Epidemiol. Infect. (2012), **140**, 1515–1524. © Cambridge University Press 2011
doi:10.1017/S0950268811001932

Association between covariates and disease occurrence in the presence of diagnostic error

F. LEWIS[1]*, M. J. SANCHEZ-VAZQUEZ[2] AND P. R. TORGERSON[1]

[1] *Vetsuisse Faculty, University of Zürich, Zürich, Switzerland*
[2] *Epidemiology Research Unit, SAC (Scottish Agricultural College), King's Buildings, West Mains Road, Edinburgh, UK*

(*Accepted 4 September 2011; first published online 23 September 2011*)

SUMMARY

Identification of covariates associated with disease is a key part of epidemiological research. Yet, while adjustment for imperfect diagnostic accuracy is well established when estimating disease prevalence, similar adjustment when estimating covariate effects is far less common, although of important practical relevance due to the sensitivity of such analyses to misclassification error. Case-study data exploring evidence for seasonal differences in *Salmonella* prevalence using serological testing is presented, in addition simulated data with known properties are analysed. It is demonstrated that: (i) adjusting for misclassification error in models comprising continuous covariates can have a very substantial impact on the resulting conclusions which can then be drawn from any analyses; and (ii) incorporating prior knowledge through Bayesian estimation can provide potentially more informative assessments of covariates while removing the assumption of perfect diagnostic accuracy. The method presented is widely applicable and easily generalized to many types of epidemiological studies.

Key words: Epidemiology, statistics, veterinary epidemiology.

INTRODUCTION

A primary objective of many epidemiological studies is to test hypothesized relationships, for example between specific covariates of interest and some response variable, denoting say the presence of, or exposure to, a pathogen or parasite [1–5]. If, however, the method of diagnosis used to classify subjects as disease positive (negative) suffers from imperfect sensitivity and/or specificity then the observed response variable is an estimate of the diagnosis positive fraction of subjects in the study population – typically referred to as apparent prevalence. In contrast, what is actually desired is an estimate of the disease positive fraction – the true prevalence. Moreover, analyses of diseases which have low prevalence represent a particular challenge because even when using a diagnostic with very high specificity, false positives may be more numerous than true positives.

A complication in performing any data analyses is that the true prevalence in a study with an imperfect diagnostic is not directly observed, but rather contained latently within the data collected. The analytical challenge for the epidemiologist is to release this latent information, and thus enable estimation of the effects of the covariates of interest, after adjusting for diagnostic misclassification. Analytical approaches for misclassification in regression models in epidemiology

* Author for correspondence: Dr F. Lewis, Vetsuisse Faculty, University of Zürich, Winterthurerstrasse 270, Zürich, Switzerland, CH 8057.
(Email: fraseriain.lewis@uzh.ch)

were introduced some years ago [6–8]; however, the use of such approaches is far from common practice.

It is intuitively obvious that additional uncertainty must be introduced into any analyses when an imperfect rather than gold-standard test is used, and thus a resulting reduction in statistical power. What may be surprising in practice is just how large an impact this may have on the conclusions which can then be drawn from any analyses. This is the key message of this article, i.e. that misclassification error, can and should, always be investigated to ensure scientifically credible conclusions of regression analyses involving imperfect diagnostics. The impact of an imperfect diagnostic is first examined using simulated data, followed by analyses which explore temporal (seasonal) fluctuations in *Salmonella* prevalence in farmed pigs. *Salmonella* is zoonotic, and one of the main risks from this pathogen to humans is through the consumption of pork products which accounts for up to 20% of human salmonellosis cases found in some European countries, and it remains the second highest (regarding occurrence) in the list of human zoonotic diseases across the European Union [9–12]. Frequentist and Bayesian estimations are both considered and viewed as complementary approaches.

There exists an extensive literature on the study and development of methods for analysing results from imperfect diagnostic tests, both from frequentist and Bayesian perspectives [13–17]. Adopting a no-gold standard (NGS) approach does, however, come with numerous caveats [15], and unlike many other statistical estimation methods the results of model fitting are not generally testable against the observed data, as the variables being estimated are latent. It is, however, very difficult to argue that an NGS approach should not at least be considered whenever study results are based on a diagnostic which is not known to be a gold-standard test against the specific study population. This is particularly relevant in analyses focused on estimation of the effect of covariates, particularly continuous covariates, as the true functional relationship between the response variable and covariates is generally unknown and hence needs to be estimated from the observed data. Any additional variance in the response variable due to misclassification error may considerably affect the accuracy with which any functional form can be estimated, with the danger of attributing erroneous covariates to a particular disease status.

Model identifiability, i.e. are there sufficient degrees of freedom available given a particular study design to estimate latent variables, is a particular challenge in analyses concerned with imperfect diagnostic testing [18]. In theory, it is highly desirable that a model is identifiable as otherwise some parameters may be entirely redundant, in which case they should arguably be removed from any model. For example, in a binomial distribution it is possible to formulate the parameter p, the probability of observing a subject with disease as $p = S_e\pi + (1 - S_p)(1 - \pi)$, where S_e and S_p are the sensitivity and specificity of the diagnostic test used, respectively, and π the true disease prevalence within the population. The terms p and π here are the apparent prevalence and true prevalence, respectively. The goodness of fit to observed data will only depend on the value of p, and no matter what combination of values are chosen for S_e, S_p, π they cannot improve the goodness of fit to the data, and hence cannot be uniquely estimated.

Much of the existing methodological NGS literature focuses on the problem of prevalence estimation using multiple imperfect tests applied to single or multiple study populations to ensure model identifiability [13]. In practice, however, it may be that for a model to be identifiable biologically untenable assumptions may be required, which is, at least in part, a practical reason for choosing a Bayesian estimation approach where identifiability is much less of a concern as prior information can be used to avoid this issue. For epidemiological studies which utilize only a single diagnostic test, which is common particularly in larger scale studies, estimation of the latent prevalence of disease is still possible but requires alternative approaches [5, 19].

In terms of previous methodological approaches to misclassification in regression models, log-linear analyses where information on error rates is provided through the availability of supplementary information, and logistic regression approaches with examples using discrete covariates have been available for many years [6, 7, 20]. However, the uptake of such approaches in practice has been very limited with a paucity of studies, particularly temporal and/or spatial studies which are a key aspect of zoonotic epidemiological research [21].

This article describes how to estimate the effects of covariates in relation to the unobserved 'true' prevalence in a population when an imperfect diagnostic test is used, and explores the resulting impact on the precision of the parameters of interest. In the analyses presented, the simplest case of a single continuous covariate is considered which is both sufficient for

illustrative purposes, and also epidemiologically relevant in a zoonotic context as demonstrated in the later section *Salmonella* case study. Models with both continuous and discrete covariates and using multiple imperfect tests are deferred to the Discussion. Of related interest is using frequentist and Bayesian estimations as complementary approaches, with the former assisting with issues of parameter stability and robustness, while the latter allows for the incorporation of valuable prior information which can be used to decrease variance/increase power.

METHODS

While Bayesian approaches to imperfect diagnostic test estimation are widespread [22], and increasingly accessible through software which implements efficient Markov chain Monte Carlo (MCMC) estimation such as WinBUGS and JAGS [23], a maximum-likelihood (ML) approach is considered alongside Bayesian estimation. An advantage of using ML estimation is that issues and difficulties around model identifiability are, generally speaking, more immediately recognizable compared to a Bayesian approach where the influence of the prior information used can be sufficient to mask a model which would be unidentifiable given the observed data alone. This is practically relevant because if a model is unidentifiable in a ML context, then in a Bayesian analyses – while WinBUGS/JAGS will probably have no problem in producing parameter estimates – any results may be very highly sensitive to the prior information used (e.g. from expert opinion) which may be undesirable.

Latent variable logistic regression

A binomial regression model with a logit link function between the latent true prevalence and covariates associated with disease occurrence can be defined as follows; for covariate pattern i,

$$\begin{pmatrix} \Pr(Y_i = y_i|n_i) = \binom{n_i}{y_i} q_i^{y_i}(1-q_i)^{n_i-y_i}, \text{ for } 0 \leqslant y_i \leqslant n_i \\ \text{where } q_i = S_e\pi_i + (1-S_p)(1-\pi_i) \\ \text{and } \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\beta, \end{pmatrix} \quad (1)$$

with q_i the probability that a subject with the ith covariate pattern tests positive (apparent prevalence) and π_i the probability that a subject with the ith covariate pattern is disease positive (true prevalence), where the latter is parameterized as a function of covariates β. The transposed vector x_i^T represents the ith row of the design matrix X, i.e. the combination of model parameters which will be used to estimate the ith covariate pattern. The parameters S_e and S_p are, respectively, the sensitivity and specificity of the diagnostic used. When $S_e = 1$ and $S_p = 1$ then the model reverts to the classical logistic regression model where $q_i = \pi_i$ (note that when $S_e < 1$ or $S_p < 1$ then q_i does not have a logit link function).

The key aspect in fitting the model in equation (1) to data is the complication that the π_is are latent parameters. The expectation maximization (EM) algorithm is a standard approach for ML estimation in the presence of unobserved variables [15, 24]. Technical implementation details of an EM algorithm for a latent variable logistic regression model can be found in the Supplementary material (available online) along with suitable R code. Model fits were assessed both visually against the data and also more formally using the ML ratio test and Akaike's Information Criterion (AIC) metric.

Later, Bayesian estimation is applied to the same form of model using MCMC via JAGS and all computer code is again provided as Supplementary material to allow easy replication of the results presented. As is good practice in Bayesian analyses a large number of different chains were run from many different initial starting points and of different lengths, including some very long runs of many millions of iterations. Trace plots for individual parameters and deviances were compared and examined to identify signs of poor behaviour such as very slow mixing, and also to assess convergence along with the use of the usual Gelman & Rubin convergence diagnostic [25].

Simulated data

Analysing simulated data where the true parameters are known with certainty provides a means of investigating the behaviour and utility of latent variable approaches. We consider the case of a single covariate where the logit of the (mean) true prevalence is modelled by a straight line where x_i^T comprises of $(1, x_{1,i})$, and $\beta^T = (\beta_0, \beta_1)$, hence $\text{logit}(\pi_i) = \beta_0 + x_{1,i}\beta_1$. In the simulated data $\beta_0 = -2$, $\beta_1 = 10$ with $x_{1,i}$ taking 24 equally spaced values from 0·2 to 0·4 and $S_e = 0{\cdot}7$ and $S_p = 0{\cdot}9$. The simulated data were created in a straightforward fashion by first generating the (linear and deterministic) mean on a logit scale, inverting to the probability scale (i.e. true prevalence π_i), then

creating an apparent probability (q_i) for each covariate pattern. Finally, Bernoulli observations were generated, e.g. coin tosses where 'heads' and 'tails' are replaced with diagnosis positive or diagnostic negative, where q_i is the probability of observing a diagnosis positive observation at covariate pattern i.

Of primary interest is exploring the impact of allowing additional variance into the regression relationship as a result of diagnostic test error. To this end two different sample sizes were considered, first with $n=100$ independent Bernoulli observations for each covariate pattern, e.g. for each value of $x_{1,i}$, and then a larger dataset with $n=2000$ Bernoulli observations per covariate pattern. This gives a total of 2400 and 48 000 individual diagnostic test results, respectively. These are deliberately large sample sizes – if the impact of misclassification error is appreciable in such relatively large sample sizes then it is reasonable to expect this to be larger in smaller datasets. Moreover, while 2400 test results might be large relative to certain types of human epidemiological studies, this is small relative to data collection in certain zoonotic contexts, e.g. the later *Salmonella* case study in farmed pigs has over 8000 observations, and in general food safety and meat inspection data can comprise very large numbers of observations. Each simulated dataset provides a potential maximum of 24 degrees of freedom (D.F.) and fitting a standard logistic regression model (where $S_e=S_p=1$) requires 2 D.F. For the latent variable model to be identifiable it requires that the data contain sufficient additional information to also allow for S_e and S_p to be estimated.

Two different sets of prior distributions are considered when modelling this data, first, where S_e and S_p have independent Beta distributions, specifically $\beta(1, 1)$ (equivalent to uniform on 0, 1), and with β_0 and β_1 having diffuse independent normal priors with mean zero and variance 1000. Next, the priors for S_e and S_p are made more informative using $\beta(70, 30)$ and $\beta(90, 10)$, respectively, where these have means of 0·7 and 0·9 and roughly equate to a total prior weighting of 100 Bernoulli observations (in a simple conjugate model with a binomial density – and equal to about 1·2% of the weight of the observed data).

Model identifiability and parameter estimation

Two aspects are of particular interest, first is the latent variable model identifiable. Assessing the identifiability of any given model in a formal mathematical setting is challenging [18] and still largely an open question. However, this can be assessed relatively easily empirically. If the algorithm used to estimate the model parameters produces different estimates, e.g. starting the algorithm from different starting points gives different estimates but which have identical maximum log-likelihood values, then that is strong evidence that the likelihood is at least in parts completely flat, and hence the model is not identifiable. Note that in practice some allowance may need to be made for numerical approximation errors in whatever algorithms are used. Even if the maximal log-likelihood values are not identical but only very similar (for different parameter estimates), then a similar issue exists since this suggests that the standard errors for at least some of the model parameters are likely be very large, and therefore give results which will not be statistically significant.

Assuming a model is identifiable, the second aspect of particular interest is estimating the uncertainty in the parameter(s) of interest, in the case of the simulated data given above this is, β_1, the covariate associated with the presence of disease. To estimate confidence intervals profile likelihood is used [26] in all (non-Bayesian) analyses presented. Profile likelihood is a method for estimating confidence intervals in the presence of 'nuisance parameters', for example, to estimate a confidence interval for β_1 we need to also take into account the uncertainty in the other unknown parameters in the model, e.g. β_0, S_p, S_e.

In the context of estimating a confidence interval for β_1 then β_0, S_p, S_e can be considered as nuisance variables, similarly when estimating β_0 then β_1, S_p, S_e become nuisance variables. We consider confidence interval estimation for β_1. First, we plot the log-likelihood function for the latent variable logistic regression model over a range of values for β_1, where for each value of β_1 the maximum possible value of the log-likelihood function is used (allowing the nuisance parameters to take any values in their range) – this is called the profile likelihood for β_1. Figure 1(c, d) shows profile likelihood functions for β_1 constructed in this way. A $(1-\alpha)$ confidence interval for β_1, where for example $\alpha=0{\cdot}05$ gives a 95% confidence interval (95% CI), is found by drawing a horizontal line at a value of $0{\cdot}5\chi^2_{d;1-\alpha}$ below the maximum value of the profile likelihood function where $\chi^2_{d;1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2 probability distribution with d degrees of freedom. For a confidence interval for a single parameter then $d=1$, for a joint confidence

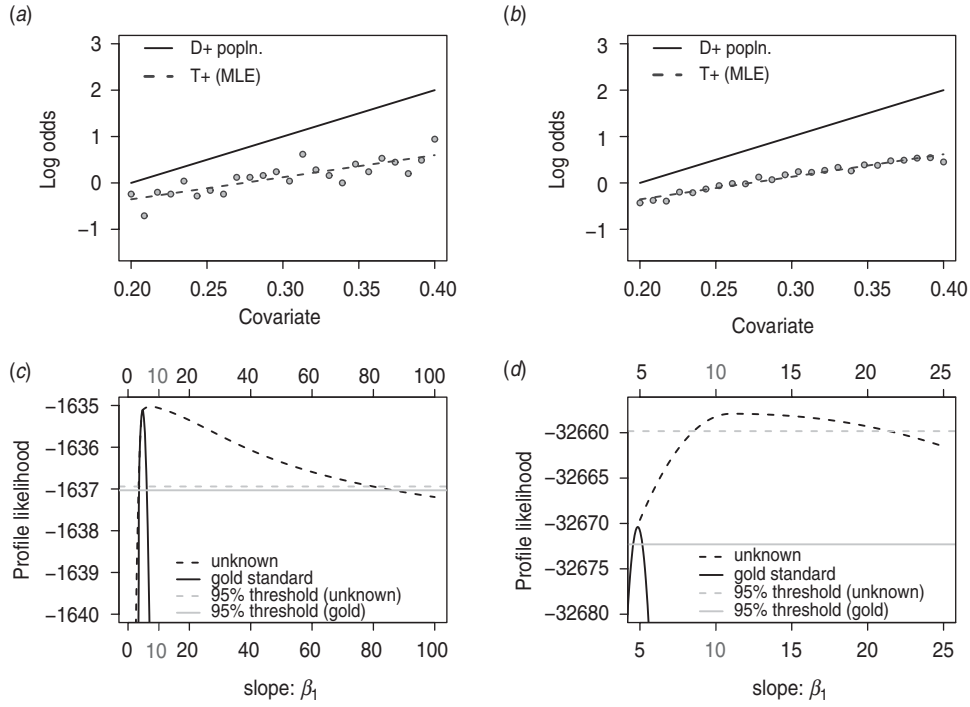

Fig. 1. Analysis of simulated data. (*a*, *b*) Raw data with corresponding fitted trend lines assuming the diagnostic test used was a gold-standard test [apparent prevalence (dashed line), true prevalence (solid line)] as a function of the covariate. As expected there is considerable difference between apparent and true prevalence. (*a*) $n=100$ per covariate pattern, test positive (T+) and true prevalence (D+); (*b*) $n=2000$ per covariate pattern. (*c*, *d*) Estimates for the slope parameters in panels (*a*) and (*b*). 95% confidence intervals (defined as where the horizontal lines cross the profile likelihood) show there is great uncertainty in the slope when the test is not a gold standard and this is still considerable even for the much larger sample size. (*c*) $n=100$ per covariate pattern; (*d*) $n=2000$ per covariate pattern.

interval of two parameters then $d=2$ and so on. Therefore, a 95% CI for β_1 is given by where a horizontal line crosses the profile likelihood function for β_1 at a distance $0{\cdot}5\chi^2_{1;0{\cdot}95} \approx 1{\cdot}92$ below the maximum value of this function (Fig. 1*c*, *d*). It is also possible to estimate joint 95% CIs in an analogous fashion, in the two-parameter case this interval (or region) is now defined by a horizontal cross-section across a profile likelihood surface (see Fig. 2).

Salmonella case study data

Data comprising ELISA serological test results for exposure to *Salmonella* from 8028 individual finishing pigs destined for human consumption in the UK were analysed. Visual inspection of the data suggests considerable seasonal fluctuations in *Salmonella* prevalence, with a strong peak in late summer and rapid decline during the winter months for the two years (2008 and 2009) for which data were available. *Salmonella* seasonal variation has been observed before, and is potentially associated with an increase of the proliferation of bacteria in the farm environment during the warmer months and an increase in *Salmonella* shedding by infected pigs resulting from heat stress [9, 27]. This is directly relevant to human health as it implies that the risk of *Salmonella* in pork products may be greater during the warmer months, which may have implications for risk assessment strategies designed to minimize *Salmonella* entering the food chain. A particular question of interest is whether analyses using standard binomial logistic regression (assuming the ELISA is a gold-standard test) support seasonal variation in prevalence, and what is the impact of removing this unsupported assumption through fitting a latent variable logistic regression model. Robust estimates of sensitivity and specificity of the ELISA used when applied to the relevant UK pig population are unavailable.

RESULTS

Simulated data

Figure 1(*a*, *b*) shows the simulated data along with corresponding fitted trend lines assuming the

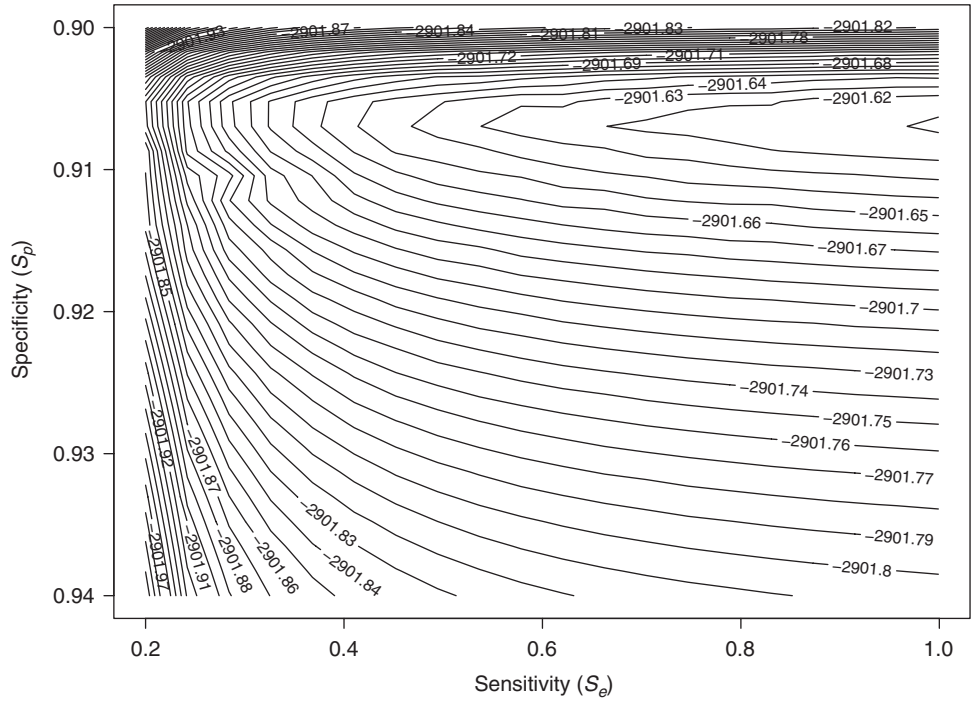

Fig. 2. Profile likelihood surface for S_e and S_p estimated from *Salmonella* data. There is great uncertainty in the estimate of S_e but relatively less in S_p. The MLE is $(S_e, S_p) = (0{\cdot}99, 0{\cdot}907)$ with the critical value defining a 95% confidence set within this surface at $-2904{\cdot}61$ (outside the limits shown).

diagnostic test used was a gold-standard test, and the true prevalence in the population as a function of the covariate. Figure 1(*c*, *d*) contains the profile likelihood for β_1 and corresponding confidence intervals for the effect of this covariate. The latent variable model is identifiable with a unique solution of $(\beta_0, \beta_1, S_e, S_p) = (-1{\cdot}38, 7{\cdot}48, 0{\cdot}76, 0{\cdot}999\,997)$ with maximum log-likelihood of $-1635{\cdot}02$ for $n = 100$, and $(\beta_0, \beta_1, S_e, S_p) = (-2{\cdot}12, 11{\cdot}54, 0{\cdot}68, 0{\cdot}953)$ with maximum log-likelihood of $-32\,657{\cdot}91$ for $n = 2000$. For the standard regression model the maximum log-likelihood values are $-1635{\cdot}11$ and $-32\,670{\cdot}38$, respectively, for $n = 100$ and $n = 2000$; note that these are not the same as those in the latent variable model as S_e and S_p are not implicitly contained within the parametrization of the standard model. While the ML solutions are unique, the likelihood surface is relatively flat resulting in comparatively low parameter precision, as can be seen in the wide confidence intervals for β_1 (Fig. 1*c*, *d*, Table 1); in contrast, 95% CIs for β_1 in the gold-standard model are relatively very precise. It is instructive to compare results using Bayesian estimation with the ML results. Table 1 shows 95% CIs for β_1 using 2·5% and 97·5% quantiles.

Salmonella case study

Figure 3*a* shows the observed data – the apparent prevalence of finishing pigs for *Salmonella* split by calendar month, along with the best-fit regression line using a standard binomial regression model. Polynomials in month (as a dummy variable 1–12) of increasing order were considered and both the ML ratio test and AIC strongly support a cubic relationship as the best-fitting model (AIC for orders 0–4: 5831·861, 5815·318, 5814·606, 5811·872, 5813·813). Treating the test as a gold standard, and considering the relatively narrow range of (cubic) trajectories which fit within the 95% prediction limits for the conditional mean (Fig. 3*a*), suggests that the magnitude of this change in prevalence is potentially worthy of further epidemiological investigation.

We now consider instead that the ELISA has not been shown to be a gold standard against this specific study population, and moreover, suppose than no reliable information is available in respect of likely true and false positive rates against the study population. Using ML estimation to fit a cubic polynomial latent variable model gives the maximum-likelihood point estimate (MLE) $(\beta_0, \beta_1, \beta_2, \beta_3, S_e,$

Table 1. *Analysis of simulated data* [*95% confidence intervals for slope parameter* ($\beta 1$) *by sample size and model type*]

Model	$n=100$ per covariate pattern	$n=2000$ per covariate pattern
ML ($S_e=S_p=1$)	(3·64, 6·18)	(4·61, 5·21)
ML ($S_e\neq S_p\neq 1$)	(3·84, 80·10)	(8·64, 21·46)
Bayesian prior1 ($S_e\neq S_p\neq 1$)	(3·95, 54·66)	(9·73, 20·19)
Bayesian prior2 ($S_e\neq S_p\neq 1$)	(3·55, 22·44)	(9·52, 14·73)

ML, Maximum-likelihood model.
Prior 1 is $\beta(1, 1)$ on both S_e and S_p; prior 2 is $\beta(70, 30)$ and $\beta(90, 10)$ for S_e and S_p, respectively.

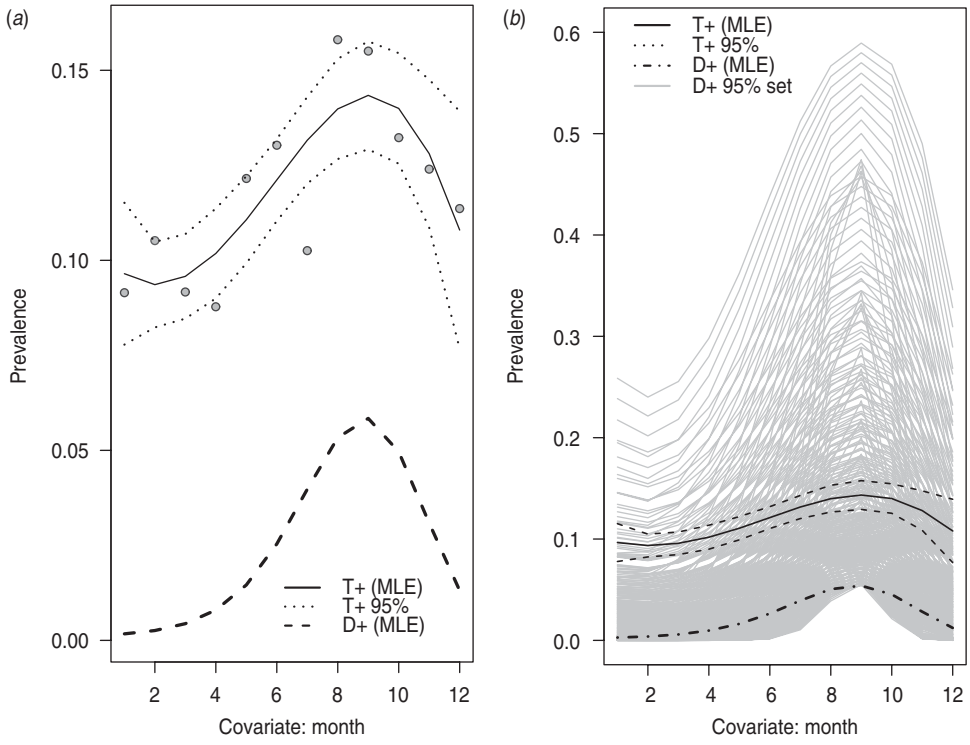

Fig. 3. (*a*) Analysis of *Salmonella* data. There are great differences between true and apparent prevalence. Total sample size of 8028 pigs, apparent prevalence (T+), latent true prevalence (D+). (*b*) After accounting for the impact of the imperfect test it is not possible to draw any conclusions as to seasonal changes in true prevalence. Range of trajectories corresponds to the joint 95% confidence set for S_e and S_p.

S_p) = (−5·98, 0·0033, 0·12, −0·0091, 0·99, 0·907), the true prevalence estimate (Fig. 3*a*) is clearly well outside the 95% range of trajectories for the gold-standard equivalent (maximum log-likelihood is −2901·61, with −2901·94 in the gold-standard model).

The existing range of gold-standard prevalence trajectories need to be adjusted to take into account the implicitly estimated sensitivity and specificity of the *Salmonella* ELISA. Figure 3*b* shows cubic polynomial trajectories where each of these corresponds uniquely to a point in the joint 95% confidence set for the sensitivity and specificity (S_e, S_p). Once this additional uncertainty has been included, not only does the MLE in the standard binomial model appear strongly biased relative to the 'true' prevalence MLE, but the additional uncertainty introduced due to ignorance of likely prior distributions for (S_e, S_p) is so large that in practical terms it is impossible to draw any substantive conclusions as to the relationship between calendar month and *Salmonella* prevalence.

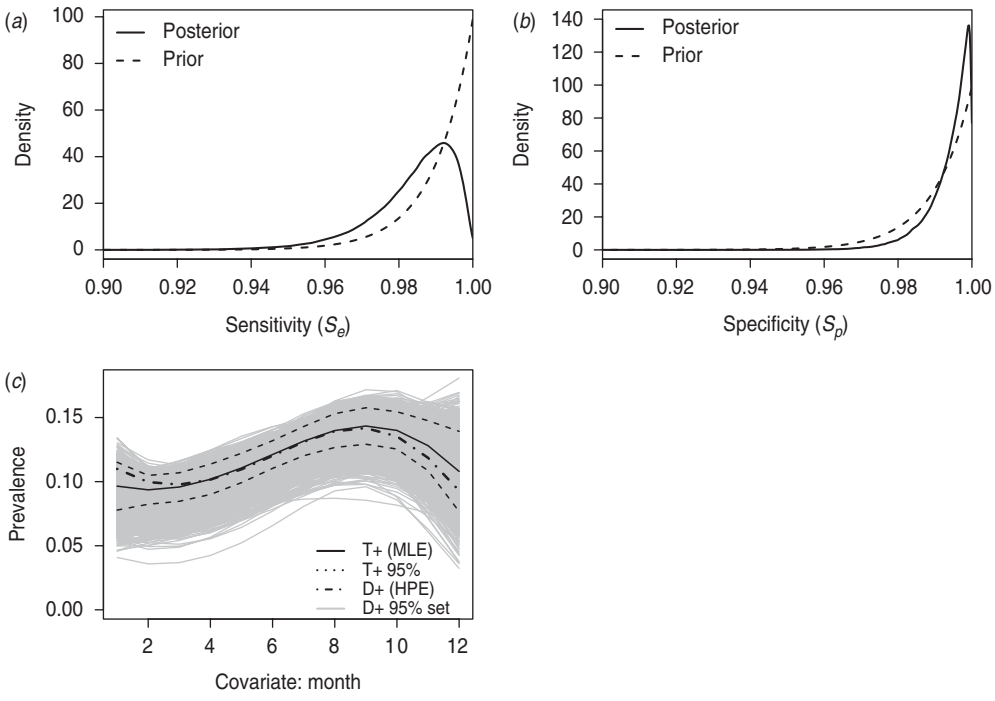

Fig. 4. Bayesian estimation of *Salmonella* data using extremely strong priors of $\beta(99, 1)$, close to perfect accuracy, for S_e and S_p. There still exists a very large amount of uncertainty in estimates of true prevalence over time and much more than assuming a gold-standard test. (*a*) Prior and posterior densities for S_e. (*b*) Prior and posterior densities for S_p. (*c*) Range of trajectories corresponding to the top 95% of log-likelihood values sampled during Markov chain estimation of the Bayesian latent variable model. The trajectory estimate with highest posterior log-likelihood is also shown.

Figure 2 shows an estimate of the joint (profile) likelihood surface for (S_e, S_p). The critical value – the value of the profile likelihood surface at the cross-section required to produce a joint 95% confidence set for (S_e, S_p) – is −2904·61, and as is clear from the contour plot that while it is not completely flat, and hence strictly speaking the model is identifiable, it is so flat that this results in the very wide range of trajectories for the monthly prevalence estimates shown in Figure 3*b* (the Supplementary online material contains an additional contour plot with wider ranges).

Given the very wide range of uncertainty in the ML estimation it is instructive to investigate how much this could be reduced by the introduction of prior information, for example sourced from a smaller scale study on the same target population whose aim was estimation of diagnostic accuracy. Figure 4 shows results from using Bayesian estimation with prior distributions for S_e and S_p of $\beta(99, 1)$ and $\beta(99, 1)$, respectively, with β_i for $i=0$, 1, 2, 3 having diffuse independent normal priors with mean zero and variance 1000. These are highly informative priors with low variance and very high accuracy for both sensitivity and specificity. Not surprisingly, the variance in trajectories for the true prevalence in the Bayesian model is extremely smaller than in the ML model (with unknown S_e and S_p), but still rather larger than the gold-standard ML model (Fig. 4*c*).

As a practically important footnote, estimating the Bayesian model required some care. In a Bayesian analyses the usual objective is to estimate probability distributions for each parameter of interest (which can then provide, for example confidence intervals), which are a combination of the data available and prior knowledge. The key part of this estimation process when using WinBUGS/JAGS is that they attempt to describe (e.g. map) a possibly very complex surface which defines the joint probability distribution for all parameters in the model. They do this by effectively 'walking' around this landscape taking different sized steps, depending on the specific algorithm used. During this process it is possible to get 'trapped' in a particular area of the surface and therefore the map of the landscape it produces will be at best incomplete, if not entirely unreliable. In the *Salmonella* analyses many different 'walks' (Markov chains) were performed from different initial starting points in the landscape, and for varying numbers of steps which is generally considered good practice in

MCMC analyses. About half of the chains appeared to get trapped indefinitely sampling around a mode with log-likelihood of about −2915, whereas other chains sampled around a mode with log-likelihood of about −2902, the latter corresponding to the ML analyses. Some chains eventually reached the ML mode but some did not even after many millions of iterations (steps). The ML analyses was valuable in assessing whether the results from the Bayesian estimation procedure were robust (technical note: chains sampling from the node at −2915 had values of about 1·0 for the usual Gelman & Rubin convergence diagnostic [25] indicating convergence, and similarly for the chains sampling around the ML node). Summary outputs and diagnostics can be found in the Supplementary material (online).

DISCUSSION

There are two key results from the analyses presented. First, that even a relatively small margin of misclassification error in the response variable in regression models can considerably increase the variance of the estimated functional relationship. Second, the methods presented for adjusting for misclassification error in regression models are relatively straightforward, especially in a Bayesian context via MCMC, e.g using either WinBUGS or JAGS (JAGS code is provided with the Supplementary material). As illustrated with the case study analyses, however, performing appropriate diagnostics and examining issues such as sensitivity to priors are then of crucial importance. The key practical implication of the work presented is that some form of adjustment for misclassification error should be considered in any study unless there is overwhelming biological evidence that the diagnostic used is error free.

The methods presented here have particular application to the study of rare diseases in populations. Although highly specific diagnostic tests may be developed, the specificity is often <1·0. This results in poor positive predictive values of the tests as the frequency of false positives in the population may exceed that of true positives. Consequently, there is a substantial risk of making inappropriate conclusions on covariates associated with disease presence as they may be substantially influenced by covariates associated with individuals who test false positive. For example a population study of human cystic echinococcosis, a highly pathogenic parasitic disease caused by *Echinococcus granulosus*, in Kazakhstan reported an estimated prevalence of 0·011 [28]. As part of the study the population was tested with an ELISA using cyst fluid antigens isolated from *E. granulosus*. This test had a specificity of 0·990 but the positive predictive value was just 0·382. Hence using standard logistic regression to evaluate covariates in the population associated with infection status, using positive test results from this ELISA is likely to produce substantial errors.

Only polynomial models in a single continuous covariate have been considered; however, the methods demonstrated are equally applicable to richer forms of regression models, e.g. with more complex variance structure, and also potentially comprising both continuous and discrete covariates. Models to identify temporal or spatial correlations with disease are particularly obvious candidates. Model identifiability is again a consideration as this is likely to become more difficult and challenging the more complex the model considered for a given dataset (see Supplementary material for a brief discussion). Also of relevance is available sample size; in the simulated data example the sample sizes used were relatively large, yet the variance in the regression coefficients was still considerable. This again, suggests that some form of informative prior may be of significant benefit even in larger studies.

NOTE

Supplementary material accompanies this paper on the Journal's website (http://journals.cambridge.org/hyg).

ACKNOWLEDGEMENTS

We are grateful to the BPEX Pig Health Scheme for providing the data for the analyses in this study.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Magalhaes RJS, *et al.*** Risk factors for methicillin-resistant *Staphylococcus aureus* (MRSA) infection in dogs and cats: a case-control study. *Veterinary Research* 2010; **41**: 55.
2. **Schweizer G, *et al.*** Prevalence of *Fasciola hepatica* in the intermediate host *Lymnaea truncatula* detected by real

time Taqman PCR in populations from 70 Swiss farms with cattle husbandry. *Veterinary Parasitology* 2007; **150**: 164–169.
3. **Leblebicioglu H, *et al.*** Outbreak of tularemia: a case-control study and environmental investigation in turkey. *International Journal of Infectious Diseases* 2008; **12**: 265–269.
4. **Katagiri S, Oliveira-Sequeira TCG.** Prevalence of dog intestinal parasites and risk perception of zoonotic infection by dog owners in Sao Paulo state, Brazil. *Zoonoses and Public Health* 2008; **55**: 406–413.
5. **Lewis FI, *et al.*** Bayesian inference for within-herd prevalence of *Leptospira interrogans* serovar Hardjo using bulk milk antibody testing. *Biostatistics* 2009; **10**: 719–728.
6. **Espeland MA, Hui SL.** A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics* 1987; **43**: 1001–1012.
7. **Magder LS, Hughes JP.** Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 1997; **146**: 195–203.
8. **Gustafsen P.** *Measurement Error and Misclassificaion in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman and Hall/CRC, 2004.
9. **Hald T, *et al.*** The occurrence and epidemiology of *Salmonella* in European pig slaughterhouses. *Epidemiology and Infection* 2003; **131**: 1187–1203.
10. **White DG, *et al.*** The isolation of antibiotic-resistant salmonella from retail ground meats. *New England Journal of Medicine* 2001; **345**: 1147–1154.
11. **Berends BR, *et al.*** Identification and quantification of risk factors in animal management and transport regarding *Salmonella* spp. in pigs. *International Journal of Food Microbiology* 1996; **30**: 37–53.
12. **Horwood J.** Infectious disease surveillance update. *Lancet Infectious Diseases* 2008; **8**: 220–220.
13. **Hui SL, Walter SD.** Estimating the error rates of diagnostic tests. *Biometrics* 1980; **36**: 167–171.
14. **Joseph L, Gyorkos TW, Coupal L.** Bayesian-estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; **141**: 263–272.
15. **Pepe MS, Janes H.** Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007; **8**: 474–484.
16. **Dendukuri N, Belisle P, Joseph L.** Bayesian sample size for diagnostic test studies in the absence of a gold standard: comparing identifiable with non-identifiable models. *Statistics in Medicine* 2010; **29**: 2688–2697.
17. **Enoe C, Georgiadis MP, Johnson WO.** Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* 2000; **45**: 61–81.
18. **Jones G, *et al.*** Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* 2010; **66**: 855–863.
19. **Brülisauer F, *et al.*** The prevalence of bovine viral diarrhoea virus infection in beef suckler herds in Scotland. *Veterinary Journal* 2010; **186**: 226–231.
20. **Qu YS, Hadgu A.** A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association* 1998; **93**: 920–928.
21. **Wang XH, *et al.*** Bayesian spatio-temporal modeling of *Schistosoma japonicum* prevalence data in the absence of a diagnostic 'gold' standard. *PLoS Neglected Tropical Diseases* 2008; **2**: e250.
22. **Branscum AJ, Gardner IA, Johnson WO.** Bayesian modeling of animal- and herd-level prevalences. *Preventive Veterinary Medicine* 2004; **66**: 101–112.
23. **Plummer M.** JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik K, *et al.*, eds. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria, 20–22 March 2003.
24. **Dempster AP, Laird NM, Rubin DB.** Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B* (*Methodological*) 1977; **39**: 1–38.
25. **Brooks SP, Gelman A.** General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; **7**: 434–455.
26. **Venzon DJ, Moolgavkar SH.** A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*) 1988; **37**: 87–94.
27. **Smith RP, Clough HE, Cook AJC.** Analysis of meat juice elisa results and questionnaire data to investigate farm-level risk factors for salmonella infection in UK pigs. *Zoonoses and Public Health* 2010; **57**: 39–48.
28. **Torgerson PR, *et al.*** Echinococcosis, toxocarosis and toxoplasmosis screening in a rural community in eastern Kazakhstan. *Tropical Medicine & International Health* 2009; **14**: 341–348.