Biostatistics (2016), **0**, 0, *pp.* 1–20
doi:10.1093/biostatistics/kxw037-supplementary-materials

Supplementary Materials to Nested Partially-Latent Class Models for Dependent Binary Data; Estimating Disease Etiology

ZHENKE WU*[,1], MARIA DELORIA-KNOLL[2], SCOTT L. ZEGER[1]

[1] *Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205*

[2] *Department of International Health, Johns Hopkins University, Baltimore, MD 21205*

zhwu@jhu.edu

Appendix S1. Mean and Covariance Structure

In this section, we present and discuss formulas for the model-based marginal observation rates and pairwise log odds ratios among cases and controls. They can be readily modified to accommodate "other" causes as discussed in Section 2.3 of the Main Manuscript.

Appendix S1.1 Marginal Observation Rate

The marginal observation rates are given by

$$\mathbb{P}(M_{i'j} = 1 \mid Y_{i'} = 1) = \pi_j \sum_{k=1}^{K} \theta_k^{(j)} \eta_k + (1 - \pi_j) \left\{ \sum_{k=1}^{K} \psi_k^{(j)} \eta_k \right\}, \tag{A1}$$

$$\mathbb{P}(M_{ij} = 1 \mid Y_i = 0) = \sum_{k=1}^{K} \psi_k^{(j)} \nu_k. \tag{A2}$$

In the context of childhood pneumonia problem, Equation (A1) indicates that the observed rate of pathogen j among cases comprises of two parts: cases whose disease is caused by pathogen

*To whom correspondence should be addressed.

j for which the observation is a true positive event, and those whose disease is caused by another pathogen for which the observation is a false positive.

The case and control mean observation rates for pathogen j are equal when either of Condition (I) or (II) below holds.

$$\text{(I) } \psi_1^{(j)} = \cdots = \psi_K^{(j)} = \psi^{(j)} \text{ and } \sum_{k=1}^{K} \theta_k^{(j)} \eta_k = \psi^{(j)};$$

$$\text{(II) } \boldsymbol{\eta} = \boldsymbol{\nu}, \text{ and } \sum_{k=1}^{K} \left[\theta_k^{(j)} \eta_k - \psi_k^{(j)} \nu_k \right] = 0.$$

The first part of Condition (I) says that the binary response on dimension j is constant across subclasses among controls, which implies independence of j-th dimension's measurement to other dimensions. The second part says, within the jth disease class, the marginal observation rate of dimension j equals the control rate.

Condition (II) means the case and control subclass weights are equal and the observation rates are also equal for the jth case class and controls. The multivariate binary distributions satisfying this condition are special cases of the non-interference submodels (Section 2.3 of the Main Manuscript).

Appendix S1.2 Marginal Pairwise Log Odds Ratios

The marginal pairwise log odds ratio $\omega_{j\ell}$ for pathogen pair (j, ℓ) among cases is given by:

$$\begin{aligned}
\omega_{j\ell} &= \log \left\{ \frac{\mathbb{P}(M_{ij}=1, M_{i\ell}=1)\mathbb{P}(M_{ij}=0, M_{i\ell}=0)}{\mathbb{P}(M_{ij}=1, M_{i\ell}=0)\mathbb{P}(M_{ij}=0, M_{i\ell}=1)} \right\} \\
&= \log \left(\sum_{c=1}^{J} \pi_c \left[\sum_{k=1}^{K} \left\{ \theta_k^{(j)} \right\}^{\mathbf{1}\{c=j\}} \left\{ \psi_k^{(j)} \right\}^{\mathbf{1}\{c\neq j\}} \left\{ \theta_k^{(\ell)} \right\}^{\mathbf{1}\{c=\ell\}} \left\{ \psi_k^{(\ell)} \right\}^{\mathbf{1}\{c\neq \ell\}} \eta_k \right] \right) \\
&\quad - \log \left(\sum_{c=1}^{J} \pi_c \left[\sum_{k=1}^{K} \left\{ 1-\theta_k^{(j)} \right\}^{\mathbf{1}\{c=j\}} \left\{ 1-\psi_k^{(j)} \right\}^{\mathbf{1}\{c\neq j\}} \left\{ \theta_k^{(\ell)} \right\}^{\mathbf{1}\{c=\ell\}} \left\{ \psi_k^{(\ell)} \right\}^{\mathbf{1}\{c\neq \ell\}} \eta_k \right] \right) \\
&\quad + \log \left(\sum_{c=1}^{J} \pi_c \left[\sum_{k=1}^{K} \left\{ 1-\theta_k^{(j)} \right\}^{\mathbf{1}\{c=j\}} \left\{ 1-\psi_k^{(j)} \right\}^{\mathbf{1}\{c\neq j\}} \left\{ 1-\theta_k^{(\ell)} \right\}^{\mathbf{1}\{c=\ell\}} \left\{ 1-\psi_k^{(\ell)} \right\}^{\mathbf{1}\{c\neq \ell\}} \eta_k \right] \right)
\end{aligned}$$

$$-\log\left(\sum_{c=1}^{J}\pi_c\left[\sum_{k=1}^{K}\left\{\theta_k^{(j)}\right\}^{\mathbf{1}\{c=j\}}\left\{\psi_k^{(j)}\right\}^{\mathbf{1}\{c\neq j\}}\left\{1-\theta_k^{(\ell)}\right\}^{\mathbf{1}\{c=\ell\}}\left\{1-\psi_k^{(\ell)}\right\}^{\mathbf{1}\{c\neq\ell\}}\eta_k\right]\right). \quad \text{(A3)}$$

Setting $K = 1$ in the formula gives log odds ratios for a locally independent model (Wu *and others*, 2016). When $K > 1$, suppose nearly all of pneumonia is caused by pathogen j: $\pi_j \approx 1$, we calculate $\omega_{j\ell}$ under two scenarios:

a) If the true positive rates for pathogen j across subclasses, i.e. $\theta_k^{(j)}, k = 1, ..., K$, are equal, then $\omega_{j\ell} \approx 0$, that is, we have approximate marginal independence between measurements on the jth pathogen and the rest among the cases;

b) If the number of subclasses $K = 2$ and true positive rates $\theta_k^{(j)}, k = 1, 2$ are very different, say, 1 versus 0 as an extreme example, we can show that $\omega_{j\ell} = \mathsf{logit}(\psi_1^{(\ell)}) - \mathsf{logit}(\psi_2^{(\ell)})$, which means the pairwise log odds ratio between pathogen j and ℓ among cases is determined by the variation of control subclass FPRs for the ℓth pathogen.

Appendix S2. Priors

Appendix S2.1 Prior Specifications

For the npLCM, we specify the prior distributions on unknown parameters as follows:

$$\boldsymbol{\pi} \sim \mathsf{Dirichlet}(a_1, \ldots, a_L), \quad \text{(A4)}$$

$$\psi_k^{(j)} \sim \mathsf{Beta}(b_{1kj}, b_{2kj}), j = 1, ..., J; k = 1, ..., \infty, \quad \text{(A5)}$$

$$\theta_k^{(j)} \sim \mathsf{Beta}(c_{1kj}, c_{2kj}), j = 1, ..., J; k = 1, ..., \infty, \quad \text{(A6)}$$

$$Z_i \mid Y_i = 1 \sim \sum_{k=1}^{\infty} U_k \prod_{s<k}[1 - U_s]\,\delta_k, \quad U_k \sim \mathsf{Beta}(1, \alpha_1), i = 1, ..., n_1, \quad \text{(A7)}$$

$$Z_i \mid Y_i = 0 \sim \sum_{k=1}^{\infty} V_k \prod_{s<k}[1 - V_s]\delta_k, \quad V_k \sim \mathsf{Beta}(1, \alpha_0), i = n_1 + 1, ..., n_1 + n_0, \quad \text{(A8)}$$

$$\alpha_0, \alpha_1 \sim \mathsf{Gamma}(0.25, 0.25), \quad \text{(A9)}$$

where δ_k is a point mass on k, and prior independence is also assumed among these parameters. As discussed in more detail by Wu *and others* (2016, p.7), the npLCM likelihood similarly has the TPRs $\boldsymbol{\Theta}$ that are not fully identified by the model likelihood and hence is partially identified (Jones *and others*, 2010). Therefore, we choose $(c_{1kj}, c_{2kj}), \forall k, j$, so that the 2.5% and 97.5% quantiles of the Beta distribution with parameters (c_{1kj}, c_{2kj}) match the prior minimum and maximum TPR values elicited from pneumonia experts (Section 5 in the Main Manuscript). Otherwise, we use the default value of 1s for the Beta hyperparameters. Hyperparameters for the etiology prior, $(a_1, ..., a_J)'$, are usually 1s to denote equal and flat prior weights for each pathogen if expert prior knowledge is unavailable.

Because our goal is to estimate the etiology fractions, $\boldsymbol{\pi}$, after marginalizing over subclass indicators (Z_i), the parameters for the dependence structure within each disease class are nuisance parameters. Therefore, rather than fixing K, we let K be a random positive integer and perform model averaging using a prior that encourages small values of K to incorporate its uncertainty into the inference about $\boldsymbol{\pi}$ in a parsimonious way. This prevents model overfitting in finite samples when the observed contingency table for the multivariate binary PERCH measurements has mostly empty cells. In (A7) and (A8), we have actually specified stick-breaking priors for both $\boldsymbol{\eta} = \left\{U_k \prod_{s<k} [1 - U_s]\right\}_{k=1,2,...}$ and $\boldsymbol{\nu} = \left\{V_k \prod_{s<k} [1 - V_s]\right\}_{k=1,2,...}$ that on average place decreasing weights on the kth subclass as k increases (Sethuraman, 1994).

Appendix S2.2 Stick-Breaking Prior

This section briefly discusses the stick-breaking priors used in the Bayesian inference for the nested partially-latent class models. A stick-breaking mixture model in theory has countably infinite number of subclasses. However, because the ν_k and η_k decrease exponentially quickly in k, *a priori*, we expect that only a small number of subclasses will be used to model the data. The expected number of subclasses from a stick-breaking prior is logarithmic in the number of

observations (Hjort *and others*, 2010). This is different than a finite mixture model, which uses a fixed number of clusters to model the data. In the stick-breaking mixture model, the actual number of clusters used to model data is not fixed, and can be automatically inferred from data using the usual Bayesian posterior inference framework (Neal, 2000).

Equations (A7)-(A9) place exchangeable prior weight on the subclasses. Following Ishwaran and James (2002), in our computations, we truncate the infinite sum to the first K^* terms with K^* sufficiently large to balance computing speed and approximating performance of the model. In our simulations and data application $K^* = 10$ is usually deemed adequate. Most subclass measurement profiles are not assigned with meaningful weights either in the simulations or in data application, so that a small number of effective subclasses are usually sufficient for approximation. Also, by placing hyperpriors on stick-breaking parameters α_0 and α_1 as in Equation (A9), we can let the data inform us about the desired sparsity level for approximating the probability contingency tables for the control and each disease class. A small value of the estimate $\widehat{\alpha}_0$ ($\widehat{\alpha}_1$) suggests that only a small number of subclasses are necessary for the controls (cases). We have chosen hyperparameters in the Gamma hyperpriors for α_1 and α_1 to be $(0.25, 0.25)$ which gives good parameter estimation performance in simulations.

Appendix S3. Gibbs Sampling Algorithm

For posterior computation involving stick-breaking priors, without truncation on the number of stick segments, Walker (2007) and Papaspiliopoulos and Roberts (2008) proposed the slice sampler and retrospective MCMC, respectively. In the following, we develop a simple and efficient blocked Gibbs sampler relying on truncation approximation to the stick-breaking prior distribution (e.g., Ishwaran and James, 2001; Gelfand and Kottas, 2002). We also include in the sampling algorithms two sets of auxiliary variables, the partially-latent individual class indicator (I_i) the nested subclass indicator (Z_i).

All model estimations are performed by the `R` package "`baker`" (https://github.com/zhenkewu/baker) that interfaces with freely available software `JAGS 3.4.0` (http://mcmc-jags.sourceforge.net/). Convergence was monitored via MCMC chain histories, auto-correlations, kernel density plots, and Brooks-Gelman-Rubin statistics (Brooks and Gelman, 1998). The statistical results below are based on $10,000$ iterations of burn-in followed by $50,000$ production samples from each of three parallel chains. Samples from every 50 iterations are retained for inference.

In the following are the MCMC sampling steps, assuming the truncation level is $K^* = K$:

1. Update the class indicator $I_{i'}$ for cases $i' = 1, ..., n_1$, from a categorical distribution with probabilities

$$\begin{aligned}\mathbb{P}(I_{i'} = j \mid \cdots) = p_{i'}^{(j)} &\propto [\mathbf{M}_{i'} \mid Z_{i'}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, I_{i'} = j][Z_{i'} \mid \boldsymbol{\eta}, I_{i'} = j][I_{i'} = j \mid \boldsymbol{\pi}] \\ &\propto \left\{\theta_{Z_{i'}}^{(j)}\right\}^{M_{i'j}} \left\{1 - \theta_{Z_{i'}}^{(j)}\right\}^{1-M_{i'j}} \prod_{l \neq j} \left\{\psi_{Z_{i'}}^{(l)}\right\}^{M_{i'l}} \left\{1 - \psi_{Z_{i'}}^{(l)}\right\}^{1-M_{i'l}} \cdot \eta_{Z_{i'}} \cdot \pi_j,\end{aligned}$$

   for $j = 1, ..., J$.

2. Update subclass indicators $Z_{i'}$ for case $i' = 1, ..., n_1$, from a categorical distribution with probabilities

$$\begin{aligned}\mathbb{P}(Z_{i'} = k \mid \cdots) = q_{i'k} &\propto [\mathbf{M}_{i'} \mid Z_{i'}, I_{i'}, \boldsymbol{\Theta}, \boldsymbol{\Psi}][Z_{i'} \mid I_{i'}, \boldsymbol{\eta}] \\ &\propto \eta_k \cdot \left\{\theta_k^{(I_{i'})}\right\}^{M_{i'I_{i'}}} \left\{1 - \theta_k^{(I_{i'})}\right\}^{1-M_{i'I_{i'}}} \prod_{l \neq I_{i'}} \left\{\psi_k^{(l)}\right\}^{M_{i'l}} \left\{1 - \psi_k^{(l)}\right\}^{1-M_{i'l}}.\end{aligned}$$

   Update subclass indicators Z_i for control $i = n_1 + 1, ..., n_1 + n_0$, from a categorical distribution with probabilities

$$\begin{aligned}\mathbb{P}(Z_i = k \mid \cdots) = q_{ik} &\propto [\mathbf{M}_i \mid Z_i = k, \boldsymbol{\Psi}][Z_i = k \mid \boldsymbol{\nu}] \\ &\propto \nu_k \cdot \prod_{j=1}^{J} \left\{\psi_k^{(j)}\right\}^{M_{ij}} \left\{1 - \psi_k^{(j)}\right\}^{1-M_{ij}}, k = 1, ..., K.\end{aligned}$$

3. Update the case subclass weights $\boldsymbol{\eta}$ for $j = 1, ..., J$ from

$$pr(\boldsymbol{\eta} \mid \cdots) \propto \prod_{i':I_{i'}=j} [Z_{i'} \mid \boldsymbol{\eta}, I_{i'}][\boldsymbol{\eta} \mid \alpha_1]$$

which can be accomplished by first setting $u_K^* = 1$ and sampling

$$u_k^* \sim \mathsf{Beta}\left(1 + z_k', \alpha_1 + \sum_{l=k+1}^{K} z_l'\right), k = 1, ..., K-1,$$

where z_k' is the number of cases assigned to subclass k in class j. We write

$$z_k' = \#\left\{i' : Y_{i'} = 1, Z_{i'} = k, I_{i'} = j\right\},$$

for $k = 1, ..., K-1$, where "$\#A$" counts the number of elements in set A. We then construct $\eta_1 = u_k^*$, $\eta_k = u_k^* \prod_{l=1}^{k-1} \{1 - u_l^*\}$, $k = 2, ..., K$.

4. Update the control subclass weights $\boldsymbol{\nu} = (\nu_1, ..., \nu_K)^T$ from

$$pr(\boldsymbol{\nu} \mid \cdots) \propto \prod_{i:Y_i=0} [Z_i \mid \boldsymbol{\nu}] \cdot [\boldsymbol{\nu} \mid \alpha_0],$$

which can be accomplished by first setting $v_K^* = 1$ and sampling

$$v_k^* \sim \mathsf{Beta}\left(1 + z_k, \alpha_0 + \sum_{l=k+1}^{K} z_k\right), k = 1, ..., K-1,$$

where z_k is the number of controls assigned to subclass k, and then constructing $\nu_1 = v_k^*$, $\nu_k = v_k^* \prod_{l=1}^{k-1}(1 - v_l^*)$, $k = 2, ..., K$.

5. Update concentration parameter α_0 and α_1 for stick-breaking prior from

$$pr(\alpha_0 \mid \cdots) \propto [\boldsymbol{\nu} \mid \alpha_0][\alpha_0] \propto \alpha_0^{K-1} \exp(-\alpha_0 \cdot r) \cdot pr(\alpha_0),$$

where $r = -\left\{\sum_{k=1}^{K-1} \log(1 - \nu_k^*)\right\}$. If conditionally conjugate prior for α_0 is used, i.e. $\alpha_0 \sim \mathsf{Gamma}(a_{\alpha_0}, b_{\alpha_0})$ with mean $a_{\alpha_0}/b_{\alpha_0}$ and variance $a_{\alpha_0}/b_{\alpha_0}^2$, then the full conditional distribution reduces to $\mathsf{Gamma}\left(a_{\alpha_0} + K - 1, b_{\alpha_0} + r\right)$. Similarly for α_1 with $\boldsymbol{\nu}$ replaced by $\boldsymbol{\eta}$ and $(a_{\alpha_0}, b_{\alpha_0})$ replaced by $(a_{\alpha_1}, b_{\alpha_1})$.

6. Update the vector of subclass TPR for $j = 1, ..., J$ from

$$pr(\boldsymbol{\theta}^{(j)} \mid \cdots) \propto \prod_{\{i':I_{i'}=j\}} [\mathbf{M}_{i'} \mid \boldsymbol{\theta}^{(j)}, Z_{i'}, I_{i'}][\boldsymbol{\theta}^{(j)}]$$

$$\propto \prod_{k=1}^{K} \left\{\theta_k^{(j)}\right\}^{m_{k1}^{(j)}} \left\{1-\theta_k^{(j)}\right\}^{m_{k0}^{(j)}} \cdot [\boldsymbol{\theta}^{(j)}],$$

where $m_{kc}^{(j)} = \#\{i' : Y_{i'} = 1, Z_{i'} = k, I_{i'} = j, M_{i'j} = c\}$, $c = 0, 1$. If prior for TPRs are independent Beta distributions, then this is a product of Beta distributions.

7. Update subclass-specific FPRs $\psi_k^{(j)}$ for $j = 1, ..., J$, $k = 1, ..., K$ from

$$pr(\psi_k^{(j)} \mid \cdots) \propto \prod_{i':Y_{i'}=1, I_{i'}\neq j, Z_{i'}=k} [M_{i'j} \mid \boldsymbol{\psi}^{(j)}, Z_{i'}, I_{i'}] \prod_{i:Y_i=0} [M_{ij} \mid \boldsymbol{\psi}^{(j)}, Z_i] \cdot [\psi_k^{(j)}]$$
$$\propto \left\{\psi_k^{(j)}\right\}^{s_{k1}^{(-j)}} \left\{1-\psi_k^{(j)}\right\}^{s_{k0}^{(-j)}} \cdot pr(\psi_k^{(j)}),$$

where $s_{kc}^{(-j)} = \#\{i' : Y_{i'} = 1, Z_{i'} = k, I_{i'} \neq j, M_{i'j} = c\} + \#\{i : Y_i = 0, Z_i = k, M_{ij} = c\}$, for $c = 0, 1$. If the prior on FPRs are $\mathsf{Beta}(a_1, b_1)$, then the above conditional distribution is $\mathsf{Beta}(a_1 + s_{k1}^{(j)}, b_1 + s_{k0}^{(j)})$.

8. Update $\boldsymbol{\pi}$ from $\mathsf{Dirichlet}\left(d_1 + t^{(j)}, ..., d_J + t^{(j)}\right)$, where $t^{(j)}$ is the number of cases assigned to class j, i.e. $t^{(j)} = \#\{i' : Y_{i'} = 1, I_{i'} = j\}$, $j = 1, .., J$.

Appendix S4. Directed Acyclic Graph for Nested Partially-Latent Class Models

This section illustrates the model structure of nested partially-latent class models using a directed acyclic graph (DAG) and provides some details on posterior inference.

Because the false positive rate (FPR) parameters $\boldsymbol{\Psi}$ are in both the control and case likelihood (2.1) and (2.2) in the Main Manuscript, their posterior depend on both the control and case models. This is referred to as "feedback" because the case model will indirectly inform $\boldsymbol{\Psi}$. If we only want the control data to inform the case model but not vice versa, we can "cut" this source of feedback through approximate conditional updating in the Gibbs sampler (Lunn *and others*, 2009). That is, we update $\psi_k^{(j)}$ by $pr(\psi_k^{(j)} \mid M_{ij}; i : Y_i = 0)$ instead of Step 7 of the Gibbs sampler (see Appendix C). It will cut the information flow from the case model to the FPR parameters $\boldsymbol{\Psi}$ and is indicated by the check-bit valves in Figure S1. It is desirable when certain parts of the joint

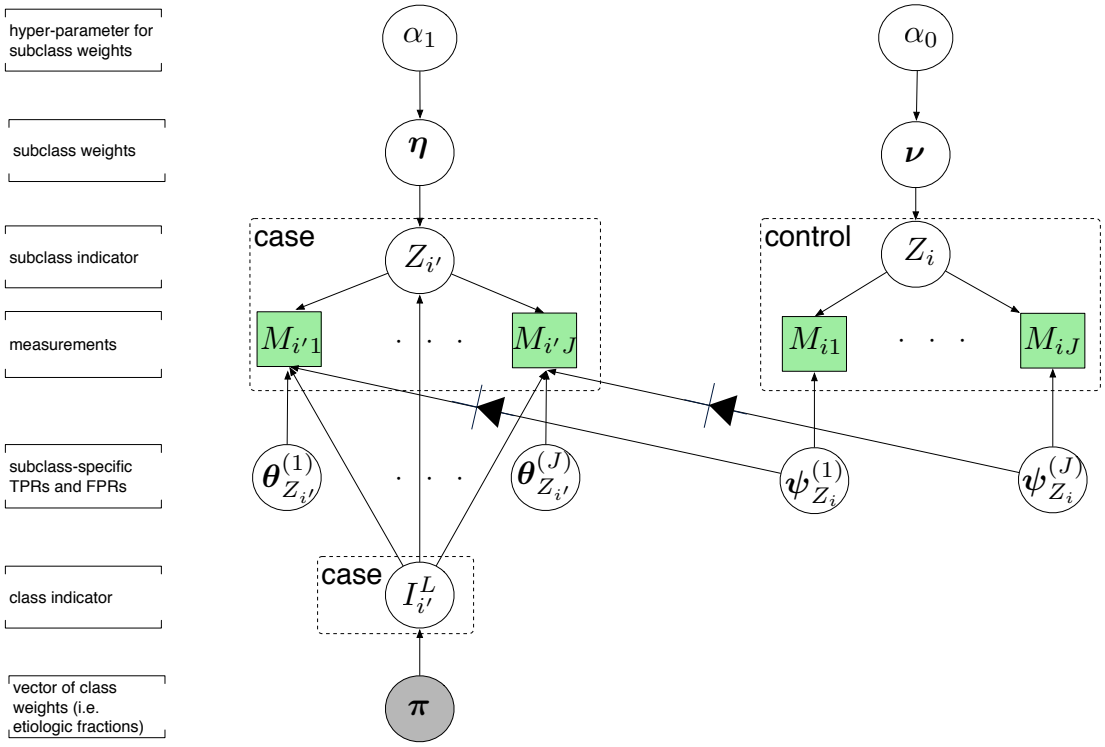

Figure S1. Directed acyclic graph (DAG) for the npLCM. Quantities in circles are unknown parameters or auxiliary variables; quantities in solid squares are observables. The etiologic fraction $\boldsymbol{\pi}$ is of primary scientific interest. The solid arrows represent probabilistic relationship between the connected variables. The "cut" valve "A ⊣▶ B" means that when updating node A in the Gibbs sampler, we drop the likelihood terms that involve node B.

model are considered not reliable to inform a subset of parameters, and can be implemented by the `cut` function in `WinBUGS 1.4`. Such "cut-the-feedback" approximate Bayesian computation has both gains in computational speed and inferential robustness, and is also suggested in other contexts (Liu *and others*, 2009; Warren *and others*, 2012; Zigler and Dominici, 2014).

Appendix S5. Simulation Studies

Appendix S5.1 Parameter Settings

We present the true parameter values and the empirical coverage rates in simulation studies (Section 4).

Scenario I		Scenario II	
$\boldsymbol{\pi}$	$= (0.5, 0.2, 0.15, 0.1, 0.05)'$	$\boldsymbol{\pi}$	$= (0.5, 0.2, 0.15, 0.1, 0.05)'$
$\boldsymbol{\Theta}^T$	$= \begin{bmatrix} 0.95 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0.95 & 0.9 & 0.9 & 0.9 & 0.9 \end{bmatrix}$	$\boldsymbol{\Theta}^T$	$= \begin{bmatrix} 0.95 & 0.95 & 0.55 & 0.95 & 0.95 \\ 0.95 & 0.55 & 0.95 & 0.55 & 0.55 \end{bmatrix}$
$\boldsymbol{\Psi}^T$	$= \begin{bmatrix} 0.25 & 0.25 & 0.2 & 0.15 & 0.15 \\ 0.2 & 0.2 & 0.25 & 0.1 & 0.1 \end{bmatrix}$	$\boldsymbol{\Psi}^T$	$= \begin{bmatrix} 0.4 & 0.4 & 0.05 & 0.2 & 0.2 \\ 0.05 & 0.05 & 0.4 & 0.05 & 0.05 \end{bmatrix}$
$\boldsymbol{\nu}$	$= (0.5, 0.5)'$	$\boldsymbol{\nu}$	$= (0.5, 0.5)'$
$\boldsymbol{\eta}$	$= (\eta_o, 1-\eta_o)'$, $0 \leqslant \eta_o \leqslant 1$	$\boldsymbol{\eta}$	$= (\eta_o, 1-\eta_o)'$, $0 \leqslant \eta_o \leqslant 1$

Appendix S5.2 Bayesian Fitting in Finite Samples

In finite samples, one can fit the larger LD model that *a priori* encourages a small number of subclasses. Extra subclasses can be used if the measurements have rich multivariate associations. Through simulations, we compare Bayes estimates of etiologic fractions obtained from the npLCM and pLCM. We generate $T = 1,000$ datasets with sample size $n_1 = n_0 = 500$ under Scenario I and II with parameters values described in the previous subsection. We fit the npLCM (truncation level $K^* = 5$ subclasses) and pLCM ($K = 1$) to each data set using informative Beta priors on the true positive rates ($\{\theta_k^{(j)}\}$) with 0.5 and 0.99 as the 2.5% and 97.5% quantiles mimicking PERCH study.

We view the Bayes estimates as functions of data and assess their frequentist properties, such as bias and variance (e.g. Efron, 2015). We define the repeated-sampling bias of the posterior mean and its mean squared error (MSE) respectively as $\lim_T T^{-1} \sum_{t=1}^T \left\{ \overline{\pi}_\ell^{(t)} - \pi_{o,\ell} \right\}$, and $\lim_T T^{-1} \sum_{t=1}^T (\overline{\pi}_\ell^{(t)} - \pi_{o,\ell})^2, \ell = A, \ldots, E$, where $\overline{\pi}_\ell^{(t)} = \mathbb{E}\{\pi_\ell \mid \mathcal{D}^{(t)}, \mathcal{M}\}$ is the posterior mean

taken with respect to the posterior distribution of $\boldsymbol{\pi}$ given the t-th simulated data set $\mathcal{D}^{(t)}$ and model $\mathcal{M}$. In finite samples, the bias multiplied by the case sample size gives the expected number of cases over- or under-attributed to a cause, which is a measure of direct interest to the motivating PERCH study.

The top panel of Table Table S1 compares the estimation biases by posterior means obtained from the two models (np and p), each using data simulated under Scenario I and II as described at the beginning of this subsection. To study the effect of LD upon the estimation of $\boldsymbol{\pi}_o$, we assess the biases at five cases' first subclass weights ($\eta_o = 0, 0.25, 0.5, 0.75, 1$). For a data set with finite sample size, estimation bias can arise from random sampling, model mis-specification or the prior, for which the first is averaged out by replication. The non-zero biases seen here reflect likelihood mis-specification and the influence of the prior. When the likelihood is correctly specified, only biases from priors remain. In Scenario II with strong LD, the npLCM performs much better. For example, the LI assumption (pLCM) results in an upward bias of 26.2% for C at $\eta_o = 0$, as well as other highlighted biases greater than 10% in magnitude. In Scenario I with weak LD, the biases from both models are negligible ($-1.9\% \sim 1.9\%$).

When the truth is close to LI, the npLCM is comparably efficient to pLCM for almost all settings. The bottom panel of Table Table S1 shows the the ratio of MSEs for pLCM versus npLCM. In Scenario I, the ratios are close to 1 indicating the npLCM has efficiently used stick-breaking to strike the balance between estimation bias and variance. In Scenario II, compared to the pLCM, the npLCM produced smaller MSEs for C at all η_o values, where the advantage is largely explained by smaller biases.

The npLCM also produces 95% credible intervals (CI) with near-nominal empirical coverage rates. For example, Appendix Table 1 highlights that the substantial under-coverages ($< 80\%$) only occurred when assuming LI. Because of the extra variability from the informative priors on the TPRs, the CIs are conservative in Scenario I for both models. The over-coverage of both

models is largely due to the assumed variances in the TPR parameters.

Table S1. Comparison of Bayes estimates of etiology fractions obtained from npLCM (np) and pLCM (p). *Top*: direct bias of the posterior mean ($\overline{\pi}_\ell - \pi_{o,\ell}$); *Bottom*: ratio of mean squared errors (MSE) for pLCM vs npLCM. All numbers are averaged across $1,000$ replications and multiplied by 100.

			Truth: Cases' First Subclass Weight (η_o)				
		Model	0	0.25	0.5	0.75	1
	Class			100×Bias(Standard Error)			
I	A	np	-0.8(0.1)	-0.5(0.1)	-0.2(0.1)	0.1(0.1)	0.4(0.1)
		p	-1.1(0.1)	-0.7(0.1)	-0.3(0.1)	-0.1(0.1)	0.0(0.1)
	B	np	-0.6(0.1)	-0.5(0.1)	-0.4(0.1)	-0.5(0.1)	-0.4(0.1)
		p	-0.6(0.1)	-0.5(0.1)	-0.6(0.1)	-0.5(0.1)	-0.3(0.1)
	C	np	1.4(0.1)	0.7(0.1)	-0.1(0.1)	-0.9(0.1)	-1.7(0.1)
		p	1.9(0.1)	0.8(0.1)	-0.1(0.1)	-0.9(0.1)	-1.9(0.1)
	D	np	-0.1(0.1)	0.1(0.1)	0.4(0.1)	0.6(0.1)	0.9(0.1)
		p	-0.2(0.1)	0.3(0.1)	0.5(0.1)	0.7(0.1)	1.1(0.1)
	E	np	0.0(0.1)	0.2(0.1)	0.3(0.1)	0.6(0.1)	0.7(0.1)
		p	0.0(0.0)	0.2(0.1)	0.5(0.1)	0.8(0.1)	1.0(0.1)
II	A	np	4.5(0.1)	5.7(0.1)	5.5(0.1)	3.5(0.1)	0.5(0.1)
		p	-3.6(0.1)	0.2(0.1)	3.0(0.1)	5.0(0.1)	5.5(0.1)
	B	np	-5.7(0.1)	-6.1(0.1)	-4.9(0.1)	-2.1(0.1)	1.9(0.1)
		p	**-13.5**(0.1)	-8.5(0.1)	-4.3(0.1)	-0.3(0.1)	4.1(0.1)
	C	np	4.5(0.1)	4.1(0.1)	2.1(0.1)	-1.0(0.1)	-6.2(0.1)
		p	**26.2**(0.1)	**13.6**(0.1)	3.7(0.1)	-4.8(0.1)	**-12.5**(0.0)
	D	np	-2.4(0.1)	-2.5(0.1)	-1.7(0.1)	-0.4(0.1)	2.1(0.1)
		p	-5.8(0.0)	-3.3(0.1)	-1.6(0.1)	-0.2(0.1)	1.3(0.1)
	E	np	-1.0(0.0)	-1.3(0.0)	-1.0(0.0)	-0.1(0.1)	1.6(0.1)
		p	-3.2(0.0)	-1.9(0.0)	-0.8(0.1)	0.4(0.1)	1.7(0.1)
	Class			100×Ratio of MSE(Standard Error)			
I	A		94(6)	115(7)	100(6)	92(6)	91(6)
	B		105(6)	94(6)	98(6)	96(6)	91(6)
	C		114(7)	101(6)	93(5)	93(5)	90(5)
	D		104(6)	105(6)	96(6)	97(6)	108(7)
	E		97(4)	96(6)	124(7)	98(6)	119(7)
II	A		82(4)	25(1)	47(2)	115(6)	221(12)
	B		516(11)	177(5)	80(3)	62(4)	140(8)
	C		2379(77)	711(26)	131(7)	268(13)	357(8)
	D		397(14)	152(6)	94(5)	79(4)	60(4)
	E		357(13)	151(6)	102(5)	95(6)	82(5)

Table S2. Comparison of the actual coverage rates of 95% credible intervals for each disease class estimated by results fitted to 1,000 replication data sets.

			Truth: Cases' First Subclass Weight (η_o)				
		Model	0	0.25	0.5	0.75	1
	Class		100×Coverage (Standard Error)				
I	A	np	98.5(0.4)	99.3(0.3)	98.5(0.4)	98.1(0.4)	97.8(0.5)
		p	97.8(0.5)	97.6(0.5)	98.3(0.4)	97.7(0.5)	96.5(0.6)
	B	np	98.8(0.3)	97.9(0.5)	97.8(0.5)	97.3(0.5)	98.4(0.4)
		p	98.5(0.4)	98.2(0.4)	97.4(0.5)	97.7(0.5)	96.8(0.6)
	C	np	96.6(0.6)	98.5(0.4)	97.7(0.5)	97.7(0.5)	94.3(0.7)
		p	93.0(0.8)	96.6(0.6)	98.6(0.4)	97.5(0.5)	95.1(0.7)
	D	np	99.0(0.3)	99.1(0.3)	98.1(0.4)	98.1(0.4)	97.6(0.5)
		p	98.3(0.4)	98.6(0.4)	98.3(0.4)	96.9(0.5)	95.8(0.6)
	E	np	98.1(0.4)	98.5(0.4)	98.2(0.4)	98.0(0.4)	97.4(0.5)
		p	98.6(0.4)	97.1(0.5)	96.6(0.6)	96.3(0.6)	95.2(0.7)
II	A	np	95.4(0.7)	88.4(1.1)	88.2(1.1)	94.0(0.8)	98.8(0.3)
		p	99.6(0.2)	100.0(0.0)	96.7(0.6)	*85.6*(1.1)	**72.3**(1.4)
	B	np	*80.4*(1.3)	*84.8*(1.1)	*86.2*(1.1)	98.3(0.4)	98.1(0.4)
		p	**9.9**(0.9)	**62.5**(1.5)	92.1(0.9)	98.9(0.3)	*82.9*(1.2)
	C	np	89.2(1.0)	89.8(1.0)	97.2(0.5)	98.0(0.4)	*84.4*(1.1)
		p	**0.0**(0.0)	**6.1**(0.8)	91.0(0.9)	**75.3**(1.4)	**0.0**(0.0)
	D	np	93.5(0.8)	90.7(0.9)	95.4(0.7)	98.0(0.4)	95.1(0.7)
		p	**53.3**(1.6)	88.7(1.0)	97.2(0.5)	98.0(0.4)	93.9(0.8)
	E	np	95.4(0.7)	94.7(0.7)	96.1(0.6)	98.5(0.4)	96.5(0.6)
		p	**56.1**(1.6)	92.1(0.9)	97.8(0.5)	98.2(0.4)	92.0(0.9)

Appendix S6. For Section 5: Analysis of PERCH Data

Full Pathogen Names and Abbreviations:

(1).HINF- *Haemophilus Influenzae*; (2). ADENO -Adenovirus; (3). HMPV-A/B - Human Metapneumovirus Type A or B; (4). PARA-1 - Parainfluenza Type 1 Virus; (5). RHINO - Rhinovirus; (6). RSV - Respiratory Syncytial Virus Type A or B.

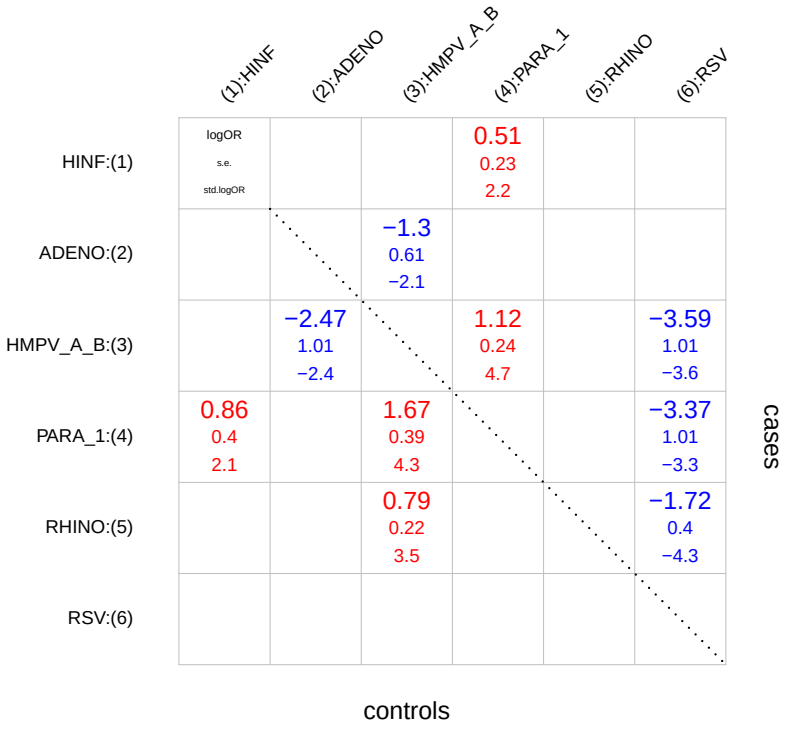

Figure S2. Matrix of significant pairwise log odds ratios (LOR) for cases (upper) and controls (lower). LOR is at the top of the cell. Below it, its standard error is in smaller type, using the same color as the LOR. Then the estimate is divided by its standard error. We put the actual value when the Z-statistics has an absolute value greater than 2; a plus (red) or minus (blue) if between 1 and 2; blank otherwise.

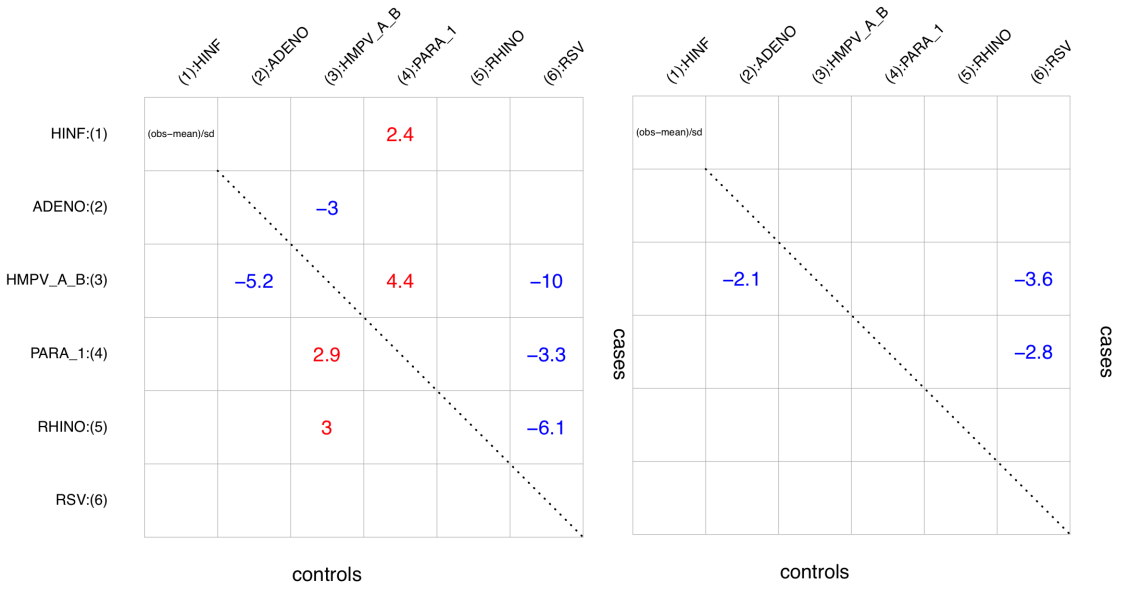

Figure S3. Posterior predictive checking for pairwise odds ratios separately for cases (upper triangle) and controls (lower triangle) with expert priors on true positive rates. *Left*: pLCM; *Right*: npLCM. Each entry is a standardized log odds ratio difference (SLORD): the observed log odds ratio for a pair of measurements minus the mean LOR for the posterior predictive distribution divided by the standard deviation of the posterior predictive distribution. The first significant digit of absolute SLORs are shown in red for positive and blue for negative values, and only those greater than 2 are shown. On average, for a well fitting model, we expect $0.05 \times \binom{6}{2} \times 2 \approx 1.5(\pm 2.4)$ non-blank cells in cases and controls, respectively.

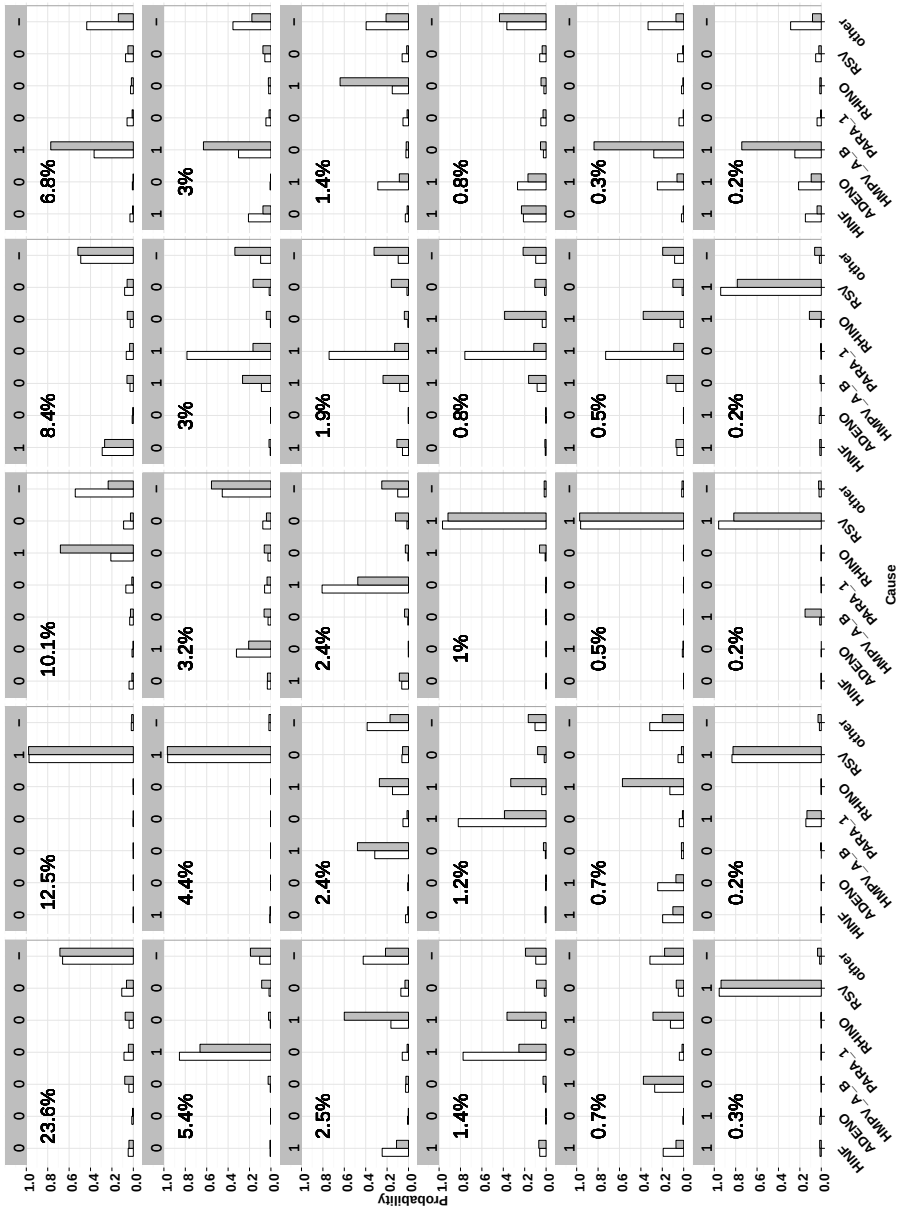

Figure S4. Each panel shows the individual etiology distribution estimated by the empirical distribution of MCMC samples of the disease class indicator. The height of a bar represents the probability of a case caused by each of the 7 causes labelled on the x-axis. For each cause, paired bars compare the estimates from the pLCM (left) and the npLCM (right); the binary codes at the top represent the NPPCR data with its observed frequency marked beneath (no measurements on "other" causes hence left as "-"). 30 most common patterns are shown here ordered by their observed frequencies.

Acknowledgments

We thank the members of the larger PERCH Study Group for discussions that helped shape the statistical approach presented herein, and the study participants. The PERCH Study Group consists of researchers from the International Vaccine Access Center, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya; ICDDR, B, Dhaka and Matlab, Bangladesh; Medical Research Council, Basse, The Gambia; Respiratory and Meningeal Pathogens Research Unit, University of the Witwatersrand, Johannesburg, South Africa; Departments of Pediatrics and Medicine, Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, MD, USA and Centre pour le Dveloppement des Vaccins (CVD-Mali), Bamako, Mali; Thailand Ministry of Public Health U.S. CDC Collaboration, Nonthaburi, Thailand; Boston University, Lusaka, Zambia; University of Otago, Christchurch, New Zealand. We also thank the members of PERCH Expert Group who provided external advice.

Research reported in this work was also partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1408-20318). The statements in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

References

Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**(4), 434–455.

Efron, Bradley. (2015). Frequentist accuracy of bayesian estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(3), 617–646.

Gelfand, Alan E and Kottas, Athanasios. (2002). A computational approach for full nonparametric bayesian inference under dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **11**(2), 289–305.

Hjort, Nils Lid, Holmes, CC, Müller, Peter and Walker, Stephen G. (2010). Bayesian nonparametrics. *AMC* **10**, 12.

Ishwaran, Hemant and James, Lancelot F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**(453), 161–173.

Ishwaran, Hemant and James, Lancelot F. (2002). Approximate Dirichlet Process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics* **11**(3), 508–532.

Jones, G., Johnson, W.O., Hanson, T.E. and Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics* **66**(3), 855–863.

Liu, Fei, Bayarri, MJ, Berger, JO *and others*. (2009). Modularization in bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis* **4**(1), 119–150.

Lunn, David, Best, Nicky, Spiegelhalter, David, Graham, Gordon and Neuenschwander, Beat. (2009). Combining mcmc with 'sequential' PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics* **36**(1), 19–38.

Neal, Radford M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**(2), 249–265.

Papaspiliopoulos, Omiros and Roberts, Gareth O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika* **95**(1), 169–186.

Sethuraman, Jayaram. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**(2), 639–650.

Walker, Stephen G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36**(1), 45–54.

Warren, Joshua, Fuentes, Montserrat, Herring, Amy and Langlois, Peter. (2012). Spatial-temporal modeling of the association between air pollution exposure and preterm birth: Identifying critical windows of exposure. *Biometrics* **68**(4), 1157–1167.

Wu, Zhenke, Deloria-Knoll, Maria, Hammitt, Laura L and Zeger, Scott L. (2016). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(1), 97–114.

Zigler, Corwin Matthew and Dominici, Francesca. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association* **109**(505), 95–107.