Multi-Test Screening and the Chances of Being Normal

William R. Phillips, MD, MPH, and Donovan J. Thompson, PhD
Seattle, Washington

Screening programs involve responsibility for appropriate action on abnormal test results. When multiple-test screening batteries are used, a simple probability formula is commonly used to predict the proportion of healthy individuals who will have one or more abnormal test results occur by chance alone. This formula is valid only when the assumptions upon which it rests are met in the population being tested. In many situations the assumptions are not met and the formula overestimates the occurrence of abnormal results in healthy populations. Data for three screening programs involving blood chemistry test batteries on 769 patients document this overestimate and its magnitude. Clinical judgment, not misapplied probability theory, should guide the physician's strategy in evaluating abnormal results of screening tests.

The clinical question is common and the answer confusing: What is the physician to do when a battery of screening tests reports an unexpected abnormal result in an apparently healthy patient?

Clinical judgment must, of course, guide the management of each case. Many authors have cautioned restraint in pursuing such test results on the grounds that their occurrence is likely due to chance alone.[1-6] They remind us that laboratory limits are set to include the central 95 percent of the healthy population and exclude the extreme 5 percent. Thus, they point out, the probability of finding an abnormal result on any test is 0.05, simply by the statistical definition of the limits, and that as more tests are performed their probabilities are multiplied to give an increasing likelihood that the patient will exceed limits on at least one test in the battery. That is, the probability p that a healthy individual will be within limits on each of the total number of tests n, each with probability s (usually 0.95), is given by the formula $p=s^n$. The probability of the individual having at least one test exceeding limits is the complement (1-p), and is given in Table 1 for common numbers of laboratory tests. Following this probability argument, the clinician is cautioned to expect 46 percent of healthy patients to have, by chance alone, at least one test exceeding limits on a standard 12-test blood chemistry screen.[7-9]

The argument is straightforward; its proponents rightly remind us of the important role statistics play in clinical medicine; and the formula is correct when the assumptions upon which it is based are met. The problem remains, however, that the resulting answer is wrong and judgments made upon it misguided. The argument leads to the illogical conclusion that if a healthy population is subjected to a large enough number of tests nearly everyone will be found to be abnormal.[6] This is the quantitative reflection of the clinical absurdity of describing a healthy patient as "one who has been

From the Department of Family Medicine, School of Medicine, and the Department of Biostatistics, School of Public Health and Community Medicine, University of Washington, Seattle, Washington. Dr. Phillips was with the Division of General Practice, Department of Community Health, School of Medicine, University of Aukland, New Zealand, during part of this work. Requests for reprints should be addressed to Dr. William R. Phillips, Department of Family Medicine, School of Medicine, RF-30, University of Washington, Seattle, WA 98195.

0094-3509/80/131095-04$01.00

Table 1. Number of Tests and Predicted Probability of Exceeding Laboratory Limits on at Least One Test

Number of Tests n	Probability of All Tests Within Limits* p	Probability of One or More Tests Exceeding Limits (1−p)
1	0.95	0.05
6	0.74	0.26
8	0.66	0.34
12	0.54	0.46
14	0.49	0.51
17	0.42	0.58
20	0.36	0.64
100	0.006	0.99

*$p=s^n$ where $s=0.95$

inadequately studied." These shortcomings have not prevented repetition of the argument in textbooks,[6,10] major journals,[1-3] national research symposia,[11] and continuing medical education publications.[5]

The reasoning rests upon at least three major assumptions: first, that the established limits for each test accurately include 95 percent of the healthy population and, if the usual parametric method of setting those limits at two standard deviations above and below the mean is used, it further requires that the values for each laboratory test follow the bell-shaped Gaussian frequency distribution; secondly, that each test is independent of all other tests in the battery; and thirdly, that the patient belongs to the same population to which the frequency distribution applies. The first and second assumptions are not true, and the third is precisely the hypothesis that is being tested when the clinician uses laboratory tests in evaluating a patient.

Most biological variables do not conform to the classic Gaussian distribution.[2,10,12] Of the 12 blood chemistry tests commonly included in automated multi-channel test batteries, only albumin fits the Gaussian model. Thus, the usual techniques of parametric statistics fail to accurately describe the distribution of these test results and the resulting laboratory limits do not faithfully divide the population into the assumed 95 and 5 percent proportions. To avoid this difficulty, some laboratories are developing alternative non-parametric or distribution-free methods of defining limits by dividing the population on the basis of percentiles or tolerance limits.[10] Even these techniques, however, are plagued with the inescapable difficulties of defining normal and abnormal in statistical terms.[1,8,10,12-15]

If the formula for combining probabilities is to hold, each test must be independent of all the others. The clearest promoters of the probability argument state this plainly enough,[1-3,5,7,16] but then go on to give the calculated probabilities without considering the actual degree of inter-relationship between common tests. Clinical reasoning alone suggests the levels of many biochemical constituents, as well as other physiologic variables, must be somehow correlated in the healthy as well as the sick. In fact, some degree of correlation must be the rule rather than the exception.[15] Common sense and evidence[16] suggest that a patient within limits on LDH must be somewhat more likely to also be within limits on SGOT than another patient who exceeds LDH limits. In the case of multiple tests following the normal distribution, correlation between tests would cause the actual proportion of patients having abnormal test results to be less than that predicted by the probability formula for independent tests. The degree of overestimation by the formula can only be demonstrated by empirical review of actual screening program experience.

The literature on multiple-test screening programs, extensive as it is, does not supply the data needed for this assessment. Many reports deal with hospital patients or other selected populations that do not represent a healthy screening population. Also, published studies usually report results from the point of view of the laboratory rather than from that of the patient. Many papers

Table 2. Number of Tests and Percentage of Population Exceeding Laboratory Limits on at Least One Test Observed in Three Screening Programs

Patient Group	Number Patients	Number Tests	Patients Exceeding Limits on at Least One Test: Observed Number	Observed Percent	Predicted Percent*
A	477	8	98	20.6	33.7
B	175	14	65	37.1	51.2
C	117	17	39	33.3	58.2

*Predicted percent = 100 × (1 − p), where (1 − p) equals probability of one or more tests exceeding limits, as given in Table 1

report the number of abnormal tests observed per patient; few report the proportion of patients with abnormal test results.[16]

The purpose of this study is to examine results from multiple-test screening programs and compare the proportion of patients with abnormal test results observed in practice with that predicted by the probability formula. Deficiencies in the probability argument predict that a discrepancy will be found between the observed and the predicted proportions and that the predicted values will prove to be overestimates of the actual experience.

Methods

This study examined existing data originally collected in actual screening programs on three populations of asymptomatic, reportedly healthy adults in one community.[17,18] Fasting blood samples were collected and processed by standard automated multi-channel analyzers at three separate major national commercial laboratories and reported with reference to sex-specific adult laboratory limits. Group A included 477 individuals with the following eight tests reported from one laboratory on each: alkaline phosphatase, creatinine, total bilirubin, uric acid, serum glutamic oxalacetic transaminase (SGOT), T_4 complement protein binding, globulins, and glucose. Group B comprised 175 individuals with 14 tests performed on each by a second laboratory, including the first seven tests listed for Group A, plus total protein, albumin, blood urea nitrogen (BUN), calcium, cholesterol, lactate dehydrogenase (LDH), and organic phosphorus. Group C comprised 117 individuals with 17 tests done on each at a third laboratory, including the first five tests listed for Group A, all seven tests listed for Group B, plus: serum sodium, potassium, chloride, carbon dioxide, and glucose.

Each represents a screening population; all individuals reported themselves to be healthy. If a small number of persons with sub-clinical or unreported disease were accidentally included the effect would be to increase the observed proportion of patients with abnormal test results and, thus, obscure any difference between the predicted and the actual occurrence of such results in a healthy screening population.

Results

Results of each of the three screening programs are shown in Table 2, contrasting the observed proportion of patients with abnormal results on one or more tests with that predicted from the probability formula $p=s^n$. In each of the three populations examined, the predicted proportion is significantly greater than that observed in actual experience (p less than 0.001).

As described above, non-parametric methods for defining limits have been developed for use where test results fail to conform to the Gaussian distribution, thereby overcoming one of the objections to the probability argument. Several methods,[7,8,19] based on the binomial probability distribution, have been described for predicting the proportion of a healthy population exceeding laboratory limits on a given number of tests when the limits are defined using these non-parametric techniques. These predictions (not shown here) also overestimate the proportion of patients with abnormal test results when compared to that actually observed in these three groups. This suggests the correlation of tests, apart from the non-Gaussian character of their frequency distributions, is an important cause of the observed discrepancy.

Comment

Empirical confirmation of the discrepancy between the predicted occurrence of abnormal screening test results and their observed occurrence—between the ideal and the real—underscores the need to review assumptions underlying the definition of "normal." Many difficulties complicate determining normality: as a statistical concept,[12,15] a laboratory definition,[1,8,10] or as a clinical classification.[20] It is essential to effective clinical decision making to bear in mind what notion of normality is being used, upon what definition of limits it is based, and what service we expect of it. If ". . . in clinical diagnosis, criteria for range of normal are generally non-existent, the standards of laboratory criteria are generally spurious."[20]

The discrepancy between the predicted and the observed occurrence of abnormal test results is large and may have implications for screening program planning. Illustrated here with biochemical laboratory tests, the point may also apply to other multiple-test screening situations; the size of the discrepancy depending upon the choice of tests and their correlation. Past lack of attention to this discrepancy may have occurred partly because much of the information used to plan screening programs for healthy populations comes, unfortunately, from populations where abnormal test results may indeed be more common and nearer the probability predictions.[16]

Documentation of this discrepancy and its magnitude suggests that the role of chance in accounting for abnormal screening test results may be smaller than previously accepted. Thus, cautions urging restraint in follow-up of abnormal test results, on the argument that they are likely due to chance alone, may be overstated. Still, despite the failure of the probability formula to accurately describe the relationship, the proportion of healthy patients with abnormal test results is likely to increase with increasing numbers of tests in a screening battery and it may be considerable in many clinical situations. Clinical judgment must remain the guiding factor in test use and appropriate follow-up care.

Screening programs must carefully review their selection of tests, their definition of limits, and their policies for test follow-up. Perhaps the clearest rule remains: Decide what is the best action if the test is normal, decide the best action if the test is abnormal, and if both actions are the same do not do the test.[4]

Acknowledgements

The authors wish to acknowledge Patricia W. Wahl, PhD, and Joanne J. Hoover, MD, of the Northwest Lipid Research Clinic, Seattle, who supplied most of the data for this study.

References

1. Korvin CC, Pearce RH: Laboratory screening: A critical survey. Can Med Assoc J 105:1157, 1971
2. Sackett DL: The usefulness of laboratory tests in health-screening programs. Clin Chem 19:366, 1973
3. Thompson RS: Approaches to prevention in an HMO setting. J Fam Pract 9:71, 1979
4. Cochrane AL: Effectiveness and Efficiency: Random Reflections on Health Service. London, Nuffield Provincial Hospitals Trust, 1972, p 43
5. Schoenberg BS: The "abnormal" laboratory result. Postgrad Med 47:151, 1970
6. Galen RS, Gambino SR: Beyond Normality: The Predictive Value and Efficiency of Medical Diagnosis. New York, John Wiley & Sons, 1975
7. Sunderman FW: Expected distributions of normal and abnormal results in multitest surveys of healthy subjects. Am J Clin Pathol 53:288, 1970
8. Werner M, Marsh WL: Normal values: Theoretical and practical aspects. CRC Crit Rev Clin Lab Sci 6:81, 1975
9. Schoen I, Brooks SH: Judgment based on 95% confidence limits: A statistical dilemma involving multitest screening and proficiency testing of multiple specimens. Am J Clin Pathol 53:190, 1970
10. Henry RJ, Reed AH: Normal values and the use of laboratory results in the detection of disease. In Henry RJ, Cannon DC, Winkleman JW (eds): Clinical Chemistry: Principles and Techniques , ed 2. New York, Harper & Row, 1974, pp 343-371
11. Daughaday WH, Erickson MM, White W, et al: Evaluation of routine 12-channel chemical profiles on patients admitted to a university general hospital. In Benson ES, Strandjord PE (eds): Multiple Laboratory Screening. New York, Academic Press, 1969, pp 181-199
12. Elveback LR, Guillier CL, Keating FR: Health, normality and the ghost of Gauss. JAMA 211:69, 1970
13. Murphy EA, Abbey H: The normal range: A common misuse. J Chron Dis 20:79, 1967
14. Murphy EA: The Logic of Medicine. Baltimore, Johns Hopkins University Press, 1976, pp 117-134
15. Oldman PD: Measurement in Medicine: The Interpretation of Numerical Data. London, English Universities Press, 1968, p 187
16. Best WR, Mason CC, Barron SS, et al: Automated twelve-channel serum screening: Part 1: What is normal? Med Clin N Am 53:175, 1969
17. Hoover J, Wahl P, Hazzard W, et al: The Pacific Northwest Bell Telephone Company Health Survey: Cholesterol and Triglyceride Distributions from Work Site Screening Study, Technical Report No. 1, Northwest Lipid Research Clinic. Seattle, University of Washington, 1977
18. Thompson DJ: The association of exposure to arsenic and nerve conduction deficits in smelter employees. Research Report, Environmental Protection Agency. Government Printing Office, 1976
19. Reed AH: Multi-test screening and ninety-five percent limits. Am J Clin Pathol 54:774, 1970
20. Feinstein A: Clinical Judgment. Baltimore, Williams & Wilkins, 1967, p 333