Prediction in Forensic and Neuropsychology

Sound Statistical Practices

Edited by

Ronald D. Franklin
St. Mary's Hospital
and Florida Atlantic University

LEA
2003
LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey London

Bayesian Inference and Belief Networks

Ronald D. Franklin[1]
St. Mary's Hospital and Florida Atlantic University

Joachim Krueger
Brown University

Clinicians routinely draw inferences from test results to the test taker's latent condition. After all, "a large part of medicine is practiced on people who do not have obvious illnesses, but rather have signs, symptoms, or findings that may or may not represent an illness that should be treated" (Eddy, 1984, p. 75). In the simplest case, both test results and latent conditions are dichotomous variables. Tests turn out either positive or negative, and test takers either do or do not have the disease in question. The inference of interest is a predictive judgment of whether a client with a positive test result has the disease. This judgment depends on several cues. Some of these cues are external, such as the information contained in the client's file (including test scores), whereas other cues are internal, such as the prevalence of the disease in the population or the clinician's experience and memory of related cases. In this chapter, we address the integration of external and internal cues in simple Bayesian inference and in more extended belief networks.

Published posthumously, Bayes' (1763) essay on the doctrine of chances became the cornerstone of a theory of probability that accounts for the interplay of internal, or subjective, and external, or objective, cues. The theory provides methods of principled inductive reasoning and it permits probabilistic predictions about individual cases. The centerpiece of this approach is a theorem stating how beliefs are to be revised in light of evidence. A practicing psychologist is interested in the probability that a person has a certain disease given the presence of a positive result on a test designed to detect this disease. To estimate this

[1] Please address correspondence to PO Box 268, Candor, NC 27229 or rdfphd@yahoo.com.

probability (i.e., $p(D_o|P_o)$),[2] which is also referred to as the *positive predictive value* of the test, the psychologist needs to consult the following probabilities. The first probability, which is externally provided by test developers, is the *sensitivity* of the test. Sensitivity is the probability that a person obtains a positive test result given that the person is known—by whatever other independent and valid method—to have the disease (i.e., $p(P_o|D_o)$). The second probability, which is also externally provided, is the *specificity* of the test. Specificity is the probability that a person obtains a negative result given that the person is known *not* to have the disease (i.e., $p(P_n|D_n)$). Together, sensitivity and specificity capture the overall accuracy or "efficiency" of the test. The third probability, which is internal, refers to the psychologist's prior estimate that the person has the disease (i.e., $p(D_o)$). If random sampling can be assumed, the prior probability of the disease corresponds to the base rate of the disease in the population. If random sampling cannot be assumed, prior estimates may vary depending to the availability of other cues (e.g., symptoms) and the psychologist's experience with interpreting these cues. The partial subjectivity of prior estimates leaves room for divergent diagnostic inferences drawn from the same test results.

$$P(Do|Po) = \frac{P(Do) * p(Po|Do)}{p(Do) * (p(Po|Do) + (p(dn) * p(Do|Dn)} \qquad (4.1)$$

Bayes' rule shows how internal prior estimates of the disease combine with the external efficiency information to yield the desired positive predictive value. The probability that a person who tested positive is sick is the product of the prior probability of a person to be sick and the so-called diagnostic ratio. The diagnostic ratio is the sensitivity of the test divided by the overall probability of a test result to be positive. The denominator of the diagnostic ratio is the sum of the probability that a person tests positive *and* is sick, and the probability that a person tests positive *and* is healthy. The first of these joint probabilities is the product of the prior probability of being sick, $p(D_o)$, and the test's sensitivity, $p(P_o|D_o)$; the second joint probability is the product of the prior probability of being healthy, $p(D_n)$, and the complement of the test's specificity, $p(P_o|D_n)$. In other words, $p(P_o)$ is itself a combination of internal and external information. The formula reads

[2] In this notation, the subscript 'o' stands for 'occurrence' (of a disease or a positive test result), whereas the subscript 'n' stands for 'nonoccurrence.'

Diagnosis and Uncertainty

Consider a hypothetical psychiatric scenario (Cohen, 1994). The base rate of schizophrenia is assumed to be low ($p(D_o) = .021$), and a test designed to diagnose schizophrenia is assumed to have excellent sensitivity ($p(P_o|D_o) = .952$) and specificity ($p(P_n|D_n) = .969$). Then, the probability of schizophrenia in a randomly tested person with a positive result is .40, namely

$$P(Do|Po) = \frac{.021 * .952}{.021 * .952 + .979 * .031} \qquad (4.2)$$

The increase in the estimated probability of schizophrenia from .021 to .4 reflects the degree to which the psychologist has become less certain that this individual is healthy. Because a categorical decision concerning the person's health status has become more, rather than less certain, further testing is indicated. Such testing is most efficient if it is conditionally independent of the initial testing, that is, if the results of the two tests are unrelated within the population of sick people and within the population of healthy people. If such independence can be assumed, the posterior probability of the disease obtained after the first test (i.e., .40) can serve as the prior probability for the second test (Winkler, 1993). In Cohen's (1994) example, confidence in the presence of the disease would rise to .95 if a positive result were obtained again and if the second test were as sensitive and as specific as the first one. If, however, the follow-up tests are not independent, confidence levels will rise more slowly. Still, sequential testing is a powerful strategy because it overcomes the psychometric limitations of single tests. In the present example, a test would have to have extraordinary sensitivity and specificity (with both $p = .999$) so that a single positive result would yield a positive predictive value of .95.

Sequential testing rapidly dilutes initial differences in clinical opinion, an effect that is often overlooked by critics of Bayesian subjectivism. Potentially divergent prior estimates enter the chain of inferences only once, whereas test results accumulate over time (Lindley, 1993; Thorndike, 1986b). For illustration's sake, suppose a more liberal psychologist approaches the diagnosis of schizophrenia with a prior estimate of .2. This belief would result in a posterior probability of .89 after the first test and a posterior of .996 after the second test, at which point the liberal's judgment hardly differs from the conservative's (i.e., by .04).

Bayesian probability estimation is not a recipe for decision making, but it offers a platform on which treatment decisions can be placed. To be able to decide whether to diagnose a suspected disease, diagnosticians must set a confidence threshold (i.e., a minimum predictive value for a positive diagnosis). The location of this threshold depends on the costs and benefits to the patient (Satz, Fennell, & Reilly, 1970). The lower the threshold, the greater the number of patients who receive a false positive diagnosis with the attendant undesirable consequences. The conservative and the liberal diagnosticians in the example would respectively expect 5% and .4% of their treatments to be wasted. For the former, the expected false positive rate would be greater than the rate indicated by the complement of the test's specificity (here: 3%); for the latter, it would be lower.

Dealing With Base-Rates

As base-rates become more extreme, their effects on diagnostic decisions become larger. This base-rate effect is crucially important because most tests are designed to detect conditions that are rare in the population. A test is most valuable if it allows the diagnosis of a specific case in a way that contravenes base-rate expectation. Critics of clinical decision making often note that people—novices and experts alike—grossly overdiagnose rare diseases (Dawes, 1994; Faust & Ziskin, 1988). Some clinicians hesitate to think probabilistically about individual cases, believing that group statistics are "of no use for the individual case" (Gigerenzer, Hoffrage, & Ebert, 1998, p. 204), or that sensitivity and specificity "predict test scores, not disorders" (Elwood, 1993, p. 230). If taken too seriously, however, this line of reasoning would also keep one from having a preference between a gun containing one bullet and a gun containing 5 bullets in a game of Russian roulette (Dawes, Faust, & Meehl, 1989). Other clinicians may try to think Bayesian but inadvertently confuse predictive accuracy with retrospective accuracy. It is indeed tempting to believe that high test sensitivity directly implies the presence of the disease given a positive result. Yielding to this temptation, many clinicians equate the posterior probability of the disease, $p(D_o|P_o)$, with the sensitivity of the test, $p(P_o|D_o)$, without proper regard for the base-rate of the disease, $p(D_o)$. This error has been

variously characterized as overconfidence, base-rate neglect, or the confusion of the inverse, and it has been attributed to fallible judgmental heuristics such as representativeness, availability, or anchoring with insufficient adjustment (see Dawes, 1988, for a review). In a classic survey, 95% of the participating physicians did not distinguish between the probability of a positive X-ray given cancer and the probability of cancer given a positive X ray (Casscells, Schoenberger, & Grayboys, 1978). In a conceptual replication study, most AIDS counselors were certain that a positive test indicated that a low-risk testee was infected (Gigerenzer et al., 1998). To be consistent, anyone who ignores the prior probability of the disease would also have to ignore its posterior probability after the first test when interpreting the results of sequential tests—hardly a desirable prospect.

The over prediction bias may arise not only from the fallibility of statistical intuitions among practitioners, but also from the way in which they acquire medical knowledge. Students "learn the signs and symptoms that occur with each disease, and most medical knowledge is organized according to disease" (Eddy & Clanton, 1982, p. 1263). Many textbooks offer the mistaken advice that positive test results indicate the presence of the condition in the tested individual regardless of the base-rate of that condition in the population (Eddy, 1982). Defenders of clinical (and other intuitive) judgment argue that practitioners reason rather well even when their judgmental task is far more complex than the judgment required in the present single-test, single-disease scenario (Eddy & Clanton, 1982). Base-rates are hardly ever completely ignored (Koehler, 1996), and judgments appear to be more rational when decision utilities are considered in addition to Bayesian probabilities (Birnbaum, 1983). Others suggest that judgments improve when clinicians are vividly reminded of the relevance of base-rates (Garb & Schramke, 1996) or when the input data are presented as frequencies (Gigerenzer & Hoffrage, 1995).

The latter recommendation is intriguing because it appears to obviate the entire Bayesian enterprise of integrating prior belief (i.e., base-rate information) with empirical evidence (i.e., test results). Cohen (1994) himself presented his numerical example in a frequency format so that "the situation may be made clearer" (p. 999). Table 4.1 shows the data. The marginal frequencies refer to the individuals with each latent status (sick vs. healthy) and each test result (positive vs. negative), and the cell frequencies refer to the four joint occurrences. Given these frequencies, the probability of the disease given a positive test is easily obtained by dividing the frequency of co-occurrence of the disease and a positive test result (here: 20) by the marginal frequency of a positive result (here: 50). No base-rate probability of the disease appears to be necessary. It is also evident that the predictive probability is much smaller than the sensitivity of the test (i.e., 20/21).

TABLE 4.1
Joint Frequencies of the Occurrence of the Disease (D_o) versus Its Nonoccurrence (D_n) and Positive (P_o) versus Negative (P_n) Test Predictions

	D_o	D_n	Total
P_o	20	30	50
P_n	1	949	950
Total	21	979	1000

But where do frequency tables originate? If clinicians were able to classify each incoming case correctly into one of the four cells of the table, test information would be superfluous. The clinician would already know if the person was really sick! Alas, such knowledge is not available in the real world, and diagnosticians cannot count on "natural sampling" of signs and diseases. Instead, they require probabilities to derive frequencies. The frequency of joint occurrence of a positive result and the disease, for example, is the sensitivity of the test times the base-rate of the disease times the total number of cases (i.e., $f(P_o \text{ and } D_o) = p(P_o|D_o) * p(D_o) * N) = .969 * .021 * 1{,}000 = 20$). The computation of frequencies is easily computerized (Sedlmeier, 1997). Once obtained, the visual display of frequencies is an effective aid to clinical inference and to the communication of these inferences to clients (Hoffrage, Lindsey, Hertwig & Gigerenzer, 2000). Nevertheless, because probability information remains essential for the construction of frequency tables, clinicians might as well compute their predictive estimates directly (Dawes, 2000; Jones, 1989).

Because knowledge of base-rates remains vital, we return to the question of what the relevant base-rates are. Although Cohen's (1994) example illustrates the mutual dependencies among conditional and unconditional probabilities, it assumes that testing occurs in a random sample of the general population. In clinical settings, however, patients are rarely sampled randomly. Clinical sampling bias is likely to increase the *available* base-rate. Many clients present themselves or are referred for assessment because other probabilistic cues (e.g., symptoms) suggesting the presence of a condition have already been observed. If, for example, the available base-rate is .5, the posterior probability of the disease lies between the values of test sensitivity and specificity. The need to compile local base-rates for local use highlights a difference between test development and test application (Meehl & Rosen, 1955). Test development can yield excellent levels of sensitivity and specificity because it operates on contrast groups of roughly equal size. The challenge of test construction is to find independent and valid criteria (i.e., a "gold standard", Elwood, 1993) for whether patients have the condition to be detected. In contrast, test application and judgments about individuals must incorporate the counterfactual idea of what the judgment would have been if no test results were in evidence. In other words, sampling bias must be assessed independently of the test results at hand.[3]

Decomposing Accuracy

Both test sensitivity and specificity must be known (along with the base-rate of the disease) for a predictive estimate to be reached. In Cohen's (1994) example, both types of accuracy are high, so their differential effects are easily overlooked. A test could be perfectly sensitive and yet diagnostically useless. Such a situation could arise if the test had no specificity so that every test taker would receive a positive result. The overall accuracy of a test is captured by the diagnostic ratio of sensitivity, $p(P_o|D_o)$, over the complement of specificity, $p(P_o|D_n)$.[4] If

[3] This requirement highlights the need for multiple assessment methods. Without circularity, results from the focal test cannot simultaneously generate a diagnosis for individuals and base-rate estimates for the available population. Bayes's Rule requires that the latter affects the estimate of the former.

[4] A drawback of this measure is that it has no upper limit. Measures of association (i.e., between clients' actual health status and their test results) that do not depend on variations in the base-rates of D_o and P_o offer useful alternatives (e.g., coefficient g, Goodman & Kruskal, 1954, or coefficients of discrimination derived from signal detection theory, Snodgrass & Corwin, 1988).

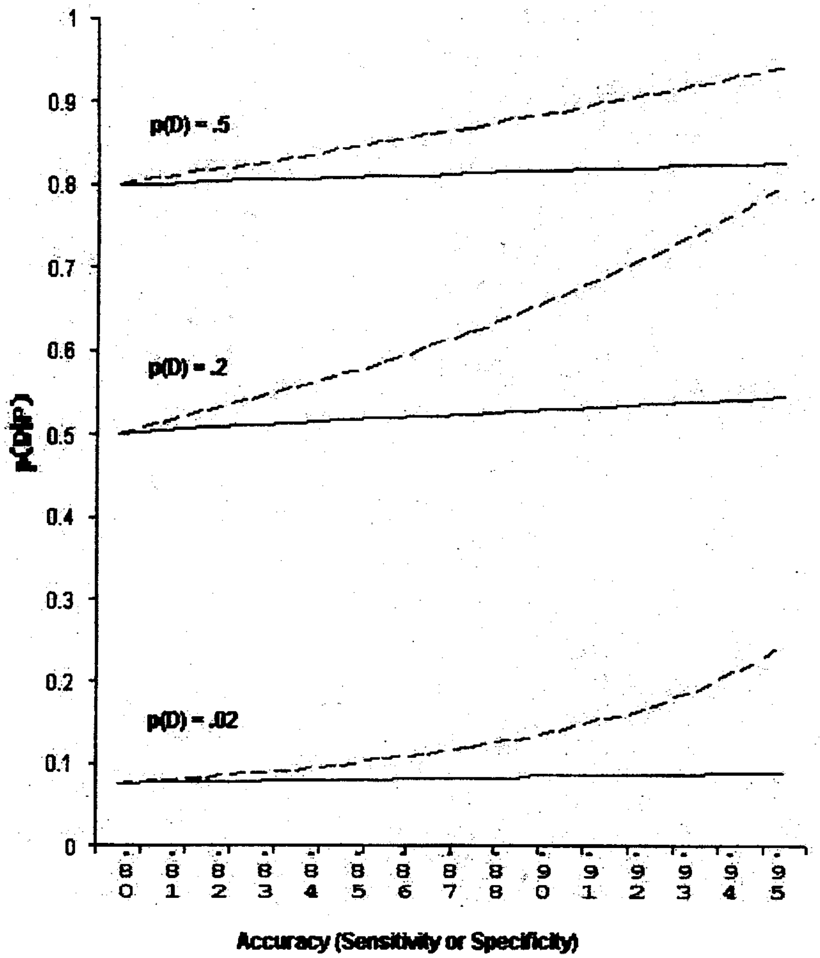


FIG. 4.1. Posterior Probabilities of Disease for Three Base Rates. The solid lines show the effect of variation in test sensitivity for a constant specificity of .8. The dashed lines show the effect of variation in test specificity for a constant sensitivity of .8.

sensitivity and specificity cannot both be maximized, test makers must choose between putting a premium either on correctly diagnosing the presence of a disease or on correctly diagnosing its absence. By

increasing sensitivity, they risk decreasing specificity, and vice versa. When a test is administered as a screening device for a rare disease, a lack of specificity may seem acceptable because most healthy clients still get a negative result. Often, test constructors' primary goal is to ensure that "if a disease is present, it should be found, even at the risk of getting a high rate of false positive results" (Feinstein, 1978, pp. 111-112).[5]

Contrary to the idea that increases in sensitivity are paramount, Bayes' rule shows that increases in specificity entail the largest increases in positive predictive value. Assuming either a high (.5, top lines), moderate (.2, center lines), or low prior probability of D_o (.02, bottom lines), Fig. 4.1 plots $p(D_o|P_o)$ against test accuracy levels ranging from .8 to .95. When specificity is constant at .8, increases in sensitivity yield hardly discernible increases in diagnostic confidence, as shown by the solid lines. When, however, sensitivity is .8, the same increases in specificity yield dramatic decreases in confidence, as shown by the dashed lines. Bayes' formula reveals why this is so. An increase in sensitivity, $p(P_o|D_o)$, affects both the numerator and the denominator of the diagnostic ratio. In contrast, an increase in test specificity entails a decrease in $p(P_o|D_n)$, which affects only the ratio's denominator. As the denominator becomes smaller, the posterior probability of D_o rises rapidly. When base-rates are low, a tolerance for low specificity reduces predictive accuracy, $p(D_o|P_o)$. Clinicians who overlook this consequence of Bayes's rule may end up making grossly inaccurate judgments.

Diagnosing Health Without Stating the Obvious

Even when a test yields a negative result for a low base-rate condition, a revision of belief is in order. In Cohen's (1994) example, a negative test result suggests that the probability of the disease has decreased from .021 to .001. To the client, a negative result may bring the desired peace of mind. This gain is particularly welcome to the extent that prior base rates of being healthy were underestimated. Aside from the disconfirmation of such mistaken expectations, diagnosing health beyond already high base rates can be useful. In genetic testing, for example, prospective parents may wish to assess the risk of having offspring with a recessive disorder. Even if only one prospective parent is tested and obtains a negative result, the probability that offspring will be affected decreases dramatically. Consider again Cohen's hypothetical data and suppose that a genetic defect will be expressed if both parents carry the

[5] This preference is not universal. When positive diagnoses lead to risky treatments, the costs of a false positive can be great.

gene. Assuming that mate selection is not affected by the presence of this gene, the probability of a defect is the product of the two base rates (i.e., $p(D_o)^2 = .02^2 = .004$). If, after testing, the probability of one parent to be a carrier can be said to be reduced to .001, the probability of risk falls to .00002 (i.e., .02 * .001). Because people are not sensitive to differences among extreme probabilities (Mellers & McGraw, 1999), test administrators may wish to express the gain as improved odds. In the present example, the negative test result suggests a 20-fold reduction of risk (i.e., from 40 to 2 in 100,000).

One of the authors experienced a counselor's difficulty of communicating the implications of a negative test result. When asked how much the risk of carrying the Tay-Sachs gene had been reduced by a negative test, the counselor insisted that the test produces few false negatives.[6] As Bayes's rule states, however, high specificity (i.e., a low $p(P_o|D_n)$) by itself does not reveal how improbable the condition is given a negative result (i.e., a low $p(D_o|P_n)$). Predictive judgments of health after a negative test result vary more with changes in specificity than with changes in sensitivity. These decreases in $p(D_n|P_n)$ are still small, however, because the posterior probability of health after a negative test result cannot be smaller than the prior probability of health, which is high when sampling is random. In sum, a loss of specificity selectively affects diagnoses of disease if, as is usually the case, the base rate of the disease in the population is below .5.

A Neuropsychological Example

Solomon et al. (1998) reported impressive efficiency data for a brief ("7 minute") screening test for Dementia of the Alzheimer Type (DAT). In a cross-validation sample, the test was 92% sensitive and 96% specific. Combined with the base-rate probabilities for DAT among the elderly (Evans et al., 1989), the test's efficiency leads to positive predictive values of .42, .84, and .95 for persons aged 65 – 74, persons aged 75 – 84, and persons older than 85, respectively (see Table 4.2, center). As Cohen's (1994) example already illustrated, a single positive result may lead to the "paradoxical consequence that *deciding on the basis of more information can actually worsen the chances of a correct decision*" (Meehl & Rosen, 1955, p. 202, emphasis in the original). When, for example, the test result is used for the youngest subgroup of the elderly, 2.76% of diagnoses are correct positives (i.e., $p(D_o) * p(P_o|D_o) * 100) = .03 * .92 * 100$) and 93.12% of diagnoses are correct negatives (i.e., $p(D_n) * p(P_n|D_n) * 100 = .97 * .96 * 100$). The sum of correct diagnoses is 95.88%, which is worse than the 97% accuracy that would be achieved if, following the base rate, everyone were judged to be healthy. To improve predictive accuracy beyond base rate accuracy, the ratio of the true positive rate (sensitivity) to the false positive rate (1-specificity) must be larger than the ratio of the negative base rate to the positive base rate (Meehl & Rosen, 1955). That is, For the young old, the diagnostic ratio and the base rate ratio are 23 and 32, respectively, indicating that additional testing remains necessary. Perfect reliability and validity cannot be expected from psychometric tests. Even if the sensitivity of the 7-minute test were to increase to from .92 to .99, with specificity remaining at .96, the positive predictive value for the young old would still be .42. Conversely, as our foregoing analysis suggested, huge losses in sensitivity are compensated by small gains in specificity when the base rate of the disease is low. A drop in sensitivity .46 would be offset by an increase in specificity to .98.

$$\frac{P(Po \mid Do)\, p(Do)}{p(Po \mid Dn)} > \frac{p(Dn)}{p(Do)} \qquad (4.3)$$

[6] Gigerenzer et al. (1998) reported that "if the client asked for clarification more than twice, the [AIDS counselors] were likely to become upset and angry, experiencing the client's insistence on clarification as a violation of social norms of communication" (p. 202).

TABLE 4.2
Bayesian Diagnosis of Alzheimer's Disease

	Age Group		
	65-74	75-84	=>85
Prior probability of DAT	.030	.190	.470
Prior probability of other dementia	.003	. 018	.045
Positive predictive value	.420	.840	.950
Revised positive predictive value	.400	.780	.870
Spoiling effect	.020	.060	.080

The Spoiling Effect

Solomon et al. (1998) constructed a screening test for DAT by contrasting a group of patients with known DAT with a community control group—of equal size—in which no one showed evidence of neuropsychological impairment. The estimate of specificity was "pure" in the sense that the available sample did not include patients who might have tested positive because of pathologies other than DAT. The drawback of the contrast-group method is that it tends to overestimate the specificity of the test in the general population (Ransohoff & Feinstein, 1978). Evans et al. (1989) estimated that 8.8% of the demented patients "have only a cause of dementia other than Alzheimer's disease" (p. 2554). Because these patients are more likely than normals to test positive, the false positive rate is higher and thus the predictive value of a positive result is lower than the test construction data suggest. Inasmuch as the base rates of different dementias are correlated, the size of the spoiling effect increases with the base rate of DAT. The illustrative values displayed in Table 4.2 show this effect. Values were computed assuming that the false positive rate forthe non-DAT group is the same as the true positive rate for the DAT group.

BAYESIAN SCORE ESTIMATION

Bayesian models offer a way of thinking through uncertainty by combining expectations with evidence in a disciplined way. The categorical prediction tasks we have considered so far are relatively simple examples from clinical practice. The range of applicability for Bayesian methods is much broader, however. Before they can make clinical diagnoses, for example, neuropsychologists often need to integrate psychometric test data with other cues to infer a person's true performance level. Similarly, they need to integrate test scores to estimate true performance levels and predict future test scores.

Thorndike (1986a) gave an example where a test with a population mean of 100 and a standard deviation of 16 is administered to the same person two years apart (Equation 4.4). The stability correlation for this interval (r = .85) determines the precision with which a score at Time 1 predicts a score at Time 2. The standard error of this prediction is $SE_{pre} = 16 * (1 - .85^2)^{.5} = 8.43$. The reliability coefficient of the test at Time 2 (r = .94) determines that the standard error of measurement at Time 2 is $SE_{meas} = 16 * (1 - .94)^{.5} = 3.92$. If a test taker scored 115 at Time 1 and 125 at Time 2, then we estimate the that the individual's true score lies a bit closer to the second test score than to the first

$$\frac{\frac{\text{Score 1}}{SE^2_{pred}} + \frac{\text{Score 2}}{SE^2_{meas}}}{\frac{1}{SE^2_{pred}} + \frac{1}{SE^2_{meas}}} \Rightarrow \frac{\frac{115}{8.43^2} + \frac{125}{3.92^2}}{\frac{1}{8.43^2} + \frac{1}{3.92^2}} = 123.5 \quad (4.4)$$

test score because the reliability coefficient at Time 2 (.94) is greater than the stability coefficient for predictions of Time 2 scores from Time 1 scores (.85). Most important, the standard error of the integrated estimate is smaller than the standard error of the Time 2 score (i.e., 3.92). Taking into account prior information (i.e., the Time 1 score), the test interpreter can revise and sharpen inferences from evidence gathered at Time 2.[7] Using Thorndike's statistical methods, the computer program "The Rev." estimates true scores from multiple test scores (and the means, standard deviations, and reliability coefficients of each test; Franklin & Allison, 1992).

Franklin and Crosby (2001) presented the application of Bayesian methods to improve the diagnostic accuracy of the neuropsychological assessment by combining data from observational, parametric, and non-parametric analysis.

Diagnostic accuracy is improved in subsequent examinations where the Posterior odds$_1$ replaces base rate of the disorder. Table 4.3 shows the effect of sequential assessment as described in Equations 4.5 - 4.7. Posterior$_1$ shows a much poorer effect (.191) than expected from a .05 significance level when the base rate (.010) is considered. Posterior$_2$ replaces Posterior$_1$ for base rate in a second testing. Because AS-NHST (see chap. 2, this volume) is used in Equations 4.5 and 4.6, probability calculations are reversed. Here, .041 reflects a much improved level of conditional probability. Equations 4.5 and 4.6 account for AS-NHST by flipping the probability ratios. Posterior$_3$ reveals after third testing that our probability has fallen to .004, a statistical finding that inspires confidence.

[7] Using the same Bayesian approach, Thorndike (1986b) showed how information from different (but correlated) tests can be integrated and how test scores may be used to predict future performance on correlated tests.

$$\text{Posterior}_1 = \frac{(1 - \text{Base Rate})}{\text{Base Rate}} * \frac{(1 - p_1)}{p_1} \qquad (4.5)$$

$$\text{Posterior}_2 = \frac{\text{Posterior Odds}_1}{(1 - \text{Posterior Odds}_1)} * \frac{p_2}{(1 - p_2)} \qquad (4.6)$$

$$\text{Posterior}_3 = \frac{\text{Posterior Odds}_2}{(1 - \text{Posterior Odds}_2)} * \frac{p_3}{(1 - p_3)} \qquad (4.7)$$

Table 4.3
Effects of Sequential Bayesian Testing

	DV	*BR/(1-BR)* *or PO/(1-PO)*	*(1-p)/p RS-H$_o$* *or p/(1-p)* *AS-H$_o$*
Symptom Base Rate	0.010		
RS-H$_o$ Test p1	0.050		
Posterior 1	***0.191***	0.010	19
AS-H$_o$ Test p2	0.850		
Posterior 2	***0.041***	0.237	5.6
AS-H$_o$ Test p3	0.910		
Posterior 3	***0.004***	0.043	10

Bayesian Networks

Bayesian networks are statistical models that evaluate relationships between sets of data, combining prior and current knowledge into one belief statement—such as diagnoses—about individual cases (e.g., Shafer, 1996). Thanks to these models "decision problems that, in the past, were hopelessly complex and unmanageable, are made both visually simple and intuitively obvious largely due to assumptions about conditional independence" (Mellers, Schwartz, & Cooke, 1998, p. 464). Networks perform three main functions. First, they represent observed phenomena. Second, they represent phenomena for which action is required without precise knowledge of the requisite data being available. Third, they integrate a priori expectations (e.g., beliefs) and outcome data (e.g., test scores) to generate lawful but fallible (i.e., probabilistic) inferences (Pearl, 1988).

In clinical settings, each new patient represents a "case" requiring a diagnostic judgment. As we have seen, a Bayesian diagnosis reflects a belief of how signs and symptoms presented by the patient correspond to signs and symptoms associated with a specific disease or disorder. The diagnosis then becomes the basis for outcome prediction, intervention selection, or additional evaluation. In Bayesian networks, clinicians serve as "experts," whose experience is captured as the initial relations between variables (often in the form of conditional probabilities). Once the initial relations are specified, Bayesian networks can be fine-tuned by adding statistical findings from prior and new cases.

Network Construction

Bayesian networks comprise two or more probabilistic variables, or *nodes*, and relations among the variables, or *links*. Nodes may contain discrete (e.g., true vs. false) or continuous (e.g., test scores) data. Links connect a source, or parent node, with a target, or child node. Parent nodes may have multiple child nodes, and child nodes may have multiple parent nodes. Links are unidirectional so that the represented relations may reflect partial or deterministic causation. Any relation between nodes can be represented in contingency tables.

Once constructed, a Bayesian network is applied to specific cases. For each known variable value, information is entered as a finding. Then, the network calculates a probabilistic inference to establish beliefs for all the other variables in the network. As is customary in Bayesian inference, the final beliefs are posterior probabilities (as opposed to the prior probabilities), which enter the network at the input level. As discussed earlier, the shift from prior to posterior probabilities reflects a revision, or updating, of belief.

The set of beliefs represented at each node reflects probabilistic inferences without changing the knowledge base (vis., original "expert" opinion) of the network. An important feature of Bayesian networks is that new findings representing "true examples" of the diagnosis can be added to the data base, thereby increasing diagnostic accuracy. Addition of new information leads to further belief revisions. An example network is presented in the Appendix.

Conclusions

The noted statistician and cognitive psychologist Ward Edwards (1998) predicted that "the 21st century [will be] the Century of Bayes" (p. 416). Several areas of psychological inquiry have been using Bayesian methods to good effect. In cognitive-experimental research, for example, Bayesian methods of hypothesis evaluation are beginning to supplement traditional significance testing (Krueger, 2001; Nickerson, 2000). Similarly, Bayes' theorem has served as a model for inductive reasoning processes that generate these data (Kahneman, Slovic, & Tversky, 1982), a development that Edwards himself pioneered (Edwards, 1961). Other academic and applied disciplines, such as medical diagnosis, forensic judgment, managerial decision making, econometrics, paternity testing, and engineering control theory, have benefited from the use of Bayesian principles as well (Press, 1989; West & Harrison, 1989).

In clinical psychology and neuropsychology, however, applications of automated Bayesian inference and the use of Bayesian belief networks lag behind the technological possibilities. Of the 36 tests marketed by American Guidance Service (AGS, 1999), for example, only 39% provide computer scoring. Slightly more than half provide narrative reports, and only two compare findings of one test with another. Although AGS is not a developer or publisher of neuropsychological tests, many of their measures are useful for the evaluation of children with neuropsychological disorders and of adults whose function is so impaired that tests standardized for "normal" adults are inappropriate.

If we consider the 53 neuropsychological tests published by the Psychological Corporation (1999), we find that only 17 provide scoring software. Only nine have the capability to produce narrative reports, and five can compare findings from one or two other tests (all of which are published in-house). Unfortunately, test publishers rarely provide interpretative software that is adequate for comparing findings from

multiple sources, even though development of user-friendly applications like *Netica* (Norsys Software, 1997) demonstrate availability and cost-effectiveness of Bayesian techniques. What is more, it is unclear which software interpretation programs, if any, include Bayesian evaluation models in their inference engines.

Edwards (1998) expressed concern that "unless psychologists learn about these new tools, they will not be able to compete in the rapidly growing market concerned with training domain experts to make the judgments they require" (p. 417). We hope that the present chapter contributes to a rising willingness among psychologists and neuropsychologists to consider these methods.

ACKNOWLEDGMENTS

We are indebted to Hal Arkes, Robyn Dawes, and Roy Poses for their generous feedback on a draft version of this chapter.

REFERENCES

American Guidance Service. (1999). *Clinical catalog.* Circle Pines, MN: Author.

Arkes, H. R., Wortmann, R. L., Saville, P. D., & Harkness, A. R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology, 66*, 252-254.

Dascells, W., Schoenberger, A., & Graybors, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299*, 999-1001.

Bayes, T. (1763). An essay toward solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society, 53*, 370-418.

Birnbaum., M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96*, 85-94.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

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

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: The Free Press.

Dawes, R. M. (2000). Proper and improper linear models. In T. Connolly, H. Arkes & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (2^{nd} ed., pp. 378-394). New York: Cambridge University Press.

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

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249-267). Cambridge, England: Cambridge University Press.

Eddy, D. M. (1984). Variations in physician practice: The role of uncertainty. *Health Affairs, 3,* 74-89.

Eddy, D. M., & Clanton, C. H. (1982). The art of diagnosis: Solving the clinicopathological exercise. *The New England Journal of Medicine, 306,* 1263-1268.

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology, 12,* 473-498.

Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist, 53,* 416-428.

Elwood, R. W. (1993). Clinical discriminations and neuropsychological tests: An appeal to Bayes' theorem. *The Clinical Neuropsychologist, 7,* 224-233.

Evans, D. A., Funkenstein, H. H., Albert, M. S., Schaer, P. A., Cook, N. R., Chowan, M. J., Hebert, L. E., Hennekens, C. H., & Taylor, J. O. (1989). Prevalence of Alzheimer's disease in a community population of older persons. *Journal of the American Medical Association, 262,* 2551-2556.

Faust, D., & Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science, 241,* 31-35.

Feinstein, A. R. (1978). On the sensitivity, specificity, and discrimination of diagnostic tests. *Clinical Biostatistics, 17,* 104-116.

Franklin, R. D., & Allison, D. B. (1992). The Rev.: An IBM BASIC program for Bayesian test interpretation. *Behavior Research Methods, Instruments, & Computers, 24,* 491-492.

Franklin, R. D., & Crosby, F. X. (2001). Early stopping rules in forensic neuropsychological evaluations. *Journal of the International Neuropsychological Society,* 7(4), 416.

Garb, H. N., & Schramke, C. J. (1996). Judgment research and neurological assessment: A narrative review and meta-analysis. *Psychological Bulletin, 120,* 140-153.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Gigerenzer, G., Hoffrage, U., & Ebert, A. (1998). AIDS counselling for low-risk clients. *AIDS Care, 10,* 197-211.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *American Statistical Association Journal, 49,* 732-764.

Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290,* 2261-2262.

Jones, W. P. (1989). A proposal for the use of Bayesian probabilities in neuropsychological assesement. *Neuropsychology, 3,* 17-22.

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

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19,* 1-53.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16-26.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15,* 22-25.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, and cutting scores. *Psychological Bulletin, 52,* 194-216.

Mellers, B. A., & McGraw, P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review, 106,* 417-424.

Mellers, B. A., Schwartz, A., & Cooke, A. D. J. (1998). Judgment and decision making. *Annual Review of Psychology, 49,* 447-477.

Nickerson, R. S (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Norsys Software Corp. (1997). *Netica manual.* Vancouver,BC, Canada: Author.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufman.

Press, S. J. (1989). *Bayesian statistics: Principles, models, and applications.* New York: Wiley.

The Psychological Corporation (1999). *The catalog for neuropsychological assessment & intervention resources.* San Antonio, TX: Author.

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluation the efficacy of diagnostic tests. *The New England Journal of Medicine, 299,* 926-930.

Satz, P., Fennell, E., & Reilly, C. (1970). Predictive validity of six neuropsychological tests: A decision theory analysis. *Journal of Consulting and Clinical Psychology, 34*, 375-381.

Sedlmeier, P. (1997). BasicBayes: a tutor system for simple Bayesian inference. *Behavior Research Methods, Instrumentats and Computers, 27,* 328-336.

Shafer, G. (1996). *Probabilistic expert systems.* Philadelphia: SIAM.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34-50.

Solomon, P. R., Hirschoff, A., Kelly, B., Relin, M., Brush, M., DeVeaux, R. D., & Pendlebury, W. W. (1999). A 7 minute neurocognitive screening battery highly sensitive to Alzheimer's Disease. *Archives of Neurology, 55,* 349-355.

Thorndike, R. L. (1986a). Bayesian concepts and test interpretation. *Journal of counseling and Development, 65,* 170-172.

Thorndike, R. L. (1986b). The role of Bayesian concepts in test development and test interpretation. *Journal of Counseling and Development, 65,* 54-56.

West, M., & Harrison, J. (1989). *Bayesian forecasting and dynamic models.* New York, NY: Springer-Verlag.

Winkler, R. L. (1993). Bayesian statistics: An overview. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical Issues* (pp. 210-232). Hillsdale, NJ: Lawrence Erlbaum Associates.

APPENDIX

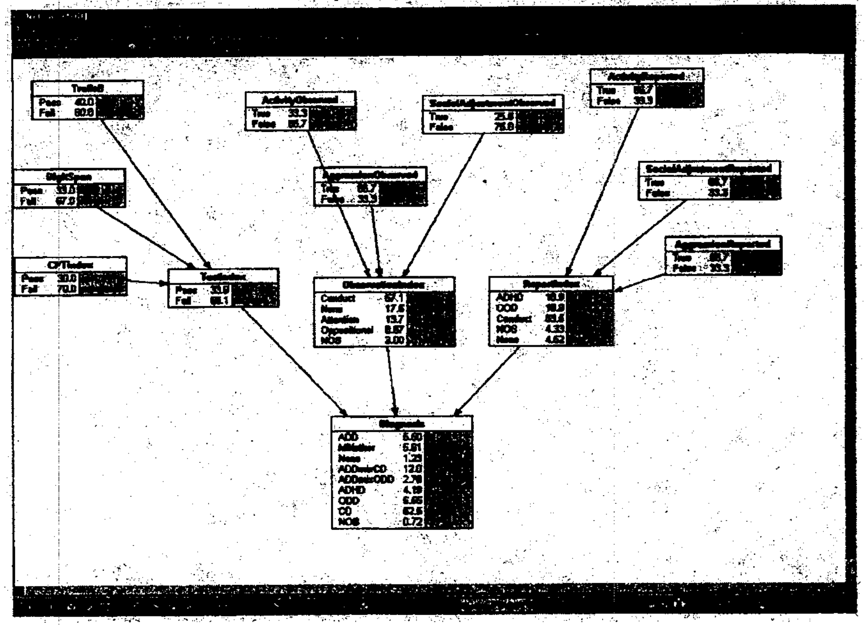
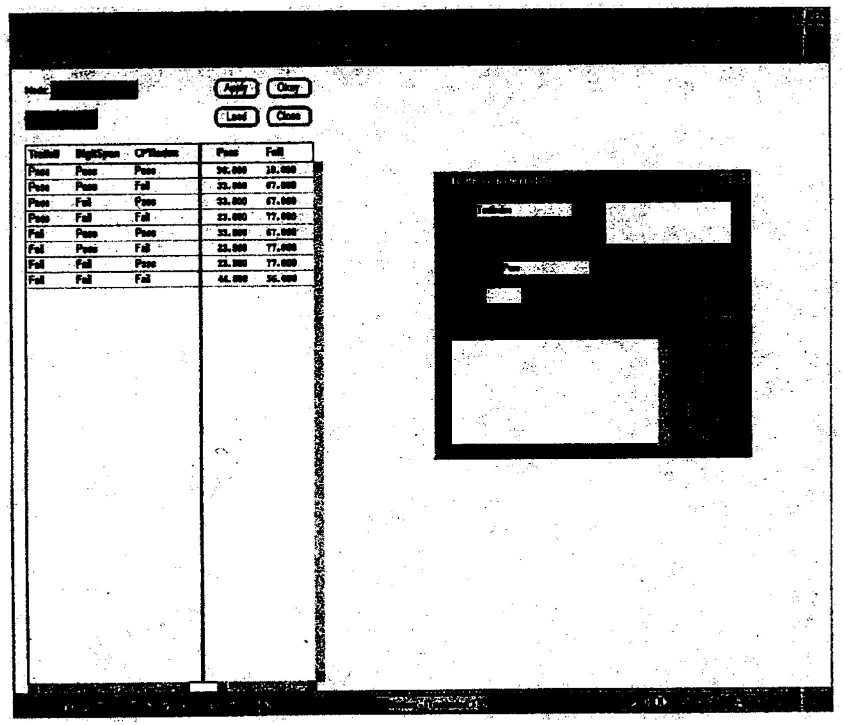
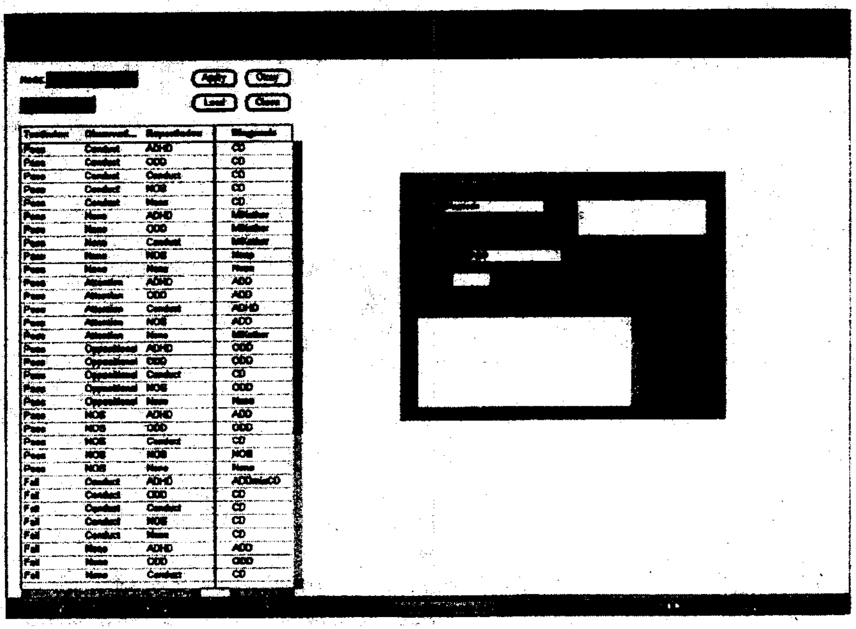


FIG. A.1. Bayes network for evaluating disorders of conduct.



TrailsB	DigitSpan	CPTIndex	Pass	Fail
Pass	Pass	Pass	90.000	10.000
Pass	Pass	Fail	33.000	67.000
Pass	Fail	Pass	33.000	67.000
Pass	Fail	Fail	23.000	77.000
Fail	Pass	Pass	33.000	67.000
Fail	Pass	Fail	23.000	77.000
Fail	Fail	Pass	23.000	77.000
Fail	Fail	Fail	44.000	56.000

FIG. A.2 Bayes network contingency table for node TestIndex.



TestIndex	Observed...	ReportIndex	Diagnosis
Pass	Conduct	ADHD	CD
Pass	Conduct	ODD	CD
Pass	Conduct	Conduct	CD
Pass	Conduct	NOS	CD
Pass	Conduct	None	CD
Pass	None	ADHD	NIOther
Pass	None	ODD	NIOther
Pass	None	Conduct	NIOther
Pass	None	NOS	None
Pass	None	None	None
Pass	Attention	ADHD	ADD
Pass	Attention	ODD	ADD
Pass	Attention	Conduct	ADHD
Pass	Attention	NOS	ADD
Pass	Attention	None	NIOther
Pass	Oppositional	ADHD	ODD
Pass	Oppositional	ODD	ODD
Pass	Oppositional	Conduct	CD
Pass	Oppositional	NOS	ODD
Pass	Oppositional	None	None
Pass	NOS	ADHD	ADD
Pass	NOS	ODD	ODD
Pass	NOS	Conduct	CD
Pass	NOS	NOS	NOS
Pass	NOS	None	None
Fail	Conduct	ADHD	ADDmixCD
Fail	Conduct	ODD	CD
Fail	Conduct	Conduct	CD
Fail	Conduct	NOS	CD
Fail	Conduct	None	CD
Fail	None	ADHD	ADD
Fail	None	ODD	ODD
Fail	None	Conduct	CD

FIG. A.3. Bayes network contingency table for node Diagnosis.