Inferences for Likelihood Ratios in the Absence of a "Gold Standard"

LAWRENCE JOSEPH, PhD, THERESA W. GYORKOS, PhD

Likelihood ratios are extensively used to evaluate the performances of diagnostic tests and to update prior odds of disease to posttest odds. Since few tests are truly 100% accurate, including many used as "gold standards," it is important to be able to estimate likelihood ratios in cases where no such standard is available. In this paper, methods to calculate point and interval estimates for likelihood ratios are described. The results numerically coincide with those reviewed by Centor when a "gold standard" is assumed available, but typically provide wider interval estimates when such a standard is not available, reflecting the increased uncertainty inherent in such situations. Unlike previous techniques, the methods do not require normal approximations or logarithmic transformations, and hence provide accurate estimates even when parameter distributions are highly skewed. The methods are illustrated using the results of two different diagnostic tests for the presence of an intestinal parasitic infection. Key *words:* Bayesian analysis; diagnostic tests; epidemiologic methods: likelihood ratios; Monte Carlo methods; statistical models. **(Med Decis Making 1990;16:412-417)**

Consider two tests, where Test 1 is a "gold standard" for detection of a disease or condition, and Test 2 is an imperfect, but more convenient, cheaper, or easier test for that disease or condition. Estimating likelihood ratios for Test 2 can help assess its diagnostic@. Interval as well as point estimates for likelihood ratios can be useful. Interval estimates such as 95% confidence intervals (CIs) describe the accuracy of point estimates, and in providing a range of plausible values for likelihood ratios aid the clinician in interpreting diagnostic tests (e.g., tests with narrower CIs provide more precise posttest odds of disease than do those with wider CIs). As a result, use of interval rather than point estimates has been increasing rapidly in recent years.[1] Gart and Nam[2] provide a review of statistical methods for point and interval estimation of the likelihood ratio.

The standard estimation techniques for diagnostic test parameters from 2 X 2 table data such as those given in table 1 cannot be used when Test **1** is not a "gold standard." The sensitivity of Test 2, ***a/(a*** + ***b)***, cannot be determined, as the number of diseased subjects is unknown in the absence of a "gold standard." Similarly, the usual point estimates of the specificity and hence of the positive and negative likelihood ratios (denoted here by LR+ and LR-, respectively) do not hold. Walter and Irwig[3] summarize the standard frequentist or classic literature for statistical inferences for diagnostic test parameters in the absence of a "gold standard." Under this framework, unless there are data from three or more tests (for example, three or more different tests are applied to each subject, or the same test is applied on three or more independent instances), simultaneous estimation of all unknown parameters of the tests is not possible. Typically, some of the unknown parameters must be assumed exactly known in order to draw inferences about the remaining parameters. For example, the sensitivity and specificity of Test 1 must be assumed exactly known in order to estimate the disease prevalence and the sensitivity and specificity of Test 2. When Test **1** is a "gold standard," its sensitivity and specificity are 100%. Otherwise, it is rare that the sensitivity and specificity of Test 1 are exactly known. Fail-

Table 1 • Observed Data from Two Diagnostic Tests

		Test 1		
		+	−	
Test 2	+	a	c	a + c
	−	b	d	b + d
		a + b	c + d	N

Received April 18, 1995, from the Department of Epidemiology and Biostatistics, McGill University, and the Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, Montreal, P.Q., Canada. Revision accepted for publication December **13, 1995.** Supported in part by the Natural Sciences and Engineering Research Council. Dr. Joseph is a research scholar of the Fonds de la Recherche en Sante du Quebec, and Dr. Gyorkos is a research scholar supported by the National Health Research and Development Program, Health Canada.

Address correspondence and reprint requests to Dr. Joseph: Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec, H3G lA4, Canada. e-mail: (joseph@binky.ri.mgh.mcgill.ca).

ure to account for this uncertainty can lead to misleading point and interval estimates for likelihood ratios. This is illustrated in the examples below.

However, by taking a Bayesian approach, Joseph et al[4] demonstrated that it is possible to estimate all unknown parameters without imposing often-unjustified constraints on a subset of the unknown parameters. The results from the Bayesian approach reduce to those of the classical approach when similar constraints are imposed, and thus the Bayesian approach can be considered a generalization of the classical methods. Like previous solutions that required a "gold standard,"',' this numerical approach proceeds iteratively. The calculations are performed using a Markov chain Monte Carlo simulation technique called the Gibbs sampler," which has recently been applied to a wide variety of estimation problems in medicine.[7-10]

In this paper, the approach of Joseph et al.[4] is used to obtain interval estimates for LR+ and LR- from serologic testing and stool examination data, used as diagnostic tests for Strongyloides infection. Neither of these tests can be considered a "gold standard" for the detection of Strongyloides. Example **1** of Centor,[5] which examines the use of creatine kinase (CK) as a diagnostic test for myocardial infarction (MI), is used to illustrate that the interval estimates given by the Bayesian approach match those given by the standard approach when a "gold standard" test is available.

Methods

In the Bayesian approach to statistical inference, the information available before the experiment about the parameters of interest is summarized in a prior distribution. The data, through the likelihood function, are then combined with the prior distribution to derive posterior distributions using Bayes' theorem. The posterior distributions contain updated beliefs about the test parameters after considering the infermation provided by the data. Gelman et al.[10] provide an introduction to, and many examples of, Bayesian data analysis.

When two diagnostic tests are applied to each subject in a given population, there are typically five unknown parameters-the prevalence 03 and the sensitivities (Sn_1, Sn_2) and specificities (Sp_1, Sp_2) of each diagnostic test. Prior distributions that summarize the available information about these parameters must be formulated as inputs to the analysis. For example, if Test *1* is considered to be a "gold standard," $Sn_1 = Sp_1 = 1$, so that the prior probability distributions for these parameters consist of point masses on the single number **1,** and zero elsewhere. At the other extreme, if no prior information is considered to be available for P, Sn_2, and Sp_2, then uniform distributions covering the interval **[0,1]** (the range of possible values) could be used. Standard methods as reviewed by **Centor**[5] are approximately numerically equivalent to a Bayesian approach that assumes these five prior distributions. However, since "gold standards" are rarely available, and diagnostic tests would not be used unless at least some information was known about their properties a priori, most prior distributions should fall in between these extremes.

When no "gold standard" is available, the calculations required by Bayes' theorem are analytically intractable, and therefore Monte Carlo algorithms such as the Gibbs sampler are employed. The output of the Gibbs sampler consists of random samples from the joint posterior density of all parameters of interest. These random samples can then be used to reconstruct the marginal densities of each parameter, or more simply, to provide summaries of these densities, such as the means, standard deviations, and interval estimates.' Since any density shape can be reconstructed in this way, assumptions concerning normality or log-normality of the distributions are not necessary. Joseph et al.[4] provided details of the Gibbs-sampler algorithm that is required to provide random samples from the sensitivity, specificity, and positive and negative predictive values of diagnostic tests using a Bayesian approach when no "gold standard" is assumed. This approach is summarized in the appendix. Since likelihood ratios are functions of sensitivity (Sn) and specificity (Sp), a random sample from the marginal posterior density of a likelihood ratio can be constructed directly from a random sample of (Sn, Sp) pairs. Therefore, the results in the next two sections were obtained by using the Gibbs-sampler algorithm to obtain random (Sn_i, Sp_i) pairs, $i = 1, \ldots, M$, where M is the size of the Monte Carlo sample. A sample from the posterior density of each likelihood ratio can then be constructed by calculating the quantities

$$LR+_i = \frac{Sn_i}{1 - Sp_i}, \quad i = 1, \ldots, M$$

and

$$LR-_i = \frac{1 - Sn_i}{Sp_i}, \quad i = 1, \ldots M$$

These samples are then used to obtain mean, median, and interval estimates for the likelihood ratios.

Table 2 • Results of Serologic and Stool Testing for Strongyloides infection of 162 Cambodian Refugees Arriving in Montreal, Canada, between July 1962 and February 1983*

		Stool Examination +	–	
Serology	+	36	87	125
		2	35	37
		40	122	182

*Based on data from Gyorkos et al.[11,12]

Table 3 • Equally-tailed 95% Probability Ranges and Coefficients of the Beta Prior Densities for the Test Parameters in the Diagnosis of Strongyloides infection*

	Sensitivity	Specificity
Stool examination		
Range	5% to 45%	90% to 100%
Beta coefficients	α = 4.44, β = 13.31	α = 71.25, β = 3.75
Serology		
Range	65% to 95%	35% to 100%
Beta coefficients	α = 21.96, β = 5.49	α = 4.1, β = 1.76

*A uniform density over the range [0, I] was used for the prior distribution for the prevalence of Strongyloides infection in the refugee population.

Estimation of Likelihood Ratios in the Absence of a "Gold Standard"

Consider the data in table 2, collected from a survey of all Cambodian refugees who arrived in Montreal between July 1982 and February 1983.[11,12] Although both the serologic test and the stool examination are standard tools for determining the presence of Strongyloides infection, neither can be considered a "gold standard." This is because serologic testing generally has low specificity due to cross reactivityz[3] or persistence of reactivity following cure,[14] and stool examination has low sensitivity.[7] Furthermore, the sensitivities and specificities of these two diaghostic tests for Strongyloides are not known with high precision. Nevertheless, the likelihood ratios for both tests can be estimated from the data presented in table 2 and the available prior information about the sensitivities and specificities of the two tests.

In consultation with a panel of experienced parasitologists from the McGill University Centre for Tropical Diseases, 95% prior credible intervals (Bayesian analogs of CIs) were elicited by consensus. Beta distributions are commonly used to represent prior distributions when the data are dichotomous.[16] Therefore, the credible intervals were converted to beta distributions by matching the midpoints of the intervals to prior means, and four times the standard deviations to the credible set ranges. Using this method of moments gave close-fitting beta prior densities. The original intervals and associated beta densities are given in table 3. The mean of a beta distribution with parameters α and β is given by $\alpha/(\alpha + \beta)$. For example, the mean of the prior distribution for the sensitivity of stool examination is $4.44/(4.44 + 13.31) = 0.25$, as desired.

Likelihood ratio estimates from the results of the analysis for each test alone and the combination of the two tests are given in table 4. Another option would be to perform serial testing, which would improve the sensitivity of stool examinations, for example, although possibly at the expense of decreased specificity. As is evident from comparing the means and medians in table 4, the distributions can be highly skewed.

Estimation of Likelihood Ratios 'in the Presence of a "Gold Standard"

Consider the data in table 5, first presented by Radack et a1.[17] Using the methods described by Gart and Nam,[2] Centor[5] reports point estimates of **1.58 (95% CI 1.17-2.00)** and 0.69 (95% CI 0.49-0.90) for *LR* + and *LR*-, respectively. Applying the Bayesian approach described above and assuming MI to be a "gold standard" also yields point estimates (means of the marginal posterior densities1 of **1.58** and 0.69

Table 4 • Means, Medians, and Lower and Upper 95% Credible Interval Limits (CILs) of the Posterior Densities for Serologic Testing (Test 1) and Stool Examinations (Test 2), Alone and in Combination

Likelihood Ratio	Mean	Median	Lower 95% CIL	Upper 95% CIL
LR+ (test 1 positive)	4.57	2.73	1.36	18.17
LR- (test 1 negative)	0.19	0.17	0.06	0.43
LR+ (test 2 positive)	9.87	7.91	3.05	28.41
LR- (test 2 negative)	0.72	0.72	0.59	0.82
LR+ + (both tests positive)	44.19	23.92	8.60	188.82
LR+- (test 1 positive, test 2 negative)	3.34	1.98	0.88	13.51
LR-+ (test 1 negative, test 2 positive)	1.86	1.32	0.34	6.45
LR- – (both tests negative)	0.13	0.12	0.05	0.30

Table 5 • Observed Data for Evaluating Serum Creatine Kinase (CK) as a Diagnostic Test for Myocardial Infarction (MI)*

	MI +	MI –	
CK> 120	26	251	279
CK< 120	23	471	494
	51	722	773

*Based on data from Radack et al.[17]

for the *LR+* and *LR-*, respectively. Ninety-five percent equal-tailed credible sets are (1.17-2.00) and (0.49-0.91) for *LR+* and *LR-*, respectively. These intervals very closely match those given by Centor.[5]

Although assuming that MI is a "gold standard" for this data set is reasonable, it is instructive to note the effects on the likelihood ratios induced by variations in this assumption. For example, if the sensitivity and specificity of MI are known exactly to be only 95% (as may occur if misclassification is possible and the exact degree to which misclassification occurs is known), then the point estimates for the positive and negative likelihood ratios change dramatically to 2.39 (95% credible interval 1.32-2.97) and 0.25 (95% credible interval 0.01-0.82), respectively.

More realistically, there will be uncertainty regarding the exact degree to which misclassification occurs, so that, for example, the prior estimates may be again centered at 95%, but the 95% prior credible set ranges from 90% to 100% for each (due to an unknown degree of misclassification). In this case, the estimates become 2.14 (95% credible interval 1.29-2.91) and 0.39 (95% credible interval 0.02-0.84), respectively. Clearly, it is crucial to closely examine assumptions concerning the degree of perfection of any purported "gold standard" test in estimating likelihood ratios.

Discussion

As Centor[5] points out, likelihood ratios will probably become standard for evaluating test results, and proper interpretation of estimated likelihood ratios from data requires interval estimates.

Based on the results in this paper, it is evident that estimates of likelihood ratios can be highly dependent upon the assumptions made about the "gold standard," in that large deviations in LR estimates can occur when the sensitivity and specificity of the standard test decrease even slightly from 100%. This is important, since very few tests can be considered perfectly accurate, and even those that may be perfect in theory can be rendered less accurate by misclassification error or equipment failures.

When a "gold standard" is not available, it' is still possible to provide point and interval estimates for likelihood ratios, using Bayesian methods. These methods calculate the best possible estimates of test parameters given the data and the available prior information. Since in a Bayesian analysis final inferences depend on the prior distributions, providing several analyses starting from a range of reasonable prior distributions is usually desirable.[4,18] In addition, different laboratories may have different test sensitivities and specificities, depending on the available equipment and level of expertise.

The methods used to produce the results of the previous sections can be easily extended to include inferences about likelihood ratios from tests with more than two outcome categories. Correlated tests can also be considered by using Dirichelet prior densities over all test parameters and multinomial likelihoods rather than independent binomial likelihoods. Correlated tests may provide less information than the same number of independent tests, leading to likelihood ratios that are closer to one.

References

1. Gardner MJ, Altman DG (eds). Statistics with Confidence: Confidence Intervals and Statistical Guidelines. London, England: British Medical Journal, 1989.
2. Gart JJ, Nam J. Approximate interval estimation of the ratio of binomial parameters: a review and corrections for skewness. Biometrics. 1988;44:323-38.
3. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. J Clin Epidemiol. 1988;41:923-37.
4. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141:263-72.
5. Centor BM. Estimating confidence intervals of likelihood ratios. Med Decis Making. 1992;12:229-33.
6. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J Am Statist Assoc. 1990;85:398-409.
7. Gilks WR, Clayton DG, Spiegelhalter DJ, et al. Modelling complexity: applications of Gibbs sampling in medicine. J R Statist Soc B. 1993;1:39-52.
8. Coursaget P, Yvonnet B, Gilks WB, et al. Scheduling of revaccinations against hepatitis B virus. Lancet. 1991;337:1180-3.
9. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol. 1993;138:430-42.
10. Gelman A, Carlin J, Stern H, Rubin D. Bayesian Data Analysis. New York: Chapman and Hall, 1995.
11. Gyorkos TW, Genta RM, Viens P, MacLean JD. Seroepidemiology of Strongyloides infection in the Southeast Asian refugee population in Canada. Am J Epidemiol. 1990;132:257-64.
12. Gyorkos TW, Frappier-Davlgnon L, MacLean JD, et al. Effect of screening and treatment on imported intestinal parasite infections: results from a randomized, controlled trial. Am J Epidemiol. 1989;129:753-61.
13. Guyatt HL, Bundy DAP. Estimation of intestinal nematode

prevalence: influence of parasite mating patterns. Parasitology. 1993;107(Part 1):99-105.
14. Gam AA, Neva FA, Krotoski WA Comparative sensitivity and specificity of ELISA and II-IA for serodiagnosis of slrongyloidiasis with larval antigens. Am J Trap Med Hyg. 1987;37:157-61.
15. Genta RM. Predictive value of an enzyme-linked immunosorbent assay (ELISA) for the serodiagnosis of Strongyloides. Am J Clin Pathol. 1988;89:391-4.
16. Lee PM. Bayesian Statistics: An Introduction. London, U.K.: Edward Arnold, 1989.
17. Radack KL, Rouan G, Hedges J. The likelihood ratio: an improved measure for reporting and evaluating diagnostic test results. Arch Pathol Lab Med. 1986;110:689-93.
18. Spiegelhalter DJ, Freedman LS, Parmer MKB. Bayesian approaches to randomized trials. J R Statist Soc A. 1994;157: 357-416.

APPENDIX

Let data be collected as in table 1. Define unobserved "latent data" $Y_1, Y_2, Y_3,$ and Y_4 to represent the unknown numbers of true-positive subjects out of the observed cell values a, *b, c,* and *d,* respectively. Any subject, whether truly possessing the disease in question or not, can test positively or negatively on each test. Therefore, there are eight possible combinations, as summarized in table 6. Table 6 also provides the likelihood of the data assuming the tests are conditionally independent, that is, the tests are independent conditional on $Y_1, Y_2, Y_3,$ and Y_4. According to Bayes' theorem, the posterior distribution is proportional to the likelihood of the data times the prior distribution. Assume that independent beta distributions can be used to represent the prior information for the unknown parameters, as in table 3. In particular, denote the prior beta parameters for the prevalence π by α_π and β_π, and similarly denote the sensitivities (Sn_1, Sn_2) and specificities (Sp_1, Sp_2). Random samples from the appropriate marginal posterior densities can then be obtained by using the Gibbs sampler. In this case, the algorithm reduces to sampling in turn from the distributions below:

Table 6 • Likelihood Contributions of All Possible Combinations of Observed and Latent Data for the Case of Two Diagnostic Tests*

No. of Subjects	Truth	Test 1 Result	Test 2 Result	Likelihood Contribution
Y_1	+	+	+	$\pi Sn_1 Sn_2$
Y_2	+	+	−	$\pi Sn_1(1 - Sn_2)$
Y_3	+	−	+	$\pi(1 - Sn_1)Sn_2$
Y_4	+	−	−	$\pi(1 - Sn_1)(1 - Sn_2)$
$a - Y_1$	−	+	+	$(1 - \pi)(1 - Sp_1)(1 - Sp_2)$
$b - Y_2$	−	+	−	$(1 - \pi)(1 - Sp_1)Sp_2$
$c - Y_3$	−	−	+	$(1 - \pi)Sp_1(1 - Sp_2)$
$d - Y_4$	−	−	−	$(1 - \pi)Sp_1Sp_2$

*The complete likelihood is proportional to the product of each entry in the last column of the table raised to the power of the corresponding entry in the first column of the table.

$$Y_1 | a, \pi, Sn_1, Sp_1, Sn_2, Sp_2 \sim \text{binomial}\left(a, \frac{\pi Sn_1 Sn_2}{\pi Sn_1 Sn_2 + (1-\pi)(1-Sp_1)(1-Sp_2)}\right) \quad (1)$$

$$Y_2 | b, \pi, Sn_1, Sp_1, Sn_2, Sp_2 \sim \text{binomial}\left(b, \frac{\pi Sn_1(1-Sn_2)}{\pi Sn_1(1-Sn_2) + (1-\pi)(1-Sp_1)Sp_2}\right) \quad (2)$$

$$Y_3 | c, \pi, Sn_1, Sp_1, Sn_2, Sp_2 \sim \text{binomial}\left(c, \frac{\pi(1-Sn_1)Sn_2}{\pi(1-Sn_1)Sn_2 + (1-\pi)Sp_1(1-Sp_2)}\right) \quad (3)$$

$$Y_4 | d, \pi, Sn_1, Sp_1, Sn_2, Sp_2 \sim \text{binomial}\left(d, \frac{\pi(1-Sn_1)(1-Sn_2)}{\pi(1-Sn_1)(1-Sn_2) + (1-\pi)Sp_1Sp_2}\right) \quad (4)$$

$$\pi | a, b, c, d, Y_1, Y_2, Y_3, Y_4, \alpha_\pi, \beta_\pi \sim \text{beta}(Y_1 + Y_2 + Y_3 + Y_4 + \alpha_\pi, N - (Y_1 + Y_2 + Y_3 + Y_4) + \beta_\pi) \quad (5)$$

$$Sn_1 | Y_1, Y_2, Y_3, Y_4, \alpha_{Sn1}, \beta_{Sn1} \sim \text{beta}(Y_1 + Y_2 + \alpha_{Sn1}, Y_3 + Y_4 + \beta_{Sn1}) \quad (6)$$

$$Sp_1 | a, b, c, d, Y_1, Y_2, Y_3, Y_4, \alpha_{Sp1}, \beta_{Sp1} \sim \text{beta}(c + d - (Y_3 + Y_4) + \alpha_{Sp1}, a + b - (Y_1 + Y_2) + \beta_{Sp1}) \quad (7)$$

$$Sn_2 | Y_1, Y_2, Y_3, Y_4, \alpha_{Sn2}, \beta_{Sn2} \sim \text{beta}(Y_1 + Y_3 + \alpha_{Sn2}, Y_2 + Y_4 + \beta_{Sn2}) \quad (8)$$

$$Sp_2 | a, b, c, d, Y_1, Y_2, Y_3, Y_4, \alpha_{Sp2}, \beta_{Sp2} \sim \text{beta}(b + d - (Y_2 + Y_4) + \alpha_{Sp2}, a + c - (Y_1 + Y_3) + \beta_{Sp2}) \quad (9)$$

To start the algorithm, Y_1, Y_2, Y_3, and Y_4 are randomly generated from binomial distributions 1 through 4, respectively, given arbitrary starting values for the other parameters. Then π is generated from the beta density 5 conditional on the Y_1 through Y_4 variates just sampled. Drawing Sn_1, Sp_1, Sn_2, and Sp_2 from beta densities given in expressions 6 through 9, respectively, using the same values of Y_1 through Y_4, completes the first iteration of the Gibbs sampler algorithm. The next iteration begins with drawing Y_1 from equation 1, using the new values for Sn_1, Sp_1, Sn_2, and Sp_2, and so on. The random samples generated by repeating the above cycle a large number of times are then used to reconstruct the marginal posterior densities of each parameter, and to find credible intervals, marginal posterior means or medians, or other inferences. To obtain the results in this paper, the algorithm was run for 20,000 iterations. The results from the first 500 iterations were used to assess convergence of the algorithm, and the remaining 19,500 were used for inferences. Similar methdds can be derived for the cases when only one test or three or more tests are applied. Full details are available in Joseph et al.[5] A computer program written in the S-PLUS statistical programming language for the cases of one, two, and three diagnostic tests is available upon request from the authors.

DECISION TREE CONSTRUCTION:

Guidelines for authors

Authors are requested to use the following guidelines in the construction of decision trees.

1. Standard notation:

   0 = chance node

   □ = choice node

   100 = quantitative outcome (utilities) inside rectangle

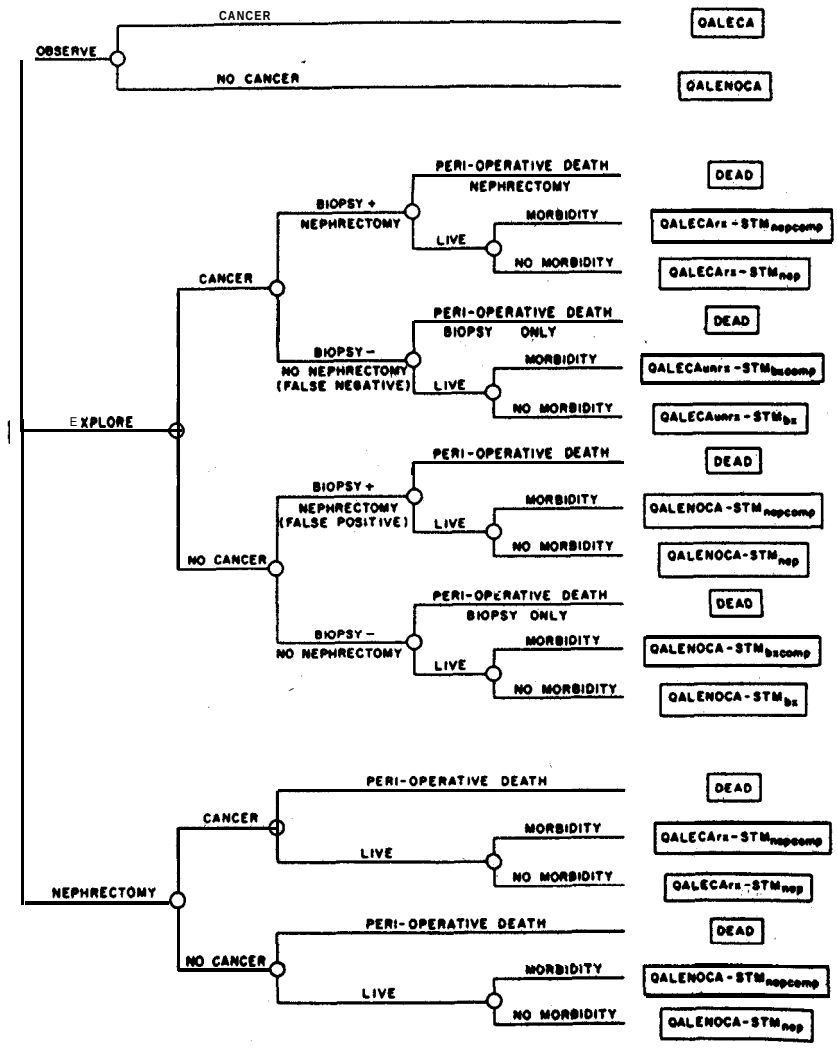
2. Use right angles at nodes

3. Analogous events should line up vertically •

4. If the tree is complicated, subtrees may be used

5. A decision tree should be oriented horizontally on an MDM page: available space is 7 inches wide by 9.5 inches long

6. Letters on the tree must be legible when it is printed on a single page; minimum point size 6 points; make lettering uniform in size and weight (legend will be set by the publisher)



An *example of a well-constructed tree*