Supplementary Information
Identifying correlates to disease in the presence of diagnostic error

version 1.0

Submitted to JournalName *June 1, 2011*

1. Expectation Maximization: Technical Details

The expectation maximization algorithm searches for an optimal solution to the log-likelihood function through iteratively maximizing the expected log-likelihood function of the complete data, that is where the latent variable denoting that a given subject is disease positive is assumed known. At step 0, an initial "guess" for the model parameters provides estimates of the probabilities of being disease positive, these probabilities are then fed into the expected log-likelihood function which is then maximized with respect to the model parameters. This proceeds iteratively until the algorithm converges to an optimal solution. From [1] if the latent variable D - true disease status (true=1, false=0) - were observed the log-likelihood for the ith subject Y_i is

$$\log L_i^c(\pi_i, \theta) = D_i \log\{\pi_i P_\theta(Y_i \mid D_i = 1)\} + (1 - D_i) \log\{(1 - \pi_i) P_\theta(Y_i \mid D_i = 0)\} \tag{1}$$

where c denotes an individual case, $P_\theta(.)$ the probability mass function with parameters $\theta = (\phi, \psi)$, where ϕ and ψ are the true and false positive rates respectively (ϕ=sensitivity, ψ=1-specificity), and π_i is the true latent prevalence of disease $P(D_i = 1)$. Given values for $\pi_i = \pi_i^\star$, e.g. $\pi_i^\star = \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\}/(1+\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\})$, where π_i is parametrized as a function of covariates $\boldsymbol{\beta}^T = (\beta_0, \ldots, \beta_m)$. The transposed vector $\boldsymbol{x}_i^T$ represents the ith row of the design matrix $\boldsymbol{X}$. The expected log-likelihood is

$$\begin{aligned} E_{\pi_i^\star, \theta^\star}(\pi_i, \theta) &= \sum_{i=1}^{n} E(\log L_i^c(\pi_i, \theta) \mid Y_i), \\ &= \sum_{i=1}^{n} \Big[P(D_i = 1 \mid Y_i, \pi_i^\star, \theta^\star)\{\log \pi_i + \log P_\theta(Y_i \mid D_i = 1)\} \\ &\quad + \ P(D_i = 0 \mid Y_i, \pi_i^\star, \theta^\star)\{\log(1 - \pi_i) + \log P_\theta(Y_i \mid D_i = 0)\} \Big], \end{aligned} \tag{2}$$

and note that $P(D_i = 1 \mid Y_i, \pi_i^\star, \theta^\star) = 1 - P(D_i = 0 \mid Y_i, \pi_i^\star, \theta^\star)$ are known constants with

$$P(D_i = 1 \mid Y_i, \pi_i^\star, \theta^\star) = \frac{P_{\theta^\star}(Y_i \mid D_i = 1)\pi_i^\star}{P_{\theta^\star}(Y_i \mid D_i = 1)\pi_i^\star + P_{\theta^\star}(Y_i \mid D_i = 0)(1 - \pi_i^\star)}$$
$$\text{where} \quad \pi_i^\star = \frac{\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\}}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\}}.$$

Using a logistic link function between π_i and $\boldsymbol{\beta}$ then

$$\begin{aligned} \log(\pi_i) &= \log\left(\frac{\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\}}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\}}\right) &=& \quad \boldsymbol{x}_i^T \boldsymbol{\beta} - \log(1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}) \\ \text{and} \quad \log(1 - \pi_i) &= \log\left(\frac{1}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}^\star\}}\right) &=& \quad -\log(1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}) \end{aligned}$$

The expected log-likelihood, the function to be maximized is therefore

$$\begin{aligned} l_E \;=\; \sum_{i=1}^{n} \Big[& c_{1i}(\boldsymbol{x}_i^T \boldsymbol{\beta} - \log(1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}) + Y_i \log\phi + (1 - Y_i)\log(1 - \psi)) \\ & + (1 - c_{1i})(- \log(1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}) + Y_i \log\psi + (1 - Y_i)\log(1 - \psi)) \Big], \end{aligned} \tag{3}$$

where $c_{1i} = P(D_i = 1 \mid Y_i, \pi_i^\star, \theta^\star)$.

At each step in the EM algorithm the function in (3) is maximized to give a new solution, $(\beta_0, \ldots, \beta_m, \phi, \psi)$, which is then used to calculate new estimates for $c_{1i} = P(D_i = 1 \mid Y_i, \pi_i, \theta)$, and then the process is repeated. Note that for this model the maximization at each iteration must be done numerically rather than analytically. A reasonably reliable numerical method for this optimization applied to the data presented in the main manuscript was the quasi-Newton method with box constraints[2].

1.1. R scripts

R scripts for running the above EM algorithm with the model $\log(\pi)/\log(1 - \pi) = \beta_0 + \beta_1 X_1$ are available in the accompanying files `functions_R.r` and `runEM_R.r`. The results in the main manuscript, including profile likelihoods, were produced using analogous code but with the functions compiled in C and dynamically loaded into R for improved computational efficiency.

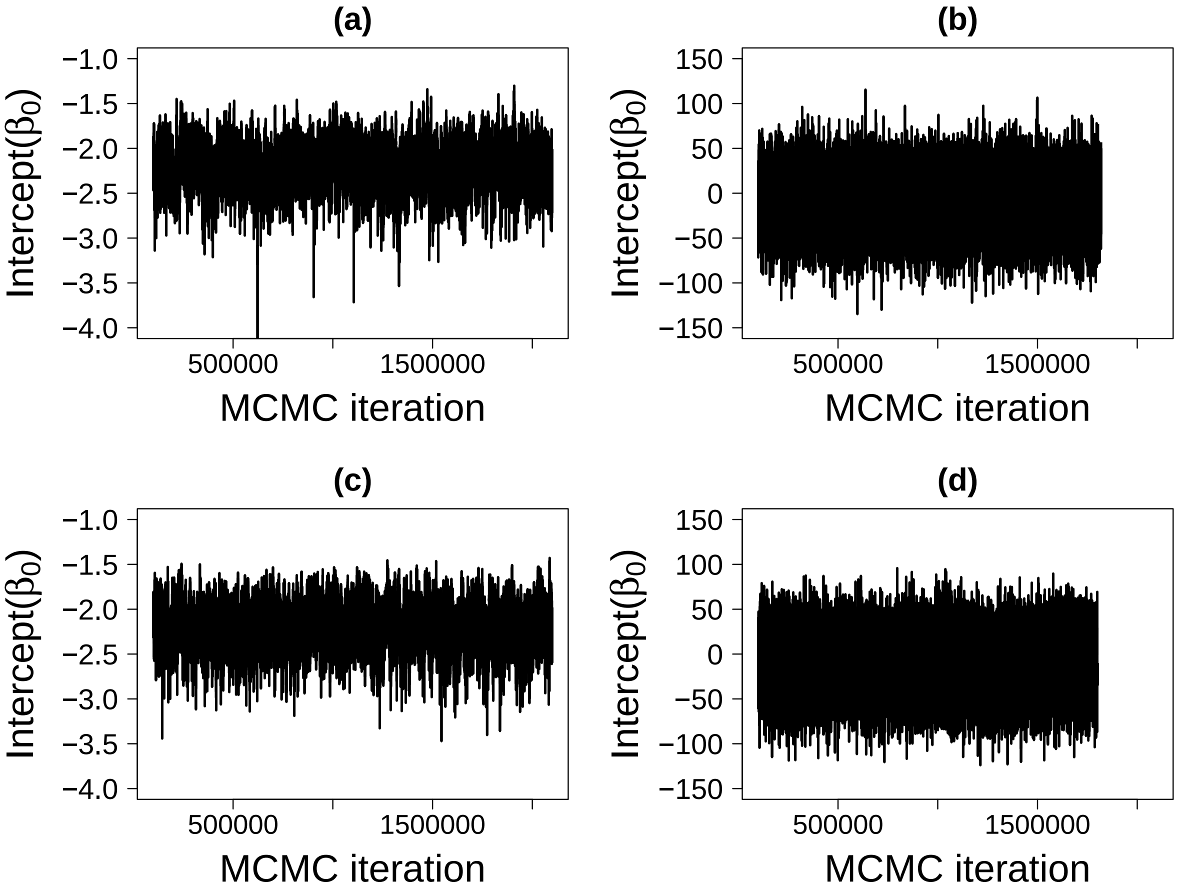
2. Bayesian Analyzes

The following files: `binom.bug; inits_1.r; script_1.r; binom_dat_n_100.r` can be run in JAGS using the command "jags script_1.r" which will run the MCMC estimation and produce output files which can then be read into R with the CODA library.

2.1. Diagnostic outputs

As mentioned in the main text some of the Markov chains got "stuck" at a sub-optimal node. Figure 1 shows four separate MCMC chains for the latent variable binomial regression model for salmonella data. An additional four chains were run and in total three sampled around one node and five around another with lower log-likelihood. Visual inspection, as can be seen from the figure, suggests adequate mixing across all the chains, and the Gelman and Rubin diagnostic was 1.00 for all the parameters in the three chains sampling around the node with highest log-likelihood, and similarly for the five chains sampling around the node with lower log-likelihood.

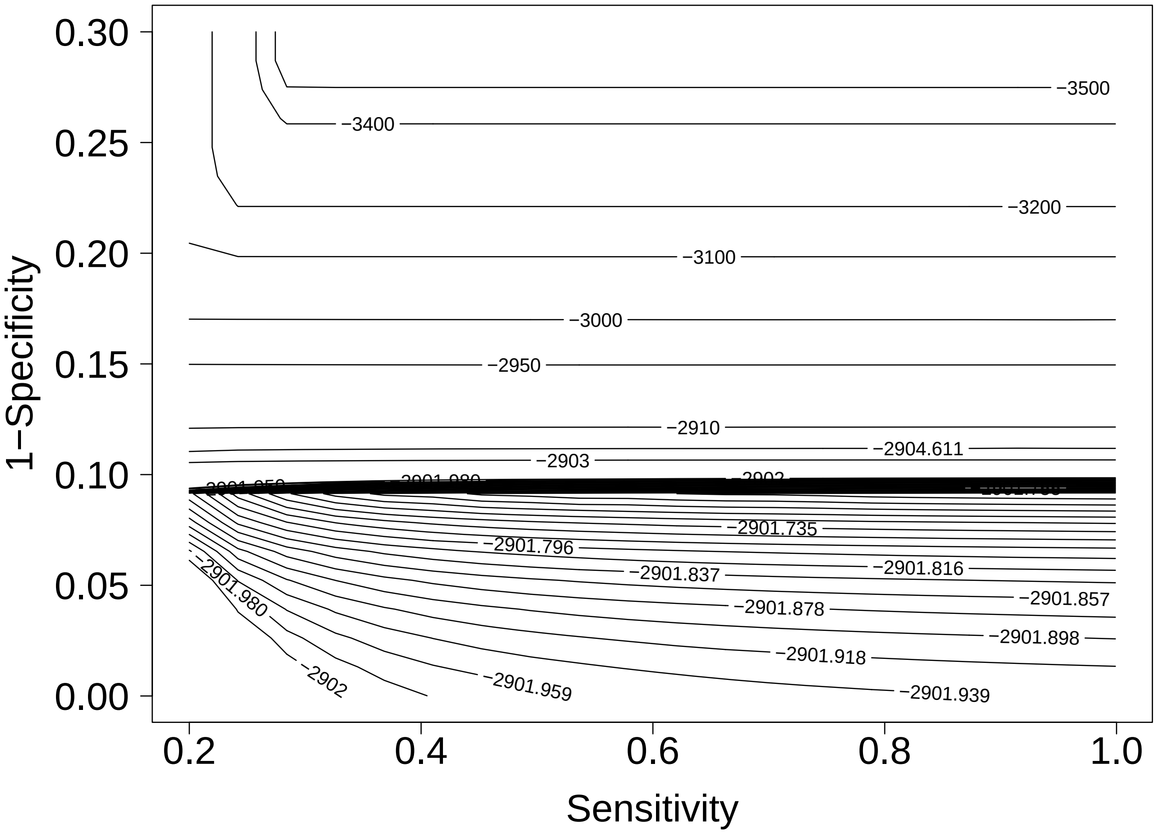
Figure 1: MCMC trace output for latent variable binomial regression model for Salmonella data. (a)-(d) four separate runs (burn-in of 1×10^6 not shown) with (a) and (c) sampling around a node with log-likelihood of approximately -2902, and (b) and (d) sampling around a node with log-likelihood of approximately -2915.



3. Additional Figures

3.1. Contour plot for $(\phi, \psi) = (S, 1 - C)$

Figure 2: Profile likelihood surface for true and false positive errors rates in Salmonella data. The MLE is $(\phi, \psi) = (0.99, 0.093)$ with the critical value for a 95% confidence set within this surface at -2904.61



4. Some comments on method applicability and model identifiability

The application and estimation of latent variable binomial regression models to epidemiological studies does requires some care and may not be suited to all types of studies which seek to identify correlates to disease. Model identifiability is an important consideration, which is essential for a ML analyzes, and desirable although not essential for Bayesian estimation. Conditions for the identifiability of latent variable models containing covariates is an open question and will likely be problem specific. For example, if time were treated as a fully continuous variable in the analyses presented - which does not make biological sense in this example - then only single Bernoulli observations would be available at each time point and it is unclear whether such a model would or would not be identifiable, as fitting a linear model would still only use the same number of parameters but there is less information available per covariate pattern, but, much more information is available at many more different patterns. As mentioned, if using a model which is not identifiable then model sensitivity to priors in a Bayesian analyzes is of particular importance. Although as demonstrated, it is likely that for such latent variable models to be of most practical use, relatively strong prior information may be necessary. Considering ML estimation as well as a Bayesian approach may be useful in diagnosing any issues of robustness with the latter.

The analyzes presented in the main manuscript have only considered a single imperfect test, however, the methods used could be readily extended to consider multiple tests along with all of the additional complications which that entails, for example covariance between tests. Estimation with multiple imperfect tests is well studied in the literature and any of the established parameterizations could be readily incorporated into the regression estimation framework (either ML or Bayesian) presented. Other complexities could include for example allowing the sensitivity and specificity of the diagnostic test to be dependent[3]. These are all obvious areas for future application.

References

1. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007;8(2):474–484.

2. Byrd RH, Lu PH, Nocedal J, et al. A limited memory algorithm for bound constrained optimization. *Siam Journal On Scientific Computing* 1995; 16(5):1190–1208.

3. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International Journal of Epidemiology* December 2005;34(6):1370–1376.