American Journal of Epidemiology
Vol. 141, No. 3
Copyright © 1995 by The Johns Hopkins University School of Hygiene and Public Health
All rights reserved
Printed in U.S.A.

Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard

Lawrence Joseph,[1-3] Theresa W. Gyorkos,[1,2,4] and Louis Coupal[2,3]

It is common in population screening surveys or in the investigation of new diagnostic tests to have results from one or more tests investigating the same condition or disease, none of which can be considered a gold standard. For example, two methods often used in population-based surveys for estimating the prevalence of a parasitic or other infection are stool examinations and serologic testing. However, it is known that results from stool examinations generally underestimate the prevalence, while serology generally results in overestimation. Using a Bayesian approach, simultaneous inferences about the population prevalence and the sensitivity, specificity, and positive and negative predictive values of each diagnostic test are possible. The methods presented here can be applied to each test separately or to two or more tests combined. Marginal posterior densities of all parameters are estimated using the Gibbs sampler. The techniques are applied to the estimation of the prevalence of *Strongyloides* infection and to the investigation of the diagnostic test properties of stool examinations and serologic testing, using data from a survey of all Cambodian refugees who arrived in Montreal, Canada, during an 8-month period. *Am J Epidemiol* 1995;141:263–72.

Bayes theorem; diagnostic tests, routine; epidemiologic methods; models, statistical; Monte Carlo method; prevalence; sensitivity and specificity

It is often the case when determining the prevalence of a medical condition through population screening or when evaluating a new medical diagnostic test that data are available on one or more tests, none of which can be considered a gold standard. In fact, one may argue that this is virtually always the situation, since few tests are considered to be 100 percent accurate. Despite these limitations, it is important for clinical and public health practices to have the best possible estimates of disease prevalence and test parameters, such as the sensitivity, specificity, and positive and negative predictive values.

For example, the data in table 1 were obtained from a survey of all Cambodian refugees who arrived in Montreal, Canada, between July 1982 and February 1983 (1, 2). The observed sample prevalence using the information from stool examinations alone is 24.7 percent, while the prevalence from serology alone is 77.2 percent, an absolute difference of more than 50 percent. In fact, the situation is even less certain than these values indicate, since the above estimates do not take into account sampling variability or the likelihood that several of the subjects may be false positives or false negatives, as neither test has perfect sensitivity or specificity. For the same reasons, inferences about the test parameters are equally contentious in the absence of a gold standard.

This problem arises from the misclassification of data. A review of frequentist (non-Bayesian) approaches to inference from data in the presence of misclassification is given by Walter and Irwig (3). In general, one can observe P different populations, each subject in each population receiving D different diagnostic tests. Here the term "diagnostic test" is used generically to denote any method of disease detection. For example, different observers of the same test or two applications of the same test on a subject over time are considered as different tests. It is of interest to estimate parameters belonging to each population, typically the prevalence of disease, as well as the parameters of each diagnostic test.

Two of the most common situations occur when $P = 1$ and $D = 1$ or $D = 2$. In the case when $P = D = 1$, there are three parameters to be estimated: the population prevalence and the sensitivity and specific-

Received for publication June 3, 1994, and in final form October 3, 1994.
[1] Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada.
[2] Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, Montreal, Canada.
[3] Centre for the Analysis of Cost Effective Care, Department of Medicine, Montreal General Hospital, Montreal, Canada.
[4] McGill Centre for Tropical Diseases, McGill University, Montreal, Canada.
Reprint requests to Dr. Lawrence Joseph, Division of Clinical Epidemiology, Department of Medicine, Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4, Canada.

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

TABLE 1. Results of serologic and stool testing for *Strongyloides* infection on 162 Cambodian refugees arriving in Montreal, Canada, between July 1982 and February 1983

		Stool examination +	Stool examination −	
Serology	+	38	87	125
	−	2	35	37
		40	122	162

ity of the test. The data possess only 1 df since, given the total sample size, the number of subjects testing positively fixes the number of subjects with negative tests. In the case of two diagnostic tests, $P = 1$, $D = 2$, there are five unknown parameters, since each test will in general have unknown sensitivities and specificities, in addition to the population prevalence. However, there are only 3 df, since knowing the total sample size and any three of four cells in the 2×2 table fixes the number in the fourth cell.

Having more parameters to estimate than degrees of freedom means that constraints have to be imposed on a subset of the parameters in order to carry out estimation procedures, such as maximum likelihood. For the case $P = D = 1$, both Quade et al. (4) and Rogan and Gladen (5) assumed that the sensitivity and specificity of the test are exactly known. When $P = 1$ and $D = 2$, estimation procedures have been described under a variety of different constraints. These have included assuming that the sensitivity and specificity of one of the two tests are completely known (6) and that the specificities, but not the sensitivities, of both tests are known (7). Estimates of the remaining unconstrained parameters are calculated, conditional on the assumed known values of the constrained parameters. However, this procedure neither estimates these latter values, which are almost always truly unknown, nor is able to account for the uncertainty in their assumed values, for example, when deriving confidence intervals for the unconstrained parameters. In fact, since all parameters are typically unknown, the division into constrained and unconstrained sets is often quite arbitrary.

The basic idea behind the Bayesian approach presented here is to eliminate the need for these constraints by first constructing a prior distribution over all unknown quantities. The data, through the likelihood function, are then combined with the prior distribution to derive posterior distributions using Bayes' theorem. This allows simultaneous inferences to be made on all parameters. The posterior distributions contain updated beliefs about the values of the model parameters, after taking into account the information provided by the data. This procedure can be viewed as a generalization of the frequentist approach, since the latter's constrained parameters can be considered to have degenerate marginal prior distributions with probability mass equal to one on their constrained values, while the lack of prior information assumed for the unconstrained parameters can be represented by a uniform or other noninformative prior distribution. Using the Bayesian approach with these prior distributions will provide numerically nearly identical point and interval estimates as the frequentist approach. However, the Bayesian approach also allows for a wide variety of other prior distributions. Since exact values for the constrained parameters are seldom if ever known, the consideration of nondegenerate prior distributions covering a range of values is more realistic.

Another advantage is that normal distribution approximations, commonly used to derive confidence intervals around unknown parameters from estimated standard errors, are not required. Since posterior distributions can be highly skewed, the use of the exact posterior marginal distributions can result in substantial improvements in the validity of interval estimates.

Direct calculation of the posterior distributions can be difficult. The Gibbs sampler (8–10) is an iterative Markov-chain Monte Carlo technique for approximating analytically intractable posterior densities. Recently, it has been used to estimate parameters in a wide variety of problems in health research (11–13). It is the goal here to demonstrate how approximate marginal posterior densities of all parameters of interest in the case of one or two diagnostic tests in the absence of a gold standard can be calculated using the Gibbs sampler.

ONE DIAGNOSTIC TEST

The problem considered in this section can be described as follows. The results of a single diagnostic test for a certain disease are available on a random sample of subjects. No gold standard test is available, either because none exists, because of measurement error, or because it cannot practically be performed. The latter situation often occurs when costs are prohibitive. The object is to draw inferences about the prevalence, π, of the disease in the population from which the sample was drawn, as well as the sensitivity, S, and specificity, C, of the test, along with the positive and negative predictive values for the population.

Let a and b be the observed number of positive and negative test results, respectively, in the sample of $a + b = N$ subjects. Let Y_1 and Y_2 be the information that is missing when there is no gold standard, that is, the number of true positive test results out of a and b,

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

respectively. Thus, Y_1 is the number of true positives, and Y_2 is the number of false negatives. See table 2. Such missing information has been termed "latent data" by Tanner and Wong (14), and analyses using such data have been referred to as "latent class analysis" by Kaldor and Clayton (15) and Walter and Irwig (3).

The likelihood function of the observed and latent data shown in table 2 is given by

$$l(a,b,Y_1,Y_2 \mid \pi,S,C) = [\pi S]^{Y_1}[\pi(1-S)]^{Y_2}[(1-\pi)(1-C)]^{a-Y_2}[(1-\pi)C]^{b-Y_2}$$
$$= \pi^{Y_1+Y_2}(1-\pi)^{N-Y_1-Y_2}S^{Y_1}(1-S)^{Y_2}C^{b-Y_2}(1-C)^{a-Y_1}.$$

Prior information in the form of a beta density will be assumed. A random variable, θ, has a beta distribution with parameters (α,β) if it has a probability density given by

$$f(\theta) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \\ \qquad 0 \le \theta \le 1, \alpha,\beta > 0, \text{ and} \\ 0, \qquad \text{otherwise,} \end{cases}$$

where $B(\alpha,\beta)$, the beta function evaluated at (α,β), is the normalizing constant. This family of distributions was selected since its region of positive density, from 0 to 1, matches the range of all parameters of interest in this study, and because it is a flexible family, in that a wide variety of density shapes can be derived by selecting different choices of α and β (16). It also has the advantage of being the conjugate prior distribution for the binomial likelihood, a property that simplifies the derivation of the posterior distributions. Let (α_π,β_π), (α_S,β_S), and (α_C,β_C) represent the prior beta parameters for π, S and C, respectively. Since by Bayes' theorem the joint posterior distribution is proportional to the product of the likelihood function and the prior distribution, it is given by

$$\pi^{Y_1+Y_2+\alpha_\pi}(1-\pi)^{N-Y_1-Y_2+\beta_\pi}S^{Y_1+\alpha_S}(1-S)^{Y_2+\beta_S}C^{b-Y_2+\alpha_C}(1-C)^{a-Y_1+\beta_C}, \quad (1)$$

up to a normalizing constant. Of course, the latent data, Y_1 and Y_2, are not observed, impeding direct use of equation 1 in calculating the marginal posterior densities of π, S, and C. However, inference is possible using a Gibbs sampler algorithm. The basic idea is as follows. Conditional on knowing the exact values of the prevalence and all diagnostic test parameters, it is possible to derive posterior distributions of the latent data Y_1 and Y_2. Conversely, if Y_1 and Y_2 are known, then deriving posterior distributions of the prevalence and diagnostic test parameters given the prior distributions requires only a straightforward application of Bayes' theorem. An algorithm that alternates between these two steps can thus be devised, similar in spirit to the expectation maximization algorithm that is commonly used in latent class analysis (3). The Gibbs sampler algorithm, described in the Appendix, provides random samples from the marginal posterior densities of each parameter of interest. These random samples can then be used to reconstruct the marginal posterior densities, or summaries of these densities, such as their means, medians, or standard deviations, as well as probability interval summaries.

TABLE 2. Observed and latent data in the case of one diagnostic test in the absence of a gold standard, presented in a 2 × 2 table

		Truth +	Truth −	
Test	+	Y_1	$a-Y_1$	a
	−	Y_2	$b-Y_2$	b
		Y_1+Y_2	$N-(Y_1+Y_2)$	N

TWO DIAGNOSTIC TESTS

The methods of the previous section can be extended to the situation where results of two diagnostic tests for the same disease are available on a randomly selected sample of subjects, where neither test can be considered a gold standard. Of interest are the marginal posterior densities of the prevalence of the disease in the population from which the sample was drawn, π, as well as the sensitivities, S_1 and S_2, specificities, C_1 and C_2, and positive and negative predictive values of each test, given the data and any available prior information. Data are collected as shown in table 3.

Let the unobserved latent data Y_1, Y_2, Y_3, and Y_4 represent the number of true positive subjects out of the observed cell values u, v, w and x, respectively, in the 2 × 2 data of table 3. Since any subject, whether truly possessing the disease in question or not, can test positively or negatively on each test, there are eight possible combinations. The situation is summarized in table 4.

The likelihood function can be derived directly from the information in table 4, and the joint posterior density is proportional to this likelihood times the prior distribution as in the previous section. The Gibbs sampler can again be used to construct the marginal

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

TABLE 3. Observed data from two diagnostic tests, in the absence of a gold standard

		Test 2 +	Test 2 −	
Test 1	+	u	v	$(u+v)$
	−	w	x	$(w+x)$
		$(u+w)$	$(v+x)$	N

posterior densities of all parameters of interest. See the Appendix for details.

PRIOR DISTRIBUTIONS

An important step in any Bayesian analysis is to obtain a prior distribution over all model parameters. This can be accomplished using past data, if available, or by drawing upon expert knowledge, or a combination of both. There is a large literature on the elicitation of prior distributions. Proposed methods have included directly matching percentiles (17) or means and standard deviations (18) to a member of a preselected family of distributions, as well as methods that use the predictive distribution of the data (19). The predictive distribution is the marginal distribution of the observable data, which is found by integrating the likelihood of the data over the prior distribution of the unknown parameters (18).

For the present problem, model parameters include the sensitivity and specificity of each diagnostic test, as well as the population prevalence. Both the stool examination and serology test are standard diagnostic tools in parasitology. It is expected that stool examinations generally underestimate population prevalence (20), while serology generally results in overestimation due to cross-reactivity (21) or persistence of reactivity following parasite cure (22). Nevertheless, the lack of a gold standard for the detection of most parasitic infections means that the properties of these tests are not known with high accuracy. In consultation with a panel of experts from the McGill Centre for Tropical Diseases, we determined equally tailed 95 percent probability intervals (i.e., 2.5 percent in each tail) for the sensitivity and specificity of each test (see table 5). These were derived from a review of the relevant literature and clinical opinion (21–28).

The particular beta prior density for each test parameter was selected by matching the center of the range with the mean of the beta distribution, given by $\alpha/(\alpha+\beta)$, and matching the standard deviation of the beta distribution, given by

$$\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}},$$

TABLE 4. 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	+	+	+	$\pi S_1 S_2$
Y_2	+	+	−	$\pi S_1(1-S_2)$
Y_3	+	−	+	$\pi(1-S_1)S_2$
Y_4	+	−	−	$\pi(1-S_1)(1-S_2)$
$u-Y_1$	−	+	+	$(1-\pi)(1-C_1)(1-C_2)$
$v-Y_2$	−	+	−	$(1-\pi)(1-C_1)C_2$
$w-Y_3$	−	−	+	$(1-\pi)C_1(1-C_2)$
$x-Y_4$	−	−	−	$(1-\pi)C_1C_2$

* The 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.

TABLE 5. Equally tailed 95% probability ranges and coefficients of the beta prior densities for the test parameters in the diagnosis of *Strongyloides* infection*

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

* A uniform density over the range [0,1] ($\alpha=1$, $\beta=1$) was used for the prior distribution for the prevalence of *Strongyloides* in the refugee population.

with one quarter of the total range. These two conditions uniquely define α and β. An alternative approach is to match the end points of the given ranges to beta distributions with similar 95 percent probability intervals. The coefficients obtained from these two approaches usually give very similar prior distributions. One way to consider a beta(α,β) distribution is to equate it with the information contained in a prior sample of $(\alpha + \beta)$ subjects, α of whom were positive. The sum $(\alpha + \beta)$ is often referred to as the "sample size equivalent" of the prior information (18).

A priori, very little was known about the prevalence of *Strongyloides* infection among the Cambodian refugees. To approximate this uncertainty, a uniform prior distribution on the range from 0 to 1 was used. While independence was assumed for all parameters a priori, this does not ensure independence of the posterior distributions.

While it is possible to use noninformative or uniform prior distributions for all test parameters, this is not necessarily desirable. In cases where there are relatively few data per parameter, drawing useful inferences may require substantive prior information. For example, if the prevalence in the population is high (low), then the data will contain relatively little information on specificity (sensitivity), since there

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

will not be many negative (positive) subjects on which to base estimates. However, previous information may indicate, for example, a high specificity, as was the case here for the stool examination. Not using this information can result in much wider interval estimates for all parameters.

STRONGYLOIDES INFECTION IN CAMBODIAN REFUGEES

The methods presented above will now be applied to the data given in table 1. Analyses were run using data from each diagnostic test alone, as well as from their combination. The prior parameters presented in table 5 were used. The results in the form of posterior medians and 95 percent equally tailed posterior credible intervals appear in table 6. Credible intervals are the Bayesian analogs of confidence intervals. Plots of the prior and marginal posterior densities for the prevalence of *Strongyloides* infection appear in figure 1. These densities were obtained by smoothing the output from the Gibbs sampler with a normal kernel (29). Similarly constructed posterior densities for the sensitivities and specificities of each test, based on the output from the Gibbs sampler using the data from both tests combined, appear in figure 2. Other techniques for density estimation in the context of the Gibbs sampler have been discussed by Gelfand and Smith (8).

As is evident from both table 6 and figure 1, the densities can be highly skewed. For example, the median of the marginal posterior distribution of the prevalence using data from serologic testing alone was 0.80, although the 95 percent credible interval was 0.23–0.99. (For nonsymmetric posterior densities, the highest posterior density (17) intervals could be used in place of equally tailed posterior credible intervals. The highest posterior density intervals result in the narrowest possible intervals with the same probability content. The 95 percent highest posterior density interval for the prevalence using data from serologic testing alone is 0.34–1.00, which is 0.10 shorter in length than the symmetric interval with the same probability content.)

Sharper inference about the prevalence of *Strongyloides* infection is gained from the combined results compared with that from stool examinations or serologic testing alone. Figure 1 supports the assertion that stool examinations underestimate and serologic testing overestimates the population prevalence, in that the posterior density from stool examinations lies more to the left than that obtained from serology, with the density from both tests combined located in between. Overall, the 95 percent posterior credible interval for the population prevalence from both tests combined was 0.52–0.91. The results confirm the low sensitivity of stool examinations (95 percent credible interval 0.22–0.44) and indicate a very high specificity (0.91–0.99). The sensitivity of serologic testing appears to be in the higher portion of the range of its prior distribution (0.80–0.95), while the posterior distribution of the specificity of serology closely matched the prior information (0.36–0.96). The latter result is partly due to the fact that the median prevalence was 76 percent, and with only $162 \times 0.24 \approx 39$ subjects typically classified as not having disease, there were limited data with which to update the prior distributions for the test specificities. The prior sample size equivalent for the specificity of stool examination was 75 subjects, about twice as large as the average number of subjects contributing to updating this parameter. Since a stool examination is positive only when the *Strongyloides* parasite is directly viewed under a microscope, false positives are rare, and the lower limit of the prior range of 90 percent was even thought by some to be

TABLE 6. Marginal prior and posterior medians and lower and upper limits of the posterior equally tailed 95% credible intervals for the prevalence (π) and sensitivities (S_1, S_2), specificities (C_1, C_2), and positive and negative predictive values (PPV_1, PPV_2, NPV_1, NPV_2) for each screening test alone and for the combination of the two tests

		Prior information		Stool examination alone		Serology alone		Both tests combined	
		Median	95% CI*	Median	95% CI	Median	95% CI	Median	95% CI
	π	0.50	0.03–0.98	0.74	0.41–0.98	0.80	0.23–0.99	0.76	0.52–0.91
Stool examination	S_1	0.24	0.07–0.47	0.30	0.21–0.47			0.31	0.22–0.44
	C_1	0.95	0.89–0.99	0.95	0.88–0.99			0.96	0.91–0.99
	PPV_1	0.84	0.10–1.00	0.95	0.74–1.00			0.98	0.88–1.00
	NPV_1	0.56	0.03–0.98	0.33	0.02–0.73			0.30	0.11–0.63
Serology	S_2	0.81	0.63–0.92			0.83	0.73–0.92	0.89	0.80–0.95
	C_2	0.72	0.31–0.96			0.58	0.22–0.94	0.67	0.36–0.95
	PPV_2	0.76	0.07–1.00			0.91	0.18–1.00	0.90	0.62–1.00
	NPV_2	0.78	0.08–1.00			0.44	0.03–0.94	0.70	0.28–0.92

* CI, credible interval.

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

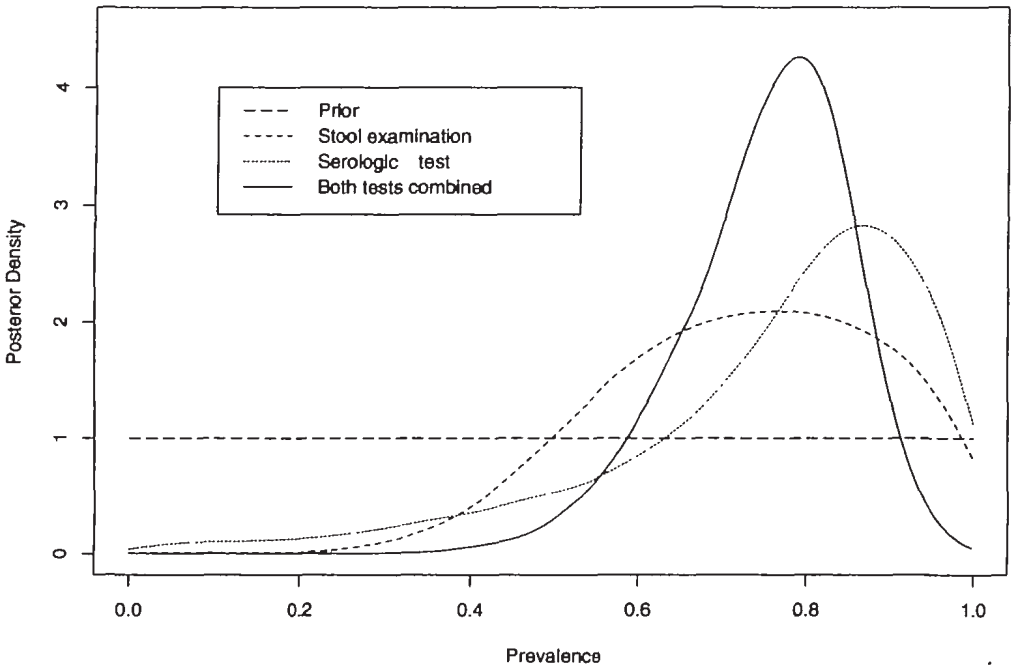


FIGURE 1. Prior density and marginal posterior density for the prevalence of *Strongyloides* infection in Cambodian refugees, using data from stool examinations and serologic tests alone, and from the two tests combined: Montreal, Canada, July 1982 to February 1983.

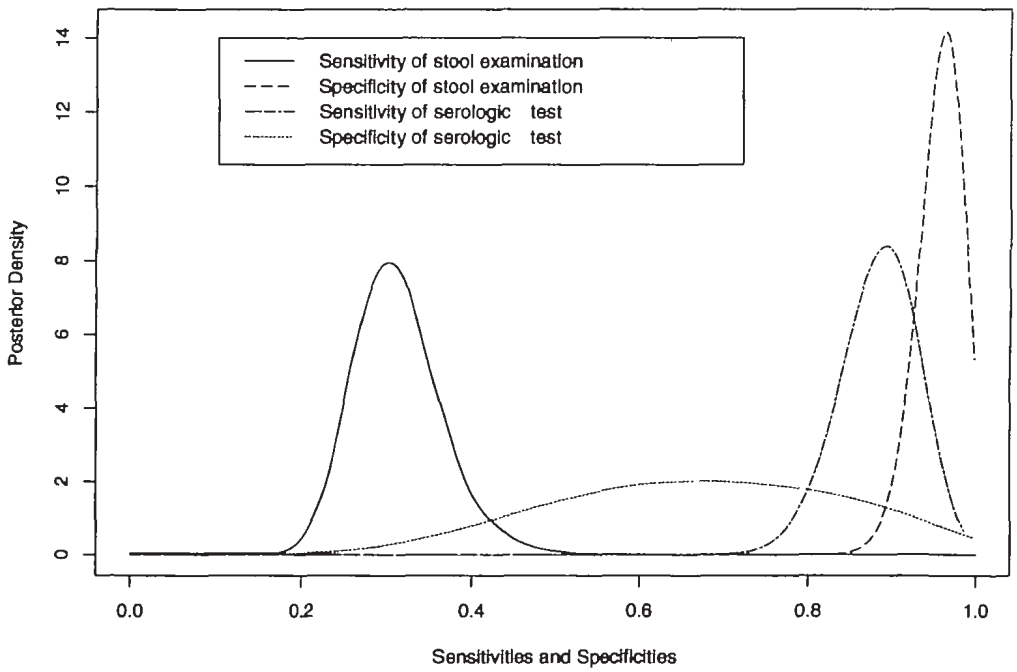


FIGURE 2. Marginal posterior density for the sensitivities and specificities of stool examinations and serologic tests for the presence of *Strongyloides* infection in Cambodian refugees, using data from both tests combined: Montreal, Canada, July 1982 to February 1983.

conservative. Of course, different posterior inferences would be drawn by anyone with less confidence in the specificity of stool examinations.

Figure 3 summarizes the marginal posterior probability functions for the latent data using the information from both tests combined. Most of the persons testing positively on stool examinations are likely to be true positives, as indicated by the histograms for Y_1 and Y_3, where high proportions of the iterations placed all such subjects as positive. It is highly likely that at least 50 of the 125 subjects testing positively on serology but negatively on stool examinations are posi-

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

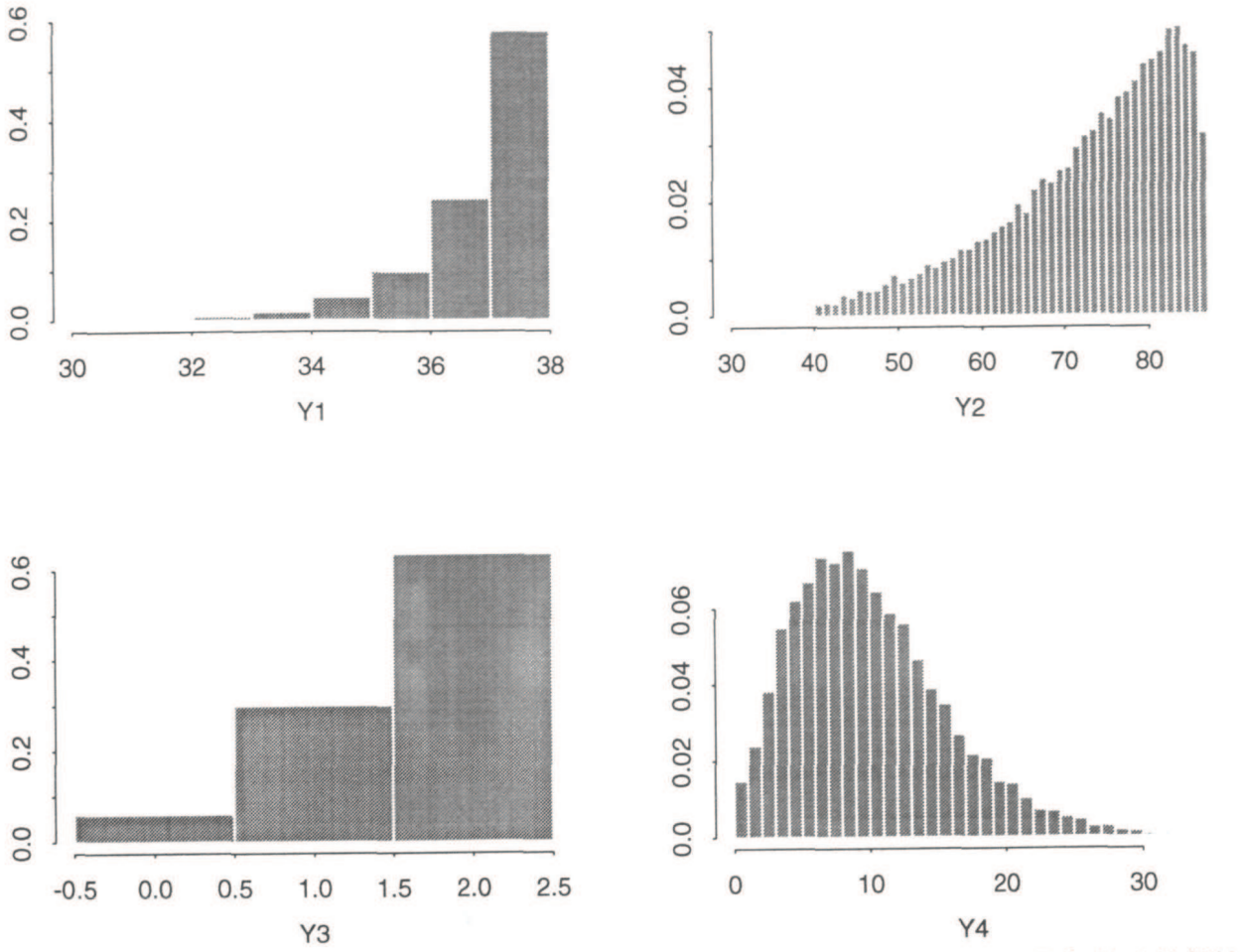


FIGURE 3. Histograms of the output from the Gibbs sampler for the number of truly positive subjects in each cell of table 1. Y_1 (Y1) is the number of truly positive subjects out of 38 with positive results on both tests, Y_2 (Y2) is the number of truly positive subjects out of 125 with positive serology but negative stool examinations, Y_3 (Y3) is the number of truly positive subjects out of two with positive stool examinations but negative serology, and Y_4 (Y4) is the number of truly positive subjects out of 35 with negative results on both tests. Data from survey of Cambodian refugees: Montreal, Canada, July 1982 to February 1983.

tive and that the number could be as high as 90, as evidenced by the histogram for Y_2. Finally, from the histogram for Y_4, it seems likely that approximately 10 subjects with negative results on both tests are, in fact, truly positive.

In general, different prior information about the parameters will lead to different posterior distributions. If there is considerable controversy, and especially if very narrow prior distributions are used, results from a range of prior distributions should be reported. An investigator may have more diffuse prior distributions, for example, with wider prior intervals by 0.10 for all parameters than those shown in table 5. In this case, the final 95 percent posterior credible interval for the prevalence is 0.41–0.91, a decreased lower interval limit compared with that resulting from the priors in table 2. The decrease in the lower limit of the latter interval is partially due to the fact that, in order to widen the prior interval for the specificities of stool examinations and serologic testing, the mean values must be lowered, since one cannot go above 100 percent. Lower specificities result in more false positives and, thus, lower numbers of true positives, given the same data. Conversely, an investigator with prior intervals narrower by 0.10 for all parameters will derive a posterior 95 percent posterior credible interval for the prevalence of 0.58–0.94, an interval that is narrower by 0.03.

DISCUSSION

The methods presented here can easily be extended to the case of three or more diagnostic tests when there is no gold standard. In this case, all parameters can be estimated by maximum likelihood without imposing constraints (30), but the Bayesian approach can still provide improved inference if there is substantive prior information or if posterior distributions are not

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

normal. An example of this occurs in population screening for asthma, where exercise tests, metacholine challenge, and a previous physician diagnosis of asthma are all in common use. Three tests result in 16 possible outcomes of the type listed in table 5, and in general, n different diagnostic tests used in combination will result in 2^{n+1} outcomes to consider. Extensions to other misclassified or latent data situations, such as those reviewed by Walter and Irwig (3), are also possible, including extensions to tests that classify individuals into more than two categories or provide continuous outcomes.

It can be argued that a subject with a greater degree of infection would be more likely to test positively on each test, so that the test sensitivities and specificities may be functions of individual subject characteristics. If this is the case, an approach in which test parameters are functions of patient characteristics may be desirable. Of course, more detailed data than those presented in table 1 would be required.

In the screening for *Strongyloides* infection, some information was obtained from either the stool examination or serologic testing used alone, but the combination of tests allowed for sharper inferences to be drawn. In general, the amount of information about population prevalences and test parameters contained in the data from any experiment is a complex function of the data and the available prior information. Not accounting for uncertainties in all parameters simultaneously can substantially affect final inferences. For example, if serology is assumed to have exactly known sensitivity and specificity values of 80 percent and 70 percent, respectively, then the final 95 percent interval for the prevalence is 78–99 percent. This total width of 21 percent can be compared with the width for prevalence from serology alone in table 6 of 76 percent, which is almost four times as wide. To recapture this uncertainty, it has been suggested that several analyses using different sets of point estimates for the sensitivity and specificity can be performed. However, this conventional sensitivity analysis is still unsatisfactory, since it provides no guidance as to how to combine the different results into overall final estimates. The methods presented here are useful in drawing the best possible inferences from diagnostic tests in the absence of a gold standard.

ACKNOWLEDGMENTS

Drs. Lawrence Joseph and Theresa Gyorkos are both research scholars, supported by the Fonds de la Recherche en Santé du Québec and the National Health Research and Development Program, respectively.

The authors thank the members of the McGill Centre for Tropical Diseases, and especially Dr. J. D. MacLean, Director, for expertise in the elicitation of the prior distributions. The authors are also grateful to Shanshan Wang and Roxane du Berger for their technical advice and to Marie-Pierre Aoun for her skills in preparing this document.

REFERENCES

1. Gyorkos TW, Genta RM, Viens P, et al. Seroepidemiology of *Strongyloides* infection in the Southeast Asian refugee population in Canada. Am J Epidemiol 1990;132:257–64.
2. Gyorkos TW, Frappier-Davignon 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.
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. Quade D, Lachenbruch PA, Whaley FS, et al. Effects of misclassifications on statistical inferences in epidemiology. Am J Epidemiol 1980;111:503–15.
5. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol 1978;107:71–6.
6. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. Am J Epidemiol 1966;83:593–602.
7. Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. Biometrics 1978;34:77–86.
8. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990;85:398–409.
9. Gelfand AE, Hills SE, Racine-Poon A, et al. Illustration of Bayesian inference in normal data using Gibbs sampling. J Am Stat Assoc 1990;85:972–85.
10. Tanner MA. Tools for statistical inference. New York: Springer-Verlag, 1991.
11. Gilks WR, Clayton DG, Spiegelhalter DJ, et al. Modelling complexity: applications of Gibbs sampling in medicine. J R Stat Soc B 1993;1:39–52.
12. Coursaget P, Yvonnet B, Gilks WR, et al. Scheduling of revaccinations against hepatitis B virus. Lancet 1991;337:1180–3.
13. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol 1993;138:430–42.
14. Tanner MA, Wong WH. The calculation of posterior densities by data augmentation (with discussion). J Am Stat Assoc 1987;82:528–50.
15. Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. Stat Med 1985;4:327–35.
16. Johnson NL, Kotz S. Distributions in statistics: continuous univariate distributions. Vol. 2. New York: John Wiley & Sons, Inc, 1970.
17. Press SJ. Bayesian statistics: principles, models, and applications. New York: John Wiley & Sons, Inc, 1989.
18. Lee PM. Bayesian statistics: an introduction. 3rd ed. New York: Halsted Press, 1992.
19. Chaloner KM, Duncan GT. Assessment of a beta prior distribution: PM elicitation. Statistician 1983;32:174–80.
20. Guyatt HL, Bundy DAP. Estimation of intestinal nematode prevalence: influence of parasite mating patterns. Parasitology 1993;107:99–105.
21. Gam AA, Neva FA, Krotoski WA. Comparative sensitivity and specificity of ELISA and IHA for serodiagnosis of strongyloidiasis with larval antigens. Am J Trop Med Hyg 1987;37:157–61.

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

22. Genta RM. Predictive value of an enzyme-linked immunosorbent assay (ELISA) for the serodiagnosis of *Strongyloides*. Am J Clin Pathol 1988;89:391–4.
23. Nutman TB, Ottesen EA, Ieng S, et al. Eosinophilia in Southeast Asian refugees: evaluation at a referral center. J Infect Dis 1987;155:309–13.
24. Genta R. Global prevalence of strongyloidiasis: critical review with epidemiologic insights into the prevention of disseminated disease. Rev Infect Dis 1989;2:755–67.
25. Carroll SM, Karthigasu KT, Grove DI. Serodiagnosis of human strongyloidiasis by an enzyme-linked immunosorbent assay. Trans R Soc Trop Med Hyg 1981;75:706–9.
26. Bailey JW. A serological test for the diagnosis of *Strongyloides* antibodies in ex-Far East prisoners of war. Ann Trop Med Parasitol 1989;83:241–7.
27. Pelletier LL Jr, Baker CB, Gam AA, et al. Diagnosis and evaluation of treatment of chronic strongyloidiasis in ex-prisoners of war. J Infect Dis 1988;157:573–6.
28. Douce RW, Brown AE, Khamboonruang C, et al. Seroepidemiology of strongyloidiasis in a Thai village. Int J Parasitol 1987;17:1343–8.
29. Silverman BW. Density estimation for statistics and data analysis. London: Chapman and Hall, 1986.
30. Walter SD. Measuring the reliability of clinical data: the case for using three observers. Rev Epidemiol Sante Publique 1984;32:206–11.

APPENDIX

Implementation of the Gibbs sampler requires the specification of the full conditional distributions of the parameters, i.e., the conditional distributions of each parameter given the values of all of the other parameters. As is often the case, the full conditional distribution of each parameter does not always depend on all of the other parameters, which leads to some further simplifications. It is straightforward to show from equation 1 that the following conditional distributions must hold:

$$Y_1 \mid a,\pi,S,C \sim \text{Binomial}\left(a,\frac{\pi S}{\pi S + (1-\pi)(1-C)}\right), \tag{A1}$$

$$Y_2 \mid b,\pi,S,C \sim \text{Binomial}\left(b,\frac{\pi(1-S)}{\pi(1-S) + (1-\pi)C}\right), \tag{A2}$$

$$\pi \mid a,b,Y_1,Y_2,\alpha_\pi,\beta_\pi \sim \text{Beta}(Y_1 + Y_2 + \alpha_\pi,\ a + b - Y_1 - Y_2 + \beta_\pi), \tag{A3}$$

$$S \mid Y_1,Y_2,\alpha_S,\beta_S \sim \text{Beta}(Y_1 + \alpha_S,\ Y_2 + \beta_S), \tag{A4}$$

and

$$C \mid a,b,Y_1,Y_2,\alpha_C,\beta_C \sim \text{Beta}(b - Y_2 + \alpha_C,\ a - Y_1 + \beta_C). \tag{A5}$$

The Gibbs sampler operates as follows. Arbitrary starting values (see paragraph on convergence below) are chosen for each parameter. A sample of size m is then drawn from each full conditional distribution, in turn. The sampled values from the previous iterations are used in the conditional distributions for subsequent iterations. A cycle of the algorithm is completed when all conditional distributions have been sampled at least once. The entire cycle is repeated a large number of times. The random samples thus generated for each parameter can be regarded as a random sample from the correct posterior marginal distribution (8).

For the above model, Y_1 and Y_2 are generated from expressions A1 and A2, respectively, given the starting values of the other parameters. Then, π is generated from equation A3 conditional on the Y_1 and Y_2 variates just sampled. Drawing S and C from densities given in expressions A4 and A5, respectively, using the same values of Y_1 and Y_2 completes the first cycle. Positive and negative predictive values can be computed after each cycle from Y_1/a and $(b-Y_2)/b$, respectively. The random samples generated by repeating the above cycle the desired number of times are then used to reconstruct the marginal posterior densities of each parameter and to find credible sets, marginal posterior means or medians, or other inferences.

For two diagnostic tests, the full conditional distributions are as follows:

$$Y_1 \mid u,\pi,S_1,C_1,S_2,C_2 \sim \text{Binomial}\left(u,\frac{\pi S_1 S_2}{\pi S_1 S_2 + (1-\pi)(1-C_1)(1-C_2)}\right), \tag{A6}$$

$$Y_2 \mid v,\pi,S_1,C_1,S_2,C_2 \sim \text{Binomial}\left(v,\frac{\pi S_1(1-S_2)}{\pi S_1(1-S_2) + (1-\pi)(1-C_1)C_2}\right), \tag{A7}$$

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013

$$Y_3 \mid w,\pi,S_1,C_1,S_2,C_2 \sim \text{Binomial}\left(w,\frac{\pi(1-S_1)S_2}{\pi(1-S_1)S_2+(1-\pi)C_1(1-C_2)}\right), \quad \text{(A8)}$$

$$Y_4 \mid x,\pi,S_1,C_1,S_2,C_2 \sim \text{Binomial}\left(x,\frac{\pi(1-S_1)(1-S_2)}{\pi(1-S_1)(1-S_2)+(1-\pi)C_1C_2}\right), \quad \text{(A9)}$$

$$\pi \mid u,v,w,x,Y_1,Y_2,Y_3,Y_4,\alpha_\pi,\beta_\pi \sim \text{Beta}(Y_1+Y_2+Y_3+Y_4+\alpha_\pi, N-(Y_1+Y_2+Y_3+Y_4)+\beta_\pi), \quad \text{(A10)}$$

$$S_1 \mid Y_1,Y_2,Y_3,Y_4,\alpha_{S1},\beta_{S1} \sim \text{Beta}(Y_1+Y_2+\alpha_{S1}, Y_3+Y_4+\beta_{S1}), \quad \text{(A11)}$$

$$C_1 \mid u,v,w,x,Y_1,Y_2,Y_3,Y_4,\alpha_{C1},\beta_{C1} \sim \text{Beta}(w+x-(Y_3+Y_4)+\alpha_{C1}, u+v-(Y_1+Y_2)+\beta_{C1}), \quad \text{(A12)}$$

$$S_2 \mid Y_1,Y_2,Y_3,Y_4,\alpha_{S2},\beta_{S2} \sim \text{Beta}(Y_1+Y_3+\alpha_{S2}, Y_2+Y_4+\beta_{S2}), \quad \text{(A13)}$$

and

$$C_2 \mid u,v,w,x,Y_1,Y_2,Y_3,Y_4,\alpha_{C2},\beta_{C2} \sim \text{Beta}(v+x-(Y_2+Y_4)+\alpha_{C2}, u+w-(Y_1+Y_3)+\beta_{C2}). \quad \text{(A14)}$$

Gibbs sampling is used to sample in turn from distribution A6 to distribution A14 in a similar fashion to the procedure used for the case of one diagnostic test outlined previously. The positive and negative predictive values for each cycle of the Gibbs algorithm are again obtained directly from the relevant fractions of the true positive or negative subjects in each cell of the 2 × 2 table to the total observed number of subjects in that cell.

Throughout, the Gibbs sampler was run for 20,500 cycles, the first 500 to assess convergence and the last 20,000 for inference. Each analysis was repeated from several different starting values, and convergence was assumed only if all runs provided very similar posterior distributions. Convergence of the algorithm here appeared to occur within the first 100–200 cycles, as evidenced by the monitoring of selected percentiles of the posterior samples. In general, the rate of convergence will depend on the starting values and the particulars of the data set and prior distributions.

A computer program written in S-PLUS implementing all of the methods described in this paper is available from the first author (E-mail address: joseph@binky.epi.mcgill.ca).

Downloaded from http://aje.oxfordjournals.org/ at Oxford University Press for SER Members on January 22, 2013