

# An Epidemiologic Critique of Current Microbial Risk Assessment Practices: The Importance of Prevalence and Test Accuracy Data

IAN A. GARDNER\*

Department of Medicine and Epidemiology, School of Veterinary Medicine, One Shields Avenue, University of California, Davis, California 95616, USA

MS 03-704: Received 25 August 2003/Accepted 12 March 2004

## ABSTRACT

Data deficiencies are impeding the development and validation of microbial risk assessment models. One such deficiency is the failure to adjust test-based (apparent) prevalence estimates to true prevalence estimates by correcting for the imperfect accuracy of tests that are used. Such adjustments will facilitate comparability of data from different populations and from the same population over time as tests change and the unbiased quantification of effects of mitigation strategies. True prevalence can be estimated from apparent prevalence using frequentist and Bayesian methods, but the latter are more flexible and can incorporate uncertainty in test accuracy and prior prevalence data. Both approaches can be used for single or multiple populations, but the Bayesian approach can better deal with clustered data, inferences for rare events, and uncertainty in multiple variables. Examples of prevalence inferences based on results of *Salmonella* culture are presented. The opportunity to adjust test-based prevalence estimates is predicated on the availability of sensitivity and specificity estimates. These estimates can be obtained from studies using archived gold standard (reference) samples, by screening with the new test and follow-up of test-positive and test-negative samples with a gold standard test, and by use of latent class methods, which make no assumptions about the true status of each sampling unit. Latent class analysis can be done with maximum likelihood and Bayesian methods, and an example of their use in the evaluation of tests for *Toxoplasma gondii* in pigs is presented. Guidelines are proposed for more transparent incorporation of test data into microbial risk assessments.

The fields of predictive microbiologic and dose-response modeling of foodborne pathogens have made substantial methodological advances in the last 5 years. Although there are areas of modeling that require refinement and further development, it is clear from an epidemiologic perspective that the more important limitation is the paucity of valid empirical data for model development and validation of models under naturally occurring conditions. Essentially, the concept of coevolution of models and data has failed.

Deficiencies exist in the amount, completeness, validity, relevance, and comparability of data used in microbial risk assessments. Risk assessors have expressed a need for high-quality, relevant, and timely data when public policy decisions hinge on findings from risk analyses. The issue of comparability of data from different sources and different studies has received less attention in the discussion of data deficiencies perhaps because there are so few data that meet the threshold for inclusion in microbial risk assessments. Noncomparability occurs when different tests and sampling schemes are used in studies. Lack of comparability may induce bias in the risk estimates if the results from the studies are treated as if the tests, sampling schemes, and population characteristics were the same.

Although much of the historic emphasis in microbial

risk assessment has been on uncertainty and variability in the collected data and how these phenomena are best modeled, the topic of data bias has been given only limited consideration. Here, I focus on bias in prevalence estimates attributable to the use of imperfect tests and describe the techniques available to risk analysts for adjustment of test-based estimates to true prevalence estimates. I also describe methods for evaluation of test accuracy in the absence of a gold standard reference test that provide a useful alternative approach to traditional methods. I demonstrate the utility of Bayesian methods and Markov chain Monte Carlo simulation for prevalence inferences for rare events and clustered data obtained from single and multiple populations and for estimation of the accuracy of correlated diagnostic tests.

## PREVALENCE ESTIMATION

Prevalence and concentration data are considered central to all microbial risk assessments regardless of the type of microbe that is evaluated (26). For microbial risk assessments, true prevalence estimates are preferable to test-based estimates because the former make data more comparable from different populations and from the same population over time. To convert test-based estimates to true prevalence estimates, estimates of test accuracy are needed. The usual practice by risk assessors has been to consider test-based (apparent) prevalence data to be equivalent to

\* Author for correspondence. Tel: 530-752-6992; Fax: 530-752-0414; E-mail: iagardner@ucdavis.edu.

true prevalence data. This assumption apparently was made in four recent risk assessments for bacterial pathogens (1, 2, 6, 39), and the author of an article outlining applications of quantitative risk assessment to food safety (37) did not mention the effects of test imperfections. In contrast, other authors have discussed the limitations and biases attributable to the tests and the implications for the results of these tests (28). To my knowledge, there are no guidelines or standards about how to describe and incorporate test-based data in an exposure assessment, although there is a need for increased transparency especially if international trade decisions are involved (22). Perhaps, because the data are scarce there has not been a pressing need to directly address this issue.

Although I have focused here on adjustment for imperfect sensitivity and specificity, many of the issues faced by risk assessors are more complex and involve prevalence data from other countries and regions, obsolete data before an intervention strategy, prevalence data collected many years ago, and seasonal effects. These factors induce additional and sometimes greater uncertainty in the estimates than that attributable to test imperfections.

The use of test-based data may be adequate to meet the objectives of some risk assessments. One example is the risk ranking of certain food types based on the frequency of isolation of pathogens such as *Salmonella* and *Listeria* spp., where the same test and test protocol are used on all food items. Even in these circumstances, the data might not be exactly comparable if the competing microflora are different and a culture detection method is used. A second example is a longitudinal study of the farm-to-table continuum, in which animals, carcasses, and meat products are tested using the same test and the goal is to compare prevalence estimates. A third scenario is a randomized trial where the effect of a mitigation on the incidence of *Salmonella* is evaluated. In all these cases, ranking or the use of ratio estimates is valid if a perfectly specific test is used. Consider the third scenario where the true risks of *Salmonella* contamination in the unmitigated and mitigated groups are 10 and 2%, respectively. If a *Salmonella* culture technique with sensitivity of 50% were used, the expected estimates in the trial would be 5 and 1%, respectively. The risk ratio estimate of 5 is unbiased, but the risk difference of 4% is a twofold underestimate of the true value.

Even when true prevalence is calculated from apparent prevalence, there is a need for improved inferences when infection or contamination is rare (0 or a few positives in  $n$  samples) and when pathogens cluster. The clustering issue has been recognized by epidemiologists at the farm level (e.g., infected and noninfected farms exist, and within infected farms prevalence in animals moved to slaughter is variable), and the idea has also been raised in the context of food products (30).

Prevalence data from surveys usually are analyzed with frequentist statistical methods, but recently there has been increasing use of Bayesian methods when incorporation of existing knowledge is appropriate and it is desirable to capture the uncertainty in estimates. In the following sections,

I describe prevalence inferences from individual samples from single and multiple clusters of sampling units. The basic principles apply also to pooled sampling techniques, which are sometimes used when infection or contamination is rare. For pooled sampling, the precision of estimates can be comparable to that of individual sample testing when the sensitivities of pooled and individual tests are equal. Methods for estimation of individual prevalence from pool prevalence have been described with an application to eggs contaminated with *Salmonella* Enteritidis (9).

**Frequentist inferences for a single population.** Assume that the presence or absence of a pathogen is denoted as  $I+$  or  $I-$ , respectively, and that a binary test ( $T$ ) is used for pathogen detection. If the test result is measured on a continuous or ordinal scale, a threshold value (cutoff) is needed for designation of a test positive ( $T+$ ) or test negative ( $T-$ ) result. The accuracy of a diagnostic test is usually measured by its sensitivity and specificity. Sensitivity,  $\text{Pr}(T+|I+)$ , is the probability that the test is positive ( $T+$ ) when an individual sampling unit (e.g., animal, carcass, or food product) is infected or contaminated ( $I+$ ). Specificity,  $\text{Pr}(T-|I-)$ , is the probability that the test is negative ( $T-$ ) when an individual sampling unit is not infected or not contaminated ( $I-$ ). The true prevalence of infection or contamination,  $\text{Pr}(I+)$ , can be estimated from test (apparent) prevalence,  $\text{Pr}(T+)$ , using a correction factor that adjusts for imperfect sensitivity and specificity of the test (33):

$$\text{Pr}(I+) = [\text{Pr}(T+) + \text{Pr}(T-|I-) - 1] / [\text{Pr}(T+|I+) + \text{Pr}(T-|I-) - 1] \quad (1)$$

When the test is perfectly specific, e.g., culture of viruses and bacteria, equation 1 simplifies to

$$\text{Pr}(I+) = \text{Pr}(T+) / \text{Pr}(T+|I+) \quad (2)$$

Point estimates for true prevalence are readily obtained using the available sensitivity and specificity estimates. For example, when estimates from published studies indicate that the sensitivity of *Salmonella* culture is 0.5, then it is straightforward to adjust an apparent prevalence estimate of 0.2 from a survey, in which 20 of 100 culture-positive samples were obtained, to a true prevalence estimate of 0.4. However, there is uncertainty in sensitivity and specificity estimates, and the sampling process has induced additional variability. The variance of the prevalence estimator in equation 1 is a function of the total sample size for the prevalence survey, the magnitude of the sensitivity and specificity estimates, and the numbers of infected and noninfected units in the original test evaluation study (16).

This variability can be incorporated in a frequentist confidence interval (CI) (16), but large-sample normal approximations may fail when sample sizes are small or prevalence is close to 0 or 1. In the latter case, confidence limits may exceed these values, which is inconsistent with the laws of probability. When prevalence estimates are close to 0 or 1, exact binomial limits should be used, but such limits do not readily allow for the incorporation of data on sensitivity and specificity.

**Bayesian inferences for a single population.** Bayes's theorem provides the basis of rational probabilistic thinking about many scientific problems (17, 27), including risk analysis (25). The Bayesian approach allows the analyst to combine prior data and/or expert opinion about prevalence and test accuracy (and the uncertainty associated with each value) with current survey data (usually termed the likelihood) to produce updated posterior inferences about prevalence. Gibbs sampling, an iterative Markov chain Monte Carlo simulation method (5, 32), is needed for prevalence inferences because there is no explicit mathematical solution to the problem.

An advantage of the Bayesian framework over traditional frequentist statistical methods is that it allows great flexibility and generality in coping with complex problems, including those with many parameters such as inferences about animal pathogens in the presence of clustering (36). The lack of easy-to-use software has been mostly resolved by the development of WinBUGS (40), which allows for easy implementation of Markov chain Monte Carlo simulation techniques such as the Gibbs sampler and the Metropolis algorithm (32).

Selection of an appropriate prior probability distribution is an important component of a Bayesian analysis. The beta distribution is appropriate for modeling binomial probabilities such as sensitivity, specificity, and prevalence (38). An appropriate beta ( $\alpha$ ,  $\beta$ ) distribution can be constructed using data from previous studies with the following relationship:  $\alpha$  = number of positives in  $n$  trials + 1, and  $\beta$  = number of negatives in  $n$  trials + 1 (38). For example, if a preliminary evaluation of the sensitivity of fecal culture for *Salmonella* found 7 test positives of 14 infected, then the beta (8, 8) distribution would be appropriate for modeling uncertainty about test sensitivity provided the animals or sampling units tested were similar in both studies. An alternative approach is to elicit expert opinion about test sensitivity by asking an expert for the best estimate, which is equated to the mode of the beta ( $\alpha$ ,  $\beta$ ) distribution, and asking for an interval in which the expert strongly believes the true value lies. When the upper and lower endpoints of the interval are equated to the 5th and 95th percentiles of a corresponding beta distribution, it is straightforward to find appropriate values for  $\alpha$  and  $\beta$  (36). The corresponding beta density is drawn and verified by an expert. A software program (BetaBuster) to implement these methods is available ([www.epi.ucdavis.edu/diagnostictests/](http://www.epi.ucdavis.edu/diagnostictests/)). In the example, a beta (8, 8) prior would be obtained if the expert indicated that the modal value was 0.5 and the 5th and 95th percentiles were 0.3 and 0.7, respectively. Triangular distributions, which are sometimes used for representing expert opinion (38), are not amenable to Gibbs sampling because they are not conjugate priors for binomial likelihoods.

**Bayesian inferences for a single population: example.** I considered a scenario that is typical of the type of inference that is needed: low prevalence or zero positive results in  $n$  samples. Suppose that feces from 100 broilers in a shed of 10,000 were cultured for *Salmonella* and none were culture positive. The sensitivity and specificity of cul-

ture are assumed to be 50 and 100%, respectively. In a frequentist analysis with a point estimate of zero, the exact upper 95% confidence limit is 3.6%, and the upper limit could be doubled to yield an upper limit of 7.2% that allowed for the fact that test sensitivity was known to be exactly 50%.

This analysis is hampered by its inability to readily incorporate uncertainty in test accuracy. To conduct a Bayesian analysis of these data, I used a beta (8, 8) for sensitivity (as described for specification of a prior distribution based on experimental data or expert opinion), a beta (9999, 1) for specificity to allow for false-positive results about 1 time in 10,000, and a noninformative prior, beta (1, 1), for prevalence to allow comparison with the frequentist analysis. WinBUGS was used for this analysis with a modification of published code (4). Inferences were based on 50,000 iterations after discarding an initial 1,000 samples as burn-in, i.e., the number of samples needed to reach convergence on the target distribution. With these inputs, the median and mean estimates of prevalence were 1.5 and 2.3%, respectively, with a 95% probability interval (PI) of 0.05 to 9.1%. This interval is comparable to but somewhat wider than the frequentist interval but does not include zero prevalence as a possible value.

If a prior study had been done on this farm, say a month ago with the same result (0 of 100), use of an informative beta (1, 101) prior for prevalence might be justified if the *Salmonella* prevalence were considered to be similar at the two sampling times. Use of this informative prior for prevalence results in posterior estimates that were approximately threefold lower; median and mean estimates were 0.5 and 0.7%, respectively, with a 95% PI of 0.02 to 2.4%.

One limitation of the above model is that it fails to account for the possibility that the tested shed of birds is free of *Salmonella*. The model can be modified by allowing prevalence to have a mixture distribution with point mass at zero with positive probability ( $\theta$ ) and a beta distribution for prevalence with probability  $1 - \theta$  (20). WinBUGS code can be readily changed for this new model (4). Assuming that the prior probability that the flock was free of infection was fixed at 50%, the posterior probability of freedom from *Salmonella* infection was estimated to be 98.0% with the same noninformative (beta 1, 1) prior for within-flock prevalence. Although a posterior distribution for prevalence (median = 49%, 95% PI = 1.5 to 97.5%) is generated from this analysis, there is essentially no information in the data to update the prior. In this model, final inferences depend strongly on the prior on within-flock prevalence. Substitution of the informative beta (1, 101) prior results in a smaller posterior probability (59.7%) because discrimination between absence of infection and low prevalence is more difficult. For this scenario, the posterior distribution of within-flock prevalence has median of 0.6% with a 95% PI of 0.2 to 3.2%.

**Frequentist and Bayesian inferences for multiple populations.** Cluster sampling designs are used commonly to study animal and foodborne pathogens. Two-stage sam-

pling that involves random selection of clusters (e.g., herds or flocks) and then random selection of animals in a population to detect a minimum threshold prevalence with high confidence is typical of cluster designs. The usual approach for interpretation of test results across multiple clusters is to define a threshold number of positive samples (usually one or more) needed to designate the cluster as positive and to generate an estimate of the proportion of positive clusters (7, 29). Within-cluster true prevalence for clusters with positive test results is estimated using equation 1. As in the single population model, the major limitation of a frequentist analysis is that it fails to capture the uncertainty in test accuracy or allow for the incorporation of prior data, if available. Negative confidence limits for within-cluster prevalence can result.

Because of the limitations of the frequentist approach, a Bayesian hierarchical model has been developed to analyze cluster-level test results and currently is being assessed for certification of pathogen-free status in animal populations (36). In the model, three levels of inference are possible: animal, herd, and aggregate of herds in a state or region or under common ownership. This model provides an extension of the single population model, and Gibbs and adaptive rejection sampling are needed to obtain an efficient solution (36).

**Frequentist and Bayesian inferences for multiple populations: example.** Suppose that the *Salmonella* study was repeated in 100 flocks, with 10 birds tested in each flock. Assume that 90 flocks had completely negative culture results and 10 flocks had culture-positive samples (4 flocks with one positive sample, 4 flocks with two positive samples, and 2 flocks with three positive samples). The selected sample size of 10 birds per flock would detect a within-flock prevalence of at least 30% with 95% confidence, assuming a perfect test. A frequentist analysis of these data yields an estimated proportion of positive flocks of 10% with an exact 95% CI of 4.9 to 17.6%. A frequentist CI that accounts for the imperfect and uncertain sensitivity of fecal culture cannot be readily generated, although it is intuitive that the true proportion of infected flocks is underestimated because of low herd sensitivity (7).

In contrast, a Bayesian analysis of these data can readily deal with its complexities. Beta (8, 8) and beta (9999, 1) are used for sensitivity and specificity distributions, respectively, as before, and the prior for within-flock prevalence is set to be weakly informative (beta 1.9, 4.61). This prior was based on expert opinion that the most likely prevalence value for infected flocks was 20% and the expert was 95% sure that prevalences in similar flocks were less than 60%. The proportion of infected flocks was modeled noninformatively with a beta (1, 1) prior. As in the single-flock scenario, prevalences are allowed to have a mixture distribution with point mass at zero and a continuous beta distribution on (0, 1) (4). Modeling involves use of prior distributions for the mean prevalence and the variability of prevalences about their mean (4, 19). Beta and gamma distributions, respectively, are used for these two parameters, and generic WinBUGS code for this problem is available

(4). The model allows estimation of the flock-level prevalence, the prevalence distribution for infected flocks, the probability that a randomly selected flock is infection free, and the predictive distribution for a new randomly selected flock from the same source population. I focus here on inferences for the first two outcomes. Based on these prior specifications and 50,000 iterations following a burn-in period of 1,000, WinBUGS estimated the median flock prevalence to be 17.0% with a 95% PI from 8 to 50.9%. Because there are 10 culture-positive flocks, updated inferences about within-flock prevalence are of interest. The posterior distribution for within-flock prevalence has a median of 23% with a 95% PI of 0.4 to 66.7%, which is consistent with the prior distribution.

Next, assume that the *Salmonella* study was repeated in the another group of 100 flocks and all 1,000 samples were culture negative. For this scenario, the estimate of primary interest might be the proportion of *Salmonella*-positive flocks. The median estimate from the WinBUGS analysis of these data using the same priors is 1.3% with a 95% PI of 0.4 to 14.9%. This result represents an almost 40-fold decrease compared with the prior median of 50%.

#### EVALUATION OF ACCURACY OF DIAGNOSTIC TESTS

The use of diagnostic tests (and interpretation of results obtained from these tests) is integral to much of the data collection related to foodborne pathogens. Low cost, ease of performance, and speed with which results can be obtained are important considerations for tests used in food-producing industries in which products are rapidly processed and distributed. Performance characteristics (sensitivity, specificity, and reproducibility) are usually of lesser concern, but this information is vital for informed interpretation of test results (12) and when test data from different sources are to be compared, used, and perhaps combined in a microbial risk assessment. The concepts of diagnostic sensitivity and specificity are related to but different from those of analytical sensitivity and specificity (34).

Estimates of test accuracy would be useful in a risk assessment for three main reasons: to improve comparability of data from different sources, for appropriate interpretation of cluster-level results, and for improved assessments of the effects of mitigations. Diagnostic accuracy using field samples can be evaluated by several methods, including the use of banked gold-standard (reference) samples, use of a new screening test and then verification of all (or a random sample of) test-positive and test-negative samples with the gold standard method, or use of a cross-sectional design in which the new test and the gold-standard test are performed on all samples. In all three situations, estimation of sensitivity and specificity and their 95% CIs is straightforward (15). Epidemiologic issues in the design of test validation studies have been discussed in detail elsewhere (15), and guidelines for validation of serologic tests have been described (23).

There is a paucity of sensitivity and specificity estimates for tests that are used commonly to detect foodborne pathogens in all foodstuffs, animals, and the environment.

Possible reasons for this lack of information include the availability of numerous tests and multiple purposes for which they can be used, the lack of universally accepted validation standards, and the nontrivial task of designing a scientifically valid study. The need for harmonized microbiologic analyses and specifically standardized approaches at the regional and international levels has gained increasing acceptance (18). For example, consider the evaluation of a new culture method for *Salmonella*. Factors affecting the sensitivity of culture include the number of organisms on a weight or volume basis, weight or volume of material tested, sample type (e.g., feces, tissues, and milk), presence and concentration of competing microflora, and laboratory techniques (e.g., preenrichment, enrichment, and plating methods). In designing a validation study, an investigator must make critical decisions about how many of these factors or combinations thereof will be assessed with available financial resources. One set of validation standards cannot be applied to all diagnostic tests and their possible applications.

Even when validated tests exist, estimates are imprecise because usually only a limited number of samples are evaluated. For some tests, such as isolation of bacteria, viruses, and parasites, an assumption of perfect specificity is justifiable in laboratories with experienced personnel and good quality-assurance protocols. In all other situations, tests should be assumed to have imperfect sensitivity and specificity until proven otherwise.

Lack of bias in sensitivity and specificity estimates is considered by most authors to be more important than imprecision in the estimates. Bias often results from flaws in the design of the study used for test evaluation and includes use of an inappropriate reference test (gold standard), unblinded evaluation of test results, and use of an inappropriate spectrum of diseased individuals (15). In the context of microbial pathogens and food products, "diseased individuals" equates to the microbe concentrations in the sampled products.

#### Evaluation of test accuracy without a gold standard.

Recently, there has been increased interest within the field of veterinary medicine in the use of latent-class methods for test evaluation. These methods do not require designation of a perfectly sensitive and specific reference test, i.e., a gold standard. A maximum likelihood (ML) approach was developed about 20 years ago (21), but it has had limited use in this context until the last decade. In part, this lack of adoption might be attributable to skepticism by medical researchers, journal editors, and reviewers about its black-box approach and their preference for standard methods that are based on the concept of a gold standard test.

Although methods exist for evaluation of multiple populations and multiple tests (3, 19), I describe here the simplest case (two tests in two populations) that is amenable to both ML and Bayesian solutions.

When two binary tests are used to determine the status of all sampling units in two populations with different prevalences, the observed count data can be represented in two  $2 \times 2$  tables (Fig. 1). Subject to the assumptions inherent

		Test 2	
		+	-
Test 1	+	a	b
	-	c	d
		Test 2	
		+	-
Test 1	+	e	f
	-	g	h

FIGURE 1. Test results classified as positive or negative, cross-classified in two  $2 \times 2$  tables, for each sampling unit in population 1 and population 2.

in the model, there are sufficient degrees of freedom to estimate six parameters: two sensitivities, two specificities, and two prevalences. Assumptions that are made in the two-test two-population model are constant sensitivity and specificity in both populations and conditional independence of the sensitivities and specificities of the two tests. Concepts of conditional independence and dependence have been described elsewhere (13). An assumption of conditional independence for two tests is reasonable when the tests measure different biologic phenomena, e.g., one test detects antibody and the other detects antigen.

ML estimates are obtained by maximizing the likelihood function and are the values that are most consistent with the observed data, in this case the counts in the eight cells (a to h) in Figure 1. Two algorithms are available for ML estimation: Newton-Raphson and expectation-maximization (11). These methods have been implemented in a Web-based program TAGS, which allows the user to readily obtain these estimates and associated CIs for conditionally independent tests (31).

The Bayesian approach essentially follows the procedure used for prevalence estimation. Prior scientific knowledge about test accuracy and prevalence, if available, is linked with information contained in the likelihood based on the observed data (Fig. 1). Means, medians, and 95% PIs are used to summarize posterior inferences about the six model parameters (11).

**Evaluation of test accuracy without a gold standard: example.** Five serologic tests for screening pigs for toxoplasmosis (*Toxoplasma gondii* infection) were evaluated using heart muscle and serum for 1,000 sows (10). The samples were collected in two batches: 1 to 463, and 464 to 1,000. The gold standard used by the authors was based on a combination of results of bioassay of heart muscle in mice (all samples) and bioassay in *T. gondii*-free cats (183 samples). When either bioassay result was positive,

**Batch 1**

		Bioassay		
		+	-	
MAT	+	37	55	92
	-	7	364	371
		44	419	463

**Batch 2**

		Bioassay		
		+	-	
MAT	+	104	26	130
	-	22	385	407
		126	411	537

FIGURE 2. Cross-classified results of modified agglutination (MAT) testing of serum and bioassay of heart muscle for 1,000 sows (10).

**Batch 1**

		ELISA		
		+	-	
MAT	+	67	25	92
	-	41	329	370
		108	354	462

**Batch 2**

		ELISA		
		+	-	
MAT	+	97	33	130
	-	36	371	407
		133	404	537

FIGURE 3. Cross-classified results of modified agglutination (MAT) and enzyme-linked immunosorbent assay (ELISA) testing of serum for 999 sows (10).

sows were considered *T. gondii* infected (170 samples), otherwise they were considered noninfected. The two most accurate tests of those studied were the modified agglutination test (MAT), which had a sensitivity of 82.9% and a specificity of 90.2% at a cutoff value of 1:20, and the enzyme-linked immunosorbent assay (ELISA), which had a sensitivity of 73.0% and a specificity of 85.9% at a cutoff value of 0.36.

The original data set (10) was used to demonstrate the ability of ML methods to obtain valid estimates. First, the true infection status information was removed from the data file. Second, two populations were created using the batch information, and the test results were cross-classified (Fig. 2). Third, these cell counts were entered into TAGS (31), and the following estimates were obtained for the MAT: sensitivity = 82.9% (95% CI = 76.5 to 87.9%) and specificity = 90.2% (95% CI = 88.0 to 92.1%). The bioassay was estimated to be perfectly sensitive and specific, which provided justification for its selection as the gold standard. The assumption of constant test accuracy in both populations was reasonable because *T. gondii* infections are sub-clinical and sows were randomly sampled. In addition, the assumption of conditional independence was biologically justified because the bioassay detects viable cysts and the MAT measures serum antibodies produced in response to the presence of cysts in body tissues. For this example, there is no ambiguity in interpretation of the latent class because of the chronic nature of *T. gondii* infections and the fact that antigen and antibody occur together. This level of certainty would not necessarily apply for comparison of results between culture and a serologic test for *Salmonella* because infections may be transient and antibodies may persist in the absence of infection for a variable period. Prevalences are also estimable with this model (although not of

primary interest) and were 9.5 and 23.5% for batches 1 and 2, respectively.

Suppose now that the goal was to obtain estimates of the sensitivity and specificity of the MAT and ELISA using the same data but in the absence of a gold standard. For this scenario, the two serologic tests will be conditionally dependent (correlated) because they both measure serum antibodies. As in the prior example, the data are classified into two populations on the basis of the batch information, and the true infection status is not incorporated into the analysis. The sample size for the first batch was reduced to 462 because 1 pig had a missing ELISA value (Fig. 3).

When ML methods that assume conditional independence are used, the sensitivity and specificity of both tests are overestimated, e.g., the MAT sensitivity and specificity are estimated to be 100 and 97.3% compared with their true values of 82.9 and 90.2%, respectively. A Bayesian analysis of the same data that assumes conditional independence also results in overestimation of both parameters (24). One approach to obtaining unbiased estimates is to include two new parameters in the model for the correlations between the sensitivities and specificities of the two tests (14). However, the addition of two more parameters precludes an ML solution because there are now more parameters than degrees of freedom. Hence, a moderate amount of prior information must be provided for two parameters in a Bayesian analysis; most likely this information would be estimates for MAT sensitivity and specificity because this test had been used for many years before the ELISA was developed.

For the Bayesian analysis of these data, which is described in detail elsewhere (14), the mode for MAT sensitivity was set at the correct value of 83%, but a diffuse distribution with the 5th percentile at 70% was allowed.

Similarly, for MAT specificity the mode was set at 90%, with the 5th percentile at 75% to account for greater certainty in that estimate. These values equated to beta (24.09, 5.73) and beta (23.05, 3.45) for sensitivity and specificity, respectively. Posterior estimates from WinBUGS were close to the true values (80.6 and 89.5% for MAT sensitivity and specificity, respectively; 71.5 and 85.5% for ELISA sensitivity and specificity, respectively), and the four 95% PIs included the true values.

An alternative approach is to use results of a third conditionally independent test (i.e., a model based on use of three tests in two populations) to yield sensitivity and specificity estimates. For the *T. gondii* example, this can be done using results of a mouse bioassay that detects viable cysts in muscle tissue and has a sensitivity of about 50% but is perfectly specific. The modeling approach and results have been described in detail elsewhere (3). The major benefit of using the three-test model over the two-test model is that the PIs for test accuracy parameters are narrower.

### GUIDELINES FOR RISK ASSESSORS

To facilitate improved exposure assessments using test-based data, standard guidelines would be helpful and would increase the transparency of data evaluation and incorporation in microbial risk assessments. Several of these suggestions have been made previously (22, 26).

1. Tests and test protocols should be described, and the analytic and diagnostic sensitivity and specificity values should be listed, if available. Sources of these estimates should be referenced, and the numbers of infected and noninfected sampling units used in the validation study should be given. In the unlikely event that multiple evaluation studies are available, these studies should be systematically reviewed using Cochrane method guidelines (8). When expert opinion is used to obtain these estimates, the name of the expert(s) and method for elicitation of the estimates should be documented.
2. For each data source, there should be a description of the sampling scheme, including a clear definition of the sampling unit(s), whether sampling was random, focused on high-risk groups, or nonrandom, and the weight or volume of sample product tested. The statistical performance characteristics of the sampling schemes should be defined (22).
3. The method of true prevalence estimation, whether Bayesian or frequentist, should be described. When Bayesian methods are used, a sensitivity analysis using different prior distributions should be performed. When apparent prevalence and rather than true prevalence estimates are used in the risk assessment, a description of the likely effects on the risk estimates should be included, e.g., under- or overestimation and implications for the study's conclusions. Similarly, when prevalence data from multiple sources that use different tests are combined, the possible biases from failing to correct for test accuracy should be described.

Unbiased accuracy estimates for tests used to collect data for microbial risk assessment will facilitate improved

interpretation of pathogen data from different studies and different populations and those collected at different times, including different seasons. Potential benefits will include better among- and within-population comparisons of prevalence, reduced underestimation of prevalence because of low-sensitivity tests, and less biased estimation of the effects of mitigation strategies. Historically, the need for such accuracy estimates has not been given high priority, but the availability of latent-class methods for conditionally independent and dependent tests provides an alternative approach that is less restrictive and may have utility especially for foodborne pathogens in animals. An additional advantage of this approach when all individuals in two populations are tested is the generation of true prevalence estimates for both populations in addition to the test accuracy estimates.

When published or expert-derived estimates of sensitivity and specificity are available, estimation of true prevalence is straightforward regardless of whether single or multiple populations are involved. Compared with ML methods, Bayesian methods are more flexible for dealing with rare events, clustered data, and multiple sources of uncertainty, all of which are considered important issues for foodborne pathogens. Bayesian methods can be readily adapted to calculate the sensitivity and specificity of testing schemes involving multiple tests that are interpreted in parallel or in series (3).

Estimation of test accuracy and true prevalence can now be done using WinBUGS (40), which allows the implementation of Markov chain Monte Carlo techniques, rather than other specialized software. As with any modeling technique, model diagnostics should be performed for complex problems, and convergence of the Markov chain Monte Carlo methods should be assessed (32). In addition, care is needed in the selection of prior distributions because if prior beliefs are biased and given excessive weight (e.g., a narrow prior distribution is specified) then the Bayesian analysis may be flawed. WinBUGS code for prevalence estimation and evaluation of test accuracy is available (<http://www.epi.ucdavis.edu/diagnostictests/>).

Additional research is needed on how best to incorporate historic prevalence data, where available, into a risk assessment. Schlosser and Ebel (35) described a herd-level simulation model that allowed for transmission of infection among herds and downweighted historic sample sizes to smaller present-day sample size equivalents.

Adoption of the suggested guidelines for risk assessors will require acceptance of the view that noncomparability of data caused by imperfect tests will persist as an issue in microbial risk assessment unless the importance of accuracy and reproducibility estimates for diagnostic tests is widely acknowledged. However, it is still unclear whether the responsibility for test evaluation rests with government agencies, test developers, test-kit marketers, or laboratories that use the tests.

### ACKNOWLEDGMENTS

This study was supported in part by the USDA-CSREES-NRI Competitive Grants Program award 99-04252. The development of methods

described here would not have been possible without the excellent assistance of Wes Johnson, Tim Hanson, Eric Suess, Pat McInturff, Marios Georgiadis, and Adam Branscum. I thank J. P. Dubey and Peter Lind for providing the toxoplasmosis data.

## REFERENCES

- Alban, L., A. Olsen, B. Nielsen, R. Sørensen, and B. Jessen. 2002. Qualitative and quantitative risk assessment for human salmonellosis due to multi-resistant *Salmonella* Typhimurium DT104 from consumption of Danish dry-cured pork sausages. *Prev. Vet. Med.* 52: 251–265.
- Bemrah, N., M. Sanaa, M. H. Cassin, M. W. Griffiths, and O. Cerf. 1998. Quantitative risk assessment of human listeriosis from consumption of soft cheese made from raw milk. *Prev. Vet. Med.* 37: 129–145.
- Branscum, A. J., I. A. Gardner, and W. O. Johnson. Estimation of diagnostic test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med.*, in press.
- Branscum, A. J., I. A. Gardner, and W. O. Johnson. Bayesian modeling of animal and herd-level prevalence. *Prev. Vet. Med.*, in press.
- Casella, G., and E. I. George. 1992. Explaining the Gibbs sampler. *Am. Stat.* 46:167–174.
- Cassin, M. H., A. M. Lammerding, E. C. D. Todd, W. Ross, and R. S. McColl. 1998. Quantitative risk assessment for *Escherichia coli* O157:H7 in ground beef hamburgers. *Int. J. Food Microbiol.* 41:21–44.
- Christensen, J., and I. A. Gardner. 2000. Herd-level interpretation of diagnostic tests. *Prev. Vet. Med.* 45:83–106.
- Cochrane Methods Group on Systematic Review of Screening and Diagnostic Test: Recommended Methods, updated 6 June 1996. Available at: <http://www.cochrane.org/cochrane/sadtdoc1.htm>. Accessed 14 January 2004.
- Cowling, D. W., I. A. Gardner, and W. O. Johnson. 1999. Comparison of methods for estimation of individual-level prevalence based on pooled samples. *Prev. Vet. Med.* 39:211–225.
- Dubey, J. P., P. Thulliez, R. M. Weigel, C. D. Andrews, P. Lind, and E. C. Powell. 1995. Sensitivity and specificity of various serologic tests for detection of *Toxoplasma gondii* infection in naturally infected sows. *Am. J. Vet. Res.* 56:1030–1036.
- Enøe, C., M. P. Georgiadis, and W. O. Johnson. 2000. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.* 45:61–81.
- Gardner, I. A. 1997. Testing to fulfill HACCP (hazard analysis critical control points) requirements: principles and examples. *J. Dairy Sci.* 80:3453–3457.
- Gardner, I. A., H. Stryhn, P. Lind, and M. T. Collins. 2002. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev. Vet. Med.* 45:107–122.
- Georgiadis, M. P., W. O. Johnson, I. Gardner, and R. Singh. 2003. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl. Stat.* 52:63–76.
- Greiner, M., and I. A. Gardner. 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.* 45:3–22.
- Greiner, M., and I. A. Gardner. 2000. Application of diagnostic tests in veterinary epidemiologic studies. *Prev. Vet. Med.* 45:43–59.
- Greenland, S. 1998. Probability logic and probabilistic interpretation. *Epidemiology* 9:322–332.
- Guissé, C. 2002. Microbiology in food and animal feeding stuffs. *ISO Bull.* March:17–19.
- Hanson, T. E., W. O. Johnson, and I. A. Gardner. 2003. Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *J. Agric. Biol. Environ. Stat.* 8:223–239.
- Hanson, T. E., W. O. Johnson, I. A. Gardner, and M. P. Georgiadis. 2003. Determining the infection status of a herd. *J. Agric. Biol. Environ. Stat.* 8:469–485.
- Hui, S. L., and S. D. Walter. 1980. Estimating the error rates of diagnostic tests. *Biometrics* 36:167–171.
- International Commission on Microbiological Specifications for Foods, Working Group on Microbial Risk Assessment. 1998. Potential application of risk assessment techniques to microbial issues related to international trade in food and food products. *J. Food Prot.* 61:1075–1086.
- Jacobson, R. H. 1998. Validation of serologic assays for diagnosis of infectious diseases. *Rev. Sci. Tech. Off. Int. Epizoot.* 17:469–486.
- Johnson, W. O., J. L. Gastwirth, and L. M. Pearson. 2001. Screening without a gold standard: the Hui-Walter paradigm revisited. *Am. J. Epidemiol.* 153:921–924.
- Kaplan, S. 1997. The words of risk analysis. *Risk Anal.* 17:407–417.
- Lammerding, A. M., and A. Fazil. 2000. Hazard identification and exposure assessment for microbial food safety risk assessment. *Int. J. Food Microbiol.* 58:147–157.
- Malakoff, D. 1999. Bayes offers a “new” way to make sense of numbers. *Science* 286:1460–1464.
- Marks, H. M., M. E. Coleman, C.-T. Jordan Lin, and T. Roberts. 1998. Topics in microbial risk assessment: dynamic flow tree process. *Risk Anal.* 18:309–328.
- Martin, S. W., M. Shoukri, and M. A. Thorburn. 1992. Evaluating the health status of herds based on tests applied to individuals. *Prev. Vet. Med.* 14:33–43.
- McNab, W. B. 1998. A general framework illustrating an approach to quantitative microbial food safety risk assessment. *J. Food Prot.* 61:1216–1228.
- Pouillot, R., G. Gerbier, and I. A. Gardner. 2002. “TAGS”, a program for the evaluation of test accuracy in the absence of a gold standard. *Prev. Vet. Med.* 53:67–81.
- Robert, C. P., and G. Casella. 1999. Monte Carlo statistical methods. Springer-Verlag, New York.
- Rogan, W. J., and B. Gladen. 1978. Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* 107:71–76.
- Saah, A. J., and D. R. Hoover. 1997. Sensitivity and specificity reconsidered: the meaning of these terms in analytical and diagnostic settings. *Ann. Intern. Med.* 126:91–94.
- Schlosser, W., and E. Ebel. 2003. Use of a Markov-chain Monte Carlo model to evaluate the time value of historical testing information in animal populations. *Prev. Vet. Med.* 48:167–175.
- Suess, E. A., I. A. Gardner, and W. O. Johnson. 2002. Hierarchical Bayesian model for prevalence inferences and determination of a country's status for an animal pathogen. *Prev. Vet. Med.* 55:155–171.
- Vose, D. J. 1998. The application of quantitative risk assessment to microbial food safety. *J. Food Prot.* 61:640–648.
- Vose, D. 2000. Risk analysis—a quantitative guide, 2nd ed. John Wiley and Sons, Chichester, UK.
- Whiting, R. C., and R. L. Buchanan. 1997. Development of a quantitative risk assessment model for *Salmonella* Enteritidis in pasteurized liquid eggs. *Int. J. Food Microbiol.* 36:111–125.
- WinBUGS Project. 1996–2000. WinBUGS, version 1.4. Department of Epidemiology and Public Health, Cambridge, UK. Available at: <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>. Accessed 14 January 2004.