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

ZHENKE WU[*,1,2] and IRENA CHEN[1]

[1] *Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA*

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

[*] *zhenkewu@umich.edu*

Summary

Optimal prevention and treatment strategies for a disease of multiple causes must be informed by the population distribution of causes among cases, or cause-specific case fraction (CSCFs) which may further depend on explanatory variables. However, the true causes are often not observed, motivating the use of non-gold-standard diagnostic tests that provide indirect etiologic evidence. Based on case-control multivariate binary data, this paper proposes a novel and unified modeling framework for estimating CSCF functions, closing the existing methodological gap in disease etiology research. With a novel shrinkage prior to encourage parsimonious approximation to a multivariate binary distribution given covariates, the model leverages critical control data for valid probabilistic cause assignment for cases. We derive an efficient Markov chain Monte Carlo algorithm for flexible posterior inference. We illustrate the inference of CSCF functions using extensive simulations and show that the proposed model produces less biased estimates and more valid inference of the overall CSCFs than an analysis that omits covariates. A regression analysis of pediatric pneumonia data reveals the dependence of CSCFs upon season, age, HIV status and disease severity.

[*]To whom correspondence should be addressed.

Key words: Bayesian methods; Case-control studies; Disease etiology; Latent class regression analysis; Measurement errors; Pneumonia; Semi-supervised learning.

1. Introduction

Assessing etiologic contributions for a disease of multiple causes is central to designing effective disease prevention and treatment strategies. The analytic target is the population cause-specific fractions among the cases defined by pre-specified clinical criteria, or *cause-specific case fractions* (CSCFs). Epidemiological factors and individual-level characteristics are often measured in these studies and may influence the CSCFs. The hierarchical Bayesian latent variable regression model proposed in this paper is motivated by the scientific need to assess: 1) the systematic variation of each CSCF as a function of explanatory variables, and 2) the *overall* CSCFs of policy interest obtained by averaging the CSCF functions over an appropriate covariate distribution.

The fundamental challenge to estimating CSCFs is due to the unobservable true causes of disease among cases defined by pre-specified clinical criteria (non-microbiological). This motivates many recent large-scale disease etiology studies to seek peripheral etiologic evidence (e.g., PERCH Study Group, 2019; Saha *and others*, 2018; Kotloff *and others*, 2013). For example, because direct sampling of the infected lung tissues is clinically infeasible, in a pneumonia etiology study (PERCH Study Group, 2019), modern microbiological diagnostic measurements such as real-time polymerase chain reaction (PCR) are made on cases' nasopharyngeal (NP) specimens, outputting presence or absence of a list of putative causes with different error rates. Despite technological advances, the diagnostic specificity remains imperfect. Positive detection of a pathogen in a pneumonia case's nasal cavity does not indicate it caused lung infection, necessitating statistical control. In addition, enormous amounts of resources have been invested to recruit and perform diagnostic tests among *controls*, who do not have the clinically-defined disease, resulting in non-gold-standard diagnostic test data from cases and controls for inferring CSCF functions.

Classical logistic regression and its variants have been the building blocks for estimating disease etiology using case-control data (e.g., Bruzzi *and others*, 1985; Blackwelder *and others*, 2012). However, Deloria Knoll *and others* (2017) identified conceptual and practical issues: 1) they do not build in the unobserved causes of diseases, 2) they are designed to estimate population attributable fractions (PAFs), an incorrect analytic target in our context, and 3) even under a special case where PAFs can be normalized to produce CSCFs, it is still unclear how to let PAFs depend on individual-level covariates.

There are two aims of our work in this article: first, to leverage critical control data when performing regression analysis of CSCFs among cases, which is not studied before in the literature; and second, to correctly assess the posterior uncertainty of the CSCF functions and the overall CSCFs $\boldsymbol{\pi}^*$ in a partially-identified model. To achieve the first aim, we design a hierarchical Bayesian latent variable regression model that explicitly builds in the *categorical* unobserved causes of diseases. The CSCFs are then defined by the population distribution of the unobserved causes among cases, which may vary by covariates. We integrate control data in the model to inform the conditional distribution of the multivariate binary diagnostic tests given each cause. Among controls, the model is based on the latent class regression formulation (Bandeen-Roche *and others*, 1997) and the probability tensor decomposition (Dunson and Xing, 2009). The control model automatically accounts for potential conditional dependence between multiple binary test results given the covariate value. In particular, we propose a covariate-dependent logistic stick-breaking process prior on the latent class profiles which allows introduction of infinitely many latent classes, with the probabilities of latent classes increasingly shrunk towards zero uniformly over covariate values as the class index increases. This allows for data-driven determination of the number of classes needed to model the control data. To integrate case and control data, we share the latent class profile parameters between the controls and the cases in each disease class but let the latent class weights be different. To achieve the second aim, we note the diagnostic

sensitivity parameters are not fully identified based on the likelihood alone. We use a single set of informative prior distributions for these non-identified parameters elicited from laboratory scientists. The proposed Bayesian regression framework overcomes the over-optimism resulting from a stratified analysis which needs multiple sets of independent priors in separate case strata defined by discrete covariates. In addition, through extensive simulation studies, we demonstrate our model produces less biased estimation and more valid inference of the overall CSCFs.

The rest of the paper is organized as follows. Section 2 introduces the motivating application and contextualizes out work. Section 3 formulates the proposed model for estimating CSCF functions based on a single source of case-control non-gold-standard data, specifies novel shrinkage priors, and derives the posterior sampling algorithm. We demonstrate the proposed method via extensive simulations in Section 4 and an application to PERCH data in Section 5. The paper concludes with a discussion on future research directions.

2. Estimating CSCFs

2.1 *Motivating Application and Scientific Challenges*

Pneumonia is a clinical condition associated with infection of the lung tissue, which can be caused by more than 30 different species of microorganisms, including bacteria, viruses, mycobacteria and fungi. The Pneumonia Etiology Research for Child Health (PERCH) study is a seven-country case-control study of the etiology of severe and very severe pneumonia and has enrolled more than $4,000$ hospitalized children under five years of age and more than $5,000$ healthy controls (PERCH Study Group, 2019). The PERCH study aims to understand how the CSCFs vary by factors such as region, season, a child's age, disease severity, nutrition status and human immunodeficiency virus (HIV) status. The cause of lung infection cannot, except in rare cases, be directly observed (Hammitt *and others*, 2017). In Section 5, we will analyze two sources of imperfect measurements of the pathogen causes (those infecting the lung): (a) case-control tests: NP PCR results that are

not perfectly sensitive or specific, referred to as "bronze-standard" (BrS) data; and (b) case-only tests: blood culture (BCX) results for a subset of pathogens that are perfectly specific but lack sensitivity, referred to as "silver-standard" (SS) data.

Valid inference about the population CSCF functions and individual cause-specific probabilities must address three salient data characteristics. First, tests lacking sensitivity may miss the true causative agent(s) which if unadjusted may underestimate the CSCFs. Second, imperfect diagnostic specificities may result in the detection of multiple pathogens that may indicate carriage but not causes of pneumonia. Determining the primary causative agent(s) must use statistical controls. Third, multiple specimens are tested among the cases with only a subset available from the controls. Other large-scale disease etiology studies have raised similar analytic needs (e.g., Saha *and others*, 2018; Kotloff *and others*, 2013).

2.2 *Related Literature*

The scientific problem of estimating CSCFs can be naturally formulated as estimating the mixing weights in a finite-mixture model where the mixture components represent distinct data generating mechanisms under different causes of diseases.

Case-only Methods. A closely related application in demography is to use verbal autopsy (VA) surveys to estimate the cause-specific mortality fractions (CSMFs) in regions without vital registry. Early methods rely on gold-standard cause-of-death information (e.g., King and Lu, 2008); McCormick *and others* (2016) proposed an unsupervised, informative Bayes implementation of a latent class model ("LCMs", Lazarsfeld, 1950), where the latent classes represent unobserved causes of death. Moran *and others* (2019) let covariates influence the conditional distribution given a cause via hierarchical factor regression models. However, these methods do not account for epidemiological factors and individual characteristics that may influence the CSMFs. Datta *and others* (2018) recognizes the variation of CSMFs by covariates and seeks transfer learning

from a source population to a target population with a few observed causes. Finally, VA data are by definition case-only.

Case-control Latent Variable Methods. Methods that use *case-control* data to estimate CSCFs remain sparse; We review the only existing methods. Wu *and others* (2016) introduced a *partially-latent class model* (pLCM) as an extension to classical LCMs. The model treats each case's true cause as a latent categorical variable with multinomial parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top$, or CSCFs, and then assumes each test provides evidence about that cause. In particular, the pLCM is a *semi-supervised* method: it assumes with probabilities $\boldsymbol{\pi}$, a case observation is drawn from a mixture of L components each representing a cause of disease, or "disease class"; Controls have no infection in the lung hence are drawn from an observed class. For a single source of case-control BrS data, each causative pathogen is assumed to be observed with a higher probability (sensitivity, or true positive rate, TPR) in case class ℓ than among the controls; A non-causative pathogen is observed with the same probability as in the controls (1 - specificity, or false positive rate, FPR). Under the pLCM, a higher observed marginal positive rate for a pathogen among cases than controls indicates its etiologic importance. Bayes rule is used to estimate $\boldsymbol{\pi}$ and other parameters via simple Markov chain Monte Carlo (MCMC) algorithms. The latent variable formulation has the unique practical advantage of integrating information from multiple data sources, including extra case-only data and multiple case-control measures, to estimate individual-level probabilities.

The pLCM assumes "local independence" (LI) which means the measurements are mutually independent given the disease class membership. Deviations from LI, or "local dependence" (LD) are testable using the control data, which is modeled by *nested* pLCM (npLCM, Wu, Deloria-Knoll and Zeger, 2017) to reduce the bias in CSCF estimation. In particular, LD is induced in an npLCM by nesting K latent subclasses within each class $\ell = 0, 1, ..., L$, where subclasses respond with distinct vectors of probabilities. The subclasses are nuisance parameters introduced to approximate complex multivariate dependence among discrete data; A truncated stick-breaking

prior for the subclass weights encourages few subclasses and side-steps the choice of the true number of subclasses by using a K deemed sufficiently large for a particular application. Finally, the npLCM is partially-identified (Wu, Deloria-Knoll and Zeger, 2017), necessitating informative priors for a subset of parameters (TPRs).

2.3 *The Central Role of Regression Models*

Covariates $\boldsymbol{X}$ such as season, age, HIV status and disease severity may influence the CSCFs. However, consider a setting with discrete covariates only: a naive *fully-stratified* analysis that fits an npLCM to the case-control data in each covariate stratum is problematic. First, sparsely-populated strata defined by many discrete covariates may lead to unstable CSCF estimates. Second, it is often of policy interest to estimate the overall CSCFs $\boldsymbol{\pi}^*$. Since the informative TPR priors are often elicited for a case population and rarely for each stratum, reusing independent prior distributions of the TPRs across all the strata during multiple npLCM fits will lead to overly-optimistic posterior uncertainty in $\boldsymbol{\pi}^*$, hampering policy decisions. Third, relative to controls, cases may have additional covariates (e.g., disease severity) to stratify upon, resulting in finer case strata nested in each control stratum (e.g., see Table 1), further complicating stratified analyses.

In addition, data from controls provide requisite information about the specificities (1-FPRs) and covariations that may depend on covariates, which must be modeled for valid probabilistic cause assignment. For example, a desired model must use control data to estimate seasonal marginal FPRs of pathogen A so that the presence of pathogen A in a case's nasal cavity does not necessarily indicate etiologic importance by itself during seasons with high asymptomatic carriage rates among controls (high FPRs).

We address all the issues above by proposing a unified regression modeling framework.

3. Proposed Model for Estimating $\boldsymbol{\pi}(\boldsymbol{X})$: a Single Source of Case-Control Data

Notations. Let $Y_i = 1$ indicate a case subject that may belong to L unobserved but pre-specified disease classes as indicated by $I_i \in \{1, \ldots, L\}$. In addition, in disease class ℓ, let $\mathcal{C}_\ell \subseteq \{1, \ldots, J\}$ specify the subset of causative agents among J agents, hence $\mathcal{C}_\ell = \{\ell\}, \ell = 1, \ldots, J$, in the special case of single-agent causes. Section 6 briefly discusses extensions to unknown L and $\{\mathcal{C}_\ell\}$. For case i, let $\boldsymbol{\iota}_i = (\iota_{i1}, \ldots \iota_{iJ})^\top \in \{0,1\}^J$ be a vector of binary values: $\iota_{ij} = 1$ if and only if $j \in \mathcal{C}_{I_i}$ (e.g., lung-infecting pathogens in the PERCH study). Because $\boldsymbol{\iota}_i$ maps one-to-one to $\mathcal{C}_{I_i}$, there are exactly L different possible patterns of $\boldsymbol{\iota}_i$. We assume each element of $\boldsymbol{\iota}_i$ is measured by each of $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top \in \{0,1\}^J$ with error (e.g., NPPCR measures from nasal cavity). Let $Y_i = 0$ indicate a control subject without disease, or equivalently $I_i = 0$, or $\boldsymbol{\iota}_i = \mathbf{0}_{J\times 1}$. Although a control subject has no disease (e.g., the lung not infected by any pathogen: $\boldsymbol{\iota}_i = \mathbf{0}_{J\times 1}$), positive response(s) in $\boldsymbol{M}_i$ may still appear due to measurement error. Let $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$ represent data, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ are covariates that may influence case i's cause-specific probabilities and $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq})^\top$ is a *possibly different* vector of covariates that may influence the distribution of a control's measurements. For notational convenience, we have ordered the continuous variables, if any, in $\boldsymbol{X}_i(\boldsymbol{W}_i)$ as the first $p_1(q_1)$ elements.

3.1 *The Hierarchical Latent Variable Regression Model*

We first characterize the proposed model likelihood by a generative process. The model assumes the measurements for a control subject i are generated in two steps:

$$\text{subclass}: \quad Z_i \mid \boldsymbol{W}_i, Y_i = 0 \sim \text{Categorical}_K\left\{\boldsymbol{\nu}_i\right\}, \boldsymbol{\nu}_i = \boldsymbol{\nu}(\boldsymbol{W}_i) \in \mathcal{S}_{K-1}, \tag{3.1}$$

$$\text{data}: \quad M_{ij} \mid Z_i = k, \iota_{ij} = 0 \sim \text{Bern}\left\{\psi_k^{(j)}\right\}, \text{ independently for item } j = 1, ..., J, \tag{3.2}$$

where, as in npLCM (Wu, Deloria-Knoll and Zeger, 2017), the subclass indicators Z_i's are nuisance quantities for inducing dependence among the multivariate binary responses $\boldsymbol{M}_i$ given covariates;

$\boldsymbol{\nu}_i = (\nu_{i1}, \ldots, \nu_{iK})^\top$ is the vector of control subclass probabilities that may depend on covariates $\boldsymbol{W}_i$, and $\mathcal{S}_n = \{\boldsymbol{r} \in [0,1]^{n+1} : \sum_{m=1}^{n+1} r_m = 1\}$ is the probability simplex. In Section 3.2.1, we propose a novel prior for the probability simplex regression $\boldsymbol{\nu}(\boldsymbol{W}_i)$ to encourage fewer effective subclasses and side-step the choice of K by setting it to a large number appropriate for specific applications. Here $\boldsymbol{\Psi} = \{\psi_k^{(j)} \in (0,1)\}$ is a $J \times K$ matrix, representing the positive response probabilities necessary for modeling the imperfect binary measurements among the controls, referred to as false positive rates (FPRs) or 1-specificity; Let $\boldsymbol{\psi}^{(j)}$ and $\boldsymbol{\psi}_k$ represent the j-th row and k-th column. FPR profile k, $\boldsymbol{\psi}_k$, receives a weight of $\nu_k(\boldsymbol{W}_i)$ for a control subject i with covariates $\boldsymbol{W}_i$, resulting in *marginal* control FPRs $\mathbb{P}(M_j = 1 \mid \boldsymbol{W}, Y = 0, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W})\psi_k^{(j)}$, $j = 1, \ldots, J$, which depend on $\boldsymbol{W}$. The control model (3.1-3.2) reduces to special cases, with covariate-independent $\nu_k(\boldsymbol{W}) \equiv \nu_k$, $k = 1, \ldots, K$, resulting in the control likelihood in a K-subclass npLCM without covariates; A further single-subclass constraint ($K = 1$) gives the control likelihood in the original pLCM. Appendix A1 in the Supplementary Materials further remarks on the control model assumption.

For cases, we add a step to generate the unobserved disease classes (causes) and then generate binary measurements with response probabilities that reflect the causes:

$$\text{disease class}: \quad I_i \mid \boldsymbol{X}_i, Y_i = 1 \sim \mathsf{Categorical}_L\left\{\boldsymbol{\pi}(\boldsymbol{X}_i)\right\}, \boldsymbol{\pi}(\boldsymbol{X}_i) \in \mathcal{S}_{L-1}, \tag{3.3}$$

$$\text{subclass}: \quad Z_i \mid \boldsymbol{W}_i, Y_i = 1 \sim \mathsf{Categorical}_K\left\{\boldsymbol{\eta}_i\right\}, \boldsymbol{\eta}(\boldsymbol{W}_i) \in \mathcal{S}_{K-1}, \tag{3.4}$$

$$\text{data}: \quad M_{ij} \mid Z_i = k, I_i = \ell \neq 0 \sim \mathsf{Bern}\left\{p_{k\ell}^{(j)}\right\}, \text{ indep. for item } j = 1, ..., J, \tag{3.5}$$

$$\text{response probabilities}: \quad p_{k\ell}^{(j)} = \begin{cases} \theta_k^{(j)}, & \iota_{ij} = 1; \\ \psi_k^{(j)}, & \iota_{ij} = 0, \end{cases} \tag{3.6}$$

where $\boldsymbol{\eta}_{ik} = (\eta_{i1}, \ldots, \eta_{iK})^\top$ is the vector of case subclass probabilities that may depend on covariates $\boldsymbol{W}_i$, $\boldsymbol{\pi}(\boldsymbol{X}_i) = (\pi_1(\boldsymbol{X}_i), \ldots, \pi_L(\boldsymbol{X}_i))^\top$ are L CSCF functions, and $p_{k\ell}^{(j)}$ is the probability of item j being positive in subclass k in a disease class $\ell \neq 0$; The form of $p_{k\ell}^{(j)}$ means that, given ℓ and k, cases respond to item $j \in \mathcal{C}_\ell$ with probability that differs from the controls. Let

$\boldsymbol{\Theta} = \{\theta_k^{(j)} \in (0,1)\}$ be a $J \times K$ matrix, where $\theta_k^{(j)}$ represents the positive response probability in subclass k if item j is causative in a disease class, referred to as true positive rate (TPR) or sensitivity; Let $\boldsymbol{\theta}^{(j)}$ and $\boldsymbol{\theta}_k$ represent the j-th row and k-th column.

We first specify $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{\nu}(\cdot)$, $\boldsymbol{\eta}(\cdot)$; In Section 3.2 we specify novel shrinkage priors to encourage parsimonious functional forms. The directed acyclic graph summarizing the generative process above and the priors is in Figure S1 of the Supplementary Materials.

3.1.1 *CSCF Regression Model.* $\boldsymbol{\pi}(\boldsymbol{X})$ is the primary target of inference. Also of interest are the *overall* CSCFs $\pi_\ell^*(\boldsymbol{X}_a; \boldsymbol{X}_b) = \int \pi_\ell(\boldsymbol{X}) \mathrm{d}G(\boldsymbol{X}_a), \ell = 1 \ldots, L$, where G is a probability or empirical distribution of $[\boldsymbol{X}_a \mid \boldsymbol{X}_b]$ and $(\boldsymbol{X}_a, \boldsymbol{X}_b)$ are non-overlapping subvectors that partition the elements in $\boldsymbol{X}$; We simply write π_ℓ^* when $\boldsymbol{X}_b = \emptyset$. We assume CSCFs depend on $\boldsymbol{X}_i$ via a classical multinomial logistic regression model:

$$\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\} / \sum_{\ell'=1}^{L} \exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}, \ell = 1, ..., L, \tag{3.7}$$

where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case i in disease class ℓ relative to L: $\log \pi_{i\ell}/\pi_{iL}$. Without specifying a baseline category, we treat all the disease classes symmetrically which simplifies prior specification. We further assume additivity in a partially linear model:

$$\phi_\ell(\boldsymbol{x}; \Gamma_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi, \tag{3.8}$$

where $\widetilde{\boldsymbol{x}}$ is the subvector of the predictors $\boldsymbol{x}$ that enters the model for all disease classes as linear predictors which may include an intercept; Let $\Gamma_\ell^\pi = [(\boldsymbol{\beta}_{\ell 1}^\pi)^\top, \ldots, (\boldsymbol{\beta}_{\ell p_1}^\pi)^\top, (\boldsymbol{\gamma}_\ell^\pi)^\top]^\top$ (Section 6 discusses non-additive extensions). For covariates such as enrollment date that serve as proxy for factors driven by seasonality, non-linear functional dependence is expected. In Section 3.2.2, we approximate unknown functions of a standardized continuous variable such as $f_{\ell j}^\pi$ (as well as f_{kj}^ν and f_{kj}^η in Section 3.1.2) via basis expansions and a shrinkage prior to encourage smoothness.

Integrating over L unobserved disease classes and K subclasses in (3.3-3.4), we obtain the

likelihood for cases:

$$L_1^{\mathsf{reg}} = \prod_{i:Y_i=1} \left\{ \sum_{\ell=1}^{L} \left[\pi_\ell(\boldsymbol{X}_i; \Gamma_\ell^\pi) \sum_{k=1}^{K} \{\eta_{ik} \cdot \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})\} \right] \right\}, \tag{3.9}$$

where $\boldsymbol{\Gamma}^\pi = \{\Gamma_\ell^\pi, \ell \leqslant L\}$ and $\Pi(\boldsymbol{m}; \boldsymbol{s}) = \prod_{j=1}^{J} \{s_j\}^{m_{ij}} \{1 - s_j\}^{1-m_{ij}}$ is the probability of observing J independent Bernoulli-distributed random variables with success probabilities $\boldsymbol{s} = (s_1, \ldots, s_J)^\top \in [0,1]^J$; In the following, we parameterize η_{ik} by logistic stick-breaking regression, which is first introduced in the control model.

3.1.2 *Covariate-dependent reference distribution.* The control data serve as reference against which test results from cases are compared when estimating cause-specific probabilities. The control model is a mixture model with covariate-dependent mixing weights in (3.1-3.2).

Control subclass weight regression. We specify ν_{ik} by logistic stick-breaking parameterization:

$$\nu_{ik} = g(\alpha_{ik}^\nu) \prod_{s<k} \{1 - g(\alpha_{is}^\nu)\}, \text{ if } k < K, \text{ and } \prod_{s<k} \{1 - g(\alpha_{is}^\nu)\} \text{ otherwise, where} \tag{3.10}$$

$$\alpha_{ik}^\nu = \alpha_k^\nu(\boldsymbol{W}_i = \boldsymbol{w}; \Gamma_k^\nu) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}^\nu(w_j; \boldsymbol{\beta}_{kj}^\nu) + \widetilde{\boldsymbol{w}}^\top \boldsymbol{\gamma}_k^\nu, \text{ for } k = 1, \ldots, K-1. \tag{3.11}$$

Here $\Gamma_k^\nu = [(\boldsymbol{\beta}_{k1}^\nu)^\top, \ldots, (\boldsymbol{\beta}_{kq_1}^\nu)^\top, (\boldsymbol{\gamma}_k^\nu)^\top]^\top$ and α_{ik}^ν is subject i's linear predictor at stick-breaking step $k = 1, \ldots, K-1$; $g(\cdot) : \mathbb{R} \mapsto [0,1]$ is a link function; In this paper, we use the logistic function $g(\alpha) = 1/\{1 + \exp(-\alpha)\}$ which is consistent with (3.7) so that the priors of the coefficients Γ_k^ν and Γ_ℓ^π can be similar. In addition, the parameterization is amenable to simple and efficient block posterior sampling via Pólya-Gamma augmentation (Linderman, Johnson and Adams, 2015). Generalization to other link functions such as the probit function is straightforward (e.g., Rodriguez and Dunson, 2011). Using the stick-breaking analogy, we begin with a unit-length stick: for a total of $K-1$ stick-breaking events, we break a fraction $g(\alpha_{ik}^\nu)$ from the remaining stick at step k, resulting in segment k of length ν_{ik}, $k = 1, \ldots, K-1$.

The likelihood for controls is: $L_0^{\mathsf{reg}} = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{\psi}_k)$.

Case subclass weight regression. $\boldsymbol{\eta}_k(\boldsymbol{W})$ is also specified via a logistic stick-breaking regression as

in the controls but with different parameters: $\eta_{ik} = g(\alpha^{\eta}_{ik}) \prod_{s<k}\{1 - g(\alpha^{\eta}_{is})\}$, $\forall k = 1, \ldots, K-1$. Given TPRs and the FPRs, $\boldsymbol{\eta}_k(\boldsymbol{W})$ fully determines the joint distribution $[\boldsymbol{M} \mid \boldsymbol{W}, I = \ell \neq 0, \boldsymbol{\Theta}, \boldsymbol{\Psi}]$, hence the measurement dependence in each disease class. We do not assume $\eta_k(\boldsymbol{w}) = \nu_k(\boldsymbol{w}), \forall \boldsymbol{w}$. Consequently, relative to the controls, the diseased individuals may have different strength and direction of dependence between the causative $\{M_j : j \in \mathcal{C}_\ell\}$ and non-causative $\{M_j : j \notin \mathcal{C}_\ell\}$ pathogens, or between the non-causative pathogens in each class. Let the k-th linear predictor $\alpha^{\eta}_{ik} = \alpha^{\eta}_k(\boldsymbol{W}_i = \boldsymbol{w}; \Gamma^{\eta}_k) = \mu_{k0} + \sum_{j=1}^{q_1} f^{\eta}_{kj}(w_j; \boldsymbol{\beta}^{\eta}_{kj}) + \widetilde{\boldsymbol{w}}^{\top}\boldsymbol{\gamma}^{\eta}_k$, where f^{η}_{kj} and f^{ν}_{kj} share basis functions but $\Gamma^{\eta}_k = [(\boldsymbol{\beta}^{\eta}_{k1})^{\top}, \ldots, (\boldsymbol{\beta}^{\eta}_{kq_1})^{\top}, (\boldsymbol{\gamma}^{\eta}_k)^{\top}]^{\top}$ differs from the control counterpart (Γ^{ν}_k). In addition, we shared the intercepts $\{\mu_{k0}\}$ from (3.11) with cases to ensure only important subclasses in the controls are used in the cases. For example, absent covariates $\boldsymbol{W}$, a large and positive μ_{k0} effectively halts the stick-breaking procedure at step k for the controls ($\nu_{k+1} \approx 0$); Applying the same intercept μ_{k0} to the cases makes $\eta_{k+1} \approx 0$.

The joint likelihood for the proposed model is $L^{\mathsf{reg}} = L^{\mathsf{reg}}_1 \times L^{\mathsf{reg}}_0$.

REMARK 3.1 Under an assumption of covariate-independence: $\forall k, \eta_k(\cdot) \equiv \eta_k$, ignoring $\boldsymbol{W}_i$ does not invalidate inference of the overall CSCFs $\boldsymbol{\pi}^*$. To see this, L^{reg}_1 and L^{reg}_0 integrate to $L^*_1 = \prod_{i:Y_i=1} \sum_{\ell=1}^{L} \pi^*_\ell \sum_{k=1}^{K} \eta_k \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$, and $L^*_0 = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu^*_k \Pi(\boldsymbol{M}_i; \boldsymbol{\psi}_k)$, where $\nu^*_k = \int \nu_k(\boldsymbol{W}) \mathrm{d}H(\boldsymbol{W})$ and H is probability or empirical distributions of $\boldsymbol{W}$. The integrated likelihood $L^*_1 L^*_0$ is now an npLCM likelihood without covariates. It can be readily shown that the above also holds when $\boldsymbol{X}$ and $\boldsymbol{W}$ do not share common covariates.

3.2 *Priors*

The number of parameters in $L^{\mathsf{reg}}(\{\boldsymbol{\Gamma}^{\pi}_\ell\}, \{\boldsymbol{\Gamma}^{\eta}_k\}, \{\boldsymbol{\Gamma}^{\nu}_k\}, \{\mu_{k0}\}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$ is $\mathcal{O}(JK + LC_{\max}p_1 + KC_{\max}q_1)$ where $C_{\max}$ is the maximum number of basis functions in $\{f^{\pi}_{\ell j}, f^{\nu}_{kj}, f^{\eta}_{kj}\}$; It easily exceeds the number of observed distinct binary measurement patterns. To overcome potential overfitting and increase model interpretability, we *a priori* encourage the following two features: (a) few non-trivial subclasses via a novel shrinkage prior over a probability simplex uniformly over covariate

values, and (b) for a continuous variable, constant $\eta_k(\cdot) = \eta_k$ and $\nu_k(\cdot) = \nu_k$, for some k, via shrinkage priors on the P-spline smoothing parameters.

3.2.1 *Uniform Shrinkage Prior for $\boldsymbol{\nu}(\boldsymbol{W})$ and $\boldsymbol{\eta}(\boldsymbol{W})$ to Encourage Few Subclasses.* We propose a novel shrinkage prior over a probability simplex $\mathcal{S}_{K-1}$ towards a subset of vertices uniformly over covariate values. When applied to $\boldsymbol{\nu}(\boldsymbol{W})$, it effectively makes higher-order subclasses *a priori* increasingly unlikely to receive substantial weights. As a result, it encourages using a small number of subclasses to approximate the observed 2^J probability contingency table for the control data in finite samples. 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 $\{g(\alpha^{\nu}_{ik})\}_{k=1}^{K-1}$ in (3.10) close to one *a posteriori* by making the posterior mean of one of $\{\alpha_{ik}\}_{k=1}^{K}$ large. We accomplish this by specifying a novel additive prior on the intercept in (3.11):

$$\mu_{k0} = \sum_{m=1}^{k} u_{km}\mu^*_{k0},\ u_{km} \geqslant 0,\quad \mu^*_{k0} \sim \mathcal{N}_+(0, \tau_{k0}),\quad \tau_{k0} \sim \mathsf{Gamma}(a_0, b_0), k = 1, \ldots, K-1.$$

where $\{u_{km}\}$ is a pre-specified triangular array. In this paper, we use $u_{km} = 1, m = 1, \ldots, k$; Other choices such as $u_{km} = \mathbf{1}\{k = m\}$ or $u_{km} = 1/k$ may be useful in other settings. Here $\mathcal{N}_+(\mu, \tau)$ represents a Gaussian distribution with mean μ, precision τ truncated to the positive half. We set shape $a_0 = \nu_0/2$, and rate $b_0 = \nu_0 s_0^2/2$. Marginalized over τ_{k0}, μ^*_{k0} has a truncated scaled-t distribution with degree of freedom ν_0 and scale s_0, which peaks at zero and admits large positive values. For a fixed ν_0, the scale parameter s_0 modulates the tendency for the prior density of $\boldsymbol{\nu}(\boldsymbol{W})$ to concentrate towards a few vertices in a probability simplex (see Figure 1).

3.2.2 *Encourage Constant $f^{\pi}_{\ell j}$, f^{ν}_{kj}, f^{η}_{k}.* We use B-splines to approximate the additive functions of a standardized continuous variable (Lang and Brezger, 2004): $f^{\nu}_{kj}(\cdot) = \sum_{c=1}^{C_j} \beta^{(c),\nu}_{kj} B^{(c)}_j(\cdot)$, $f^{\eta}_{kj}(\cdot) = \sum_{c=1}^{C_j} \beta^{(c),\eta}_{kj} B^{(c)}_j(\cdot)$ where $\{B^{(c)}_j(\cdot) : c = 1, \ldots, C_j\}$ are shared C_j cubic B-spline bases; We assume distinct coefficients: $\boldsymbol{\beta}^{\nu}_{kj} = \{\beta^{(c),\nu}_{kj}, c \leqslant C_j\}$ and $\boldsymbol{\beta}^{\eta}_{kj} = \{\beta^{(c),\eta}_{kj}, c \leqslant C_j\}$. In addition, $f^{\pi}_{\ell j}(\cdot) = \sum_{c=1}^{C^{\pi}_j} \beta^{(c),\pi}_{kj} B^{(c),\pi}_j(\cdot)$ where $\{B^{(c),\pi}_j(\cdot) : c = 1, \ldots, C^{\pi}_j\}$ are also cubic B-spline bases and

$\boldsymbol{\beta}^{\pi}_{\ell j} = \{\beta^{(c),\pi}_{\ell j}, c \leqslant C^{\pi}_{j}\}$. 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 $M+4$ basis functions; It readily extends to different numbers of basis functions in the above. We restrict $\{f^{\pi}_{\ell j}, j \leqslant p_1\}$, $\{f^{\nu}_{kj}, f^{\eta}_{kj}, j \leqslant q_1\}$ to have zero means for statistical identifiability.

Since the prior specifications below apply to $\boldsymbol{\beta}^{\pi}_{\ell j}$, $\boldsymbol{\beta}^{\nu}_{kj}$ and $\boldsymbol{\beta}^{\eta}_{kj}$, for notational simplicity, we omit the superscripts π, ν, η and use "$\bullet$" as a placeholder. We specify Gaussian random walk priors on the basis coefficients via Bayesian P-splines (Lang and Brezger, 2004): $\boldsymbol{\beta}^{\bullet}_{kj} \mid \tau^{\bullet}_{kj} \sim \mathcal{N}(\mathbf{0}_{C^{\bullet}_j \times 1}, \tau^{\bullet}_{kj}\boldsymbol{K}^{\bullet})$, where the symmetric penalty matrix $\boldsymbol{K}^{\bullet} = (\Delta^{\bullet}_1)^{\top}\Delta^{\bullet}_1$ is constructed from the first-order difference matrix $\Delta^{\bullet}_1$ of dimension $(C^{\bullet}_j - 1) \times C^{\bullet}_j$ that maps adjacent B-spline coefficients to $\beta^{(c),\bullet}_{kj} - \beta^{(c-1),\bullet}_{kj}$, $c = 2, \ldots, C^{\bullet}_j$, and $\tau^{\bullet}_{kj}$ is the smoothing parameters with large values leading to smoother fit of $f^{\bullet}_{kj}(\cdot)$ (constant when $\tau^{\bullet}_{kj} = \infty$). The precision matrix $\boldsymbol{K}^{\bullet}$ is not full rank and leaves the prior of $\beta^{(1),\bullet}_{kj}$ unspecified; we assume independent priors $\beta^{(1),\bullet}_{kj} \sim \mathcal{N}(0, k_{\beta})$. We specify a mixture prior for smoothing parameter $\tau^{\bullet}_{kj}$:

$$\tau^{\bullet}_{kj} \sim \xi^{\bullet}_{kj}\mathsf{Gamma}(a_{\tau}, b_{\tau}) + (1 - \xi^{\bullet}_{kj})\mathsf{IP}(a'_{\tau}, b'_{\tau}),\ \xi^{\bullet}_{kj} \sim \mathsf{Bern}(\rho^{\bullet}),\ \rho^{\bullet} \sim \mathsf{Beta}(a^{\bullet}_{\rho}, b^{\bullet}_{\rho}), \tag{3.12}$$

where the Gamma-distributed component concentrates near smaller values and the inverse-Pareto component $\mathsf{IP}(\tau; a, b) = \frac{a}{b}\left(\frac{\tau}{b}\right)^{a-1}, a > 0, 0 < \tau < b$, prefers larger values; This bimodal mixture distribution creates a sharp separation between flexible and constant fits (Morrissey *and others*, 2011; Ni, Stingo and Baladandayuthapani, 2015).

3.2.3 *Prior Distributions for Other Parameters* We assume independent Gaussian priors $\mathcal{N}(0, \kappa_{\gamma})$ for each element of $\boldsymbol{\gamma}^{\pi}_{\ell}$, $\boldsymbol{\gamma}^{\nu}_{kj}$ and $\boldsymbol{\gamma}^{\eta}_{kj}$. See Appendix A2 in the Supplementary Materials for the choice of hyperparameters (a_{τ}, b_{τ}), (a'_{τ}, b'_{τ}), $(a^{\nu}_{\rho}, b^{\nu}_{\rho})$, $(a^{\eta}_{\rho}, b^{\eta}_{\rho})$, $(a^{\pi}_{\rho}, b^{\pi}_{\rho})$, ν_0, s_0, κ_{β}, κ_{γ}. The npLCM regression model is partially-identified (Jones *and others*, 2010). We assume independent $\theta^{(j)}_k \sim \mathsf{Beta}(a_j, b_j)$, $j \leqslant J$. Hyperparameters (a_j, b_j) are chosen so that the 2.5% and 97.5% quantiles match a prior range elicited from laboratory scientists (Deloria Knoll *and others*, 2017).

3.3 *Posterior Inference and Software*

Appendix A3 in the Supplementary Materials derives the Markov chain Monte Carlo (MCMC) algorithm that draws posterior samples of the unknowns to approximate their joint posterior distribution (Gelfand and Smith, 1990). Flexible posterior inferences about any functions of the model parameters and individual latent variables are available by plugging in the posterior samples of the unknowns. All the models in this paper are fitted using a free and publicly available R package `baker` (`https://github.com/zhenkewu/baker`).

4. Simulations

We simulate case-control BrS measurements along with observed continuous and/or discrete covariates under multiple combinations of true parameter values and sample sizes that mimic the motivating PERCH study. In Simulation I, we illustrate flexible statistical inferences about the CSCF functions. In Simulation II, we focus on the overall CSCFs; Let π_ℓ^* be an empirical average of $\pi_\ell(\boldsymbol{X})$, $\ell = 1, \ldots, L$. We compare the frequentist properties of the posterior mean $\boldsymbol{\pi}^*$ obtained from analyses with or without covariates upon repeated use across replications (Little *and others*, 2011); We compare the proposed model against npLCMs without covariates, because the latter is the only available method for estimating CSCFs using case-control data. Regression analyses reduce estimation bias, retain efficiency and provide more valid frequentist coverage of the 95% credible intervals (CrIs). The relative advantage varies by the true data generating mechanism and sample sizes.

In all analyses below, we use independent Beta(7.13,1.32) TPR prior distributions that match 0.55 and 0.99 with the lower and upper 2.5% quantiles, respectively; The priors for other parameters are specified in Section 3.2.

<u>*Simulation I.*</u> We demonstrate posterior inference of true CSCF functions $\{\pi_\ell^0(\boldsymbol{X})\}$. We simulate $N_d = 500$ cases and $N_u = 500$ controls for each of two levels of S (a discrete covariate) and

uniformly sample the subjects' enrollment dates over a period of 300 days. Appendix A4 in the Supplementary Materials specifies the true data generating mechanism. Based on the simulated data, pathogen A has a bimodal case positive rate curve mimicking the trends observed of RSV in one PERCH site; other pathogens have overall increasing case positive rate curves over enrollment dates. We set the simulation parameters in a way that the *marginal* control rate may be higher than cases for small t's. Row 2 of Figure 2 shows that for the 9 causes (by column), the posterior means and 95% CrIs for the CSCF functions $\pi_\ell(\cdot)$ well recover the simulation truths $\pi_\ell^0(\cdot)$. Figure S2 in the Supplementary Materials further demonstrates well-recovered subclass weight curves; Appendix A5 in the Supplementary Materials provides additional simulation results that shows the true $\pi_\ell^0(X)$ is well-recovered for a discrete covariate X.

Simulation II. We show the regression model accounts for population stratification by covariates hence reduces the bias of the posterior mean $\{\widehat{\pi}_\ell^*\}$ in estimating the overall CSCFs ($\boldsymbol{\pi}^*$) and produces more valid 95% CrIs. We illustrate the advantage of the regression approach under simple scenarios with a single two-level covariate $X \in \{1, 2\}$; We let $W = X$. We perform npLCM regression analysis with $K = 3$ for each of $R = 200$ replication data sets simulated under each scenario detailed in Appendix A4 in the Supplementary Materials corresponding to distinct numbers of causes, sample sizes, relative sizes of CSCF functions (rare versus popular causes), signal strengths (more discrepant TPRs and FPRs indicate stronger signals, Wu *and others* (2016)), and effects of W on $\{\nu_k(W)\}$ and $\{\eta_k(W)\}$.

In estimating π_ℓ^*, we evaluate the bias $\widehat{\pi_\ell^*} - \pi_\ell^{0*}$, where $\pi_\ell^{0*} = N_d^{-1} \sum_{i:Y_i=1} \pi_\ell^0(\boldsymbol{X}_i)$ is the true overall CSCF, and $\widehat{\pi_\ell^*} = N_d^{-1} \sum_{i:Y_i=1} \widehat{\pi}_\ell(\boldsymbol{X}_i)$ is an empirical average of the posterior mean CSCFs at $\boldsymbol{X}_i$. We also evaluate the empirical coverage rates of the 95% CrIs.

The proposed model incorporates covariates and performs better in estimating $\boldsymbol{\pi}^*$ than a model that omits covariates. For example, Figure 3(a) shows for $J = 6$ that, relative to no-covariate npLCM analyses, regression analyses produce posterior means that on average have

negligible relative biases (percent difference between the posterior mean and the truth relative to the truth) for each pathogen across simulation scenarios. As expected, we observe slight relative biases from the regression model in the bottom two rows of Figure 3(a), because the informative TPR prior Beta(7.13,1.32) has a mean value lower than the true TPR 0.95; A more informative prior further reduces the relative bias; See additional simulations in Appendix A5 in the Supplementary Materials on the role of informative TPR priors. Regression analyses also produce 95% CrIs for π_ℓ^* that have more valid empirical coverage rates in all the scenarios (Figure 3(b)). Misspecified models without covariates concentrate the posterior distribution away from the true overall CSCFs, resulting in severe under-coverage.

5. Results from PERCH Study

We restrict attention in this regression analysis to 518 cases and 964 controls from one of the PERCH study sites in the Southern Hemisphere that collected information on enrollment date (t, August 2011 to September 2013; standardized), age (dichotomized to younger or older than one year), HIV status (positive or negative), disease severity for cases (severe or very severe), and presence or absence of seven species of pathogens (five viruses and two bacteria, representing a subset of pathogens evaluated) in NPPCR; We also include in the analysis the BCX results for two bacteria from cases only. Detailed analyses of the entire data are reported in PERCH Study Group (2019). Table 1 shows the observed frequencies in the $\boldsymbol{W} = (\mathsf{age},\ \mathsf{HIV\ status})$ strata for controls and $\boldsymbol{X} = (\mathsf{age},\ \mathsf{HIV\ status},\ \mathsf{disease\ severity})$ strata for cases. The two case strata with the most subjects are severe pneumonia children who were HIV negative, under or above one year of age. Table 1 has small cell counts that preclude fitting npLCMs by stratum. Figure S5 in the Supplementary Materials shows summary statistics for the NPPCR (BrS) and BCX (SS) data including the positive rates in the cases and the controls and the conditional odds ratio (COR) contrasting the case and control rates adjusting for the presence or absence of other pathogens

(NPPCR data only).

For NPPCR, pathogens RSV and *Haemophilus influenzae* (HINF) are detected with the highest positive rates among cases: 29.3% and 34.1%, respectively, which are higher than the corresponding control rates (3.1% and 21.7%). The CORs are large, 14 (95%CI: $9.4, 21.6$) for RSV and 1.8 (95%CI: $1.3, 2.3$) for HINF, indicating etiologic importance. Adenovirus (ADENO) also has a statistically significant COR of 1.5 (95%CI: $1.1, 2.2$). Human metapneumovirus type A or B (HMPV_A_B) and Parainfluenza type 1 virus (PARA_1) have larger positive and statistically significant CORs of 2.6 (95%CI: $1.5, 4.4$) and 6.4 (95%CI: $2.3, 20.3$). However, the two pathogens rarely appear in cases' nasal cavities (HMPV_A_B: 6.8%, PARA_1: 2.3%), which in light of high sensitivities $(50 \sim 90)\%$ means non-primary etiologic roles. For the rest of pathogens, we observed similar case and control positive rates as shown by the statistically non-significant CORs (RHINO (case: 21.4%; control: 19.9%) and *Streptococcus pneumoniae* (PNEU) (case: 14.4%; control: 9.9%). Similar to Wu, Deloria-Knoll and Zeger (2017), we integrate case-only SS measurements for HINF and PNEU by using informative priors of the sensitivities to adjust the CSCF estimates in a coherent Bayesian framework. It is expected that the rare detection of the two bacteria, 0.4% for HINF and 0.2% for PNEU from SS data, will lower their CSCF estimates relative to the ones obtained from an NPPCR-only analysis.

To fit the model, we include in the regression analysis seven single-pathogen causes $\mathcal{C}_\ell = \ell$, $\ell = 1, \ldots, J(= 7)$ and a "Not Specified (NoS)" cause denoted by $\mathcal{C}_{\mathsf{NoS}} = \emptyset$ to account for other non-targeted causative agents. We incorporate the prior knowledge about the TPRs of the NPPCR measures from laboratory experts. We set the Beta priors for sensitivities by $a_\theta = 126.8$ and $b_\theta = 48.3$, so that the 2.5% and 97.5% quantiles match the lower and upper ranges of plausible sensitivity values of 0.5 and 0.9, respectively. We use a working number of subclasses $K = 5$; Results under larger Ks remain nearly the same. In the presence of SS data for a subset of $J^{\mathsf{SS}} = 2$ pathogens (e.g., only bacteria can be cultured from blood), we multiple L^{reg} by SS data

likelihood with LI assumption and specify the $\mathsf{Beta}(7.59,58.97)$ prior for the two TPRs of SS measurements with a lower range of $5-20\%$ based on existing vaccine probe trials (e.g., Feikin, Scott and Gessner, 2014). For SS data, we assume perfect specificity. In the CSCF regression model $f^{\pi}_{\ell j}(t)$, we use 7 d.f. for B-spline expansion of the additive function for the standardized enrollment date t at uniform knots along with three binary indicators for age older than one, HIV positive, very severe pneumonia; In the subclass weight regression model, we use 5 d.f. for the standardized enrollment date t with uniform knots and two indicators: $\mathbf{1}\{\mathsf{age}_i \geqslant 1\ \mathsf{year}\}$ and $\mathbf{1}\{\mathsf{HIV}_i = 1\}$. The prior distributions for other parameters follow the specification in Section 3.2.

The regression analysis produces seasonal estimates of the CSCF function for each cause that varies in trend and magnitude among the eight case strata defined by age, HIV status and disease severity. Figure 4 shows among two age-HIV-severity strata the posterior mean curve and 95% pointwise credible bands of the CSCF functions $\pi_\ell(t, \mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ as a function of t. For example, among the younger, HIV negative and severe pneumonia children (Figure 4(a)), the CSCF curve of RSV is estimated to have a prominent bimodal temporal pattern that peaked at two consecutive winters in the Southern Hemisphere (June 2012 and 2013), prioritizing preventative measures and treatment algorithms for RSV. Other single-pathogen causes HINF, PNEU, ADENO, HMPV_A_B and PARA_1 have overall low and stable CSCF curves across seasons. As a result, the estimated CSCF curve of NoS shows a trend with a higher level of uncertainty that is complementary to RSV. In contrast, Figure 4(b) shows a lower degree of seasonal variation of the RSV CSCF curve among the older, HIV negative and severe pneumonia children.

The regression model also performs individual-level cause-specific probability assignment given a case's measurements and automatically use covariate values during assignment. Figure S6 in the Supplementary Materials show distinct cause-specific probabilities for two cases (one older and the other younger than one) with all-negative NPPCR results.

Given age, HIV status and disease severity, we quantify the overall CSCFs $\boldsymbol{\pi}^*(\mathsf{t}; \mathsf{age}, \mathsf{HIV}, \mathsf{severity})$

by averaging the CSCF function estimates over the empirical distribution of enrollment date. Contrasting the results in the two age-HIV-severity strata in Figure 4(a) and 4(b), the case positive rate of RSV among the older children drops from 39.3% to 17.9% but the control positive rates remain similar (from 3.0% to 4.1%). The overall CSCF of RSV (π^*_{RSV}) is estimated to drop from 47.7 (95% CrI : $37.6, 61.5$)% to 17.3 (95% CrI : $8.0, 29.1$)%; The CSCF for NoS (π^*_{NoS}) is estimated to increase from 37.6 (95% CrI : $20.3, 51.9$)% to 56.1 (95% CrI : $29.5, 79.3$)%; The overall CSCFs for other causes remain similar between the younger and older pneumonia children.

6. Discussion

In disease etiology studies where gold-standard data are infeasible to obtain, epidemiologists need to integrate multiple sources of data of distinct quality to draw inference about the population CSCFs and individual cause-specific probabilities that may depend on covariates. While the only existing methods for case-control data based on npLCM account for imperfect diagnostic sensitivities and specificities, complex measurement dependence and missingness, they do not describe the relationship between covariates and the CSCFs. This paper fills this analytic gap by extending npLCM to a unified hierarchical Bayesian regression modeling framework that for the first time estimates CSCF functions and is amenable to efficient posterior computation. We also propose novel shrinkage priors to encourage parsimonious approximation to a multivariate binary distribution given covariate values that may have broader applications.

The proposed approach has three distinguishing features: 1) It allows an analyst to specify a model for the functional dependence of the CSCFs upon important covariates; Assumptions such as additivity further improves estimation stability for sparsely-populated strata defined by many discrete covariates. 2) The model incorporates control data to infer the CSCF functions. The posterior algorithm estimates a parsimonious covariate-dependent reference distribution of the diagnostic measurements from controls, which is critical for correctly assigning cause-specific

probabilities for individual cases. Finally, 3) the model uses informative priors of the sensitivities (TPRs) only once in the entire target population for which these priors were intended. Relative to a fully-stratified npLCM analysis that reuses these priors, the proposed regression analysis avoids overly-optimistic uncertainty estimates for the overall CSCFs.

We have shown via simulations that the regression approach accounts for population stratification by important covariates and, as expected, reduces estimation biases and produces 95% credible intervals that have more valid empirical coverage rates than an npLCM analysis that omits covariates. In addition, the proposed regression analysis can readily integrate multiple sources of diagnostic measurements with distinct levels of diagnostic sensitivities and specificities, a subset of which are only available from cases (SS data). Our regression analysis integrates the BrS and SS data from one PERCH site and reveals prominent dependence of the CSCFs upon seasonality and a pneumonia child's age, HIV status and disease severity.

Future work may further expand the utility of the proposed methods. First, flexible and parsimonious alternatives to the additive models may capture important interaction effects (e.g., Linero, 2018). Second, in the presence of many covariates, class-specific predictor selection methods for $\pi_\ell(\boldsymbol{X}_i)$ may provide further regularization and improve interpretability (Gustafson and Lefebvre, 2008). Third, when the subsets of pathogens $\{\mathcal{C}_\ell\}$ that have caused the diseases in the population are unknown, the proposed method can be combined with subset selection procedures (Gu and Xu, 2019).

Supplementary Materials

The reader is referred to the on-line Supplementary Materials for technical appendices, additional simulation results and supplemental figures referenced in the Main Paper.

Acknowledgments

This work was supported by the Patient-Centered Outcomes Research Institute (PCORI) Award [ME-1408-20318 to Z.W.]; and the National Institutes of Health grants [P30CA046592 to Z.W., I.C., U01CA229437 to Z.W.]. We thank the PERCH study team led by Katherine O'Brien for providing the data and scientific advice, Scott Zeger, Maria Deloria-Knoll, Christine Prosperi and Qiyuan Shi for valuable feedback about `baker` and Jing Chu for preliminary simulations. We also thank Michael Elliott and Abhirup Datta for constructive comments. *Conflict of Interest*: None declared.

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.

Blackwelder, W. C., Biswas, K., Wu, Y., Kotloff, K. L., Farag, T. H., Nasrin, D., Graubard, B. I., Sommerfelt, H. and Levine, M. M. (2012). Statistical methods in the global enteric multicenter study (GEMS). *Clinical Infectious Diseases* **55**(suppl_4), S246–S253.

Bruzzi, P., Green, S., Byar, D., Brinton, L. and Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology* **122**(5), 904–914.

Datta, A., Fiksel, J., Amouzou, A. and Zeger, S. (2018). Regularized Bayesian transfer learning for population level etiological distributions. *arXiv preprint arXiv:1810.10572*.

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., Murdoch, D. R., Madhi, S. A., M, T. D., Brooks, W. A., Kotloff, K. L., Li, M., Park, D. E., Lin, W., Levine, O. S.,

O'Brien, K. L. *and others*. (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.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104**(487), 1042–1051.

Erosheva, E. A., Fienberg, S. E. and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics* **1**(2), 346–384.

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

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**(410), 398–409.

Gu, Y. and Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, In press.

Gustafson, P. and Lefebvre, G. (2008). Bayesian multinomial regression with class-specific predictor selection. *The Annals of Applied Statistics* **2**(4), 1478–1502.

Hammitt, L. L., Feikin, D. R., Scott, J. A. G., Zeger, S. L., Murdoch, D. R., O'brien, K. L. and Deloria Knoll, M. (2017). Addressing the analytic challenges of cross-sectional pediatric pneumonia etiology data. *Clinical infectious diseases* **64**(suppl_3), S197–S204.

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.

King, G. and Lu, Y. (2008). Verbal autopsy methods with multiple causes of death. *Statistical Science* **23**(1), 78–91.

Kotloff, K. L., Nataro, J. P., Blackwelder, W. C., Nasrin, D., Farag, T. H., Panchalingam, S., Wu, Y., Sow, S. O., Sur, D., Breiman, R. F. *and others*. (2013). Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet* **382**(9888), 209–222.

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

Lazarsfeld, P. F. (1950). *The logical and mathematical foundations of latent structure analysis*, Volume IV, Chapter The American Soldier: Studies in Social Psychology in World War II. Princeton, NJ: Princeton University Press, pp. 362–412.

Linderman, S., Johnson, M. and Adams, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation. In: *Advances in Neural Information Processing Systems*. pp. 3456–3464.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association* **113**(522), 626–636.

Little, R. *and others*. (2011). Calibrated bayes, for statistics in general, and missing data in particular. *Statistical Science* **26**(2), 162–174.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K. and Clark, S. J. (2016). Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association* **111**(515), 1036–1049.

Moran, K. R., Turner, E. L., Dunson, D. and Herring, A. H. (2019). Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *arXiv preprint arXiv:1908.07632*.

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* **71**(3), 585–595.

PERCH Study Group. (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* **392**(10200), 757–779.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**(1), 145–177.

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. *and others*. (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.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., Zeger, S. L. and the PERCH Study Team. (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.

Wu, Z., Deloria-Knoll, M. and Zeger, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics (Oxford, England)* **18**, 200–213.

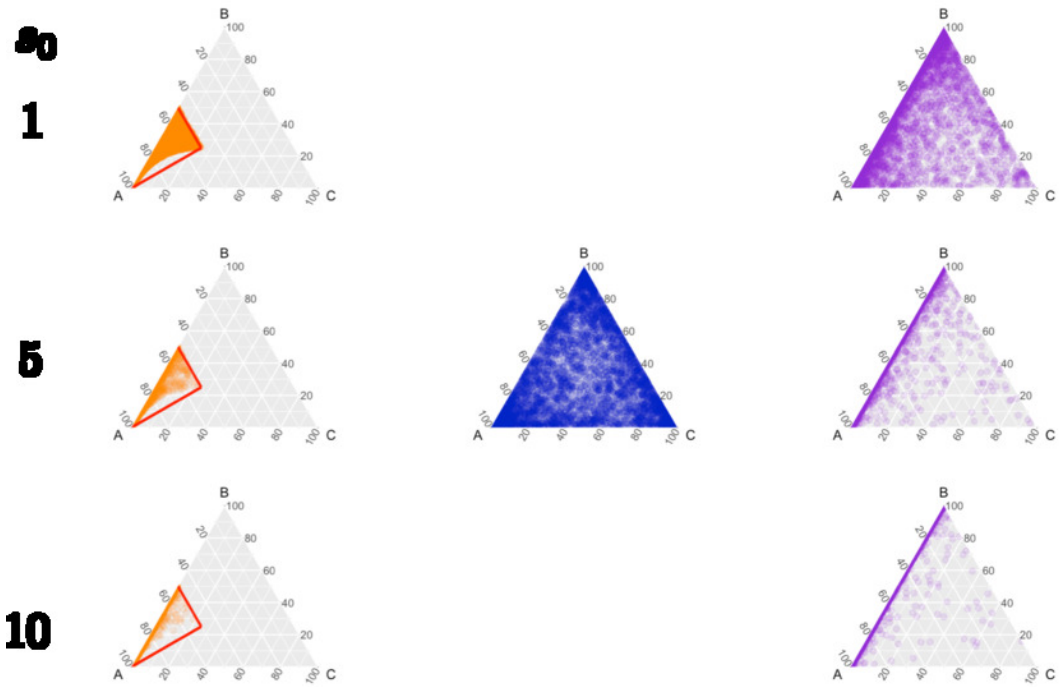


Fig. 1: Simulated draws from the shrinkage prior for three subclass weights $(\nu_A, \nu_B, \nu_C)^\top$; Here we focus on the role of μ_{k0} and do not include covariates. From the top to the bottom, the three rows represent the increasing scale parameter $s_0 = 1, 5, 10$ while fixing $a_0 = 0.5$. We display a three-dimensional non-negative vector that sums to 1 within ternary diagrams. In each row are $5,000$ random draws from the prior of $(\nu_A, \nu_B, \nu_C)^\top$, with $g(\alpha^\nu_{ik}) = g(\mu_{k0} + \gamma^\nu_{k1})$, $k = A, B$, when α^ν_{ik} equals μ^ν_{k0} (Left), γ^ν_{k1} (Middle), or $\mu^\nu_{k0} + \gamma^\nu_{k1}$ (Right; computed from the samples in the left and middle). We show only one ternary diagram in the middle because the prior of γ^ν_{k1} does not depend on s_0; The γ^ν_{k1}, $k = A, B$, here have a prior precision of $\kappa_\gamma = 1/4$, prior means of -1.07 and 0, respectively; these chosen values result in approximately evenly-distributed draws in a ternary diagram (middle).

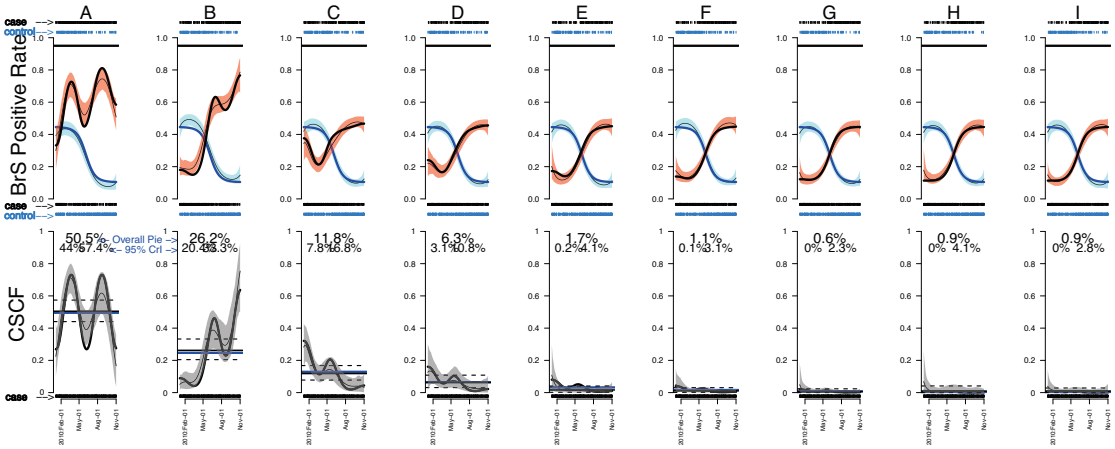
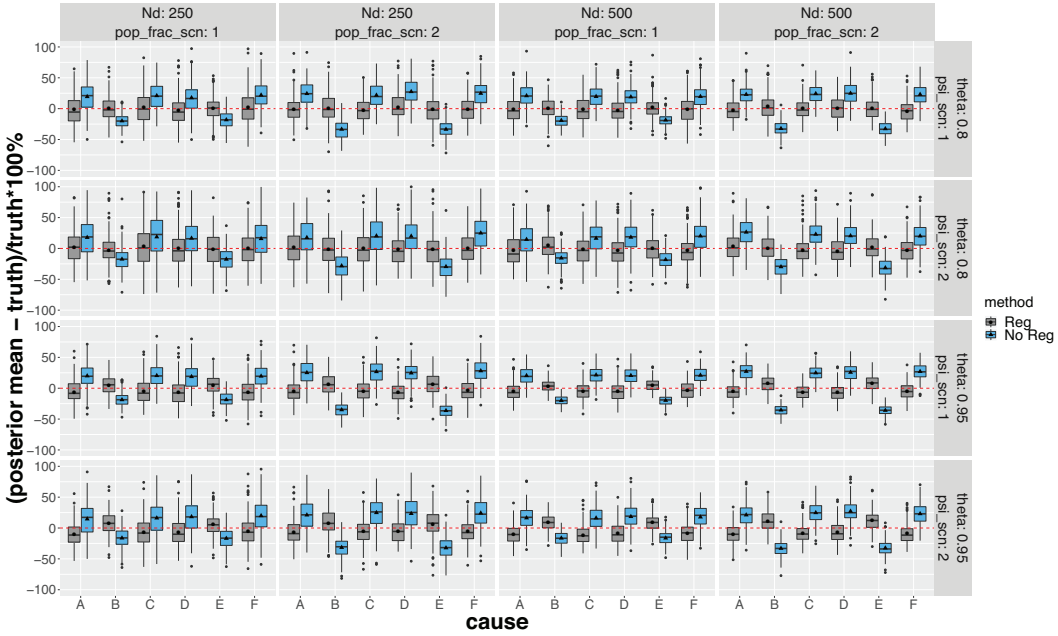
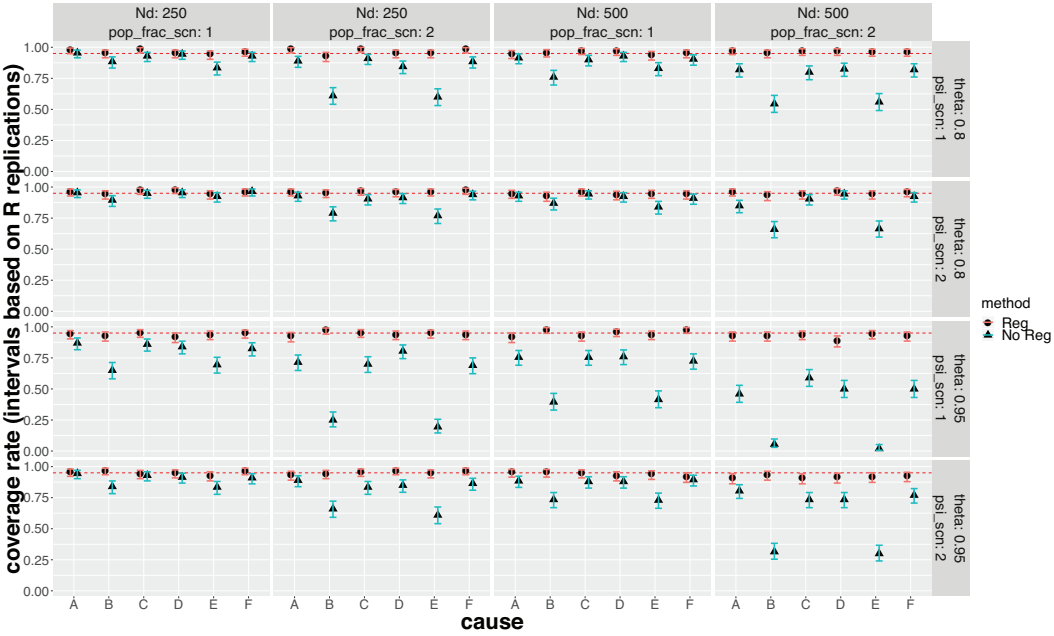


Fig. 2: Row 2) For each of the 9 single-agent causes (by column) in Simulation I, the posterior mean (thin black curves) and pointwise 95% CrIs (gray) for the CSCF curves $\pi_\ell(x)$ are close to the simulation truths $\pi_\ell^0(x)$ (thick black curves). In row 1), the fitted case (red, $\widehat{\mathbb{P}}(M_\ell = 1 \mid \boldsymbol{X}, \boldsymbol{W}, Y = 1)$) and control (blue, $\widehat{\mathbb{P}}(M_\ell = 1 \mid \boldsymbol{W}, Y = 0)$) positive rate curves are shown with the posterior mean curves (thin black curves) along with pointwise 95% credible bands; The rug plots show the positive (top) and negative (bottom) measurements made on cases and controls on the enrollment dates. The solid horizontal lines in row 1 indicate the true TPRs.

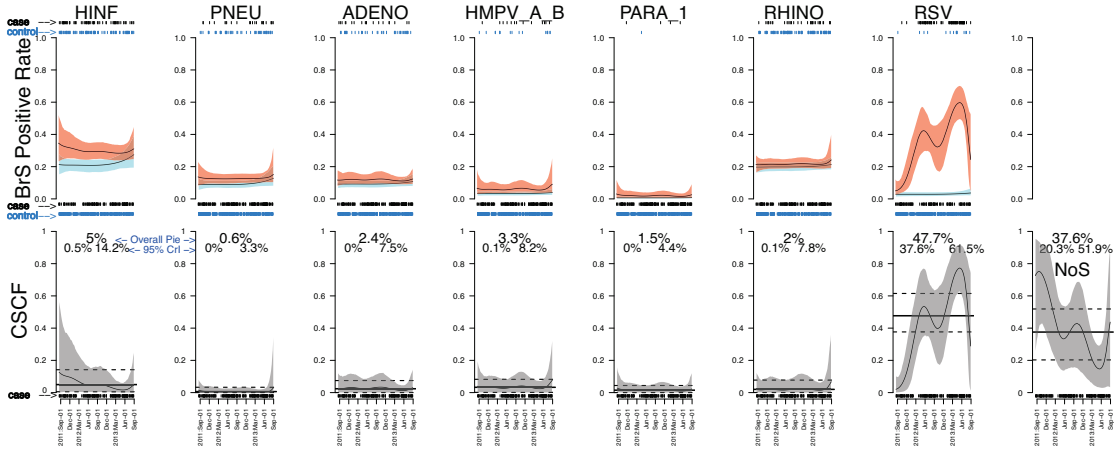


(a) Percent relative bias

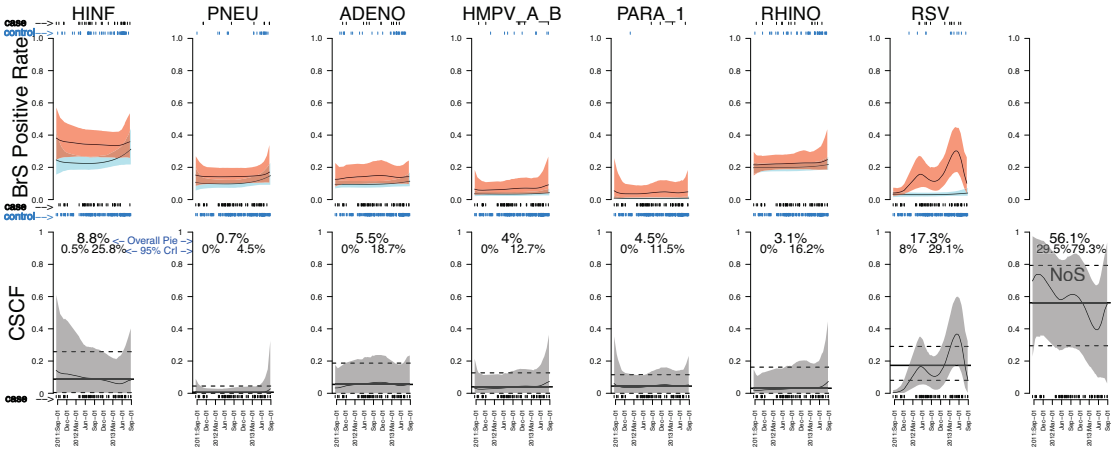


(b) Empirical coverage rates

Fig. 3: The regression analyses produce less biased posterior mean estimates and more valid empirical coverage rates for π_ℓ^* in Simulation II. Each panel corresponds to one of 16 combinations of true parameter values and sample sizes. *Top*) Each boxplot (left: regression; right: no regression) shows the distribution of the percent relative bias of the posterior mean in estimating the overall CSCF π_ℓ^* for six causes (A - F); "- - -" indicates zero bias. *Bottom*) The empirical coverage rates of the 95% CrIs with regression (●) or without regression (▲); "- - -" indicates the nominal 95% level. Since each coverage rate for π_ℓ^* is computed from $R = 200$ binary observations of the true π_ℓ^{0*} being covered or not, a 95% CI is also shown.



(a) **Age** $\leqslant 1$ **year**, HIV negative, severe pneumonia



(b) **Age** > 1 **year**, HIV negative, severe pneumonia

Fig. 4: Estimated seasonal CSCF for two most prevalent age-HIV-severity case strata under single-pathogen causes (HINF, PNEU, ADENO, HMPVAB, PARA1, RHINO, RSV) and a "Not Specified" cause. In each age-HIV-severity case stratum and for each cause ℓ:
Row 2): Temporal trend $\widehat{\pi}_\ell(\mathsf{t}; \mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ enveloped by pointwise 95% CrIs (gray). The horizontal solid line shows the estimated overall CSCF $\widehat{\pi}^*_\ell(\mathsf{t}; \mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ averaged over cases in the present stratum (dashed black lines: 95% CrI). The rug plot on the x-axis indicates cases' enrollment dates.
Row 1) shows the fitted temporal case (red) and control (blue) positive rate curves enclosed by the pointwise 95% CrIs; The two rug plots at the top (bottom) indicate the dates of the cases and controls being enrolled and tested positive (negative) for the pathogen.

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 observed marginal rates are shown at the bottom. Enrollment date (t) is not stratified upon here.

age $\geqslant 1$	HIV positive	# controls (%) total: 964 (100)	very severe (VS) (case-only)	# cases (%) total: 518 (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%	17.2%		34.7%	
control: 34.3%	14.1%		-	