Supplementary Materials to "Regression Analysis of Dependent Binary Data for Estimating Disease Etiology from Case-Control Studies"

Zhenke Wu[1,2] and Irena Chen[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA; E-mail: zhenkewu@umich.edu.

[2]Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, USA

Summary

The Supplementary Materials contain the technical details, a remark, extra simulation results and figures referenced in Main Paper. Section A1 provides the technical specifications of the proposed prior distributions. Section A2 remarks on model assumptions with covariates. Section A3 details convergence checks for valid posterior inference. Section A4 presents more details of the simulations in Main Paper. Section A5 presents additional simulation results. Finally, Section A6 contains Supplemental Figures.

A1 Prior distributions

The unknown parameters include the regression coefficients in the etiology regression ($\{\mathbf{\Gamma}^{\pi}_{\ell}\}$), the parameters in the subclass weight regression for the cases ($\{\mathbf{\Gamma}^{\eta}_{k}\}$) and the controls ($\{\mathbf{\Gamma}^{\nu}_{k}\}$), the true and false positive rates ($\mathbf{\Theta} = \{\theta^{(j)}_{k}\}, \mathbf{\Psi} = \{\psi^{(j)}_{k}\}$). To mitigate potential overfitting and increase model interpretability, we *a priori* place substantial probabilities on models with the following two features: (a) Few non-trivial subclasses via a novel additive half-Cauchy prior for the intercepts $\{\mu_{k0}\}$, and (b) for a continuous variable, smooth regression curves $\pi_{\ell}(\cdot)$, $\nu_k(\cdot)$ and $\eta_k(\cdot)$ by Bayesian Penalized-splines (P-splines, Lang and Brezger, 2004) combined with shrinkage priors on the spline basis coefficients (Ni et al., 2015) to encourage towards constant values.

A1.1 Subclass Weight Regression: Encourage Few Subclasses

We propose a novel prior to encourage a small number of subclasses of non-trivial weights in finite samples, or "simplex regression shrinkage prior". We parameterize the intercepts $\{\mu_{k0}\}$ so that *a priori* the higher-order subclasses are less likely to receive non-trivial weights. We let $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu^{*}_{k0}$ where $u_{kj}, 1 \leq j \leq k \leq K-1$ is a pre-specified triangular array of positive values. Upon heavy-tailed priors on μ^{*}_{k0} with positive supports, we will *a priori* make higher-order subclasses increasingly less likely to receive substantial weights. In this paper, we use $u_{kj} = 1, j = 1, \ldots, k$; Other choices such as $u_{kj} = \mathbb{I}\{k = j\}$ or $u_{kj} = 1/k$ may be useful in other settings. We specify the prior distributions of μ^{*}_{k0} to be heavy-tailed. In this paper we use Cauchy distribution with scale $s_0 = 10$. Since our control model take a classical latent class regression model form (Bandeen-Roche et al., 1997) (the generic term "class" here corresponds to control "subclass" in an npLCM), the proposed prior for the subclass weight $\nu_k(\boldsymbol{W}) = h_k(\boldsymbol{W}; \mathbf{\Gamma}^{\nu}_{k}), k = 1, \ldots, K-1$ is also useful for a classical LCM regression analysis where the number of classes is unknown. Unlike a logistic stick-breaking specification $h_k(\boldsymbol{W}; \cdot)$ without the intercepts $\{\mu_{k0}\}$, the proposed priors on the intercepts $\{\mu_{k0}\}$ encourage few subclasses and well recovers the true subclass weights. Using the same data simulated in Simulation I, Section 3 of Main Paper, Figure S2 shows the proposed prior propagates into the posterior distribution and estimates 2 non-trivial subclasses from a working number of 7 subclasses.

At stick-breaking step k, the prior allows taking away nearly the entire stick segment currently left. Our basic idea is to have one of $\{g(\alpha_{ik})\}_{k=1}^{K-1}$ close to one *a posteriori* by making the posterior mean of one of $\{\alpha_{ik}\}_{k=1}^{K}$ large. We accomplish this by designing a novel prior on the intercept $\mu_{k0} = \sum_{j=1}^{k} u_{kj}\mu_{k0}^{*}$ where

$$\mu_{k0}^{*} \sim N^{+}(0, \tau_{0k}^{-1}), \quad \tau_{0k} \sim \mathsf{Gamma}(a_0, b_0), k = 1, \ldots, K-1.$$

The first level has a mean-zero Gaussian distribution truncated to the positive half. At the second-level, the precision (inverse variance) is Gamma distributed with shape $a_0 = \nu/2$, and rate $b_0 = \nu s_0^2/2$; it has the interpretation of ν prior independent sample(s) with a mean sample variance of s_0^2. Large values of τ_{0k}^{-1} help to stop stick-breaking at subclass k forcing weights for ensuing subclasses $\nu_{k'} \approx 0$, $k' > k$, while small values let the stick-breaking scheme continue to step $k+1$. This type of prior sparsity, which we call "selective stopping" or shrinkage over a simplex $\mathcal{S}_{K-1}$ uniformly over covariates, effectively encourages using a small number of subclasses to approximate the observed 2^J probability contingency table for the control measurements in finite samples.

We accomplish selective stopping by the heavy right tail of μ_{k0}^{*}'s marginal prior. It has a truncated scaled-t distribution with degree of freedom ν and scale s_0, and consequently peaks at zero and admits large positive values. Given other parameters in $\alpha_{ik}^{\nu} = \alpha_{k}^{\nu}(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\nu})$, a near-zero intercept takes the stick-breaking procedure to the next step, while a large positive intercept effectively halts it. The tendency to stop at step k is *a priori* modulated by the scale parameter s_0. Because, given the degree-of-freedom ν, the prior probability $P(g(\alpha_{1k}) > C \mid \nu, s_0)$, $\forall C \in (0.5, 1)$ approaches 1 as the scale parameter s_0 increases.

In our simulations and applications, we choose hyperparameters $\nu = 1$ and $s_0 = 10$ for the intercept, and $k_\beta = 4$ for the first B-spline coefficients $\boldsymbol{\beta}_{kj}^{(1),\nu}$ in the prior (Equation S1, Section A1.2). We have chosen our hyperparameters based on the interpretations on the probability (inverse-link) scale; see similar prior elicitations for regression coefficients in other applications (e.g., Bedrick et al., 1996; Witte et al., 1998) and for automatic, stabilized and weakly-informative fitting of generalized linear models (Gelman et al., 2008). We choose the hyperparameters for the intercepts that put most prior mass of $g(\mu_{10})$ within $(0.5, 1-10^{-9})$, because $1 - 10^{-9}$ is sufficiently close to 1 which means the stick-breaking is stopped at Step $k = 1$. In contrast, we choose the first B-spline coefficient's hyperparameter $k_\beta = 4$ that puts most prior mass of $g(\beta_{kj}^{(1),\nu})$ within $(0.02, 0.98)$, a range for the weight of a non-trivial subclass to break from the rest of the stick at Step k. Figure S5 shows a sharp separation between the priors for $g(\mu_{k0}^*)$ and $g(\beta_{kj}^{(1),\nu})$. The shapes of the priors again highlight the different roles played by the intercept and the B-spline coefficients: the former decides whether to continue the stick-breaking procedure to induce complex conditional dependence given covariates, and if so, the latter computes the fraction to break from the remaining length of the stick. The intercepts in the controls $\{\mu_{k0}\}$ are shared with the case subclass weight regression $\eta_k(\boldsymbol{W}) = h_k(\boldsymbol{W}; \boldsymbol{\Gamma}_k^\eta)$; We set the same prior distributions for other elements of $\boldsymbol{\Gamma}_k^\eta$, $k = 1, \ldots, K-1$.

A1.2 Encourage Smooth f_{kj}^π and f_{kj}

We use penalized B-splines to model the additive functions of a continuous variable in etiology regression (f_{kj}^π), subclass weight regression for the cases and the controls (f_{kj}) (Lang and Brezger, 2004). We expand $f_{kj}^\bullet(\cdot) = \sum_{c=1}^{C} \beta_{kj}^{(c)} B_j^{(c)}(\cdot)$, with $\{B_j^{(c)}(\cdot) : c = 1, \ldots, C\}$ being the shared C cubic B-spline bases. We let $f_{\ell j}^\pi$, f_{kj} in the case subclass weight regression and f_{kj} in the control subclass weight regression have distinct coefficients: $\{\beta_{kj}^{(c),\pi}, k = 1, \ldots, L\}$, $\{\beta_{kj}^{(c),\eta}, k = 1, \ldots, K-1\}$ and $\{\beta_{kj}^{(c),\nu}, k = 1, \ldots, K-1\}$, respectively. With M interior equally-spaced knots $\boldsymbol{\kappa} = (\kappa_0, \ldots, \kappa_{M+1})^\top$: $\min_i(x_{ij}) = \kappa_0 < \kappa_1 < \cdots < \kappa_M < \kappa_{M+1} = \max_i(x_{ij})$, there are $C = M + 4$ basis functions. It readily extends to let f_{kj}^π and f_{kj} have different numbers of basis functions.

Since the specification below applies to $f_{kj}^\pi(x; \boldsymbol{\beta}_{kj}^\pi)$, $f_{kj}(w; \boldsymbol{\beta}_{kj}^\nu)$ and $f_{kj}(w; \boldsymbol{\beta}_{kj}^\eta)$ for any centered and standardized continuous variable, for simplicity we omit the superscripts π, ν, η and subscript j .

The Penalized-splines in our formulation bypass the choice of the number and placement of knots $\boldsymbol{\kappa}$ by using a large number of knots deemed sufficient to capture the curves and imposing smoothing penalty on the coefficients for basis functions to prevent overfitting. The Gaussian random walk priors on basis coefficients are good choices for fitting Bayesian

P-splines (Lang and Brezger, 2004):

$$\boldsymbol{\beta}_k \mid \tau_k, \lambda_k \sim N(\mathbf{0}_{C\times 1}, (\tau_k \boldsymbol{K})^{-1}), \tag{S1}$$

where the symmetric penalty matrix $\boldsymbol{K} = \Delta_1^\top \Delta_1$ is constructed from the first-order difference matrix Δ_1 of dimension $(C-1)\times C$ that maps adjacent B-spline coefficients to $\beta_k^{(c)} - \beta_k^{(c-1)}$, $c = 2, \ldots, C$ (`diff(diag(C),differences = 1)` in R language), and τ_k is the smoothing parameters with large values leading to smoother fit of $f_k(x)$ (constant when $\tau_k = \infty$) and interpolation when near zero. This first-order random walk prior above uses a precision matrix $\boldsymbol{K}$ of rank $C-1$ to model the adjacent differences. This leaves the prior of β_{k1} unspecified, for which we further assign an independent prior $\beta_{k1} \sim N(0, k_\beta^{-1})$. We discuss the hyperparameter k_β in the next subsection.

We use a mixture prior with two well-separated component distributions with one favoring small and the other large smoothing parameters τ_{kj}:

$$\tau_{kj} \sim \xi_{kj}\mathsf{Gamma}(\cdot \mid a_\tau, b_\tau) + (1-\xi_{kj})\mathsf{InvPareto}(\cdot \mid a'_\tau, b'_\tau), \tag{S2}$$

$$\mathsf{InvPareto}(\tau; a, b) = \frac{a}{b}\left(\frac{\tau}{b}\right)^{a-1}, a > 0, 0 < \tau < b, \tag{S3}$$

where the Gamma-distributed component ($a_\tau = 3$, $b_\tau = 2$) concentrates near smaller values while the inverse-Pareto component prefers larger values ($a'_\tau = 1.5$, $b'_\tau = 400$). This bimodal mixture distribution creates a sharp separation between flexible and smooth fits (Morrissey et al., 2011; Ni et al., 2015). Because we use the first-order random walk prior, the most smooth fit is of degree 0, i.e., constant functions. The random smoothness indicator ξ_{kj} represents a flexible (1) or constant (0) shape of $f_k(\cdot)$. We let $\xi_{kj} \sim \mathsf{Bernoulli}(\rho)$ with success probability ρ and then put a hyperprior $\rho \sim \mathsf{Beta}(a_\rho, b_\rho)$ to let data inform the degree of smoothness.

In this paper we use $a_\rho = 0.5$, $b_\rho = 1$ for each set of the B-spline basis coefficients for the cases ($\{\boldsymbol{\beta}_{kj}^{(c),\eta}, c = 1, \ldots, C\}$) and the controls ($\{\boldsymbol{\beta}_{kj}^{(c),\nu}, c = 1, \ldots, C\}$) to *a priori* give slight preference for constant curves, $k = 1, \ldots, K-1$, $j = 1, \ldots, q_1$; We use $a_\rho^\pi = 1$, $b_\rho^\pi = 0.5$ for the set of basis coefficients ($\{\boldsymbol{\beta}_{\ell j}^{(c),\pi}, c = 1, \ldots, C\}$) to *a priori* give slight preference for flexible etiology regression functions, $\ell = 1, \ldots, L$, $j = 1, \ldots, p_1$. In the presence of highdimensional covariates, the Beta prior with other hyperparameters can also allow a prior spread that lets the fraction of constant functions $\rho = \rho_p$ to approach 0 as $p \to \infty$.

A1.3 Informative Prior Distributions for TPRs and FPRs

The npLCM regression model is partially-identified (Jones et al., 2010). We assume independent informative priors for the TPRs in the BrS data likelihood: $\theta_k^{(j)} \sim \mathsf{Beta}(a_j^{\mathsf{BrS}}, b_j^{\mathsf{BrS}})$, $j = 1, \ldots, J$, where $(a_j^{\mathsf{BrS}}, b_j^{\mathsf{BrS}})$ are chosen so that the 2.5% and 97.5% quantiles match a prior range elicited from laboratory scientists (Deloria Knoll et al., 2017). In the presence of SS data for a subset of pathogens (e.g., culturing bacteria from blood), we similarly set the hyperparameters for the Beta distribution of the TPRs of the SS measures where ranges can be computed from existing vaccine probe trials (e.g., Feikin et al., 2014). Since the control data provide direct estimates of the FPRs, we specify independent priors for $\psi_k^{(j)} \sim \mathsf{Beta}(1, 1), j = 1, \ldots, J, k = 1, \ldots, K$.

A2 Remark on the Control Model with Covariates

The proposed model for the control data with covariates $\boldsymbol{W}$ is a generative model where we first draw a subclass indicator $Z \mid \boldsymbol{W} \sim \mathsf{Categorical}_K\{\boldsymbol{\nu}(\boldsymbol{W})\}$, and generate measurements $M_j \mid Z = k$ according to a Bernoulli distribution with positive rate $\psi_k^{(j)}$, independently for $j = 1, ..., J$. By assuming mutually independent measurements $M_1, \ldots, M_J$ given subclass Z and $Y = 0$, we let the covariates influence the dependence structure of the measurement only through the unobserved Z. As a result, upon integrating over Z, the proposed model does *not* assume marginal independence $\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, Y = 0) = \prod_{j=1}^{J} \mathbb{P}(M_j \mid \boldsymbol{W}, Y = 0)$ in contrast to a kernel-based extension of the pLCM that makes this assumption (Saha et al., 2018, Supplementary appendix). Our approach to incorporating covariates to model control data follows Bandeen-Roche et al. (1997); For other approaches, see examples in the study of particulate matter (Gryparis et al., 2007), HIV population size estimation (Bartolucci and Forcina, 2006), and alcoholic and drug addiction (Chung et al., 2006).

A3 Convergence Checks

In simulations and data analysis, we ran three MCMC chains each with a burn-in period of $10,000$ iterations followed by $10,000$ iterations stored for posterior inference. We look for potential non-convergence in terms of Gelman-Rubin statistic (Brooks and Gelman, 1998) that compares between-chain and within-chain variances for each model parameter where a large difference ($R_c > 1.1$) indicates non-convergence; We also used Geweke's diagnostic (Geweke and Zhou, 1996) that compare the observed mean for each unknown variable using the first 10% and the last 50% of the stored samples where a large Z-score indicates non-convergence ($|Z| > 2$). In our simulations and data analyses, we observed fast convergence (many satisfied convergence criteria within $2,000$ iterations) that led to well recovered regression curves, TPRs and FPRs.

A4 Additional Information about Simulations of Main Paper

Simulation I. we let $\pi_\ell(\cdot)$, $\nu_k(\cdot)$ and $\eta_k(\cdot)$ depend on the two covariates $\boldsymbol{X} = \boldsymbol{W} = (S, T)$, S and enrollment date (T), so that regression adjustments are necessary (see Remark 1, Main Paper). We simulate BrS measurements on $J = 9$ pathogens and assume the number of potential single-pathogen causes $L = J = 9$. To specify etiology regression functions that satisfy the constraint $\sum_{\ell=1}^{L} \pi_\ell(\boldsymbol{x}) = 1$, we use stick-breaking parameterization with $L = 9$ segments. In particular, we let $\mathsf{logit}\,\{g_1(s,t)\} = \beta_1\,\mathbb{I}(s = 1) + \sin(8\pi(t - 0.5)/7)$, $\mathsf{logit}\,\{g_2(s,t)\} = \beta_2\,\mathbb{I}(s = 1) + 4\exp(3t)/(1 + \exp(3t)) - 0.5$, $\mathsf{logit}(g_\ell) = \beta_8\,\mathbb{I}(s = 1)$ for $\ell > 2$; Let the PEF functions $\pi_\ell(s,t) = g_\ell(s,t)\prod_{j<\ell}\{1 - g_j(s,t)\}, \ell = 1, \ldots, L(= 9)$, where $\beta_\ell = 0.1, \ell = 1, \ldots, 8$. The true control distribution depend on covariates with $K = 2$ subclass weight functions: $\nu_1(s,t) = \mathsf{logit}^{-1}\{\gamma_1^{\nu}\,\mathbb{I}(s = 1) + 4\exp(3t)/(1 + \exp(3t)) - 0.5\}$ and $\nu_2(s,t) = 1 - \nu_1(s,t)$. We specify $\eta_k(s,t) = \nu_k(s, -t), k = 1, 2$, highlighting the need for using different subclass weights among cases and controls in an npLCM analysis. We set the true TPRs $\theta_k^{(j)} = 0.95$ and the FPRs $\psi_1^{(j)} = 0.5$ and $\psi_2^{(j)} = 0.05$.

In the regression analyses, we set $\phi_\ell(\boldsymbol{X})$ to be an additive model of a $\mathbb{I}\{S = 2\}$ indicator and a B-spline expansion with 7 degrees of freedom (d.f.) for standardized enrollment date t. We use $K^* = 7$ and specify the regression formula for subclass weights $\nu_k(\cdot)$ and $\eta_k(\cdot)$ by additive models of the $\mathbb{I}\{S = 2\}$ indicator and a B-spline expansion with 5 d.f. for standardized enrollment date.

Simulation II. We consider $L = J = 3, 6, 9$ causes, under single-pathogen-cause assumption, BrS measurements made on N_d cases and N_u controls for each level of X where $N_d = N_u = 250$ or 500. The functions $\phi_\ell(X) = \beta_{0\ell} + \beta_{1\ell}\,\mathbb{I}\{X = 2\}$ take two sets of values to reflect how variable the PEFs are across the two X levels: i) $\boldsymbol{\beta}_0^{\mathsf{i}} = (0, 0, 0, 0, 0, 0)$ and $\boldsymbol{\beta}_1^{\mathsf{i}} = (-1.5, 0, -1.5, -1.5, 0, -1.5)$ where causes have uniform PEFs when $X = 1$ and causes B and E dominate when $X = 2$, or ii) $\boldsymbol{\beta}_0^{\mathsf{ii}} = (1, 0, 1, 1, 0, 1)$ and $\boldsymbol{\beta}_1^{\mathsf{ii}} = (-1.5, 1, -1.5, -1.5, 1, -1.5)$ to mimic the scenario where pathogens B and E have lower PEFs when $X = 1$ and occupy more fractions when $X = 2$. We further let the measurement error parameters take distinct values of the TPRs $\theta_k^{(j)} = 0.95$ or 0.8 and the FPRs $(\psi_1^{(j)}, \psi_2^{(j)}) \in \{(0.5, 0.05), (0.5, 0.15)\}$, for $j = 1, \ldots, J$. Finally, we set the truth $\nu_k(W) = \eta_k(W) = \mathsf{logit}^{-1}(\gamma_{k0} + \gamma_{k1}\,\mathbb{I}\{W = 2\})$ where $(\gamma_{10}, \gamma_{11}) = (-0.5, 1.5)$ and $(\gamma_{20}, \gamma_{21}) = (1, -1.5)$.

Simulation II*: a randomly chosen replication.* Here we illustrate the inferences about the stratum-specific and overall PEFs that are available to an analyst by considering a two-level covariate $X = W$ with $J = 6$ measurements. Under the single-pathogen cause assumption, we can estimate $12 = (2\times 6)$ PEFs, six per level of X as well as six overall PEFs. For example, based on a single data set simulated under the scenario $\{L = 6,\ N_d = 500,\ K = 2,\ \theta_k^{(j)} = 0.8,\ (\psi_1^{(j)}, \psi_2^{(j)}) = (0.5, 0.05),\ (\boldsymbol{\beta}_0^{\mathsf{ii}}, \boldsymbol{\beta}_1^{\mathsf{ii}})\}$, Supplemental Figure S3 shows the posterior distribution of the stratum-specific etiology fractions $\pi_\ell(X = s)$ for $(s = 1, 2)$ by row and $L(= J)$ causes $(\ell = 1, \ldots, 6)$ by column with the true values indicated by the blue vertical dashed lines; The bottom row shows the posterior distribution of $\pi_\ell^* = \sum_s w_s \pi_\ell(X = s)$ for L causes with empirical weights $w_s = N_d^{-1} \sum_{i:Y_i=1} \mathbb{I}\{X_i = s\}$, $s = 1, 2$. The true stratum-specific and overall PEFs are covered by their respective 95% CrIs.

A5 Additional Simulation Results

A5.1 Estimating $\pi_\ell(X)$

We use simulation studies to show the frequentist performance of the npLCM regression model in recovering stratum-specific PEFs; The results below are based on a single discrete covariate that influence the PEFs but not the subclass weights in the cases or controls.

In this simulation study, we simulate 500 cases and 500 controls for each of 7 sites. Every subject is measured on 6 pathogens A to F; The causes of disease are single-pathogen causes A-F. First, we let the PEFs vary by site which are shown in Table S1. Second, we simulate the data using $K = 1$ subclass.

Table S1: True PEFs for seven sites (boldfaced numbers indicate the highest PEFs within each stratum).

site\cause	A	B	C	D	E	F
1	**0.5**	0.2	0.15	0.05	0.05	0.05
2	0.2	**0.5**	0.15	0.05	0.05	0.05
3	0.2	0.15	**0.5**	0.05	0.05	0.05
4	0.2	0.15	0.05	**0.5**	0.05	0.05
5	0.2	0.15	0.05	0.05	**0.5**	0.05
6	0.2	0.15	0.05	0.05	0.05	**0.5**
7	0.05	0.2	0.15	**0.5**	0.05	0.05

We simulate data under two TPR scenarios (I) strong signal with $\theta_1^{(j)} = 0.99$ and $\psi_1^{(1)} = 0.01$ where data are expected to provide strong information about the PEFs, and (II) weak signal with $\theta_1^{(j)} = 0.55$ and $\psi_1^{(1)} = 0.45$ where it is easy to confuse true and false positive results and the data do not provide strong information about the PEFs. In both scenario (I) and (II) , we used a Beta(6,2) distribution as a prior for the TPRs of the BrS measurements. We set the true TPRs and FPRs to be the same across sites and pathogens. In fitting the regression models, we use the etiology regression formulation by specifying $L - 1$ sets of regression parameters with site dummy variables as the predictors in $\phi_\ell(\cdot)$. Since our goal is to infer $S = 7$ sets of PEFs, we can also specify $S = 7$ sets of symmetric Dirichlet priors with hyperparameter α ($\mathsf{Dir}(\alpha)$); We use $\alpha = 1$ here. The package `baker` (https://github.com/zhenkewu/baker) provides an option to use Dirichlet priors when the PEFs depend on discrete covariates only.

A5.1.1 Scenario I: Strong Signal

Over $R = 100$ replications, the top half of Table S3 summarizes the coverage rates of the 95% credible intervals (CrIs) for the PEFs across all the sites. We observed excellent recovery of the true values across all causes and sites with the 95% CrIs covered the true values between 90% to 100% of the time. Panel I of Table S3 also shows for site 1 the posterior mean PEFs, posterior standard deviations (sd's) of the PEFs, and posterior mean squared errors (PMSEs, estimated by $B^{-1}\sum_{b=1}^{B}\sum_{i:Y_i=1}\{\pi_\ell(X_i = s; \boldsymbol{\gamma}^{\pi,(b)}) - \pi_\ell^0(X_i = s)\}$ with B retained posterior samples $\{\boldsymbol{\gamma}^{\pi,(b)}\}$) averaged over R replications. The posterior means provide excellent estimation of the PEFs with small average PMSEs.

A5.1.2 Scenario II: Weak Signal

Using data simulated under less discrepant TPRs and FPRs than those in Scenario I, the 95% CrIs cover the truths well for most site-cause pairs, but undercover the truths for causes with the highest PEF in each site (see Table S2). This is expected because when the signal from the data is weak, the model relies more heavily on the uniform prior distribution for the PEFs (symmetric Dirichlet prior with hyper-parameter 1).

Table S2: Number of times (out of 100 replications) that the true value is covered by the 95% CrIs (Scenario II, Beta(6,2) prior for the TPRs). Boldfaced numbers indicate the highest PEFs (0.5) within each stratum.

site\cause	A	B	C	D	E	F
1	**73**	100	100	99	100	100
2	100	**79**	100	100	100	99
3	100	100	**83**	98	100	100
4	100	100	100	**73**	100	99
5	99	100	100	100	**85**	100
6	100	100	100	99	100	**88**
7	100	100	100	**81**	100	99

More Informative TPR Priors (II).* We further investigate the model performance when we change the TPR prior distributions from the Beta(6,2) to a Beta distribution that has

95% of its mass between 0.525 and 0.575 and is around the true TPRs (Beta(835.95, 683.79); `beta_parms_from_quantiles(c(0.525,0.575))` using `baker`). Panel II^* of Table S3 shows dramatic improvements in the coverage rates. These results suggest that changing the prior distributions of the TPRs so that it is more tightly concentrated around plausible values can improve inferences of the stratum-specific PEFs in the presence of high levels of noises. Relative to Scenario I, the average PMSEs are larger across sites and pathogens reflecting the weaker signal in this setting.

In summary, in the simulation study where the PEFs are influenced by a discrete covariate, the regression model recovers the true values well under high signals (high sensitivities and low FPRs). Under lower sensitivities and higher FPRs, the noisier simulated data are less informative about the PEFs which are then more influenced by the prior distributions of the TPRs and PEFs. In practice, we recommend eliciting quality informative TPR priors from domain scientists as in the PERCH study and perform sensitivity analyses to understand the robustness of the results with respect to the prior distributions.

A5.2 Valid inference of π^*_ℓ omitting covariates

Under assumption (A1) in Remark 1 of Main Paper, the case subclass weights $\boldsymbol{\eta}_k(\boldsymbol{W}) = \eta_k$, $k = 1, \ldots, K$, we conduct a simulation study to show that an npLCM analysis omitting covariates is able to provide valid inference about the overall PEFs ($\boldsymbol{\pi}^*_\ell$). The simulation settings are exactly the same as in Simulation II, Section 4 of Main Paper, except that we set $\gamma_{20} = \gamma_{21} = 0$ to satisfy assumption (A1). Figure 5(a) shows the percent relative biases are similarly negligible in all the 16 scenarios with 6 disease classes; Figure 5(b) shows excellent empirical coverage rates of the 95% CrIs for $\{\pi^*_\ell\}$.

Table S3: Scenario I and II*: coverage rates of the 95% CrIs; For Site 1, the posterior means, standard deviations (s.d.'s) and PMSE of the stratum-specific PEFs averaged over $R = 100$ replications are also shown. Boldfaced numbers indicate the highest PEFs (0.5) within each stratum.

		site \cause	A	B	C	D	E	F
I	coverage	1	**99**	93	97	94	96	90
		2	97	**90**	96	97	95	94
		3	100	95	**98**	98	95	96
		4	93	94	96	**95**	92	99
		5	96	94	96	97	**95**	98
		6	96	97	98	99	95	**96**
		7	96	97	91	**100**	95	96
	posterior summary	truth (Site 1)	0.5	0.2	0.15	0.05	0.05	0.05
		average of post. mean	0.495	0.197	0.152	0.053	0.053	0.051
		average of post. s.d.	0.023	0.018	0.016	0.01	0.01	0.01
		average PMSE	0.0010	0.0007	0.0005	0.0002	0.0002	0.0002
II*	coverage	1	**98**	89	98	99	100	100
		2	97	**95**	96	100	100	99
		3	93	98	**91**	99	99	100
		4	95	98	100	**95**	99	100
		5	94	94	99	99	**91**	100
		6	95	97	100	99	99	**90**
		7	100	95	94	**96**	100	99
	posterior summary	truth (Site 1)	0.5	0.2	0.15	0.05	0.05	0.05
		average post. mean	0.417	0.163	0.138	0.091	0.086	0.106
		average post. s.d.	0.27	0.174	0.162	0.135	0.13	0.141
		average PMSE	0.131	0.067	0.056	0.034	0.031	0.042

A6 Supplemental Figures

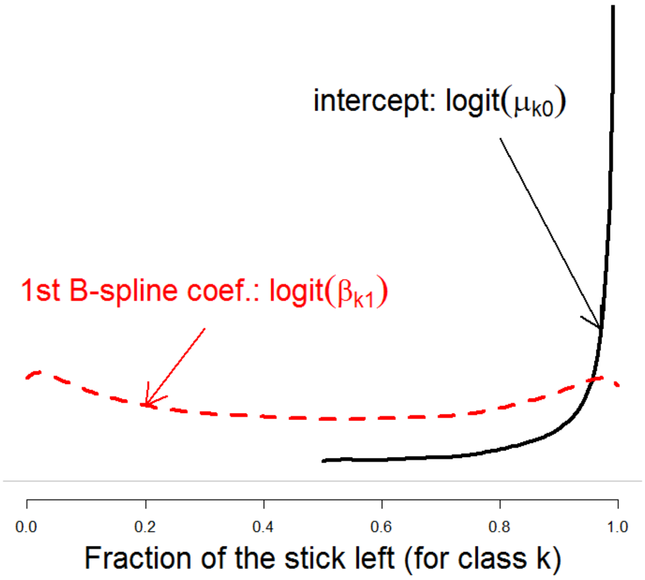
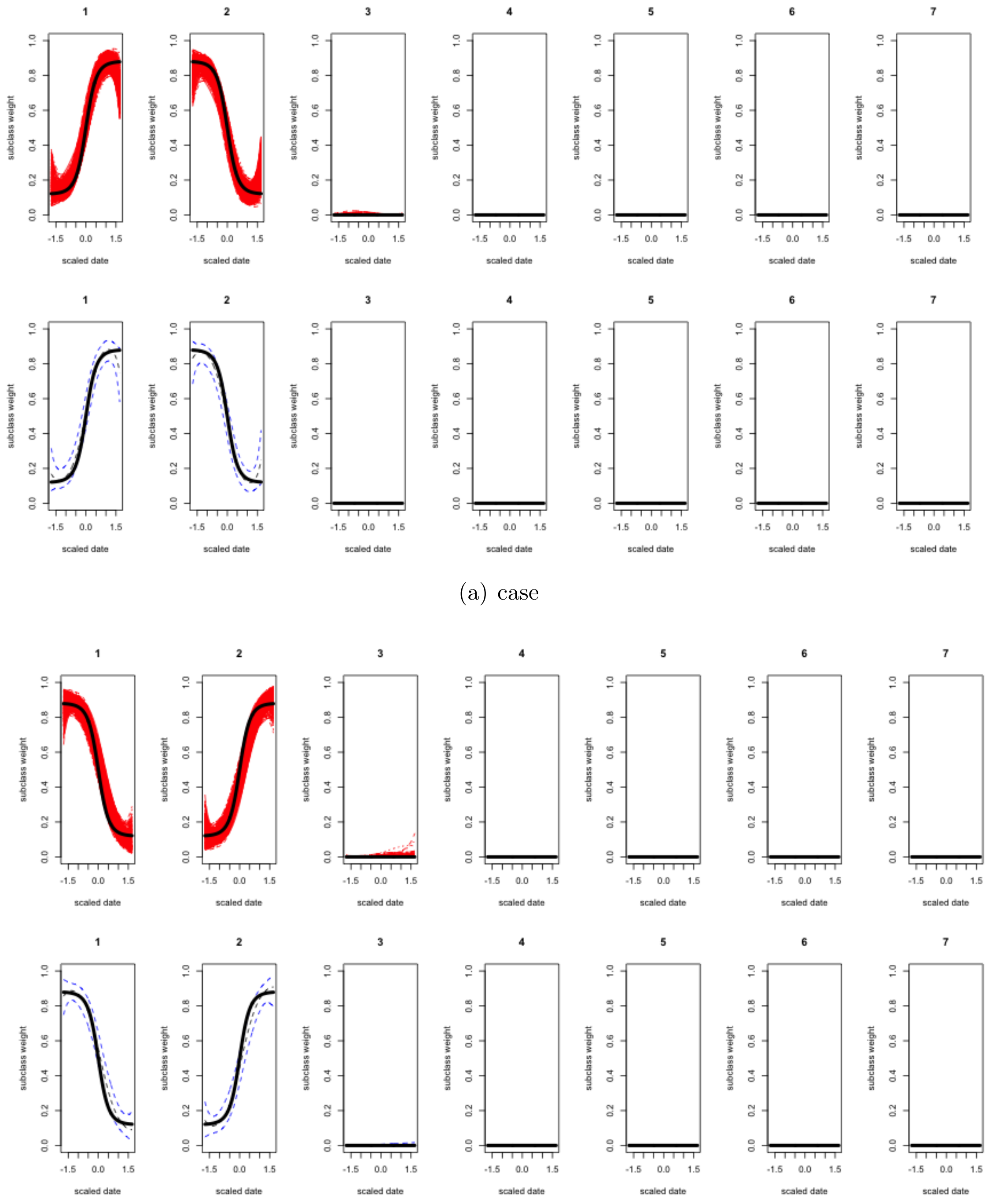


Figure S1: Prior densities for $\mathsf{logit}(\alpha_{ik}^{\nu})$, the fraction to be broken for subclass k from the stick currently left, when α_{ik} equals: 1) the intercept μ_{k0}^{*} (*black, solid line*) or 2) the first B-spline coefficient $\beta_{kj}^{(1),\nu}$ (*red, broken line*). The former concentrates near 1 because μ_{k0}^{*} has a scaled-t distributed prior that puts substantial mass at the right tail; much less so for the latter.



(a) case

(b) control

Figure S2: By propagating the prior that encourages few subclasses, the algorithm correctly infers two subclasses from the simulated data in Simulation I, Section 4 of Main Paper. Estimated case (top) and control (bottom) subclass weight curves for seven subclasses over one continuous covariate $\widehat{\nu}_k(t)$ (central blue dashed lines enclosed by the 95% credible regions; the red curves are posterior samples) compared against the simulation truths ($\nu_k^0(t)$, black solid lines). The number of subclasses is bounded by seven during model fitting.

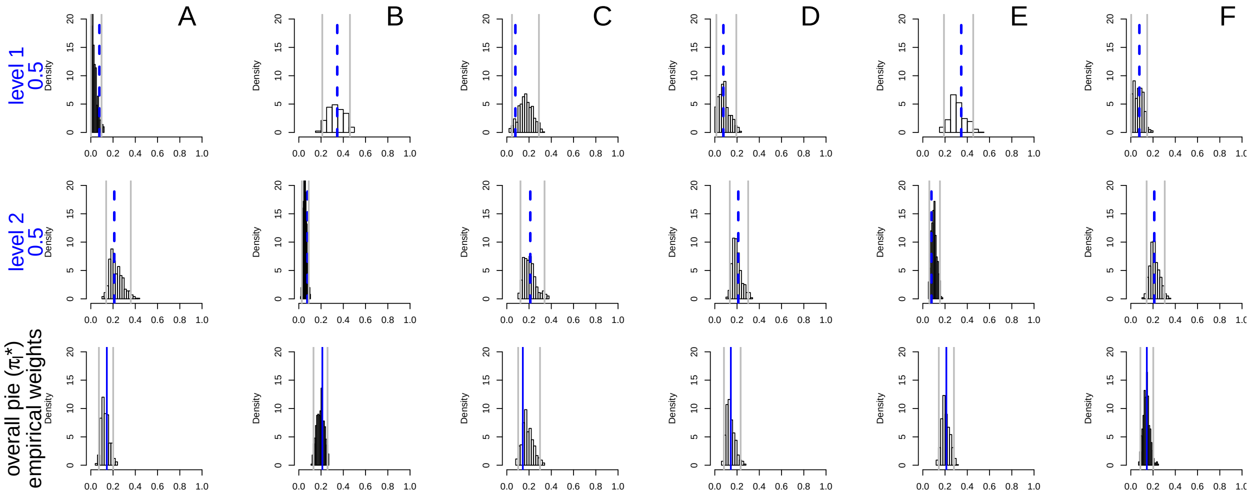
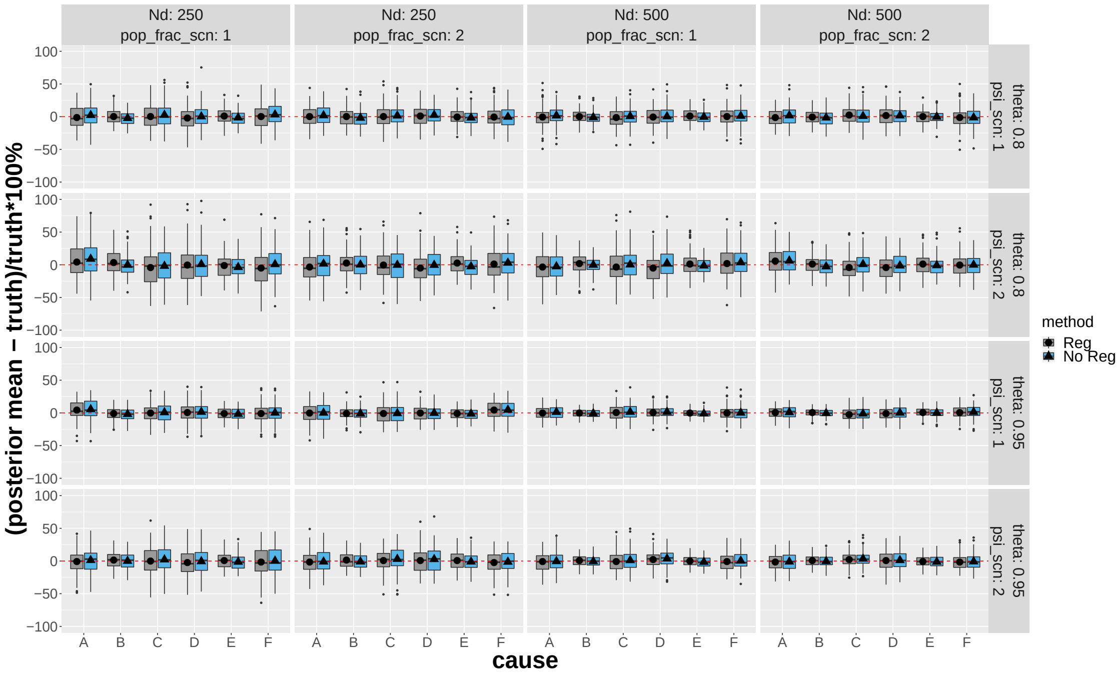
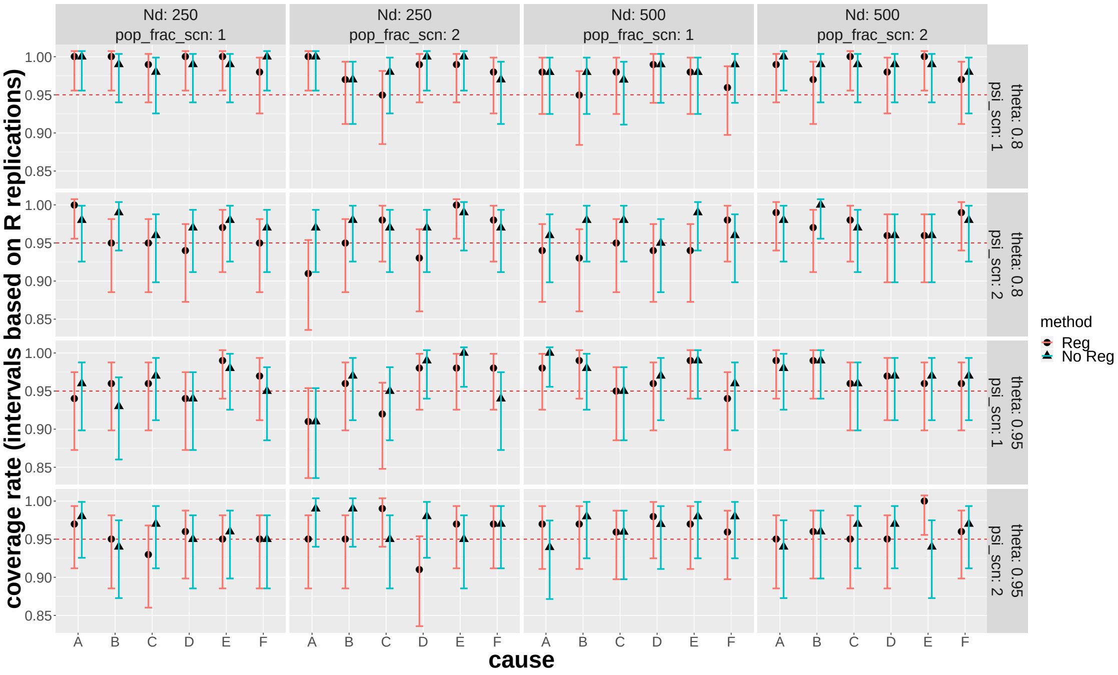


Figure S3: Posterior distributions of the stratum-specific (Row 1 and 2) and the overall (Bottom Row) PEFs based on a simulation with a two-level discrete covariate and $L = J = 6$ causes. The vertical gray lines indicate the 2.5% and 97.5% posterior quantiles, respectively; The truths are indicated by vertical blue dashed lines. *Row 1-2*) PEFs by stratum (level = 1,2) and cause (A-F); *Bottom*) π^*_ℓ: overall population etiologic fraction for cause A-F (empirical average of the two PEFs above).



(a)



(b)

Figure S4: NPLCM analyses with or without regression perform similarly in terms of percent relative bias (top) and empirical coverage rates (bottom) over $R = 100$ replications in simulations where the case and control subclass weights *do not* vary by covariates. Each panel corresponds to one of 16 combinations of true parameter values and sample sizes. See Figure 3 in Main Paper for detailed descriptions of the figure.

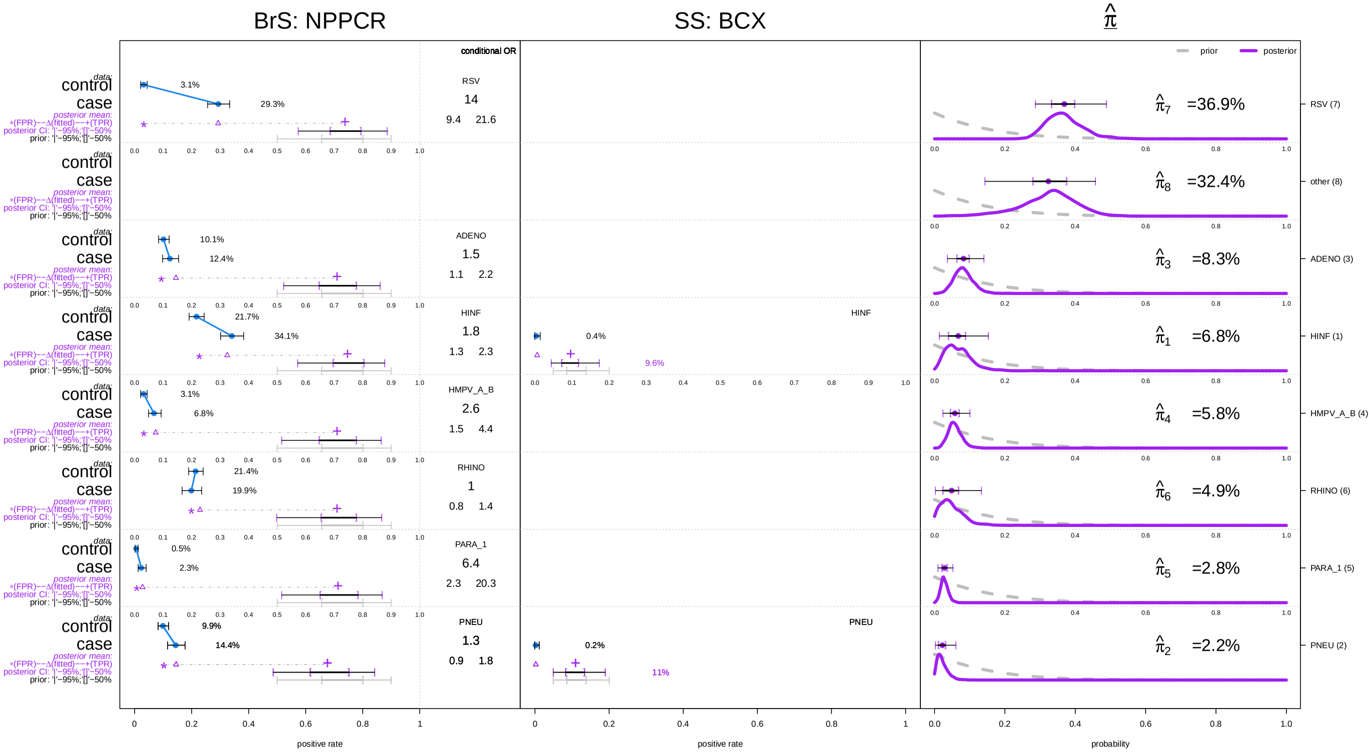
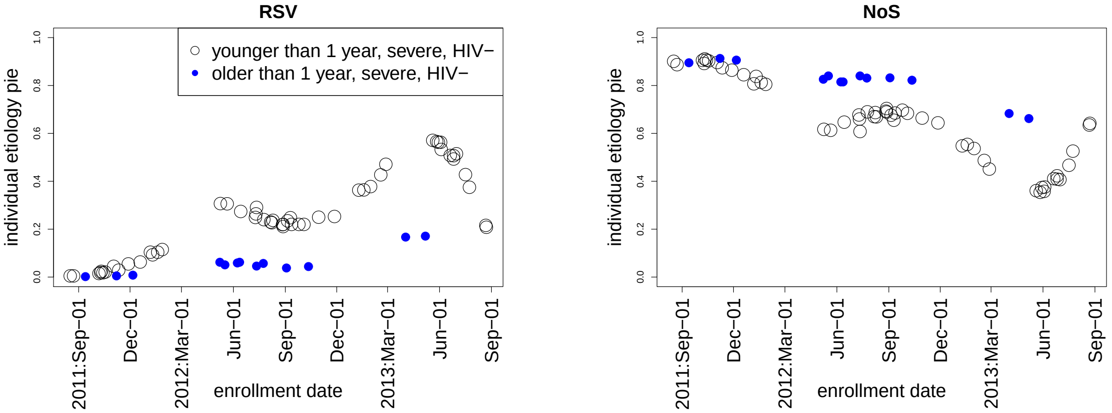


Figure S5: Panel plot with BrS, SS and Etiology Pies obtained from an npLCM analysis omitting covariates ($K = 5$). For each of the 7 pathogens, a summary of the BrS and SS data analyzed in Section 5 of Main Paper is shown in the left two columns, along with some of the intermediate model results; and the prior and posterior distributions for the PEFs on the right (rows ordered by posterior means). *Left*) The observed BrS rates (with 95% confidence intervals, CI) for cases and controls are shown on the far left with solid dots. The conditional odds ratio (COR) contrasting the case and control rates given the other pathogens is listed with 95% CI in the box to the right of the BrS data summary. Below the case and control observed rates is a horizontal line with a triangle. From left to right, the line starts at the estimated false positive rate (FPR, $\widehat{\psi}_j^{\mathsf{BrS}}$) and ends at the estimated true positive rate (TPR, $\widehat{\theta}_j^{\mathsf{BrS}}$), both obtained from the model. Below the TPR are 95% and 50% intervals summarizing its posterior (top) and prior (bottom) distributions for that pathogen. These intervals show how the prior assumption influences the TPR estimate as expected given the identifiability constraints. The triangle on the line is the model estimate of the case rate to compare to the observed value above it. *Middle*) The SS data are shown in a similar fashion to the right of the BrS data. By definition, the FPR is 0.0 for SS measures and there is no control data. The observed rate for the cases is shown with its 95% CI. The estimated SS TPR ($\widehat{\theta}_j^{\mathsf{SS}}$) with prior and posterior distributions is shown as for the BrS data. *Right*) The marginal posterior and prior distributions of the etiologic fraction for each pathogen. We appropriately normalized each density to match the height of the prior and posterior curves. The posterior mean with 50% and 95% CrIs are shown above the density.



(a) Cause: RSV

(b) Cause: NoS

Figure S6: Individual etiology fraction estimates for RSV (left) and NoS (right) differ by age and season among HIV negative and severe pneumonia cases for whom the seven pathogens were *all tested negative* in the nasopharyngeal specimens.

References

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Bartolucci, F. and Forcina, A. (2006). A class of latent marginal models for capture–recapture data with continuous covariates. *Journal of the American Statistical Association*, 101(474):786–794.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, 91(436):1450–1460.

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

Chung, H., Flaherty, B. P., and Schafer, J. L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):723–743.

Deloria Knoll, M., Fu, W., Shi, Q., Prosperi, C., Wu, Z., Hammitt, L. L., Feikin, D. R., Baggett, H. C., Howie, S. R., Scott, J. A. G., et al. (2017). Bayesian estimation of pneumonia etiology: epidemiologic considerations and applications to the pneumonia etiology research for child health study. *Clinical infectious diseases*, 64(suppl_3):S213–S227.

Feikin, D., Scott, J., and Gessner, B. (2014). Use of vaccines as probes to define disease burden. *The Lancet*, 383(9930):1762–1770.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

Gryparis, A., Coull, B. A., Schwartz, J., and Suh, H. H. (2007). Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater boston area. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(2):183–209.

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

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Morrissey, E. R., Juárez, M. A., Denby, K. J., and Burroughs, N. J. (2011). Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully bayesian spline autoregression. *Biostatistics*, 12(4):682–694.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2015). Bayesian nonlinear model selection for gene regulatory networks. *Biometrics*.

Saha, S. K., Schrag, S. J., El Arifeen, S., Mullany, L. C., Islam, M. S., Shang, N., Qazi, S. A., Zaidi, A. K., Bhutta, Z. A., Bose, A., et al. (2018). Causes and incidence of community-acquired serious infections among young children in south asia (anisa): an observational cohort study. *The Lancet*, 392(10142):145–159.

Witte, J. S., Greenland, S., and Kim, L.-L. (1998). Software for hierarchical modeling of epidemiologic data. *Epidemiology*, 9(5):563–566.