.4)

Note: N = 12 in each cell.

TABLE V

ANALYSIS OF VARIANCE ON CHNGCONF

Factor	SS	df	MS	F ratio	p
Result	63.375	1	63.375	13.06	<.001
Sensitivity	2.042	1	2.042	0.42	N.S.
Specificity	2.667	1	2.667	0.55	N.S.
Res X Sens	20.167	1	20.167	4.16	<.05
Res X Spec	0.042	1	0.042	0.01	N.S.
Sens X Spec	5.042	1	5.042	1.04	N.S.
Res X Sens X Spec	0.167	1	0.167	0.03	N.S.
Error	427.000	88	4.852		

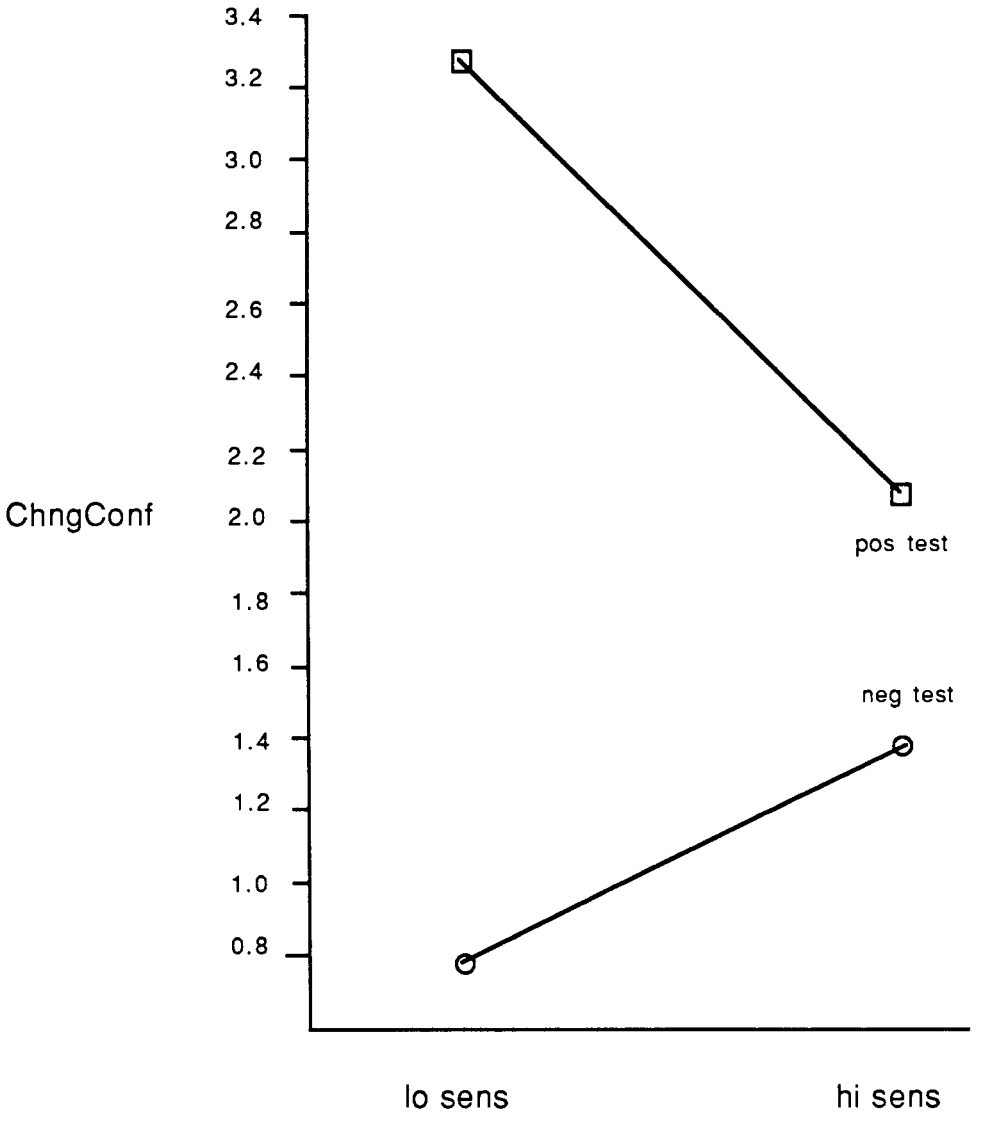


Figure 5. Interactive effects of test result and sensitivity on change in confidence (ChngConf).

A Pearson product moment correlation was calculated between the 96 subjects' change in confidence (ChngConf) and the degree of accuracy of the posterior probability estimate (|Accuprob-1.000|). The result was not significant ($r = -.15$, 95% C.I. = -.34 to +.06, $z = -1.41$).

Forty three subjects responded to the question regarding preference for a test with high sensitivity or one with high specificity. Thirty five were able to choose, although many were uncomfortable with the idea of a forced choice: of these, 18 chose high sensitivity and 17 chose high specificity. Seven subjects responded that it depends on the purpose of the test. One subject insisted that both are necessary.

All subjects were asked whether sensitivity and specificity values were routinely available to them in practice. Sixty two said the values were not routinely available; 22 subjects said they were. Four subjects said the values were available if they looked for them. Four said it depends on the test and two said either sensitivity or specificity was available but not both. Two subjects said they didn't need sensitivity and specificity values.

DISCUSSION

The scatterplot in Figure 2 shows clearly that subjects in this study did not revise their probabilities as Bayesians, based on new information from the mammogram readings. The few subjects whose posterior probability estimates agreed with the Bayesian values are indicated by data points on the diagonal. Subjects who overestimated are above the diagonal and subjects who underestimated are below the diagonal.

The results of the three-way analysis of variance performed on ChangeP also show that subjects did not process test results by the Bayesian model. In fact, they performed quite differently from the model, being influenced by test result and by nothing else. As seen by the analysis on the magnitude of the normative change (NormChng), there should be significant effects of sensitivity and specificity, with the higher levels of each one bringing about a greater change. The interactions between result X sensitivity and result X specificity show the different importances of sensitivity and specificity in the case of positive or negative results. There should be no effect of test result on the absolute value of the magnitude of the change. For the subjects in this study, however, the magnitude of probability adjustment appeared to be only a function of whether the test was positive or negative and had nothing to do with the

validity of the test.

The analysis of AccuProb shows how well subjects estimated probability, relative to what they should have done by the Bayesian formula. These results show, as do the analyses on ChangeP, that subjects tended to overadjust or underadjust probability, depending on whether the test result was positive or negative. It would appear that different heuristics operate in each case.

When test results were positive, subjects tended to overestimate the probability of disease. This would appear to be consistent with representativeness bias, in which outcomes are predicted as a function of their similarity to new information (Kahnemann & Tversky, 1973). When results were negative, however, probability revisions were underestimated. The phenomenon of "cognitive conservatism" is another example of non-Bayesian probability revision, and appears consistent with the behavior of these subjects. Kozielecki (1970) discussed cognitive conservatism as a two-part phenomenon: (1) posterior probability is almost always adjusted in the proper direction from the prior probability, and (2) the change is almost always underestimated. In their study of nurses' revised estimates of the probability of disease, Hammond et al. (1967) also showed that subjects were conservative in their probability revisions. These authors used a measure of accuracy similar to the one used in this study (AccuProb), which they called the "accuracy ratio," the subjective log likelihood ratio divided by the Bayesian log likelihood ratio.

The two different behaviors -- overadjustment when a test is positive and underadjustment when a test is negative -- result in

both cases in an overestimated posterior probability of disease. This is consistent with a tendency to avoid false negative errors in medical diagnosis, and supports the theory that avoidance of this type of error is an important non-Bayesian component of medical diagnosis (Scheff, 1971). Bergman and Pantell (1986) showed that physicians increased their estimates of the posterior probability of disease after reading an article which discussed the complications of that disease but did not give any information which would change the probability values which they had already used in diagnosing. The results of the present study suggest that the high "cost" of a missed diagnosis of breast cancer affects the estimate of its presence. In other words, outcome values are being included in the estimate of probability.

The results of this study suggest that the likelihood ratio fallacy (Fischoff & Beyth-Marom, 1983) may also create a serious problem in medical diagnosis, as the effect of specificity on AccuProb indicates that subjects are not making normative use of specificity. Subjects with low specificity results overestimated the probability of malignancy relative to the Bayesian model (mean AccuProb = 1.156), while subjects who had high specificity results underestimated the probability (mean AccuProb = 0.849). Since there was no significant effect due to the levels of sensitivity, we may conclude that subjects understood better the statistical implications of sensitivity than they did of specificity. Since they were estimating the probability that cancer was present, and since the specificity is the complement of P(H|-D) (the probability that a hypothesis is true, given the nonoccurrence of a datum), the likelihood ratio fallacy is shown in

the context of a commonly demanded professional task.

Among subjects who received positive test results, the error of overestimation was significantly greater with low-specificity tests. This is because subjects tended to revise their probabilities the same, even though lower specificity causes a lower value in the denominator of AccuProb, resulting in a higher ratio. In other words, specificity appears to be ignored in the case of a positive test result. Specificity may also be ignored in the case of a negative test result, with lack of significant difference due to less impact of specificity on the denominator. It is important to note that the ANOVA on the normative change in probability (NormChng) shows that the normative differences were small between conditions. This is because the ten-point range between sensitivity and specificity values of .91 and .81 did not provide wide variations in their effects. The AccuProb differences in this study would conceivably be even greater if the sensitivity and specificity values were farther apart, although this was not feasible in a mammogram scenario.

The question of clinical importance is whether departure from Bayesian reasoning has any effect on a physician's treatment plan. The scatterplot in Figure 2 shows that 74 subjects (77%) fall either in one of the "true" quadrants or on the diagonal at the treatment threshold (0.10). We could assume that the treatment plans of these subjects would be appropriate for their normative probabilities as well as for their own estimates of posterior probability. For the remaining 22 subjects (23%), however, there could conceivably be a change in treatment plan based on more accurate probability determination. Three subjects (3.1%) had NormP values of 0.10; all

of these subjects had estimated posterior probabilities greater than 0.10. For these physicians, there is a danger of treating when it is not indicated. Twelve subjects (12.5%) estimated posterior probabilities of 0.10. Seven of these had NormP values less than 0.10, again suggesting a danger of over treatment, while 3 had NormP values greater than 0.10, suggesting danger of undertreatment. For all of these physicians, understanding that they were at the treatment threshold might help them design better treatment plans. Seven subjects (7.3%) fell into the "false" quadrants--5 in the "false treat" and 2 in the "false wait." Therefore, using the median value of 10% as a treatment threshold, it appears that a better understanding of probabilistic reasoning could have a favorable impact on the treatment plans of 23% of the subjects in this study. It should be noted that moving the threshold toward 50% would result in more subjects in the "false" quadrants. A more central threshold would not be inconceivable in the case of less serious disease, which suggests that the impact of poor probabilistic reasoning may even be underestimated in this study.

A closer look at the 33 subjects who indicated treatment thresholds indicates that better probability judgments could influence the treatment plans in 21% of that subgroup. Three of these subjects estimated posterior probabilities on the opposite side of the treatment threshold from their normative probability. In all three cases, their estimated probability was higher than the normative, indicating a "false treat" error. Of the four subjects who indicated treatment threshold values that were the same as their estimated posterior probabilities, three had overestimated the

posterior probability of malignancy, and one had underestimated. Thus in six subjects (18%) there was an indication of overtreatment and in one (03%) subject an indication of undertreatment, due to lack of understanding of the statistics of probability. Based on these and the above findings, we conclude that at least 20% of physicians would benefit from better understanding of probabilistic reasoning in medical diagnosis.

Although subjects did not use test results normatively to revise their estimates of the probability of malignancy, the mean value of confidence in their estimates increased after the results were received. This increase was almost two points on a scale of ten. The change in confidence was significantly affected by the type of test result, with subjects who had received a positive result being more confident in their revised estimates. This is despite the fact that they significantly overestimated on the basis of the positive result. The fact that there were no main effects on change of confidence due to sensitivity or specificity is inconsistent with the findings by Kahnemann and Tversky (1973), who found that increasing the perceived accuracy of new information increased subjects' confidence in their probability estimates. The fact that subjects who received positive results had significantly greater confidence increases when the sensitivity level of the test was low may be a spurious finding, attributable to the fact that almost 1/20 of all significance tests will have $p<.05$ due to chance alone.

One of the most curious findings of this study is the wide variability of subjects' estimates of the prior probability of disease. The pilot study of part one of the scenario suggested that estimates

would generally fall within the range of 40% to 60%. Although the mean value of subjects in the study was the desired 50%, the range from 02% to100% was surprising. Figure 1 shows that subjects were distributed throughout the range. Two possible explanations for this variability are that the predictive values of historical and clinical indicators are not agreed upon, and that different physicians weight the factors differently in making their clinical judgments. This is clearly a topic for further exploration.

How valid is a case-scenario study in determining professional performance by clinicians? One's ability as an intuitive statistician seems to vary with the context in which it is being used (Beyth-Marom & Fischoff, 1983; Evans, Brooks, & Pollard, 1985). It is worthwhile to consider the conclusions of Goran, Williamson and Gonnella (1973), in which they point out that high scores on case-scenario questions do not always mean better clinical ability, although low scores may very well indicate poor clinical ability. If anything, the findings of the present study may be an underestimate of errors in the real diagnostic setting.

Practical Implications

Available Information. Perhaps part of the problem is the fact that sensitivity and specificity concepts are not used much in an actual clinical setting. Only 23% of the subjects in this study felt that sensitivity and specificity values were routinely available to them in practice. This supports the findings of others -- that sensitivity and specificity values may be hard to get, and often the information

doesn't exist at all (Benish, 1985). In fact, in their experiment which demonstrated overreliance on test result information, Balla et al. (1983) gave only the sensitivity value of the test and withheld the specificity value, claiming that although "the data on test diagnosticity were formally insufficient, [this presentation was] consistent with common, everyday presentation of this information." (p. 25).

Information in medical literature is also often of poor quality. Sheps and Schechter (1984) examined 129 articles in the medical literature using the following Medline criteria: the articles were written in English; they were published in 1982; the words "sensitivity" and "specificity" both appeared in the abstract, title, or key words; they were in journals cited in the abridged Index Medicus. The authors located 129 of the 151 articles disclosed by the Medline search, and assessed them against seven methodological criteria for test validation. Along with other methodological flaws, they found that in 21% of them the terms "sensitivity" and "specificity" were used incorrectly.

Eddy (1982) cites numerous examples from the literature which do not contain sufficient information for diagnosticity. This author also points out that retrospective accuracy (sensitivity and specificity) is often confused in the medical literature with predictive accuracy (positive and negative predictive values), and that medical authors often do not appear to understand the impact of prior probability (disease prevalence or prior probability for the patient) on the posterior probability of disease. Another problem in the medical literature is ambiguity of terms. For example, such terms as

"true positive rate" and "percent true positive" are often used without specifying the denominator (Dolan & Mushlin, 1985; Griner et al., 1981). Is the denominator all positive tests, all diseased individuals, or all tests administered?

It has been estimated (Billings & Bernstein, 1985) that at least one article per month in the major medical journals requires understanding of probability concepts. The bad examples certainly do not constitute the whole of medical literature. There have been excellent contributions which recognize and explain the relationships between probability elements in medical diagnosis in terminology common to physicians with general-level knowledge of statistics (Connell & Koepsell, 1985; Griner et al., 1981; Sox, 1986). In fact, shortly after Sheps and Schechter published the findings discussed above (1984), Human Pathology cited the study and committed in an editorial to seek only papers that "adhere to the critical approach required for validity" (Wagner, 1985, p. 325). The editor claimed that, out of 14 papers which had been accepted in 1984 about new tests using primarily immunocytochemical methods, only six were accepted because of consideration of the issues raised by Sheps and Schechter.

Training. Although subjects in this study may not have understood how to use sensitivity and specificity, they appeared to appreciate that those values were important. Those who were able to make a forced choice between high sensitivity and high specificity were evenly divided between their preferences. Some clearly understood that the two differ in importance according to the test's purpose (e.g. screening, where the prior probability is low, versus

confirmation of a diagnosis, where the prior probability is high). It is important to note, however, that some of these subjects had the importances of high sensitivity and high specificity reversed, advocating high sensitivity for screening and high specificity for confirmation of diagnosis.

If physicians do not handle the concepts of conditional probability well, would this be helped by further training? The medical literature varies in its suggestions concerning this matter. Casscells et al. (1978) recommend instruction to physicians in the theory of test interpretation. Billings and Bernstein (1985) found that residents and medical students performed better than practicing physicians at determining the probability of disease after obtaining results of a diagnostic test, although the difference may have been overestimated due to the methodological flaws discussed earlier in this manuscript. In light of their findings, the authors suggested that probability concepts may be hard to retain when they are not actively used. Other physicians have concluded that there are benefits to statistical training. Borak and Veilleux (1982) found that statistically sophisticated physicians (SP) gave the highest percent of correct answers in probability judgment tasks, compared to practicing physicians, clinical nurses, and hospital laborers. The authors stressed, however, that the majority of responses were still incorrect in the SP group, and that their errors were distributed in patterns identical to the other three groups. The results of their study may suggest that only a few individuals benefit from conventional statistical training. A further study of interest would be to compare training techniques and learning styles of those who

retain statistical information with those who do not. Indeed, Doubilet and McNeil (1985) have suggested that we need more research into cognitive processes, in order to design effective training programs for physicians in handling probabilities.

There are those, however, who would argue against the effectiveness of training (Balla et al., 1985). Bergman and Pantell (1986) found that recent training had no effect on how well physicians processed probability. Their study compared family practitioners, pediatric residents, and board-certified pediatricians. Another group of medical educators who studied the use of formalized decisions in clinical medicine (Schwartz et al, 1973) stated that their teaching experience showed that, although some physicians in training welcomed an analytical approach, others denied its usefulness, claiming that they could make good decisions informally. Borak and Veilleux (1982) felt that statistical training is not the best answer because intuitive reasoning skills are not made better by more education. Indeed, Tversky and Kahneman (1974) found that experienced researchers were subject to the same judgment errors as lay persons.

A particularly ingenious training device was described by Ledley and Lusted in their seminal article on the use of Bayesian reasoning in medical diagnosis (1959). This was a system of cards, each representing a possible diagnosis, sorted on indications by a pin-sort method similar to that which was used by many pre-computer cataloguing systems. Since diagnostic possibilities were dependent on previous diagnoses, the authors recommended that only "carefully evaluated or definitely verified diagnoses should be

used in making up the deck, or at least there should be a provision for review and removal of incorrect diagnoses" (p. 20), lest the process become too cumbersome. The authors claimed that although their invention was "essentially an experimental tool,...undoubtedly more sophisticated forms of the device could be further developed" (p. 20). Indeed, more sophisticated forms have been developed, and we will discuss these below.

Decision Aids. Beyth-Marom and Fischoff (1983) concluded that P(D|-H) will be processed properly in updating probability, if it is presented with equal status as P(D|H). Indeed, it has been recommended that the likelihood ratio, the figure which incorporates both of these values, be promoted as an aid to decision making (Doubilet, 1983). Pascoe (1986) suggests that physicians should be taught to request the likelihood ratio along with test results, reminding readers of three important advantages to physicians: (1) Like sensitivity and specificity, it is calculated independently of the prior probability in the validation population. (2) It can be calculated at different levels of continuous-outcome test results, and thus can be used to make the decision regarding cutoff point between "positive" and "negative" results. (3) It can be used to construct graphic decision aids. It should be noted, however, that use of the likelihood ratio alone results in loss of information of the separate magnitudes of sensitivity and specificity. This information may be important in cases of very high or very low prior probabilities.

Some medical literature pleads the case for large medical data bases to help with probability assessment (Doubilet & McNeil, 1985).

Others in the field have proposed mathematical aids. Connell and Koepsell (1985) designed an aid to deciding whether or not to test, which considers prior probability, test sensitivity and test specificity. Moller-Petersen (1985) recommended a rearranged formula to calculate the posterior probability of disease in "less than a minute" on a pocket calculator. The formula does not appear any less confusing than Bayes' equation, however, and in fact it seems less intuitively obvious. Furthermore, one who understands the Bayesian formula should be able to calculate a posterior probability on a pocket calculator in less than 30 seconds.

In their early article about the use of decision analysis in medicine, Schwartz et al. (1973) maintained that a computerized approach was the most promising direction in aids to clinical decision making, as it will continue to remain unlikely that clinicans will have time to perform elaborate analyses themselves. Borak and Veilleux (1982) also concluded that the best answer is probably to design ways of making probability calculations for clinicans, instead of trying to train them to make their own. This may, indeed, be the best hope for improving the clinical processing of probabilies, and is a growing trend in the literature. For example, one medical journal recently published a simple program written in BASIC for revising probability of disease after receiving a test result, on the basis of sensitivity, specificity, and pre-test probability of disease (Schechter & Sheps, 1985). Using the computer would certainly open the way to more sophisticated use of available diagnostic information. For example, Doubilet and McNeil (1985) point out that the use of the computer will make possible a "probabilistic sensitivity analysis",

providing the potential for stating with certainty that a selected test or test strategy is optimal.

Graphic Aids. Benish (1985) affirmed that a graphic aid "presents a last refuge for those physicians who want to practice good medicine without walking around with a calculator on their belts" (p. 34), and then presented one to calculate posterior probability of disease, given the prior probability, test result, test sensitivity, and test specificity. A less sophisticated one had been proposed eleven years earlier by Katz (1974), who designed a graphic aid to calculate the posterior probability of disease based on a positive result from an "extremely sensitive" test (defined as being "close to 1.0"), as a function of prior probability and test specificity. Specificity was actually stated as its complement, the "false positive frequency" of the test. A nomogram for calculating posterior probability of disease based on the prior probability of disease and the likelihood ratio of a given positive or negative test result has also been published in the medical literature several times in recent years (Fagan, 1975; Moller-Petersen, 1985; Pascoe, 1986).

Graphic aids to diagnosis may also lead the way for more sophisticated decision-making aids which could be carried in busy clinicians' pockets. Doubilet (1983) suggests that mathematic and graphic models would be useful to help the physician make the following decisions, based on probabilities: (1) Should a test be done?; (2) Which test should be done, if more than one test is available? (3) In a test with continuous results, what is the best cutoff point for dichotomization of the results (for example, into "positive" and "negative")? Furthermore, by using graphic aids to

make these decisions, the way is opened for the incorporation of outcome values into the clinical decision making process, bringing closer the use of more comprehensive aids such as decision trees.

Fischoff and Beyth-Marom (1983) have enumerated three ways decision makers may deal with the complexity brought by the conditional nature of events, each one resulting in a fallacy: (1) ignoring the conditional nature of the events, (2) confusing the roles of the datum and the hypothesis, and (3) just feeling confused. The first fallacy is analogous in medical diagnosis to ignoring the separate roles of sensitivity and specificity in the accuracy of the test result, as a function of whether the disease is actually present or absent. The second fallacy is analogous to confusing sensitivity and specificity with posterior probabilities of disease. The third may relate to a rejection of these statistical procedures in clinical medicine. It may be the third consequence which leads to admonishments against viewing the Bayesian model as a substitute for clinical skill (Balla et al., 1985).

Eddy (1982) cites examples of medical maxims which illustrate the diversity of opinions and feelings of medical professionals regarding the use of statistics in clinical medicine. Those who advocate Bayesian reasoning might say "Common things occur most commonly," "Follow Sutton's law: go where the money is" or "When you hear hoofbeats, think of horses, not of zebras." On the other hand, those who oppose Bayesian inference in clinical medicine might say "The patient is a case of one" or "Statistics are for dead men" (p. 259). The battle to introduce normative probability revision

techniques into the process of medical diagnosis may not be without its allies, but it will probably be hard won. In any case, this study shows that there is much room for improvement, and that such improvement would have a significant impact on the effectiveness of many physicians' treatment plans.

REFERENCES

Alloy, L. B. & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. Psychological Review, 91(1), 112-149.

Anderson, B.F., Gaffuri, A., & Morris, R.K. (1986). Treatment-condition bias in the judgment of covariation. Paper presented at the meeting of the Washington Psychological Association, Seattle, Wa.

Balla, J.I., Elstein, A., & Gates, P. (1983). Effects of prevalence and test diagnosticity upon clinical judgments of probability. Methods of Information in Medicine, 22 (1), 25-28.

Balla, J.I., Iansek, R., & Elstein, A. (1985, February 9). Bayesian diagnosis in presence of pre-existing disease. The Lancet, 326-329.

Benish, W.A. (1985). A graphical aid to medical decision making. Henry Ford Hospital Medical Journal, 33 (1), 34-35.

Bergman, D.A. & Pantell, R.H. (1986, May). The impact of reading a clinical study on treatment decisions of physicians and residents. Journal of Medical Education, 61, 380-386.

Beyth-Marom, R. & Fischoff, B. (1983). Diagnosticity and pseudodiagnosticity. Journal of Personality and Social Psychology, 45 (6), 1185-1195.

Billings, P.R. & Bernstein, M.S. (1985, September 6). Physicians poor at prevalence and positive predictive value [Letter to the editor]. JAMA, 254 (9), 1173-1174.

Borak, J. & Veilleux, S. (1982). Errors of intuitive logic among physicians. Social Science and Medicine, 16, 1939-1947.

Casscells, W., Schoenberger, A., & Graboys, T.B. (1978, November 2). Interpretation by physicians of clinical laboratory results. The New England Journal of Medicine, 299 (18), 999-1001.

Connell, F.A. & Koepsell, T.D. (1985). Measures of gain in certainty from a diagnostic test. American Journal of Epidemiology, 121 (5), 744-753.

Dillman, D.A. (1978). Mail and Telephone Surveys: The Total Design Method. New York: John Wiley & Sons.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. Acta Psychologica, 43, 111-121.

Dolan, J.G. & Mushlin, A.I. (1985, November). Routine laboratory testing for medical disorders in psychiatric inpatients. Archives of Internal Medicine, 145, 2085-2088.

Doubilet, P. (1983). A mathematical approach to interpretation and selection of diagnostic tests. Medical Decision Making, 3 (2), 177-195.

Doubilet, P. & McNeil, B.J. (1985, May). Clinical Decisionmaking. Medical Care, 23 (5), 648-662.

Eddy, D.M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, D., Slovic, P., and Tversky, A. (Eds), Judgment Under Uncertainty: Heuristics and Biases (pp. 249-267). Cambridge: Cambridge University Press.

Evans, J.S.B.T., Brooks, P.G., & Pollard, P. (1985). Prior beliefs and statistical inference. British Journal of Psychology, 76, 469-477.

Fagan, T.J. (1975, July 31). Nomogram for Bayes's theorem [Letter to the editor]. The New England Journal of Medicine, 293 (5), 257.

Fischoff, B. & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. Psychological Review, 90 (3), 239-260.

Goran, M.J., Williamson, J.W., & Gonnella, J.S. (1973, February). The validity of patient management problems. Journal of Medical Education, 48, 171-177.

Griner, P.F., Mayewski, R.J., Mushlin, A.I., & Greenland, P. (1981, April). Selection and interpretation of diagnostic tests and procedures: Principles and applications. Annals of Internal Medicine, 94 (4, part 2), 453-600.

Hammond, K.R., Kelly, K.J., Schneider, R.J., & Vancini, M. (1967, Winter). Clinical inference in nursing: Revising judgments. Nursing Research, 16 (1), 38-45.

Kahneman, D. & Tversky, A. (1973, July). On the psychology of prediction. Psychological Review, 80 (4), 237--251.

Katz, M.A. (1974, November 21). A probability graph describing the predictive value of a highly sensitive diagnostic test. The New England Journal of Medicine, 291 (21), 1115-1116.

Kozielecki, J. (1970). Psychological characteristics of probabilistic inference. Acta Psychologica, 34, 480-488.

Ledley, R.S. & Lusted, L.B. (1959, July 3). Reasoning foundations of medical diagnosis. Science, 130 (3366), 9-21.

Moller-Petersen, J. (1985, February 9). Nomogram for predictive values and efficiencies of tests [Letter to the editor]. The Lancet, 8424 (1), p. 348.

Mushlin, A.I. (1985, July). Diagnostic tests in breast cancer: Clinical strategies based on diagnostic probabilities. Annals of Internal Medicine, 103 (1), 79-85.

NWA STATPAK (1985, Version 3.2). Northwest Analytical, Inc. Portland, Or.

Pascoe, J.M. (1986). Use of the likelihood ratio in the management of the young child with fever. The Journal of Family Practice, 22 (4), 349-352.

Schechter, M.T. & Sheps, S.B. (1985, April 1). Diagnostic testing revisited: Pathways through uncertainty. Canadian Medical

Association Journal, 132, 755-760.

Scheff, T.J. (1971). Decision rules and types of error, and their consequences in medical diagnosis. In Freidson, E. (Ed), Medical Men and Their Work (pp. 309-323). Chicago: Aldine Publishing Co.

Schwartz, W.B., Gorry, G.A., Kassirer, J.P., & Essig, A. (1973, October). Decision analysis and clinical judgment. The American Journal of Medicine, 55, 459-472.

Sheps, S.B. & Schechter, M.T. (1984, November 2). The assessment of diagnostic tests: A survey of current medical research. JAMA, 252 (17), 2418-2422.

Sox, H.C. (1986, January). Probability theory in the use of diagnostic tests: An introduction to critical study of the literature. Annals of Internal Medicine, 104 (1), 60-66.

Swets, J.A., Feehrer, C.E., Greenes, R.A., & Bynum, T.E. (1986). Use of probability estimates in medical communications and decisions. Methods of Information in Medicine, 25 (1), 35-42.

SYSTAT (1985, Version 3.0). SYSTAT, Inc., Evanston, Il.

Tversky, A. & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Von Winterfeldt, D. & Edwards, W. (1986). Decision Analysis and Behavioral Research. Cambridge: Cambridge University Press.

Wagner, B.M. (1985, April). When is a test diagnostic? [Editorial]. Human Pathology, 16 (4), 325.

APPENDIX A.
LETTER TO MEMBERS OF SUBJECT POOL

THE OREGON
HEALTH SCIENCES UNIVERSITY
^D 3181 S.W. Sam Jackson Park Road Portland, Oregon 97201 (503) 279-7590, 279-5321

Department of Family Medicine
School of Medicine

Dr. name and address

Dear Dr. name:

Within a week or two, you may receive a telephone call from Portland as part of a research project to study how physicians make clinical judgments. As these processes become better understood, it will become possible to improve our methods of professional education and to design aids for diagnostic reasoning for physicians in training and in practice.

We are writing in advance of our telephone call because we have found that many people appreciate being advised that a research study is in process and that they will be called. You are one of 200 active members of the Oregon Academy of Family Physicians who have been randomly selected for a pool of potential subjects. Physicians will be contacted from this pool until 96 interviews have been completed.

When our interviewer calls, she or he will ask to speak to you in connection with this study. If you are busy at the time and would prefer the interviewer to call back, please convey that message through your receptionist. We will be happy to call back at whatever time would be convenient for you. Altogether, the interview should be take less than ten minutes. Your name will not be recorded by the interviewer, so your responses to the questions will be anonymous.

Your help and that of the other physicians being asked to participate in this study of clinical decision making is essential to the project's success. We greatly appreciate it.

If you have any questions, please do not hesitate to ask our interviewer. If you prefer, you may contact me by telephone at 279-7855 or by mail.

Sincerely,

Ann Sinclair
Senior Research Assistant

as.007

APPENDIX B
TELEPHONE SCENARIO, PART ONE

SUBJECT NUMBER:_____YOB:_____TELEPHONE NUMBER:_____

INTERVIEWER:_____DATE:_____TIME BEGAN:_____TIME ENDED:_____

****

"THIS IS _____. I'M CALLING FROM THE OREGON HEALTH SCIENCES UNIVERSITY DEPARTMENT OF FAMILY MEDICINE. MAY I PLEASE SPEAK WITH DR. _____?"

[CALL BACK APPOINTMENT:_____]

[OTHER NOTES:_____]

"THIS IS _____, FROM THE OREGON HEALTH SCIENCES UNIVERSITY [OHSU] DEPARTMENT OF FAMILY MEDICINE. WE SENT YOU A LETTER ABOUT A TELEPHONE SURVEY WE'RE CONDUCTING ON CLINICAL DECISION MAKING. THE QUESTIONS INVOLVE A HYPOTHETICAL CASE SCENARIO, AND SHOULD TAKE LESS THAN 5 MINUTES. IS THIS A CONVENIENT TIME FOR YOU?"

[CALL BACK APPOINTMENT:_____]

[OTHER NOTES:_____

_____]

"THIS IS A 50 YEAR OLD WHITE FEMALE PATIENT WHOM YOU HAVE NEVER SEEN. SHE HAS COME TO SEE YOU BECAUSE OF A BREAST LUMP THAT SHE HAS NOTICED WITHIN THE LAST COUPLE OF WEEKS. THE LUMP IS PAINLESS. SHE HAS NO FAMILY HISTORY OF BREAST CANCER. SHE IS MARRIED AND HAS ONE ADULT SON.

ON PHYSICAL EXAMINATION, YOU FIND A ONE-CENTIMETER, HARD MASS WITH WELL-DEFINED, REGULAR BORDERS. THE MASS IS LOCATED IN THE UPPER OUTER QUADRANT OF HER LEFT BREAST.

AT THIS POINT, I'D LIKE TO ASK YOU WHAT YOU THINK IS THE PROBABILITY THAT THE MASS IS MALIGNANT?"

[INTERVIEWER MAY REPEAT ANY OF THE ABOVE INFORMATION. ANY ADDITIONAL INFO IS "NOT AVAILABLE TO YOU AT THIS TIME."]

_____[MUST BE EXPRESSED NUMERICALLY]

"HAVING ESTIMATED THE PROBABILITY OF MALIGNANCY, HOW CONFIDENT ARE YOU THAT YOUR ESTIMATE IS CORRECT? PLEASE EXPRESS THIS VALUE ON A SCALE FROM 1 TO 10, WHERE '1' IS JUST A GUESS AND '10' IS ABSOLUTELY CERTAIN:

_____[SCALE OF 1 TO 10]

APPENDIX C
TELEPHONE SCENARIO, PART TWO (EXAMPLE)

"YOU HAVE DECIDED TO ORDER A MAMMOGRAM FOR YOUR PATIENT. ONCE THIS IS COMPLETED, THE RADIOLOGIST SENDS YOU A REPORT STATING THAT THE LESION APPEARS BENIGN. THE REPORT ALSO STATES THAT, IN YOUR HOSPITAL, 91% OF SUBSEQUENTLY PROVEN BREAST CANCERS ARE READ ON MAMMOGRAM AS "MALIGNANT" AND 91% OF SUBSEQUENTLY PROVEN NONCANCEROUS LESIONS ARE READ AS "BENIGN." IN OTHER WORDS, THE SENSITIVITY IS 91% AND THE SPECIFICITY IS 91%.

NOW THAT YOU HAVE THESE RESULTS, I'D LIKE YOU TO ESTIMATE AGAIN THE PROBABILITY THAT THE LESION IS MALIGNANT.

[MAY REPEAT ANY OF THE ABOVE INFORMATION, INCLUDING PART 1]

_____[MUST BE EXPRESSED NUMERICALLY]

"ONCE AGAIN, HOW CONFIDENT ARE YOU THAT YOUR ESTIMATE IS CORRECT? PLEASE EXPRESS ON A SCALE FROM 1 TO 10, WHERE '1' IS JUST A GUESS AND '10' IS ABSOLUTELY CERTAIN:"

_____[SCALE OF 1 TO 10]

***

"ALTHOUGH YOUR ANSWERS WILL BE CODED ANONYMOUSLY, I'D LIKE TO ASK YOU A FEW DEMOGRAPHIC QUESTIONS TO HELP US DESCRIBE OUR DATA SAMPLE."

WHAT YEAR DID YOU GRADUATE FROM MEDICAL SCHOOL?_____

HOW MANY YEARS HAVE YOU BEEN IN PRACTICE?_____

ARE SENSITIVITY AND SPECIFICITY VALUES ROUTINELY AVAILABLE TO YOU?

__NO __YES

COMMENTS?_____

_____

_____

[OK TO ANSWER QUESTIONS AT THIS POINT. STUDY IS LOOKING AT HOW INFORMATION ABOUT THE SENSITIVITY AND SPECIFICITY OF A LABORATORY VALUE IS USED, ALONG WITH THE RESULT OF THE TEST. IN THIS CASE, THE TEST IS MAMMOGRAPHY, ALTHOUGH WE'RE INTERESTED IN OTHER LABORATORY TESTS AS WELL.]

"WOULD YOU BE INTERESTED IN RECEIVING A REPORT OF THE STUDY WHEN IT IS WRITTEN"

___NO ___YES--NAME:_____

"THE LAST THING I'D LIKE TO ASK YOU IS NOT TO DISCUSS THIS INTERVIEW WITH ANY OF YOUR COLLEAGUES FOR AT LEAST A MONTH, BECAUSE THEY MAY BE ON OUR LIST OF POTENTIAL SUBJECTS TO CALL."

"THANK YOU VERY MUCH FOR YOUR TIME."