Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

The Web Appendix for “G-scores: A method for identifying disease-causing pathogens with application to lower respiratory tract infections” by Zhang et al.

Statistical Inference of the proposed models in Section 3

In this section, we illustrate in details how one can utilize the data augmentation procedure and Markov chain Monte Carlo (MCMC) algorithm to implement the statistical inference of the proposed models. We focus on demonstrating how to make inference for the model with baseline covariates. The model without adjustment of baseline covariates can be similarly implemented.

The joint likelihood function of the model is

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2;\boldsymbol{X},\boldsymbol{Z}) = \prod_{i=1}^{N}\Bigg[&\Big\{1-\lambda(\boldsymbol{Z}_i)+\lambda(\boldsymbol{Z}_i)\Phi\Big(\frac{3-\mu(\boldsymbol{Z}_i)}{\sigma}\Big)\Big\}^{I(X_i=0)} \\ &\Big\{\frac{\lambda(\boldsymbol{Z}_i)}{\sigma}\varphi\Big(\frac{X_i-\mu(\boldsymbol{Z}_i)}{\sigma}\Big)\Big\}^{I(X_i>0)}\Bigg] \end{aligned} \tag{1}$$

For notation convenience, we shall denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \sigma^2)^T$. Following the classical likelihood theory, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is asymptotically normally distributed:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \mathcal{I}_{\boldsymbol{\theta}_0}^{-1}),$$

where $\boldsymbol{\theta}_0$ is the true parameters corresponding to data generating process and $\mathcal{I}_{\boldsymbol{\theta}_0}$ is the Fisher information matrix. $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ and an estimate of $\mathcal{I}_{\boldsymbol{\theta}_0}$ can be used to construct confidence intervals of $\boldsymbol{\theta}_0$. However, computation of these quantities is intensive, and complicated by the fact that the density function is a mixture distribution and hence the likelihood function could have multiple local maxima. We propose to use Markov chain Monte Carlo method to overcome this difficulty.

For a given prior distribution $\pi(\boldsymbol{\theta})$, for instance $\pi(\boldsymbol{\theta}) \propto 1/\sigma^2$, we can calculate the posterior distribution of $\boldsymbol{\theta}$ through Bayes’ rule

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z}) = \frac{L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

This conditional density function defines a random probability measure, where randomness comes from $\boldsymbol{X}$ and $\boldsymbol{Z}$. We compare this with a random measure which corresponds to $N(\hat{\boldsymbol{\theta}}_{\text{MLE}}, \mathcal{I}_{\boldsymbol{\theta}_0}^{-1}/N)$. Following Bernstein-von Mises theorem (see chap. 10 in [1]), we have

$$\left\| p(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z}) - N(\hat{\boldsymbol{\theta}}_{\text{MLE}}, \frac{1}{N}\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}) \right\| \xrightarrow{P} 0, \tag{2}$$

where $\| \cdot \|$ stands for the total variation norm, i.e. for any signed measure m, $\|m\| \triangleq \sup_{A \in \mathcal{B}} |m(A)|$, and $\mathcal{B}$ is the collection of all Borel sets. This suggests that one can use $N \cdot \text{Var}(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z})$ to estimate $\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}$. Furthermore, the location functional, i.e. the expectation of the probability measure, is continuous with respect to the total variation norm. By applying the continuous mapping theorem on equation (2), we have

$$E(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{Z}) - \hat{\boldsymbol{\theta}}_{\text{MLE}} \xrightarrow{P} 0.$$

Hence, we just need to figure out how to sample from the posterior distribution. This can be achieved through data augmentation and Markov chain Monte Carlo as follows.

We introduce a vector of binary latent variables, $\boldsymbol{I} = (I_1, I_2, \ldots, I_N)$, where I_i indicates whether patient i is bacterial pathogen carrier. Hence

$$P(I_i = 1 \mid \boldsymbol{Z}_i) = \lambda(\boldsymbol{Z}_i), \quad P(I_i = 0 \mid \boldsymbol{Z}_i) = 1 - \lambda(\boldsymbol{Z}_i).$$

Given $\boldsymbol{Z}_i$, let $X_i^* = 0$ if $I_i = 0$ and $X_i^* \sim N(\mu(\boldsymbol{Z}_i), \sigma^2)$ if $I_i = 1$. $\boldsymbol{X^*} = (X_1^*, X_2^*, \ldots, X_N^*)$ could be thought of as the ideal qrt-LAMP measurements if they were not censored at 3. Let $\boldsymbol{X} = \boldsymbol{X^*} \times I(\boldsymbol{X^*} \geq 3)$, then $(\boldsymbol{X}, \boldsymbol{Z})$ would be the observed data. The joint distribution of $(\boldsymbol{X}, \boldsymbol{Z})$ leads to the likelihood in Equation (1). With the augmented data $(\boldsymbol{I}, \boldsymbol{X^*}, \boldsymbol{Z})$, we have a much simpler likelihood:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2; \boldsymbol{I}, \boldsymbol{X^*}, \boldsymbol{Z}) = \prod_{i=1}^{N} \left[\lambda(\boldsymbol{Z}_i)^{I_i} \big\{1 - \lambda(\boldsymbol{Z}_i)\big\}^{1-I_i} \left\{ \frac{1}{\sigma} \varphi\Big(\frac{X_i^* - \mu(\boldsymbol{Z}_i)}{\sigma} \Big) \right\}^{I_i} \right] \tag{3}$$

We can implement a Gibbs sampler by sequentially sampling from the conditional distribution $P\big((\boldsymbol{I}, \boldsymbol{X^*}) \mid \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}\big)$, $P\big((\boldsymbol{\gamma}, \sigma^2) \mid \boldsymbol{\beta}, \boldsymbol{I}, \boldsymbol{X^*}, \boldsymbol{Z}\big)$, and $P\big(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2, \boldsymbol{I}, \boldsymbol{X^*}, \boldsymbol{Z}\big)$. The second and third conditional densities are reduced to linear regression model and logistic regression model respectively, which can be derived as in Gelman et al. [2]. With the non-informative prior $\pi(\boldsymbol{\theta}) \propto 1/\sigma^2$, we demonstrate details as follows.

1. *Conditional posterior distribution of* $(\boldsymbol{I}, \boldsymbol{X^*})$. The conditional posterior distribution can be factorized as follows:

$$P\big\{(\boldsymbol{I}, \boldsymbol{X^*}) \mid \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}\big\} = \prod_{i=1}^{N} P\big\{(I_i, X_i^*) \mid \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}\big\}$$
$$= \prod_{i=1}^{N} P\big\{(I_i, X_i^*) \mid \boldsymbol{\theta}, X_i, \boldsymbol{Z}_i\big\} = \prod_{i=1}^{N} \Big\{ P\big(I_i \mid \boldsymbol{\theta}, X_i, \boldsymbol{Z}_i\big) P\big(X_i^* \mid \boldsymbol{\theta}, X_i, \boldsymbol{Z}_i, I_i\big) \Big\}$$

Hence, we first sample I_i from $P(I_i \mid \boldsymbol{\theta}, X_i, \boldsymbol{Z}_i)$, and then sample X_i^* from $P(X_i^* \mid \boldsymbol{\theta}, X_i, \boldsymbol{Z}_i, I_i)$. It is only necessary to show this when $X_i = 0$, since $(I_i, X_i^*) = (1, X_i)$ almost surely conditioning on $(\boldsymbol{\theta}, X_i > 0, \boldsymbol{Z}_i)$. Given $X_i = 0$, I_i can be updated according to

$$P\big(I_i = 1 \mid \boldsymbol{\theta}, X_i = 0, \boldsymbol{Z}_i\big) = \frac{\lambda(\boldsymbol{Z}_i) \Phi\big(\frac{3 - \mu(\boldsymbol{Z}_i)}{\sigma}\big)}{1 - \lambda(\boldsymbol{Z}_i) + \lambda(\boldsymbol{Z}_i) \Phi\big(\frac{3 - \mu(\boldsymbol{Z}_i)}{\sigma}\big)}.$$

Afterward, one can sample X_i^* from

$$p\big(X_i^* = x_i^* \mid \boldsymbol{\theta}, X_i = 0, \boldsymbol{Z}_i, I_i\big) = (1 - I_i)\delta_0 + I_i \frac{\frac{1}{\sigma} \varphi\big(\frac{x_i^* - \mu(\boldsymbol{Z}_i)}{\sigma}\big) I(x_i^* \leq 3)}{\Phi\big(\frac{3 - \mu(\boldsymbol{Z}_i)}{\sigma}\big)},$$

which is the mixture distribution between a degenerate distribution and a truncated normal distribution.

2. *Conditional posterior distribution of* (γ, σ^2). Conditional on the augmented data $(\boldsymbol{I}, \boldsymbol{X}^*, \boldsymbol{Z})$ and parameter $\boldsymbol{\beta}$, the posterior distribution is

$$p\big\{(\gamma, \sigma^2) \mid \boldsymbol{\beta}, \boldsymbol{I}, \boldsymbol{X}^*, \boldsymbol{Z}\big\} \propto \frac{1}{\sigma^2} \prod_{i=1}^{N} \bigg[\frac{1}{\sigma}\varphi\Big\{\frac{X_i^* - \mu(\boldsymbol{Z}_i)}{\sigma}\Big\}\bigg]^{I_i}, \tag{4}$$

This coincides with the posterior distribution of the linear regression model $E(X^* \mid \boldsymbol{Z}) = \mu(\boldsymbol{Z})$ with weight I and common random normal errors. In another word, we can filter out data with $I_i = 0$ and run regression model $\boldsymbol{X}^*$ on $\boldsymbol{Z}$ for the remaining data. With the non-informative prior we use, the posterior distribution can be written in an explicit form [2]. Let n be $\sum_{i=1}^{N} I_i$. Denote $\boldsymbol{X}^*$ by $\boldsymbol{W}$ and $\boldsymbol{Z}$ by $\boldsymbol{V}$ after removing entries with $I_i = 0$, where $\boldsymbol{W}$ is a $n \times 1$ vector and $\boldsymbol{V}$ is a $n \times 5$ matrix. It can be seen from (4) that $P\big\{(\gamma, \sigma^2) \mid \boldsymbol{\beta}, \boldsymbol{I}, \boldsymbol{X}^*, \boldsymbol{Z}\big\}$ only depends on $\boldsymbol{W}$ and $\boldsymbol{V}$, so we can write $P\big\{(\gamma, \sigma^2) \mid \boldsymbol{\beta}, \boldsymbol{I}, \boldsymbol{X}^*, \boldsymbol{Z}\big\}$ as

$$P\big\{(\gamma, \sigma^2) \mid \boldsymbol{W}, \boldsymbol{V}\big\} \propto \sigma^{-n-2} \exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{W} - \boldsymbol{V}\gamma)^T(\boldsymbol{W} - \boldsymbol{V}\gamma)\Big\}. \tag{5}$$

Thus

$$\begin{aligned} P\big\{(\gamma, \sigma^2) \mid W, V\big\} &\propto \sigma^{-n-2} \exp\big\{-\frac{1}{2\sigma^2}(\boldsymbol{W} - \boldsymbol{V}\hat{\gamma} + \boldsymbol{V}\hat{\gamma} - \boldsymbol{V}\gamma)^T(\boldsymbol{W} - \boldsymbol{V}\hat{\gamma} + \boldsymbol{V}\hat{\gamma} - \boldsymbol{V}\gamma)\big\} \\ &= \sigma^{-n-2} \exp\Big\{-\frac{1}{2\sigma^2}\big[vs^2 + (\gamma - \hat{\gamma})^T \boldsymbol{V}^T \boldsymbol{V}(\gamma - \hat{\gamma})\big]\Big\} \\ &= \Big\{(\sigma^2)^{-(n-5)/2} \exp\big(-\frac{vs^2}{2\sigma^2}\big) \times (\sigma^2)^{-1}\Big\} \times \Big\{(\sigma^2)^{-5/2} \exp\big(-\frac{(\gamma - \hat{\gamma})^T \boldsymbol{V}^T \boldsymbol{V}(\gamma - \hat{\gamma})}{2\sigma^2}\big)\Big\}, \end{aligned}$$

where

$$\begin{aligned} \hat{\gamma} &= (\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T\boldsymbol{W} & \qquad v &= n - 5 \\ s^2 &= (\boldsymbol{W} - \hat{\boldsymbol{W}})^T(\boldsymbol{W} - \hat{\boldsymbol{W}})/v & \qquad \hat{\boldsymbol{W}} &= \boldsymbol{V}\hat{\gamma}. \end{aligned}$$

Therefore,

$$\begin{aligned} \sigma^2 \mid \boldsymbol{W}, \boldsymbol{V} &\sim \text{Inv-}\chi^2(v, s^2) \\ \boldsymbol{\gamma} \mid \sigma^2, \boldsymbol{W}, \boldsymbol{V} &\sim N\big(\hat{\gamma}, \sigma^2(\boldsymbol{V}^T\boldsymbol{V})^{-1}\big), \end{aligned}$$

where Inv-χ^2 denotes the scaled inverse-χ^2 distribution. σ^2 and $\boldsymbol{\gamma}$ can then be sampled in a two-step manner.

3. *Conditional posterior distribution of* $\boldsymbol{\beta}$. Conditional on the augmented data $(\boldsymbol{I}, \boldsymbol{X}^*, \boldsymbol{Z})$ and parameter $\boldsymbol{\beta}$, its posterior distribution is

$$p\big(\boldsymbol{\beta} \mid \gamma, \sigma^2, \boldsymbol{I}, \boldsymbol{X}^*, \boldsymbol{Z}\big) \propto \prod_{i=1}^{N} \Big[\lambda(\boldsymbol{Z}_i)^{I_i}\big\{1 - \lambda(\boldsymbol{Z}_i)\big\}^{1-I_i}\Big], \tag{6}$$

which corresponds to exactly a logistic regression model, $E(I \mid \boldsymbol{Z}) = \lambda(\boldsymbol{Z})$. Though the posterior distribution can not be directly sampled from, one can fit the weighted least square regression on pseudo-data $\hat{u}$ with their corresponding variances $\hat{v}^2$, which are defined as

$$\hat{u_i} = \hat{\eta_i} + \frac{(1 + e^{\hat{\eta_i}})^2}{e^{\hat{\eta_i}}}\left(I_i - \frac{e^{\hat{\eta_i}}}{1 + e^{\hat{\eta_i}}}\right),$$

$$\hat{v_i}^2 = \frac{(1 + e^{\hat{\eta_i}})^2}{e^{\hat{\eta_i}}}, \qquad \hat{\eta_i} = \boldsymbol{Z}_i\hat{\boldsymbol{\beta}}.$$

Denote $\boldsymbol{W}$ as the pseudo data vector, $\boldsymbol{V}$ as a diagonal matrix with $\hat{v_i}^2$ as its elements, and $\boldsymbol{Z}$ as the design matrix. The weighted least square solution $\hat{\boldsymbol{\beta}}$ follows multivariate normal distribution

$$N\big\{(\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{W}, (\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\big\}. \tag{7}$$

As shown in Gelman et al. [2], multivariate normal distribution (7) approximates the posterior conditional distribution. Therefore, we can add a single Metropolis-Hastings step in the Gibbs sampler iterations. Denote $p^{\mathrm{N}}(\cdot)$ as the density function of (7). With $\boldsymbol{\beta}^{(t-1)}$, $\boldsymbol{\gamma}^{(t)}$, $\sigma^{2(t)}$, $\boldsymbol{I}^{(t)}$, $\boldsymbol{X}^{*(t)}$ and $\boldsymbol{Z}$, we generate $\boldsymbol{\beta}'$ from (7), and calculate

$$r \triangleq \min\{1, \frac{p(\boldsymbol{\beta}' \mid \boldsymbol{\gamma}^{(t)}, \sigma^{2(t)}, \boldsymbol{I}^{(t)}, \boldsymbol{X}^{*(t)}, \boldsymbol{Z}) p^{\mathrm{N}}(\boldsymbol{\beta}^{(t-1)})}{p(\boldsymbol{\beta}^{(t-1)} \mid \boldsymbol{\gamma}^{(t)}, \sigma^{2(t)}, \boldsymbol{I}^{(t)}, \boldsymbol{X}^{*(t)}, \boldsymbol{Z}) p^{\mathrm{N}}(\boldsymbol{\beta}')}\}.$$

We then assign $\boldsymbol{\beta}^{(t)}$ as

$$\boldsymbol{\beta}^{(t)} = \begin{cases} \boldsymbol{\beta}^{(t-1)} & \text{with probability } 1 - r \\ \boldsymbol{\beta}' & \text{with probability } r \end{cases}$$

The rejection probability for the Metropolis step is around 10%, indicating that the posterior conditional distribution is well approximated by (7).

The algorithm was implemented in C++. Random numbers were drawn with the default random number generator in the R stand-alone library. We ran 12000 iterations for the Gibbs sampler with the initial 2000 iterations as the burn-in period. The point estimates and interval estimates of the parameters were based on the later 10000 iterations. Diagnosis plots are shown below in this appendix.

References

1. Van der Vaart A. *Asymptotic statistics*. Cambridge Univ Pr: Cambridge, 2000.
2. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian data analysis*. 2nd edn., Champan and Hall/CRC: Boca Raton, 2004.

Table 1. P-Values for Model Checking. The left column shows p-values of the exact multinomial test for the null hypothesis that the observed SCM data follow zero-inflated binomial distribution with parameters estimated from Section 3. The entry for Pathogen Sma is blank due to insufficient positive results. The right column shows p-values of Kolmogorov-Smirnov test for the null hypothesis that the observed qrt-LAMP data follow zero-inflated truncated normal distribution with parameters estimated from Section 3.

Pathogen	Exact Multinomial Test	Kolmogorov-Smirnov Test
Spn	0.854	0.869
Sau	0.017	0.840
Eco	0.041	0.817
Kpn	0.053	0.886
Pae	0	0.974
Aba	0.283	0.994
Sma		0.864
Hin	0.746	0.808

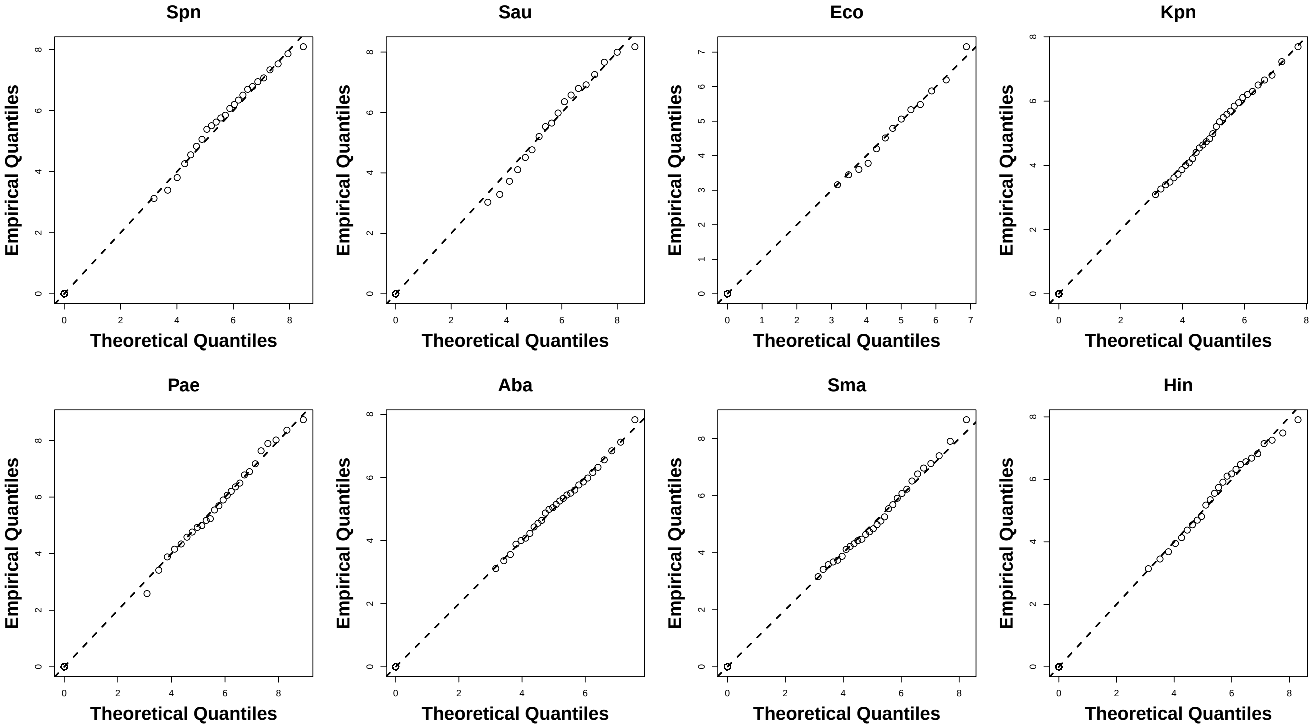


Figure 1. Q-Q Plots for Model Checking on qrt-LAMP Data of All Eight Pathogens. Circles are drawn at theoretical quantiles versus the corresponding empirical quantiles. Dashed lines are the 45° diagonal line.

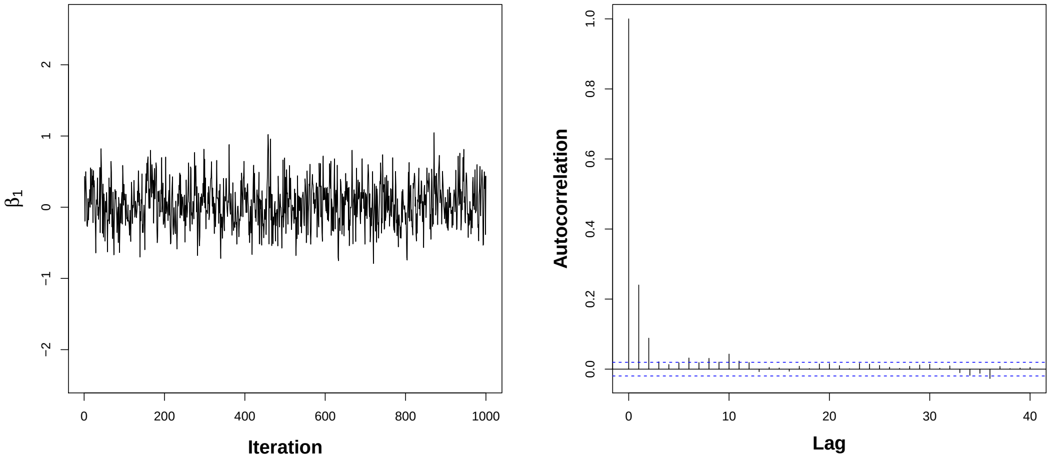
Table 2. Sensitivity and False Negative Rate. Panel (a) shows sensitivities of qrt-LAMP test for different pathogens estimated from both the marginal model and the regression model. Panel (b) shows false negative rates of qrt-LAMP test for different pathogens estimated from both models.

(a) Sensitivity

Pathogen	Marginal		Regression	
	Est.	SE.	Est.	SE.
Spn	0.960	0.013	0.955	0.018
Sau	0.909	0.024	0.917	0.027
Eco	0.819	0.047	0.823	0.051
Kpn	0.847	0.028	0.843	0.025
Pae	0.940	0.016	0.934	0.021
Aba	0.919	0.019	0.915	0.026
Sma	0.831	0.029	0.838	0.032
Hin	0.937	0.017	0.925	0.020

(b) False Negative Rate

Pathogen	Marginal		Regression	
	Est.	SE.	Est.	SE.
Spn	0.006	0.002	0.006	0.003
Sau	0.010	0.003	0.010	0.003
Eco	0.016	0.005	0.015	0.006
Kpn	0.034	0.007	0.035	0.007
Pae	0.010	0.003	0.010	0.004
Aba	0.015	0.004	0.016	0.005
Sma	0.036	0.007	0.034	0.008
Hin	0.010	0.003	0.012	0.003



(a) Trace Plot of β_1

(b) Autocorrelation Plot of β_1

Figure 2. Diagnosis for the Convergence of MCMC. We ran 12000 iterations in total with the first 2000 iterations as burn-in samples. Panel (a) shows the last 1000 iterations of Monte Carlo samples of β_1. Panel (b) shows the autocorrelation plot of 10000 Monte Carlo samples of β_1 without those 2000 burn-in samples.