Psychology, **2010, 1: 194-198**
doi:10.4236/psych.2010.13026 Published Online August 2010 (http://www.SciRP.org/journal/psych)

Scientific Research

Using Generalizability Theory to Evaluate the Applicability of a Serial Bayes Model in Estimating the Positive Predictive Value of Multiple Psychological or Medical Tests

Clarence D. Kreiter

Office of Consultation and Research in Medical Education, Department of Family Medicine, University of Iowa, Iowa City, USA.
Email: clarence-kreiter@uiowa.edu

Received May 10th, 2010; revised June 14th, 2010; accepted June 16th, 2010.

ABSTRACT

***Introduction*:** *It is a common finding that despite high levels of specificity and sensitivity, many medical tests are not highly effective in diagnosing diseases exhibiting a low prevalence within a clinical population. What is not widely known or appreciated is how the results of retesting a patient using the same or a different medical or psychological test impacts the estimated probability that a patient has a particular disease. In the absence of a 'gold standard' special techniques are required to understand the error structure of a medical test. Generalizability can provide guidance as to whether a serial Bayes model accurately updates the positive predictive value of multiple test results.* ***Methods*:** *In order to understand how sources of error impact a test's outcome, test results should be sampled across the testing conditions that may contribute to error. A generalizability analysis of appropriately sampled test results should allow researchers to estimate the influence of each error source as a variance component. These results can then be used to determine whether, or under what conditions, the assumption of test independence can be approximately satisfied, and whether Bayes theorem accurately updates probabilities upon retesting.* ***Results*:** *Four hypothetical generalizability study outcomes are displayed as variance component patterns. Each pattern has a different practical implication related to achieving independence between test results and deriving an enhanced PPV through retesting an individual patient.* ***Discussion*:** *The techniques demonstrated in this article can play an important role in achieving an enhanced positive predictive value in medical and psychological diagnostic testing and can help ensure greater confidence in a wide range of testing contexts.*

***Keywords*:** *Generalizability Theory*, *Bayes*, *Serial Bayes Estimation*, *Positive Predictive Value*, *Psychological Testing*, *Serial Medical Testing*

1. Introduction

When a medical disease's prevalence and a medical test's specificity and sensitivity are known, an equations based on Bayes Theorem provides useful information related to the diagnostic power of a medical test. It is a common finding that despite high levels of specificity and sensitivity, many medical tests are not highly effective in diagnosing diseases with a low prevalence within a clinical population [1]. Since a large number of diseases occur only in a small proportion of the population (*i.e.* have low prevalence), the low positive predictive value (PPV) of medically diagnostic tests is of obvious concern to physicians attempting to identify the presence of a low prevalence disease. To provide an example, let's suppose a physician is attempting to determine whether a patient has a disease that occurs in 1% of a defined patient population. When the test is performed on patients with the disease, it yields a positive test result indicating the presence of the disease in 90% of the patients (sensitivity equals .90). When the test is performed on patients without the disease, it correctly identifies 98% of those patients as disease free (specificity equals .98). An equation based on Bayes Theorem can be used to calculate the probability that a patient with a positive test result actually has the disease. The simple equation for calculating this probability is:

$$P(A \mid B) = P(B \mid A) * P(A) / P(B) \quad (1)$$

Equation (1) describes the probability that a patient

has the disease given a positive test result [P (A | B)], and equals the probability of a positive test result given the patient has the disease [P (B | A) - *sensitivity*] multiplied by the probability of the disease [P (A) - *prevalence*] divided by the overall probability of a positive test result within the population [P (B)]. The denominator in Equation (1), the overall prior probability of a positive test result, is derived as shown in Equation (2), where *j* is 1, 2… and takes on as many values as there are hypotheses. In the case being discussed in this example problem, there are just two possible hypotheses (Ho_1: the patient has the disease – Ho_2: the patient does not have the disease) and hence in this example the sum is taken over just two levels. Hence, the overall probability of a positive test result is the sum of the probabilities of a positive test in those with (sensitivity) and without (1 – specificity) the disease each multiplied by their prevalence in the population.

$$P (B) = [\Sigma_j P (B | A_j) P (A_j)] \quad (2)$$

Equation (3) displays the calculation using the levels of specificity, sensitivity and prevalence discussed in our example. Despite high levels of specificity and sensitivity, the patient with a positive test result has only a 31% chance of actually having the disease. This is a common and well known type of finding related to medical testing designed to detect low prevalence diseases.

$$P (A | B) = .90 * .01 / ((.90 * .01) + (.02 * .99)) = .31 \quad (3)$$

What is not widely known or appreciated is how the results of retesting a patient using the same or different test will impact the estimated probability that the patient has the disease. There is little guidance in the medical or psychological literature regarding whether or how the results from serial testing improve the ability to diagnosis disease when the structure or cause of the dependence between tests is uncertain. However, it is clearly important for clinicians to understand how the PPV changes when a patient is administered a second or third medical or psychological test. When the assumption of test independence applies, a serial Bayes model may provide guidance within contexts like those presented in the example just discussed.

When probabilities from a previous Bayes calculation are used to update estimates of the prior probability [P (A)], and when independence is confirmed, we can use a Bayes serial calculation to derive the probability that a patient has the disease given a second test result. Equation (4) presents the next step in the context of our example using a Bayes serial calculation for a second consecutive positive test under the assumption that the two tests are independent. With a second positive result, the probability of having the disease goes from .31 to .95, and our confidence in the diagnosis appears to improve dramatically. It should be noted that under the assumption of independence, parallel testing may also yield an outcome similar to serial testing. So, although the focus of this paper is on sequential or serially administered tests, when time or the occasion of the test is not an important factor in determining test independence, what is reported and discussed here may also apply to parallel testing.

$$P (A | B) = .90 * \mathbf{.31} / ((.90 * \mathbf{.31}) + (.02 * \mathbf{.69})) = .95 \quad (4)$$

From the outcome presented in Equation (4), it appears that the PPV of tests used to detect low prevalence diseases may be dramatically improved simply by administering the test a second or third time. However as mentioned, such positive outcomes rely on an independence assumption that is critical to the valid application of the serial Bayes probability model and implies that the error rate for each test is independent. Therefore, to determine whether an enhancement of PPV can be achieved by retesting, it is necessary to first establish the primary source(s) of test error and whether, or under what conditions, each medical test can be regarded as independent.

When a "gold standard" is available for determining the accuracy of a fallible test, establishing the independence between two test administrations is straight forward. One needs simply to twice calculate the specificity and sensitivity for the second test administration, once for the group of patients who test positive on the first test and once for the group of patients who tested negative on the first test. If the two calculations are in close agreement, the assumption of independence is satisfied. Unfortunately, a "gold standard" method for checking test accuracy is often not available, and other procedures are required.

Independence between test results can be achieved when clinicians randomly sample from the test-related variables that contribute to error and when each disease positive patient is equally likely to display a false negative test result and when each disease negative patient is equally likely to display a false positive test result. Indeed, when the conditions leading to test independence are understood, the utility of testing in a low prevalence disease context can often be dramatically enhanced by a simple random replication of a testing process that samples from the variables contributing to error. To ascertain under what conditions an independence assumption is satisfied, researchers must first investigate and understand the error structure of medical or psychological test outcomes. Given the potential for dramatically enhanced diagnostic accuracy, such research is critically important in improving the utility of certain tests with low PPV.

Within many testing contexts, it is often not possible to establish the accuracy of a fallible test by comparing it

to a more accurate "gold standard" testing methodology [2,3]. Although methods for estimating disease prevalence with the use of multiple tests, none of which are gold standard, have been developed [4], and latent class analysis has been used to estimate prevalence, specificity and sensitivity in the absense of a gold standard [5-7], there is little guidance for revising diagnostic predictions for a specific patient when evaluating multiple fallible test results. When a "gold standard" test procedure is unavailable, too risky, too invasive, and/or a violation of ethical research standards, an alternate and efficient method for establishing a test's error structure and the appropriateness of a serial Bayes-based revision of disease probability can be achieved using Generalizability (G) analysis [8]. The strength of G theory analysis is that it requires only a record of the fallible test outcomes and does not require "gold standard" testing. When outcomes for the fallible test are appropriately sampled and analyzed, precise quantification of relevant sources of error can be achieved.

Generalizability analysis is by far the most powerful method for quantifying sources of measurement error. In order to determine if, or under what conditions, a serial Bayes calculation is appropriate, a G study can analyze sampled test results and quantify and describes the error structure of a test. For example, in the context of medical testing, sources of error might be attributable to the laboratory, the clinician administering the test (e.g. psychiatric diagnosis), the occasion on which the test was administered, or some unobservable but consistent attribute of the patient. Each of these error sources can potentially lead to dependence between two tests results performed on a single patient and hence can be a source of dependent error leading to a violation of the independence assumption on which a Bayes serial testing model depends. To conduct a G study analysis, it is necessary to first collect test outcomes for randomly sampled administrations of the test. It is important to randomly sample across variables which naturally vary in the administration of a specific test and which might contribute to error in the test results. Such studies allow researchers to estimate each specified error source and establish whether, or under what conditions, a serial Bayes probability model appropriately updates patient probabilities upon retesting.

2. Methods

To illustrate how a G theory-based analysis might improve testing accuracy, let's further develop our example of the hypothetical medical test. Suppose that the medical test in the example problem involves a laboratory analysis of a specimen provided by a patient. Suppose further that a team of expert medical researchers identify three variables or potential sources of error over which the collection of test results tend to vary and which might be relevant to the outcome of the medical test in question. The first identified potential source of error concerns the occasion on which the patient is tested. Specifically, the researchers suspect that short-term fluctuations in patient values may lead to inconsistent test results. Hence, the test outcome may depend on when the patient was tested. For purposes of this illustration, we will designate this type of error as "Error Type 1". The second hypothesized source of error (Error Type 2) relates to an unobservable and temporally stable patient attribute that causes certain patients to be consistently more or less likely to generate false positive or false negative test results. The third identified error source (Error Type 3) is related to laboratory processing. In particular, the researchers hypothesized that variation in laboratory procedure may contribute to the generation of false negative or false positive test results.

In order to understand how these sources of error influence the test's outcome, the researchers design an experiment that samples test results from across the variables that tend to vary in the real world administration of the test within the population and that are hypothesized to contribute to error. The experiment draws a large random sample of patients from the clinical population of interest. Each patient in the sample is scheduled for multiple random appointment times at a clinic where specimens are collected. After each clinic visit, the collected specimen is divided into sub samples and sent for processing at multiple randomly selected laboratories. In G study terminology, the experiment's object of measurement is patient (p), and the two study variables over which sampling occurred, usually referred to as facets, are occasion (o) and laboratory (l). For purposes of analysis, the test's outcomes are analyzed using analysis of variance (ANOVA) with each cell containing the result of a single test outcome (*i.e.* either positive or negative, 0/1, or a continuous variable with or without a threshold value). Equation (5) displays the G study model for the decomposition of the total observed score variance $\sigma^2(X_{pol})$ into seven variance components that are estimated using ANOVA-based mean squares to derive estimates of the quantitative effects that compose a single test outcome (X_{pol}).

$$\sigma^2(X_{pol}) = \sigma^2(p) + \sigma^2(o) + \sigma^2(l) + \sigma^2(po) + \sigma^2(pl) + \sigma^2(ol) + \sigma^2(pol) \quad (5)$$

The ANOVA-based research model is a fully crossed person (p)-by-occasion (o)-by-laboratory (l) [$p \times o \times l$] random model. However, unlike typical ANOVA applications which focus on F tests and follow-up significance testing of certain mean effects within the model, G studies estimate variance components (VCs) for a single outcome and quantify all sources of variance. (It should

be noted that in some medical testing applications both ANOVA-based significance testing and VC estimation might prove useful.) This G study model estimates seven VCs. There are three main effects: *p*, *o* and *l*, and four interactions: *po*, *pl*, *ol*, and *pol*. It is useful here to consider what each VC conveys about the test results. The VCs can be verbally described as follows:

p – the degree to which test results tend to yield a consistent outcome for patients across occasions and laboratories (may contain Error Type 2),

o – the degree to which certain sequential occasions are more or less likely to detect a positive or negative result (contributes to Error Type 1, but in this example it should logically be zero),

l – the degree to which certain laboratories are more or less likely to detect positive or negative results (contributes to Error Type 3),

po – the degree to which a patient's test status tends to change depending on the occasion on which the sample was collected (contributes to Error Type 1),

pl – the degree to which a patient's test status tends to change depending on the particular lab to which the specimen was sent (contributes to Error Type 3),

ol – the degree to which the probability of a positive test result for a particular occasion tends to vary depending on the laboratory processing the specimen (this should logically be zero),

pol – the degree to which the patient's test status depends on the occasion / laboratory combination and other un-modeled or residual sources of error (also indicates the degree to which the G study model fails to capture relevant error sources).

3. Results

Interpreting the magnitude of the VCs from the G study can determine whether, or under what conditions, the assumption of test independence is satisfied and whether enhanced prediction upon retesting is achieved. The total variance in the G study model [$\sigma^2(X_{pol})$] is simply the sum of the all the variance components. Suppose in our example problem the test yields dichotomous data (negative/positive test results) and is summarized as the proportion of positive test results (ρ). Therefore, model variance is estimated as approximately: (ρ) * (1 – ρ); and hence is equal to the proportion of positive tests observed multiplied by the proportion of negative tests obtained across all tests in the sample. Hence, the first result of interest from the experimentally sampled data in our example problem is the proportion of positive test results observed within the sample. If the random sample is of adequate size it should yield an accurate estimate of the population proportion.

If, as in our example, there are established estimates of disease prevalence, and test specificity and sensitivity the researcher should examine the congruence between sample results and expected population values. Although a G study can productively proceed if a sample disagrees with established estimates of population prevalences and the test's specificity and sensitivity, for the purpose of simplicity in illustrating our example problem, let's assume that the proportion of positive tests obtained from our sample is in close agreement with the expected population proportion. Hence, 31% of patients testing positive in our sample reflect a patient's true positive disease status and 69% of the positive test results represent false positive results in a sample with a disease prevalence of .01. The expected proportion of positive results is .029 and the total model variance will sum to .0279.

To further illustrate, **Table 1** displays four hypothetical G study outcomes from the fully crossed G study model. Each outcome has a different practical implication related to achieving test independence. Assuming the sample is approximately consistent with expectations estimated from the population, let's focus on the VC outcomes in **Table 1**. For Outcome 1, 30% of the variance is found to be related to the patient and the largest source of the error is attributable to a patient by occasion interaction. Since the specificity, sensitivity and prevalence are known, Equation (3) suggests that in the absence of patient Error Type 2, patient variance will account for approximately 30% of the total variance in the model. Outcome 1 appears to agree closely with this expectation and hence would suggest Error Type 2 does not make a major contribution to measurement error. Further, since 60% of the variance is related to the person-by-occasion interaction (po), test results from a specimen collected on a single occasion and submitted to multiple laboratories are unlikely to provide independent information. The obvious recommendation resulting from Outcome 1 would be that to maximize the information from retesting and to insure that results exhibit test independence, specimens should be collected on different occasions.

As another illustration, let's suppose Outcome 2 as reported in **Table 1** was the G study result of our sampling experiment. Here patient variance is significantly higher than might be expected in the absence of Error Type 2. With 60% of overall variance attributed to patient variance, this outcome dramatically exceeds what one would expect in the absence of Error Type 2. In this situation, the practical implication is that a Bayes serial calculation would always be inappropriate even if specimens were collect on multiple occasions and sent to multiple labs. This result suggests that some patients are consistently more likely to generate false positive test results.

Outcome 3 in **Table 1** displays a G study outcome where most of the error is attributable to the patient-by-

Table 1. Percent of total variance for seven variance components for each of four hypothetical G study outcomes

Effect	VC Outcome 1	Outcome 1 %	Outcome 2 %	Outcome 3 %	Outcome 4 %
p	0084	30	60	30	30
o	.0000	0	0	0	0
l	.0000	0	10	2	2
po	.0167	60	2	2	2
pl	.0003	1	10	50	6
ol	.0006	2	8	0	0
pol	.0020	7	10	16	60
TOT	.0279	100	100	100	100

laboratory interaction (*pl*) (Error Type 3). To achieve enhanced prediction/accuracy through the use of serial testing, a single occasion would likely suffice as long as the specimen was sent to multiple laboratories. For Outcome 4, the three way interaction term (*pol*) explains most of the variance and illustrates a possible failure to specify and sample across relevant sources of error. Since the three-way interaction contains un-modeled error as well as the three-way interaction, this outcome may indicate that the variables investigated are not related to observed variation in test results.

4. Discussion

Although the testing problem presented within this hypothetical example focused on the interpretation of a hypothetical diagnostic biomedical test, G theory methodology coupled with Bayes serial estimations has much broader application. For example, many concerned constituents are currently attempting to assure the accurate and fair use of tests in employment, sports eligibility, and in making sanction decisions. In all of these contexts, issues of fairness have arisen due to the large proportion of false positive results and the high stakes nature of the test results. There is considerable interest in increasing the accuracy of test evidence for making important decision or a diagnosis. In addition, in many instances the data for such analyses may already exist since medical testing companies when seeking FDA approval for a particular test must submit the test to a series of trials.

It is obvious from governing board recommendations and from published legal advice that test users are aware that retesting might reduce error. However, recommendations for retesting are usually made without statistically estimation of the utility of retesting. Suggestions that samples be divided into multiple collection tubes, or that the test be repeated implies an expectation of increased precision with repeated testing. Unfortunately, when the sources of error are not systematically estimated, the usefulness of a particular retesting protocol is currently unknown.

REFERENCES

[1] D. L. Katz "Clinical Epidemiology and Evidence Based Medicine," Sage Publications, Inc., Thousand Oaks, 2001.

[2] G. A. Diamond, A. Rozanski, J. S. Forrester, D. Morris, B. H. Pollack, H. M. Staniloff, D. S. Berman and H. J. C. Swan, "A Model for Assessing the Sensitivity and Specificity of Tests Subject to Selection Bias," *Journal of Chronic Disease*, Vol. 39, No. 5, 1986, pp. 343-355.

[3] R. M. Henkelman, I. Kay and M. J. Bronskill, "Receiver Operating Characteristic (ROC) Analysis without Truth," *Medical Decision Making*, Vol. 10, No. 1, 1990, pp. 24-29.

[4] L. Joseph, T. W. Gyorkos and L. Coupal, "Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard," *American Journal of Epidemiology*, Vol. 141, No. 3, 1995, pp. 3546-3553.

[5] D. Rindskopf and W. Rindskopf, "The Value of Latent Class Analysis in Medical Diagnosis," *Statistics in Medicine*, Vol. 5, No. 1, 1986, pp. 21-27.

[6] T. A. Alonza and M. Pepe, "Using a Combination of Reference Tests to Assess the Accuracy of a New Diagnostic Test," *Statistics in Medicine*, Vol. 18, No. 22, 1999, pp. 2987-3003.

[7] S. V. Faraone and M.T. Tsuang, "Measuring Diagnostic Accuracy in the Absence of a 'Gold Standard'," *American Journal of Psychiatry*, Vol. 151, No. 5, 1994, pp. 650-657.

[8] R. L. Brennan, "Generalizability Theory," Springer Verlag, Inc., New York, 2001.

