Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

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

Zhenke Wu*[1,2] | Irena Chen[1]

[1]Department of Biostatistics, University of Michigan, Michigan, USA
[2]Michigan Institute for Data Science, University of Michigan, Michigan, USA

**Correspondence**
*Zhenke Wu, 1415 Washington Heights, Ann Arbor, MI48109, USA. Email: zhenkewu@umich.edu

Funding Information
This research was supported by the Patient-Centered Outcomes Research Institute (PCORI), Grant/Award Number: ME-1408-20318; National Institutes of Health grants, Grant/Award Number P30CA046592, U01CA229437

**Abstract**

Optimal prevention and treatment strategies for a disease of multiple causes, such as pneumonia, must be informed by the population distribution of causes among cases, or cause-specific case fractions (CSCFs). CSCFs may further depend on additional explanatory variables. Existing methodological literature in disease etiology research does not fully address the regression problem, particularly under a case-control design. Based on multivariate binary non-gold-standard diagnostic data and additional covariate information, this paper proposes a novel and unified regression modeling framework for estimating covariate-dependent CSCF functions in case-control disease etiology studies. 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 analyses that omit covariates. A regression analysis of pediatric pneumonia data reveals the dependence of CSCFs upon season, age, HIV status and disease severity. The paper concludes with a brief discussion on model extensions that may further enhance the utility of the regression model in disease etiology research.

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

## 1 | INTRODUCTION

### 1.1 | Motivating Application

Pneumonia is a clinical condition associated with infection of the lung tissue, which can be caused by more than 30 different species of pathogens. In studies of pneumonia etiology, a cause is the subset of one or more pathogens infecting the lung. Knowledge about population-level cause-specific etiologic contributions can help prioritize prevention programs and design treatment algorithms. 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.[1] The primary aim of the study is to estimate the etiologic contributions quantified by *cause-specific case fractions* (CSCFs), which may vary by individual-level factors such as age, enrollment date, disease severity, nutrition status and human immunodeficiency virus (HIV) status. We will refer to the covariate-dependent CSCFs as *CSCF functions*.

### 1.2 | Inferential Challenges: Imperfect Measurements and Additional Control Data

The fundamental challenge to estimating CSCFs is due to the unobservable true causes of disease among cases defined by clinical (non-microbiological) criteria. In PERCH study, tabulating case frequencies by cause is infeasible, because the lung-infecting pathogen(s) can rarely be directly observed due to potential clinical complications associated with invasive lung aspiration procedure.[2] Alternatively, non-invasive real-time polymerase chain reaction (PCR) test was made on each case's nasopharyngeal (NP) specimen (referred to as NPPCR), outputting presence or absence of a list of pathogens in the nasal cavity. The NP multivariate binary measurements are imprecise indicators for what infected the lung. In particular, detecting a pathogen in a case's nasal cavity does not indicate it caused lung infection. To provide statistical control for false positive detections, PERCH study performed NPPCR tests on pneumonia-free controls.

Valid inference must take into account two salient characteristics: imperfect sensitivities and specificities. First, tests lacking sensitivity may miss the true causative pathogen(s) which if unadjusted may produce inferior estimates of the CSCF functions. Second, imperfect diagnostic specificities may result in false positive detection of pathogens that are not causes of pneumonia. Control data in PERCH study provide the requisite information for estimating false positive rates and must be integrated.

### 1.3 | Primary Aim: Regression Modeling with the Goal of Estimating CSCF Functions

Motivated by the case-control, non-gold-standard multivariate binary diagnostic test data from the PERCH study, statistical models and inferential algorithms have been successfully developed to estimate CSCFs without covariates.[3,4] However, when individual-level explanatory variables are also available, the question of how to estimate CSCF functions remains. Data of similar structure have been collected by other large-scale disease etiology studies,[5,6] raising acute needs for regression modeling.

The primary aim of our work is to leverage critical control diagnostic test data and covariate information when performing regression modeling of CSCFs among the cases, which is not studied before in the statistical literature. This is partly because such case-control studies of disease etiology using multiple diagnostic tests are completed only recently to provide new clinical and microbiological data to inform future prevention and treatment strategies. The models for no-covariate analysis are also developed recently, requiring extensions to regression settings to enable characterization of covariate-dependent disease etiology. See Section 7 for additional inferential challenges associated with small cell counts of cases and controls in covariate strata. To achieve this aim, we design a hierarchical Bayesian latent variable regression model for case-control multivariate binary responses and derive an efficient posterior inference algorithm. In addition, through extensive simulation studies, we demonstrate that our regression model produces more valid inference than competing case-control models that cannot use covariates.

The rest of the paper is organized as follows. Section 2 describes the data structure and introduces notations. Section 3 contextualizes our work by reviewing related literature. Section 4 reviews existing models without covariates. Section 5 formulates the proposed regression model, specifies priors, and derives the posterior sampling algorithm. We demonstrate the proposed method via extensive simulations in Section 6 and an application to PERCH data in Section 7. The paper concludes with a discussion on future research directions.

## 2 | DATA STRUCTURE AND NOTATIONS

Let $Y_i = 1$ indicate a case subject with the clinically-defined disease and $Y_i = 0$ indicate a control subject without disease. Let $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top \in \{0,1\}^J$ represent the multivariate binary case-control, non-gold-standard diagnostic test results from subject $i$. Let $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \dots, N\}$ represent data, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\top$ are the $p$ primary covariates in CSCF functions and hence must be available for cases, and $\boldsymbol{W}_i = (W_{i1}, \dots, W_{iq})^\top$ are $q$ covariates that are available in the cases and the controls. $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ may be identical, overlapping or completely different. $\boldsymbol{X}_i Y_i = \boldsymbol{X}_i$ for a case $Y_i = 1$; $\boldsymbol{X}_i Y_i$ is a vector of zeros for a control subject. For notational convenience, we have ordered the continuous variables, if any, in $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ as the first $p_1$ and $q_1$ elements, respectively. In this paper, we focus on pre-specified $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ and discuss the important problem of variable selection in Section 8.

# 3 | 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. In this paper, "cause" does not represent an intervention as in classical causal inference literature. Rather, it describes a specific mechanism leading up to the disease and follows the use in demography and disease etiology.[7,3,8]

*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[7,8] proposed an unsupervised, informative Bayes implementation of a latent class model ("LCMs"),[9] where the latent classes represent unobserved causes of death and the survey responses are mutually independent given each latent class. Moran et al.[10] let covariates influence the conditional distribution of the survey responses 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 et al.[11] recognized the variation of CSMFs by covariates and studied transfer learning from a source population to a target population with a few deaths with observed causes. Finally, VA data are by definition case-only, hence are not applicable in our case-control setting.

*Case-control Latent Variable Methods.* Methods that use *case-control* data to estimate CSCFs remain sparse. We review the only existing methods. Wu et al.[3] introduced a *partially-latent class model* (pLCM) as an extension to classical LCMs. In particular, the pLCM is a *semi-supervised* method: it assumes with probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_L)^\top$, or CSCFs, 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. Each causative pathogen is assumed to be observed with a higher probability (sensitivity, or true positive rate, TPR) in case class $\ell$ 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. Like classical LCMs, the pLCM assumes "local independence" (LI) which means the measurements $(M_{i1}, \dots, M_{iJ})$ 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)[4] to reduce the bias in CSCF estimation. The importance of modeling LD to reduce the biases in estimating the mixing weights and response probabilities has also been extensively studied before but in the context of classical LCMs.[12,13] Finally, the npLCM is partially-identified,[4] necessitating informative priors for a subset of parameters (TPRs), which in PERCH study are obtained from external vaccine probe studies.

# 4 | PRELIMINARIES: EXISTING MODELS WITHOUT COVARIATES

## 4.1 | Additional Notations for Unobserved Causes-of-disease: "Latent States"

We first introduce notations for the unobserved causes of disease for the case subjects. Suppose a total of $J$ "agents" or "items" are measured by the diagnostic tests. Let a binary variable $\iota_{ij}$ indicate whether or not the $j$-th agent caused case $i$'s disease, that is, $\iota_{ij} = 1$ if $j$-th agent caused the disease, $\iota_{ij} = 0$ otherwise. We also allow more than one agent to cause the disease. We therefore have $\boldsymbol{\iota}_i = (\iota_{i1}, \dots, \iota_{iJ})^\top \in \{0, 1\}^J$ which is a vector of multiple binary indicators that represent the causes for subject $i$. We will also refer to $\boldsymbol{\iota}_i$ as "latent states" for case subject $i$. Note that we allow the all-zero latent states $\boldsymbol{\iota}_i = \mathbf{0}_{J\times 1}$. This is to represent a case subject with a "Not Specified" (NoS) cause which, in PERCH study, represents the subgroup of cases whose diseases are caused by agents not specified as molecular targets in the diagnostic PCR tests. Cases can thus be classfied by distinct multivariate binary patterns of $\boldsymbol{\iota}_i$. We will refer to cases having the same pattern of $\boldsymbol{\iota}_i$ as belonging to the same disease class.

In this paper, we assume that there are $L$ classes of *pre-specified* latent state patterns (non-zero or all-zero) among the case subjects. Let a set $\mathcal{A}$ comprise the pre-specified $L$ distinct multivariate binary patterns so that $|\mathcal{A}| = L$, where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. We discuss the important problem of unknown subsets in Section 8. We then introduce disease class indicators by arbitrarily labeling elements in $\mathcal{A}$ from 1 to $L$. We can now use $I_i$ that takes its value from $\{1, \dots, L\}$ to indicate case subject $i$'s class. We also let $\mathcal{C}_\ell = \{j : \iota_{ij} = 1, I_i = \ell, j = 1, \dots, J\}$ represent the subset of causative agents for disease class $\ell$; for NoS class, we have $\mathcal{C}_{\mathsf{NoS}} = \emptyset$.

For a control subject $i'$ without disease, we use $I_{i'} = 0$ to indicate $\boldsymbol{\iota}_{i'} = \mathbf{0}_{J\times 1}$ which means no measured pathogen is a cause. For a case or a control subject, the value of $I_i$ thus corresponds to a particular latent state pattern; we denote this correspondence by $\boldsymbol{\iota}_i = \boldsymbol{\iota}_i(I_i)$ which will be handy for specifying the models.

To illustrate the scientific meaning of the notations, consider a hypothetical list of $J = 5$ species of pathogens that are targeted by the NPPCR tests and possible disease-causing agents in the context of PERCH study. First, under an assumption of single-pathogen causes and no NoS class, we have $L = J = 5$ disease classes with distinct patterns of $\boldsymbol{\iota}$: $\mathcal{A} = \{(1,0,0,0,0)^\top, (0,1,0,0,0)^\top, \cdots, (0,0,0,0,1)^\top\}$. We can label the five disease classes by $1,\ldots,L = 5$, so that, for example, $I_i = 2$ corresponds to $\boldsymbol{\iota}_i = (0,1,0,0,0)^\top$ and $\mathcal{C}_2 = \{2\}$, $I_i = 5$ corresponds to $\boldsymbol{\iota}_i = (0,0,0,0,1)^\top$ and $\mathcal{C}_5 = \{5\}$. Second, under a less restrictive assumption of single- or double-pathogen causes (still no NoS class), we have $L = \binom{J}{1} + \binom{J}{2} = 5 + 10 = 15$ disease classes. For example, cases with the first and third pathogen infecting the lung are represented by $\boldsymbol{\iota}_i = (1,0,1,0,0)^\top$ which has the subset of causes $\mathcal{C}_\ell = \{1,3\}$ where $I_i = \ell$ is an arbitrary integer label of the disease class.

We first briefly review existing models without covariates, because they are the basis for our proposed regression model. We will describe the likelihood functions via generative processes, first for control data and then for case data. We then explain the rationales along with each step to provide scientific motivations for the model structures.

## 4.2 | PLCM: No covariate, conditional independence

We first review partially-latent class models (pLCM)[3], which assumes that a subset of subjects have observed states. In PERCH, this means assuming the control subjects have no pathogen infecting the lungs. Although a control subject has no disease, positive responses among $\boldsymbol{M}_i$ may result from imperfect specificities of the diagnostic tests. In particular, the data generating process for control data is from $J$ independent Bernoulli trials with distinct success probabilities:

$$\text{control data}: \quad M_{ij} \mid \iota_{ij} = 0 \sim \text{Bern}\left\{\psi_j\right\}, \text{ independently for item } j = 1, ..., J, \tag{1}$$

where the parameters $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_J)^\top$ represent the positive response probabilities absent disease, referred to as "false positive rates" (FPRs), or 1-specificity. The data generating process for cases is an $L$-component finite mixture model. In the following, Step (2) generates a disease class indicator for a case subject that takes value from $\{1,\ldots,L\}$. The lung-infecting pathogens are represented by $\boldsymbol{\iota}_i$, which is found by the correspondence between the disease class indicator $I_i$ and $\boldsymbol{\iota}_i$. Given $\boldsymbol{\iota}_i$, Step (3) generates measurements of $\boldsymbol{\iota}_i$ resulting in error-prone multiple responses $\{M_{i1},\ldots,M_{iJ}\}$ with positive response probabilities $p_\ell^{(j)}$ that takes the value of $\theta_j$ or $\psi_j$ according as whether the $j$-th agent caused the disease. We have

$$\text{disease class}: \quad I_i \mid Y_i = 1 \sim \text{Categorical}_L\left\{\boldsymbol{\pi}\right\}, \boldsymbol{\pi} \in \mathcal{S}_{L-1}, \text{ where} \tag{2}$$

$$\mathcal{S}_{L-1} = \{\boldsymbol{r} \in [0,1]^L : \sum_{\ell=1}^{L} r_\ell = 1\} \text{ is the probability simplex;}$$

$$\text{convert class to causes}: \quad \boldsymbol{\iota}_i = \boldsymbol{\iota}_i(I_i) \in \mathcal{A};$$

$$\text{case data}: \quad M_{ij} \mid \iota_{ij}, I_i = \ell \sim \text{Bern}\left\{p_\ell^{(j)}\right\}, \text{ independently for item } j = 1, ..., J, \text{ where} \tag{3}$$

$$\text{response probabilities}: \quad p_\ell^{(j)} = \begin{cases} \theta_j, & \iota_{ij} = 1; \\ \psi_j, & \iota_{ij} = 0, \end{cases} \quad \ell = 1,\ldots,L. \tag{4}$$

We term the parameter $\theta_j$ as "true positive rate" (TPR), or sensitivity. It is assumed to be larger than the FPR $\psi_j$ ($\theta_j > \psi_j$). For the case model, the pLCM makes a key "non-interference" assumption in (4) that disease-causing pathogen(s) are more frequently detected among cases than controls and the non-causative pathogens are observed with the same rates among cases as in controls[4]. Under the single-pathogen-cause assumption, pLCM uses $J$ TPRs $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_J)^\top$ and $J$ FPRs $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_J)^\top$. Note that (1) and (3) assume conditional independence of the measurements given the lung status $I_i$; this is referred to as local independence (LI).

## 4.3 | Nested PLCM: No covariate, conditional dependence

The proposed regression model in this paper will build upon an extension to pLCM, referred to as "nested pLCM".[4] We therefore provide a review of the npLCM model structure and rationale. The npLCM is designed to reduce potential estimation bias in $\boldsymbol{\pi}$ under large degrees of deviations from LI assumed in the original pLCM. It achieves the aim by characterizing residual correlations among $J$ binary measurements $\boldsymbol{M}_i$ even after conditioning on $I_i$, e.g., the controls ($I_i = 0$) or a case class ($I_i = \ell$,

$\ell \neq 0$).[4] The extension is motivated by the flexibility of the classical LCM formulation[9] to approximate any joint distribution of multivariate discrete responses.[14]

Compared to the pLCM, the data generating process for control data now assumes an additional step of generating "subclasses". Scientifically, the subclasses in the controls (and cases in the following) can represent subgroups of children with different levels of immunity with positive rates varying across subclasses. Importantly, the conditional dependence of $[\boldsymbol{M}_i \mid I_i]$ arises after the subclass indicators are integrated out. More specifically, we assume

$$\text{control subclass}: \quad Z_i \mid Y_i = 0 \sim \text{Categorical}_K\{\boldsymbol{\nu}\}, \boldsymbol{\nu} \in \mathcal{S}_{K-1}, \tag{5}$$

$$\text{control data}: \quad M_{ij} \mid \iota_{ij} = 0, Z_i = k \sim \text{Bern}\left\{\psi_k^{(j)}\right\}, \text{ independently for item } j = 1, ..., J, \tag{6}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^\top$ is the vector of subclass probabilities and lies in a probability simplex. The original pLCM results if $K = 1$. Let $\boldsymbol{\Psi} = \{\psi_k^{(j)} \in (0,1)\}$ be a $J \times K$ matrix comprising FPRs, which are necessary for modeling the imperfect binary measurements among the controls. Let $\boldsymbol{\psi}^{(j)}$ and $\boldsymbol{\psi}_k$ represent the $j$-th row and $k$-th column. The data generating process for cases is as follows, with an additional Step (8) for drawing a subclass indicator $Z_i$ for each case subject:

$$\text{disease class}: \quad I_i \mid Y_i = 1 \sim \text{Categorical}_L\{\boldsymbol{\pi}\}, \boldsymbol{\pi} \in \mathcal{S}_{L-1}, \tag{7}$$

$$\text{case subclass}: \quad Z_i \mid Y_i = 1 \sim \text{Categorical}_K\{\boldsymbol{\eta}\}, \boldsymbol{\eta} \in \mathcal{S}_{K-1}, \tag{8}$$

$$\text{convert class to causes}: \quad \boldsymbol{\iota}_i = \boldsymbol{\iota}_i(I_i) \in \mathcal{A};$$

$$\text{case data}: \quad M_{ij} \mid \iota_{ij}, Z_i = k, I_i = \ell, \sim \text{Bern}\left\{p_{k\ell}^{(j)}\right\}, \text{ independently for item } j = 1, ..., J, \tag{9}$$

$$\text{response probabilities}: \quad p_{k\ell}^{(j)} = \begin{cases} \theta_k^{(j)}, & \iota_{ij} = 1; \\ \psi_k^{(j)}, & \iota_{ij} = 0, \end{cases} \quad k = 1, \ldots, K, \text{ and } \ell = 1, \ldots, L. \tag{10}$$

At Step (8), the npLCM introduces $K$ unobserved subclasses with weights $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^\top$. The weights are shared across $L$ disease classes. 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. We also refer to $\theta_k^{(j)}$ as TPR or sensitivity as in pLCM. Let $\boldsymbol{\theta}^{(j)}$ and $\boldsymbol{\theta}_k$ represent the $j$-th row and $k$-th column. In Step (10), $p_{k\ell}^{(j)}$ represents the positive response probability of $M_{ij}$ in subclass $k$ of disease class $\ell$, which equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise. We collect all the positive response probabilities for subclass $k$ in disease class $\ell$ into $\boldsymbol{p}_{k\ell} = (p_{k\ell}^{(1)}, \ldots, p_{k\ell}^{(J)})^\top$.

### 4.3.1 | Subclasses: Nuisance Parameters for Characterizing Residual Correlations Given $I_i$

Different from what are of primary interest (the disease class indicators $I_i$ and its population distribution CSCFs $\boldsymbol{\pi}$), the latent subclasses $Z_i$ nested in each class $\ell = 0, 1, ..., L$ are *nuisance* parameters introduced to approximate complex multivariate dependence among discrete data. Importantly, the conditional dependence of $[\boldsymbol{M}_i \mid I_i]$ arises after the subclass indicators are integrated out with respect to subclass weights ($\boldsymbol{\nu}$ or $\boldsymbol{\eta}$). In particular, Wu et al.[4] used truncated stick-breaking priors for the subclass weights to encourage few subclasses and to side-step the choice of the true number of subclasses. However, nuisance parameter does not mean it is not important for improving the inference about the primary substantive parameters - the CSCFs. To the contrary, Wu et al.[4] showed via asymptotic bias calculations and simulations that the use of subclasses reduces the estimation bias of CSCFs in the context without covariates and improves the coverage of the credible intervals.

Scientifically, the subclasses in the cases and controls can represent subgroups of children with different levels of immunity with positive rates $\boldsymbol{\theta}_k$ and $\boldsymbol{\psi}_k$ varying across subclasses.

*Remark 1.* Similar to the pLCM, the FPRs $\boldsymbol{\Psi}$ in the npLCM are shared between the controls and the case classes for non-causative agents (via $p_{kl}^{(j)}$). Different from the pLCM, the subclass mixing weights may differ between the cases ($\boldsymbol{\eta}$) and the controls ($\boldsymbol{\nu}$). The special case of $\eta_k = \nu_k, k = 1, \ldots, K$, means the covariation patterns among the non-causative pathogens in a disease class is not different from the controls. However, relative to the controls, cases may have different strength and direction of observed pairwise measurement dependence in each disease class. By allowing the subclass weights to differ between the cases and the controls, npLCM is more flexible than pLCM in referencing cases' measurements against the controls.

# 5 | PROPOSED REGRESSION EXTENSION

## 5.1 | Overview

We now describe the likelihood of the proposed regression model by a generative process building on npLCM. Only Steps (11), (12) and (13) below are results of the proposed extension; other parts of npLCM remains the same.

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. The proposed model assumes the control subclass weights are covariate-dependent:

$$\text{Extend (5)} - \text{control subclass}: \quad Z_i \mid \boldsymbol{W}_i, Y_i = 0 \sim \text{Categorical}_K\left\{\boldsymbol{\nu}_i\right\}, \boldsymbol{\nu}_i = \boldsymbol{\nu}(\boldsymbol{W}_i) \in \mathcal{S}_{K-1}, \tag{11}$$

where, as in npLCM[4], the subclass indicators $Z_i$'s are nuisance quantities for inducing dependence among the multivariate binary responses $\boldsymbol{M}_i$, but now given covariates. $\boldsymbol{\nu}_i = \left(\nu_{i1}, \dots, \nu_{iK}\right)^\top$ is the vector of control subclass probabilities that now may depend on $\boldsymbol{W}_i$. Scientifically, we are not interested in how the subclass probabilities are associated with covariates. We introduce $\boldsymbol{\nu}(\boldsymbol{W})$ here because, upon integrating over the distribution of $Z_i$ in (11), it helps define a flexible conditional distribution of $\boldsymbol{M}_i$ given covariates $\boldsymbol{W}_i$.

FPR subclass profile $k$, $\boldsymbol{\psi}_k = (\psi_k^{(1)}, \dots, \psi_k^{(J)})^\top$, receives a weight of $\nu_k(\boldsymbol{W}_i)$ for a control subject $i$ with covariates $\boldsymbol{W}_i$. After integrating out $Z_i$, we obtain *marginal* control FPRs: $\mathbb{P}(M_{ij} = 1 \mid \boldsymbol{W}_i, Y_i = 0, \boldsymbol{\Psi}) = \boldsymbol{\nu}(\boldsymbol{W}_i)^\top \boldsymbol{\psi}^{(j)}$, $j = 1, \dots, J$, which depend on $\boldsymbol{W}_i$. For example, this enables us to use control data to estimate seasonal marginal FPRs of pathogen A. During seasons with frequent asymptomatic carriage among the controls, we may estimate the FPRs to be high. This means that presence of pathogen A in a case's nasal cavity (the body site where the diagnostic tests in PERCH study were done) during these seasons does not necessarily indicate etiologic importance.

In Section 5.4.1, 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 that is appropriate for specific applications.

A reviewer raised the question of why formulating the model with subclass weights but not the TPRs and FPRs depending on covariates $\boldsymbol{W}_i$. Our modeling choice is primarily driven by observed conditional dependence among the multivariate binary measurements even after conditional on the covariates, which in the context of PERCH study means that diagnostic measurements of pathogens still exhibit mutual inhibition or stimulation even after stratifying on individual-level covariates, e.g, age. If we directly regressed TPRs and FPRs on covariates without introducing subclasses, a more restrictive assumption of independent measurements given covariates in each disease class would result. See Appendix A1 in the Supplementary Materials for further remarks on the control model assumption and the use of subclasses in this paper. In addition, the biological motivation for introducing subclasses is to characterize pathogen mutual inhibitions or stimulations by distinct the response profiles in the subclasses. The model structure based on subclasses is also amenable to posterior computations via data augmentation.

For cases, we follow the case model for npLCM, but extend in two aspects: let CSCFs depend on covariates $\boldsymbol{X}_i$ and let case subclass weight depend on covariates $\boldsymbol{W}_i$. That is,

$$\text{Extend (7)} - \text{disease class}: \quad I_i \mid \boldsymbol{X}_i, Y_i = 1 \sim \text{Categorical}_L\left\{\boldsymbol{\pi}_i\right\}, \boldsymbol{\pi}_i = \boldsymbol{\pi}(\boldsymbol{X}_i) \in \mathcal{S}_{L-1}, \tag{12}$$

$$\text{Extend (8)} - \text{case subclass}: \quad Z_i \mid \boldsymbol{W}_i, Y_i = 1 \sim \text{Categorical}_K\left\{\boldsymbol{\eta}_i\right\}, \boldsymbol{\eta}_i = \boldsymbol{\eta}(\boldsymbol{W}_i) \in \mathcal{S}_{K-1}, \tag{13}$$

where $\boldsymbol{\pi}(\boldsymbol{X}_i) = (\pi_1(\boldsymbol{X}_i), \dots, \pi_L(\boldsymbol{X}_i))^\top$ are CSCF functions evaluated at $\boldsymbol{X}_i$, and $\boldsymbol{\eta}(\boldsymbol{W}_i) = \left(\eta_{i1}(\boldsymbol{W}_i), \dots, \eta_{iK}(\boldsymbol{W}_i)\right)^\top$ is the vector of case subclass probabilities evaluated at $\boldsymbol{W}_i$. Both $\boldsymbol{\pi}_i$ and $\boldsymbol{\eta}_i$ are quantities from probability simplexes.

In this paper, we have adopted the approach of specifying an upper bound of working number of subclasses, and using sparsity priors to encourage towards a few subclasses for fitting the data. In Simulation I in Section 6, we illustrate a scenario where in truth two subclasses are present. During the model fitting, agnostic to the truth, we specify an upper bound of 7. Figure S1 in the Supplementary Materials shows that the model can learn from the data and figure out that two subclasses are actually needed. If the working number of subclasses $K$ is too small, the model may not be flexible enough to characterize the dependence among the measurements. We therefore recommend starting with a number, e.g., $5 - 10$, and then if computational resource is ample, trying larger $K$s. We usually observe that results remained similar for larger $K$. We can then report the final result based on a small $K$ that gives stable results.

## 5.2 | Targets of Inference

In this paper, CSCF functions $\boldsymbol{\pi}(\boldsymbol{X})$ is the primary target of inference. In addition, let $\boldsymbol{X} = (\boldsymbol{X}_a^\top, \boldsymbol{X}_b^\top)^\top$, where $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ are *pre-specified* partition of $\boldsymbol{X}$. The partition is a generic notation, and is useful to represent, e.g., $\boldsymbol{X}_b = \{\mathtt{I}(\mathsf{age} > 1), \mathtt{I}(\mathsf{HIV\ positive}), \mathtt{I}(\mathsf{very\ severe\ pneumonia})\}$ and $X_a$ can be enrollment date. It is also of policy interest to infer the overall CSCFs given a specific age-HIV-severity stratum, while integrating over the study period.[1] We define the *overall* CSCFs in a stratum $\boldsymbol{X}_b = \boldsymbol{x}_b$ by integrating CSCF functions over the subset of covariates $\boldsymbol{X}_a$ while holding $\boldsymbol{X}_b$ fixed at specified covariate values $\boldsymbol{x}_b$:

$$\pi_\ell^*(\boldsymbol{X}_b = \boldsymbol{x}_b) = \int \pi_\ell(\boldsymbol{X}_a, \boldsymbol{X}_b = \boldsymbol{x}_b)\mathrm{d}G(\boldsymbol{X}_a), \ell = 1 \ldots, L, \tag{14}$$

where $G$ is a probability or empirical distribution of $[\boldsymbol{X}_a \mid \boldsymbol{X}_b]$. Here $[A \mid B]$ represents the conditional distribution of random quantity $A$ given another random quantity $B$. When $B = \emptyset$, it represents the distribution of $A$. We simply write $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ when $\boldsymbol{X}_b = \emptyset$. Following the definition, $\boldsymbol{\pi}^*\{\mathtt{I}(\mathsf{age} > 1) = 0, \mathtt{I}(\mathsf{HIV\ positive}) = 0, \mathtt{I}(\mathsf{very\ severe\ pneumonia}) = 0\}$ characterizes the overall CSCFs in the specific age-HIV-severity stratum. Inferences made about $\boldsymbol{\pi}(\boldsymbol{X})$ and $\boldsymbol{\pi}^*(\boldsymbol{X}_b)$ are demonstrated in Section 7.

## 5.3 | Detailed Regression Specifications

We first specify the functional forms of $\boldsymbol{\pi}(\cdot)$, $\boldsymbol{\nu}(\cdot)$, $\boldsymbol{\eta}(\cdot)$, based on which we complete the joint distribution specification by describing the priors in Section 5.4.

### 5.3.1 | Primary Component: CSCF Regression Model

We assume the CSCFs depend on $\boldsymbol{X}_i$ via a classical multinomial logistic regression model:

$$\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\} / \sum_{\ell'=1}^{L} \exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}, \ell = 1, ..., L, \tag{15}$$

where $\phi_\ell(\boldsymbol{X}_i) - \phi_L(\boldsymbol{X}_i)$ is the log odds of case $i$ in disease class $\ell$ relative to $L$: $\log \pi_{i\ell}/\pi_{iL}$. The choice of not fixing a baseline category with zero regression coefficients is based on convenient prior specification so that all categories are treated symmetrically. Although the actual values of regression coefficients are not identifiable, contrasts between parameters in distinct categories remain identifiable, which are then summarized in the posterior inference. We further assume additivity in a partially linear model:

$$\phi_\ell(\boldsymbol{x}; \Gamma_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi, \tag{16}$$

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, and $\Gamma_\ell^\pi = [(\boldsymbol{\beta}_{\ell 1}^\pi)^\top, \ldots, (\boldsymbol{\beta}_{\ell p_1}^\pi)^\top, (\boldsymbol{\gamma}_\ell^\pi)^\top]^\top$ is the vector of the regression coefficients for disease class $\ell$. Section 8 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 5.4.2, we approximate unknown functions of a standardized continuous variable such as $f_{\ell j}^\pi$ via basis expansions and along with a prior on the basis coefficients to encourage smoothness.

Integrating over $L$ unobserved disease classes and $K$ subclasses in (13-12), we obtain the likelihood for cases:

$$L_1^{\mathsf{reg}} = \prod_{i:Y_i=1} \left\{ \sum_{\ell=1}^{L} \left[ \pi_\ell(\boldsymbol{X}_i; \Gamma_\ell^\pi) \sum_{k=1}^{K} \left\{ \eta_{ik} \cdot \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell}) \right\} \right] \right\}, \tag{17}$$

where $\Pi(\boldsymbol{m}; \boldsymbol{s}) = \prod_{j=1}^{J} \{s_j\}^{m_{ij}} \{1 - s_j\}^{1-m_{ij}}$ 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 $\eta_{ik}$ by a logistic stick-breaking regression technique, which we first introduce in the control model.

### 5.3.2 | Nuisance Components: Covariate-dependent Reference Distribution

The control data serve as reference against which diagnostic test results obtained from the cases are compared when estimating cause-specific probabilities. The control model is a mixture model with covariate-dependent mixing weights $\boldsymbol{\nu}_i = \boldsymbol{\nu}(\boldsymbol{W}_i)$.

*Control subclass weight regression.* We specify $\nu_{ik}$ by logistic stick-breaking parameterization:

$$\nu_{ik} = g(\alpha^{\nu}_{ik}) \prod_{s<k} \{1 - g(\alpha^{\nu}_{is})\}, \text{ if } k < K, \text{ and } \prod_{s<k} \{1 - g(\alpha^{\nu}_{is})\} \text{ otherwise, where} \tag{18}$$

$$\alpha^{\nu}_{ik} = \alpha^{\nu}_{k}(\boldsymbol{W}_i = \boldsymbol{w}; \Gamma^{\nu}_{k}) = \mu_{k0} + \sum_{j=1}^{q_1} f^{\nu}_{kj}(w_j; \boldsymbol{\beta}^{\nu}_{kj}) + \widetilde{\boldsymbol{w}}^{\top} \boldsymbol{\gamma}^{\nu}_{k}, \text{ for } k = 1, \ldots, K-1. \tag{19}$$

Let $\Gamma^{\nu}_{k} = [(\boldsymbol{\beta}^{\nu}_{k1})^{\top}, \ldots, (\boldsymbol{\beta}^{\nu}_{kq_1})^{\top}, (\boldsymbol{\gamma}^{\nu}_{k})^{\top}]^{\top}$ be the regression coefficients in the $k$-th subclass, and $\alpha^{\nu}_{ik}$ 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 (15) so that the priors of the coefficients $\Gamma^{\nu}_{k}$ and $\Gamma^{\pi}_{\ell}$ can be similar. In addition, the parameterization is amenable to simple and efficient block posterior sampling via Pólya-Gamma augmentation[15]. Generalization to other link functions such as the probit function is straightforward.[16] 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 $g(\alpha^{\nu}_{ik})$ from the remaining stick at step $k$, resulting in segment $k$ of length $\nu_{ik}$; this procedure is repeated for $k = 1, \ldots, K-1$.

It is readily derived that the likelihood for controls is: $L^{\text{reg}}_0 = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu_{ik} \Pi(\boldsymbol{M}_i; \boldsymbol{\psi}_k)$.

*Case subclass weight regression.* The case subclass weight curve $\boldsymbol{\eta}_k(\boldsymbol{W})$ is also specified via a logistic stick-breaking regression as in the controls but with different linear predictors $\alpha^{\eta}_{ik}$: $\eta_{ik} = g(\alpha^{\eta}_{ik}) \prod_{s<k} \{1 - g(\alpha^{\eta}_{is})\}$, $\forall k = 1, \ldots, K-1$; $\eta_{iK} = \prod_{s<K} \{1 - g(\alpha^{\eta}_{is})\}$. Given $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$, $\boldsymbol{\eta}_k(\boldsymbol{W})$ fully determines the joint distribution $[\boldsymbol{M} \mid \boldsymbol{W}, I = \ell \neq 0, \boldsymbol{\Theta}, \boldsymbol{\Psi}]$. We do not assume $\eta_k(\boldsymbol{w}) = \nu_k(\boldsymbol{w}), \forall \boldsymbol{w}$. Consequently, relative to the controls, the individuals in disease class $\ell$ may have different strength and direction of observed dependence between the causative $\{M_j : j \in C_\ell\}$ and non-causative $\{M_j : j \notin C_\ell\}$ pathogens, or between the non-causative pathogens. Let the $k$-th linear predictor

$$\alpha^{\eta}_{ik} = \alpha^{\eta}_{k}(\boldsymbol{W}_i = \boldsymbol{w}; \Gamma^{\eta}_{k}) = \mu_{k0} + \sum_{j=1}^{q_1} f^{\eta}_{kj}(w_j; \boldsymbol{\beta}^{\eta}_{kj}) + \widetilde{\boldsymbol{w}}^{\top} \boldsymbol{\gamma}^{\eta}_{k}, \tag{20}$$

where $f^{\eta}_{kj}$ and $f^{\nu}_{kj}$ (from the control model) share the basis functions but the regression coefficients $\Gamma^{\eta}_{k} = [(\boldsymbol{\beta}^{\eta}_{k1})^{\top}, \ldots, (\boldsymbol{\beta}^{\eta}_{kq_1})^{\top}, (\boldsymbol{\gamma}^{\eta}_{k})^{\top}]^{\top}$ differ from the control counterpart ($\Gamma^{\nu}_{k}$). In addition, we have used the same intercepts $\{\mu_{k0}\}$ in (19) to ensure only important subclasses in the controls are used in the cases. For example, absent covariates $\boldsymbol{W}$, a large and positive $\mu_{k0}$ effectively halts the stick-breaking procedure at step $k$ for the controls. This is because the $k$-th stick-breaking will take almost the entire remaining stick, resulting in $\nu_{k+1}$ that is approximately zero. Applying the same intercept $\mu_{k0}$ to the cases makes $\eta_{k+1} \approx 0$.

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

*Remark 2.* Sometimes ignoring $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ does not invalidate inference of the overall CSCFs $\boldsymbol{\pi}^*$. For example, assuming covariate-independence: $\forall k, \eta_k(\cdot) \equiv \eta_k$, $L^{\text{reg}}_1$ and $L^{\text{reg}}_0$ 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 distribution of $\boldsymbol{W}$. The integrated likelihood $L^*_1 L^*_0$ is now an npLCM likelihood without covariates. In addition, it can be readily shown that no-covariate inference for $\boldsymbol{\pi}^*$ is also valid when $\boldsymbol{X}$ and $\boldsymbol{W}$ do not share common elements.

## 5.4 | Prior Distribution

The number of parameters in $L^{\text{reg}}(\{\Gamma^{\pi}_{\ell}\}, \{\Gamma^{\eta}_{k}\}, \{\Gamma^{\nu}_{k}\}, \{\mu_{k0}\}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$ is $\mathcal{O}(LC_{\max}p_1 + KC_{\max}q_1 + JK)$ 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 uniformly over $\boldsymbol{W}_i$ values, and (b) constant subclass weights over $\boldsymbol{W}_i$ values $\eta_k(\cdot) = \eta_k$ and $\nu_k(\cdot) = \nu_k$.

### 5.4.1 | Priors: Encourage Few Subclasses

Based on the parameterization of $\boldsymbol{\nu}(\boldsymbol{W})$ in (18) and (19), we propose a novel prior for $\boldsymbol{\nu}(\boldsymbol{W})$ over a probability simplex $\mathcal{S}_{K-1}$. Because the linear predictor $\alpha^{\nu}_{ik}$ depends on $\mu_{k0}$ and other regression coefficients, priors on $\mu_{k0}$ and the coefficients induce a prior distribution for $g(\alpha^{\nu}_{ik})$ and thus a prior for $\nu(\boldsymbol{W})$ according to (18). For example, consider $g(\alpha^{\nu}_{ik}) = g(\mu_{k0} + \gamma^{\nu}_{k1})$, for subclass denoted as $k = A, B$. The prior for $\mu_{k0}$ and a Gaussian prior for $\gamma^{\nu}_{k1}$ will induce a prior for $g(\alpha^{\nu}_{ik})$ and thus for $(\nu_A, \nu_B, 1-\nu_A-\nu_B)^{\top}$.

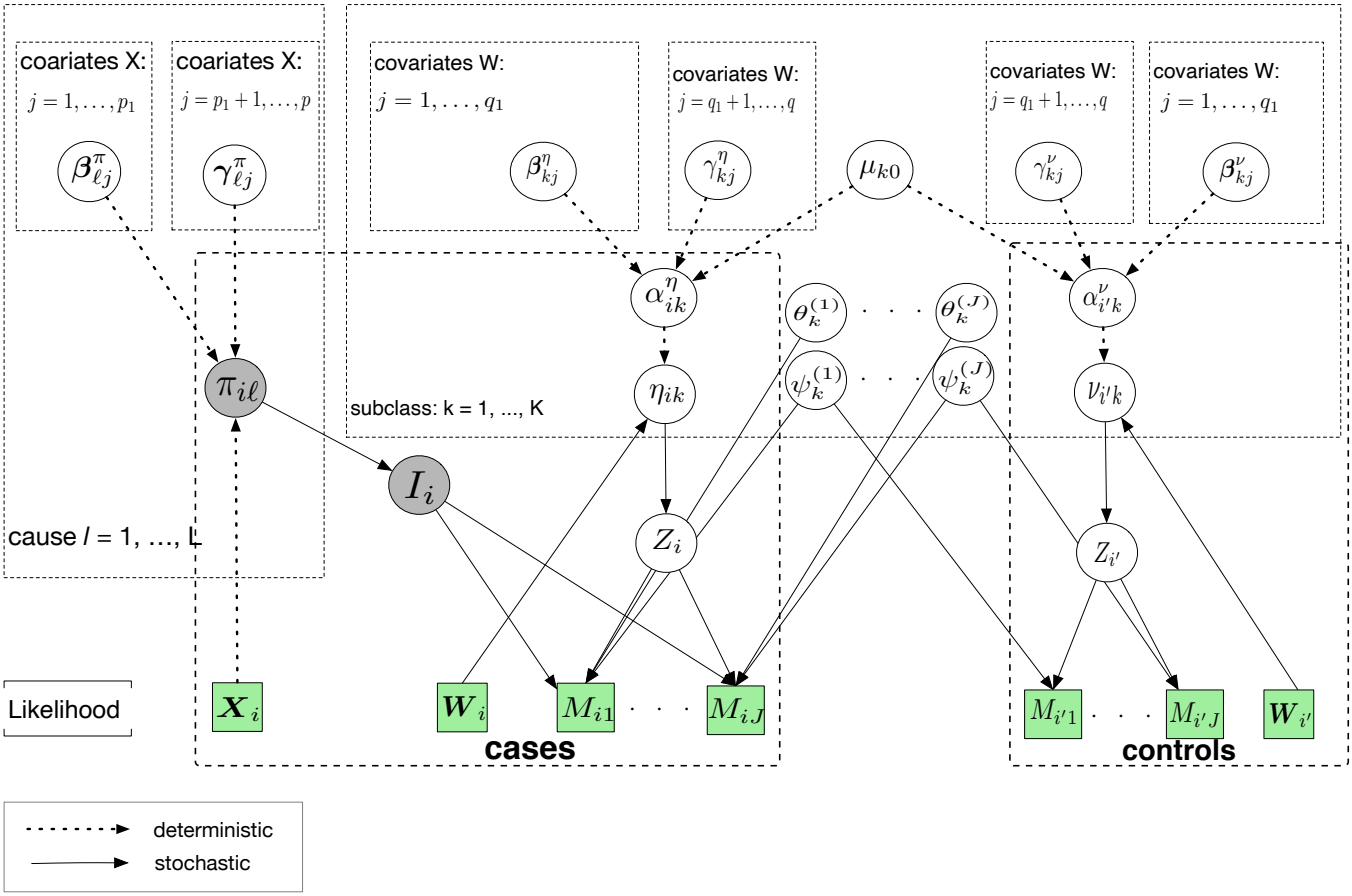



**FIGURE 1** The directed acyclic graph (DAG) representing the structure of the model likelihood. The quantities in squares are either data or hyperparameters; the unknown quantities are shown in the circles. The arrows connecting variables indicate that the parent parameterizes the distribution of the child node (solid lines) or completely determines the value of the child node (dotted arrows). The rectangular "plates" where the variables are enclosed indicate that a similar graphical structure is repeated over the index; The index in a plate indicate subjects, causes, covariates or subclasses. Figure S6 in the Supplementary Materials presents the complete DAG with prior specification.

Specifically, our basic idea is to have one of $\{g(\alpha^{\nu}_{ik})\}_{k=1}^{K-1}$ in (18) 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 (19):

$$\mu_{k0} = \sum_{m=1}^{k} u_{km}\mu^*_{k0},\ u_{km} \geq 0,\ \ \mu^*_{k0} \sim \mathcal{N}_+(0, \tau_{k0}),\ \ \tau_{k0} \sim \text{Gamma}(a_0, b_0), k = 1, \ldots, K-1. \tag{21}$$

where $\{u_{km}, m = 1, \ldots, k,$ and $k = 1, \ldots, K-1\}$ 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 $\mu$, precision $\tau$ 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 $\tau_{k0}$, $\mu^*_{k0}$ has a truncated scaled-$t$ distribution with degree of freedom $\nu_0$ and scale $s_0$, which peaks at zero and has a heavy right tail. A large positive value from the heavy-tailed prior for $\mu^*_{k0}$ at stick-breaking step $k$ produces a large $\mu_{k0}$ and takes away nearly the entire stick segment currently left. We will detail the Gaussian priors for other regression coefficients in the next two subsections.

For a fixed $\nu_0$, 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 2). Under the prior (21), it makes higher-order subclasses *a priori* increasingly unlikely to receive substantial weights. As a result, once applied to the control model (11), the prior encourages using a small number of subclasses to approximate the observed $2^J$ covariate-dependent probability contingency table for the control data $\{\boldsymbol{M}_i, \boldsymbol{W}_i : Y_i = 0\}$ in finite samples.

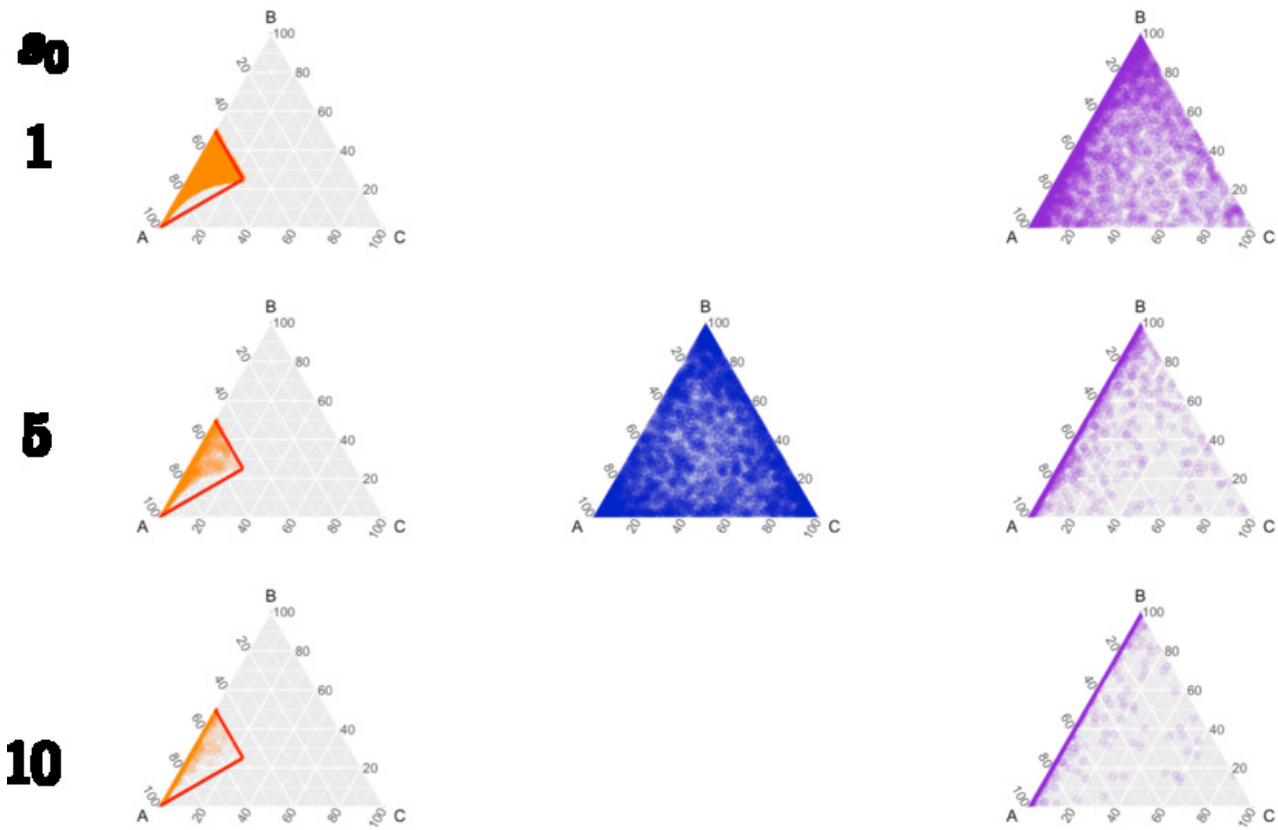



**FIGURE 2** Random draws from the prior for three subclass weights $(\nu_A, \nu_B, \nu_C)^\top$ shown in ternary diagrams. The scale parameter $s_0$ increases from 1 to 10 from the top to the bottom while fixing $\nu_0 = 1$. The random samples from the prior are increasingly concentrated towards the first subclass $A$. In each row, the three columns correspond to random samples when $\alpha^\nu_{ik}$ equals $\mu_{k0}$ in the left, $\gamma^\nu_{k1}$ in the middle, and $\mu_{k0} + \gamma^\nu_{k1}$ in the right. Here we show a single ternary diagram in the middle because the Gaussian prior of $\gamma^\nu_{k1}$ does not depend on $s_0$: prior means are $-1.07$ and 0 for $\gamma^\nu_{k1}$, $k = A, B$, respectively, and the prior precision is $\kappa_\gamma = 1/4$. These values were chosen to produce approximately evenly-distributed draws in a ternary diagram (middle).

### 5.4.2 | Priors: Encourage Constant Additive Regression Functions: $f^\pi_{\ell j}$, $f^\nu_{kj}$, $f^\eta_{kj}$

We use B-splines to approximate the additive functions of a standardized continuous variable[17]: $f^\nu_{kj}(\cdot) = \sum_{c=1}^{C_j} \beta^{(c),\nu}_{kj} B^{(c)}_j(\cdot)$, $f^\eta_{kj}(\cdot) = \sum_{c=1}^{C_j} \beta^{(c),\eta}_{kj} B^{(c)}_j(\cdot)$, where $\{B^{(c)}_j(\cdot) : c = 1, \dots, C_j\}$ are $C_j$ cubic B-spline bases that are shared between the cases and the controls for the $j$-th covariate $\boldsymbol{W}_j$ in subclass $k$. We assume distinct coefficients: $\boldsymbol{\beta}^\nu_{kj} = \{\beta^{(c),\nu}_{kj}, c \le C_j\}$ and $\boldsymbol{\beta}^\eta_{kj} = \{\beta^{(c),\eta}_{kj}, c \le C_j\}$. In addition, we assume $f^\pi_{\ell j}(\cdot) = \sum_{c=1}^{C^\pi_j} \beta^{(c),\pi}_{kj} B^{(c),\pi}_j(\cdot)$ where $\{B^{(c),\pi}_j(\cdot) : c = 1, \dots, C^\pi_j\}$ are also cubic B-spline bases and $\boldsymbol{\beta}^\pi_{\ell j} = \{\beta^{(c),\pi}_{\ell j}, c \le C^\pi_j\}$. With $M$ interior equally-spaced knots $\boldsymbol{\kappa} = (\kappa_0, \dots, \kappa_{M+1})^\top$. For example, for covariate $W_{ij}$ $\min_i(W_{ij}) = \kappa_0 < \kappa_1 < \cdots < \kappa_M < \kappa_{M+1} = \max_i(W_{ij})$, there are $M + 4$ basis functions. It readily extends to different numbers of basis functions. We further restrict $\{f^\pi_{\ell j}, j \le p_1\}$, $\{f^\nu_{kj}, f^\eta_{kj}, j \le q_1\}$ to have zero means for statistical identifiablity.

The choice of the number of bases (or, degrees of freedom) is crucial for the estimation of the CSCF functions and subclass weight regression functions. For example, larger values of $C_j$ and $C^\pi_j$ define a richer class of functions that can accommodate abrupt seasonal effect. In the context of PERCH study, more parsimonious models are preferred. Further improvements in knot selection is possible using knots on a nonequidistant grid so that more knots are placed where cases and controls are dense. Our practical suggestion is to place 5 to 10 equally spaced knots. A more computationally intensive approach is to add or delete knots using split-merge type of algorithms.[18]

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 $\pi, \nu, \eta$ and use "•" as a placeholder. We specify Gaussian random walk priors on the basis coefficients via Bayesian P-splines.[17] Let

$\boldsymbol{\beta}^{\bullet}_{kj} \mid \tau^{\bullet}_{kj} \sim \mathcal{N}(\mathbf{0}_{C^{\bullet}_j \times 1}, \tau^{\bullet}_{kj}\boldsymbol{K}^{\bullet})$, 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 $(C^{\bullet}_j - 1) \times C^{\bullet}_j$. It maps adjacent B-spline coefficients to $\beta^{(c),\bullet}_{kj} - \beta^{(c-1),\bullet}_{kj}$, $c = 2, \ldots, C^{\bullet}_j$. The precision matrix $\boldsymbol{K}^{\bullet}$ is not full rank and hence leaves the prior of $\beta^{(1),\bullet}_{kj}$ unspecified. We assume independent priors $\beta^{(1),\bullet}_{kj} \sim \mathcal{N}(0, k_{\beta})$.

Importantly, the $\tau^{\bullet}_{kj}$ is the smoothing parameters. Large values of $\tau^{\bullet}_{kj}$ lead to smoother fit of $f^{\bullet}_{kj}(\cdot)$ (constant when $\tau^{\bullet}_{kj} = \infty$). We specify a mixture prior for smoothing parameter $\tau^{\bullet}_{kj}$:

$$\tau^{\bullet}_{kj} \sim \xi^{\bullet}_{kj}\mathsf{Gamma}(a_{\tau}, b_{\tau}) + (1 - \xi^{\bullet}_{kj})\mathsf{IP}(a'_{\tau}, b'_{\tau}),\ \xi^{\bullet}_{kj} \sim \mathsf{Bern}(\rho^{\bullet}),\ \rho^{\bullet} \sim \mathsf{Beta}(a^{\bullet}_{\rho}, b^{\bullet}_{\rho}), \tag{22}$$

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, and $\xi^{\bullet}_{kj}$ is a binary indicator for the Gamma component. This bimodal mixture distribution creates a sharp separation between flexible and constant fits.[19,20]

### 5.4.3 | 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 $\nu_0$ and $s_0$ in (21), $(a_{\tau}, b_{\tau})$, $(a'_{\tau}, b'_{\tau})$, $(a^{\nu}_{\rho}, b^{\nu}_{\rho})$, $(a^{\eta}_{\rho}, b^{\eta}_{\rho})$ and $(a^{\pi}_{\rho}, b^{\pi}_{\rho})$ in (22), $\kappa_{\beta}$, and $\kappa_{\gamma}$. The npLCM regression model is partially-identified.[21] We assume independent $\theta^{(j)}_k \sim \mathsf{Beta}(a_j, b_j)$, $j \leq J$. In real data applications, hyperparameters $(a_j, b_j)$ are chosen so that the 2.5% and 97.5% quantiles match an elicited prior range.[22] For FPRs, we assume a flat prior $\psi^{(j)}_k \sim \mathsf{Beta}(1, 1)$ because they are empirically estimable from the control data.

## 5.5 | Posterior Inference and Software

Figure 1 summarizes the generative process above and the priors in a directed acyclic graph (DAG). Appendix A3 in the Supplementary Materials uses the DAG to derive the Markov chain Monte Carlo (MCMC) algorithm that draws posterior samples of the unknowns to approximate their joint posterior distribution.[23] 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).

# 6 | SIMULATIONS

We simulate case-control non-gold-standard diagnostic test data 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 $\boldsymbol{\pi}^*$ in (14) with $G(\cdot)$ being the empirical distribution of $\{\boldsymbol{X}_i\}$. We compare the frequentist properties of the posterior mean $\boldsymbol{\pi}^*$ obtained from analyses with or without covariates upon repeated use across independent replications.[24] We compare the proposed model with 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, to mimic weakly informative prior for TPRs, we use independent Beta(7.13,1.32) TPR prior distributions that match a wide interval 0.55 to 0.99 with the lower and upper 2.5% quantiles, respectively. We tested other priors that matched slightly wider or narrower ranges and observed similar advantages of the proposed model. The priors for other parameters are specified in Section 5.4.

*Simulation I.* We demonstrate posterior inference of true CSCF functions $\{\pi^0_{\ell}(\boldsymbol{X})\}$. 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 2). 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 earlier enrollment dates. Row 2 of Figure 3 shows that for the 9 causes in the columns, the posterior means and 95% CrIs for the CSCF functions $\pi_{\ell}(\cdot)$ well recover the simulation truths $\pi^0_{\ell}(\cdot)$ in the Stratum $S = 1$; similarly good recovery is observed for stratum $S = 2$. Figure S1 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 $X$.

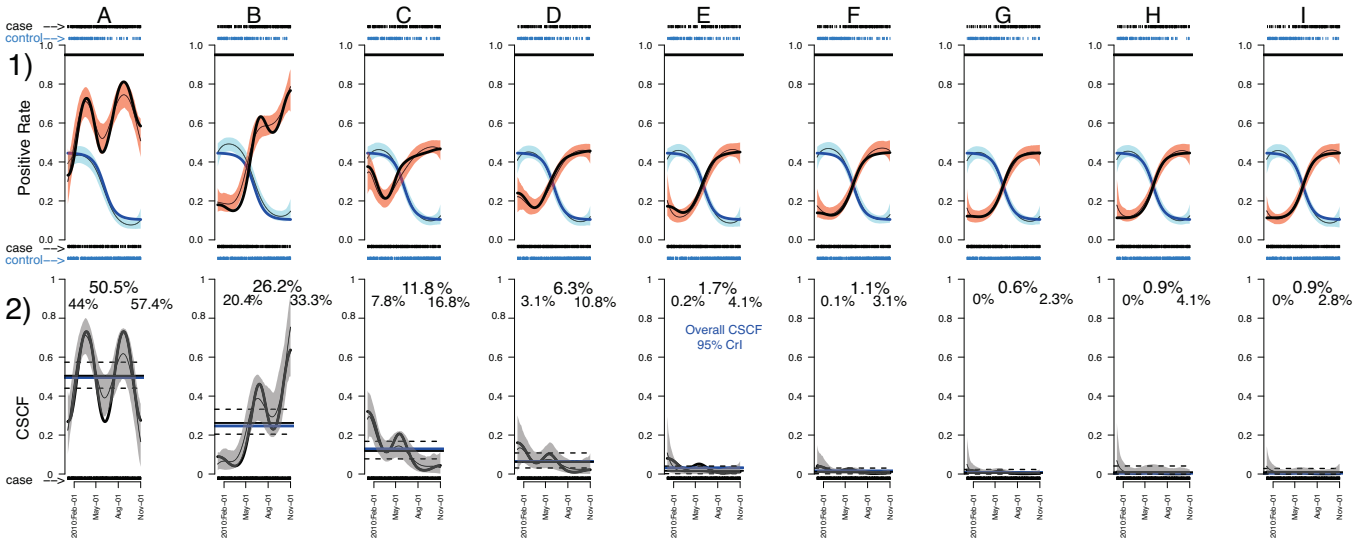



**FIGURE 3** Row 2) For each of the 9 single-agent causes (by column) in Simulation I, the posterior mean (thin black curves) and pointwise 95% CrIs (gray) for the CSCF curves $\pi_\ell(x)$ are close to the simulation truths $\pi_\ell^0(x)$ in the stratum of $S = 1$ (thick black curves). The overall CSCF in the particular stratum is shown along with the 95% CrI on both sides.[25] 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.

*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 set $W = X$. 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), and effects of $W$ on $\{\nu_k(W)\}$ and $\{\eta_k(W)\}$.

In estimating $\pi_\ell^*$, we focus on evaluating the marginal 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 4(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 4(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 centered closer to the true TPR would further reduce 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 $\pi_\ell^*$ that have more valid empirical coverage rates in all the scenarios (Figure 4(b)). Misspecified models without covariates concentrate the posterior distribution away from the true overall CSCFs, resulting in severe under-coverage.

The proposed model is flexible in incorporating covariates which enables the borrowing of information from subjects with similar covariates. As expected, we observed improved prediction performance in simulation settings. As a more realistic test

of the proposed method, we will discuss in more detail the comparison of the relative predictive performances of the proposed method against a closely related stratified analytic approach on the real data analysis in Section 7.

The hyperparameter choice is critical for adequate model performance, e.g., in terms of producing small posterior mean squared error (PMSE). We conducted sensitivity analysis for the hyperparameter that controls the sparseness of subclass weights at values $\nu_0 = 0.01, 0.1, 1, 10, 100$. We observed that the influence of hyperparameter upon the PMSE for CSCFs is stronger when the parameters settings correspond to weaker signals, e.g., when $\theta_k^{(j)}$ is closer to $\psi_k^{(j)}$; the results are stable for different hyperparameter choices under simulation scenarios corresponding to stronger signals.

## 7 | 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 more complete 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.

The names of the pathogens and the abbreviations are (i) bacteria: *Haemophilus influenzae* (HINF) and *Streptococcus pneumoniae* (PNEU), (ii) viruses: adenovirus (ADENO), human metapneumovirus type A or B (HMPV_A_B), parainfluenza type 1 virus (PARA_1), rhinovirus (RHINO), and respiratory syncytial virus (RSV).

We also include in the analysis the case-only, perfectly specific but imperfectly sensitive blood culture (BCX) diagnostic test results for two bacteria from cases only. For BCX data, we assume perfect specificity which is guided by the fact that if a pathogen did not infect the lung, it cannot be cultured from the blood (so we do not need control data to estimate the specificities). Detailed analyses of the entire data are reported elsewhere.[1] Table 1 shows the observed frequencies in the $\boldsymbol{W} = (\text{age}, \text{HIV status})$ strata for controls and $\boldsymbol{X} = (\text{age}, \text{HIV status}, \text{disease severity})$ 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.

Regression modeling techniques must be used to deal with small strata of cases and controls and obtain uncertainty quantification. Consider the 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 be further stratified by additional covariates (e.g., disease severity), resulting in finer case strata nested in each control stratum (e.g., see Table 1). Because every npLCM fit requires a case and control sample, a control stratum would have to be reused for every finer case strata. We use the proposed regression model to address these issues.

Figure S4 in the Supplementary Materials shows summary statistics for the NPPCR and BCX 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).

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 $\mathcal{C}_{\text{NoS}} = \emptyset$ 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 $a_\theta = 12.68$ and $b_\theta = 4.83$, 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 $K$s remain nearly the same. In the presence of BCX data for a subset of two bacteria (only bacteria can be cultured from blood), because BCX data have very low prevalence, we multiple $L^{\text{reg}}$ by BCX data likelihood in the simpler pLCM (2)-(4) and specify the Beta(7.59,58.97) prior for the two TPRs of BCX measurements matching the range of $5 - 20\%$ based on existing vaccine probe trials.[26] 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: $\mathbf{1}\{\text{age}_i \geq 1 \text{ year}\}$ and $\mathbf{1}\{\text{HIV}_i = 1\}$. The prior distributions for other parameters follow the specification in Section 5.4.

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 5 shows for two age-HIV-severity strata the

posterior mean CSCF curves and 95% pointwise credible bands of the $\pi_\ell(t, \mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ as a function of $t$. For example, among the younger, HIV negative and severe pneumonia children (Figure 5(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), suggesting prioritization of 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 5(b) shows a lower degree of seasonal variation of the RSV CSCF curve among the older, HIV negative and severe pneumonia children.

The inferential algorithm based on the regression model can also perform individual-level cause-specific probability assignment given a case's measurements and automatically use covariate values during assignment. Figure S5 in the Supplementary Materials show distinct cause-specific probabilities for two cases (one older than one and the other younger than one) with all-negative NPPCR results.

We estimate the overall CSCFs $\boldsymbol{\pi}^*(\mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ in every age-HIV-severity stratum by averaging the CSCF function estimates over the empirical distribution of enrollment date in each stratum. For example, contrasting the two age-HIV-severity strata in Figure 5(a) and 5(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 estimate for RSV ($\pi^*_{\mathsf{RSV}}(\mathsf{age} = 0, \mathsf{HIV} = 0, \mathsf{severity} = 0)$) drops from 47.7 (95% CrI : 37.6, 61.5)% to 17.3 (95% CrI : 8.0, 29.1)%. The CSCF estimate for NoS ($\pi^*_{\mathsf{NoS}}(\mathsf{age} = 1, \mathsf{HIV} = 0, \mathsf{severity} = 0)$) increases 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, HIV negative severe pneumonia children.

*Comparison with Other Methods.* The most relevant comparison method is based on stratum-specific npLCM analysis. Because in the data analysis, the enrollment date is continuous, stratification requires categorization of the enrollment date into "time periods", which after discussion with scientists in PERCH study is not straightforward. We decided to only stratify on the three binary covariates used in the main analysis, $\mathtt{I}(\mathsf{age} > 1)$, $\mathtt{I}(\mathsf{HIV\ positive})$, $\mathtt{I}(\mathsf{very\ severe\ pneumonia})$. Because in the PERCH study, case and control enrollment are frequency matched, leaving out enrollment date still provides a reasonable comparison against the main method. We re-emphasize that, however, the stratified analysis cannot output a CSCF as a function of continuous covariates such as enrollment date.

We compare the proposed and the stratum-specific analysis via prediction of the multivariate binary responses $\{\boldsymbol{M}_i\}$ from the diagnostic tests. We define the prediction based on the posterior mean of the positive response probabilities given all the data $\mathcal{D}$. We predict a response $M_{ij}$ to be 1 if the posterior mean of the positive response probability given the individual's covariate values exceeds 0.5, i.e. $\mathbb{E}[\sum_{\ell=1}^{L} \pi_\ell(\boldsymbol{x}_i) \sum_{k=1}^{K} \eta_k(\boldsymbol{w}_i) p_{k\ell}^{(j)} \mid \mathcal{D}] > 0.5$; otherwise, we let the prediction be 0. The prediction error is defined as the fraction of incorrectly predicted responses. We found that the proposed regression method has gained additional ability to predict the observed responses compared to the stratum specific method. This is partly due to the ability of the regression model to borrow information across subjects with different covariate values, e.g., through additive assumptions and smoothing over the continuous enrollment date. We also conducted a no-covariate analysis and compared its prediction performance; the complete no-covariate analysis results are shown in Figure S4 in the Supplementary Materials. Among the three methods, proposed regression model, stratified analysis and no-covariate analysis, the prediction errors are 10.3%, 17.8%, 24.7%, respectively, with the smallest prediction error based on the proposed regression model.

## 8 | DISCUSSION

In disease etiology studies where gold-standard data are infeasible to obtain, epidemiologists need to integrate case and control data to draw inference about the population CSCFs and individual cause-specific probabilities that may depend on covariates. The only existing methods for case-control data based on npLCM 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. The model-based inferential algorithm estimates CSCF functions and is amenable to efficient posterior computation.

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. Our regression analysis uses data from one PERCH site with more complete covariate information and reveals prominent dependence of the CSCFs of the virus RSV upon seasonality and a pneumonia child's age, HIV status and disease severity.

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 TPR priors, the proposed regression analysis avoids overly-optimistic uncertainty estimates for the overall CSCFs.

The covariate-stratified approach have the advantage of being simple-to-understand and easy to implement given the previous generation of the nested partially-latent class model. However, the pain points are 1) its requirement on the sample size for each covariate stratum, and 2) the controls may be reused as statistical comparison group for multiple case stratum. As shown in Table 1, both issues appeared. First, small sample sizes in many strata (e.g., the final row has only three cases: age older than 1, HIV positive, very severe) result in a stratified analysis that is not statistically stable. Turning to the second issue of selecting the control group, a naive stratified analysis would need to use the 51 controls twice, one for the not very severe case group (24 cases) and the other for very severe case group (3 cases). Reusing control data twice may lead to incorrect variance estimate of the CSCFs. The above two issues prevented producing reliable results based on the stratified analysis. However, we provided comparison results with only discrete covariates in the regression framework and showed advantage of the proposed regression approach that can handle both continuous and discrete covariates.

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.[27] 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[28]. Third, when the subsets of pathogens $\{C_\ell\}$ that have caused the diseases in the population are unknown, the proposed method can be combined with subset selection procedures[29]. Finally, wide applicability of posterior sampling algorithms have increased the use of complex and multilevel hierarchical models like the proposed model in this paper. Other methods for hyperparameter choice may be marginal likelihood maximization, specifying another layer of hyperpriors, or empirical Bayes approach. A more thorough comparative study for different methods of hyperparameter choice is warranted. We leave this for future work.

## ACKNOWLEDGMENTS

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 are grateful for comments from the Editor, Associate Editor and two reviewers that greatly improved the presentation. *Conflict of Interest*: None declared.

## References

1. PERCH Study Group . 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; 392(10200): 757–779.

2. Hammitt LL, Feikin DR, Scott JAG, et al. Addressing the analytic challenges of cross-sectional pediatric pneumonia etiology data. *Clinical infectious diseases* 2017; 64(suppl_3): S197–S204.

3. Wu Z, Deloria-Knoll M, Hammitt LL, Zeger SL, PERCH Study Team t. Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2016; 65(1): 97–114.

4. Wu Z, Deloria-Knoll M, Zeger SL. Nested partially latent class models for dependent binary data; estimating disease etiology.. *Biostatistics (Oxford, England)* 2017; 18: 200–213. doi: 10.1093/biostatistics/kxw037

5. Saha SK, Schrag SJ, El Arifeen S, et al. Causes and incidence of community-acquired serious infections among young children in south Asia (ANISA): an observational cohort study. *The Lancet* 2018; 392(10142): 145–159.

6. Kotloff KL, Nataro JP, Blackwelder WC, et al. 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* 2013; 382(9888): 209–222.

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

8. McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association* 2016; 111(515): 1036–1049.

9. Lazarsfeld PF. The logical and mathematical foundations of latent structure analysis. In: Stouffer S., ed. *The American Soldier: Studies in Social Psychology in World War II*. IV. Princeton, NJ: Princeton University Press. 1950 (pp. 362-412).

10. Moran KR, Turner EL, Dunson D, Herring AH. Bayesian Hierarchical Factor Regression Models to Infer Cause of Death From Verbal Autopsy Data. *arXiv preprint arXiv:1908.07632* 2019.

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

12. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; 60(2): 427–435.

13. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 2007; 8(2): 474–484.

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

15. Linderman S, Johnson M, Adams RP. Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation. In: ; 2015: 3456–3464.

16. Rodriguez A, Dunson DB. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* 2011; 6(1): 145–177.

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

18. Dimatteo I, Genovese CR, Kass RE. Bayesian curve-fitting with free-knot splines. *Biometrika* 2001; 88(4): 1055-1071. doi: 10.1093/biomet/88.4.1055

19. Morrissey ER, Juárez MA, Denby KJ, Burroughs NJ. Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression. *Biostatistics* 2011; 12(4): 682–694.

20. Ni Y, Stingo FC, Baladandayuthapani V. Bayesian nonlinear model selection for gene regulatory networks. *Biometrics* 2015; 71(3): 585-595.

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

22. Deloria Knoll M, Fu W, Shi Q, et al. Bayesian estimation of pneumonia etiology: epidemiologic considerations and applications to the Pneumonia Etiology Research for Child Health study. *Clinical Infectious Diseases* 2017; 64(suppl_3): S213–S227.

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

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

25. Louis TA, Zeger SL. Effective communication of standard errors and confidence intervals. *Biostatistics* 2009; 10(1): 1–2.

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

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

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

29. Gu Y, Xu G. Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research* 2019; 20(115): 1–58.

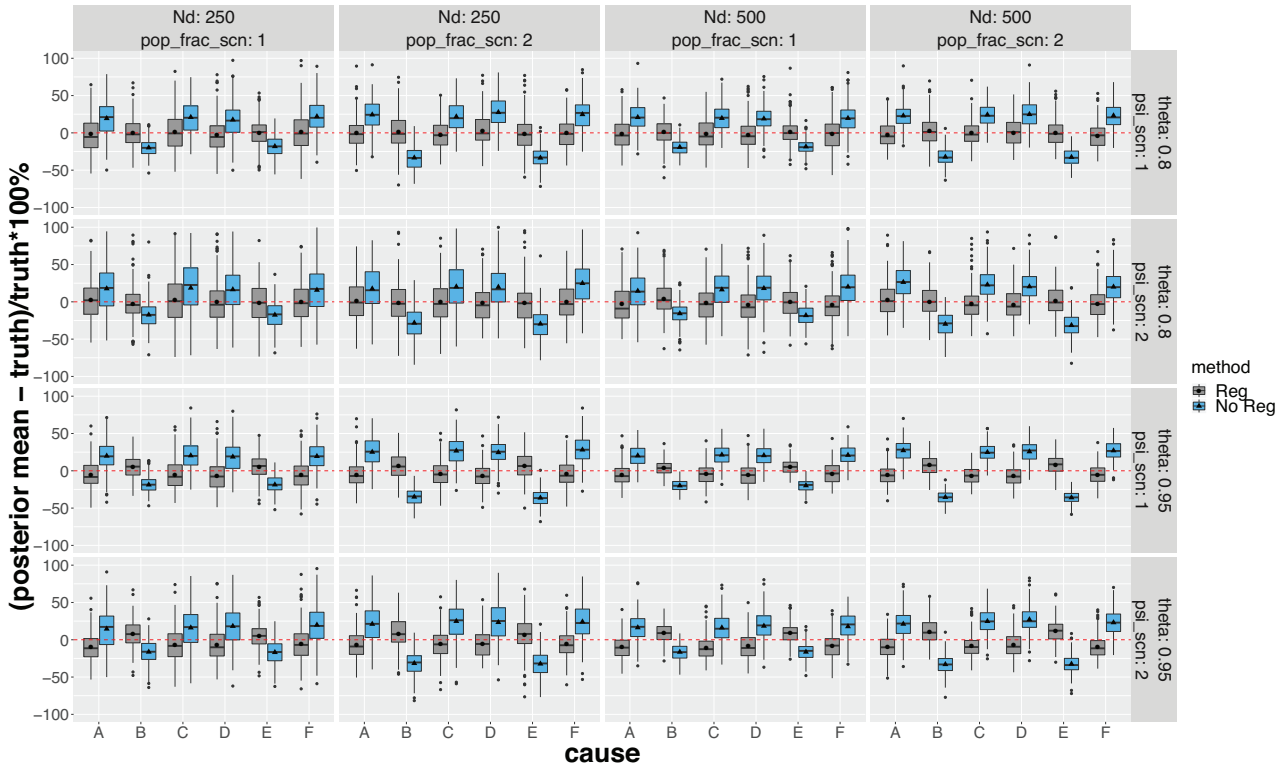
## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

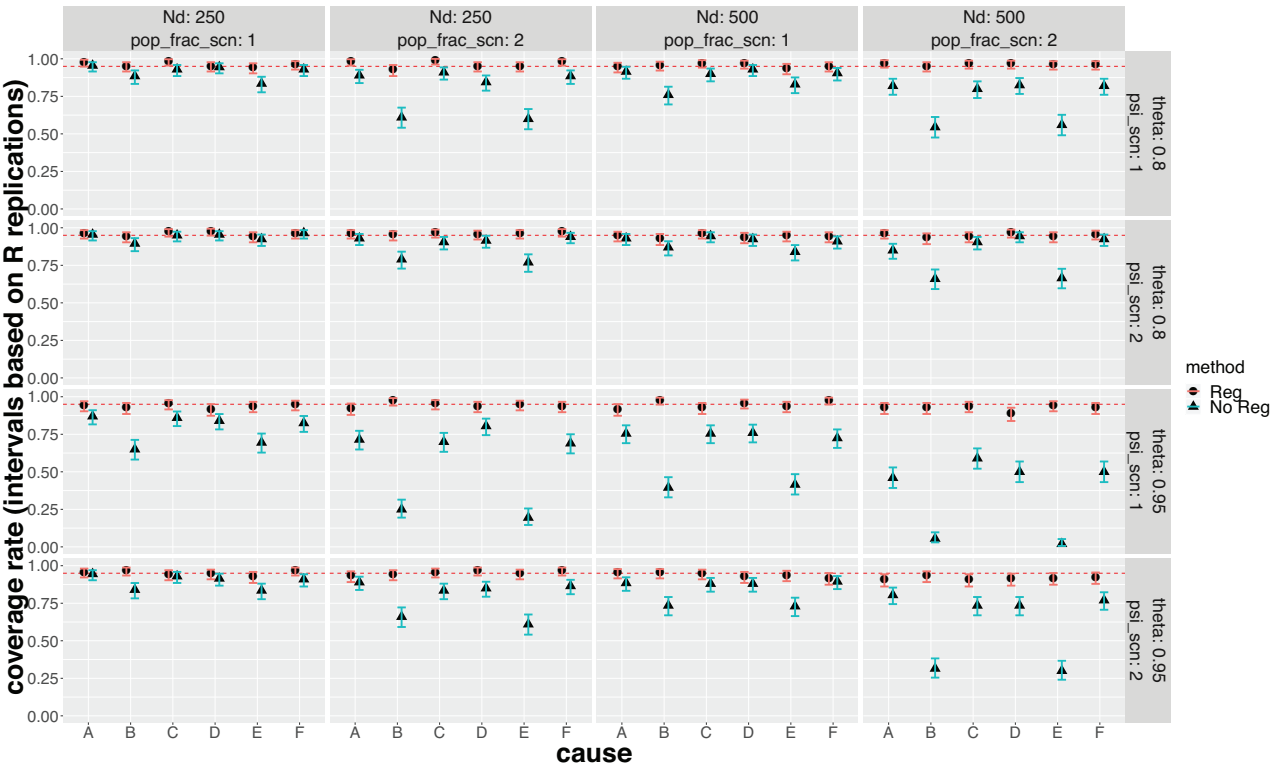
**TABLE 1** The observed counts (frequencies) of controls by age and HIV status; Case counts are further stratified by disease severity (1: yes; 0: no). The observed marginal rates are shown at the bottom. Enrollment date ($t$) is not stratified upon here.

| age $\geq$ 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% | | - | |

**How to cite this article:** Z. Wu and I. Chen (2020+), Probabilistic Cause-of-disease Assignment using Case-control Diagnostic Tests: A Hierarchical Bayesian Latent Variable Regression Approach, *Statistics in Medicine*, *20XX;XX:X–X*.

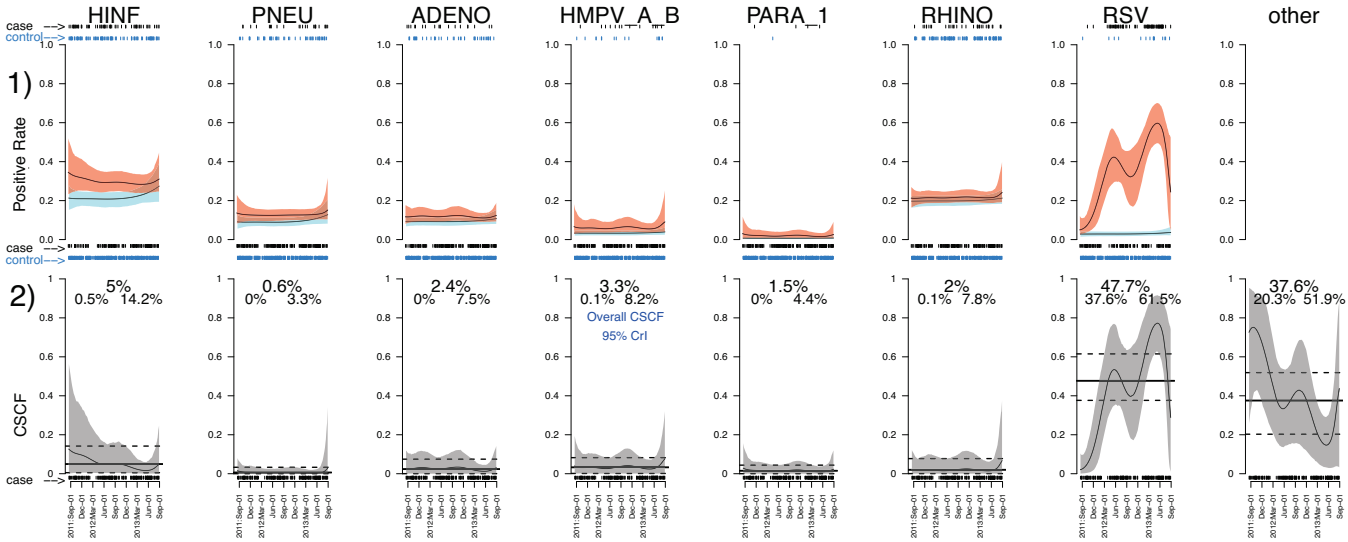



(a) Percent relative bias

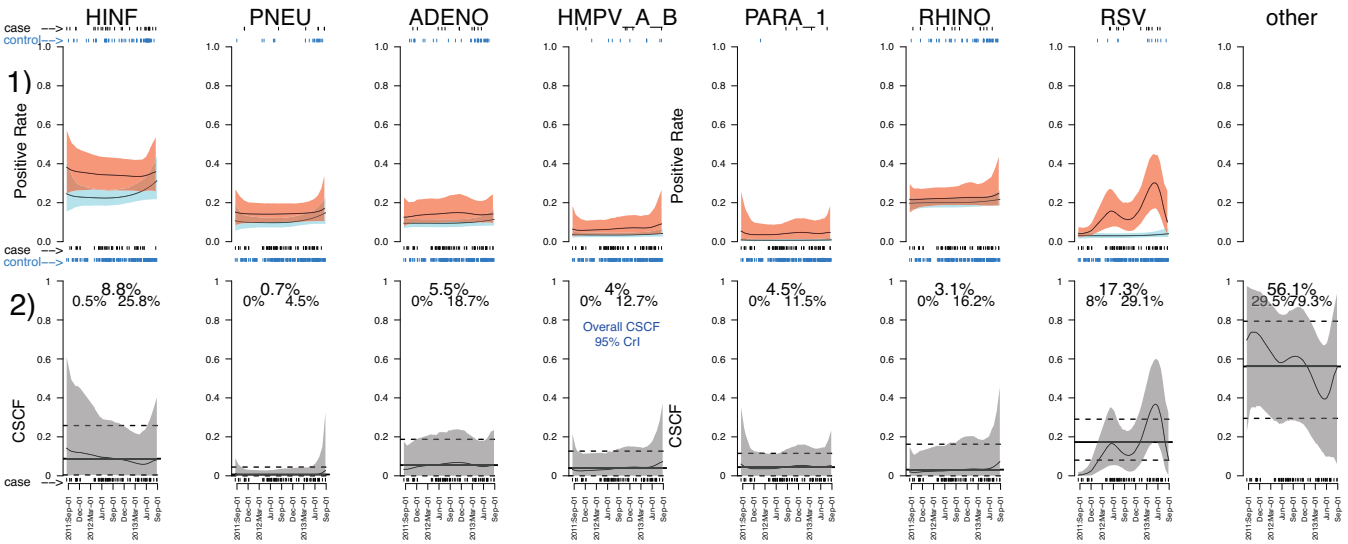



(b) Empirical coverage rates

**FIGURE 4** The regression analyses produce less biased posterior mean estimates and more valid empirical coverage rates for $\pi_\ell^*$ 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 $\pi_\ell^*$ 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 $\pi_\ell^*$ is computed from $R = 200$ binary observations of the true $\pi_\ell^{0*}$ being covered or not, a 95% CI is also shown.




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




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

**FIGURE 5** 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 $\ell$:

Row 2): Temporal trend $\widehat{\pi}_\ell(\mathrm{t}; \mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ enveloped by pointwise 95% CrIs (gray). The horizontal solid line shows the estimated overall CSCF $\widehat{\pi}^*_\ell(\mathsf{age}, \mathsf{HIV}, \mathsf{severity})$ 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.