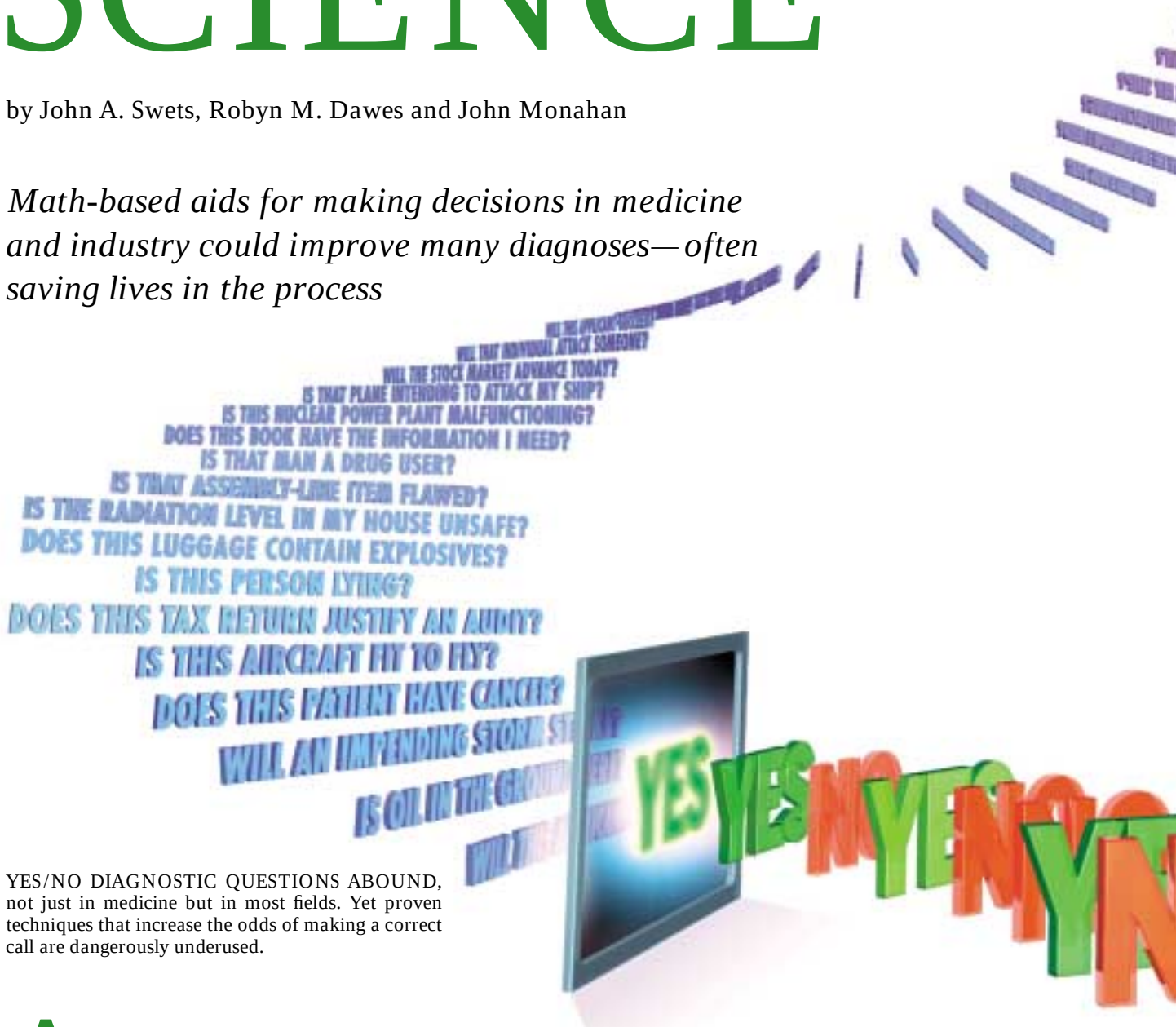
Better DECISIONS through SCIENCE

by John A. Swets, Robyn M. Dawes and John Monahan

Math-based aids for making decisions in medicine and industry could improve many diagnoses—often saving lives in the process



YES/NO DIAGNOSTIC QUESTIONS ABOUND, not just in medicine but in most fields. Yet proven techniques that increase the odds of making a correct call are dangerously underused.

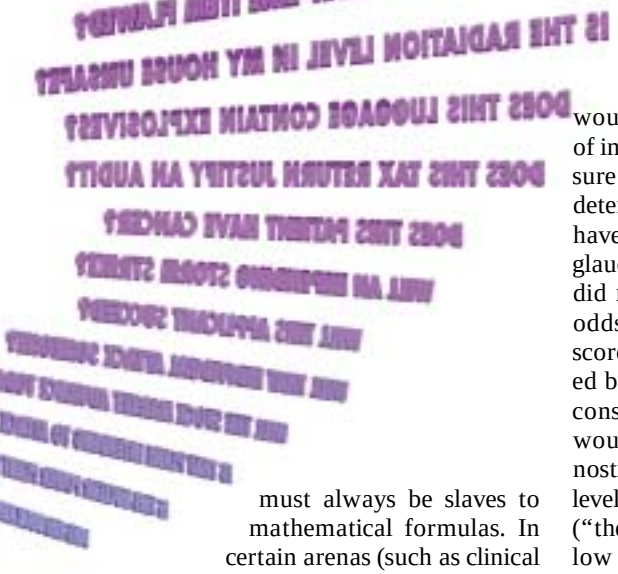
SLIM FILMS

A physician stares at a breast x-ray, agonizing over whether an ambiguous spot is a tumor. A parole board weighs the release of a potentially violent criminal. A technician at an airport worries over a set of ultrasound readings: do they suggest a deadly crack in an airplane's wing?

All these people are grappling with diagnostic decisions. In spite of incomplete or ambiguous evidence, they must determine whether or not a certain condition exists (or will occur). Such problems abound in health care, public safety, business, environment, justice, education, manufacturing, information processing, the military and government. And the stakes can be high. In many cases, a wrong verdict means that people will die.

Perhaps surprisingly, the diagnostic decision-making process turns out to be essentially the same across fields. Hence, methods that improve the process in one industry can usually serve in others. At least two such methods are already available. Sadly, though, they remain unknown or unused in many realms. One increases accuracy, enhancing the odds that any given decision will be the correct one. The other improves the "utility" of a decision-making approach, ensuring that the number of true cases found does not come at the cost of an unreasonable number of false positive diagnoses ("false alarms"). These methods are statistical, but math phobics have nothing to fear; the basic logic is easy to grasp.

No one is saying that diagnosticians



must always be slaves to mathematical formulas. In certain arenas (such as clinical medicine and weather forecasting), objective tools may function best as "second opinions" that inform a reviewer's decisions but do not have the final word. In other fields, however, statistical analyses have frequently been found to be more accurate than subjective judgments, even those made by highly experienced professionals.

We focus in this article on diagnoses that hinge on a choice between just two alternatives—yes or no (Is a tumor present? Is an airplane wing defective?). Certainly the world is full of problems involving a wider range of options, but serious yes/no decisions are prevalent.

Tools of the Trade

If diagnostic tests always produced straightforward answers, no one would need statistical decision-making tools. In reality, though, the raw results of diagnostic tests usually have to be interpreted. In a simple example, the fluid pressure in the eye is measured to detect whether a person has glaucoma, which robs vision by damaging the optic nerve and other parts of the eye. A very low score clearly means the eye is healthy, and a high score signifies glaucoma. But scores in between are ambiguous, unable to indicate which patients have the condition and which do not.

Statistics can clear some of that fog. For argument's sake, assume that pressure is the only diagnostic measure available for glaucoma. Assume, too, that pressures below 10 on the standard measuring scale always signify good health, pressures over 40 always signify disease, and readings between 10 and 40 can occur in affected as well as healthy eyes.

To cope with this ambiguity, analysts would first identify a large population of individuals whose scores on the pressure test were known. Then they would determine which people went on to have vision problems characteristic of glaucoma within a set period and which did not. And they would calculate the odds that people having each possible score will have glaucoma. Finally, guided by those probabilities (and by other considerations we will discuss), they would set a rational cut point, or diagnostic threshold: scores at or above that level would yield a positive diagnosis ("the patient has glaucoma"); scores below would yield a negative diagnosis ("the patient does not have glaucoma").

Of course, single diagnostic tests may not be as informative as a combination. To enhance the accuracy of a diagnosis, analysts can combine data from many tests that each provide unique information, giving greater weight to measurements that are most predictive of the condition under study. The mathematical algorithms that specify the best tests to include in a diagnostic workup and that calculate the likelihood, based on the combined results, that a condition is present are known as statistical prediction rules (SPRs).

Totally objective data, such as pressure readings, are not the only features that can be incorporated to enhance the accuracy of statistical prediction rules; subjective impressions can be quantified and included as well. They can be objectified, for instance, by making an explicit list of perceptual criteria (such as the size and irregularity of a possibly malignant mole) that can be rated according to a scale, perhaps from one to five.

If more than one statistical prediction rule is available, decision makers have to determine which ones are most accurate. This challenge, too, can be met objectively. The overall accuracy of prediction rules can be evaluated by reviewing what are called ROC (receiver operating characteristic) curves. Such curves were first applied to assess how well radar equipment in World War II distinguished random interference (noise) from signals truly indicative of enemy planes.

Programs that generate ROC curves consider what will happen if a particular raw score on a diagnostic test (or set of tests) is selected as the diagnostic threshold for a yes/no decision. What percent of individuals who truly have the condition in question will correctly be deemed to have it (true positive decisions, or "hits")? And what percent of individuals free of the condition will mistakenly be deemed to have it (false positive decisions, or false alarms)?

Then, for each threshold, the programs plot the percentage of true positives against the percentage of false positives. The result is a bowed curve, rising from the lower left corner, where both percentages are zero, to the upper right corner, where both are 100. The more sharply the curve bends, the greater the accuracy of the rule, because the number of hits relative to the number of false alarms is higher.

Obviously, true positives and false positives are not the only outcomes possible. A yes/no diagnosis based on any particular threshold will also generate true negatives (individuals are correctly deemed to be free of the condition being evaluated) and false negatives, or "misses" (individuals are incorrectly deemed to be free of the condition). But these results are the exact complements of the others and thus can be ignored when constructing ROC curves. A true positive rate of 80 percent, for instance, automatically means that the miss rate is 20 percent.

Given that few diagnostic methods are perfect at sorting individuals who have a condition from individuals who do not, institutions have to decide how important it is to find all or most true positives—because more true positives come at the cost of more false alarms. That is, they need to set a threshold that makes good sense for their particular situation.

Returning to our glaucoma example, clinicians who looked only at pressure could find virtually every case of glaucoma if they chose a very "lenient" diagnostic cutoff—say, a score of 10. After all, the test sample revealed that virtually everyone with glaucoma has a score above that level. Yet that cutoff would result in many healthy people being told they were ill; those people would then be subjected unnecessarily to both worry and treatment. To minimize such errors, clinicians could instead set a rather strict diagnostic threshold—an eye pressure of 35, perhaps; very few healthy people in the sample had pressures that high. But this strict criterion would miss more than half of all affected individuals, denying them treatment.

In setting a threshold, decision makers weigh such issues as the consequences of misses and false alarms and

the prevalence of the problem under consideration in the population being tested. Fortunately, some rules of thumb and mathematical aids for finding the optimal cutoff point have been developed. For instance, a high prevalence of a problem in a population or a large benefit associated with finding true cases generally argues for a lenient threshold; conversely, a low prevalence or a high cost for false alarms generally calls for a strict threshold.

Rules Come to Life

Although statistical prediction rules and ROC curves are often sorely underused by diagnosticians, real-life examples of their value abound. One of the most dramatic illustrations comes from psychiatry.

Increasingly, psychiatrists and clinical psychologists are asked to determine whether incarcerated or disturbed individuals are likely to become violent. People who seem most likely to endanger others need to be identified and treated for their own good and for others' safety. At the same time, interfering in the lives of people who do not need care is unacceptable.

Disconcertingly, in 1993 the most sophisticated study of clinicians' unaided assessments uncovered a startling lack of accuracy. Clinicians who diagnosed consecutive patients coming to the emergency department of a metropolitan psychiatric hospital proved no more accurate than chance at predicting which female patients would commit violence in the community within the next six months. Their success rate with male patients was only modestly better.

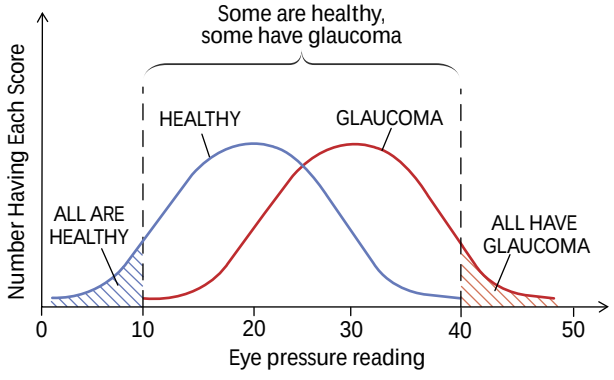
In response to such findings, a number of statistical prediction rules were developed for assessing the probability of violence. One of the most studied is the Violence Risk Appraisal Guide (VRAG), which measures 12 variables, among them scores on a checklist of features indicative of psychopathy and assessments of maladjustment in elementary school.

In a test of the rule's ability to predict whether criminals being discharged from a maximum-security hospital would commit violent acts over the next several years, the VRAG divided the subjects into two categories of risk: "high" and "low." Fifty-five percent of the high-risk group but only 19 percent of the low committed a new violent offense—an accuracy level well above that of chance. And a newer statistical prediction rule proved to be even better at forecasting violence in noncriminals about to be discharged from psychiatric facilities. Nevertheless, interested parties continue to disagree over whether

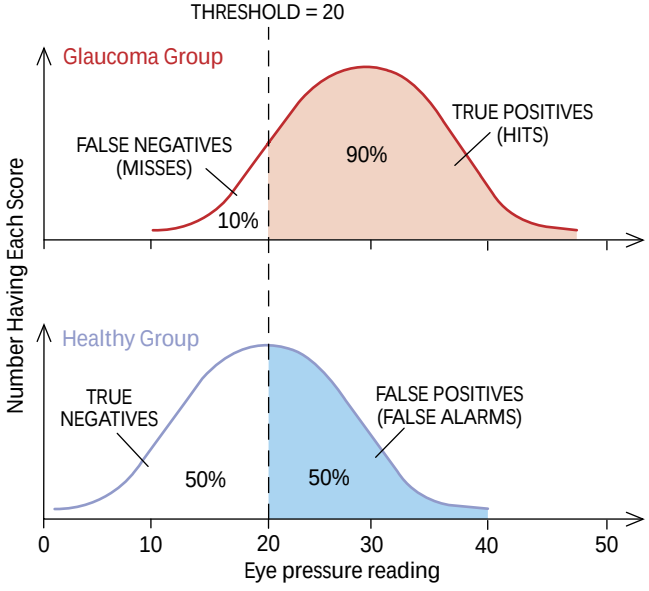
Better Decision Making, Step by Step

How can decision makers ensure that the diagnostic tests they use are as accurate as possible, doing the best job of distinguishing individuals who have a condition from those who do not? A major way involves constructing so-called ROC (receiver operating characteristic) curves. This approach is best described by example. Imagine the steps an analyst might take to evaluate how well glaucoma is diagnosed by measuring the fluid pressure in patients' eyes.

STEP 1 Find a large sample population of people whose eye pressure level and glaucoma status are known. Separate those who are healthy from those with glaucoma, and plot the number of individuals with each pressure level. The graph for this hypothetical population reveals that pressure readings in the 10 to 40 range cannot conclusively distinguish healthy people from those with glaucoma.



STEP 2 Calculate the probability that a "yes" diagnosis at or above any given score, or threshold, would be correct for a new patient. Find such probabilities by determining the fraction of patients in the sample population who would have been properly diagnosed if that threshold were applied. Below, the area under the curves represents 100 percent of each population. If the threshold were 20, 90 percent of people who truly had glaucoma would be diagnosed correctly (true positives), and 50 percent of healthy people would be incorrectly diagnosed as having the condition (false positives).



JENNIFER JOHANSEN

clinicians should treat such rules as advisory or make decisions based solely on the statistics.

Better Cancer Diagnoses

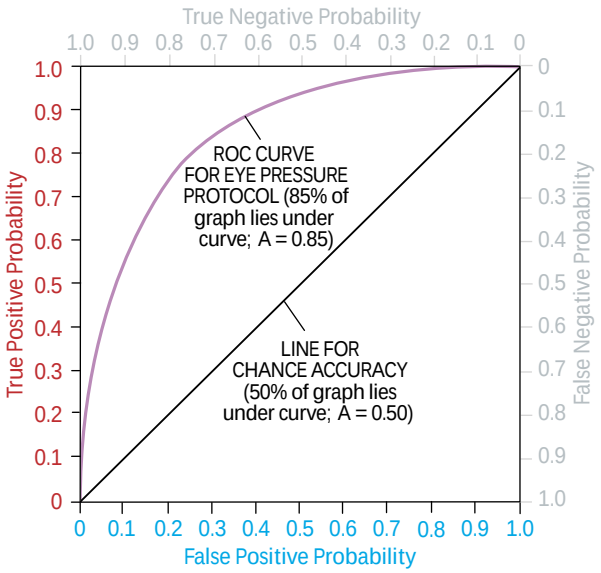
Statistical prediction rules have also had impressive success in studies aimed at helping radiologists diagnose breast cancer. In one such investigation, radiologists in community hospitals evaluated mammograms in their usual, subjective way. Months later they examined the same mammograms according to a checklist of perceptual features (such as how fuzzy the borders of a mass seem to be) developed by radiologists who specialize in reviewing mammograms. Then a statistical prediction rule converted the ratings into probability assessments indicating the likelihood for each patient that breast cancer was present. The radiologists reviewed these probabilities but ultimately made their own judgments. The extra data helped considerably. General radiologists who took the statistical data into account became more accurate, reaching the precision of specialists who had used the checklist.

Physicians who treat prostate cancer are already making extensive use of statistical prediction rules. One rule in particular is getting a serious workout. Once a man is "clinically" deemed to have cancer of the prostate gland (on the basis of a checkup, a simple needle biopsy and noninvasive tests), the question of the best treatment arises [see "Combating Prostate Cancer," by Marc B. Garnick and William R. Fair; Scientific American, December 1998]. Neither surgery to remove the affected gland nor radiation focused tightly on it (to limit side effects) will eliminate the tumor if it has grown beyond the gland or has spread to other parts of the body. Hence, physicians strive to determine the status of the tumor before any treatment is attempted. Unfortunately, a great many tumors that initially seem to be confined to the prostate later turn out to have been more advanced.

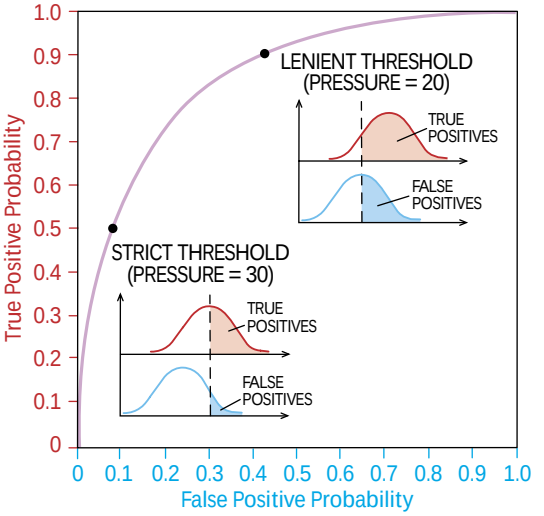
For years, doctors had few good ways of predicting which patients truly had confined disease and which did not. More recently, however, doctors and patients have been able to gain a clearer picture by consulting probability tables published in the May 14, 1997, issue of the *Journal of the American Medical Association*.

The researchers who created the tables knew that three assessments each had independent predictive value: the tumor's "clinical stage" (a determination, based on noninvasive tests, of tumor size and spread), the level in the blood of a specific protein (PSA, or prostate-specific anti-

STEP 3 Construct an ROC curve by plotting, for each potential threshold, the rate of true positives against the rate of false positives. A straight line would signify that the diagnostic test had 50/50 odds of making a correct diagnosis (no better than flipping a coin). As curves bow more to the left, they indicate greater accuracy (a higher ratio of true positives to false positives). Accuracy (A) is indexed more precisely by the amount of area under the curve, which increases as the curves bend. Our glaucoma protocol is moderately accurate.



STEP 4 If the accuracy is acceptable, select a threshold for yes/no diagnoses. Choose a threshold that yields a good rate of true positives without generating an unacceptable rate of false positives. Each point on the curve represents a specific threshold, moving from the most strict at the bottom left to the most lenient at the top right. Strict thresholds (*bottom inset*) limit false positives at the cost of missing many affected individuals; lenient thresholds (*top inset*) maximize discovery of affected individuals at a cost of many false positives. Which threshold is optimal for a given population depends on such factors as the seriousness of the condition being diagnosed, the prevalence of the condition in a population, the availability of corrective measures for those who are diagnosed, and the financial, emotional and other costs of false alarms.



gen) and the Gleason score (an indicator of tumor aggressiveness, based on microscopic analyses of a biopsy sample). The investigators therefore developed a statistical prediction rule that looked at virtually every combination of results for these three variables and calculated the odds that the initial diagnosis of "no spread" would be correct. Then they listed the probabilities in a user-friendly, tabular form.

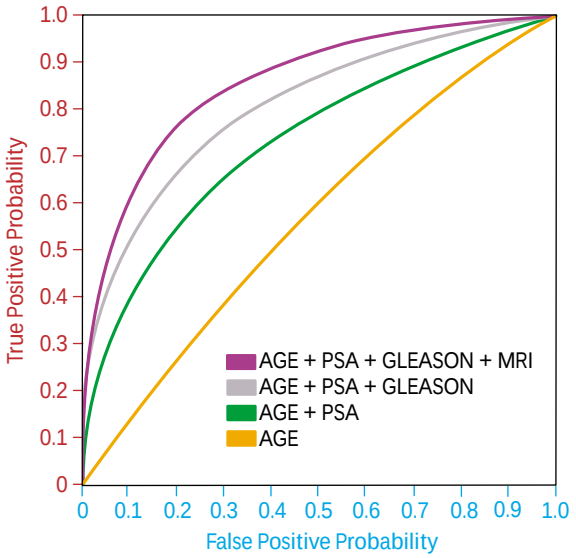
Good Chance of Rain

It would be a mistake to think that only medical practitioners use statistical prediction rules. In fact, meteorologists adopted the tools for weather forecasting more then 25 years ago.

The National Weather Service routinely feeds weather-related data into statistical programs designed to estimate the likelihood that tornadoes, hurricanes, heavy rains and other hazards will arise in different parts of the nation. The weather service then conveys these objective predictions to meteorologists in local areas, who modify the predictions in light of new information or of factors they think the computer programs did not address adequately.

Other groups have embraced the techniques as well—among them, graduate admissions committees at universities. In a typical example, a committee will project first-year grades from two variables—undergraduate grades and graduate school aptitude exams, on the assumption that students scoring above some preselected high level should generally be admitted and those scoring below a specified lower level should generally be rejected. Then the committee will more subjectively evaluate the credentials of applicants who have not been admitted or rejected by the school's statistical prediction rule.

One law school objectively rates two variables that were formerly assessed subjectively: the quality of the student's undergraduate institution and the extent of grade inflation at that institution. Along with the student's grade point average and scores on the aptitude exam required for law school, it considers the mean exam score of all students from the applicant's college who took the test and the mean grade point average of students from that college who applied to law school. The revised formula predicts first-year law-school grades significantly better than the two-variable scheme.



ROC CURVES compared the accuracy achievable by measuring one or more of the following variables to assess whether prostate cancer is advanced: patient age, blood level of PSA (prostate-specific antigen), tumor aggressiveness (represented by the Gleason score) and tumor appearance as judged by magnetic resonance imaging (MRI). The scheme that included all four variables (*top curve*) proved best.

Thorny Thresholds

So far we have highlighted success stories. But the merit of statistical analyses may be best illustrated by examples of failure to apply them for setting rational diagnostic thresholds—such as for tests that detect the human immunodeficiency virus (HIV), the cause of AIDS.

HIV screening relies initially on a relatively simple test that detects the presence of anti-HIV antibodies, molecules produced when the immune system begins to react against HIV. Sometimes these antibodies arise for reasons other than the presence of HIV, however. Hence, if the outcome (based on some antibody threshold) is positive, laboratories will run a different, more sophisticated test. This two-test requirement is meant to help limit false positives. The antibody tests are particularly problematic in that, illogically, the several approved tests differ in their accuracies and thresholds. Varied thresholds would make sense if each test were aimed at a distinct population, but that is not the case.

The thresholds are disturbing in another way as well. They were originally set to distinguish clean from tainted donated blood; then they were left unchanged when they were enlisted to identify people infected with the virus. Throwing out a pint of uncontaminated blood because of a false positive is a cheap mistake; sending an alarmed, uninfected person for further HIV testing is not. Worse still, the original thresholds have been applied mindlessly to low-risk blood donors, high-risk donors, military recruits and methadone-clinic visitors—groups whose infection rates vary over an enormous range. For the high-risk groups, the threshold should be set more leniently than for the low-risk populations (to maximize discovery), even if the price is a higher rate of false positives.

Recent years have seen the introduction of confirmatory tests that are more accurate and of HIV therapies that prolong life and health. Consequently, false positive diagnoses are rare these days, and people who are infected with HIV benefit much more from being diagnosed than was true in the past. These advances mean that the diagnostic problem has shifted from whom to call positive to whom to test.

The time has come for doctors to lower their thresholds for deciding when to test; they should not be waiting until patients show up with obvious symptoms of infection. We would even argue that almost every adult should be screened and that federal agencies should take the lead in encouraging such testing.

Objective methods for establishing thresholds are also being dangerously underused in parts of the aerospace industry. This industry must constantly diagnose conditions that are serious but arise relatively infrequently, among them cracked wings and life-threatening hazards during flights. The costs of missing a cracked wing are large and obvious: many passengers may die if the plane crashes. On the other hand, a false-positive decision takes a plane out of service unnecessarily, potentially causing inconvenience and lost income. At first blush, the benefits and costs point toward a lenient threshold, favoring lives over dollars. Yet such cracks occur rarely; therefore a lenient threshold yields an unworkable number of false posi-

JENNIFER JOHANSEN

tives. Unfortunately, no one has yet tackled this issue with the available statistical techniques.

Purchasers of cockpit alarms (such as airlines and the military) have similarly failed to come to grips with how best to set decision thresholds. Alarms go off in flight under many circumstances—when sensing devices determine that another plane is too close, that the plane is getting too near to the ground, that an engine is dying or that wind shear is threatening the landing area. But they cry wolf too often, largely because the sensors are only moderately accurate and because the thresholds set for them are rather lenient. Pilots are reluctant to act on the warnings unnecessarily, because doing so can be quite disruptive. This situation has raised fears that the high number of false alarms will cause pilots to ignore or respond slowly to a real emergency. To date, though, no one has forced manufacturers to consider the false positive rate when they establish alarm thresholds.

A Plea

Clearly, statistical prediction rules can often raise the accuracy of repetitive diagnostic decisions, and formulas for setting decision thresholds can improve the utility of those decisions. But these tools provide other advantages as well. By standardizing the features that are assessed to make a diagnosis, the prediction rules can hasten the speed with which professionals recognize key diagnostic features. They also give decision makers a way to communicate more easily and precisely about impressionistic features. And they can help teach newcomers to a field.

Yet they are often met with resistance, especially if they are seen as replacing or degrading clinicians. Further, diagnosticians want to feel that they understand their own diagnoses and recommendations and that they can give a narrative of their thought processes. The results of a statistical prediction rule may be hard to include in such an account, particularly if the logic behind the analysis is not self-evident.

We understand all these concerns. Nevertheless, the benefits that statistical tools provide surely justify consideration by decision makers who hold others' lives and futures in their hands. SA

A QUESTION OF TASTE

Architects and connoisseurs of wine have invented two of the more offbeat applications of statistical prediction rules. The architectural rule applies to opera houses and was developed by having conductors rate the overall sound quality of 23 facilities. The conductors favored the houses in Buenos Aires, Dresden, Milan and Tokyo. Next, acoustical engineers physically measured several individual acoustical properties in each of the 23 buildings—such as the time delays between directly received and reflected sound, and the diffusion of sound waves caused by irregularities in walls and ceilings. Statistical analyses then revealed which properties combined to give the favored opera houses their exceptional sound and which of the acoustic characteristics were most important. The resulting rule can now guide the construction of future facilities.

COURTESY OF TAK ARCHITECTS, YANAGISAWA, TOKYO

OPERA HOUSE in Tokyo's New National Theater has stellar acoustics that can be emulated, thanks to a statistical prediction rule.

The wine rule predicts the eventual quality of red Bordeaux wines (as measured by auction price) when they are still young and undrinkable. Classically, experts have attempted to predict later quality "clinically," by smelling and tasting the new product. But about 10 years ago, researchers noted that years marked by a dry August and September and a warm growing season yield excellent wines if those years also follow a wet winter.

They then formulated the "Bordeaux equation," a statistical prediction rule that combines weather conditions and years of aging to predict the probability that wine quality will be great years ahead. That equation works quite well, accounting for 83 percent of the variance in price of mature Bordeaux red wines at auction. But it has not met with universal acclaim. "Somewhere between violent and hysterical" is how the reaction of the wine-tasting industry was described in a newspaper report soon after the equation was unveiled. *—J.A.S., R.M.D. and J.M.*

The Authors

JOHN A. SWETS, ROBYN M. DAWES and JOHN MONAHAN recently collaborated on a more technical paper on this topic published in the inaugural issue of a journal, from the American Psychological Society, that reviews psychological research on pressing issues of broad importance (see "Further Information"). Swets is chief scientist emeritus at BBN Technologies in Cambridge, Mass., senior research associate in radiology at the Brigham and Women's Hospital in Boston and lecturer on health care policy at Harvard Medical School. Dawes is the Charles J. Queenan, Jr., University Professor in the department of social and decision sciences at Carnegie Mellon University and author of *Rational Choice in an Uncertain World*. Monahan, a psychologist, holds the Doherty Chair in law at the University of Virginia and is director of the Research Network on Mental Health and the Law of the John D. and Catherine T. MacArthur Foundation.

Further Information

Combination of Prostate-Specific Antigen, Clinical Stage, and Gleason Score to Predict Pathological Stage of Localized Prostate Cancer. A. W. Partin et al. in *Journal of the American Medical Association*, Vol. 277, No. 18, pages 1445–1451; May 14, 1997.

Think HIV: Why Physicians Should Lower their Threshold for HIV Testing. Kenneth A. Freedberg and Jeffrey H. Samet in *Archives of Internal Medicine*, Vol. 159, No. 17, pages 1994–2003; September 27, 1999.

Psychological Science Can Improve Diagnostic Decisions. John A. Swets, Robyn M. Dawes and John Monahan in *Psychological Science in the Public Interest* (supplement to *Psychological Science*), Vol. 1, No. 1, pages 1–26; May 2000.