Biometrics XX, 1–25 DOI: 10.1111/XXXX
XXXX XXX

Regression Analysis of Dependent Binary Data for Estimating Disease Etiology from Case-Control Studies

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

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

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

**email:* zhenkewu@umich.edu

Summary: In large-scale disease etiology studies, epidemiologists often need to use multiple binary measures of unobserved causes of disease that are not perfectly sensitive or specific to estimate cause-specific case fractions, referred to as "population etiologic fractions" (PEFs). Despite recent methodological advances, the scientific need of incorporating control data to estimate the effect of explanatory variables upon the PEFs, however, remains unmet. In this paper, we build on and extend nested partially-latent class model (npLCMs, Wu et al., 2017) to a general framework for etiology regression analysis in case-control studies. Data from controls provide requisite information about measurement specificities and covariations, which is used to correctly assign cause-specific probabilities for each case given her measurements. We estimate the distribution of the controls' diagnostic measures given the covariates via a separate regression model and *a priori* encourage simpler conditional dependence structures. We use Markov chain Monte Carlo for posterior inference of the PEF functions, cases' latent classes and the overall PEFs of policy interest. We illustrate the regression analysis with simulations and show less biased estimation and more valid inference of the overall PEFs than an npLCM analysis omitting covariates. Regression analysis of data from a childhood pneumonia study site reveals the dependence of pneumonia etiology upon season, age, disease severity and HIV status.

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

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

In epidemiologic studies of disease etiology, one important scientific goal is to assess the effect of explanatory variables upon disease etiology. Based on multiple binary non-gold-standard diagnostic measurements made on a list of putative causes with different error rates, this paper develops and demonstrates a regression analytic approach for drawing inference about the cause-specific fractions among the case population that depend on covariates. We illustrate the analytic needs raised by a study of pediatric pneumonia etiology.

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 (Scott et al., 2008). 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 goal of the PERCH study is to estimate the population fractions of cases due to the pathogen causes, referred to as "population etiologic fractions" (PEFs) and to assign cause-specific probabilities for each pneumonia child given her measurements, termed as "individual etiologic fractions" (IEFs). The PERCH study also aims to understand the variation of the PEFs as a function of 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 et al., 2017). The PERCH study tests the presence or absence of a list of pathogens using specimens in peripheral compartments including the blood, sputum, pleural fluid and nasopharyngeal (NP) cavity (Crawley et al., 2017). In this paper, we focus on two sources of imperfect measurements: (a) NP Polymerase Chain Reaction (NPPCR) results from cases and controls that are not perfectly sensitive or specific, referred to as "bronze-standard"

(BrS) data; and (b) blood culture (BCX) results from cases only that are perfectly specific but lack sensitivity, referred to as "silver-standard" (SS) data.

Valid inference about the population and individual etiologic fractions must address three salient characteristics of the measurements. First, tests lacking sensitivity such as NPPCR and BCX may miss true causative agent(s) which if unadjusted may underestimate the PEFs. Second, imperfect diagnostic specificities may result in the detection of multiple pathogens in NPPCR that may indicate asymptomatic 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 and challenges of integrating multiple sources of imperfect measurements of multiple pathogens to produce an accurate understanding of etiology (e.g., Saha et al., 2018; Kotloff et al., 2013).

To address the analytic needs, Wu et al. (2016) introduced a *partially-latent class model* (pLCM) as an extension to classical latent class models (LCMs Lazarsfeld, 1950; Goodman, 1974) that uses case-control data to estimate the PEFs. This prior work shows the capacity of the multivariate specimen measurements to inform the distribution of unobserved, or "latent" health status for an individual and the population. PLCM is a finite mixture model with $L+1$ components for multivariate binary data where 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 assumed drawn from an observed class. The pLCM is a *semi-supervised* method for learning the unobserved classes, where the "label" (cause of disease) is observed for only a subset of subjects. Let $I_i \in \{1, \ldots, L\}$ represent case i's disease class which is categorically distributed with probabilities equal to the PEFs $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top$ in the $(L-1)$-dimensional simplex $\mathcal{S}_{L-1} = \{\boldsymbol{\pi} : \sum_{\ell=1}^{L} \pi_\ell = 1, 0 \leqslant \pi_\ell \leqslant 1\}$. A case class can represent a single- or multiple-pathogen cause of pneumonia, or pathogen causes not

targeted by the assays which we refer to as "Not Specified (NoS)". PLCM uses a vector of J response probabilities to specify the conditional distribution of the measurements in each class. PLCM is an example of *restricted* LCMs (RLCMs, Wu et al., 2019) which restrict how the response probabilities differ by class to reflect the scientific knowledge that causative pathogens are more likely to appear in the upper respiratory tract in a pneumonia child than a healthy control. In particular, each causative pathogen is assumed to be observed with a higher probability in case class ℓ (sensitivity or true positive rate, TPR) than among the controls; A non-causative pathogen is observed with the same probability as in the controls (1 - specificity or false positive rate). Under the pLCM, a higher observed marginal positive rate of pathogen j among cases than controls indicates etiologic significance.

In a Bayesian framework, measurements of differing precisions can be optimally combined under a pLCM to generate stronger evidence about $\boldsymbol{\pi}$. The pLCM is partially-identified (Jones et al., 2010; Gu and Xu, 2019b). There exist two sets of values of a subset of model parameters (here the TPRs) that the likelihood function alone cannot distinguish even with infinite samples; Bounds on the parameters however are available (e.g., Wu et al., 2016, Equation 6). Informative prior distributions for the TPRs elicited from laboratory experts or estimated from vaccine probe studies for a subset of pathogens (Feikin et al., 2014) can be readily incorporated to improve inference (Gustafson, 2015).

The pLCM assumes "local independence" (LI) which means the BrS measurements are mutually independent given the class membership. This classical assumption is central to mixture models for multivariate data, because the estimation procedures essentially find the optimal partition of observations into subgroups so that the LI approximately holds in each subgroup. Deviations from LI, or "local dependence" (LD) are testable using the control BrS data, which can be accounted for by an extension of pLCM, called *nested partially-latent class model* (npLCM, Wu et al., 2017). In each class, the npLCM uses the classical

LCM formulation that has the capacity to describe complex multivariate dependence among discrete data (Dunson and Xing, 2009). For example, it assumes the within-class correlations among NPPCR tests are induced by unobserved heterogeneity in subjects' propensities for pathogens colonizing the nasal cavities. 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. In a Bayesian framework with a prior that encourages few important subclasses, the npLCM reduces the bias in estimating $\boldsymbol{\pi}$, retains estimation efficiency and offers more valid inference under substantial deviation from LI.

Extensions to incorporate covariates in an npLCM are critical for two reasons. Firstly, covariates such as season, age, disease severity and HIV status may directly influence $\boldsymbol{\pi}$. Secondly, in an npLCM without covariates, the relative probability of assigning a case subject to class ℓ versus class ℓ' depends on the FPRs (Wu et al., 2016) which are estimable using the control data. However, the FPRs may vary by covariates which if not modeled will bias the assignment of cause-specific probabilities for each case subject. For example, pathogen A found in a case's nasal cavity less likely indicates etiologic significance than a colonization during seasons with high asymptomatic carriage rates, and much more so when the same pathogen rarely appears in healthy subjects.

Adapting existing no-covariate methods to account for discrete covariates, one may perform a *fully-stratified analysis* by fitting an npLCM to the case-control data in each covariate stratum. Like pLCM, the npLCM is partially-identified in each stratum (Wu et al., 2017), necessitating multiple sets of *independent* informative priors across multiple strata. There are two primary issues with this approach. First, sparsely-populated strata defined by many discrete covariates may lead to unstable PEF estimates. Second, it is often of policy interest to quantify the overall cause-specific disease burdens in a population. Let the overall PEFs $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ be the empirical average of the stratum-specific PEFs. 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 will lead to overly-optimistic posterior uncertainty in $\boldsymbol{\pi}^*$, hampering policy decisions.

Estimating disease etiology across discrete and continuous epidemiologic factors needs new methods in a general modeling framework. In this paper, we extend the npLCM to perform regression analysis in case-control disease etiology studies that (a) incorporates controls to estimate the PEFs, (b) specifies parsimonious functional dependence of $\boldsymbol{\pi}$ upon covariates such as additivity, and (c) correctly assesses the posterior uncertainty of the PEF functions and the overall PEFs $\boldsymbol{\pi}^*$ by applying the TPR priors just once.

The rest of the paper is organized as follows. Section 2 overviews the npLCM without covariates. Section 3 builds on the npLCM and makes the regression extension. We demonstrate the estimation of disease etiology regression functions $\pi_\ell(\cdot)$ through simulations in Section 4; We also show superior inferential performance of the regression model in estimating the overall PEFs $\boldsymbol{\pi}^*$ relative to an analysis omitting the covariates. In Section 5, we characterize the effect of seasonality, age, HIV status upon the PEFs by applying the proposed npLCM regression model to the PERCH data. The paper concludes with a discussion.

2. Overview of npLCMs without Covariates

Let binary BrS measurements $\boldsymbol{M}_i = (M_{i1}, ..., M_{iJ})^\top$ indicate the presence or absence of J pathogens for subject $i = 1, \ldots, N$. Let Y_i indicate a case (1) or a control (0) subject. If $Y_i = 1$, let $I_i \in \{1, \ldots, L\}$ represent case i's unobserved disease class; Otherwise, let $I_i = 0$ because a control subject's class is known (in PERCH, no lung infection). In this paper, we simplify the presentation of models by focusing on single-pathogen causes (hence $L = J$). The npLCM readily extends to $L > J$ for including additional pre-specified multi-pathogen and/or "Not Specified" (NoS) causes (Wu et al., 2017).

The likelihood function for an npLCM has three components: (a) PEFs or cause-specific

case fractions: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top = \{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1}$; (b) $\boldsymbol{P}_{1\ell} = \{\boldsymbol{P}_{1\ell}(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = \ell, Y = 1)\}$: a table of probabilities of making J binary observations $\boldsymbol{M} = \boldsymbol{m}$ in a case class $\ell \neq 0$; (c) $\boldsymbol{P}_0 = \{\boldsymbol{P}_0(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = 0, Y = 0)\}$: the same probability table but for controls. Since cases' disease classes are unobserved, the distribution of cases' measurements $\boldsymbol{P}_1 = \mathbb{P}(\boldsymbol{M} \mid Y = 1)$ is a finite-mixture model with weights $\boldsymbol{\pi}$ for the L disease classes: $\boldsymbol{P}_1 = \sum_{\ell=1}^{L} \pi_\ell \boldsymbol{P}_{1\ell}$.

Models in this section differ by how $\boldsymbol{P}_0$ and $\{\boldsymbol{P}_{1\ell}\}$ are specified; Regression models in Section 3 further incorporate covariate into the specifications ($\boldsymbol{\pi}$ as well). More specifically, the likelihood of an npLCM (Wu et al., 2017) is a product of case (L_1) and control (L_0) likelihood functions

$$L = L_1 \cdot L_0 = \left\{ \prod_{i:Y_i=1} \sum_{\ell=1}^{L} \pi_\ell \cdot \boldsymbol{P}_{1\ell}(\boldsymbol{M}_i; \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\eta}) \right\} \times \prod_{i':Y_{i'}=0} \boldsymbol{P}_0(\boldsymbol{M}_{i'}; \boldsymbol{\Psi}, \boldsymbol{\nu}), \tag{1}$$

where $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ are sensitivity and specificity parameters necessary for modeling the imperfect measurements; The rest of parameters $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^\top \in \mathcal{S}_{K-1}$. Existing methods for estimating $\boldsymbol{\pi}$ in the framework of npLCM can be classified by whether or not $\boldsymbol{P}_0$ and $\boldsymbol{P}_{1\ell}$ assumes local independence (LI) which means measurements are independent of one another given the class ($I_i = \ell = 0, 1, \ldots, L$). In Equation (1), LI results if and only if $\nu_1 = \eta_1 = 1$; Otherwise, $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ account for deviations from LI given a control or disease class.

PLCM. $\boldsymbol{P}_0(\boldsymbol{m})$ under the original pLCM (Wu et al., 2016) satisfies LI and equals a product of J probabilities: $\boldsymbol{P}_0(\boldsymbol{m}) = \prod_{j=1}^{J} \{\psi_j\}^{m_j} \{1 - \psi_j\}^{1-m_j} = \Pi(\boldsymbol{m}; \boldsymbol{\psi})$, where $\Pi(\boldsymbol{m}; \boldsymbol{s}) = \prod_{j=1}^{J} \{s_j\}^{m_{ij}} \{1 - s_j\}^{1-m_{ij}}$ is the probability mass function for a product Bernoulli distribution given the success probabilities $\boldsymbol{s} = (s_1, \ldots, s_J)^\top$, $0 \leqslant s_j \leqslant 1$ and the parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_J)^\top$ represent the positive rates absent disease, referred to as "false positive rates" (FPRs). For example, in the PERCH data, Respiratory Syncytial Virus (RSV) has a

low observed FPR because of its rare appearance in controls' NPs; Other pathogens such as Rhinovirus (RHINO) have higher observed FPRs.

For $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$, the pLCM makes a key "non-interference" assumption 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 (Wu et al., 2017). The "non-interference" assumption says that $\boldsymbol{P}_{1\ell}(\boldsymbol{m})$ in a case class $\ell \neq 0$ is a product of the probabilities of measurements made (a) on the *causative* pathogen ℓ, $\mathbb{P}(M_\ell \mid I = \ell, Y = 1, \boldsymbol{\theta}) = \{\theta_\ell\}^{M_\ell}\{1-\theta_\ell\}^{1-M_\ell}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^\top$ and (b) on the *non-causative* pathogens $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \boldsymbol{\psi}_{[-\ell]}) = \Pi(\boldsymbol{M}_{[-\ell]}; \boldsymbol{\psi}_{[-\ell]})$, where $\boldsymbol{a}_{[-\ell]}$ represents all but the ℓ-th element in a vector $\boldsymbol{a}$. The parameter θ_ℓ is termed "true positive rate" (TPR) and may be larger than the FPR ψ_ℓ; Under the single-pathogen-cause assumption, pLCM uses J TPRs $\boldsymbol{\theta}$ for $L = J$ causes and J FPRs $\boldsymbol{\psi}$.

<u>*NPLCM.*</u> To reduce estimation bias in $\boldsymbol{\pi}$ under deviations from LI, the "nested pLCM" or npLCM extends the original pLCM to describe residual correlations among J binary pathogen measurements in the controls ($I_i = 0$) and in each case class ($I_i = \ell$, $\ell \neq 0$) (Wu et al., 2017). The extension is motivated by the ability of the classical LCM formulation (Lazarsfeld, 1950) to approximate any joint multivariate discrete distribution (Dunson and Xing, 2009).

For $\boldsymbol{P}_0(\boldsymbol{m})$ in the controls, the npLCM introduces K subclasses; The original pLCM results if $K = 1$. Given a subclass k, the probability of observing J binary measurements $\boldsymbol{M} = \boldsymbol{m}$ among controls is $\boldsymbol{P}_0^{(k)}(\boldsymbol{m}) = \mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid Z = k, I = 0, Y = 0, \{\psi_k^{(j)}\}) = \Pi(\boldsymbol{m}; \boldsymbol{\Psi}_k)$, where $\boldsymbol{\Psi}_k$ is the k-th column of a J by K FPR matrix $\boldsymbol{\Psi} = \{\psi_k^{(j)}\}$. Since we do not observe controls' subclasses, $\boldsymbol{P}_0$ is a weighted average of $\boldsymbol{P}_0^{(k)}$ according to the subclass probabilities $\{\nu_k\}$: $\boldsymbol{P}_0 = \sum_k^K \nu_k \boldsymbol{P}_0^{(k)}$.

For $\boldsymbol{P}_{1l}(\boldsymbol{m})$ in case class $\ell \neq 0$, the npLCM again introduces K unobserved subclasses

and assumes $\boldsymbol{P}_{1\ell}$ is a weighted average of $\boldsymbol{P}_{1\ell}^{(k)}$ according to the case subclass weights $\{\eta_k\}$: $\boldsymbol{P}_{1\ell} = \sum_{k=1}^{K} \eta_k \boldsymbol{P}_{1\ell}^{(k)}$. In particular, the npLCM assumes the probability of observing $\boldsymbol{M}$ in subclass k in disease class $\ell \neq 0$, $\boldsymbol{P}_{1l}^{(k)} = \mathbb{P}(\boldsymbol{M} \mid Z = k, I = \ell, Y = 1)$, is a product of the probabilities of making an observation (a) on the *causative* pathogen ℓ: $\mathbb{P}(M_\ell \mid Y = 1, Z = k, I = \ell, \theta_k^{(\ell)}) = \{\theta_k^{(\ell)}\}^{M_\ell}\{1 - \theta_k^{(\ell)}\}^{1-M_\ell}$ and (b) on *non-causative* pathogens $\mathbb{P}(\boldsymbol{M}_{[-\ell]} \mid Y = 1, Z = k, I = \ell, \boldsymbol{\Psi}_k^{([-\ell])}) = \Pi(\boldsymbol{M}_{[-\ell]}; \boldsymbol{\Psi}_k^{([-\ell])}) = \prod_{j\neq\ell}\{\psi_k^{(j)}\}^{m_j}\{1 - \psi_k^{(j)}\}^{1-m_j}$, where $\boldsymbol{\Psi}_k^{([-\ell])}$ is the k-th column of Ψ excluding the ℓ-th row. We collect the TPRs in a J by K TPR matrix $\boldsymbol{\Theta} = \{\theta_k^{(j)}\}$. We summarize the preceding specification by $\boldsymbol{P}_{1l}^{(k)} = \Pi(\boldsymbol{M}; \boldsymbol{p}_{k\ell}), \ell \neq 0$, where the vector $\boldsymbol{p}_{k\ell} = \{p_{k\ell}^{(j)}, j = 1, \ldots, J\}$ represents the positive rates for J measurements in subclass k of disease class ℓ: $p_{k\ell}^{(j)} = \left\{\theta_k^{(j)}\right\}^{\mathbb{I}\{j=\ell\}} \cdot \left\{\psi_k^{(j)}\right\}^{1-\mathbb{I}\{j=\ell\}}$ which equals the TPR $\theta_k^{(j)}$ for a causative pathogen and the FPR $\psi_k^{(j)}$ otherwise; Here $\mathbb{I}\{A\}$ is an indicator function that equals 1 if the statement A is true and 0 otherwise.

The likelihood for npLCM results upon substituting $\boldsymbol{P}_0$ and $\boldsymbol{P}_{1\ell}$ above into Equation (1): $L = L_1 \cdot L_0 = (\prod_{i:Y_i=1} \sum_{\ell=1}^{L} \pi_\ell[\cdot \sum_{k=1}^{K}\{\eta_k \cdot \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})\}]) \times \prod_{i':Y_{i'}=0} \sum_{k=1}^{K} \nu_k \cdot \Pi(\boldsymbol{M}_{i'}; \boldsymbol{\Psi}_k)$. Setting $\nu_1 = \eta_1 = 1$ and $\nu_k = \eta_k = 0, k \geqslant 2$, the special case of pLCM results.

Similar to the pLCM, the FPRs $\boldsymbol{\Psi}$ in the npLCM are shared among controls and case classes over non-causative pathogens (via $\boldsymbol{p}_{kl}$). Different from the pLCM, the subclass mixing weights may differ between cases ($\boldsymbol{\eta}$) and 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 no different from the controls. However, relative to controls, diseased individuals may have different strength and direction of 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 controls.

3. Regression Analysis via npLCM

We extend npLCM to perform regression analysis of data $\mathcal{D} = \{(\boldsymbol{M}_i, Y_i, \boldsymbol{X}_i Y_i, \boldsymbol{W}_i), i = 1, \ldots, N\}$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ are covariates that may influence case i's etiologic fractions and $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iq})^\top$ is a possibly different vector of covariates that may influence the subclass weights among the controls and the cases; Let the continuous covariates comprise the first p_1 and q_1 elements of $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$, respectively. A subset of $\boldsymbol{X}_i$ may be available from the cases only. We let $\boldsymbol{X}_i Y_i = \boldsymbol{0}_{p\times 1}$ if $Y_i = 0$ so that all the covariates for a control subject are included in $\boldsymbol{W}_i$; Let $\boldsymbol{X}_i Y_i = \boldsymbol{X}_i$ for a case subject. For example, healthy controls have no disease severity information. We let three sets of parameters in an npLCM (1) depend on the observed covariates: (a) the etiology regression function among cases, $\{\pi_\ell(\boldsymbol{x}), \ell \neq 0\}$, which is of primary scientific interest, (b) the conditional probability of measurements $\boldsymbol{m}$ given covariates $\boldsymbol{w}$ in case classes: $\boldsymbol{P}_{1\ell}(\boldsymbol{m}; \boldsymbol{w}) = [\boldsymbol{M} = \boldsymbol{m} \mid \boldsymbol{W} = \boldsymbol{w}, I = \ell]$, $\ell = 1, \ldots, L$, (c) and in the controls $\boldsymbol{P}_0(\boldsymbol{m}; \boldsymbol{w}) = [\boldsymbol{M} = \boldsymbol{m} \mid \boldsymbol{W} = \boldsymbol{w}, I = 0]$; We keep the specifications for the TPRs and FPRs ($\boldsymbol{\Theta}$, $\boldsymbol{\Psi}$) as in the original npLCM.

3.1 *Disease Etiology Regression*

$\pi_\ell(\boldsymbol{X})$ is the primary target of inference. Recall that $I_i = \ell$ represents case i's disease being caused by pathogen ℓ. We assume this event occurs with probability $\pi_{i\ell}$ that depends upon covariates. In our model, we use a multinomial logistic regression model $\pi_{i\ell} = \pi_\ell(\boldsymbol{X}_i) = \exp\{\phi_\ell(\boldsymbol{X}_i)\} / \sum_{\ell'=1}^{L} \exp\{\phi_{\ell'}(\boldsymbol{X}_i)\}$, $\ell = 1, ..., L$, 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 additive models for $\phi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi) = \sum_{j=1}^{p_1} f_{\ell j}^\pi(x_j; \boldsymbol{\beta}_{\ell j}^\pi) + \widetilde{\boldsymbol{x}}^\top \boldsymbol{\gamma}_\ell^\pi$, where $\widetilde{\boldsymbol{x}}$ is the subvector of the predictors $\boldsymbol{x}$ that enters the model for all disease classes as linear predictors and $\boldsymbol{\Gamma}_\ell^\pi = (\boldsymbol{\beta}_{\ell j}^\pi, \boldsymbol{\gamma}_\ell^\pi)$ collects all the parameters. For covariates such as enrollment date that serves as a proxy for factors driven by seasonality, nonlinear functional dependence is expected. We

use B-spline basis expansion to approximate $f^{\pi}_{\ell j}(\cdot)$ and use P-spline for estimating smooth functions (Lang and Brezger, 2004). Finally, we specify the distribution of case measurements $\boldsymbol{M}$ given disease class I, covariates $\boldsymbol{X}$ and $\boldsymbol{W}$. We extend the case likelihood L_1 in an npLCM (1) to let the subclass weights depend on covariates $\boldsymbol{W}$: $P(\boldsymbol{M} \mid \boldsymbol{W}, I = \ell, Y = 1) = \sum_{k=1}^{K} \eta_k(\boldsymbol{W}) \cdot \Pi\left(\boldsymbol{M}; \boldsymbol{p}_{k\ell}\right), \ell = 1, \ldots, L$. Integrating over L unobserved disease classes, we obtain the likelihood function for the cases that incorporates covariates $\{\boldsymbol{X}_i, \boldsymbol{W}_i\}$:

$$L_1^{\mathsf{reg}} = \prod_{i:Y_i=1} \left\{ \sum_{\ell=1}^{L} \left[\pi_\ell(\boldsymbol{X}_i; \boldsymbol{\Gamma}^{\pi}_{\ell}) \sum_{k=1}^{K} \{\eta_{ik} \cdot \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})\} \right] \right\}, \tag{2}$$

where $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}^{\eta}_k)$ and $\boldsymbol{\Gamma}^{\eta}_k$ are the regression parameters; The form of h_k is introduced in the model for controls.

3.2 *Covariate-dependent reference distribution*

Data from controls provide requisite information about the specificities and covariations at distinct covariate values, necessitating adjustment in an npLCM analysis. For example, factors such as enrollment date is a proxy for season and may influence the background colonization rates and interactions of some pathogens that circulate more during winter (Obando-Pacheco et al., 2018; Nair et al., 2011). We propose a novel approach to estimating the reference distribution of measurements that may depend on covariates using control data.

The regression model for a control subject is a mixture model with covariate-dependent mixing weights $\nu_k(\boldsymbol{W})$: $\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, Y = 0) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}) \Pi(\boldsymbol{M}; \boldsymbol{\Psi}_k)$, where FPRs $\boldsymbol{\Psi}_k = (\psi_k^{(1)}, \ldots, \psi_k^{(J)})^{\top}$ do not depend on covariates and the vector $\boldsymbol{\nu}(\boldsymbol{W}) = (\nu_1(\boldsymbol{W}), \ldots, \nu_K(\boldsymbol{W}))^{\top}$ lies in a $(K-1)$-simplex $\mathcal{S}_{K-1}$. We discuss the FPRs $\{\boldsymbol{\Psi}_k\}$ and the subclass weight functions $\{\nu_k(\boldsymbol{W})\}$ in order.

Firstly, constant FPR profiles $\{\boldsymbol{\Psi}_k\}$ enable coherent interpretation across individuals with different covariate values (Erosheva et al., 2007). FPR profile k receives a weight of $\nu_k(\boldsymbol{W}_i)$ for a control subject i with covariates $\boldsymbol{W}_i$. The *marginal* FPRs in the controls $\mathbb{P}(\boldsymbol{M}_j = 1 \mid \boldsymbol{W}, Y = 0, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \nu_k(\boldsymbol{W}) \psi_k^{(j)} \in [\min_k \psi_k^{(j)}, \max_k \psi_k^{(j)}]$, $j = 1, \ldots, J$, also depend on $\boldsymbol{W}$.

Consequently, observed marginal control positive curve for a pathogen informs how different the FPRs $\boldsymbol{\Psi}_k^{(j)}$ are across the subclasses. For example, if the NPPCR measure of pathogen A shows strong seasonal trends among the controls, the estimated FPRs will be more variable across the subclasses. And the subclass with a high FPR will receive a larger weight during seasons with higher carriage rates in controls. The control model reduces to special cases, with covariate-independent $\nu_k(\boldsymbol{W}) \equiv \nu_k$, $k = 1, \ldots, K$, resulting in the $\boldsymbol{P}_0$ in a K-subclass npLCM without covariates; A further single-subclass constraint ($K = 1$) gives the $\boldsymbol{P}_0$ in the original pLCM.

Secondly, we parameterize the case and control subclass weight regressions $\eta_k(\boldsymbol{W})$ and $\nu_k(\boldsymbol{W})$ using the same regression form $h_k(\boldsymbol{W}; \cdot)$ but different parameters.

Control subclass weight regression. We rewrite the subclass weights $\nu_k(\cdot), k = 1, \ldots, K$, using a stick-breaking parameterization. Let $g(\cdot) : \mathbb{R} \mapsto [0, 1]$ be a link function. Let α_{ik} be subject i's linear predictor at stick-breaking step $k = 1, \ldots, K-1$. Using the stick-breaking analogy, we begin with a unit-length stick, break a segment of length $g(\alpha_{i1}^{\nu})$ and continue breaking a fraction $g(\alpha_{i2}^{\nu})$ from the remaining $\{1 - g(\alpha_{i1}^{\nu})\}$ and so on; At step k, we break a fraction $g(\alpha_{ik}^{\nu})$ from what is left in the preceding $k-1$ steps resulting in the k-th stick segment k of length $\eta_{ik} = g(\alpha_{ik}^{\nu}) \prod_{s<k} \{1 - g(\alpha_{is}^{\nu})\}$; We stop until K sticks of variable lengths result. In this paper, we use the logistic function $g(\alpha) = 1/\{1 + \exp(-\alpha)\}$ which is consistent with the multinomial logit regression for $\pi_\ell(\cdot)$ so that the priors of the coefficients $\boldsymbol{\Gamma}_k^{\nu}$ and $\boldsymbol{\Gamma}_\ell^{\pi}$ can be similar (Supplementary Materials A1.2). Generalization to other link functions such as the probit function is straightforward (e.g., Rodriguez and Dunson, 2011). We use this parameterization to introduce a novel shrinkage prior on a simplex for the subclass weights $\{\nu_k(\boldsymbol{W})\}$ (see Supplementary Material A1.1) which encourages fewer than K effective subclasses, or "m-sparse" shrinkage prior on the simplex. This provides parsimonious approximation to the

conditional distribution of control measurements $\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, Y = 0, \{\nu_k(\cdot)\}, \boldsymbol{\Psi})$ using a few subclasses.

In our analysis, we use generalized additive models (Hastie and Tibshirani, 1986) for the k-th linear predictor $\alpha_{ik}^{\nu} = \alpha_k^{\nu}(\boldsymbol{W}_i = \boldsymbol{w}; \boldsymbol{\Gamma}_k^{\nu}) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(w_j; \boldsymbol{\beta}_{kj}^{\nu}) + \widetilde{\boldsymbol{w}}^{\top}\boldsymbol{\gamma}_k^{\nu}$, for $k = 1, \ldots, K-1$. We have parameterized the possibly nonlinear $f_{kj}(\cdot)$ using B-spline basis expansions with coefficients $\boldsymbol{\beta}_{kj}^{\nu}$; $\widetilde{\boldsymbol{w}}^{\top}\boldsymbol{\gamma}_k^{\nu}$ are the linear effects of a subset of predictors which can include an intercept and $\widetilde{\boldsymbol{w}}$ is a subvector of predictors $\boldsymbol{w}$; Let $\boldsymbol{\Gamma}_k^{\nu} = \{\mu_{k0}, \{\boldsymbol{\beta}_{kj}^{\nu}\}, \boldsymbol{\gamma}_k^{\nu}\}$ collect all the regression parameters. Following Lang and Brezger (2004), we constrain $\{f_{kj}, j = 1, \ldots, J\}$ to have zero means for statistical identifiability. Supplementary Material A1.2 provides the technical details about the parameterization of f_{kj}.

The subclass-specific intercepts $\{\mu_{k0}\}$ globally control the magnitudes of the linear predictors. We hence propose priors on $\{\mu_{k0}\}$ to *a priori* encourage few subclasses (see Supplementary Materials A1.1). In particular, a large positive intercept μ_{k0} makes $g(\alpha_{ik}^{\nu}) \approx 1$ and hence breaks nearly the entire remaining stick after the $(k-1)$-th stick-breaking. Since the stick-breaking parameterization one-to-one maps to a classical latent class regression model formulation for the control data, the linear predictor α_{ik}^{ν} and the sum $\mu_{k0}+\gamma_{k0}^{\nu}$ are identifiable except in a Lebesgue zero set of parameter values, or "generic identifiability" (Huang and Bandeen-Roche, 2004). Consequently, even if the intercept μ_{k0} is not statistically identified if $\widetilde{\boldsymbol{w}}$ includes an intercept γ_{k0}^{ν}, the MCMC samples of the statistically identifiable functions can provide valid posterior inferences (Carlin and Louis, 2009). We write the control likelihood with covariates $\boldsymbol{W}_i$ as $L_0^{\mathsf{reg}} = \prod_{i:Y_i=0} \sum_{k=1}^{K} h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\nu}) \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$. Supplementary Materials A2 provides further remarks on the assumption for introducing covariates into the control model.

Case subclass weight regression. The subclass weight regression functions for cases $\{\eta_k(\boldsymbol{W})\}$ are also specified via a logistic stick-breaking regression as in the controls but with different

parameters: $\eta_{ik} = g(\alpha^{\eta}_{ik}) \prod_{s<k}\{1 - g(\alpha^{\eta}_{is})\}$, $k = 1, \ldots, K-1$. Since given the TPRs and the FPRs, the subclass weights fully determine the joint distribution $[\boldsymbol{M} \mid \boldsymbol{W}, I = \ell \neq 0]$ hence the measurement dependence in each class, we let $\eta_k(\boldsymbol{w})$ and $\nu_k(\boldsymbol{w})$ be different between cases and controls for any $\boldsymbol{w}$.

Let the k-th linear predictor $\alpha^{\eta}_{ik} = \alpha^{\eta}_{k}(\boldsymbol{W}_i = \boldsymbol{w}; \boldsymbol{\Gamma}^{\eta}_k) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(w_j; \boldsymbol{\beta}^{\eta}_{kj}) + \widetilde{\boldsymbol{w}}^{\top}\boldsymbol{\gamma}^{\eta}_k$, where $\boldsymbol{\Gamma}^{\eta}_k = \{\mu_{k0}, \{\boldsymbol{\beta}^{\eta}_{kj}\}, \boldsymbol{\gamma}^{\eta}_k\}$ are the regression parameters that differ from the control counterpart ($\boldsymbol{\Gamma}^{\nu}_k$). In particular, we approximate $f_{kj}(\cdot), j = 1, \ldots, J$, here using the same set of B-spline basis functions as in the controls but estimate a different set of basis coefficients $\boldsymbol{\beta}^{\eta}_{kj}$. In addition, we have directly used the intercepts $\{\mu_{k0}\}$ from the control model 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 $\eta_{k+1} \approx 0$.

Combining the case (L^{reg}_1) and control likelihood (L^{reg}_0) with covariates, we obtain the joint likelihood for the regression model $L^{\mathsf{reg}} = L^{\mathsf{reg}}_1 \times L^{\mathsf{reg}}_0$.

Remark 1: Under an assumption (A1): the case subclass weights are constant over covariates: $\eta_k(\cdot) \equiv \eta_k$, $k = 1, \ldots, K$, the regression model reduces to an npLCM model without covariates upon integration over a distribution of covariates $\boldsymbol{X}$. To see this, the case and control likelihood functions 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 $L^{*}_0 = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu^{*}_{k} \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$, respectively; Here $\pi^{*}_{\ell} = \int \pi_{\ell}(\boldsymbol{X}) \mathrm{d}G(\boldsymbol{X})$ and $\nu^{*}_{k} = \int \nu_k(\boldsymbol{W}) \mathrm{d}H(\boldsymbol{W})$ where G and H are probability or empirical distributions of $\boldsymbol{X}$ and $\boldsymbol{W}$, respectively. The mathematical equivalence enables valid inference about the overall PEFs $\boldsymbol{\pi}^{*}$ omitting $\boldsymbol{X}$ and $\boldsymbol{W}$ (see Supplementary Materials A5.2 for an example). The no-covariate analysis becomes deficient under deviations from (A1); Section 4 provides examples.

3.2.1 *Priors and Posterior Inference.* The unknown parameters include the coefficients in the etiology regression ($\{\boldsymbol{\Gamma}^{\pi}_{\ell}\}$), the subclass mixing weight regression for the cases ($\{\boldsymbol{\Gamma}^{\eta}_{k}\}$) and

the controls ($\{\boldsymbol{\Gamma}_k^{\nu}\}$), the true and false positive rates ($\boldsymbol{\Theta} = \{\theta_k^{(j)}\}, \boldsymbol{\Psi} = \{\psi_k^{(j)}\}$). With typical samples sizes about 500 controls and 500 cases in each study site, the number of parameters in controls likelihood L_0 ($> JKCp$) easily exceeds the number of distinct binary measurement patterns observed. To overcome potential overfitting and increase model interpretability, we *a priori* place substantial probabilities on models with the following two features: (a) Few non-trivial subclasses via a novel additive half-Cauchy prior for the intercepts $\{\mu_{k0}\}$, and (b) for a continuous variable, smooth regression curves $\pi_\ell(\cdot)$, $\nu_k(\cdot)$ and $\eta_k(\cdot)$ by Bayesian Penalized-splines (P-splines) (Lang and Brezger, 2004) combined with shrinkage priors on the spline coefficients (Ni et al., 2015) to encourage towards constant values, $\eta_k(\cdot) = \eta_k, \nu_k(\cdot) = \nu_k, k = 1, \ldots, K$, which reduces to the original npLCM. Supplementary Material A1 details the prior specifications.

We use the Markov chain Monte Carlo (MCMC) algorithm to draw 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. For example, the posterior samples of the case positive rate curve for pathogen j help evaluate model fit. The red bands in Row 1 of Figure 1 are posterior 95% credible bands obtained by substituting relevant parameters with their sampled values across MCMC iterations in $\mathbb{P}(M_\ell = 1 \mid \boldsymbol{x}, \boldsymbol{w}, Y = 1) = \pi_\ell(\boldsymbol{w}; \boldsymbol{\Gamma}_\ell^{\pi}) \sum_{k=1}^{K} h_k(\boldsymbol{w}; \boldsymbol{\Gamma}_k^{\eta}) \theta_k^{(\ell)} + \{1 - \pi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^{\pi})\} \sum_{k=1}^{K} h_k(\boldsymbol{w}; \boldsymbol{\Gamma}_k^{\eta}) \psi_k^{(\ell)}$. The npLCMs with or without covariates are fitted using a free and publicly available R package `baker` (`https://github.com/zhenkewu/baker`). `Baker` calls an external automatic Bayesian model fitting software `JAGS 4.2.0` (Plummer et al., 2003) from within R and provides functions to visualize the posterior distributions of the unknowns (e.g., the PEFs and cases' latent disease class indicators) and perform posterior predictive model checking (Gelman et al., 1996). Supplementary Materials A3 details the convergence diagnostics.

4. Simulations

We simulate case-control bronze-standard (BrS) measurements along with observed continuous and/or discrete covariates under multiple combinations of true model parameter values and sample sizes that mimic the motivating PERCH study. In Simulation I, we illustrate flexible statistical inferences about the PEF functions $\{\pi_\ell(\cdot)\}$. In Simulation II, we focus on the overall PEFs that quantify the overall cause-specific disease burdens in a population which are of policy interest. 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 covariate (Little et al., 2011). Regression analyses reduce estimation bias, retain efficiency and provide more valid frequentist coverage of the 95% CrIs. The relative advantage varies by the true data generating mechanism and sample sizes.

In all analyses here, we use a working number of K^* subclasses, with 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; We specify Beta(1,1) for the identifiable FPRs. The priors for the regression coefficients follow the specifications in Supplementary Materials A1.

Simulation I. We demonstrate that the inferential algorithm recovers the true PEF functions $\{\pi^0_\ell(\boldsymbol{X})\}$. 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. Supplementary Materials A4 specifies the true data generating mechanism and the regression specifications. Based on the simulated data, pathogen A has a bimodal positive rate curve mimicking the trends observed of RSV in one PERCH site; other pathogens have overall increasing 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 (impossible under the more restrictive pLCM). Row 2 of Figure 1 visualizes for the 9 causes (by column), the posterior means (thin black line) and 95% CrIs (gray bands) for the etiology regression

curves $\pi_\ell(\cdot)$ are close to the simulation truths $\pi_\ell^0(\cdot)$. Supplementary Materials A4 provides additional simulation results to assess the recovery of the true $\pi_\ell^0(X)$ for a discrete covariate X.

[Figure 1 about here.]

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 PEFs ($\boldsymbol{\pi}^*$) 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 of 48 scenarios detailed in Supplementary Materials A4 that correspond to distinct numbers of causes, sample sizes, relative sizes of PEF functions (rare versus popular etiologies), signal strengths (more discrepant TPRs and FPRs indicate stronger signals, Wu et al. (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 $\pi_\ell^{0*} = N_1^{-1} \sum_{i:Y_i=1} \pi_\ell^0(\boldsymbol{X}_i)$ is the true overall PEF, and $\widehat{\pi_\ell^*} = N_1^{-1} \sum_{i:Y_i=1} \widehat{\pi}_\ell(\boldsymbol{X}_i)$ is an empirical average of the posterior mean PEFs at $\boldsymbol{X}_i$. We also evaluate the empirical coverage rates of the 95% CrIs.

[Figure 2 about here.]

The regression model incorporates covariates and performs better in estimating $\boldsymbol{\pi}^*$ than a model omitting covariates. For example, Figure 2(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 2(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

Supplementary Materials A5 on the role of informative TPR priors. Figure 2(b) regression analyses also produce 95% CrIs for π_ℓ^* that have more valid empirical coverage rates in all scenarios. Misspecified models without covariates concentrate the posterior distribution away from the true overall PEFs, resulting in large biases that dominate the posterior uncertainty of π_ℓ^* which is evident from the more severe undercoverages with higher TPRs and lower FPRs (row 3 and 4 versus row 1 and 2, Figure 2).

5. Regression Analysis of PERCH Data

We restrict attention in this regression analysis to 494 cases and 944 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), disease severity for cases (severe or very severe), HIV status (positive or negative) and presence or absence of seven species of pathogens (five viruses and two bacteria, representing a subset of pathogens evaluated) in nasopharyngeal (NP) specimens tested with polymerase chain reaction (PCR), or NPPCR (bronze-standard, BrS); We also include in the analysis the blood culture (BCX, silver-standard, SS) 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 case and control frequencies by age, disease severity and HIV status. The two strata with the most subjects are severe pneumonia children who were HIV negative and under or above one year of age. Some low or zero cell counts preclude fitting npLCMs by stratum. Regression models with additive assumptions among the covariates can borrow information across strata and stabilize the PEF estimates. Supplemental Figure S5 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 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 et al. (2017), we integrate case-only SS measurements for HINF and PNEU by using informative priors of the sensitivities (e.g., from vaccine probe studies e.g., Feikin et al. (2014)) to adjust the PEF 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 PEF estimates relative to the ones obtained from an NPPCR-only analysis.

We include in the regression analysis a cause "Not Specified (NoS)" to account for true pathogen causes other than the seven pathogens. 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 specify the Beta(7.59,58.97) prior for the two TPRs of SS measurements similarly but with a lower range of $5 - 20\%$. We use a working number of subclasses $K = 5$. In the etiology 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, very severe pneumonia, HIV positive; In the subclass weight regression model $h_k(\boldsymbol{W};\cdot)$, we use 5 d.f. for the standardized enrollment date t with uniform knots and two indicators for age older than one and HIV positive. The prior distributions for the etiology and subclass weight regression parameters follow the specification in Supplementary Materials A1.

[Table 1 about here.]

The regression analysis produces seasonal estimates of the PEF function for each cause that varies in trend and magnitude among the eight strata defined by age, disease severity and HIV status. Figure 3 shows among two age-HIV-severity strata the posterior mean curve and 95% pointwise credible bands of the etiology regression functions $\pi_\ell(t, \mathsf{age}, \mathsf{severity}, \mathsf{HIV})$ as a function of t. For example, among the younger, HIV negative and severe pneumonia children (Figure 3(a)), the PEF 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). Other single-pathogen causes HINF, PNEU, ADENO, HMPV_A_B and PARA_1 have overall low and stable PEF curves across seasons. The estimated PEF curve of NoS shows a trend with a higher level of uncertainty that is complementary to RSV because given any enrollment date the PEFs of all the causes sum to one. In contrast, Figure 3(b) shows a lower degree of seasonal variation of RSV PEF curve among the older, HIV negative and severe pneumonia children.

[Figure 3 about here.]

The regression model accounts for stratification of etiology by the observed covariates and assigns cause-specific probabilities for two cases who have identical measurements but different covariate values. Supplemental Figure S6 shows for two cases with all negative NPPCR results (the most frequent pattern among cases), the older case has a lower posterior probability of her disease caused by RSV and higher probability of being caused by NoS. Indeed, contrasting older and younger children while holding the enrollment date, HIV,

severity constant, the estimated difference in the log odds (i.e., log odds ratio) of a child being caused by RSV versus NoS is negative: -1.82 (95% CrI : $-2.99, -0.77$).

Given age, severity and HIV status, we quantify the overall cause-specific disease burdens $\boldsymbol{\pi}^*$ by averaging the PEF function estimates by the empirical distribution of the enrollment dates. Contrasting the results in the two age-severity-HIV strata in Figure 3(a) and 3(b), since the case positive rate of RSV among the older children reduces from 39.3% to 17.9% but the control positive rates remain similar (from 3.0% to 4.1%), the overall PEF of RSV (π^*_{RSV}) decreases from 47.7 (95% CrI : $37.6, 61.5$)% to 17.3 (95% CrI : $8.0, 29.1$)% and attributing a higher total fraction of cases to NoS (π^*_{NoS}) from 37.6 (95% CrI : $20.3, 51.9$)% to 56.1 (95% CrI : $29.5, 79.3$)%; The overall PEFs for other causes remain similar.

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 and individual etiologic fractions. While the existing methods 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 PEFs. This paper addresses this analytic need by extending npLCM to a general regression modeling framework using case-control multivariate binary data to estimate disease etiology.

The proposed approach has three distinguishing features: 1) It allows analysts to specify a model for the functional dependence of the PEFs upon important covariates. And with assumptions such as additivity, we can improve estimation stability for sparsely populated strata defined by many discrete covariates. 2) The model incorporate control data for the inference of PEF curve. The posterior inferential algorithm estimates a parsimonious covariate-dependent reference distribution of the diagnostic measurements from controls. Finally, 3) the model uses informative priors of the sensitivities (TPRs) only once in a population for

which these priors were elicited. Relative to a fully-stratified npLCM analysis that reuses these priors, the proposed regression analysis avoids overly-optimistic etiology uncertainty estimates.

We have shown by 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 omitting covariates. In addition, the proposed regression analysis can readily integrate multiple sources of diagnostic measurements of distinct levels of diagnostic sensitivities and specificities, a subset of which are only available from cases (SS data), to further reduce the posterior uncertainty of the etiology estimates. Our regression analysis integrates the BrS and SS data from one PERCH site and reveals prominent dependence of the PEFs upon seasonality and a pneumonia child's age, HIV status and disease severity.

Future work may improve 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 et al., 2008). Third, when the subsets of pathogens that have caused the diseases in the population is unknown, the proposed method can be combined with subset selection procedures (Wu et al., 2019; Gu and Xu, 2019a). Finally, scalable posterior inference for multinomial regression parameters (e.g., Zhang and Zhou, 2017) will likely improve the computational speed in the presence of a large number of disease classes and covariates.

Supplementary Materials

The supplementary materials contain the technical details on prior specifications, a remark, additional simulation results and supplemental figures referenced in Main Paper.

Acknowledgment

We thank the PERCH study team led by Kathernine O'Brien for providing the data and scientific advice, Scott Zeger, Maria Deloria-Knoll, Christine Prosperi and Qiyuan Shi for insightful comments and valuable feedback about `baker` and Jing Chu for preliminary simulations. The research was partly supported by the Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1408-20318, ZW), NIH grants P30CA046592 (National Cancer Institute Cancer Center Support Grant Development Funds, Rogel Cancer Center; ZW and IC), U01CA229437 (ZW, IC) and an Investigator Award from Precision Health Initiative and an MCubed Award from University of Michigan (ZW).

References

Carlin, B. and Louis, T. (2009). *Bayesian methods for data analysis*, volume 78. Chapman & Hall/CRC.

Crawley, J., Prosperi, C., Baggett, H. C., Brooks, W. A., Deloria Knoll, M., Hammitt, L. L., Howie, S. R., Kotloff, K. L., Levine, O. S., Madhi, S. A., et al. (2017). Standardization of clinical assessment and sample collection across all perch study sites. *Clinical infectious diseases* **64,** S228–S237.

Dunson, D. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 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,** 346.

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

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* pages 398–409.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6,** 733–760.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

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

Gu, Y. and Xu, G. (2019b). Partial identifiability of restricted latent class models. *Annals of Statistics* page In press.

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*, volume 140. CRC Press.

Gustafson, P., Lefebvre, G., et al. (2008). Bayesian multinomial regression with class-specific predictor selection. *The Annals of Applied Statistics* **2,** 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,** S197–S204.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1,** 297–318.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* **69,** 5–32.

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,** 855–863.

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., et al. (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,**

209–222.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics* **13,** 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, pages 362–412. Princeton, NJ: Princeton University Press.

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

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

Nair, H., Brooks, W. A., Katz, M., Roca, A., Berkley, J. A., Madhi, S. A., Simmerman, J. M., Gordon, A., Sato, M., Howie, S., et al. (2011). Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet* **378,** 1917–1930.

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

Obando-Pacheco, P., Justicia-Grande, A. J., Rivero-Calle, I., Rodríguez-Tenreiro, C., Sly, P., Ramilo, O., Mejías, A., Baraldi, E., Papadopoulos, N. G., Nair, H., et al. (2018). Respiratory syncytial virus seasonality: a global overview. *The Journal of infectious diseases* **217,** 1356–1364.

PERCH Study Group (2019). Aetiology of severe hospitalised pneumonia in hiv-uninfected children from africa and asia: the pneumonia aetiology research for child health (perch) case-control study. *Lancet* .

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed*

Statistical Computing, volume 124.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian analysis (Online)* **6,**.

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

Scott, J. A. G., Brooks, W. A., Peiris, J. M., Holtzman, D., and Mulhollan, E. K. (2008). Pneumonia research to reduce childhood mortality in the developing world. *The Journal of clinical investigation* **118,** 1291.

Wu, Z., Casciola-Rosen, L., Rosen, A., and Zeger, S. L. (2019). A bayesian approach to restricted latent class models for scientifically-structured clustering of multivariate binary outcomes. *arXiv preprint arXiv:1808.08326* .

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., Zeger, S. L., and for Child Health Core Team, P. E. R. (2016). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65,** 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.

Zhang, Q. and Zhou, M. (2017). Permuted and augmented stick-breaking bayesian multinomial regression. *The Journal of Machine Learning Research* **18,** 7479–7511.

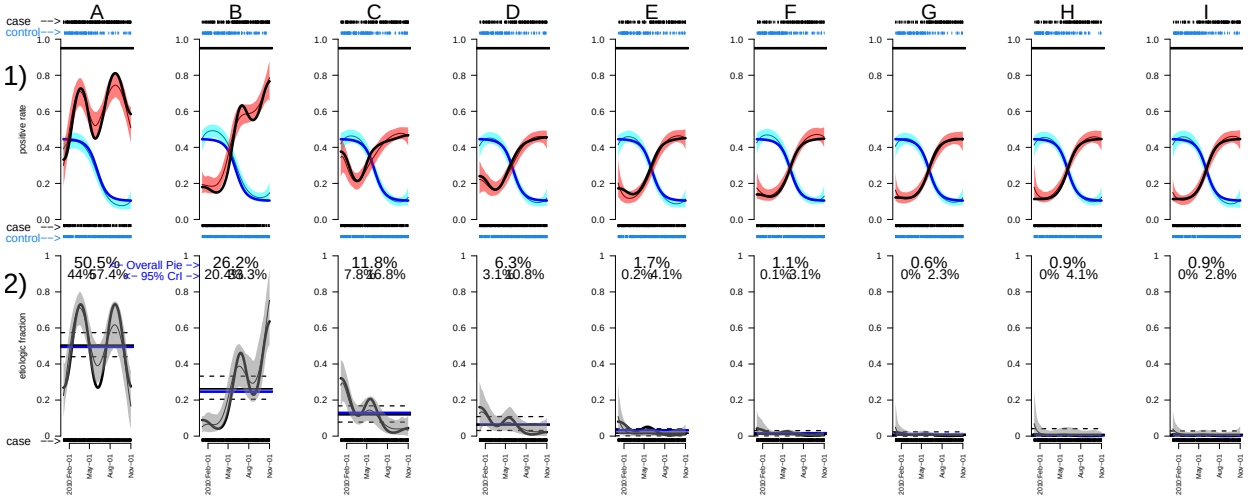
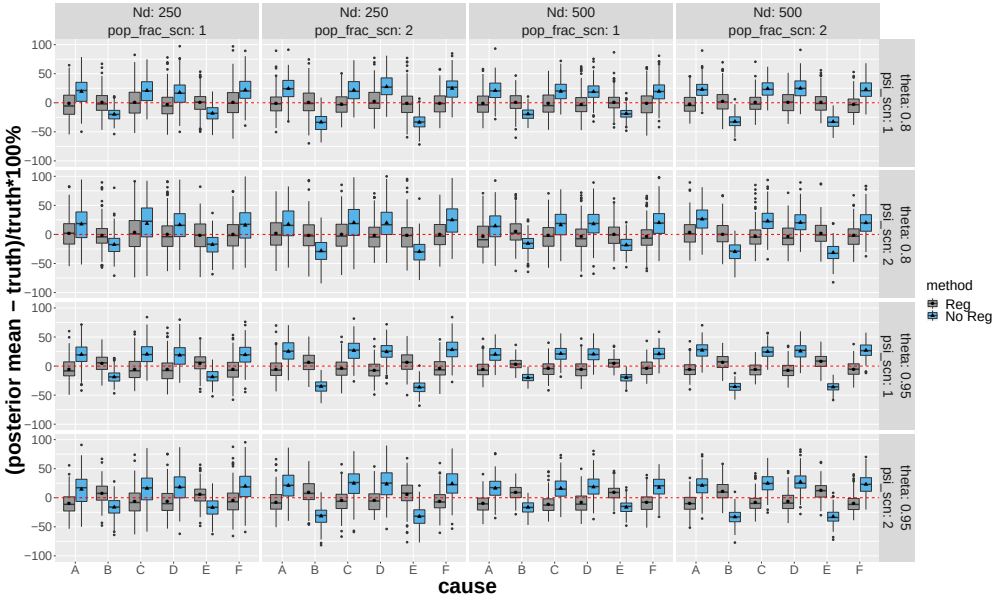
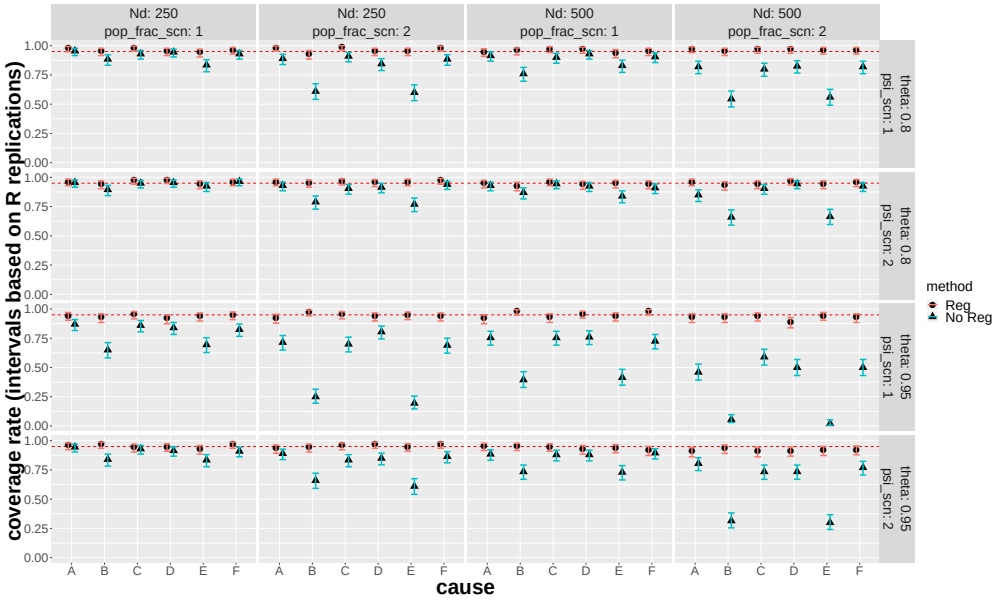


Figure 1: Row 2) For each of the 9 causes (by column) in Simulation I, the posterior mean (thin black curves) and pointwise 95% credible bands (gray bands) for the etiology regression curves $\pi_\ell(x)$ are close to the simulation truths $\pi_\ell^0(x)$. In row 1), the fitted case (red) and control (blue) positive rate curves are shown with the posterior mean curves (solid black curves) and pointwise 95% credible bands (shaded); 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.

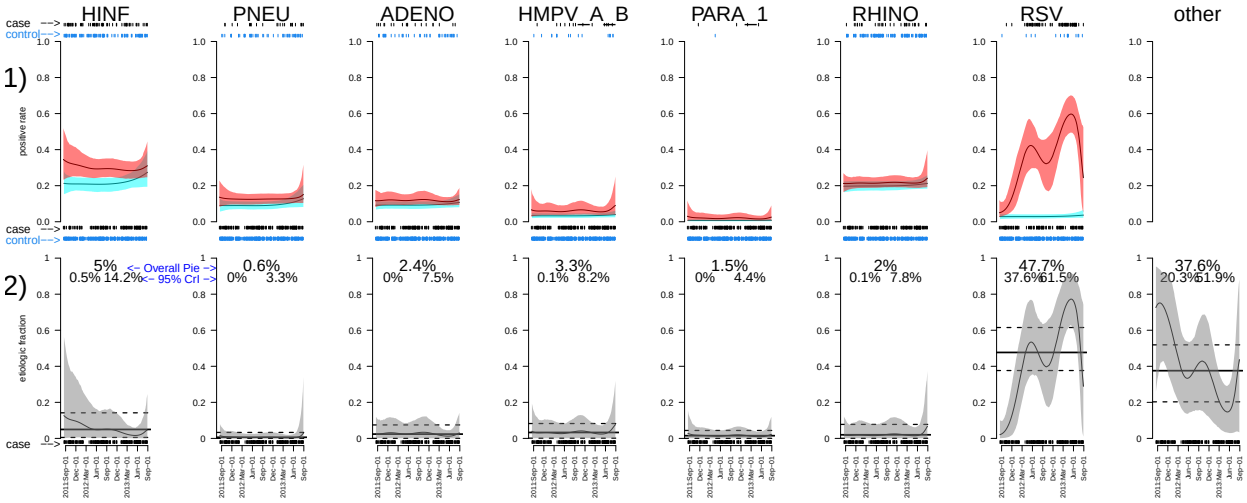


(a) Percent relative bias

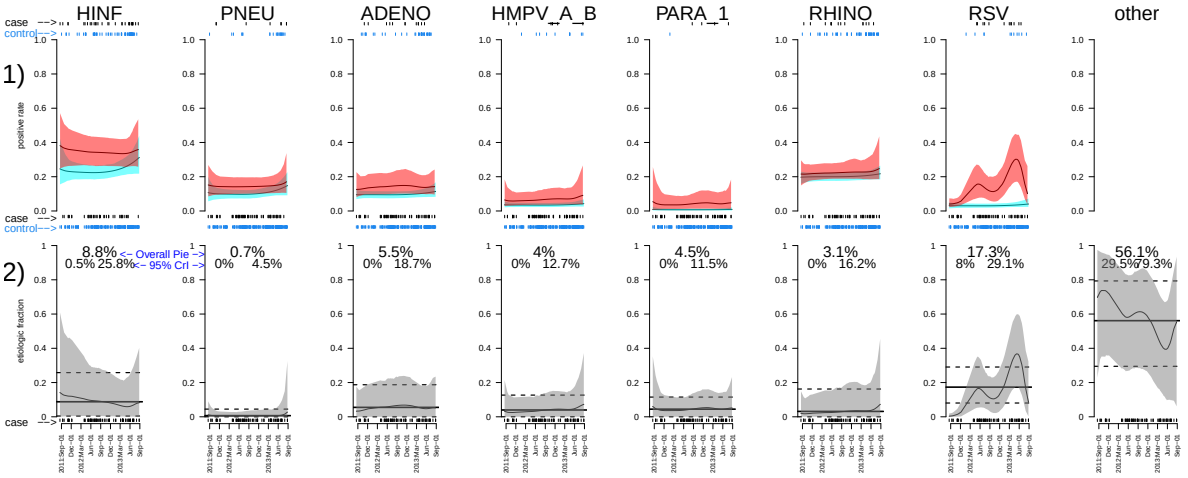


(b) Empirical coverage rates

Figure 2: The regression analyses produce less biased posterior mean estimates and more valid empirical coverage rates for π_ℓ^* over $R = 200$ replications in Simulation II with $J = 6$. 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 PEF π_ℓ^* for six causes (A - F); The red horizontal dashed lines indicate zero bias. *Bottom*) Each dot or triangle indicates the empirical coverage rate of the 95% CrIs produced by analyses with regression (•) or without regression (▲); The nominal 95% rate is marked by horizontal red dashed lines. Since each coverage rate for π_ℓ^* is computed from $R = 200$ binary observations, the truth being covered or not, a 95% CI is also shown.



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



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

Figure 3: Estimated seasonal PEF $\widehat{\pi}_\ell(\mathsf{date}, \mathsf{age}, \mathsf{severity}, \mathsf{HIV})$ for two most prevalent age-severity-HIV strata: **younger** (a) or **older** (b) than one, with severe pneumonia, HIV negative; Here the results are obtained from a model assuming seven single-pathogen causes (HINF, PNEU, ADENO, HMPV.A.B, PARA.1, RHINO, RSV) and an "Not Specified" cause. In an age-severity-HIV stratum and for each cause ℓ:
Row 2) shows the temporal trend of $\widehat{\pi}_\ell$ which is enveloped by pointwise 95% credible bands shown in gray. The estimated overall PEF $\widehat{\pi^*_\ell}$ averaged among cases in the present stratum is shown by a horizontal solid line, below and above which are two dashed black lines indicating the 2.5% and 97.5% posterior quantiles. 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 count (frequency) of cases and controls by age, disease severity and HIV status (1: yes; 0: no). The marginal fractions among cases and controls for each covariate are shown at the bottom. Results from the regression analyses are shown in Figure 3 for the first two strata.

age $\geqslant 1$	very severe (VS) (case-only)	HIV positive	# cases (%) total: 524 (100)	# controls (%) total: 964 (100)
0	0	0	208 (39.7)	545 (56.5)
1	0	0	72 (13.7)	278 (28.8)
0	1	0	116 (22.1)	0
1	1	0	33 (6.3)	0
0	0	1	37 (7.1)	85 (8.8)
1	0	1	24 (4.5)	51 (5.3)
0	1	1	25 (4.8)	0
1	1	1	3 (0.6)	0
case: 25.2%	34.5%	17.0%		
control: 34.3%	-	14.1%		