Probability

Meaning

Long Run Proportion
Estimate of (Un)certainty
Amount prepared to bet

Use

Describe likely behaviour of data
Communicate (un)certainty
Measure how far data are from some hypothesized model

How Arrived At

Subjectively

Intuition, Informal calculation, consensus

Empirically

Experience (actuarial, ...)

Pure Thought

Elementary Statistical Principles

If necessary, breaking Complex outcomes into simpler ones

Advanced Statistical Theory

calculus e.g. Gauss' Law of Errors

References

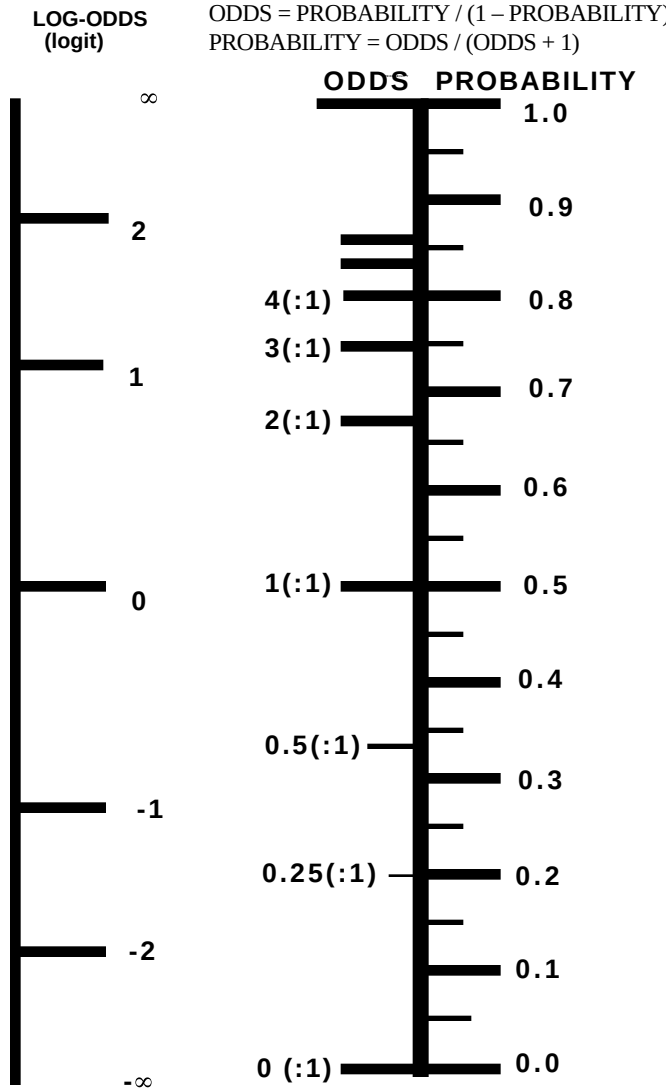
• WMS5, Chapter 2 • Moore & McCabe Chapter 4 •Colton, Ch 3
• Freedman et al. Chapters 13,14,15 •Armitage and Berry, Ch 2
• Kong A, Barnett O, Mosteller F, and Youtz C. "How Medical Professionals Evaluate Expressions of Probability" NEJM 315: 740-744, 1986 ... ***on reserve***

• Death and Taxes • Rain tomorrow • Cancer in your lifetime • Win lottery in single try • Win lottery twice • Get back 11/20 pilot questionnaires • Treat 14 patients get 0 successes • Duplicate Birthdays • Canada will use $US before the year 2010

- **OJ murdered his wife**
- **DNA matched**
- **OJ murdered wife | DNA matched**

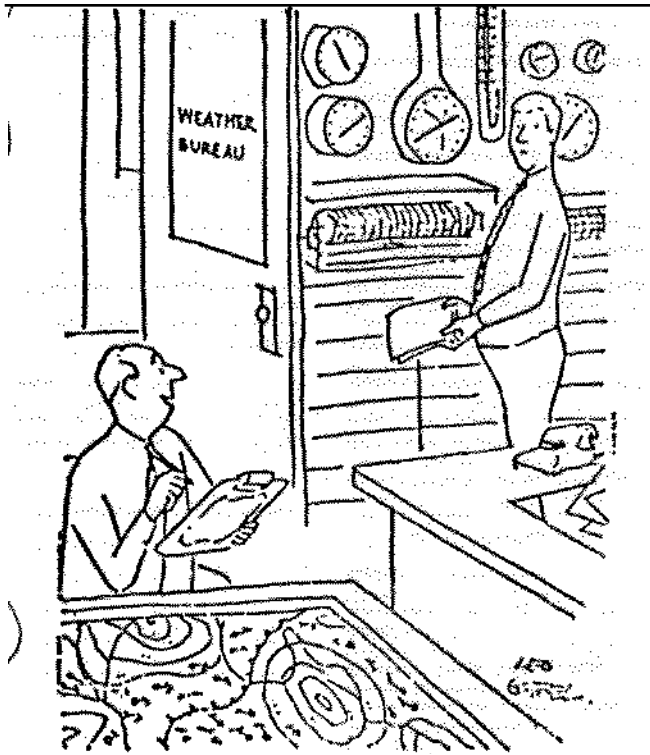
" | " is shorthand for "given that.."

Probability Scales



- **50 year old has colon ca**
- **50 year old with +ve haemoccult test has colon ca**
- **child is Group A Strep B positive**
- **8 yr old with fever & v. inflamed nodes is Gp A Strep B positive**
- **There is life on Mars**

How to calculate probabilities

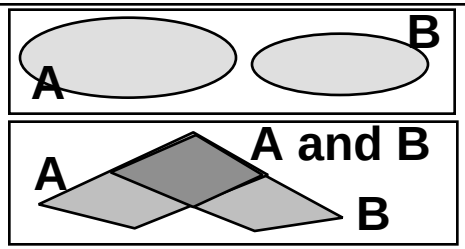


Wall Street Journal

"I figure there's a 40% chance of showers, and a 10% chance we know what we're talking about"

Probability Calculations

Basic Rules



Probabilities add to 1

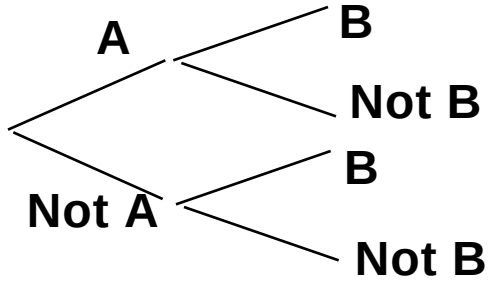
Prob(event) =
1 - Prob(complement)

"PARALLEL"

ADDITION FOR "EITHER A OR B"

If mutually exclusive
P(A or B) = P(A) + P(B)

If overlapping
P(A or B) = P(A) + P(B) - P(A and B)



"SERIAL"

MULTIPLICATION FOR "A AND B" OR "A THEN B"

If independent
P(A and B) = P(A) • P(B)

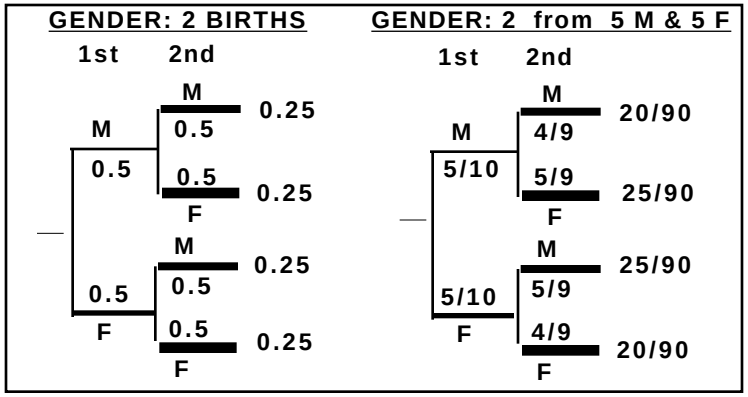
If dependent
P(A and B) = P(A) • P(B | A)

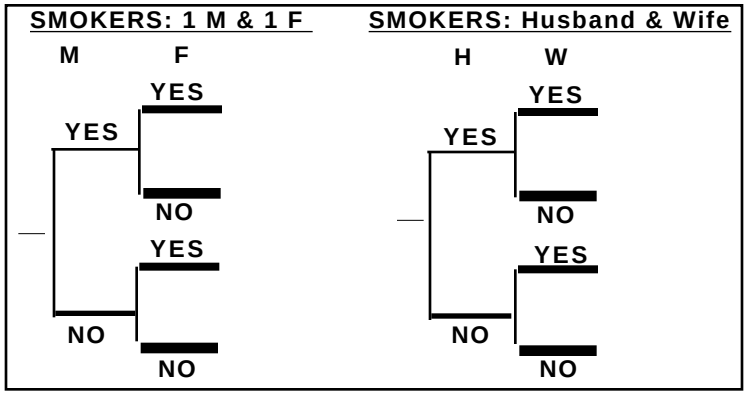
Conditional Probability P(B | A) = Probability of B "**given** A" or "conditional on A"

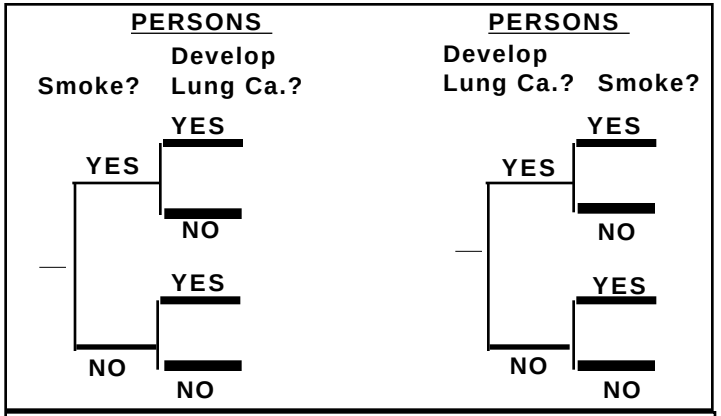
More Complex:

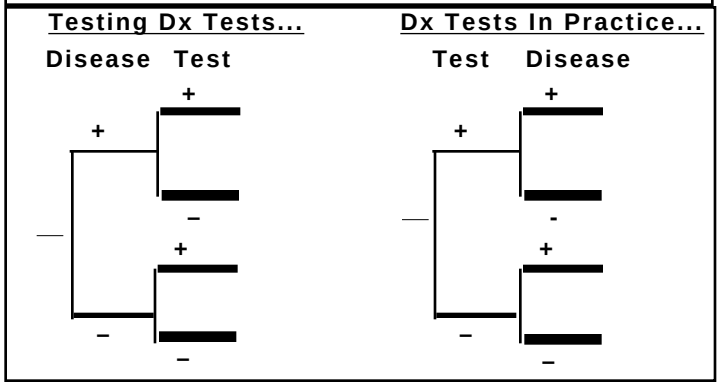
- Break up into elements
- Look for already worked-out calculations
- Beware of intuition, especially with "after the fact" calculations for non-standard situations

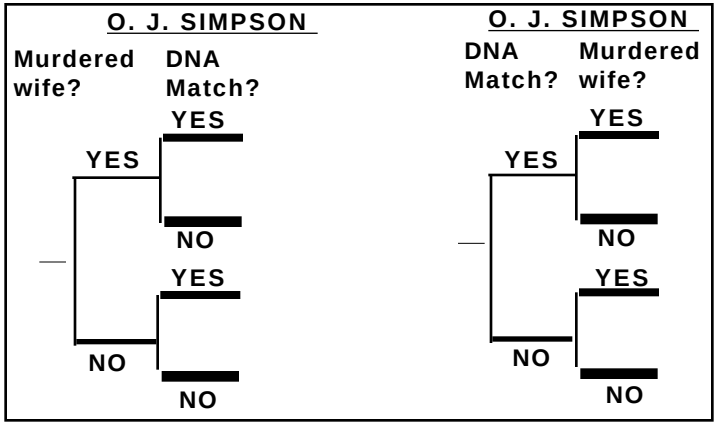
Examples of Conditional Probabilities...



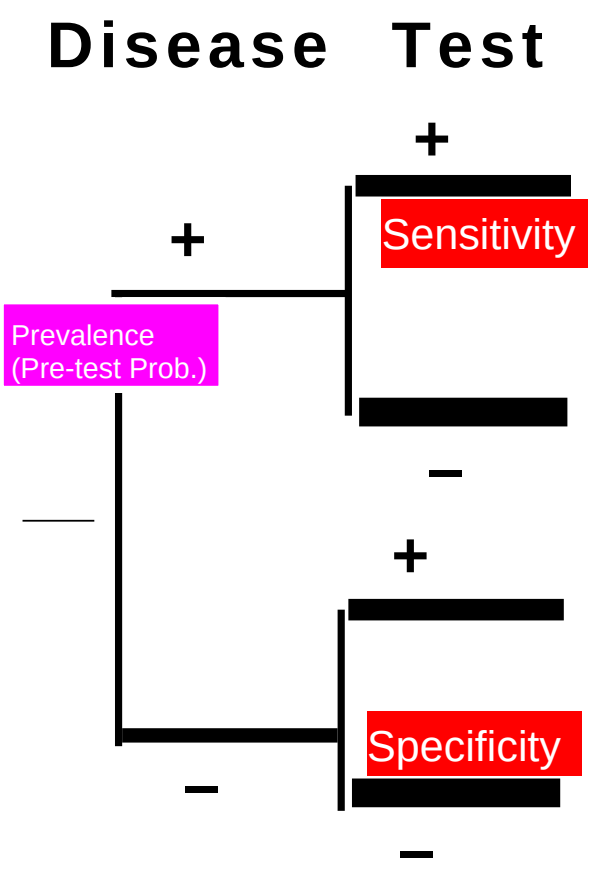




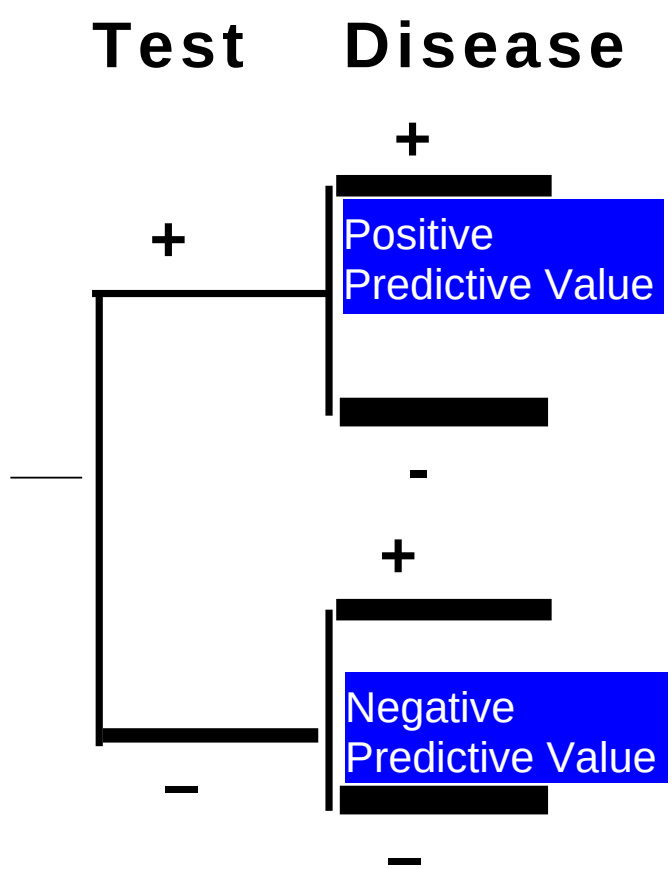




Testing Dx Tests...



Dx Tests In Practice...



Positive Predictive Value: Prob[Disease+ | Test +]

$$PPV = \frac{PriorProb[D] \times Sens}{PriorProb[D] \times Sens + (1 - PriorProb[D]) \times (1 - Spec)}$$

Negative Predictive Value: Prob[Disease - | Test -]

$$NPV = \frac{(1 - PriorProb[D]) \times Spec}{(1 - PriorProb[D]) \times Spec + PriorProb[D] \times (1 - Sens)}$$

Post-Test Odds of D+ after positive (+) test ["Pre-test" = Prior]

$$\text{Post-test Odds} = \text{Likelihood Ratio}_{(+)} \times \text{Pre-test Odds}$$

$$= \frac{\text{Sensitivity}}{\text{1 - Specificity}} \times \frac{\text{Pre-test Probability}}{\text{1 - Pre-test Probability}}$$

$$= \frac{\text{True Positive Fraction}}{\text{False Positive Fraction}} \times \frac{\text{Pre-test Probability}}{\text{1 - Pre-test Probability}}$$

Post-Test Odds of D+ after negative (-) test

$$\text{Post-test Odds} = \text{Likelihood Ratio}_{(-)} \times \text{Pre-test Odds}$$

$$= \frac{\text{1 - Sensitivity}}{\text{Specificity}} \times \frac{\text{Pre-test Probability}}{\text{1 - Pre-test Probability}}$$

$$= \frac{\text{False Negative Fraction}}{\text{True Negative Fraction}} \times \frac{\text{Pre-test Probability}}{\text{1 - Pre-test Probability}}$$

The odds formulation separates the characteristics of test (LR) from the context (PriorProb[D]).

Reverse Probabilities: Probability[data | Hypothesis] ≠ Probability[Hypothesis | data]

U.S. National Academy of Sciences under fire over plans for new study of DNA statistics:

Confusion leads to retrial in UK.

[NATURE p 101-102 Jan 13, 1994]

... He also argued that one of the prosecution's expert witnesses, as well as the judge, had **confused two different sorts of probability.**

One is the probability that DNA from an individual selected at random from the population would match that of the semen taken from the rape victim, a calculation generally based solely on the frequency of different alleles in the population.

The other is the separate probability that a match between a suspect's DNA and that taken from the scene of a crime could have arisen simply by chance [1] ***-- in other words that the suspect is innocent despite the apparent match. This probability depends on the other factors that led to the suspect being identified as such in the first place.***

During the trial, a forensic scientist gave the first probability in reply to a question about the second. Mansfield convinced the appeals court that the error was repeated by the judge in his summing up, and that this slip -- widely recognized as a danger in any trial requiring the explanation of statistical arguments to a lay jury -- justified a retrial.

In their judgement, the three appeal judges, headed by the Lord Chief Justice, Lord Farquharson, explicitly stated that their decision "should not be taken to indicate that DNA profiling is an unsafe source of evidence".

Nevertheless, with DNA techniques being increasingly used in court cases, some forensic scientists are worried that flaws in the presentation of their statistical significance could, as in the Deen case, undermine what might otherwise be a convincing demonstration of a suspect's guilt.

Some now argue, for example, that quantified statistical probabilities should be replaced, wherever possible, by a more descriptive presentation of the conclusions of their analysis. "The whole issue of statistics and DNA profiling has got rather out of hand," says one.

Others, however, say that the Deen case has been important in revealing the dangers inherent in the '**prosecutor's fallacy**'. They argue that this suggests the need for more sophisticated calculation and careful presentation of statistical probabilities.

"The way that the prosecution's case has been presented in trials involving DNA-based identification has often been very unsatisfactory," says David Balding, lecturer in probability and statistics at Queen Mary and Westfield College in London. "Warnings about the prosecutor's fallacy should be made much more explicit. After this decision, people are going to have to be more careful."

---

[1] Underlining is mine (JH). The wording of the singly-underlined phrase is imprecise; the doubly-underlined wording is much better .. if you read 'despite' as "given that" or "conditional on the fact of" JH

Reverse Probabilities: Probability[data | Hypothesis] ≠ Probability[Hypothesis | data]

"The prosecutor's fallacy"

Who's the DNA fingerprinting pointing at?

New Scientist, 29 Jan. 1994, 51-52. David Pringle

Pringle describes the successful appeal of a rape case where the primary evidence was DNA fingerprinting. In this case the statistician Peter Donnelly opened a new area of debate. He remarked that

forensic evidence answers the question

"What is the probability that the defendant's DNA profile matches that of the crime sample, assuming that the defendant is innocent?"

while the jury must try to answer the question

"What is the probability that the defendant is innocent, assuming that the DNA profiles of the defendant and the crime sample match?"

(JH) Donnelly's words make the contrast of the two types of probability much "crisper". The fuzziness of the wording on the previous page is sadly typical of the way statistical concepts often become muddied as they are passed on.

Apparently, Donnelly suggested to the Lord Chief Justice and his fellow judges that they imagine themselves playing a game of poker with the Archbishop of Canterbury. If the Archbishop were to deal himself a royal flush on the first hand, one might suspect him of cheating. Assuming that he is an honest card player (and shuffled eleven times) the chance of this happening is about 1 in 70,000.

But if the judges were asked whether the Archbishop were honest, given that he had just dealt a royal flush, they would be likely to place the chance a bit higher than 1 in 70,000 *.

The error in mixing up these two probabilities is called the "the prosecutor's fallacy", and it is suggested that newspapers regularly make this error.

Apparently, Donnelly's testimony convinced the three judges that the case before them involved an example of this and they ordered a retrial

from Vol 3.02 of Chance News

*** (JH) This is a very nice example of the advantages of Bayesian over Frequentist inference .. it lets one take one's prior knowledge (the fact that he is the Archbishop) into account.**

SOUNDING BOARD

SCREENING FOR HIV: CAN WE AFFORD THE FALSE POSITIVE RATE?

We are a testing culture: we test our urine for drugs; we test our sweat for lies. It is not surprising that we should also test our blood for the acquired immunodeficiency syndrome (AIDS). But before we screen low-risk groups for antibody to the human immunodeficiency virus (HIV), we should consider what the results will mean. Tests for HIV antibody appear to be characterized by extraordinarily low false positive rates. Even so, positive initial and confirmatory tests in someone at low risk of HIV infection are by no means synonymous with infection, because of the possibility of false positive results. Furthermore, any increase in the false positive rate could turn a screening program into a social catastrophe.

Whatever its scientific merits, widespread HIV-antibody testing is becoming a political reality. Blood banks screen potential donors; the armed forces test recruits and personnel on active duty; the State Department tests Foreign Service officers and their dependents; and the Peace Corps and Job Corps test their applicants. Soon, screening of immigrants, prisoners in federal penitentiaries, and perhaps veterans will begin. Pregnant women have been advised to undergo testing in both the first and third trimesters.[1] President Reagan has suggested that applicants for marriage licenses should also be screened.[2]

Plans to test low-risk populations for HIV antibody generally ignore the possibility of false positive results. When screening of blood donors began two years ago, decontaminating the blood supply was an urgent need; it justified the assumption that confirmatory testing could identify most, or at least enough, of the testing errors. But before we establish a public policy of widespread screening, we should consider whether testing that is justified in the blood bank is also justified in other settings. If the false positive rate is not virtually zero, screening a population in which the prevalence of HIV is low will unavoidably stigmatize and frighten many healthy people. How will these mistakes change the lives of the unfortunate persons who are incorrectly identified as infected? Will such screening affect the course of the AIDS epidemic? Does the benefit of identifying infected persons justify the personal and social burden of false positive tests?

Characteristics of the Tests

The central issue is the false positive rate of tests for HIV infection. Current screening programs use a sequence of tests, starting with an enzyme immunoassay. Serum samples yielding repeatedly positive results on enzyme immunoassay are subjected to more complicated and expensive confirmatory testing, typically with a Western blot. A positive confirmatory test is considered evidence of HIV infection.

The results of screening among blood donors allow us to deduce an upper limit for the false positive rate in testing conducted to date. In 1985 and 1986, 0.01 percent of female blood donors in Atlanta and of both male and female blood donors in the northeastern Netherlands had antibody to HIV on both enzyme immunoassay and Western blot assay.[3,4] In the worst case, if none of those blood donors were truly infected, then the highest possible false positive rate for the pair of tests would be 0.01 percent. Because some of those blood donors were truly infected, the false positive rate was almost certainly even lower. If we make the best-case assumption that the probability of a false positive Western blot is independent of the probability of a false positive enzyme immunoassay, or if we have data about the false positive rate on Western blot tests among patients with false positive enzyme immunoassays, the joint false positive rate of the two tests in sequence will equal the product of their false positive rates. One recent study found that the false positive rates of six commercial enzyme immunoassay kits used to test blood from donors ranged from zero to 0.42 percent.[5] Another study noted variations in false positive rates of enzyme immunoassays, even among different batches of one manufacturer's kit.[6] Other investigators have found that the false positive rate of enzyme immunoassays can be as high as 6.8 percent among hospitalized patients.[7]

Confirmatory tests are intended to distinguish false positive results of enzyme immunoassays from those that truly represent HIV infection. Here, variations in the false positive rate may be even more important. The Western blot, the most common confirmatory test for HIV antibody and a standard against which new techniques are evaluated, is complex and very labor intensive. Its techniques have not been standardized, and the magnitude and consequences of interlaboratory variations have not been measured. Its results require interpretation, and the criteria for this interpretation vary not only from laboratory to laboratory but also from month to month. When widespread Western blot confirmation of positive findings on enzyme immunoassays began in 1985, a band indicating the presence of antibody to a protein of 24,000 to 25,000 daltons was regarded as evidence of infection. Some laboratories report this as a 24-kd band, whereas others report it as a 25-kd band. Within a year, many investigators had concluded that apparent bands in this region could represent artifacts and that even a definite band there was not specific for HIV infection.[8]

By mid-1986, the U.S. Army had adopted criteria that required either a band at 41 kd or bands at both 24 and 55 kd. But when investigators from the Army HIV-testing program sent panels of 15 serum samples from healthy adults at low risk to five large commercial firms offering HIV Western blot testing, six different specimens were classified as positive. All samples had yielded repeatedly negative results at the Walter

(209)

Reed Army Institute of Research. Three laboratories considered 1 of 15 specimens positive; one considered 3 positive.[9]

Within several months of the report from Walter Reed, investigators in both Sweden and Paris reported what they considered false positive results on Western blot tests despite the presence of both 25- and 55-kd bands. Their conclusion was based on the absence of risk factors in the individual blood donors and of concordant findings on confirmatory tests in research laboratories.[10,11] Reactivity to the cultured human cells in which the virus had grown served to explain two unexpectedly positive Western blots.[12,13] To find that explanation, one patient's serum was examined in three research laboratories. Other investigators have reported instances in which one specimen from a patient yielded results on a Western blot that were interpreted as positive, whereas subsequent specimens from the same patient yielded negative results.[14,15] Several abstracts presented at the recent Third International Conference on AIDS described extensive retesting and follow-up of "atypical positive" results that would clearly be considered negative according to the U.S. Army criteria published a year earlier.[16-20] Another study described very sensitive Western blot tests that even showed reactivity in the 41-kd region to serum from normal donors at low risk for HIV infection.[21] Thus, the lack of standardization persists.

A recent Army study compared the interpretation of the first Western blot performed with the final classification of the specimens after more extensive investigation. Among specimens that were repeatedly positive on enzyme immunoassay, the false positive rate was 1.17 percent.[22] If the false positive rate of enzyme immunoassays is about 0.4 percent, the joint false positive rate of the two tests performed sequentially should be about 0.005 percent. A pair of tests with a joint false positive rate of 1 per 20,000 is unusual in clinical medicine.

These reports reflect the difficulty, uncertainty, and even disagreement that characterize testing for antibody to HIV. They suggest that positive results from low-risk populations deserve thoughtful interpretation and perhaps further testing. Despite these technical difficulties, laboratories testing blood donors and military recruits have achieved a very high standard of performance. However, specimens collected in more widespread screening programs might not all be analyzed in reference laboratories or with the same techniques. Decentralized testing might further compromise standardization. Smaller laboratories could not offer the research methods that are sometimes used to verify positive Western blot findings in persons at low risk. Technicians processing the specimens might not be as skilled as those who have developed the technique, and laboratories performing a large number of tests might be less inclined to scrutinize positive results. Interlaboratory variation in test characteristics may increase as a new generation of tests (under development by more than 25 companies) becomes available.[23] Some new tests have been proposed to be used as a one-stage procedure, thus eliminating the extra protection of an independent confirmatory test.[24,25]

Prevalence of Infection

What do we know about the prevalence of HIV infection? Perhaps 50 percent of homosexual men in San Francisco have serologic evidence of the infection. The prevalence of seropositivity among intravenous drug abusers and among patients with hemophilia who received factor VIII concentrate pooled before the advent of heat inactivation is similar.[3,8] At somewhat lower risk are patients who received repeated transfusions of red cells, platelets, and plasma before routine HIV testing of donated blood began in 1985. Antibody testing of one group of patients with leukemia treated between 1978 and 1985 showed that about 5 percent became seropositive. The patients who became seropositive had received an average of 164 units of blood products.[26]

Other segments of the population are at much lower risk. Screening of military recruits has shown 0.16 percent of the men and 0.06 percent of the women to be seropositive.[27] When antibody screening of donated blood began in 1985, 1 unit of blood in 2500 had HIV antibody.[28] At that rate, the chance of infection from 2 units of blood donated before antibody screening began would be about 0.08 percent. Among female blood donors, as noted, the reported prevalence of seropositivity is 0.01 percent. Some of these donors may have had sexual contact with members of known high-risk groups; among women without such contact, the prevalence of infection may be even lower than 0.01 percent.

Meaning of Positive Tests

Test sensitivity is not the issue here, and to emphasize our concern with the false positive rate, our analysis makes the best-case assumption that the combination of enzyme immunoassay and Western blot testing for HIV is 100 percent sensitive, identifying all persons who are infected. The meaning of positive tests will depend on the joint false positive rate. Because we lack a gold standard, we do not know what that rate is now. We cannot know what it will be in a large-scale screening program. However, we can be fairly sure that without careful quality control, it will rise.

Bayes' rule allows us to calculate the probability that a person with positive tests is infected.[29] Imagine testing 100,000 people, among whom the prevalence of disease is 0.01 percent. Of the 100,000, 10 are infected; 99,990 are not. A combination of tests that is 100 percent sensitive will correctly identify all 10 who are infected. If the joint false positive rate is 0.005 percent, the tests will yield false positive results in 5 of the 99,990 people who are not infected. Thus, of the 15 positive results, 10

2.10

will come from people who are infected and 5 from people who are not infected, and the probability that infection is present in a patient with positive tests will be 67 percent.

Figure 1 shows the consequences of screening in four populations. The implications of positive test results depend on the joint false positive rate. The horizontal axis shows a range of joint false positive rates from 0 to 0.5 percent. If the prevalence of infection is 5 percent or higher, more than 90 percent of persons with positive tests will truly be infected, whether the joint false positive rate is 0 or 0.5 percent. Unfortunately, this is not true in populations at lower risk. The probability that infection is present in a male army recruit with positive tests is 97 percent if the joint false positive rate is 0.005 percent, and 94 percent if the joint rate is 0.01 percent, but it will be only 62 percent if the joint rate rises to 0.1 percent. The probability that infection is present in a female blood donor with positive tests is about 67 percent if the joint false positive rate is 0.005 percent, and about 50 percent if the joint rate is 0.01 percent, but it will be only 9 percent if the joint rate rises to 0.1 percent. In other words, at this higher joint false positive rate, 10 women without HIV infection will be falsely identified as infected for each truly infected blood donor found. If the joint false positive rate increases to 0.5 percent, as might occur in a single-stage testing program, then 50 women without HIV infection will be stigmatized for every truly infected person identified.

The joint false positive rate may rise if single-stage testing is introduced into physicians' offices; a false positive rate of 0.6 percent was recently reported for such a test.[24,25] The joint rate will rise if tests are performed and interpreted less carefully when the amount of testing increases substantially. Finally, it will rise if criteria for defining a positive Western blot test are less stringent than those observed by the military and the Red Cross.

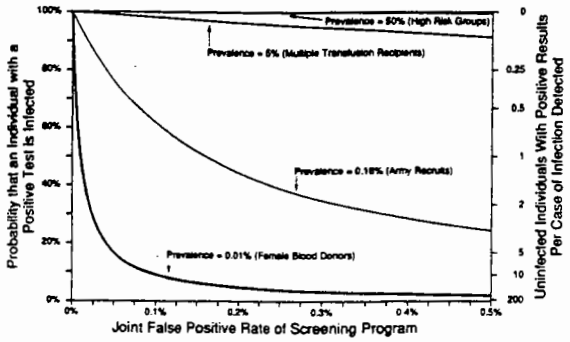


Figure 1. Meaning of Positive Screening Tests for HIV.

The horizontal axis shows the joint false positive rate of the tests. The left vertical scale shows the probability that HIV infection is present in a person with positive tests. The right vertical scale shows the number of uninfected persons falsely classified as infected for every infected person correctly identified. Sensitivity is assumed to be 100 percent. The four lines correspond to four populations that might be screened, each of which has a different prevalence of HIV infection. The boldface line represents low-prevalence populations such as those in which screening has recently been proposed.

Consequences of Widespread Screening

How many cases of infection can we hope to prevent by screening groups at low risk? It is not clear how many of the few infected persons identified would have transmitted the virus to their sexual partners and children, or that testing will substantially reduce the transmission rate.[8,30-34] Screening blood donors prevents transmission because we do not transfuse the blood. But how much does screening change behavior? By no means all seropositive persons are persuaded to practice "safer sex."[35-37] Apparently only a minority abstain from childbearing.[38] What can we expect to happen when we screen other populations? We do not know what changes it would make in public health and our society.

Before we test, we should think again about the ethics of screening and about the social consequences of positive tests for HIV antibody. The first proposals to screen blood donors elicited widespread discussion of the potential threat to individual privacy. Special procedures were devised to ensure that this sensitive information remained private. The statutory requirement of HIV testing would in all likelihood eliminate such protection. The Secretary of Education has suggested that positive test results should be reported not only to public health authorities but also to the sexual partners of the person tested.[39]

Despite educational efforts, public understanding of the epidemic is limited. As we contemplate recommendations and regulations, we should remember that most people consider a "positive AIDS test" to be a sentence to ghastly suffering and death. Patients with such results will take little comfort in Bayes' rule and will be offered little reassurance by their insurers, employers, and acquaintances.

A Time for Caution

The AIDS epidemic frightens us all. But we should not allow our fear to cloud our judgment. Hasty and indiscriminate screening for antibody to HIV is imprudent and potentially dangerous, whether we suggest the tests to young women, require them of engaged couples, or impose them on our veterans. Although screening of blood donors and military recruits appears to have generated few false positive results, we do not know whether this performance can continue if the testing programs are expanded. Standardization and quality control should come first. These will take time and money; monitoring laboratory performance will require continuing effort, expenditure, and regulation.

Nor will our problems be purely technical. HIV screening poses questions that are at once scientific, political, legal, and philosophical. If laws are to link our fates to test results, should not due process be brought to the benches where those tests are performed? We will need guarantees not only of the confi-

dentiality of test results but also of the quality of the testing procedure. Should everyone be subjected to tests of uniform sensitivity and specificity, or should performance characteristics be tailored to the clinical situation? Should screening programs in the general population sacrifice specificity by adopting the highly sensitive tests designed to protect the blood supply? In the past, inexplicably positive results in persons at no apparent risk of HIV infection prompted extensive investigation of the specimens in research laboratories. Wider screening will inevitably yield more unanticipated positive results — perhaps far more than researchers can review. How will we decide whose positive results we scrutinize? Who will weigh the scientific evidence against the skepticism of the person who does not believe his positive test results? Will we recognize the results of tests performed in other countries? How often will we retest and reclassify on the basis of technical advances or because of the passage of time?

If we want to test each other, we should make a deliberate choice of the threshold probability of infection above which we will screen. We should make explicit the trade-offs implicit in any testing program. How many engagements should end to prevent one infection? How many jobs should be lost? How many insurance policies should be canceled or denied? How many fetuses should be aborted and how many couples should remain childless to avert the birth of one child with AIDS?

New England Medical Center
Boston, MA 02111

KLEMENS B. MEYER, M.D.
STEPHEN G. PAUKER, M.D.

Supported in part by a training grant (7044) and a research grant (4493) from the National Library of Medicine, Bethesda, Md.

REFERENCES

1. Gruson L. AIDS toll in children is called 'deadly crisis.' New York Times. April 9, 1987:B8.
2. Boffey PM. Reagan urges wide AIDS testing but does not call for compulsion. New York Times. June 1, 1987:A1.
3. Francis DP, Chin J. The prevention of acquired immunodeficiency syndrome in the United States: an objective strategy for medicine, public health, business, and the community. JAMA 1987; 257:1357-66.
4. Sibinga CTS, de Vries B, McShine RL, Das PC, Schooley RT. HIV antibody screening in a low-risk population (blood donors). Transfusion 1987; 27:215-6.
5. Reesink HW, Lelie PN, Huisman JG, et al. Evaluation of six enzyme immunoassays for antibody against human immunodeficiency virus. Lancet 1986; 2:483-6.
6. Füst G, Ujhelyi E, Héjjas M, Hollán SR. Traps of HIV serology: independent changes in sensitivity and specificity of ELISA kits. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:49.
7. Cockerill FR, Edson RS, Chase RC, Katzmann JA, Taswell HF. "False-positive" antibodies to human immunodeficiency virus (HIV) detected by an enzyme-linked immunosorbent assay (ELISA) in patients at low risk for acquired immunodeficiency syndrome (AIDS). Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:34.
8. Institute of Medicine, National Academy of Sciences. Confronting AIDS: directions for public health, health care and research. Washington, D.C.: National Academy Press, 1986:51,69,306.
9. Burke DS, Redfield RR. False-positive Western blot tests for antibodies to HTLV-III. JAMA 1986; 256:347.
10. Couroucé A-M, Muller J-Y, Richard D. False-positive western blot reactions to human immunodeficiency virus in blood donors. Lancet 1986; 2:921-2.
11. Biberfeld G, Bredberg-Rådén U, Böttiger B, et al. Blood donor sera with false-positive western blot reactions to human immunodeficiency virus. Lancet 1986; 2:289-90.
12. Saag MS, Britz J. Asymptomatic blood donor with a false positive HTLV-III Western blot. N Engl J Med 1986; 314:118.
13. Roy S, Portnoy J, Wainberg MA. Need for caution in interpretation of Western blot tests for HIV. JAMA 1987; 257:1047.
14. Sandstrom EG, Schooley RT, Ho DD, et al. Detection of human anti-HTLV-III antibodies by indirect immunofluorescence using fixed cells. Transfusion 1985; 25:308-12.
15. Stoneburner RL, Chiasson MA, Solomon K, Rosenthal S. Risk factors in military recruits positive for HIV antibody. N Engl J Med 1986; 315:1355.
16. Shriver K, Klanieki J, Houghton R, Masinovsky R, McClure J, Watson AJ. Analysis of sera exhibiting atypical reactions with HIV. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:16.
17. McGrath K, Mijch A, Maskell W, et al. Follow-up of western blot positive blood donors — Victoria, Australia. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:103.
18. Thomas D, Mundon FK, Zimmerman D, Larson D, Gowan L, Wilhelm S. Analysis of discrepant anti-HIV ELISA reactives. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:103.
19. Busch M, Shiota J, Nason M, Samson S, Vyas G, Perkins H. Serologic and culture follow-up study of anti-HIV reactive blood donors. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:108.
20. Dock NL, Lamberson HV, O'Brien TA, Petteway SR, Alexander S, Poiesz BJ. Evaluation of HIV antibody reactivity in blood donors with atypical western blot patterns. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:148.
21. Tan PL, Kay JWD, Munjal D. Studies of normal donor specimens causing various reactivity patterns in sensitive western blot assays. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:151.
22. Burke DS, Brandt BL, Redfield RR, et al. Diagnosis of human immunodeficiency virus infection by immunoassay using a molecularly cloned and expressed virus envelope polypeptide: comparison to Western blot on 2707 consecutive serum samples. Ann Intern Med 1987; 106:671-6.
23. Blakeslee S. Pressure for wider AIDS testing fuels search for better methods. New York Times. June 9, 1987:C14.
24. Foreman J. Faster, cheaper AIDS test reported. Boston Globe. June 5, 1987:1.
25. Quinn TC, Francis H, Kline R, et al. Evaluation of a latex agglutination assay using recombinant envelope polypeptides for detection of antibody to HIV. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:166.
26. Human immunodeficiency virus infection in transfusion recipients and their family members. MMWR 1987; 36:137-40.
27. Human T-lymphotropic virus type III/lymphadenopathy-associated virus antibody prevalence in U.S. military recruit applicants. MMWR 1986; 35:421-4.
28. Schorr JB, Berkowitz A, Cumming PD, Katz AJ, Sandler SG. Prevalence of HTLV-III antibody in American blood donors. N Engl J Med 1985; 313: 384-5.
29. Pauker SG, Kassirer JP. Decision analysis. N Engl J Med 1987; 316:250-8.
30. Selwyn PA, Schoenbaum EE, Feingold AR, et al. Perinatal transmission of HIV in intravenous drug abusers (IVDAs). Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:157.
31. Willoughby A, Mendez H, Minkoff H, et al. Human immune deficiency virus in pregnant women and their offspring. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:158.
32. Blanche S, Rouzioux C, Veber F, Le Deist F, Mayaux MJ, Griscelli C. Prospective study on newborns of HIV seropositive women. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:158.
33. Braddick M, Kreiss JR, Quinn T, et al. Congenital transmission of HIV in Nairobi, Kenya. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:158.
34. Nzilambi N, Ryder RW, Behets F, et al. Perinatal HIV transmission in two African hospitals. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:158.
35. Willoughby B, Schechter MT, Boyko WJ, et al. Sexual practices and condom use in a cohort of homosexual men: evidence of differential modification between seropositive and seronegative men. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:5.
36. Pollak M, Schiltz MA, Lejeune B. Safer sex and acceptance of testing: results of the nationwide annual survey among French gay men. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:5.
37. Pesce A, Negre M, Cassuto JP. Knowledge of HIV contamination modalities and its consequence on seropositive patients behaviour. Presented at the 3rd International Conference on AIDS, Washington, D.C., June 1–5, 1987:60.
38. Wofsy CB. Human immunodeficiency virus infection in women. JAMA 1987; 257:2074-6.
39. Werner LM. Education chief presses AIDS tests. New York Times. May 1, 1987:A18.

DISTINGUISHING POPULATIONS WITH DIFFERENT MEAN *BIRTHWEIGHTS*

The entries in the 4 panels below represent birthweights, recorded to the nearest 10 grams, but with the ending 0 removed to save space. Thus the very first entry of 336 in Panel A represents a birthweight of 3360 grams or 3.36 Kg. The birthweights in a panel are all from infants of the same sex, but different panels may be from different sexes. The standard deviation of the entries in each panel is approximately SD = 43 (430 grams).

By eye, by comparing all the entries in a panel with all of those in another, you may be able to discern if two panels have different means. But what can you conclude if you take just a sample from each of 2 panels and perform a formal test of significance on the difference in the sample means? **Details for exercise are explained on p 5.**

PANEL A

336	357	338	379	386	362	277	340	404	300
295	340	264	317	303	342	340	400	348	327
294	390	347	346	294	407	408	380	343	413
346	360	321	379	338	345	377	362	318	341
428	346	354	358	353	401	338	283	356	275
366	303	351	378	413	381	319	312	298	281
372	380	282	303	345	282	445	304	339	357
314	264	380	389	264	325	327	298	334	347
299	428	338	277	268	310	345	316	396	381
400	318	341	321	328	370	336	371	371	449

PANEL B

397	399	306	371	356	368	362	396	338	326
331	411	422	413	381	399	385	333	293	311
319	349	268	383	398	328	385	373	274	467
328	377	300	341	386	387	265	411	378	358
373	336	366	325	322	283	329	323	327	401
292	313	340	424	311	363	335	350	343	364
348	298	314	401	384	362	370	375	373	312
399	355	435	437	362	316	371	340	315	359
414	302	317	407	432	334	428	386	406	388
325	334	448	344	373	296	301	347	361	294

PANEL C

344	382	358	429	398	336	406	366	385	357
258	346	401	315	430	373	377	346	378	357
346	406	425	346	367	347	388	348	300	326
333	397	355	282	360	421	416	346	370	329
366	360	282	393	329	352	450	371	379	323
430	397	349	321	334	369	367	274	427	355
349	393	295	372	283	313	316	268	334	413
322	397	309	348	376	345	497	343	361	391
327	374	344	354	322	277	287	396	323	389
391	303	319	314	368	389	343	342	330	369

PANEL D

262	328	363	399	328	375	310	417	278	346
340	350	364	299	318	339	307	381	314	388
355	290	331	304	351	333	382	310	331	287
370	356	394	265	368	288	448	416	350	333
306	360	236	273	381	435	332	323	349	354
294	337	390	408	299	345	375	428	273	353
407	419	333	331	330	387	303	275	334	335
391	348	348	302	356	370	374	353	352	432
353	346	356	342	382	293	348	332	375	350
346	407	339	364	288	389	282	434	380	378

Key
Cailíní[céad/deireadh -- trí céad, daiched is a trí/seacht]
Buachaillí [-- trí céad, deich is daichead, is a sé]

DISTINGUISHING POPULATIONS WITH DIFFERENT MEAN *ADULT HEIGHTS*

The entries in the 4 panels below represent adult heights, recorded to the nearest centimetre. Thus the 1st entry (188) in Panel A represents a height of 188 cm or 1.68m. The birthweights in a panel are all from adults of the same sex, but different panels may be from different sexes. The standard deviation of the entries in each panel is approximately SD = 6cm.

By eye, by comparing all the entries in a panel with all of those in another, you may be able to discern if two panels have different means. But what can you conclude if you take just a sample from each of 2 panels and perform a formal test of significance on the difference in the sample means? **Details for exercise are explained on p 5.**

PANEL A

188	178	175	168	169	171	170	166	161	171
180	178	184	174	168	176	175	167	182	177
181	183	185	178	165	172	178	176	164	186
176	179	169	169	184	169	173	173	173	177
177	170	179	183	183	172	189	181	174	171
170	182	163	171	176	176	183	181	174	175
171	167	175	175	174	168	170	175	185	181
183	180	178	170	174	173	176	173	175	173
165	172	175	183	167	171	176	182	174	170
187	185	167	169	168	178	182	178	171	175

PANEL B

156	159	169	161	157	158	171	166	169	170
168	170	175	171	167	168	160	170	173	165
160	162	156	150	168	157	168	167	159	168
159	165	165	165	164	163	159	169	176	176
166	155	164	162	172	172	156	166	166	161
165	162	177	162	160	171	164	174	164	173
174	160	164	163	171	172	159	157	159	168
161	166	160	167	168	162	158	154	159	167
166	163	166	177	168	172	177	169	175	166
158	156	165	161	162	157	168	163	167	166

PANEL C

171	175	178	168	181	177	185	174	177	177
169	174	184	173	182	179	178	167	186	175
176	172	176	174	174	170	184	173	174	174
179	177	177	176	171	161	172	168	177	176
186	172	173	184	167	161	166	171	180	163
181	176	179	176	170	172	165	178	174	182
169	179	176	183	172	172	170	178	179	178
179	166	174	184	169	164	177	180	183	172
183	164	178	166	177	186	174	179	175	179
183	165	174	173	172	171	176	188	181	169

PANEL D

165	161	168	155	172	160	176	170	162	161
167	158	155	163	158	159	174	179	161	157
176	171	160	164	167	173	174	163	162	157
155	167	161	163	169	168	158	166	160	167
163	162	165	167	169	161	174	164	154	174
171	168	162	173	164	172	170	166	165	163
166	168	158	161	175	164	164	164	167	173
162	164	161	169	170	157	164	169	161	166
174	168	174	168	156	160	153	167	167	156
176	165	161	164	161	163	168	161	173	166

Key
Fir [ar clé -- céad, deich is trí fichid, cúig]
Mná [-- céad, trí fichid, cúig]

"Homegrown" Exercises around M&M Chapter 6

• **Birthweight**:
Perform a test of each of the following 4 (obviously competing, so not independent) contrasts; use **new** samples of size n=4 and n=4 for each of the 4 tests; use a z-test (σ is given) with alpha = **0.10** (two-sided, so z_{alpha}=1.645) for each. []

1. μ_A vs. μ_B 2. μ_C vs. μ_D 3. μ_A vs. μ_D 4. μ_B vs. μ_C

• **Adult heights**:
test the following 4 contrasts*, again using n = 4 vs n = 4.

1. μ_A vs. μ_C 2. μ_B vs. μ_D 3. μ_A vs. μ_D 4. μ_B vs. μ_C

* NB: 1 and 2 are not the same as 1 and 2 for birthweight above.

To save you time, the structure of the tests is laid out below.

To help with rapid compilation of results in class, circle below which contrasts yielded "statistically significant" differences and BRING YOUR 8 DECISIONS TO CLASS.

Birthweights	A vs. B	C vs. D	A vs. D	B vs. C
Adult Heights	A vs. C	B vs. D	A vs. D	B vs. C

"Arithmetic" of Testing if 2 panels have same mean

H_0: $\mu_1 = \mu_2$ [same sex] α = **0.10** (2-sided) H_{alt}: $\mu_1 \neq \mu_2$ [different sexes]

Reject H_0 (i.e. infer that $\mu_1 \neq \mu_2$) if

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > 1.645 \quad \text{or} \quad \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -1.645$$

(use z-test since σ is given)

i.e. conclude "different sexes" if

$$|\bar{x}_1 - \bar{x}_2| > 1.645\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad ***$$

σ is given, so we can work out ahead of time (from ***) what difference between $\bar{x}_1$ and $\bar{x}_2$ would lead us to conclude "different sexes"... the average birthweights need to be > 50 (ie 500g) apart, and average heights > 7 cm apart.

[with t-tests, we don't have σ, and in fact have to calculate s from sample)

-3- Exercise to Illustrate Type I Errors and Statistical Power

	Value of $1.645\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	
	BIRTHWEIGHTS σ = 43 g x 10	ADULT HEIGHTS σ = 6 cm
$n_1 = n_2 =$ 4	**50** g x 10	**7.0** cm

Just for interest, here is what is is for other sample sizes...

$n_1 = n_2 =$ 8	35.4 g x 10	4.9 cm
$n_1 = n_2 =$ 16	25.0 g x 10	3.5 cm

On class, I will 'play god' and tell you which contrasts belong in which rows. In practice, you may not be able to unequivocally determine the truth -- or it may take a lot more work. And determining how big a difference is takes even more work.

Results of statistical tests [columns] performed by students in relation to real situations[rows]

BIRTHWEIGHT	"Can't say" p > 0.10 ("negative") ("N.S")	"different" p < 0.10 ("positive") ("Stat. sig.")	No. of Tests
same sex			
different sexes			

ADULT HEIGHT	"Can't say" p > 0.10 ("negative") ("N.S")	"different" p < 0.10 ("positive") ("Stat. sig.")	No. of Tests
same sex			
different sexes			

Results of statistical tests [columns] performed by students in relation to real situations [rows]

BIRTHWEIGHT

	"Can't say" p > 0.10 ("negative") ("N.S")	"different" p < 0.10 ("positive") ("Stat. sig.")	No. of Tests
3. A vs. D (=) same sex 4. B vs. C (=)	KD KD AB AB AC AC AS AS BF TE TE NW NW NW NW NW NW NW NW PW PW RB RB AS AS BM AS SP SP MPS MPS VC VC SR SR RP RP AE JB JB MS MS KO KO CF CF US1 US1 LR LR BM MA MA BMcG BMcG USZ USZ VS VS KR JS JS JMcK JMcK TdiP TdiP AM AR AR SF SF GF GF 73	AS MPS MS MM MM KR AM BF 8	81
1. A vs. B (≠) different sexes 2. C vs. D (≠)	BM AS AS SP SP MPS MPS MS VC VC SR SR RP RP AE JB JB MS MS KO KO CF CF USI US1 LR LR MA MA BMcG BMcG US2 US2 MM MM VS VS KR JS JS JMcK JMcK SF SF AM AM AR AR TdiP TdiP PW PW RB RB AS AS NW NW NW NW NW NW NW BF BF TE TE AC AC AS AB AB KD KD GF GF 76	BM MS AS KR NW AS 6	82

Results of statistical tests [columns] performed by students in relation to real situations [rows]

ADULT HEIGHT

	"Can't say" p > 0.10 ("negative") ("N.S")	"different" p < 0.10 ("positive") ("Stat. sig.")	No. of Tests
3. A vs. C (=) same sex 4. B vs. D (=)	BM BM SP AS AE AE RP SR SR MS MS USI CF CF KO KO JB JB KD KD AS AS AC AC TE TE BF BF NW NW NW NW NW NW NW NW AS RB PW PW PW SF SF AR AM AM TdiP TdiP JMcK JMcK JS JS KR KR VS MM MM US2 USZ BMcG BMcG MA MA LR LR GF GF 67	AS RB AR VS VC VC SP AS 8	75
1. A vs. D (≠) different sexes 2. B vs. C (≠)	EB NW NW NW JMcK MA MA US1 KO JB MS SP GF 13	KD KD EB AS AS AC AC TE TE BF BF NW NW NW BMcG AS BMcG AS RB RB PW PW SF SF AR AR AM AM TdiP JS TdiP JS KR KR VS VS MM MM JMcK CF CF KO JB AE RP SR US2 US2 NW NW LR LR USI SR VC VC MS SP AS AS BM BM GF 63	76

"Operating" Characteristics of a Statistical Test

As with diagnostic tests, there are 2 ways statistical test can be wrong:

1) **The null hypothesis was in fact correct but the sample was genuinely extreme and the null hypothesis was therefore (wrongly) rejected.**

2) **The alternative hypothesis was in fact correct but the sample was not incompatible with the null hypothesis and so it was not ruled out.**

The probabilities of the various test results can be put in the same type of 2x2 table used to show the characteristics of a diagnostic test.

		Result of Statistical Test	
		"**Negative**" (do not reject H_0)	"**Positive**" (reject H_0 in favour of H_a)
TRUTH	H_0	$1 - \alpha$	α
	H_a	β	$1 - \beta$

The quantities $(1 - \beta)$ and $(1 - \alpha)$ are the "sensitivity (power)" and "specificity" of the statistical test. Statisticians usually speak instead of the complements of these probabilities, the false positive fraction (α) and the false negative fraction (β) as "Type I" and "Type II" errors respectively [It is interesting that those involved in diagnostic tests emphasize the correctness of the test results, whereas statisticians seem to dwell on the errors of the tests; they have no term for $1-\alpha$].

Note that all of the probabilities start with (i.e. are conditional on knowing) the truth. This is exactly analogous to the use of sensitivity and specificity of diagnostic tests to describe the performance of the tests, conditional on (i.e. given) the truth. As such, they describe performance in a "what if" or artificial situation, just as sensitivity and specificity are determined under 'lab' conditions.

So just as we cannot interpret the result of a Dx test simply on basis of sensitivity and specificity, likewise we cannot interpret the result of a statistical test in isolation from what one already thinks about the null/alternative hypotheses.

Interpretation of a "positive statistical test"

It should be interpreted n the same way as a "positive diagnostic test" i.e. in the light of the characteristics of the subject being examined. The lower the prevalence of disease, the lower is the post-test probability that a positive diagnostic test is a "true positive". Similarly with statistical tests. We are now no longer speaking of sensitivity = Prob(test + | H_a) and specificity = Prob(test - | H_0) but rather, the other way round, of Prob(H_a | test +) and Prob(H_0 | test -), i.e. of positive and negative predictive values, both of which involve the "background" from which the sample came.

A Popular Misapprehension: It is not uncommon to see or hear seemingly knowledgeable people state that

> "the P-value (or alpha) is the probability of being wrong if, upon observing a statistically significant difference, we assert that a true difference exists"

Glantz (in his otherwise excellent text) and Brown (Am J Dis Child 137: 586-591, 1983 -- on reserve) are two authors who have made statements like this. For example, Brown, in an otherwise helpful article, says (italics and strike through by JH) :

> ~~*"In practical terms, the alpha of .05 means that the researcher, during the course of many such decisions, accepts being wrong one in about every 20 times that he thinks he has found an important difference between two sets of observations"*~~ [1]

[1] [Incidentally, there is a second error in this statement : it has to do with equating a "statistically significant" difference with an important one... minute differences in the means of large samples will be statistically significant]

But if one follows the analogy with diagnostic tests, this statement is like saying that

> *"1-minus-specificity is the probability of being wrong if, upon observing a positive test, we assert that the person is diseased".*

We know [from dealing with diagnostic tests] that we cannot turn Prob(test | H) into Prob(H | test) without some knowledge about the unconditional or a-priori Prob(H) ' s.

The influence of "background" is easily understood if one considers an example such as a testing program for potential chemotherapeutic agents. Assume a certain proportion P are truly active and that statistical testing of them uses type I and Type II errors of α and β respectively. A certain proportion of all the agents will test positive, but what fraction of these "positives" are truly positive? It obviously depends on α and β, but it also depends in a big way on P, as is shown below for the case of α = 0.05, β = 0.2.

P -->	0.001	.01	.1	.5
TP = P(1- β) -->	.00080	.0080	.080	.400
FP = (1 - P)(α)->	.04995	.0495	.045	.025
Ratio TP : FP -->	≈ 1 : 62	≈ 1: 6	≈ 2 : 1	≈ 16 : 1

Note that the post-test odds TP:FP is

$$P(1-\beta) : (1-P)(\alpha) = \{ P : (1-P) \} \times \left[\frac{1-\beta}{\alpha} \right]$$

PRIOR × function of TEST's characteristics

i.e. it has the form of a "**prior odds**" P : (1 - P), the "background" of the study, multiplied by a "**likelihood ratio positive**" which depends only on the characteristics of the statistical test. Text by Oakes helpful here

Are All Significant P Values Created Equal? The Analogy Between Diagnostic Tests and Clinical Research

Warren S Browner, MD MPH and Thomas B Newman, MD MPH †

Just as diagnostic tests are most helpful in light of the clinical presentation, statistical tests are most useful in the context of scientific knowledge. Knowing the specificity and sensitivity of a diagnostic test is necessary, but insufficient: the clinician must also estimate the prior probability of the disease. In the same way, *knowing* the P value and power, or the confidence interval, for the results of a research study is necessary but insufficient: the reader must estimate the prior probability that the research hypothesis is true. Just as a positive diagnostic test does not mean that a patient has the disease, especially if the clinical picture suggests otherwise, a significant P value does not mean that a research hypothesis is correct, especially if it is inconsistent with current knowledge. Powerful studies are like sensitive tests in that they can be especially useful when the results are negative. Very low P values are like very specific tests; both result in few false-positive results due to chance. This Bayesian approach can clarify much of the confusion surrounding the use and interpretation of statistical tests. (JAMA 1987;257:2459-2463)

IN THE four ORIGINAL CONTRIBUTIONS in this issue of THE JOURNAL, the authors report the results of statistical tests of 76 hypotheses.[1-4] Of these, 32 had significant P values (P<.05). But do these P values imply that the 32 hypotheses are true? Or that 95% of them are true? Are all significant P values created equal?

† From the Departments of Medicine (Dr Browner), Pediatrics (Dr Newman), and Epidemiology and International Health (Drs Browner and Newman), School of Medicine, University of California at San Francisco, and the Clinical Epidemiology Program, Institute for Heath Policy Studies, San Francisco (Drs Browner and Newman).

The answer to these questions is "No!" What then is a P value? It is the likelihood of observing the study results under the assumption that the null hypothesis of no difference is true. Probably because this definition is elusive and intimidating, understanding P values (and other statistical concepts like power, confidence intervals, and multiple hypothesis testing) is often left to experts in the field. It is easier just to check whether a P value is .05 or less, call the result "statistically significant," regard the tested hypothesis as probably true, and move on to the next paragraph.

Readers of medical literature need not give up quite so quickly, however. As Diamond and Forrester[5] pointed out, many statistical concepts have remarkably similar analogues in an area familiar to clinicians - the interpretation of diagnostic tests. In the diagnosis of Cushing's syndrome, for example, most clinicians recognize that an elevated serum cortisol level is more useful than an elevated blood glucose level, and that an elevated cortisol level is more likely to be due to Cushing's syndrome in a moon-faced patient with a buffalo hump and abdominal striae than in an overweight patient with hypertension.[6-7] Why?? Because the interpretation of a test result depends on the characteristics of both the test and the patient being tested.[8-13]

The same type of reasoning - called *Bayesian analysis* after Thomas Bayes, the mathematician who developed it more than 200 years ago[14] - can also be used to clarify the meaning of the P value and other statistical terms. Although this application of Bayes' ideas has been discussed in epidemiologic and statistical literature,[15-18] it has received less attention in the journals read by clinicians. In this article, we begin with the basic aspects of the analogy between research studies and diagnostic tests, such as the similarity between the power of a study and the sensitivity of a test, and then examine more challenging issues, such as how a study with multiple hypotheses resembles a serum chemistry panel.

THE ANALOGY

An overview of the analogy between research studies and diagnostic tests is shown in Table 1. In this analogy, a clinician obtains diagnostic data to test for the presence of a disease, such as breast cancer, and an investigator collects study data to determine the truth of a *research hypothesis* such as that the efficacies of two drugs differ in the treatment of peptic ulcer disease. (The research hypothesis is often called the *alternative hypothesis* in standard terminology.) The absence of a *disease* (no breast cancer) is like the *null hypothesis* of no difference in the efficacy of the two drugs.

The term "positive" is used in its usual sense: to refer to diagnostic tests that are consistent with the presence of the disease and to studies that have statistically

significant results. Similarly, "negative" refers to diagnostic tests consistent with the absence of disease and research results that fail to reach statistical significance. Thus there are four possible results whenever a patient undergoes a diagnostic test. Consider the use of fine-needle aspiration in the evaluation of a breast mass, for example (Table 2). If the patient has breast cancer, there are two possibilities: the test result can either be correctly positive or incorrectly negative. On the other hand, if the patient actually does not have cancer, then the result will either be correctly negative or incorrectly positive. Similarly, there are four possible results whenever an investigator studies a research hypothesis (Table 3). If the efficacies of the two drugs really do differ, there are two possibilities: the study can be correctly positive if it finds a difference or incorrectly negative if it misses the difference. If the two drugs actually have the same efficacy, then the study can either be correctly negative if it finds no difference or incorrectly positive if it does find one.

Table l.-The Analogy Between Diagnostic Tests and Research Studies

Diagnostic Test	Research Study
Absence of disease	Truth of null hypothesis
Presence of disease	Truth of research (altematve hypothesis
Positive result (outside normal limits)	Positive result (reject null hypothesis)
Negative result (within normal limits)	Negative result (fail to reject null hypothesis)
Sensitivity	Power
False-positive rate (I - specificity)	P value
Prior probability of disease	Prior probability of research hypothesis
Predictive value of a positive (or negative) test result	Predictive value of a positive (or negafive) study

The relationships between the four possible outcomes of a diagnostic test are usually expressed as the *sensitivity* and *specificity* of the test, which are determined by assuming that the presence or absence of the disease is known. Sensitivity is the likelihood that a test result will be positive in a patient with the disease. Specificity is the likelihood that a test result will be negative in a patient without the disease. If the result from a fine-needle aspiration is positive in 80 of 100 women with breast cancer, and negative in 95 of 100 women without cancer, the test would have a sensitivity of 80% and a specificity of 95%. There is another term that is useful in the analogy: the false-positive rate (1-specificity), which is the likelihood that a test result will be (falsely) positive in someone without the disease. In this example, the false-positive rate is 5%: of 100 women without breast cancer, five will have falsely positive test results.

Table 2.—The Four Possible Results of a Diagnostic Test

		If Breast Mass is actually:	
		Malignant	Benign
And Result of Fine-Needle Aspirate is:	Positive	This is a true-positive test: result is correct	This is a false-positive test: result is incorrect
	Negative	This is a false-negative test: result is incorrect	This is a true-negative test result is correct

Similarly, the relationships between the four possible outcomes of a research study are usually expressed as the *power* and *P value* of the study, which are determined by assuming that the truth or falsity of the null hypothesis is known. Power is the likelihood of a study being positive if the research hypothesis is true (and the null hypothesis is false); it is analogous to the sensitivity of a diagnostic test. The P value is the likelihood of a study being positive when the null hypothesis is true; it is analogous to the false-positive rate (1 - specificity) of a diagnostic test. A study comparing two drugs in the treatment of ulcers that has an 80% chance of being correctly positive if there really is a difference in their efficacies wuould have a power of 0.80. A study with a 5% chance of being incorrectly positive if there is no difference between the drugs would have aP value of .05. (Conventionally, when the P value is less than a certain predetermined "level of statistical significance," usually .01 or .05, the results are said to be "statistically sign)ficant.")

Table 3.—The Four Possible Results of a Research Study

		If Research Hypothesis is actually:	
		True	False
		(Efficacy of Drug A and Drug B differ in treatment of ulcer disease)	(Drug A has same efficacy as B in treatment of ulcer disease)
And Result of Study is:	Positive	This is a true-positive study: result is correct	This is a false-positive study: result is incorrect
	Negative	This is a false-negative study: result is incorrect	This is a true-negative study result is correct

Knowing the sensitivity and specificity of a test is not aufficient, however, to interpret its results: that interpretation also depends on the charactertistics of the patient being tested. If the patient is a 30-year-old woman uith several soft breast

masses, a positive result from a fine-needle aspiration (even with a false-positive rate of only 5%) would not suffice to make a diagnosis of cancer. Similarly, if the patient is a 60-year-old woman with a firm solitary breast mass, a negative aspirate result (with a sensitivity of 80%) would not rule out malignancy[19] Clinicians use these sorts of patient characteristics to estimate the prior probability of the disease—the likelihood that the patient has the disease, made prior t to knowing the test results. The prior probability of a disease is based on the history and physical findings, previous experience with similar patients, and knowledge of alternative diagnostic explanations. It can be very high (breast cancer in the 60-year old woman with a single firm mass), very low (breast cancer in the younger woman), or somewhere in between. Although they may not realize it, clinicians express prior probabilities when using phrases such as "a low index of suspicion" or "a strong clinical impression."

In the same way, knowing the power and the P value of a study is not sufficient to determine the truth of the research hypothesis. That determination also depends on the characteristics of the hypothesis being studied. Suppose one drug is diphenhydramine hydrochloride (Benadryl) and the other is chlorpheniramine maleate (Chlor-Tri-meton): a positive study (at P=0.05) would not ensure that one of the drugs is effective in the treatment of ulcers Similarly, if one drug was ranitidin hydrochloride (Zantac) and the other placebo, a negative study (even with power of 0.80) would not establish the ineffectiveness of ranitidine. The characteristics of a research hypothesis determine its prior probability—an estimate of the likelihood that the hypothesis is true, made prior to knowing the study results. The prior probability of a hypothesis is based on biologic plausibility, previous experience with similar hypotheses. and knowledge of alternative scientific explanations Analogous to the situation with diagnostic tests, the prior probability of a research hypothesis can be very high (that an H_2-blocker, such as ranitidine is more effective than placebo in the treatment of ulcers), very low (that the efficacies of two H1-blockers, such as diphenhydramine and chlorpheniramine, differ in the treatment of ulcer disease), or somewhere in between. Authors of research reports indicate prior probabilities with terms like "unanticipated" or "expected" when they discuss their results.

The advantage of Bayesian analysis in interpreting diagnostic tests is that it can determine what the clinician really wants to know—the likelihood that the patient has the disease, given a certain test result. Bayesian analysis combines the characteristics of the patient (expressed as the prior probability of disease), the characteristics of the test (expressed as sensitivity and specificity), and the test result (positive or negative) to determine the predictive value of a test result. The predictive value of a positive diagnostic test is the probability that given a positive result, the patient actually has the disease. (The predictive value of a negative test is the probability that given a negative result, the patient does not have the disease.)

As an example, recall the 60-year-old woman with a firm breast mass. The prior probability that the mass is malignant is moderate, say 50%. A positive result from a fine-needle aspirate (with a specificity of 95% and a sensitivity of 80% for cancer) results in a very high predictive value for malignancy, about 94% (Figure). Next, consider the 30 year-old woman with multiple soft masses. The prior probability of cancer is low, say 1%. Even given a positive aspirate result, the likelihood that she has breast cancer is still small (about 14%).

A Bayesian approach can also be used to determine what the reader of a research study really wants to know—the likelihood that the research hypothesis is true, given the study results. It combines the characteristics of the hypothesis (expressed as prior probability), the characteristics of the study (expressed as power and the P value), and the study results (positive or negative) to determine the predictive value of a study. The predictive value of a positive study is the probability that given a positive result, the research hypothesis is actually true. (The predictive value of a negative study is the probability that given a negative result, the research hypothesis is false.)

The predictive value of a research study, however, is usually harder to estimate than the predictive value of a diagnostic test (see "Limitations" section). Nonetheless, the basic analogy remains valid: the prior probability of the hypothesis must be combined with the power and the P value of the study to determine the likelihood that the research hypothesis is true. In the next section, we discuss how this analogy can be used to understand several statistical concepts.

IMPLICATIONS

Specificity and the P Value

How low must a P value be for it to be accepted as evidence of the truth of a research hypothesis? This question is analogous to asking: how high must the specificity of a test be to accept a positive test result as evidence of a disease? Requiring that a P value be less than 0.05 before it is "significant" is as arbitrary as requiring that a diagnostic test have a specificity of at least 95%. A more important criterion, but one that is not as easy to quantitate, is whether the results of the study combined with the prior probability of the research hypothesis are sufficient to suggest that the hypothesis is true. Consider the hypothesis, tested in the Lipid Research Clinics Primary Prevention Trial [20] that cholestyramine resin decreases the incidence of coronary heart disease in hypercholesterolemic men. This research hypothesis had at least a low to moderate prior probability, based on previous evidence. Even with a "nonsignificant" P value of .094 (the two-sided equivalent of the controversial one-sided P=.047 reported by the investigators), the hypothesis is likely to be true.

It is also a mistake to believe a research hypothesis just because a P value is statistically significant. Consider a study that found that drinking two or more cups of coffee a day was associated with pancreatic cancer (P<.06).2' This hypothesis had a very low prior probability: the authors called the association "unexpected." Thus,

finding a significant P value did not establish the truth of the hypothesis; subsequent studies, including one by the same authors, failed to confirm the association. [22-27]

Of course, many diagnostic test results are not simply reported as "positive"; they also indicate how abnormal the result is. The more abnormal that result, the less likely that it is just a chance finding in a normal person. If the upper limit of normal for a serum thyroxine level at a specificity of 95% is 12.0 µg/dL (154 nmol/L), then a thyroxine level of 18.0 µg/L (232 nmol/L) is almost certainly abnormal. The question becomes whether it represents hyperthyroidism, another disease, or a laboratory error. By analogy, if the cutoff for calling a study positive is a P value less than .05, then a P value of.0001 means chance is an extremely unlikely explanation for the findings. The question becomes whether the results indicate the truth of the research hypothesis or are a result of confounding or bias (see "Laboratory Error and Bias" and "Alternative Diagnoses and Confounding Explanations" sections). Because the P value is analogous to the false-positive rate (1- specificity), a study with a very low P value is like a test with very high specificity: both give few false-positive results due to chance, but may require careful consideration of other possible explanations.

Sensitivity and Power

When the result of a diagnostic test that has a high sensitivity is negative, such as a urinalysis in the diagnosis of pyelonephritis, it is especially useful for ruling out a disease. Similarly, when a powerful research study is negative, it strongly suggests that the research hypothesis is false. However, if the sensitivity of a test is low, such as a sputum smear in a patient with possible tuberculosis, then a negative result does not rule out the disease.[9] In the same way, a negative study with inadequate power cannot disprove a research hypothesis,[28,29]

Laboratory Error and Bias

When unexpected or incredible results on a diagnostic test are found, such as a serum potassium level of 9.0 mEqlL (mmol/L) in an apparently well person, the first possibility to consider is laboratory error: Was the test adequately performed? Did the sample hemolyze? Was the specimen mislabeled? Similarly, readers of a research study, such as a trial of biofeedback in the treatment of hypertension, must always consider the possibility of bias, especially if the study yields surprising results: Was the study adequately designed and executed? Did the investigators assign subjects randomly? Was blood pressure measured blindly? [30] Improperly performed tests and biased studies do not ~ ield reliable information, no matter how specific or significant their results.

Alternative Diagnoses and Confounding Explanations

Even if a diagnostic test is adequately performed, there may be several explanations for the result. An elevated serum amylase level, for example. has a high specificity to distinguish patients who have pancreatitis from those with nonspecific abdominal pain. However, there are extrapancreatic diseases (such as bowel infarction) that elevate the amylase level and that must be considered in the differential diagnosis. In the same way, although a low P value may indicate an association between an exposure and a disease (like the association between carrying matches and lung cancer), a confounder (cigarette smoking) may actually be responsible. Readers of research studies should always keep in mind potential confounding explanations for significant P values.

Better Tests and Bigger Studies

Increasing the sample size in a research study is similar to using a better diagnostic test. Better diagnostic tests can have more sensitivity or specificity or both, Iarge studies can have greater power or lower levels of statistical significance or both. Often the choice of a diagnostic test is a matter of practicality: biopsies are not feasible in every patient for every disease. Similarly, power or the significance level may be determined by practical considerations, since studies of 20 000 or more subjects cannot be done for every research question. Of course, bigger studies may find smaller differences, just like better tests may detect less advanced cases of a disease. A small but statistically significant difference in a research study is like a subtle but definite abnormality on a diagnostic test; its importance is a matter of judgment.

Intentionally Ordered Tests and Prospective Hypotheaea

A positive result on a single intentionally ordered test is more likely to indicate disease than the same result that turns up on a set of routine admission laboratory tests. Similarly, the P value for a research hypothesis stated in advance of a study is usually more meaningful than the same P value for a hypothesis generated by the data. The reason is that clinicians usually order tests and investigators state hypotheses in advance when the prior probability is moderate or high. Thus the predictive values of positive results are generally greater for intentionally ordered tests and prospectively stated hypotheses.

Not all unexpected results however, have low prior probabilities. Occasionally, clinicians or investigators are just not smart or lucky enough to consider the diagnosis or hypothesis in advance. For example, a house officer caring for a patient with fatigue and vague abdominal symptoms might ignore a serum calcium level of 10.5 mg/dL (2.62 mmol/L) until the attending physician mentions the possibility of hyperparathyroidism in rounds the next morning. Similarly, researchers might disregard the association between smoking and cervical cancer until a plausible biologic explanation is suggested.[31-34] Estimating the prior probability of a

hypothesis on the basis of whether it was considered prospectively is a useful, but not infallible, method. The truth, elusive though it sometimes may be, does not depend on when a hypothesis is first formulated.

Multiple Tests and Multiple Hypothesea

Most of us are intuitively skeptical when one of 50 substances on a checklist is associated with a disease at P<.05 because of the likelihood of finding such an association by chance alone. A standard technique for dealing with this problem of testing multiple hypotheses is to use a more stringent level of statistical significance, thus requiring a lower P value.[35, 36] This approach is simple and practical, but it leads to some unsatisfying situations. It seems unfair, for example, to reduce the required significance level for a reasonable hypothesis just because other, perhaps ridiculous, hypotheses were also tested. What if the disease was mesothelioma and one of the exposures was asbestos: should a more stringent level of statistical significance be required because 49 other substances were also included? Should the level of significance be reduced when testing the main hypothesis of a study whenever additional hypotheses are considered? Need statistical adjustments for multiple hypothesis testing be made only when reporting all of the hypotheses in a single publication?

This vexing problem of multiple hypothesis testing resembles the interpretation of a serum chemistry panel. When a clinician evaluates a patient with a swollen knee, a serum uric acid level of 10.0 mg/dL (0.6 mmol/L) has the same meaning no matter how many other tests were also performed on the specimen by the autoanalyzer. However, an unanticipated abnormal value on another test in the panel is likely to be a false-positive: that is because the diseases it might represent usually have low prior probabilities, not because several tests were performed on the same sample of serum. Similarly, testing multiple hypotheses in a single study causes problems because the prior probabilities of such hypotheses tend to be low: when investigators are not sure, what they are looking for, they test many possibilities. The solution is to recognize that it is not the number of hypotheses tested, but the prior probability of each of them, that determines whether a result is meaningful.

Confirmatory Tests and Pooled Studies

When a single diagnostic test is insufficient to make a diagnosis, additional tests are often ordered, some results of which may be positive and some negative. The clinician revises the probability of the disease by combining these results, often weighting them by the tests' characteristics. In a patient with a swollen leg, for example, a normal result from a Doppler study would lowerwer the probability of deep venous thrombosis, but an abnormal result of a fibrinogen scan might raise it aufficiently to make the diagnosis. In the same way, it may be necessary to combine the results of several research studies weighting them by the characteristics of each study. This process, known as *pooling*, allows studies with both significant and nonsignificant P values to change incrementally the likelihood that a research hypothesis is true. Howeever, just as only those tests that are relevant to the diagnosis in question should be combined, only those research studies that address the same research hypothesis should be pooled.

Confidence Intervals

There is no ready diagnostic test analogy for confidence intervals from research studies (the concept of test precision comes closest). But because confidence intervals are commonly mis-interpreted as expressions of predictive value, they merit a short discussion. The term "confidence interval" is unfortunate, because it leads many people to believe that they can be confident that the interval contains the true value being estimated. Actually, confidence intervals are determined entirely by the study data: the prior probability that the true value lies within that interval is not at all considered in the calculations. A 95% confidence interval is simply tne range of values that would not differ from the estimate provided by the study at a statistical significance level of 0.05 [38,39].

Confidence intervals are useful because they define the upper and lower limits consistent with a study's data. But they do not estimate the likelihood that the results of the research are correct. A confidence interval provides no more information about the likelihood of chance as an explanation for a finding than does a P value.[40] As an example, suppose a well-designed study finds that joggers are twice as likely as non-joggers to develop coronary heart disease, with a 95% confidence interval for the relative risk of 1.01 to 3.96. (This is equivalent to rejecting the null hypothesis of no association between jogging and heart disease at P=.05). Despite a 95% confidence interval that excludes 1.0, there is obviously not a 95% likelihood that joggers are at an increased risk of coronary heart disease. There are many other studies that have found that exercise is associated with a reduced risk of heart disease. Given the low prior probability of the hpothesis that jogging increases the risk of coronary heart disease, chance (or perhaps bias) would be a more likely explanation for the results.

LIMITATIONS

While it provides several useful insights the analogy between diagnostic tests and clinical research is not perfect. It is easier to determine the prior probability of a disease, based on the prevalence of the disease in similar patients, than the prior probability of a hypothesis, based on the prevalence of the truth of similar hypotheses. Similarity in patients can be defined by characteristics known to be associated with a disease, such as age, sex, and symtoms.[11] But what defines similar hypotheses? Thus the prior probability of most research hypotheses tends to be a subjective estimate (although, in practice estimates of the prior probability of a disease are generally subjective as well).

Second, as long as there is a gold standard for its diagnosis, a disease is either present or absent: there are only these two possibilities. If a group of patients known to have the disease is assembled, a single value for the sensitivity of a test can be determined empirically. But there is no single value for the power of a research study: it depends on the sample size, as well as the magnitude of the actual difference between the groups being compared. A study comparing IQ in internists and surgeons for example, might have a power of only 50% to detect a difference between them if surgeons actually scored five points higher than internists, but a power of 98% if surgeons actually scored ten points higher. Since the actual difference is unknown, a unique value for power cannot be calculated.

CONCLUSIONS

Clinicians do not simply decide that a patient has a disease when a diagnostic test result is positive or rule out the disease when the test result is negative. They also consider the sensitivity and specificity of the test and the characteristics of the patient being tested. In the same way, readers should not believe or disbelieve the research hypothesis of a study on the basis of whether the results were statistically significant. They should also take into account the study's power and P value and the characteristics of the hypothesis being tested.

Thus, all signficant P values are not created equal. Just as the accuracy of a diagnosis depends on how well the clinician has estimated the prior probability and considered alternative diagnoses and laboratory errors, the interpretation of a research study depends on how well the reader has estimated the prior probability and considered confounders and biases. Knowing the power and P value (or the confidence interval) for a study's results, like knowing the sensitivity and specificity of a diagnostic test, is necessary but not sufficient. This Bayesian approach requires the active participation of the reader and emphasizes the importance of scientific context in the interpretation of research.

This project was supported by a grant from the Andrew W. .Mellon Foundation.

REFERENCES

1. Schade DS, Mitchell WJ, Griego G: Addition of sulfonylurea to insulin treatment in poorly controlled type II diabetes: A double-blind, randomized clinical trial. JAMA 1987,257:2441-2445.

2. Cramer DW, Goldman MB, Schiff I, et al: The relationship of tubal infertility to barrier method and oral contraceptive use. JAMA 1987;257: 2446-2450.

3. Bennett KJ, Sackett DL, Haynes RB, et al: A controlled trial of teaching critical appraisal of the clinical literature to medical students. JAMA 1987,257:2451-2454.

4. Chaiken BP, WilliamsNM, Preblud SR, et al: The effect of a school entry law on mumps activity in a school district. JAMA 1987;257:2455-2458.

5. Diamond GA, Forrester JS: Clinical trials and statistical verdicts: Probable grounds for appeal. Ann Intern Med 1983,98:385-394.

6. Nugent CA, Warner HR, DunnJT, et al: Probability theory in the diagnosis of Cushing's syndrome. J Clin Endocrinol Metab 1964;24:621-627.

7. Crapo L: Cushing's syndrome: A review of diagnostic tests. Metabolism 1979;28:955-977.

8. Vecchio TJ: Predictive value of a single diagnostic test in unselected populations. N Engl J Med 1966,274:1171-1173.

9. Boyd JC, Marr JJ: Decreasing reliability of acidfast smear techniques for detection of tuberculosis. Ann Intern Med 1975;82:489-492.

10. Jones RB: Bayes' theorem, the exercise ECG, and coronary artery disease. JAMA 1979;242: 1067-1068.

11. Diamond GA, Forrester JS: Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. N Engl J Med 1979;300:1350-1358.

12. Griner PF, Mayewski RJ. Mushlin AI, et al: Selection and interpretation of diagnostic tests and procedures: Principles and applications. Ann Intern Mled 1981:94:553-600.

13. Havey RJ, Krumlovsky F, delGreco F, et al: Screening for renovascular hypertension: Is renal digital-subtraction angiography the preferred noninvasive test? JAMA 1985:254:388-393.

14. Bayes T: An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond 1763:53:370-418.

13. Phillips LD: Bayesian statistics for Social Scientists. New York, Crowell. 1974.

16. Donner A: A Bayesian approach to the interpretation of subgroup results in clinical trials. J Chronic Dis 1982;35:429-435

17. Pater JL, Willan AR: Clinical trials as diagnostic tests. Controlled Clin Trials 1984:5:107-113.

18. Thomas DC, Siemiatychi J Dewar R. et al: The problem of multiple inferences in studies designed to generate hypotheses. Am J Epidemiol 1985 122:1080-1095.

19. Mushlin A1: Diagnostic tests in breast cancer: Clinical strategies based on diagnostic probabilities. Ann Intern Med 1985:103:79-85.

20. Lipid Research Clinics Program: The Lipid Research Clinics Coronary Primary Prevention Trial results: I. Reduction in incidence of coronary heart disease. JAMA 1984;251:351-364.

21. MacMahon B, Yen S, Trichopolous D, et al: Coffee and cancer of the pancreas. N Engl J Med 1981;304:630-633.

22. Jick H, Dinan BJ: Coffee and pancreatic cancer. Lancet 1981 2:92.

23. Goldstein HR: No association found between coffee and cancer of the pancreas. N Engl J Med 1982;306:997.

24. Wynder EL, Hall NEL, Polansky M: Epidemiology of coffee and pancreatic cancer. Cancer Res 1983;43:3900-3906.

25. Kinlen LJ, McPherson K: Pancreas cancer and coffee and tea consumption: A case-control study. Br J Cancer 1984;49:9~96.

26. Gold EB, Gordis L, Diener MD, et al: Diet and other risk factors for cancer of the pancreas. Cancer 1985,55:460-467.

27. Hsieh C, MacMahon B, Yen S, et al: Coffee and pancreatic cancer (chapter 2) N Engl J Med 1986;315:587-589.

28. Frieman JA, Chalmers TC, Smith H Jr, et al: The importance of beta, type II errors and sample size in the randomized control trial: Survey of 71 'negative' trials. N Engl J Med 1978;299:690-694.

29. Young MJ, Bresnitz EA, Strom BL: Sample size nomograms for interpreting negative clinical studies. Ann Intern Med 1983;99:248-251.

30. Sackett DL: Bias in analytic research. J Chronic Dis 1979,32:51-63.

31. Winkelstein W Jr: Smoking and cancer of the uterine cervix: Hypothesis. Am J Epidemiol 1977,106:257-259.

32. Wright NH, Vessey MP, Kenward B, et al: Neoplasia and dysplasia of the cervix uteri and contraception: A possible protective effcet of the diaphragm. Br J Cancer 1978;38:273-279.

33. Harris RWC, Brinton LA, CowdelI RH, et al: Characteristics of women with dysplasia or carcinoma in situ of the cervix uteri. Br J Cancer 1980;42:359-369.

34. Lyon JL, GardnerJW, West DW, et al: Smoking and carcinoma in situ of the uterine cervix. Am J Public Health 1983,73:558-562.

35. Godfrey K: Comparing the means of several groups. N Engl J Med 1985;313:1450-1456.

36. Cupples LA, Heeren T, Sehatzkin A, et al: Multiple testing of hypotheses in comparing two groups. Ann Intern Med 1984;100:122-129.

37. Cole P: The evolving case-control study. J Chronic Dis 1979,32:15-27.

38. Fleiss JL: Statistical Methods for Rates and Proportions, ed 2. New York, John Wiley & Sons Inc, 1981, p 14.

39. Rothman K: A show of confidence. N Engl J Med 1978;299:1362-1363.

40. Browner WS, Newman TB: Confidence intervals. Ann Intern Med 1986;105:973 974.