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

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

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

Abstract

In large-scale disease etiology studies, epidemiologists often need to use multiple imperfect binary measures of the unobserved true causes of disease to estimate the cause-specific case fractions, or "population etiologic fractions" (PEFs). Despite recent advances in statistical methods, the scientific need of estimating the effect of explanatory variables upon the PEFs in the presence of control data remains unmet. In this paper, we start with and extend the nested partially-latent class model (npLCM, Wu et al., 2016b) to a general framework for etiology regression analysis in case-control studies. Data from controls provide requisite information about measurement specificities and covariations 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 dependence structures. We use Markov chain Monte Carlo for posterior inference of the PEF functions, cases' latent classes and the overall PEFs of policy import. We illustrate the regression analysis via 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.

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

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). Knowing which pathogen has caused a pneumonia case is crucial for choosing effective treatment. Knowing the distribution of infecting pathogens in a pneumonia population in each region and stratum informs prioritizing vaccine development and manufacture.

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 (Levine et al., 2012). 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 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 samples and tests peripheral compartments including the blood, sputum, pleural fluid and nasopharyngeal (NP) cavity and determines the presence or absence of a list of pathogens in each sample (Crawley et al., 2017). In this paper, we focus on measurements obtained from two specimen-test pairs: NP Polymerase Chain Reaction (NPPCR) results from cases and controls and blood culture (BCX) results from cases only.

Valid inference about the population and individual etiologic fractions must address three characteristics of the measurements. First, imperfect diagnostic specificities may result in

the detection of multiple pathogens. For example, NPPCR may detect multiple colonizing pathogens in the nasal cavity that are not causes of a case's pneumonia. Determining the primary causative agent(s) among the detected pathogen(s) requires statistical controls. Second, multiple specimens are tested among the cases with only a subset available from controls. Third, tests with imperfect sensitivity such as NPPCR and BCX may miss true causative agent(s) which if unadjusted may underestimate the PEFs. Other large-scale disease etiology studies have raised similar analytic needs and challenges of integrating multiple sources of imperfect measurements to draw inference about complex disease etiology (e.g., Saha et al., 2018; Kotloff et al., 2013).

Wu et al. (2016a) introduced a *partially-latent class model* (pLCM) as an extension to classical latent class models (LCMs Lazarsfeld, 1950; Goodman, 1974) for using 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 $\{\boldsymbol{M}_i = (M_{i1}, \ldots, M_{iJ})^\top\}$ 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. 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 $\mathcal{S}_{L-1} = \{\boldsymbol{\pi} : \sum_{\ell=1}^{L} \pi_\ell = 1, 0 \leq \pi_\ell \leq 1\}$. 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 specifies the control distribution using J "false positive rates" (FPRs) because controls have no lung infection ($I_i = 0$). Given a case class $\ell = 1, \ldots, L$, a pLCM lets the conditional distribution of the diagnostic measures be completely characterized by a vector of J response probabilities, a subset of which may differ from the controls: each causative pathogen is observed with a higher probability in case class ℓ (sensitivity or true positive rate, TPR) than among the controls and a non-causative pathogen is observed with the same probability as in the controls (1 - specificity or FPR). Under pLCM, a higher observed marginal positive rate for pathogen j among cases than controls indicates etiologic importance.

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 able to estimate $\boldsymbol{\pi}$ using case-control measurements with imperfect sensitivities and specificities, referred to as "bronze-standard" (BrS) data. Case-only measurements from sterile sites with imperfect sensitivity but perfect specificity ("silver-standard", SS) can also be incorporated. 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, although bounds on the parameters may be available (e.g., Wu et al., 2016a, 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 makes a "local independence" (LI) assumption that the BrS measurements are stochastically independent given a subject's class membership. This classical assumption is central to mixture models for multivariate data, because the estimation procedures essentially find the optimal partition of observations so that the LI approximately holds in each subgroup. However, we observed pairwise correlations among the NPPCR measures from the controls. Deviations from LI, or "local dependence" (LD) can be directly modeled by an extension of pLCM, called *nested partially-latent class model* (npLCM, Wu et al., 2016b). The npLCM is motivated by the capacity of the classical LCM formulation to describe the complex dependence among discrete data (Dunson and Xing, 2009). It assumes the within-class correlations among BrS tests are induced by unobserved heterogeneity in subjects' propensities for pathogens colonizing their nasal cavities. In particular, LD is induced in an npLCM by nesting a small number of latent subclasses $Z = 1, \ldots, K$, within each class $\ell = 0, 1, ..., L$, where subclasses respond with distinct vectors of probabilities. In a Bayesian framework with stick-breaking process priors for Z 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., 2016a), which can be directly

estimated from 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 is more likely a colonization than indicating the cause of lung infection during seasons with high background colonization rates, and much less so when the pathogen is rarely found in controls.

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., 2016b), necessitating multiple sets of *independent* informative priors across stratum. 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 $\boldsymbol{\pi}^* = (\boldsymbol{\pi}_1^*, \ldots, \boldsymbol{\pi}_L^*)^\top$ 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. Correct assessment of the posterior uncertainty then must "spread" a set of informative priors across strata or adjust up the posterior standard deviations of $\boldsymbol{\pi}^*$ obtained from a fully-stratified analysis. Neither is ideal given their *ad hoc* nature.

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. The proposed approach 1) allows users to specify parsimonious (e.g., additive on the linear predictor scale) functional dependence of $\boldsymbol{\pi}$ upon covariates, and 2) correctly assesses the posterior uncertainty of the PEF functions and the overall PEFs $\boldsymbol{\pi}^*$ by applying the TPR priors just once.

We fit the model using Markov chain Monte Carlo (MCMC) which simulates correlated samples of the unknown parameters to approximate their posterior distributions (Gelfand and Smith, 1990). The inferential algorithms for the family of npLCMs with or without covariates are implemented in a free and publicly available `R` package `baker` at `https://github.com/zhenkewu/baker`.

The rest of the paper is organized as follows. Section 2 overviews the npLCM with-

out 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 simulation studies 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 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 where each tested pathogen corresponds to a cause (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., 2016b).

The likelihood function for an npLCM has three components:

- PEFs $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^\top = \{\pi_\ell = \mathbb{P}(I = \ell \mid Y = 1), \ell = 1, \ldots, L\} \in \mathcal{S}_{L-1}$: cause-specific case fractions;
- $\boldsymbol{P}_{1\ell} = \{\boldsymbol{P}_{1\ell}(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = \ell, Y = 1)\}$: a table of probabilities of making J binary observations $\boldsymbol{M} = \boldsymbol{m}$ in a case class $\ell \neq 0$;
- $\boldsymbol{P}_0 = \{\boldsymbol{P}_0(\boldsymbol{m})\} = \{\mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid I = 0, Y = 0)\}$: the probability table above 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 L disease classes:

$$\boldsymbol{P}_1 = \sum_{\ell=1}^{L} \pi_\ell \boldsymbol{P}_{1\ell}. \tag{1}$$

$\boldsymbol{P}_0$, $\boldsymbol{P}_{1\ell}$ and $\boldsymbol{P}_1$ are three different probability tables with 2^J rows where each row specifies the probability of making binary observations $\boldsymbol{M} = \boldsymbol{m}$ among controls, ℓ-th case disease class and among all cases, respectively. We write the likelihood function as a product of case (L_1) and control (L_0) likelihood functions

$$L = L_1 \cdot L_0 = \prod_{i:Y_i=1} \boldsymbol{P}_1(\boldsymbol{M}_i; \boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\eta}) \times \prod_{i':Y_{i'}=0} \boldsymbol{P}_0(\boldsymbol{M}_{i'}; \boldsymbol{\Psi}, \boldsymbol{\nu}) \tag{2}$$

and seek to infer $\boldsymbol{\pi}$ from data $\{(\boldsymbol{M}_i, Y_i), i = 1, \ldots, N\}$ (Wu et al., 2016b); Here $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ are sensitivity and specificity parameters necessary for modeling the imperfect measurements. The rest of parameters $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^\top$ in L_0 and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^\top$ in L_1 are used to induce residual measurement correlations given a control or disease class.

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 the following, under $\nu_1 = \eta_1 = 1$ under LI; $\nu_1, \eta_1 \in (0, 1)$ otherwise.

PLCM. The original pLCM (Wu et al., 2016a) assumes LI so that any $\boldsymbol{P}_0(\boldsymbol{m})$ is a product of J probabilities: $\boldsymbol{P}_0(\boldsymbol{m}) = \prod_{j=1}^J \{\psi_j\}^{m_j}\{1-\psi_j\}^{1-m_j}$. The parameters $\boldsymbol{\psi} = \{\psi_j\}$ represent the positive rates absent disease and are termed "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.

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., 2016b). 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 1) 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 2) on the non-causative pathogens $\mathbb{P}(\boldsymbol{M}_{i[-\ell]} \mid I_i = \ell, Y_i = 1, \boldsymbol{\psi}_{[-\ell]}) = \prod_{j\neq\ell}\{\psi_j\}^{M_j}\{1-\psi_j\}^{1-M_j}$, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_L)^\top$, and $\boldsymbol{a}_{[-\ell]}$ 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; $\boldsymbol{\psi}_{[-\ell]}$ are a subset of the FPRs $\boldsymbol{\psi}$ that specify $\boldsymbol{P}_0$. Under the single-pathogen-cause assumption, pLCM uses J TPRs $\boldsymbol{\theta}$ for

$L = J$ causes and J FPRs $\boldsymbol{\psi}$.

The posterior inferential algorithm based on the pLCM optimally clusters cases into L subgroups so that LI approximately holds in each subgroup. Relative to the controls, the causative pathogens are observed with higher positive rates in each subgroup. We estimate each case's individual etiologic fractions (IEFs) by the empirical frequencies with which a case's disease class indicator $I_i = \ell$, $\ell = 1, \ldots, L$, in the posterior samples; The PEFs are approximately an average of the IEFs.

NPLCM. 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., 2016b). 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).

The npLCM introduces K subclasses for the controls and for each of the L disease classes among cases; 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 $\boldsymbol{P}_0^{(k)}(\boldsymbol{m}) = \mathbb{P}(\boldsymbol{M} = \boldsymbol{m} \mid Z = k, I = 0, Y = 0, \{\psi_k^{(j)}\}) = \prod_{j=1}^{J} \{\psi_k^{(j)}\}^{m_j} \{1 - \psi_k^{(j)}\}^{1-m_j}$; We use a J by K FPR matrix $\boldsymbol{\Psi} = \{\psi_k^{(j)}\}$ to represent the FPRs. Since we do not observe controls' subclasses, $\boldsymbol{P}_0$ is a weighted average of $\boldsymbol{P}_0^{(k)}$ according to the subclass probabilities $\{\nu_k\}$: $\boldsymbol{P}_0 = \sum_k^K \nu_k \boldsymbol{P}_0^{(k)}$.

The model specification for a case subclass in the npLCM follows the pLCM by assuming $\boldsymbol{P}_{1l}^{(k)} = \mathbb{P}(\boldsymbol{M} \mid Z = k, I = \ell, Y = 1)$, $\ell \neq 0$, the probability of observing $\boldsymbol{M}$ in subclass k in disease class ℓ, to be a product of the probabilities of making an observation 1) 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 2) on non-causative pathogens $\mathbb{P}(\boldsymbol{M}_{[-\ell]} \mid Y = 1, Z = k, I = \ell, \boldsymbol{\Psi}_k^{([-\ell])}) = \prod_{j \neq \ell} \{\psi_k^{(j)}\}^{m_j} \{1 - \psi_k^{(j)}\}^{1-m_j}$. We denote the TPRs by a J by K TPR matrix $\boldsymbol{\Theta} = \{\theta_k^{(j)}\}$. To simplify notation, we summarize the preceding specification by letting $\boldsymbol{P}_{1l}^{(k)} = \Pi(\boldsymbol{M}; \boldsymbol{p}_{k\ell}), \ell \neq 0$, 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 $\boldsymbol{s} = (s_1, \ldots, s_J)^\top$, $0 \leq s_j \leq 1$; The column 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 or 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. Since cases' subclasses are unobserved, we obtain $\boldsymbol{P}_{1\ell}$ as a weighted average of $\boldsymbol{P}_{1\ell}^{(k)}$ according the case subclass weights $\{\boldsymbol{\eta}_k\}$: $\boldsymbol{P}_{1\ell} = \sum_{k=1}^{K} \eta_k \boldsymbol{P}_{1\ell}^{(k)}$. Setting $\nu_1 = \eta_1 = 1$ and $\nu_k = \eta_k = 0, k \geq 2$ in the npLCM likelihood function (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. Different from pLCM, the subclass mixing weights may differ between the cases ($\boldsymbol{\eta}$) and the controls ($\boldsymbol{\nu}$). The special case of identical subclass mixing weights 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 the 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 $\boldsymbol{X}_1$ 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 $\boldsymbol{W}_i$; $\boldsymbol{X}_i Y_i = \boldsymbol{X}_i$ for a case subject. For example, healthy controls have no disease severity information. We let three sets of parameters in an npLCM (2) depend on the observed covariates: 1) the etiology regression function $\{\pi_\ell(\boldsymbol{X}_i)\}$ among cases which is of primary scientific interest, 2) the subclass weights in the case likelihood $\{\eta_k(\boldsymbol{W}_i)\}$, 3) the subclass weights in the control likelihood $\{\nu_k(\boldsymbol{W}_i)\}$.

Regression model overview. Firstly, we let the PEFs $\{\pi_\ell\}$ vary by covariates $\boldsymbol{X}$, e.g., through an additive multinomial logistic regression. Secondly, controls ($I = 0$) provide the covariate-dependent reference distribution $[\boldsymbol{M} \mid \boldsymbol{W}, I = 0]$ against which we assess pathogens' etiologic importance (Wu et al., 2016a), where $[A \mid B]$ represents the conditional distribution of a stochastic variable or vector A conditional on B. Since $[\boldsymbol{M} \mid \boldsymbol{W}, I = 0] = \sum_{k=1}^{K} \nu_k [\boldsymbol{M} \mid Z =$

$k, \boldsymbol{W}, I = 0]$, we let the distribution of control measurements depend on $\boldsymbol{W}$ through subclass weight regressions $\nu_k(\boldsymbol{W}) = h_k(\boldsymbol{W}; \boldsymbol{\Gamma}_k^\nu)$, $k = 1, \ldots, K-1$, where $h_k(\boldsymbol{W}; \cdot)$ represents a parameterized form for the function $\nu_k(\cdot)$ and $\boldsymbol{\Gamma}_k^\nu$ are the parameters ($K-1$ sets of parameters due to the simplex constraint). We let the subclass weights $\nu_k(\boldsymbol{W}) = \mathbb{P}(Z = k \mid \boldsymbol{W}, I = 0)$, $k = 1, \ldots, K$, vary with covariates according to an additive logistic stick-breaking regression. We propose a novel prior for the logistic stick-breaking regression that *a priori* encourages few subclasses of non-trivial weights uniformly over covariate values to approximate $[\boldsymbol{M} \mid \boldsymbol{W}, I = 0]$. For each disease class $\ell \neq 0$, we also assume a subclass weight regression $\eta_k(\boldsymbol{W}) = \mathbb{P}(Z = k \mid \boldsymbol{W}, I = \ell \neq 0) = h_k(\boldsymbol{W}; \boldsymbol{\Gamma}_k^\eta)$, $k = 1, \ldots, K-1$, as in the controls with the same parameterized form, the same number of subclasses and covariates but with different regression parameters $\boldsymbol{\Gamma}_k^\eta \neq \boldsymbol{\Gamma}_k^\nu$. Finally, we assume the TPR of observing a causative pathogen and the FPR of observing a non-causative pathogen in each *subclass* to be constant across covariate values. Of note, each *marginal* FPR in the control class may vary by covariates, because it is a weighted average of the subclass-specific FPRs $\{\psi_k^{(j)}, k = 1, \ldots, K\}$ with covariate-dependent weights $\{\nu_k(\boldsymbol{W}), k = 1, \ldots, K\}$; So do the marginal TPR in each case class.

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 $f_{\ell j}^\pi(\cdot)$ and use P-spline for estimating smooth functions (Lang and Brezger, 2004). Finally, we specify the distribution of case measurements given disease class I, covariates $\boldsymbol{X}$ and $\boldsymbol{W}$. We extend the case likelihood L_1 in an npLCM

(2) 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:

$$L_1^{\mathsf{reg}} = \prod_{i:Y_i=1} \left\{ \sum_{\ell=1}^{L} \left[\pi_\ell(\boldsymbol{X}_i; \boldsymbol{\Gamma}_\ell^\pi) \sum_{k=1}^{K} \{\eta_{ik} \cdot \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})\} \right] \right\}, \tag{3}$$

where $\eta_{ik} = h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^\eta)$ and $\boldsymbol{\Gamma}_k^\eta$ are the regression parameters for cases' subclass weights. We first introduce the parameterized form $h_k(\boldsymbol{W}_i; \cdot)$ by specifying control likelihood with covariates L_0^{reg}.

3.2 Covariate-dependent reference distribution for disease classes

The distribution of control measurements provides 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 that depends on the 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, the degree to which the observed marginal control positive rates depend on covariates informs how different the FPRs $\{\boldsymbol{\Psi}_k\}$ 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. The control distribution $\boldsymbol{P}_0$ in the original pLCM results upon a further single-subclass constraint ($K = 1$).

Secondly, we parameterize the case and control subclass weight regressions $\eta_k(\boldsymbol{W})$ (Equation 3) 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 the linear predictor for subject i at stick-breaking step $k = 1, \ldots, K$. Using the stick-breaking analogy, we begin with a unit-length stick, we break a segment of length $g(\alpha^\nu_{i1})$ and continue breaking a fraction $g(\alpha^\nu_{i2})$ from $\{1 - g(\alpha^\nu_{i1})\}$ that is left and so on; At step k, we break a fraction $g(\alpha^\nu_{ik})$ from what is left in the preceding $k-1$ breaking events resulting in a sticking segment k of length $\eta_{ik} = g(\alpha^\nu_{ik}) \prod_{s<k}\{1 - g(\alpha^\nu_{is})\}$; 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 etiology regression $\pi_\ell(\cdot)$ so that the priors for the coefficients can be similar. Generalization to other link functions such as the probit function is straightforward, but with a different posterior sampling algorithm involving latent Gaussian variables (e.g., Albert and Chib, 1993; Rodriguez and Dunson, 2011). We use this parameterization to introduce a novel shrinkage prior on a simplex of the subclass weights $\{\nu_k(\boldsymbol{W})\}$ (see Supplementary Material A1.1) which encourages parsimonious approximation to the conditional distribution of control measurements $\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, Y = 0, \{\nu_k(\cdot)\}, \boldsymbol{\Psi})$.

In our analysis, we use generalized additive models (Hastie and Tibshirani, 1986) for the k-th linear predictor $\alpha^\nu_{ik} = \alpha^\nu_k(\boldsymbol{W}_i = \boldsymbol{w}, \boldsymbol{\Gamma}^\nu_k) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(w_j; \boldsymbol{\beta}^\nu_{kj}) + \widetilde{\boldsymbol{w}}^\top \boldsymbol{\gamma}^\nu_k$, for $k = 1, \ldots, K-1$. We have parameterized the possibly nonlinear $f_{kj}(\cdot)$ using B-spline basis expansions with coefficients $\boldsymbol{\beta}^\nu_{kj}$; $\widetilde{\boldsymbol{w}}^\top \boldsymbol{\gamma}^\nu_k$ are the linear effects of a subset of predictors which can include an intercept where $\widetilde{\boldsymbol{w}}$ is a subvector of predictors $\boldsymbol{w}$; We let $\boldsymbol{\Gamma}^\nu_k = \{\mu_{k0}, \{\boldsymbol{\beta}^\nu_{kj}\}, \boldsymbol{\gamma}^\nu_k\}$ collect all the regression parameters. Following Lang and Brezger (2004), we constrain f_{kj} 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 linear predictors. We hence will propose priors on $\{\mu_{k0}\}$ to *a priori* encourage few subclass. In particular, a large positive intercept μ_{k0} makes $g(\alpha_{ik}^{\nu}) \approx 1$ and hence breaks nearly the entire stick that is left 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 as $L_0^{\mathsf{reg}} = \prod_{i:Y_i=0} \sum_{k=1}^{K} h_k(\boldsymbol{W}_i; \boldsymbol{\Gamma}_k^{\nu}) \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$.

Remark 1. *The proposed model for the control data with covariates $\boldsymbol{W}$ is a generative model where we first draw a subclass indicator $Z \mid \boldsymbol{W} \sim \mathsf{Categorical}_K\{\boldsymbol{\nu}(\boldsymbol{W})\}$, and generate measurements $M_j \mid Z = k$ according to a Bernoulli distribution with positive rate $\psi_k^{(j)}$, independently for $j = 1, ..., J$. By assuming mutually independent measurements $M_1, \ldots, M_J$ given subclass Z and $Y = 0$, we let the covariates influence the dependence structure of the measurement only through the unobserved Z. As a result, upon integrating over Z, the proposed model does not assume marginal independence $\mathbb{P}(\boldsymbol{M} \mid \boldsymbol{W}, Y = 0) = \prod_{j=1}^{J} \mathbb{P}(M_j \mid \boldsymbol{W}, Y = 0)$ in contrast to a kernel-based extension of the pLCM that make this assumption (Saha et al., 2018, Supplementary appendix).*

Case subclass weight regression. The subclass weight regression for cases $\eta_k(\boldsymbol{W})$ is also specified via a logistic stick-breaking regression as in the controls but with different parameters: $\eta_{ik} = g(\alpha_{ik}^{\eta}) \prod_{s<k}\{1 - g(\alpha_{is}^{\eta})\}$. Since given the TPRs and the FPRs, the subclass weights fully determine the measurement dependence in each class, we let the case and control subclass weight functions $\eta_k(\boldsymbol{w})$ and $\nu_k(\boldsymbol{w})$ be different for any $\boldsymbol{w}$.

Let the k-th linear predictor in the subclass weight regression for case subject i depend on covariates $\boldsymbol{W}_i$ via $\alpha_{ik}^{\eta} = \alpha_k^{\eta}(\boldsymbol{W}_i = \boldsymbol{w}; \boldsymbol{\Gamma}_k^{\eta}) = \mu_{k0} + \sum_{j=1}^{q_1} f_{kj}(w_j; \boldsymbol{\beta}_{kj}^{\eta}) + \widetilde{\boldsymbol{w}}^{\top} \boldsymbol{\gamma}_k^{\eta}$, where $\boldsymbol{\Gamma}_k^{\eta} = \{\mu_{k0}, \{\boldsymbol{\beta}_{kj}^{\eta}\}, \boldsymbol{\gamma}_k^{\eta}\}$ are the regression parameters. We use $\boldsymbol{\Gamma}_k^{\eta}$ that is different from the regression parameters in the controls ($\boldsymbol{\Gamma}_k^{\nu}$). In particular, we approximate $f_{kj}(\cdot)$ here using the same set of B-spline basis functions as in the controls but estimate a different set of

basis coefficients $\boldsymbol{\beta}_{kj}^{\eta}$. In addition, we have directly used the intercepts $\{\mu_{k0}\}$ from controls' subclass weight regressions 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_1^{reg}) and control likelihood (L_0^{reg}) with covariates, we obtain the joint likelihood for the regression model $L^{\mathsf{reg}} = L_1^{\mathsf{reg}} \times L_0^{\mathsf{reg}}$.

Remark 2. *The regression model reduces to an npLCM model without covariates upon integration over a distribution of covariates* $\boldsymbol{X}$ *under an assumption that* (A1)*: the case subclass weights are constant over covariates:* $\eta_k(\cdot) \equiv \eta_k$, $k = 1, \ldots, K$. *The likelihood function for cases' measurements* L_1^{reg} *integrates to* $L_1^* = \prod_{i:Y_i=1} \sum_{\ell=1}^{L} \pi_\ell^* \sum_{k=1}^{K} \eta_k \Pi(\boldsymbol{M}_i; \boldsymbol{p}_{k\ell})$, *where* $\pi_\ell^* = \int \pi_\ell(\boldsymbol{X}) \mathrm{d}G(\boldsymbol{X})$ *and* G *is a probability or empirical distribution of* $\boldsymbol{X}$. *The integrated control likelihood function is* $L_0^* = \prod_{i:Y_i=0} \sum_{k=1}^{K} \nu_k^* \Pi(\boldsymbol{M}_i; \boldsymbol{\Psi}_k)$, *where* $\nu_k^* = \int \nu_k(\boldsymbol{W}) \mathrm{d}H(\boldsymbol{W})$ *and* H *is a probability or empirical distribution of* $\boldsymbol{W}$. *The product of the integrated case and control likelihood* ($L_1^* L_0^*$) *is equivalent to an npLCM without covariates. The mathematical equivalence means we can perform an npLCM analysis omitting covariates* $\boldsymbol{X}$ *and* $\boldsymbol{W}$ *and obtain valid inference about the overall PEFs* $\boldsymbol{\pi}^*$. *The equivalence is evident once we enforce non-interference assumption (Wu et al., 2016b) by setting* $\eta_k(\boldsymbol{W}) = \nu_k(\boldsymbol{W})$ *for any* k *and* $\boldsymbol{W}$ *(or equivalently* $\boldsymbol{\Gamma}_k^{\eta} = \boldsymbol{\Gamma}_k^{\nu}$ *for any* k *under the parameterization* $h_k(\boldsymbol{W}; \cdot)$*), assumption* (A1) *is equivalent to* (A2)*:* $\eta_k(\boldsymbol{w}) = \nu_k(\boldsymbol{w}) = \nu_k, k = 1, \ldots, K$. *Under deviations from* (A1)*, an npLCM analysis omitting covariates however will result in substantial estimation bias by the posterior mean of* π_ℓ^* *and* 95% *CrIs that undercover the truth; Section 4 provides examples.*

Priors. The unknown parameters include the regression coefficients in the etiology regression ($\{\boldsymbol{\Gamma}_\ell^{\pi}\}$), the parameters in subclass mixing weight regression for the cases ($\{\boldsymbol{\Gamma}_k^{\eta}\}$) 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 for a

continuous variable (b) 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; For example, constant $\eta_k(\cdot) = \eta_k, \nu_k(\cdot) = \nu_k, k = 1, \ldots, K$ reduce the regression model to an npLCM without covariates. Supplementary Material A1 provides the details of the prior specifications.

Posterior Inference. We use the Markov chain Monte Carlo (MCMC) algorithm to draw samples of the unknowns to approximate the joint posterior distribution. The posterior inference is flexible and provide inferences about any functions of the model parameters and individual latent variables. For example, we may obtain the posterior distribution of the case positive rate curve for pathogen j (see red bands, Row 1, Figure 1) by plugging in the posterior samples of relevant parameters into $\mathbb{P}(M_\ell = 1 \mid \boldsymbol{x}, \boldsymbol{w}, Y = 1) = \pi_\ell(\boldsymbol{w}; \boldsymbol{\Gamma}_\ell^\pi) \sum_{k=1}^K h_k(\boldsymbol{w}; \boldsymbol{\Gamma}_k^\eta) \theta_k^{(\ell)} + \{1 - \pi_\ell(\boldsymbol{x}; \boldsymbol{\Gamma}_\ell^\pi)\} \sum_{k=1}^K h_k(\boldsymbol{w}; \boldsymbol{\Gamma}_k^\eta) \psi_k^{(\ell)}$. In the following, we fit npLCMs with or without covariates using a free and publicly available R package `baker` (`https://github.com/zhenkewu/baker`). It 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).

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 PEF π_ℓ^* as an empirical average of $\pi_\ell(\boldsymbol{X})$, $\ell = 1, \ldots, L$, which quantify the overall cause-specific disease burdens in a population and are often of health policy interest. We assess the frequentist properties of the posterior means of $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_L^*)^\top$ obtained from analyses with or without regression. Relative to npLCM analyses without covariates, the proposed 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.

We simulate BrS measurements made on J pathogens among N_1 cases and N_0 controls. We perform two analyses for each simulated data with and without covariates. We repeat the simulation and analyses for $R = 200$ independent replications to empirically assess the frequentist performance of the Bayesian procedures (Little et al., 2011). In estimating π^*_ℓ, we evaluate the bias $\widehat{\pi}^*_\ell - \pi^{0*}_\ell$ where $\widehat{\pi}^*_\ell = N_1^{-1}\sum_{i:Y_i=1}\widehat{\pi}_\ell(\boldsymbol{X}_i)$ and the true overall PEF $\pi^{0*}_\ell = N_1^{-1}\sum_{i:Y_i=1}\pi^0_\ell(\boldsymbol{X}_i)$, the posterior standard deviation $\mathbb{V}^{-1/2}\{\pi^*_\ell(\boldsymbol{\Gamma}^\pi_\ell) \mid \mathcal{D}\}$, the empirical coverage rates of the 95% credible intervals (CrIs) with the lower and upper end points being the 2.5% and 97.5% quantiles of the posterior distribution $[\pi^*_\ell(\boldsymbol{\Gamma}^\pi_\ell) \mid \mathcal{D}]$.

Simulation I. We demonstrate 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. In the true data generating mechanism, we let $\pi_\ell(\cdot)$, $\nu_k(\cdot)$ and $\eta_k(\cdot)$ depend on the two covariates $\boldsymbol{X} = (S, T)$, S and enrollment date (T), so that regression adjustments are necessary (see Remark 2). We simulate BrS measurements on $J = 9$ pathogens and assume the number of potential single-pathogen causes $L = J = 9$. To specify etiology regression functions that satisfy the constraint $\sum_{\ell=1}^{L}\pi_\ell(\boldsymbol{x}) = 1$, we use stick-breaking parameterization with $L = 9$ segments. In particular, we let $\mathsf{logit}\ \{g_1(s,t)\} = \beta_1\,\mathbb{I}(s = 1) + \sin(8\pi(t - 0.5)/7)$, $\mathsf{logit}\ \{g_2(s,t)\} = \beta_2\,\mathbb{I}(s = 1) + 4\exp(3t)/(1 + \exp(3t)) - 0.5$, $\mathsf{logit}(g_\ell) = \beta_8\,\mathbb{I}(s = 1)$ for $\ell > 2$; Let the PEF functions $\pi_\ell(s,t) = g_\ell(s,t)\prod_{j<\ell}\{1 - g_j(s,t)\}, \ell = 1,\ldots,L(= 9)$, where $\beta_\ell = 0.1, \ell = 1,\ldots,8$. During model fitting, we use B-spline expansion in the multinomial logistic regression for $\pi_\ell(\cdot)$ for t during estimation. To let the control distribution depend on covariates, we use $K = 2$ subclass weight functions for controls: $\nu_1(s,t) = \mathsf{logit}^{-1}\{\gamma^\nu_1\,\mathbb{I}(s = 1) + 4\exp(3t)/(1 + \exp(3t)) - 0.5\}$ and $\nu_2(s,t) = 1 - \nu_1(s,t)$. We specify case subclass weight functions that are different from the controls via $\eta_k(s,t) = \nu_k(s,-t), k = 1,2$, highlighting the need for using different subclass weights among cases and controls in an npLCM fitted to the data. We set the TPRs to be $\theta^{(j)}_k = 0.95$ and the FPRs to be $\psi^{(j)}_1 = 0.5$ and $\psi^{(j)}_2 = 0.05$. In our analyses with or without covariates, we use a working number of $K^* = 3$ subclasses, with independent $\mathsf{Beta}(7.13,1.32)$ TPR prior distributions which have the lower and upper 2.5% quantiles that match 0.55 and 0.99, respectively; We specify $\mathsf{Beta}(1,1)$ for the FPRs. The priors for the coefficients in the regression analyses follows the specifications in Supplementary Materials A1.

The solid lines in the top row of Figure 1 shows the TPRs for all $J = 9$ pathogens. Pathogen A has a bimodal positive rate curve mimicking the trends observed of pathogen 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 large t's; This is the extra modeling flexibility offered by the npLCM than the pLCM. This is an example where subclass with a low FPR ($k = 2$) is more heavily weighted in the controls than the cases: $\nu_2(s,t) > \eta_2(s,t)$. We then perform regression analysis by assuming $\phi_\ell(\boldsymbol{X})$ in the PEF regression specification to be an additive model of a $\mathbb{I}\{S = 2\}$ indicator and a B-spline expansion with 7 degrees of freedom for enrollment date t; The regression formula for subclass weights $\nu_k(\cdot)$ and $\eta_k(\cdot)$ are additive models of the $\mathbb{I}\{S = 2\}$ indicator and a B-spline expansion with 5 degrees of freedom for the enrollment date. 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 $\widehat{\pi}_\ell(\cdot)$ are close to the simulation truths $\pi^0_\ell(\cdot)$. See Supplementary Materials A3.1 for more simulation results to assess the recovery of the truth $\pi^0_\ell(X)$ for a discrete covariate X.

Simulation II. We show the regression model accounts for population stratification by covariates hence reduces the bias of the posterior mean $\{\widehat{\pi}^*_\ell\}$ in estimating the overall 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 of each $R = 200$ replication data sets simulated under each of 48 scenarios below that characterize combinations of 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. (2016a)), and effects of W on $\{\nu_k(W)\}$ and $\{\eta_k(W)\}$. In particular, we consider $L = J = 3, 6, 9$ causes, under single-pathogen-cause assumption, BrS measurements made on N_d cases and N_u controls for each level of X where $N_d = N_u = 250$ or 500. The functions $\phi_\ell(X) = \beta_{0\ell} + \beta_{1\ell}\mathbb{I}\{X = 2\}$ take two sets of values to reflect how variable the PEFs are across the two X levels: i) $\boldsymbol{\beta}^{\mathsf{i}}_0 = (0, 0, 0, 0, 0, 0)$ and $\boldsymbol{\beta}^{\mathsf{i}}_1 = (-1.5, 0, -1.5, -1.5, 0, -1.5)$ where causes have uniform PEFs when $X = 1$ and causes B and E dominate when $X = 2$, or ii) $\boldsymbol{\beta}^{\mathsf{ii}}_0 = (1, 0, 1, 1, 0, 1)$ and $\boldsymbol{\beta}^{\mathsf{ii}}_1 = (-1.5, 1, -1.5, -1.5, 1, -1.5)$ to mimic the scenario where pathogens B and E have lower PEFs when $X = 1$ and occupy

more fractions when $X = 2$. We further let the measurement error parameters take distinct values of the TPRs $\theta_k^{(j)} = 0.95$ or 0.8 and the FPRs $(\psi_1^{(j)}, \psi_2^{(j)}) \in \{(0.5, 0.05), (0.5, 0.15)\}$, for $j = 1, \ldots, J$. Finally, we specify control and case subclass weight regression functions to be the same: $\nu_k(W) = \eta_k(W) = \mathsf{logit}^{-1}(\gamma_{k0} + \gamma_{k1}\,\mathbb{I}\{W = 2\})$ where $(\gamma_{10}, \gamma_{11}) = (-0.5, 1.5)$ and $(\gamma_{20}, \gamma_{21}) = (1, -1.5)$.

Based on a single data set simulated under the scenario $\{L = 6,\ N_d = 500,\ K = 2,\ \theta_k^{(j)} = 0.8,\ (\psi_1^{(j)}, \psi_2^{(j)}) = (0.5, 0.05),\ (\boldsymbol{\beta}_0^{\mathsf{ii}}, \boldsymbol{\beta}_1^{\mathsf{ii}})\}$, Supplemental Figure S3 shows the posterior distribution of the stratum-specific etiology fractions $\pi_\ell(X = s)$ for $(s = 1, 2)$ by row and $L(= J)$ causes $(\ell = 1, \ldots, 6)$ by column with the true values indicated by the blue vertical dashed lines; The bottom row shows the posterior distribution of $\pi_\ell^* = \sum_s w_s \pi_\ell(X = s)$ for L causes with empirical weights $w_s = N_d^{-1} \sum_{i:Y_i=1} \mathbb{I}\{X_i = s\}$, $s = 1, 2$.

We also observe superior performance of the regression method upon repeated applications of the posterior inferential algorithm across simulation scenarios. Figure 2(a) shows for $J = 6$ that, relative to no-covariate npLCM analyses, regression analyses produce posterior means that on average have smaller absolute relative biases (the percent difference between the posterior mean and the truth relative to the truth) for each pathogen across simulation scenarios. The regression analyses also produce 95% CrIs for π_ℓ^* that have more valid empirical coverage rates in all combinations of the parameters than npLCM analyses omitting covariates for which we observe undercoverage. Misspecified models without covariates result in large biases that dominate the posterior uncertainty of π_ℓ^* and become more deficient under larger sample sizes and stronger signals. For example, contrast the more severe undercoverages in the bottom two rows with higher TPRs than the top two rows with lower FPRs in Figure 2. The regression analyses also perform evidently better in our simulations for $J = 3$ and $J = 9$ (results not shown).

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 (August 2011 to September 2013), 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; We also include in the analysis the blood culture (BCX) results for two bacteria from cases only. Detailed analyses of the entire data are reported in PERCH Study Group (2019).

Table 1 shows the observed 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 stratum-specific fitting of npLCM. 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).

Pathogens RSV and *Haemophilus influenzae* (HINF) are detected with the highest positive rates among pneumonia children: 29.3% and 34.1%, respectively, which are higher than the corresponding control rates (3.1% and 21.7%). The CORs are 14 (95%CI: $9.4, 21.6$) and 1.8 (95%CI: $1.3, 2.3$) are large and indicate 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, detection of HMPV_A_B and PARA_1 are less frequent in cases' nasal cavities than RSV and HINF (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%).

Additional imperfectly sensitive but highly specific blood culture measurements are available for HINF and PNEU. Similar to Wu et al. (2016b), we incorporate such additional measurements and informative priors on 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 extremely rare detection from blood culture of the two bacteria, 0.4% for HINF and 0.2% for PNEU, will lower their PEF estimates obtained from an analysis that

only uses NPPCR data.

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 from laboratory experts. We set the Beta priors for sensitivities by $a_\theta = 126.8$ and $b_\theta = 48.3$ 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 range of $5 - 20\%$. We use a working number of subclasses $K = 5$. In the etiology regression model $\phi_\ell(\boldsymbol{X})$, we use 7 degrees of freedom for B-spline expansion of the additive function for the standardized enrollment date 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 degrees of freedom for the standardized enrollment date 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 following the specification in Supplementary Materials A1.

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 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 by setting other covariates to particular levels for two strata with the most cases: severe pneumonia, HIV negative and younger than one (Figure 3(a)) or older than one (Figure 3(b)). Among the younger, severe pneumonia and HIV negative children, 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 population etiologic fractions of all the causes sum to one.

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 covariates. For example, consider two pneumonia cases with negative results on the seven pathogens (the most frequent pattern among cases' and controls' NPPCR mea-

surements) but one is under one year of age and the other is not. The older pneumonia case have a lower posterior probability of her disease caused by RSV (solid dots below the empty dots, Supplemental Figure S6 and higher probability of being caused by NoS (solid dots above the empty dots, Supplemental Figure S6). Indeed, comparing older and younger children while holding the enrollment date, HIV, severity enrollment 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 : $-2.99, -0.77$). As a result, compared with younger children, the older children have overall lower RSV and higher NoS PEF estimates.

Given age, severity and HIV status, we quantify the overall cause-specific disease burden by averaging the PEF estimates by the empirical distribution of enrollment dates. The posterior means are shown along with the 95% CrIs above the etiology regression functions. 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 etiologic fraction of 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 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 methods are motivated by a study of pediatric pneumonia etiology (PERCH Study Group, 2019) and can be applied to other large-scale disease etiology studies of neonatal infections (Saha et al., 2018) and diarrheal diseases (Kotloff et al., 2013). Similar analytic needs and challenges have been raised by different scientific areas such as estimating cause-

specific mortality rates from verbal autopsy data in demography (McCormick et al., 2016), subgrouping disease in medicine (Wu et al., 2019) and identifying the underlying mechanisms of learning difficulties in psychology (Gu and Xu, 2019a).

The proposed approach has three distinguishing features: 1) It allows analysts to specify a model for the 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 posterior inferential algorithm estimates a parsimonious covariate-dependent reference distribution of the diagnostic measurements from controls, against which the measurements made on a case with similar covariate values are compared to assign cause-specific probabilities given her measurements. 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 reuse these priors, the proposed regression analysis avoids overly-optimistic etiology uncertainty estimates.

On estimating the overall PEFs $\boldsymbol{\pi}^*$ that characterize the overall cause-specific disease burdens in a population, 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.

Similar to an npLCM analysis without covariates, 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, to further reduce the posterior uncertainty of the etiology estimates.

Future analyses of data from large scale disease etiology studies may benefit from four improvements to the proposed method. First, although the proposed model and posterior inferential algorithm accommodate a regression specification to arbitrary levels of interactions, we chose to fit additive models to the PERCH data. In so doing, we avoid statistical instability of fully-stratified etiology estimates due to sparsely-populated strata at the expense of introducing some bias when interaction effects truly exist. Bayesian additive regression tree with variable selection (e.g., Linero, 2018) may provide a parsimonious alternative for characterizing interactions; Here "additive" means a regression function is assumed to be a random sum of decision trees each of which is flexible to capture part of the non-linear

and interaction effects. Second, in the etiology regression model $\{\pi_\ell(\boldsymbol{X}), \ell = 1, \ldots, L\}$, this paper assumes the predictors are commons to all the disease classes. Class-specific predictor selection methods (Gustafson et al., 2008) may provide useful regularization in the presence of a large number of predictors to further stabilize and improve the interpretability of the PEF function estimates. Third, when the number of causes $1 \leq L \leq 2^J$ and the subset of pathogen-cause combinations in the population is unknown, combining the proposed method with subset selection procedures (Wu et al., 2019; Gu and Xu, 2019a) may be fruitful. 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.

Acknowledgment

We thank the PERCH study team led by Kathernine O'Brien for providing the data and scientific advice, in particular Scott Zeger, Maria Deloria-Knoll, Christine Prosperi and Qiyuan Shi for insightful comments when the idea was conceived and for valuable feedback that made `baker` better. We also thank Jing Chu for preliminary simulation studies. 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) and an Investigator Award from Precision Health Initiative and an MCubed Award from University of Michigan (ZW).

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

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(suppl_3):S228–S237.

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

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The annals of applied statistics*, 1(2):346.

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

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

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):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(4):1478–1502.

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

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):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(1):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(3):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(9888):209–222.

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

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

Levine, O., O'Brien, K., Deloria-Knoll, M., Murdoch, D., Feikin, D., DeLuca, A., Driscoll, A., Baggett, H., Brooks, W., Howie, S., et al. (2012). The pneumonia etiology research for child health project: A 21st century childhood pneumonia etiology study. *Clinical Infectious Diseases*, 54(suppl 2):S93–S101.

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

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

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

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(9807):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(9):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(1).

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

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(4):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. (2016a). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):97–114.

Wu, Z., Deloria-Knoll, M., and Zeger, S. L. (2016b). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics*, 18(2):200–213.

Zhang, Q. and Zhou, M. (2017). Permuted and augmented stick-breaking bayesian multinomial regression. *The Journal of Machine Learning Research*, 18(1):7479–7511.

Figures and Tables

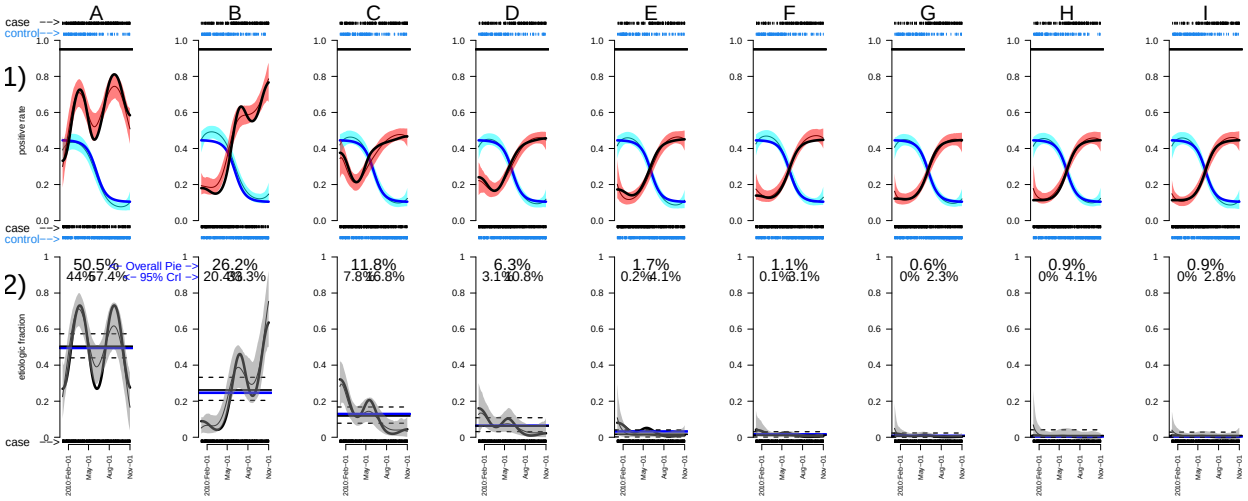
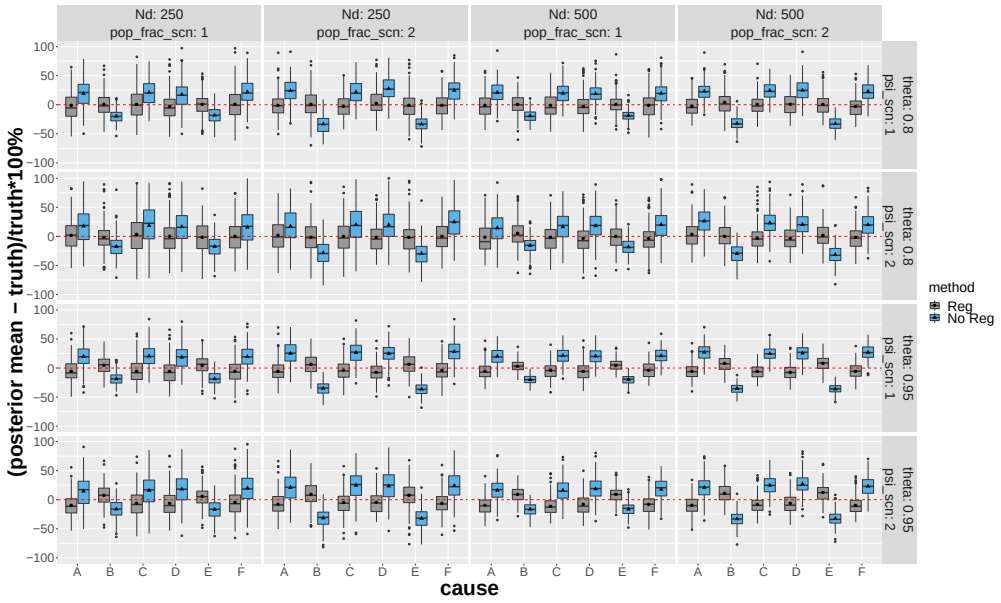
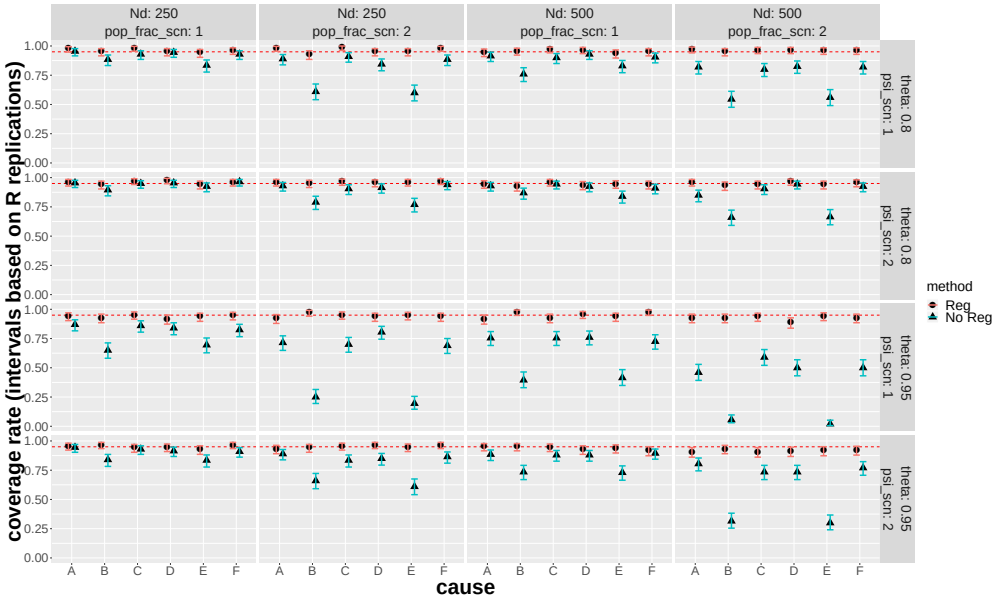


Figure 1: For each of the 9 causes (by column) in the Simulation I, the posterior mean (thin black curves) and pointwise 95% credible bands (gray bands) for the etiology regression curves $\pi_\ell(x)$ are close to the simulation truths $\pi_\ell^0(x)$. The fitted case (red) and control (blue) positive rate curves are shown with the posterior mean curves and pointwise 95% credible bands; The rug plots show the positive (top) and negative (bottom) measurements made on cases and controls on the enrollment dates.

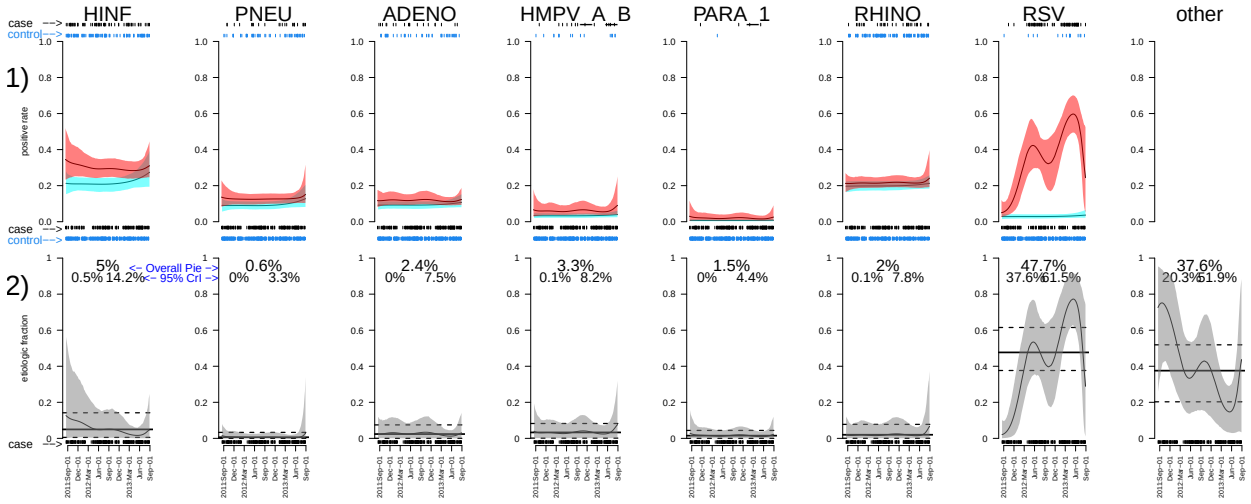


(a) Percent relative bias

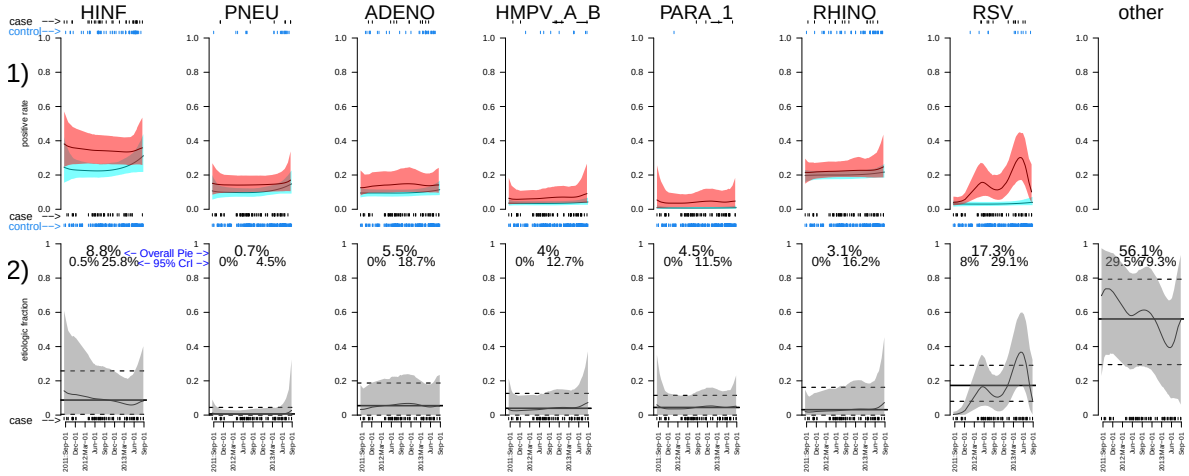


(b) Empirical coverage rates

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% confidence interval is also shown.



(a) **Age** ≤ 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 percentages among cases and controls for each covariate are shown at the bottom. Results from the regression analyses are shown for the first two strata.

age ≥ 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%		