Regression Analysis for Probabilistic Cause-of-disease Assignment using Case-control Diagnostic Tests: A Hierarchical Bayesian Approach

Irena Chen* and Zhenke Wu | Department of Biostatistics and Michigan Institute for Data Science | irena@umich.edu

Why should you care?

- **Scientific Goal:** For a disease with multiple causes (*not directly observed*):
  1. Assess the effect of explanatory variables on *cause-specific case fractions (CSCFs),* $\boldsymbol{\pi}(\boldsymbol{X}_i)$, for L causes
     - important for optimizing prevention and treatment strategies
  2. Assess the *overall* CSCFs ($\boldsymbol{\pi}(\boldsymbol{X}_i)$ averaged over the empirical distribution of covariates, $\boldsymbol{\pi}^* = \int \boldsymbol{\pi}(\boldsymbol{X}_i) dG(\boldsymbol{X}_i)$)

- **Data Setting:** *Case-control*, multiple *binary* diagnostic measurements ($\boldsymbol{M}_i$) of disease causes (*with error*)
- **Current approaches to including covariates fall short:**
  - Fully stratified analysis breaks down for sparsely populated strata (Table 1)
  - Unable to **quantify how explanatory variables influence the probabilities** of the unobserved causes

What's the regression model?

- **Our method:** A flexible Bayesian model for incorporating regression covariates in a latent class framework
- $\boldsymbol{W}_i$ = vector of covariates that may influence controls ($\boldsymbol{\nu}_k$) and cases ($\boldsymbol{\eta}_k$)
- $\boldsymbol{X}_i$ = vector of covariates that may influence CSCFs ($\boldsymbol{\pi}$)

Subclass Weight Regression:
$\boldsymbol{\nu_k}(\boldsymbol{W_i}), \boldsymbol{\eta_k}(\boldsymbol{W_i})$ **:**

- $$h_k(W_i; \Gamma_k^c) = \begin{cases} g(\alpha_{ik}^c) \Pi_{s<k}, \{1 - g(\alpha_{is}^c)\} & k < K \\ \Pi_{s<k}\{1 - g(\alpha_{is}^c)\}, & k = K \end{cases}$$
  - c designates control ($\boldsymbol{\nu}$) or cases ($\boldsymbol{\eta}$) subclass weights
  - **Stick-breaking parameterization**
- α_{ik}^c is obtained via **Generalized Additive Models:**

$$\alpha_{ik}^c = \alpha_k^c(W_i; \Gamma_k^c) = \mu_{k0} + \sum_{j=1}^{q_1} f_{jk}(W_i; \beta_{kj}^c) + \widetilde{W_i^T} \gamma_k^c$$

CSCF regression:

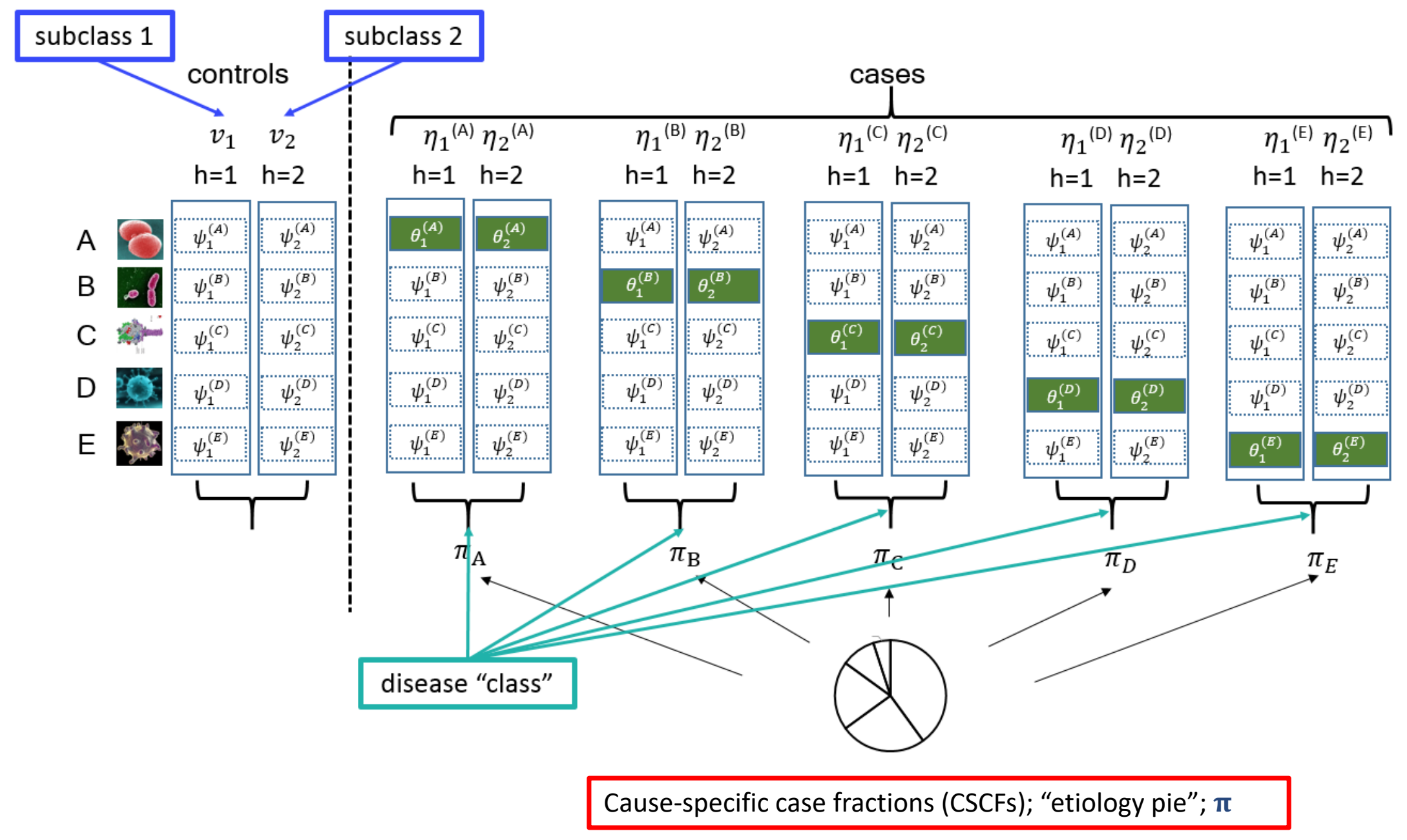
$$\pi_l(\boldsymbol{X}_i) = \frac{\exp(\phi_l(\boldsymbol{X}_i))}{\sum_{l'=1}^{L} \exp(\phi_{l'}(\boldsymbol{X}_i))}$$;$\phi_l(X_i) - \phi_L(X_i)$ = log odds of case *i* in disease class l relative to disease class L

Model $\phi_l(X_i)$ as **additive models:**

- $\phi_l(X_i; \boldsymbol{\Gamma}_l^{\pi}) = \sum_{j=1}^{p_1} f_{lj}^{\pi}(X_i; \beta_{lj}^{\pi}) + \widetilde{X_i^T} \gamma_l^{\pi}$

Specify shrinkage priors on $\mu_{k0}, f_{jk}, f_{lj}^{\pi}$ to encourage parsimonious regressions with few effective subclasses (not shown here)

Existing nested partially-latent class models (npLCM)


subclass 1
subclass 2
controls
cases
ν_1 ν_2
h=1 h=2
$\eta_1^{(A)}$ $\eta_2^{(A)}$
$\eta_1^{(B)}$ $\eta_2^{(B)}$
$\eta_1^{(C)}$ $\eta_2^{(C)}$
$\eta_1^{(D)}$ $\eta_2^{(D)}$
$\eta_1^{(E)}$ $\eta_2^{(E)}$
A
B
C
D
E
$\psi_1^{(A)}$ $\psi_2^{(A)}$
$\theta_1^{(A)}$ $\theta_2^{(A)}$
π_A
π_B
π_C
π_D
π_E
disease "class"
Cause-specific case fractions (CSCFs); "etiology pie"; π

- Estimate $\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\psi}$ (CSCFs, true and false positive rates) via **Markov chain Monte Carlo (MCMC)**

- **Proposed regression extension:** let $\boldsymbol{\nu}_k, \boldsymbol{\eta}_k$, (subclass weights) and $\boldsymbol{\pi}$ (CSCFs) depend on observed covariates

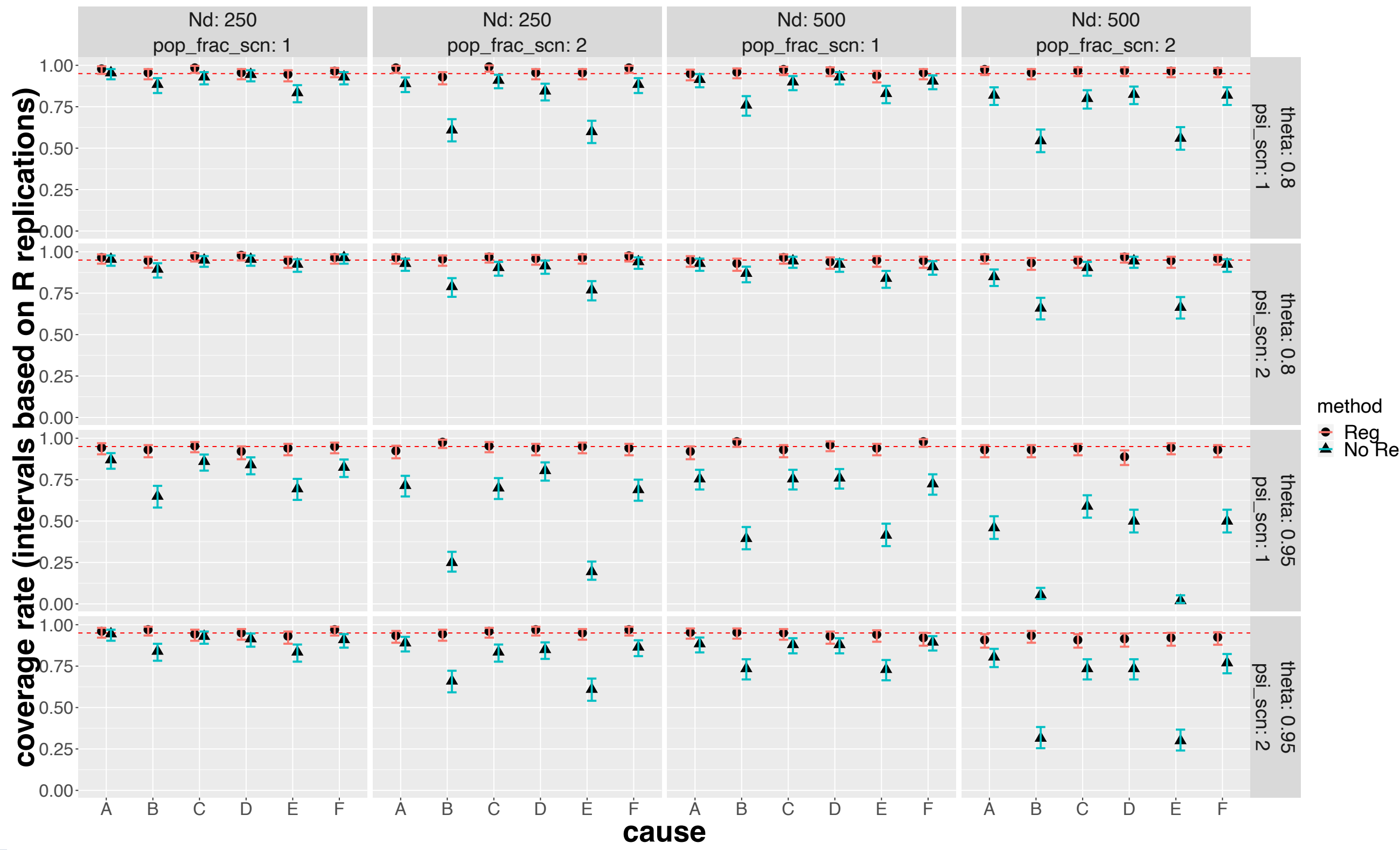
Summary

- **Likelihood for controls:**
- $L_0^{reg} = \Pi_{\mathrm{i:\, Y_i=0}} \sum_{\mathrm{k=1}}^{\mathrm{K}} \nu_{\mathrm{ik}}(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\nu}) \Pi(\boldsymbol{m}; \boldsymbol{\psi}_{\mathrm{k}})$, where $\Pi(\boldsymbol{m}; \boldsymbol{p})$ is the probability of observing $\boldsymbol{m}$ for *J* independent $m_j \sim Bernoulli(p_j)$ and $\boldsymbol{m} \in \{0,1\}^J$

- **Likelihood for cases:**
- $L_1^{reg} = \Pi_{\mathrm{i:\, Y_i=1}} \sum_{\mathrm{l=1}}^{\mathrm{L}} \pi_{\mathrm{l}}(\boldsymbol{X}_i; \boldsymbol{\Gamma}_{\mathrm{l}}^{\pi}) \sum_{\mathrm{k=1}}^{\mathrm{K}} \eta_{\mathrm{ik}}(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\eta}) \Pi(\boldsymbol{m};\ \boldsymbol{\Theta_k}, \boldsymbol{\psi}_{\mathrm{k}})$

- **Unknown** parameters:
  - Etiology regression coefficients: $\boldsymbol{\Gamma}_l^{\pi}$
  - Subclass weights: $\{\boldsymbol{\Gamma}_k^{\eta}\}$ (cases), $\{\boldsymbol{\Gamma}_k^{\nu}\}$ (controls)
  - True/false positive rates: $(\boldsymbol{\Theta} = \{\theta_{\mathrm{k}}^{(\mathrm{j})}\}, \boldsymbol{\Psi} = \{\psi_{\mathrm{k}}^{(\mathrm{j})}\})$
- Use **MCMC** to approximate posterior distribution

Simulation: Improved coverage of 95% credible intervals (CrI) for π_l^*



Childhood pneumonia etiology study

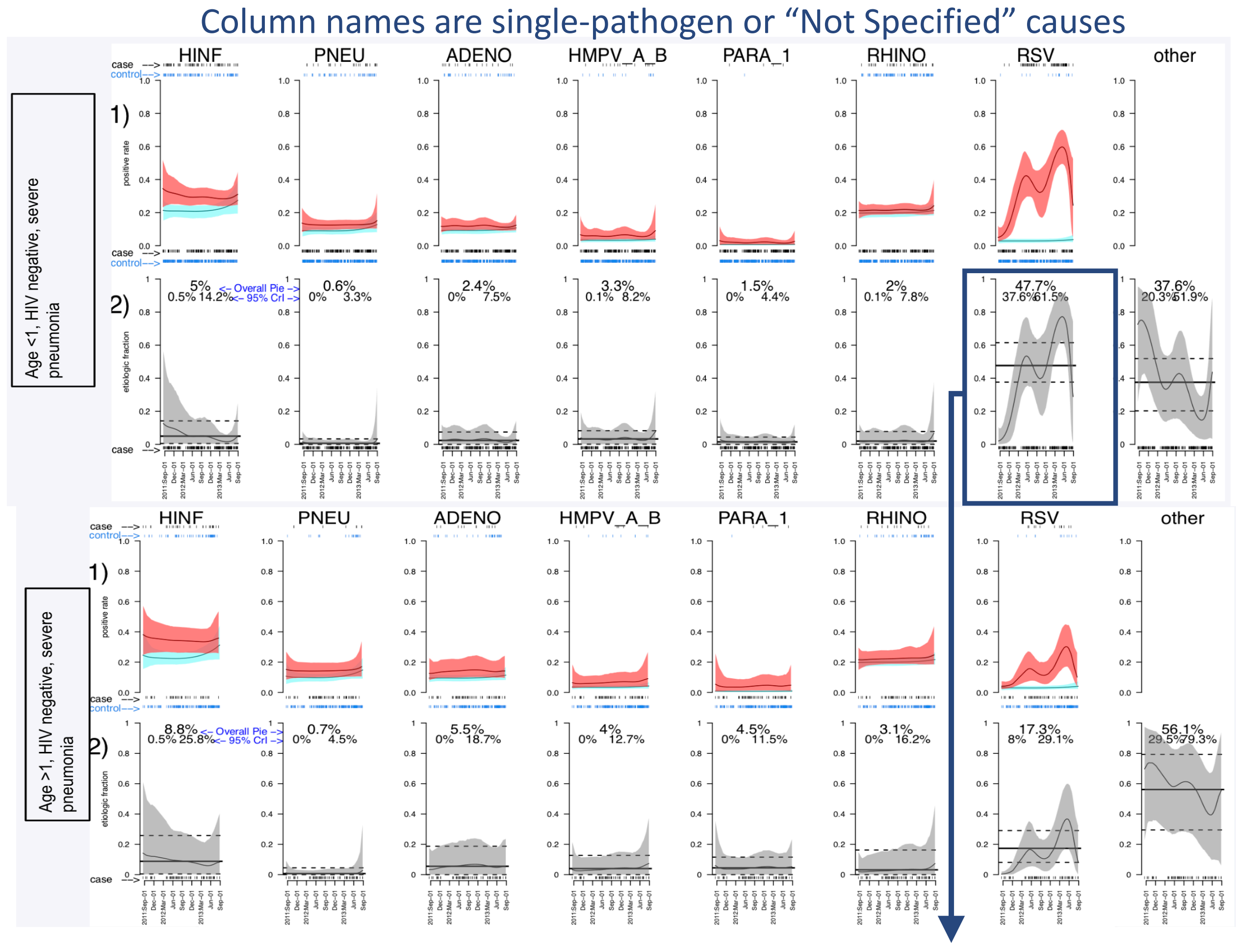
- $\boldsymbol{X}_i$= (age, gender, HIV status, disease severity, enrollment date); $\boldsymbol{W}_i$= $\boldsymbol{X}_i$ minus the disease severity (case-only)
- **Goal:** Evaluate $\pi_l(\boldsymbol{X}_i)$ of seven single-pathogen and "Not Specified" (NoS) causes of lung infection using nasal pharyngeal polymerase chain reaction (NPPCR) tests.

Table 1. The observed counts (frequencies) of controls by age and HIV status; Case counts are further stratified by disease severity (1: yes; 0: no). The marginal case-control positive fractions for each covariate are shown at the bottom. Enrollment date (t) is not stratified upon here.

Age $\geq$ 1	HIV +	# of controls Total: 964 (100)	very severe (case-only)	# of cases Total: 964 (100)
0	0	548 (56.8)	0	208 (40.2)
			1	120 (23.2)
1	0	280 (29.0)	0	69 (13.3)
			1	32 (6.2)
0	1	85. (8.8)	0	37 (7.1)
			1	25 (4.8)
1	1	51 (5.3)	0	24 (4.6)
			1	3 (0.6)
Case: 24.7% Control: 34.3%	17.2% 14.1%		34.7% ---	

Population-level CSCF estimates for two strata

Column names are single-pathogen or "Not Specified" causes



- ▨ 95% credible bands of positive rate curves for controls
- ▨ 95% credible bands of positive rate curves for cases
- ▨ 95% credible bands of CSCF functions (cases)
- ━ Estimated curve (posterior mean)

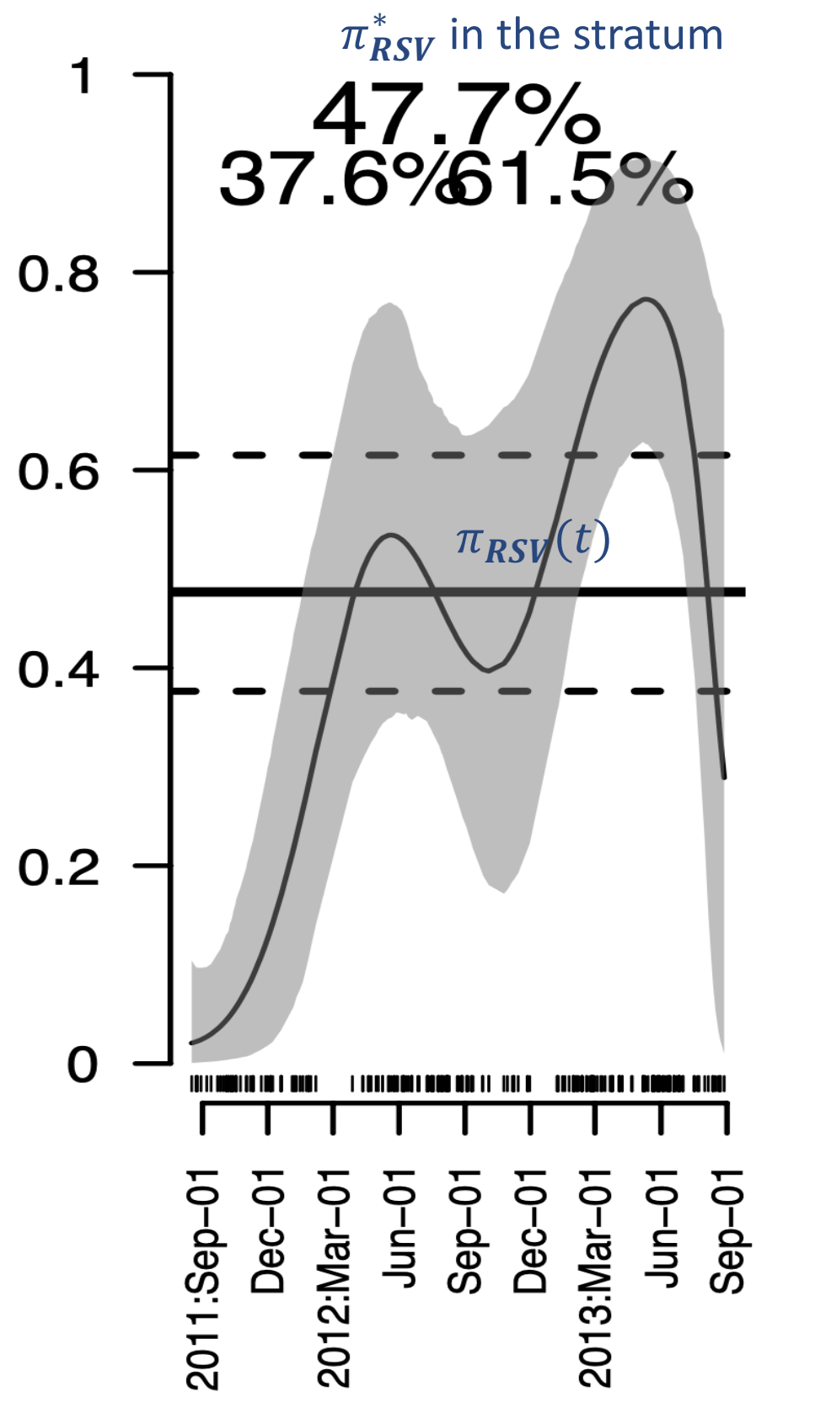
Figure 1: Estimated seasonal CSCFs for two age-HIV-severity strata.

Row 1)

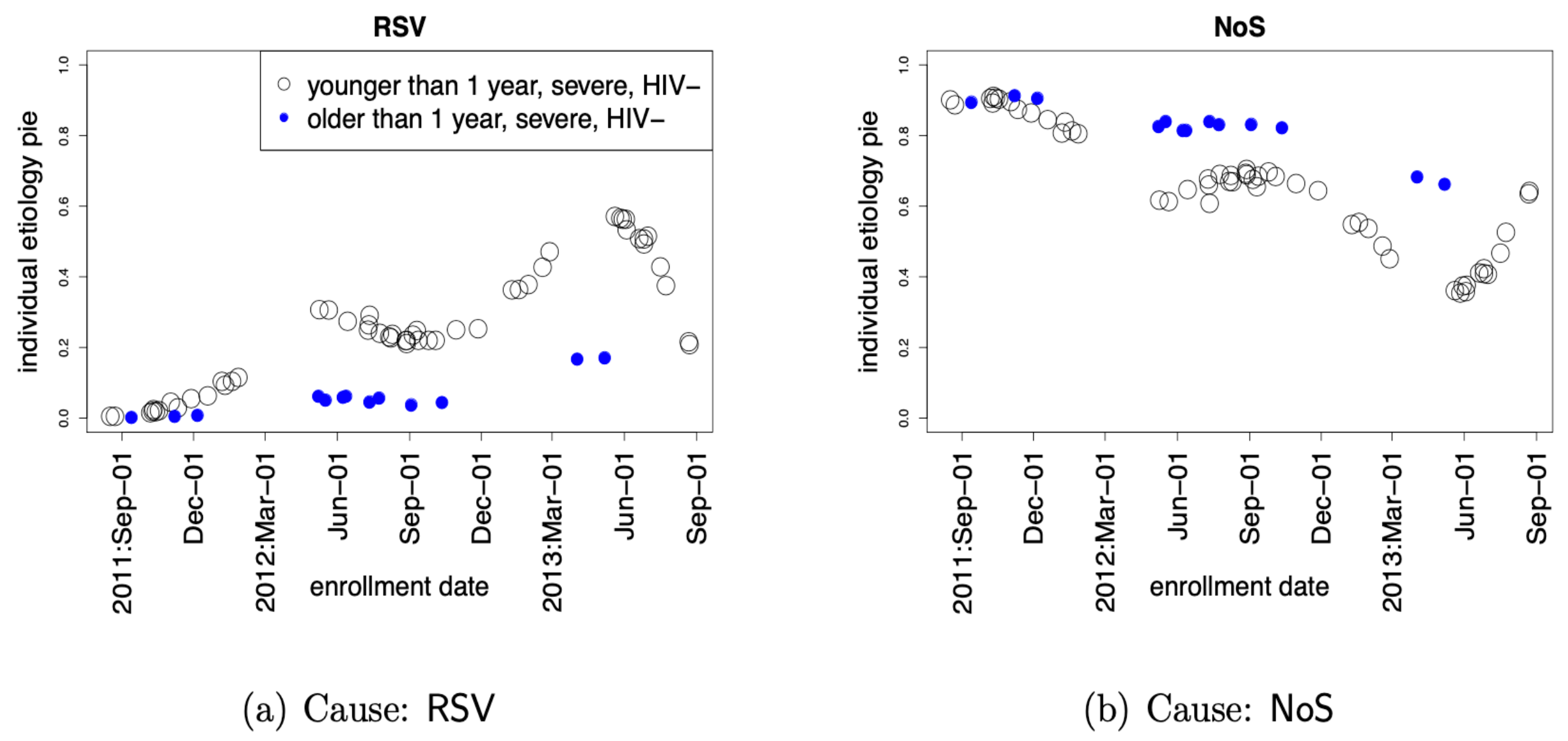
- fitted temporal case (red) and control (blue) positive rate curves enclosed by the pointwise 95% CrIs
- Two rug plots at the top (bottom) indicate the dates of the cases and controls being enrolled and tested positive (negative) for the pathogen.

Row 2)

- Temporal trend of $\hat{\pi}_l$ which is enveloped by pointwise 95% credible bands shown in gray. The estimated overall $\pi *$ averaged among cases is shown by a horizontal solid line
- Two dashed black lines indicating the 2.5% and 97.5% posterior quantiles.



Individual-level estimated probabilities of causes



(a) Cause: RSV

(b) Cause: NoS

Figure 2. Individual probability of cause 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.

Why does this method matter?

- **π regression**: specifies functional dependence of the CSCFs (**π**) upon important covariates
  - improves estimation stability for sparsely populated strata (see Table 1) using assumptions such as additivity

- **Utilizes case-control data**: estimate covariate-dependent reference distribution from controls
  - critical for assigning cause-specific probabilities to cases.

- **Correctly assess posterior uncertainty:** Uses informative priors (**Θ**) only once in the elicited population
  - avoids overly-optimistic uncertainty estimates for **π** (unlike stratified npLCM that reuses these priors)

Future directions

- Explore more flexible regression models (e.g. Bayesian Additive Regression Trees, or BART)
- Extend to not-prespecified causes (combinatorial space)
- Applications to survey data such as *verbal autopsy*

Open-source software (R package)

baker: Bayesian Analysis Kit for Etiology Research
https://github.com/zhenkewu/baker

References:

- **Wu Z. and Chen I** (2019+). Regression Analysis of Dependent Binary Data for Estimating Disease Etiology from Case-Control Studies. Submitted. https://doi.org/10.1101/672808
- **Wu Z** and Zeger SL (2018+). A Bayesian Approach to Restricted Latent Class Models for Scientifically-Structured Clustering of Multivariate Binary Outcomes. https://doi.org/10.1101/400192
- **Wu Z,** Deloria-Knoll M, Zeger S.L, Nested partially latent class models for dependent binary data; estimating disease etiology, Biostatistics, Volume 18, Issue 2, April 2017, Pages 200–213. https://doi.org/10.1093/biostatistics/kxw037
- O'Brien et al. (2019). Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *The Lancet (2019): 394 (10200): 757-779.* https://doi.org/10.1016/S0140-6736(19)30721-4.