5-31-2014

Partially-Latent Class Models (pLCM) for Case-Control Studies of Childhood Pneumonia Etiology

Zhenke Wu
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, zhwu@jhu.edu

Maria Deloria-Knoll
Department of International Health, Johns Hopkins Bloomberg School of Public Health, mknoll2@jhu.edu

Laura L. Hammitt
Department of International Health, Johns Hopkins Bloomberg School of Public Health, lhammitt@jhu.edu

Scott L. Zeger
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, sz@jhu.edu

Partially-Latent Class Models (pLCM) for Case-Control Studies of Childhood Pneumonia Etiology

Zhenke Wu*[1], Maria Deloria-Knoll[2], Laura L. Hammitt[2], Scott L. Zeger[1]

for the PERCH Core Team

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[2]Department of International Health, Johns Hopkins Bloomberg School of Public Health

May 31, 2014

Abstract

In population studies on the etiology of disease, one goal is the estimation of the fraction of cases attributable to each of several causes. For example, pneumonia is a clinical diagnosis of lung infection that may be caused by viral, bacterial, fungal, or other pathogens. The study of pneumonia etiology is challenging because directly sampling from the lung to identify the etiologic pathogen is not standard clinical practice in most settings. Instead, measurements from multiple peripheral specimens are made. This paper considers the problem of estimating the *population etiology distribution*and the *individual etiology probabilities*. We formulate the scientific problem in statistical terms as estimating the posterior distribution of mixing weights and latent class indicators under a partially-latent class model (pLCM) that combines

*Email: zhwu@jhu.edu

heterogeneous measurements with different error rates obtained from a case-control study. We introduce the pLCM as an extension of the latent class model. We also introduce graphical displays of the population data and inferred latent-class frequencies. The methods are illustrated with simulated and real data sets. The paper closes with a brief description of extensions of the pLCM to the regression setting and to the case where conditional independence among the measures is relaxed.

Keywords: Bayesian method; Case-control; Etiology; Latent class; Measurement error; Pneumonia

1 Introduction

Identifying the pathogens responsible for infectious diseases in a population poses significant statistical challenges. Consider the measurement problem in the Pneumonia Etiology Research for Child Health (PERCH), a case-control study that has enrolled $9,500$ children from 7 sites around the world. Pneumonia is a clinical syndrome that develops because of an infection of the lung tissue by bacteria, viruses, mycobacteria or fungi (Levine et al., 2012). The appropriate treatment and public health control measures vary by pathogen. Which pathogen is infecting the lung usually cannot be directly observed and must therefore be inferred from multiple peripheral measurements with differing error rates. The primary goals of the PERCH study are to integrate the multiple sources of data to: (1) aid the attribution of which pathogen or pathogens have caused a particular case's lung infection, and (2) estimate the prevalences of the etiologic pathogens in a population of children.

The basic statistical framework of the problem is pictured in Figure 1. Let Y_i represent whether the child is a pneumonia case ($Y_i = 1$) or control ($Y_i = 0$). For a child with pneumonia, let I_i^L indicate which pathogen causes the lung infection. I_i^L takes values in $\{0, 1, 2, ...J\}$ where 0 represents no infection (control) and $I_i^L = j, j = 1, ..., J$, represents the jth pathogen from a pre-specified cause-of-pneumonia or pneumonia etiology list. Among the J candidate pathogens being tested, we assume only one is the primary cause. Because, for most cases, it is not possible to directly sample the lung, we do not know with certainty which pathogen infected the lung, so we seek to infer the infection status I_i^L based upon a series of laboratory measurements of specimens from various body fluids and body sources S (M_i^S).

The measurement error rates differ by type of measurement. In the motivating PERCH application and the following discussions, the error rates refer to *epidemiologic* error rates

that characterize the probability of the pathogen's presence/absence in specimen tests given whether it infected the lung. For this and possibly other applications, it is convenient to categorize measures into three subgroups referred to as "gold", "silver", and "bronze" standard measurements. A gold-standard (GS) measurement is assumed to have both perfect sensitivity and specificity. A silver-standard (SS) measurement is assumed to have perfect specificity, but imperfect sensitivity. Culturing bacteria from blood samples (B-Cx) is an example of silver standard measurements in PERCH. Finally, bronze-standard (BrS) measurements are assumed to have imperfect sensitivity and specificity. Polymerase chain reaction (PCR) evaluation of bacteria and viruses from nasopharyngeal samples is an example. In the PERCH study, both SS and BrS measurements are available in all cases. BrS measures are also available for controls. A goal of this study is to develop a statistical model that combines GS and SS measurements from cases, with bronze data from cases and controls to estimate the distribution of pathogens in the population of pneumonia cases, and the conditional probability that each of the J pathogens is the primary cause of an individual child's pneumonia given her or his set of measurements. Even in applications where GS data is not available, a flexible modeling framework that can accommodate GS data is useful for both the evaluation of statistical information from BrS data (Section 3) and the incorporation of GS data if it becomes available as measurement technology improves.

Latent class models (LCM) (Goodman, 1974) have been successfully used to integrate multiple diagnostic tests or raters' assessments to estimate a binary latent status $D \in \{0, 1\}$ for all study subjects (Hui and Walter, 1980; Qu and Hadgu, 1998; Albert et al., 2001; Albert and Dodd, 2008). (In these applications, $D = 1$ if $I^L > 0$.) In the LCM framework, conditional distributions $[\boldsymbol{M}|D = j], j = 0, 1$, are specified to use multivariate measurements $\boldsymbol{M}$ to maximize the likelihood as a function of the disease prevalence, sensitivities

and specificities. This framework has also been extended to infer ordinal latent status (Wang et al., 2011).

There are three salient features of the PERCH childhood pneumonia problem that require extension of the typical LCM approach. First, we have *partial* knowledge of the latent lung state I^L for some subjects as a result of the case-control design. In the standard LCM approach, the study population comprises subjects with completely unknown class membership D. In this study, the latent etiology $I^L = 0$ is applied to all controls because absent clinical disease, the lung is assumed to be non-infected. Also, were gold standard measurements available from the lung for some cases, their latent variable would be directly observed. As the latent state is known for a non-trivial subset of the study population, we refer to the model posited below as a partially-Latent Class Model or pLCM.

Second, in most LCM applications, the number of diagnostic test results on a subject is much larger than the number of latent state categories. Here, the number of diagnostic tests is of the same order, and often equal to the number of categories that I^L can assume. For example, if we consider only the PERCH study BrS data, we simultaneously observe the presence/absence of J pathogens for each child. Even with additional control data, the larger number of latent categories of I^L leads to weak model identifiability as is discussed in more detail in Section 2.1.

Lastly, measurements with differing error rates (i.e. GS, SS, BrS) need to be integrated in this application. Understanding the relative value of each level of measurements is important to optimally invest resources into data collection (number of subjects, type of samples) and laboratory assays. An important goal is therefore to estimate the relative information from each type of measurements about the population and individual etiology distributions. Albert and Dodd (2008) studied a model where some subjects are selected to verify their latent status (i.e. collect from them GS measurements) with the probability of

verification depending on the previous test results or completely at random. They showed GS data can make model estimates more robust to model misspecifications. We quantify how much GS data reduces the variance of model parameter estimates for design purposes. Also, they considered binary latent status and did not have available control data. Another related literature that uses both GS and BrS data is on verbal autopsy (VA) in the setting where no complete vital registry system is established in the community (King and Lu, 2008). Quite similar to the goal of inferring pneumonia etiology from lab measurements, the goal of VA is to infer the cause of death (I^D) from a pre-specified list by asking close family members questions about the presence/absence of K symptoms. King and Lu (2008) proposed estimating the cause-of-death distribution in community $P(I^D = j), j = 1, ...J$, (similar to etiology) using data on K dichotomous symptoms and GS data from the hospital where cause-of-death and symptoms are both recorded. However, their method involves nonparametric estimation of J K-way probability contingency tables and therefore requires a sizable sample of GS data, especially when the number of symptoms is large. In addition, a key difference between VA and most infectious disease etiology studies is that the VA studies are by definition case-only.

Another approach previously used with case and control data is to perform logistic regression of case status Y on laboratory measurements $\boldsymbol{M}$ and then to calculate point estimates of population attributable risks for each pathogen (Bruzzi et al., 1985; Blackwelder et al., 2012). This method does not account for imperfect laboratory measurements and cannot use GS data if available. Also, zero prevalence is assigned to pathogens whose estimated odds ratios are smaller than 1, without taking account of their statistical uncertainty.

In this paper, we define and apply a partially-latent class model (pLCM) with conditional independent assumptions to incorporate these three features: known infection status

for controls, a large number of latent classes, and multiple types of measurements. We use a hierarchical Bayesian formulation to estimate: (1) the *population etiology distribution* or *etiology fraction* —the frequency with which each pathogen "causes" clinical pneumonia in the case population. and (2) the *individual etiology probabilities*—the probabilities that a case is "caused" by each of the candidate pathogens, given observed specimen measurements for that individual. Shrinkage of the individual's predictive distribution toward the population etiology distribution is controlled in a natural way by the estimated case pathogen prevalences, and the differences in the estimated true positive rates relative to false positive rates (Section 3); and (3) the relative information content of GS, SS, and BrS data (Section 3 and 4).

The remainder of this paper proceeds as follows. In section 2, we formulate the pLCM and the Gibbs sampling algorithms for implementation. In Section 3, we evaluate our method through simulations tailored for the childhood pneumonia application. Section 4 presents the application of our methodology to a subsample of the PERCH data to demonstrate its applicability. Lastly, Section 5 concludes with a discussion of results and limitations, a few natural extensions of the pLCM also motivated by the PERCH data, as well as future directions of research.

2 A partially-latent class model for multiple indirect measurements

We develop pLCM to address two characteristics of the motivating pneumonia problem: (1) a partially-latent state variable because the pathogen infection status is known for controls but not cases; and (2) multiple categories of measurements with different error rates across classes. As shown in Figure 1, let I_i^L, taking values in $\{0, 1, 2, ...J\}$, represent the true state

of child i's lung ($i = 1, ..., N$) where 0 represents no infection (control) and $I_i^L = j, j = 1, ..., J$, represents the jth pathogen from a pre-specified cause-of-pneumonia list that is assumed to be exhaustive. Let $\boldsymbol{M}_i^S$ represent the $J \times 1$ vector of binary indicators of the presence/absence of each pathogen in the measurement at site S, where, in our application S can be nasopharyngeal (NP), blood (B), or lung (L). Let $\boldsymbol{m}_i^S$ be the actual observed values. In the following, we replace S with BrS, SS, or GS, because they correspond to the measurement types at NP, B, and L, respectively.

Let $Y_i = y_i \in \{0, 1\}$ represent the indicator of whether child i is a control or case. Note $I_i^L = 0$ given $Y_i = 0$. To formalize the pLCM, we define three sets of parameters:

- $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)^T$ for the probability $\Pr(I^L = j \mid Y = 1, \boldsymbol{\pi}), j = 1, ..., J$
- $\psi_j^S = \Pr(M_j^S = 1 | I^L = 0)$, the marginal false positive rate (FPR) for measurement j at site S
- $\theta_j^S = \Pr(M_j^S = 1 | I^L = j)$, the marginal true positive rate (TPR) for measurement j at site S for a person whose lung is infected by pathogen j.

We further let $\boldsymbol{\psi}^S = (\psi_1^S, ..., \psi_J^S)^T$ and $\boldsymbol{\theta}^S = (\theta_1^S, ..., \theta_J^S)^T$. Using these definitions, we have FPR $\psi_j^{\mathsf{BrS}} = 0$ and TPR $\theta_j^{\mathsf{BrS}} = 1$ for GS measurements, so that $M_j^{\mathsf{GS}} = 1$ if and only if $I_i^L = j$ (perfect sensitivity and specificity). Let δ_i be the binary indicator of a case i having GS measurements; it equals 1 if the case has available GS data and 0 otherwise. For SS measurements, FPR $\psi_j^{\mathsf{SS}} = 0$ so that $M_j^{\mathsf{SS}} = 0$ if $I_i^L \neq j$ (perfect specificity).

We formalize the model likelihood for each type of measurement. We first describe the model for BrS measurement $\boldsymbol{M}^{\mathsf{BrS}}$ for a control or a case. For control i, positive detection of the jth pathogen is a false positive representation of the non-infected lung. Therefore, we assume $M_{ij}^{\mathsf{BrS}} \mid \boldsymbol{\psi}^{\mathsf{BrS}} \sim \text{Bernoulli}(\psi_j^S), j = 1, ..., J$, with conditional independence, or

equivalently,

$$P_i^{0,\mathsf{BrS}} = \Pr(\boldsymbol{M}_i^{\mathsf{BrS}} = \boldsymbol{m} \mid \boldsymbol{\psi}^{\mathsf{BrS}}) = \prod_{j=1}^{J} \left(\psi_j^{\mathsf{BrS}}\right)^{m_j} \left(1-\psi_j^{\mathsf{BrS}}\right)^{1-m_j}, \boldsymbol{m} = \boldsymbol{m}_i^{\mathsf{BrS}} \quad (2.1)$$

For a case infected by pathogen j, the positive detection rate for the jth pathogen in BrS assays is θ_j^{BrS}. Since we assume a single cause for each case, detection of pathogens other than j will be false positives with probability equal to marginal FPR as in controls: ψ_l^{BrS}, $l \neq j$. This nondifferential misclassification across the case and control populations is the essential assumption of the latent class approach because it allows us to borrow information from control BrS data to distinguish the true cause from background colonization. We further discuss it in the context of the pneumonia etiology problem in the final section. Then,

$$\begin{aligned} P_{i'}^{1,\mathsf{BrS}} &= \Pr(\boldsymbol{M}_{i'}^{\mathsf{BrS}} = \boldsymbol{m} \mid \boldsymbol{\pi}, \boldsymbol{\theta}^{\mathsf{BrS}}, \boldsymbol{\psi}^{\mathsf{BrS}}) \\ &= \sum_{j=1}^{J} \pi_j \cdot \left(\theta_j^{\mathsf{BrS}}\right)^{m_j} \left(1-\theta_j^{\mathsf{BrS}}\right)^{1-m_j} \prod_{l\neq j} \left(\psi_l^{\mathsf{BrS}}\right)^{m_l} \left(1-\psi_l^{\mathsf{BrS}}\right)^{1-m_l}, \boldsymbol{m} = \boldsymbol{m}_{i'}^{\mathsf{BrS}} \end{aligned} \quad (2.2)$$

is the likelihood contributed by BrS measurements from case i'. Convenient for Gibbs sampler, we introduce the latent lung infection state $I_{i'}^L$ and represent (2.2) by the following two-stage sampling scheme:

(i) multinomial sampling of lung infection state among cases: $I_{i'}^L \mid \boldsymbol{\pi}, Y_{i'} = 1 \sim \text{Multinomial}(\boldsymbol{\pi})$,

(ii) measurement stage given lung infection state:

$M_{i'j}^{\mathsf{BrS}} \mid I_{i'}^L, \boldsymbol{\theta}^{\mathsf{BrS}}, \boldsymbol{\psi}^{\mathsf{BrS}} \sim \text{Bernoulli}\left(\mathbf{1}_{\{I_{i'}^L=j\}}\theta_j^{\mathsf{BrS}} + \left(1-\mathbf{1}_{\{I_{i'}^L=j\}}\right)\psi_j^{\mathsf{BrS}}\right), j = 1, ..., J,$ conditionally independent, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function and equals one if the statement in $\{\cdot\}$ is true; otherwise, zero.

Similarly, likelihood contribution from a case i''s SS measurements can be written as

$$P_{i'}^{1,\mathrm{SS}} = \Pr(\boldsymbol{M}_{i'}^{\mathsf{SS}} = \boldsymbol{m} \mid \boldsymbol{\pi}, \boldsymbol{\theta}^{\mathsf{SS}}) = \sum_{j=1}^{J'} \pi_j \cdot \left(\theta_j^{\mathsf{SS}}\right)^{m_j} (1-\theta_j^{\mathsf{SS}})^{1-m_j} \mathbf{1}_{\left\{\sum_{l=1}^{J'} m_l \leq 1\right\}}, \quad (2.3)$$

for $\boldsymbol{m} = \boldsymbol{m}_{i'}^{\mathsf{SS}}$, noting the perfect specificity of SS measurements, where $J' \leq J$ represents the number of actual SS measurements on each case, and $\boldsymbol{\theta}^{\mathsf{SS}} = \left(\theta_1^{\mathsf{SS}}, ...\theta_{J'}^{\mathsf{SS}}\right)$. SS measurements only test for a subset of all J pathogens, e.g., blood culture only detects bacteria and J' is the number of bacteria that are potential causes. Finally, GS measurement $M_{i'}^{\mathsf{GS}}$ that accurately indicates the actual cause for case i', is assumed to follow multinomial distribution with likelihood:

$$P_{i'}^{1,\mathrm{GS}} = \Pr\left(M_{i'}^{\mathsf{GS}} = \boldsymbol{m} \mid \boldsymbol{\pi}\right) = \prod_{j=1}^{J} \pi_j^{\mathbf{1}\{m_j=1\}} \mathbf{1}_{\{\sum_j m_j = 1\}}, \boldsymbol{m} = \boldsymbol{m}_{i'}^{\mathsf{GS}} . \tag{2.4}$$

Combining likelihood components (2.1)—(2.4), the total model likelihood for BrS, SS, and GS data across independent cases and controls is

$$L(\boldsymbol{\gamma}; \mathcal{D}) = \prod_{i:Y_i=0} P_i^{0,\mathsf{BrS}} \prod_{i':Y_{i'}=1,\delta_{i'}=1} P_{i'}^{1,\mathsf{BrS}} \cdot P_{i'}^{1,\mathsf{SS}} \cdot P_{i'}^{1,\mathsf{GS}} \prod_{i'':Y_{i''}=1,\delta_{i''}=0} P_{i''}^{1,\mathsf{BrS}} \cdot P_{i''}^{1,\mathsf{SS}} \tag{2.5}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\theta}^{\mathsf{BrS}}, \boldsymbol{\psi}^{\mathsf{BrS}}, \boldsymbol{\theta}^{\mathsf{SS}}, \boldsymbol{\pi})^T$ stacks all unknown parameters, and $\mathcal{D} = \left\{\left\{\boldsymbol{m}_i^{\mathsf{BrS}}\right\}_{i:Y_i=1}\right\} \cup \left\{\left\{\boldsymbol{m}_{i'}^{\mathsf{BrS}}, \boldsymbol{m}_{i'}^{\mathsf{GS}}, \boldsymbol{m}_{i'}^{\mathsf{SS}}\right\}_{i':Y_{i'}=1,\delta_{i'}=1}\right\} \cup \left\{\left\{\boldsymbol{m}_{i''}^{\mathsf{BrS}}, \boldsymbol{m}_{i''}^{\mathsf{SS}}\right\}_{i'':Y_{i''}=1,\delta_{i''}=0}\right\}$ collects all the available measurements on study subjects. Our primary statistical goal is to estimate the posterior distribution of the population etiology distribution $\boldsymbol{\pi}$, and obtain individual etiology (I_*^L) prediction given a case's measurements ($\boldsymbol{m}_*^{\mathsf{BrS}}, \boldsymbol{m}_*^{\mathsf{SS}}$), i.e., $\Pr(I_*^L = j \mid \boldsymbol{m}_*^{\mathsf{BrS}}, \boldsymbol{m}_*^{\mathsf{SS}}, \mathcal{D}), j = 1, ..., J$.

To enable Bayesian inference, prior distributions on model parameters are specified as follows: $\boldsymbol{\pi} \sim \mathrm{Dirichlet}(a_1, \ldots, a_J)$, $\psi_j^{\mathsf{BrS}} \sim \mathrm{Beta}(b_{1j}, b_{2j})$, $\theta_j^{\mathsf{BrS}} \sim \mathrm{Beta}(c_{1j}, c_{2j}), j = 1, ..., J$, and $\theta_j^{\mathsf{SS}} \sim \mathrm{Beta}(d_{1j}, d_{2j})$, $j = 1, ..., J'$. Hyperparameters for etiology prior, $a_1, ..., a_J$, are usually 1s to denote equal and non-informative prior weights for each pathogen if expert prior knowledge is unavailable. The FPR for the jth pathogen, ψ_j^{BrS} , generally can be well estimated from control data, thus $b_{1j} = b_{2j} = 1$ is the default choice. For TPR parameters θ_j^{BrS} and θ_j^{SS} , if prior knowledge on TPRs is available, we choose (c_{1j}, c_{2j}) so

that the 2.5% and 97.5% quantiles of Beta distribution with parameter (c_{1j}, c_{2j}) match the prior minimum and maximum TPR values elicited from pneumonia experts . Otherwise, we use default value 1s for the Beta hyperparameters. Similarly we choose values of (d_{1j}, d_{2j}) either by prior knowledge or default values of 1. We finally assume prior independence of the parameters as $[\boldsymbol{\gamma}] = [\boldsymbol{\pi}][\boldsymbol{\psi}^{\mathsf{BrS}}][\boldsymbol{\theta}^{\mathsf{BrS}}][\boldsymbol{\theta}^{\mathsf{SS}}]$, where $[A]$ represents the distribution of random variable or vector A. These priors represent a balance between explicit prior knowledge about measurement error rates and the desire to be as objective as possible for a particular study. As described in the next section, the identifiability constraints on the pLCM require specifying a reasonable subset of parameter values to identify parameters of greatest scientific interest.

2.1 Model identifiability

Potential non-identifiability of LCM parameters is well-known. For example, an LCM with four observed binary indicators and three latent classes is not identifiable despite providing 15 degree-of-freedom to estimate 14 parameters (Goodman, 1974). In principle, the Bayesian framework avoids the non-identifiability problem in LCMs by incorporating prior information about unidentified parameter subspaces (Garrett and Zeger, 2000). Many authors point out that the posterior variance for non-identifiable parameters does not decrease to zero as sample size approaches infinity (e.g., Kadane (1974); Gustafson et al. (2001); Gustafson (2005)). For scientific investigations, when data are not fully informative about a parameter, an identified set of parameter values consistent with the observed data shall, nevertheless, be valuable in a complex application (Gustafson, 2009) like PERCH.

This identifiability issue for the pLCM only occurs in the absence of GS data. Here we restrict attention to the scenario with only BrS data for simplicity but similar arguments pertain to the BrS + SS scenario. The problem can be understood from the form of the

marginal positive measurement rates for pathogens among cases. In the pLCM likelihood for BrS data (only retaining components in (2.5) with superscripts BrS), the marginal positive rate for pathogen j is a convex combination of the TPR and FPR:

$$\Pr\left(M_{i'j}^{\mathsf{BrS}} = 1 \mid \pi_j, \theta_j^{\mathsf{BrS}}, \psi_j^{\mathsf{BrS}}\right) = \pi_j \theta_j^{\mathsf{BrS}} + (1-\pi_j)\psi_j^{\mathsf{BrS}}, \tag{2.6}$$

where the left-hand side of the above equation can be estimated by the observed marginal positive rate of pathogen j among cases. Although the control data provide ψ_j^{BrS} estimates, the two parameters, π_j and θ_j^{BrS}, are not both identified. GS data, if available, identifies π_j and resolves the lack of identifiability. Otherwise, we need to incorporate prior scientific information on one of them, usually the TPR (θ_j^{BrS}), derived from infectious disease and laboratory experts (Murdoch et al., 2012) and/or from vaccine probe studies (Feikin et al., 2014). If the observed case marginal positive rate is much higher than the rate in controls (ψ_j^{BrS}), only large values of TPR (θ_j^{BrS}) are supported by the data making etiology estimation more precise (Section 2.2).

In more generality, the full model identification can be characterized by inspecting the Jacobian matrix of the transformation (F) from model parameters ($\boldsymbol{\gamma}$) to the distribution of the observables ($\boldsymbol{p}$): $\boldsymbol{p} = F(\boldsymbol{\gamma})$. Let $\boldsymbol{\gamma} = (\boldsymbol{\theta}^{\mathsf{BrS}}, \boldsymbol{\psi}^{\mathsf{BrS}}, \pi_1, ..., \pi_{J-1})^T$ represent the $3J-1$-dimensional unconstrained model parameters. The pLCM defines the transformation $(\boldsymbol{p}_1, \boldsymbol{p}_0)^T = F(\boldsymbol{\gamma})$, where $\boldsymbol{p}_1$ and $\boldsymbol{p}_0$ are the two contingency probability distributions for the BrS measurements in the case and control populations. It can be shown that the Jacobian matrix $\boldsymbol{\Gamma}(\boldsymbol{\gamma})$ has $J-1$ of its singular values being zero, which means model parameters $\boldsymbol{\gamma}$ are not fully identified from the data. The FPRs ($\psi_j^{\mathsf{BrS}}, j = 1, ..., J$) in pLCM are, however, identifiable parameters that can be estimated from control data. Therefore, pLCM is termed partially identifiable (Jones et al., 2010).

2.2 Parameter estimation and individual etiology prediction

The parameters in likelihood (2.5) include the population etiology distribution ($\boldsymbol{\pi}$), TPRs and FPRs for BrS measurements ($\boldsymbol{\psi}^{\mathsf{BrS}}$ and $\boldsymbol{\theta}^{\mathsf{BrS}}$), and TPRs for SS measurements ($\boldsymbol{\theta}^{\mathsf{SS}}$). The posterior distribution of these parameters can be estimated by constructing approximating samples from the joint posterior via Gibbs sampler. The full conditional distributions for the Gibbs sampler are detailed in Section 1 of the supplementary material.

We use freely available software `WinBUGS 1.4`, to fit the partially-latent class model. Convergence was monitored via Markov chain Monte Carlo (MCMC) chain histories, autocorrelations, kernel density plots, and Brooks-Gelman-Rubin statistics (Brooks and Gelman, 1998). The statistical results below are based on $10,000$ iterations of burn-in followed by $10,000$ production samples from each of three parallel chains.

The Bayesian framework naturally allows individual within-sample classification (infection diagnosis) and out-of-sample prediction. This section describes how we calculate the etiology probabilities for an individual with measurements $\boldsymbol{m}_*$. We focus on the more challenging inference scenario when only BrS data are available; the general case follows directly.

The within-sample classification for case i' is based on the posterior distribution of latent indicators given the observed data, i.e. $\Pr(I^L_{i'} = j \mid \mathcal{D})$, $j = 1, ..., J$, which can be obtained by averaging along the cause indicator ($I^L_{i'}$) chain from MCMC samples. For a case with new BrS measurements $\boldsymbol{m}_*$, we have

$$\Pr(I^L_{i'} = j \mid \boldsymbol{m}_*, \mathcal{D}) \;=\; \int \Pr(I^L_{i'} = j \mid \boldsymbol{m}_*, \boldsymbol{\gamma}) \Pr(\boldsymbol{\gamma} \mid \boldsymbol{m}_*, \mathcal{D}) \mathrm{d}\boldsymbol{\gamma}, j = 1, ...J, \qquad (2.7)$$

where the second factor in the integrand can be approximated by the posterior distribution given current data, i.e., $\Pr(\boldsymbol{\gamma} \mid \mathcal{D})$. For the first term in the integrand, we explicitly obtain the model-based, one-sample conditional posterior distribution, $\Pr(I^L_{i'} = j \mid \boldsymbol{m}_*, \boldsymbol{\gamma}) =$

$\pi_j \ell_j(\boldsymbol{m}_*;\boldsymbol{\gamma}) \Big/ \sum_m \pi_{\boldsymbol{r}_m} \ell_m(\boldsymbol{m}_*;\boldsymbol{\gamma})$, $j = 1, ..., J$, where $\ell_m(\boldsymbol{m}_*;\boldsymbol{\gamma}) = \left(\theta_j^{\mathsf{BrS}}\right)^{m_{*j}} \left(1-\theta_j^{\mathsf{BrS}}\right)^{1-m_{*j}} \prod_{l\neq j} \left(\psi_l^{\mathsf{BrS}}\right.$ is the mth mixture component likelihood function evaluated at $\boldsymbol{m}_*$. The log relative probability of $I_i^L = j$ versus $I_i^L = l$ is

$$R_{jl} = \log\left(\frac{\pi_j}{\pi_l}\right) + \log\left\{\left(\frac{\theta_j^{\mathsf{BrS}}}{\psi_j^{\mathsf{BrS}}}\right)^{m_{*j}} \left(\frac{1-\theta_j^{\mathsf{BrS}}}{1-\psi_j^{BrS}}\right)^{1-m_{*j}}\right\} + \log\left\{\left(\frac{\psi_l^{\mathsf{BrS}}}{\theta_l^{\mathsf{BrS}}}\right)^{m_{*l}} \left(\frac{1-\psi_l^{\mathsf{BrS}}}{1-\theta_l^{\mathsf{BrS}}}\right)^{1-m_{*l}}\right\}.$$

The form of R_{jl} informs us about what is required for correct diagnosis of an individual. Suppose $I_i^L = j$, then averaging over $\boldsymbol{m}_*$, we have $E[R_{jl}] = \log(\pi_j/\pi_l) + I(\theta_j^{\mathsf{BrS}}; \psi_j^{\mathsf{BrS}}) + I(\psi_l^{\mathsf{BrS}}; \theta_l^{\mathsf{BrS}})$, where $I(v_1, v_2) = v_1 \log(v_1/v_2) + (1-v_1)\log((1-v_1)/(1-v_2))$ is the information divergence (Kullback, 2012) that represents the expected amount of information in $m_{*j} \sim$ Bernoulli(v_1) for discriminating against $m_{*j} \sim$ Bernoulli(v_2). If $v_1 = v_2$, then $I(v_1; v_2) = 0$. The form of $E[R_{jl}]$ shows that there is only additional information from BrS data about an individual's etiology in the person's data when there is a difference between θ_j^{BrS} and ψ_j^{BrS}, $j = 1, ..., J$.

Following (2.7), we average $\Pr(I_{i'}^L = j \mid \boldsymbol{m}_*, \boldsymbol{\gamma})$ over MCMC iterations with $\boldsymbol{\gamma}$ replaced by its similated values $\boldsymbol{\gamma})^*$ at each iteration. Repeating for $j = 1, ..., J$, we obtain a J probability vector, $\boldsymbol{p}_{i'} = (p_{i'1}, ..., p_{i'J})^T$, that sums to one. This scheme is especially useful when a newly examined case has a BrS measurement pattern not observed in $\mathcal{D}$, which often occurs when J is large. The final decisions regarding which pathogen to treat can then be based upon estimated $\widehat{\boldsymbol{p}}_{i'}$. In particular, the pathogen with largest posterior value might be selected. It is Bayes optimal under mean misclassification loss. Individual etiology predictions described here generalize the positive/negative predictive value (PPV/NPV) from single to multivariate binary measurements and can aid diagnosis of case subjects under other user-specified misclassification loss functions.

3 Simulation for three pathogens case with GS and BrS data

One key question for studies like PERCH is what fraction of the total evidence about etiology derives from the BrS sources relative to from GS or SS sources if available. In this simulation, we illustrate the extent to which BrS case-control data can supplement observation of the etiologic agent directly from the site of infection. We discuss the role of SS measurements in Section 4 through application to the PERCH data set.

We simulate BrS data sets with 500 cases and 500 controls for three pathogens, A, B, and C using pLCM specifications. We focus on three states to facilitate viewing of the $\boldsymbol{\pi}$ estimates and individual predictions in the 3-dimensional simplex $\mathcal{S}^2$. We use the ternary diagram (Aitchison, 1986) representation where the vector $\boldsymbol{\pi} = (\pi_A, \pi_B, \pi_C)^T$ is encoded as a point with each component being the perpendicular distance to one of the three sides. The parameters involved are fixed at TPR = $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_C)^T = (0.9, 0.9, 0.9)^T$, FPR = $\boldsymbol{\psi} = (\psi_A, \psi_B, \psi_C)^T = (0.6, 0.02, 0.05)^T$, and $\boldsymbol{\pi} = (\pi_A, \pi_B, \pi_C)^T = (0.67, 0.26, 0.07)^T$. We focus on BrS and GS data here and drop the "BrS" superscript on the parameters for simplicity. We further let the fraction of cases with GS measurements (Δ) be either 1% or 10%. Although GS measurements are rare in the PERCH study, we investigate a large range of Δ to understand in general how much statistical information is contained in BrS measurements relative to GS measurements.

For any given data set, three distinct subsets of the data can be used: BrS-only, GS-only, and BrS+GS, each producing its posterior mean of $\boldsymbol{\pi}$, and 95% credible region (Bayesian confidence region) by transformed Gaussian kernel density estimator for compositional data (Chacón et al., 2011). To study the relative importance of the GS and BrS data, the primary quantity of interest in the simulations is the relative sizes of the credible regions

for each data mix. Here, we use uniform priors on $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, and Dirichlet$(1, ..., 1)$ prior for $\boldsymbol{\pi}$. The results are shown in Figure 2.

First, in Figures 2(a) (1% GS) and 2(b) (10% GS), each region covers the true etiology $\boldsymbol{\pi}$. In data not shown here, the nominal 95% credible regions covers slightly more than 95% of 100 simulations. Credible regions narrow in on the truth as we combine BrS and GS data, and as the fraction of subjects with GS data (Δ) increases. Also, the posterior mean from the BrS+GS analysis is a result of optimal balance between information contained in the GS and BrS data.

We then fix $\boldsymbol{\psi}$ and $\boldsymbol{\pi}$, while varying the TPR θ on the grid $(0.6, 0.7, 0.8, 0.9, 0.95, 0.99)$ to test the estimation performance under a variety of signal-to-noise ratios, measured by the difference between the TPRs and FPRs. At each (θ, Δ) grid point, we run analyses on each of 100 simulated data sets. We quantify the gain in precision by adding the BrS data to the GS data following Xu and Zeger (2001). For pathogen A, let $g_A(\theta) = \left(d_A^0 - d_A^{\mathsf{BrS}+\mathsf{GS}}(\theta)\right) / \left(d_A^0 - d_A^{\mathsf{GS}}(\theta)\right)$, where d_A^0, $d_A^{\mathsf{GS}}(\theta)$ and $d_A^{\mathsf{BrS}+\mathsf{GS}}(\theta)$ are the length of 95% highest density interval from the prior, length of 95% credible interval using GS data, and length of the 95% credible interval using BrS and GS data, respectively. This quantity ($g_A(\theta)$) is the ratio of the reduction of the 95% interval widths with and without the BrS data at TPR value θ. If $g_A(\theta) = 1$, then there is no additional gain in the precision of π_A when BrS data is added to GS data. When $\Delta = 1\%$, we observe the expected increase in g_A as TPR θ approaches 1. For pathogen A, $g_A(0.8)$ has mean value 1.7 across 100 simulated data sets with standard error 0.3; $g_A(0.95)$ further increase to 3.0(standard error 0.3). Similar patterns are also observed for pathogen B and C.

Using the same simulated data sets, Figures 2(a) and 2(b) also show individual etiology predictions for each of the $8(= 2^3)$ possible BrS measurements $(m_A, m_B, m_C)^T, m_j = 0, 1$, obtained by the methods from Section 2.2. Consider the example of a newly enrolled case

without GS data and with no pathogen observed in her BrS data: $\boldsymbol{m} = (0, 0, 0)$. Suppose she is part of a case population with 10% GS data. In the case illustrated in Figure 2(b), her posterior predictive distribution has highest posterior probability (0.76) on pathogen A reflecting two competing forces: the FPRs that describe background colonization (colonization among the controls) and the population etiology distribution; Given other parameters, $\boldsymbol{m} = (0, 0, 0)$ gives the smallest likelihood for $I_i^L = A$ because of its high FPR that reflects its background colonization rate, $\psi_A = 0.6$. However, prior to observing $(0, 0, 0)$, π_A is well estimated to be much larger than π_B and π_C. Therefore the posterior distribution for this case is heavily weighted towards pathogen A.

For a case with observation $(1, 1, 1)$, because it is rare to observe pathogen B in a case whose pneumonia is not caused by B, the prediction favors B. Although B is not the most prevalent cause among cases, the presence of B in the BrS measurements gives the largest likelihood when $I_i^L = B$. For any measurement pattern with a single positive, the case is always classified into that category in this example.

Most predictions are stable with increasing Δ. Only 000 cases have predictions that move from near the center to the corner of A. This is mainly because that TPR $\boldsymbol{\theta}$ and etiology fractions $\boldsymbol{\pi}$ are not as precisely estimated in GS-scarce scenarios relative to GS-abundant ones. Averaging over a wider range of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ produces 000 case predictions that are ambiguous, i.e. near the center. As Δ increases, parameters are well estimated, and precise predictions result.

4 Analysis of PERCH data

The Pneumonia Etiology Research for Child Health (PERCH) study is a standardized and comprehensive evaluation of etiologic agents causing severe and very severe pneumonia among hospitalized children aged 1-59 months in seven low and middle income countries.

The study sites include countries with a significant burden of childhood pneumonia and a range of epidemiologic characteristics (Levine et al., 2012). PERCH is a case-control study that has enrolled over $4,000$ patients hospitalized for severe or very severe pneumonia and over $5,000$ controls selected randomly from the community frequency-matched on age in each month. More details about the PERCH design are available in Deloria-Knoll et al. (2012).

To illustrate the application of pLCM model for the analysis of PERCH study data, we have focused on preliminary data from one site with good availability of both SS and BrS laboratory results. Results for all 7 countries will be reported elsewhere upon study completion. Included in the current illustrative analysis are BrS data (nasopharyngeal specimen with PCR detection of pathogens) for 432 cases and 479 frequency-matched controls on 11 species of pathogens (7 viruses and 4 bacteria with their abbreviations in Figure 3, and full names in Section 2 of the supplementary material), and SS data (blood culture results) on the 4 bacteria for only the cases.

In PERCH, prior scientific knowledge of measurement error rates is incorporated into the analysis. The TPR of our BrS measurements, θ_j^{BrS} is assumed to be in the range of $90\% - 97\%$ (Murdoch et al., 2012). Observations from vaccine probe studies—randomized clinical trials of pathogen-specific vaccines in which non-specific clinical endpoints such as clinical pneumonia are evaluated thereby revealing the contribution of the pathogen to the burden of that syndrome— illustrate that the total number of clinical pneumonia cases prevented by the vaccine is much larger than the few laboratory-confirmed cases prevented. Comparing the total preventable disease burden to the number of blood culture (SS) positive cases prevented provides information about the TPR of the bacterial blood culture measurements, $\theta_j^{\mathsf{SS}}, j = 1, ..., 4$. In our analysis, we use the range $10 - 20\%$ for the SS TPRs of four bacteria. We set Beta priors that match these ranges (Section 2) and

assumed Dirichlet$(1, ..., 1)$ prior on etiology fractions $\boldsymbol{\pi}$.

In latent variable models like the pLCM, key variables are not directly observed. It is therefore essential to picture the model inputs and outputs side-by-side to better understand the analysis performed. In this spirit, Figure 3 displays for each of the 11 pathogens, a summary of the BrS and SS data in the left two columns, along with some of the intermediate model results; and the prior and posterior distributions for the etiology fractions on the right (rows ordered by posterior means). The observed BrS rates (with 95% confidence intervals) for cases and controls are shown on the far left with solid dots. The conditional odds ratio contrasting the case and control rates given the other pathogens is listed with 95% confidence interval in the box to the right of the BrS data summary. Below the case and control observed rates is a horizontal line with a triangle. From left to right, the line starts at the estimated false positive rate (FPR, $\widehat{\psi}_j^{\mathsf{BrS}}$) and ends at the estimated true positive rate (TPR, $\widehat{\theta}_j^{\mathsf{BrS}}$), both obtained from the model. Below the TPR are two boxplots summarizing its posterior (top) and prior (bottom) distributions for that pathogen. These box plots show how the prior assumption influences the TPR estimate as expected given the identifiability constraints discussed in Section 2.1. The triangle on the line is the model estimate of the case rate to compare to the observed value above it. As discussed in Section 2.1, the model-based case rate is a linear combination of the FPR and TPR with mixing fraction equal to the estimated etiology fraction. Therefore, the location of the triangle, expressed as a fraction of the distance from the FPR to the TPR, is the model-based point estimate of the etiologic fraction for each pathogen. The SS data are shown in a similar fashion to the right of the BrS data. By definition, the FPR is 0.0 for SS measures and there is no control data. The observed rate for the cases is shown with its 95% confidence interval. The estimated SS TPR ($\widehat{\theta}_j^{\mathsf{SS}}$) with prior and posterior distributions is shown as for the BrS data, except that we plot 95% and 50% credible intervals for SS TPR above

its prior distribution boxplot.

On the right side of the display are the marginal posterior and prior distributions of the etiologic fraction for each pathogen. We appropriately normalized each density to match the height of the prior and posterior curves. The posterior mean with 50% and 95% credible intervals are shown above the density.

Figure 3 shows that respiratory syncytial virus (RSV), *Streptococcus pneumoniae* (PNEU), rhinovirus (RHINO), and human metapneumovirus (HMPV_A_B) occupy the greatest fractions of the etiology distribution, from 10% to 30% each. That RSV has the largest estimated mean etiology fraction reflects the large discrepancy between case and control positive rates in the BrS data: 25.3% versus 0.8% (marginal odds ratio 38.5 (95%CI $(18, 128.7)$) as shown on the left of the display. RHINO has marginal case and control rates that are close to each other, yet its estimated mean etiology fraction is 15.9%. This is because the model considers the joint distribution of the pathogens, not the marginal rates. The conditional odds ratio of case status with RHINO given all the other pathogen measures is estimated to be 1.5 $(1.1, 2.1)$ as compared to the marginal odds ratio close to 1 $(0.8, 1.3)$.

As discussed in Section 2.1, the data alone cannot precisely estimate both the etiologic fractions and TPRs absent prior knowledge. This is evidenced by comparing the prior and posterior distributions for the TPRs in the BrS boxes for each pathogen (i.e. left hand column of Figure 3). The posteriors are similar to their priors indicating little else about TPR is learned from the data. The posteriors for some pathogens making up $\boldsymbol{\pi}$ (i.e. shown in the right hand column of Figure 3) are likely to be sensitive to the prior specifications of the TPRs.

We performed sensitivity analyses using multiple sets of priors for the TPRs. At one extreme, we ignored background scientific knowledge and let the priors on the FPR and TPR be uniform for both the BrS and SS data. The results are shown in Figure 7. Ignoring

prior knowledge about error rates lowers the etiology estimates of the bacteria PNEU and *Haemophilus influenzae* (HINF). The substantial reduction in the etiology fraction for PNEU, for example, is a result of the difference in the TPR prior for the SS measurements. In the original analysis (Figure 3), the informative prior on the SS sensitivity (TPR) place 95% mass between $10-20\%$. Hence the model assumes almost 85% of the PNEU infections are being missed in the SS sampling. When a uniform prior is substituted (Figure 4), the fraction assumed missed is greatly reduced. For RSV, its posterior mean etiology fraction increases from 27.3% to 31.7%. The etiology estimates for other pathogens are fairly stable, with changes in posterior means between -0.4% and 3.3%.

Under the original priors for TPR, PARA1 has an estimated etiologic fraction of 5.2%, even though it has conditional odds ratio 5.8 $(2.5, 15)$. In general, pathogens with larger conditional odds ratios have larger etiology fraction estimates. Also, a pathogen still needs a reasonably high observed case positive rate to be allocated a high etiology fraction. The posterior etiology fraction estimate of 5.2% for PARA1 results because the prior for the TPR takes values in the range of $0.9 - 0.97$. By Equation (2.6), the TPR weight in the convex combination with FPR (around 1.5%) has to be very small to explain the small observed case rate 5.5%. When a uniform prior is placed on TPR instead, the PARA1 etiology fraction increases to 10.2% with a wider 95% credible interval (Figure 4).

Furthermore, when uniform priors on TPR and FPR are used, PARA1 is still allocated a smaller etiology fraction than RHINO despite PARA1 having a larger conditional odds ratio. This is related to the dependence structure among case measurements. RHINO has the highest negative association with RSV among cases (standardized log odds ratio -14). Under the conditional independence assumption of the pLCM, this dependence is partly induced by multinomial correlation among the latent cause indicators: $I_i^L = \text{RSV}$ versus $I_i^L = \text{RHINO}$ that is $-\pi_{\text{RSV}}\pi_{\text{RHINO}}$. RSV has strong evidence as a frequent cause

with a stable estimate $\widehat{\pi}_{\text{RSV}}$ around 30%. The strong negative association in the cases' measurements between RHINO and RSV is contributing to the increased etiologic fraction estimate $\widehat{\pi}_{\text{RHINO}}$ relative to other pathogens that have less or no association with RSV among the cases. The conditional independence assumption is leveraging information from the associations between pathogens in estimation of the etiologic fractions.

We have checked the model in two ways by comparing the characteristics of the observed measurements joint distribution with the same characteristic for the distribution of new measurements generated by the model from a population of the same size. By generating the new data characteristics at every iteration of the MCMC chain, we can integrate the predictive distribution over the posterior distribution of the parameters as discussed in Garrett and Zeger (2000). Figure 5 displays the observed frequency of the 10 most common measurement outcomes for the BrS data, separately for cases and controls to compare to the predictive distributions based upon the model. Among the cases, the 95% predictive interval includes the observed values in all but two of the BrS patterns and even there the fits are reasonable. Among the controls, there is evidence of lack of fit for the most common BrS pattern with only PNEU and HINF. There are fewer cases with this pattern observed than predicted under the pLCM. This lack of fit is due to associations of pathogen measurements in control subjects. Note that the FPR estimates remain consistent regardless of such correlation as the number of controls increases, however posterior variances for them may be underestimated.

Figure 6 presents standardized log odds ratios (SLORs) for cases (lower triangle) and controls (upper triangle). Each entry is the observed log odds ratio for a pair of BrS measurements minus the mean LOR for the predictive data distribution value divided by the standard deviation of the LOR predictive distribution. The first significant digit of the absolute SLOR is shown in blue for negative and red for positive values. Absolute SLORs

less than 2 are omitted from the table for graphical effect. We see two large deviations among the cases: RSV with RHINO and RSV with HMPV. These are caused by strong seasonality in RSV that is out of phase with weaker seasonality in the other two. Otherwise, the associations are roughly what is expected under the assumed model.

An attractive feature of using MCMC to estimate posterior distributions is the ease of estimating posteriors for functions of the latent variables and/or parameters. One interesting question from a clinical perspective is whether viruses or bacteria are the major cause and among each subgroup, which species predominate. Figure 7 shows the posterior distribution using expert TPR prior for viruses versus bacteria on the top, and then the conditional distributions of the two leading bacteria (viruses) among bacterial (viral) causes below. The posterior shape of the viral etiologic fraction is more concentrated compared to the prior shape, with mode around 63% and 95% credible interval $(54\%, 71\%)$. Of all viral cases, RSV is estimated to cause about 43% $(36\%, 51\%)$, and RHINO about 25% $(17\%, 34\%)$. PNEU accounts for most bacterial cases (71% $(48\%, 87\%)$), and HINF accounts for 19% $(4\%, 42\%)$. In both the viral and bacterial categories, the 95% credible intervals for the first most common pathogen does not overlap that of the second most common one.

5 Discussion

In this paper, we estimated the frequency with which pathogens cause disease in a case population using a partially-latent class model (pLCM) to allow for known states for a subset of subjects and for multiple types of measurements with different error rates. In a case-control study of disease etiology, measurement error will bias estimates from traditional logistic regression and attributable fraction methods. The pLCM avoids this pitfall and more naturally incorporates multiple sources of data. Here we considered three levels of

measurement error rates.

Absent GS data, we show that the pLCM is only partially identified because of the relationship between the estimated TPR and prevalence of the associated pathogen in the population. Therefore, the inferences are sensitive to the assumptions about the TPR. Uncertainty about their values persists in the final inferences from the pLCM regardless of the number of subjects studied.

The current model provides a novel solution to the analytic problems raised by the PERCH Study. This paper illustrates the design and application of the pLCM using a preliminary and limited set of data from one PERCH study site. Confirmatory laboratory testing, incorporation of additional pathogens, and adjustment for various factors are likely to change the scientific findings that will be reported in the complete analysis of the study results.

An essential assumption relied upon in the pLCM is that the probability of detecting one pathogen at a peripheral body site depends on whether that pathogen is infecting the child's lung, but is unaffected by the presence of other pathogens in the lung, that is, the non-differential misclassification error assumption, $[\boldsymbol{M}_{ij}^{\mathsf{SS}} \mid I_i^L = l] = [\boldsymbol{M}_{ij}^{\mathsf{SS}} \mid I_i^L = k]$, $\forall l, k \neq j$. We have formulated the model to include GS measures even though they are infrequently available from PERCH cases. In general, the availability of GS measures makes it possible to test this assumption as has been discussed by Albert and Dodd (2008).

Several extensions have potential to improve the quality of inferences drawn and are being developed for PERCH. First, because the control subjects have known class, we can model the dependence structure among the BrS measurements and use this to avoid aspects of the conditional independence assumption central to most LCM methods. The approach is to extend the pLCM to have K subclasses within each of the current disease classes. These subclasses can introduce correlation among the BrS measurements given

the true disease state. An interesting question is about the bias-variance trade-off for different values of K. This ideas follows previous work on the PARAFAC decomposition of probability distribution for multivariate categorical data (Dunson and Xing, 2009). This extension will enable model-based checking of the standard pLCM.

Second, in our analyses to date, we have assumed that the pneumonia case definition is error-free. Given new biomarkers and availability of chest radiograph that can improve upon the clinical diagnosis of pneumonia, one can introduce an additional latent variable to indicate true disease status and use these measurements to probabilistically assign each subject as a case or control. Finally, regression extensions of the pLCM will allow PERCH investigators to study how the etiology distributions vary with HIV status, age group, and season.

Acknowledgments

We thank the members of the larger PERCH Study Group for discussions that helped shape the statistical approach presented herein, and the study participants. We also thank the members of PERCH Expert Group who provided external advice.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.

Albert, P. and Dodd, L. (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, 103(481):61–73.

Albert, P., McShane, L., and Shih, J. (2001). Latent class modeling approaches for assessing

diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57(2):610–619.

Blackwelder, W., Biswas, K., Wu, Y., Kotloff, K., Farag, T., Nasrin, D., Graubard, B., Sommerfelt, H., and Levine, M. (2012). Statistical methods in the global enteric multicenter study (gems). *Clinical infectious diseases*, 55(suppl 4):S246–S253.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

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

Chacón, J., Mateu-Figueras, G., and Martín-Fernández, J. (2011). Gaussian kernels for density estimation with compositional data. *Computers & Geosciences*, 37(5):702–711.

Deloria-Knoll, M., Feikin, D., Scott, J., OBrien, K., DeLuca, A., Driscoll, A., Levine, O., et al. (2012). Identification and selection of cases and controls in the pneumonia etiology research for child health project. *Clinical Infectious Diseases*, 54(suppl 2):S117–S123.

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

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

Garrett, E. and Zeger, S. (2000). Latent class model diagnosis. *Biometrics*, 56(4):1055–1067.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111–140.

Gustafson, P. (2009). What are the limits of posterior distributions arising from nonidentified models, and why should we care? *Journal of the American Statistical Association*, 104(488):1682–1695.

Gustafson, P., Le, N., and Saskin, R. (2001). Case–control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*, 57(2):598–609.

Hui, S. and Walter, S. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171.

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.

Kadane, J. (1974). The role of identification in bayesian theory. *Studies in Bayesian Econometrics and Statistics*, pages 175–191.

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

Kullback, S. (2012). *Information theory and statistics*. Courier Dover Publications.

Levine, O., OBrien, 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.

Murdoch, D., OBrien, K., Driscoll, A., Karron, R., Bhat, N., et al. (2012). Laboratory methods for determining pneumonia etiology in children. *Clinical Infectious Diseases*, 54(suppl 2):S146–S152.

Qu, Y. and Hadgu, A. (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association*, 93(443):920–928.

Wang, Z., Zhou, X., and Wang, M. (2011). Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. *Biostatistics*, 12(3):567–581.

Xu, J. and Zeger, S. (2001). The evaluation of multiple surrogate endpoints. *Biometrics*, 57(1):81–87.

A Full conditional distributions in Gibbs sampler

In this section, we provide analytic forms of full conditional distributions that are essential for Gibbs sampling algorithm. We use data augmentation scheme by introducing latent lung state I_i^L into the sampling chain and we have the following full conditional distributions:

- $\left[I_i^L \mid \text{others}\right]$. If M_i^{GS} is available, $\Pr\left(I_i^L = j \mid \text{others}\right) = 1$, if $M_{ij}^{\mathsf{GS}} = 1$ and $M_{il}^{\mathsf{GS}} = 0$, for $l \neq j$; otherwise zero. If M_i^{GS} is missing, according as whether M_i^{SS} is available, the full conditional is given as

$$\begin{aligned}\Pr(I_i^L = j \mid \text{others}) \propto & \\ & \left(\theta_j^{\mathsf{BrS}}\right)^{M_{ij}^{\mathsf{BrS}}} \left(1-\theta_j^{\mathsf{BrS}}\right)^{1-M_{ij}^{\mathsf{BrS}}} \prod_{l\neq j}\left(\psi_l^{\mathsf{BrS}}\right)^{M_{il}^{\mathsf{BrS}}} \left(1-\psi_l^{\mathsf{BrS}}\right)^{1-M_{il}^{\mathsf{BrS}}} \\ & \cdot \left[\left(\theta_j^{\mathsf{SS}}\right)^{M_{ij}^{\mathsf{SS}}} (1-\theta_j^{\mathsf{SS}})^{1-M_{ij}^{\mathsf{SS}}} \mathbf{1}_{\left\{\sum_{l\neq j} M_{il}^{\mathsf{SS}} = 0\right\}}\right]^{\mathbf{1}_{\{j\leq J'\}}} \cdot \pi_j; \end{aligned} \tag{A.1}$$

  if SS measurement is not available for case i, we remove terms involving M_{ij}^{SS}.

- $\left[\psi_j^{\mathsf{BrS}} \mid \text{others}\right] \sim \text{Beta}\left(N_j + b_{1j}, n_1 - \sum_{i:Y_i=1} \mathbf{1}_{\{I_i^L=j\}} + n_0 - N_j + b_{2j}\right)$, where n_1 and n_0 are number of cases and controls, respectively, and $N_j = \sum_{i:Y_i=1, I_i^L\neq j} M_{ij}^{\mathsf{BrS}} + \sum_{i:Y_i=0} M_{ij}^{\mathsf{BrS}}$ is the number of positives at position j for cases with $I_i^L \neq j$ and all controls.

- $\left[\theta_j^{\mathsf{BrS}} \mid \text{others}\right] \sim \text{Beta}\left(S_j + c_{1j}, \sum_{i:Y_i=1} \mathbf{1}_{\{I_i^L=j\}} - S_j + c_{2j}\right)$, where $S_j = \sum_{i:Y_i=1, I_i^L=j} M_{ij}^{\mathsf{BrS}}$ is the number of positives for cases with jth pathogen as their causes.

- $\left[\theta_j^{\mathsf{SS}} \mid \text{others}\right] \sim \text{Beta}\left(T_j + d_{1j}, \sum_{i:Y_i=1,\mathsf{SS}\text{ available}} \mathbf{1}_{\{I_i^L=j\}} - T_j + d_{2j}\right)$, where

$$T_j = \sum_{i:Y_i=1, I_i^L=j, \mathsf{SS}\text{ available}} M_{ij}^{\mathsf{SS}}.$$

  When no SS data is available, this conditional distribution reduces to $\text{Beta}(d_{1j}, d_{2j})$, the prior.

- $\left[\boldsymbol{\pi} \mid I_i^L, i : Y_i = 1\right] \sim \text{Dirichlet}(a_1 + U_1, ..., a_J + U_J)$, where $U_j = \sum_{i:Y_i=1} \mathbf{1}_{\{I_i^L=j\}}$.

B Pathogen names and their abbreviations

Bacteria: HINF- *Haemophilus influenzae*; PNEU-*Streptococcus pneumoniae*; SASP-*Salmonella* species; SAUR-*Staphylococcus aureus*.

Viruses: ADENOVIRUS-adenovirus; COR_43-coronavirus OC43; FLU_C-influenza virus type C; HMPV_A_B-human metapneumovirus type A or B; PARA1-parainfluenza type 1 virus; RHINO-rhonovirus; RSV_A_B-respiratory syncytial virus type A or B.

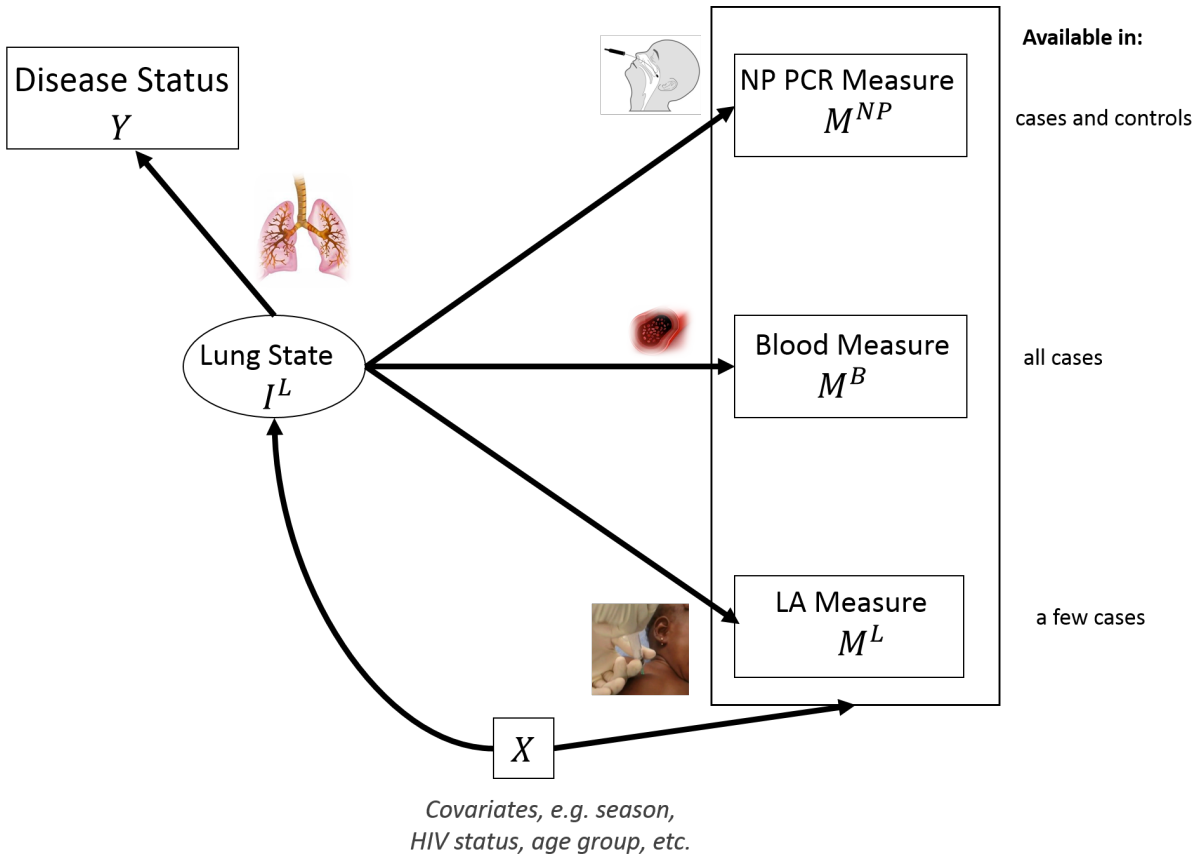
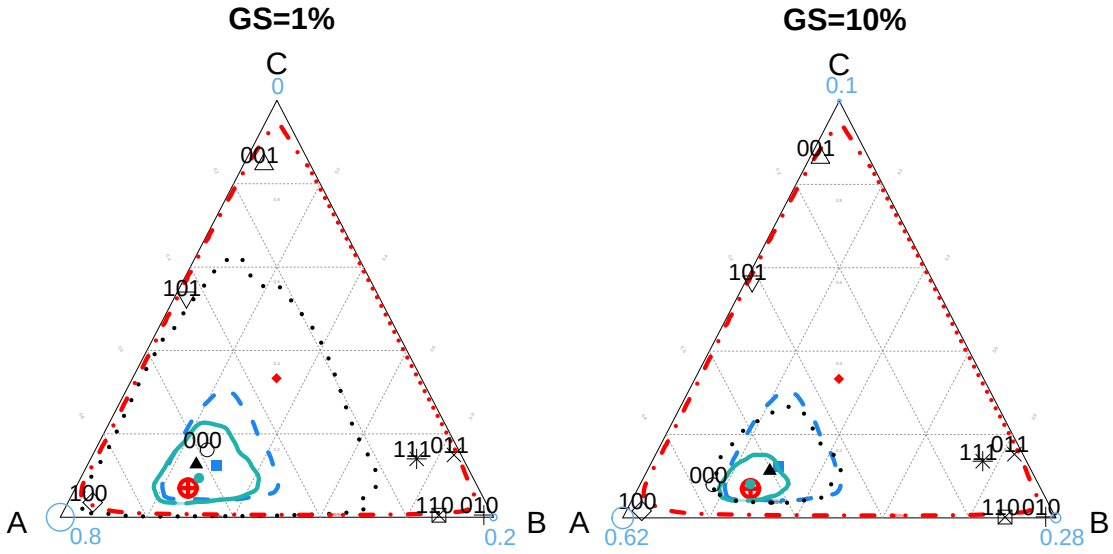


Figure 1: Directed acyclic graph (DAG) illustrating relationships among lung infection state (I^L), imperfect lab measurements on the presence/absence of each of a list of pathogens at each site(M^{NP}, M^B and M^L), disease outcome, and covariates (X). For a subject missing one or more of the three types of measurements, we remove the corresponding measurement component(s). For example, if a case does not have lung aspirate (LA) measurement, we remove M^L from the DAG.



(a) (b)

Figure 2: Population and individual etiology estimations for a single sample with 500 cases and 500 controls with true $\boldsymbol{\pi} = (0.67, 0.26, 0.07)^T$ and either $1\%N = 5$) or $10\%(N = 50)$ GS data on cases. In (a) or (b), *Red circled plus* shows the true population etiology distribution $\boldsymbol{\pi}$. The *closed curves* are 95 percent credible regions: *blue dashed lines* "- - -", *light green solid lines* "—", *black dotted lines* "$\cdots$" correspond to analysis using BrS data only, BrS+GS data, GS data only, respectively; *Solid square/dot/triangle* are corresponding posterior means of $\boldsymbol{\pi}$; The 95 percent highest density region of uniform prior distribution is also visualized by red "$\cdot - \cdot$—" for comparison. $8(= 2^3)$ BrS measurement patterns and predictions for individual children are shown with different shapes, with measurement patterns attached to them. The radii of circles and numbers at the vertices show empirical frequencies GS measurements belonging to A, B, or C.

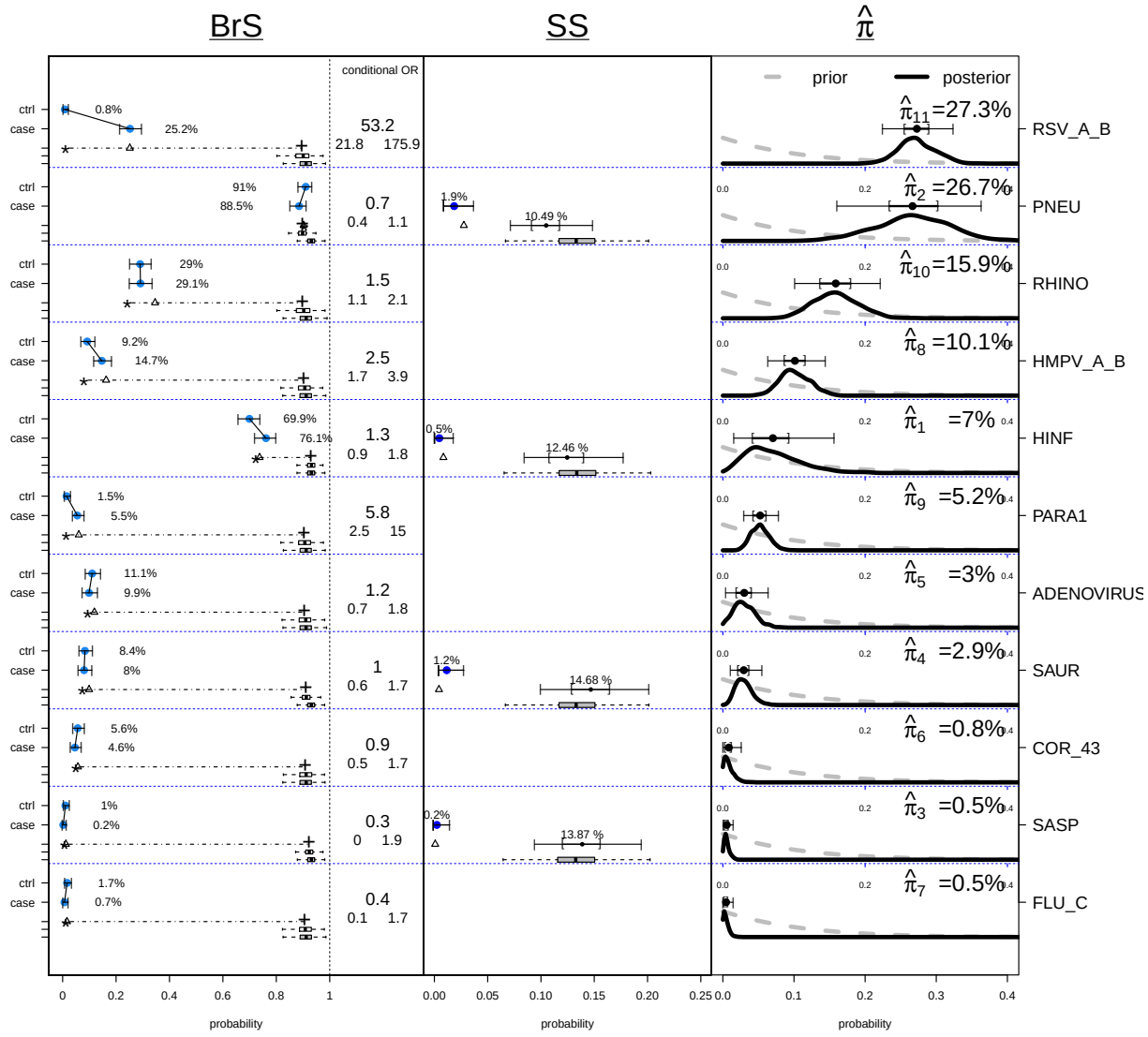


Figure 3: Results using expert priors on TPRs. The observed BrS rates (with 95% confidence intervals) for cases and controls are shown on the far left. The conditional odds ratio given the other pathogens is listed with 95% confidence interval in the box to the right of the BrS data summary. Below the case and control observed rates is a horizontal line with a triangle. From left to right, the line starts at the estimated false positive rate (FPR, $\widehat{\psi}_j^{\mathsf{BrS}}$) and ends at the estimated true positive rate (TPR, $\widehat{\theta}_j^{\mathsf{BrS}}$), both obtained from the model. Below the TPR are two boxplots summarizing its posterior (top) and prior (bottom) distributions. The location of the triangle, expressed as a fraction of the distance from the FPR to the TPR, is the model-based point estimate of the etiologic fraction for each pathogen. The SS data are shown in a similar fashion to the right of the BrS data. The observed rate for the cases is shown with its 95% confidence interval. The estimated SS TPR ($\widehat{\theta}_j^{\mathsf{SS}}$) with prior and posterior distributions is shown as for the BrS data, except that we plot 95% and 50% credible intervals for SS TPR above the boxplot for its prior distribution. See Appendix for pathogen name abbreviations.

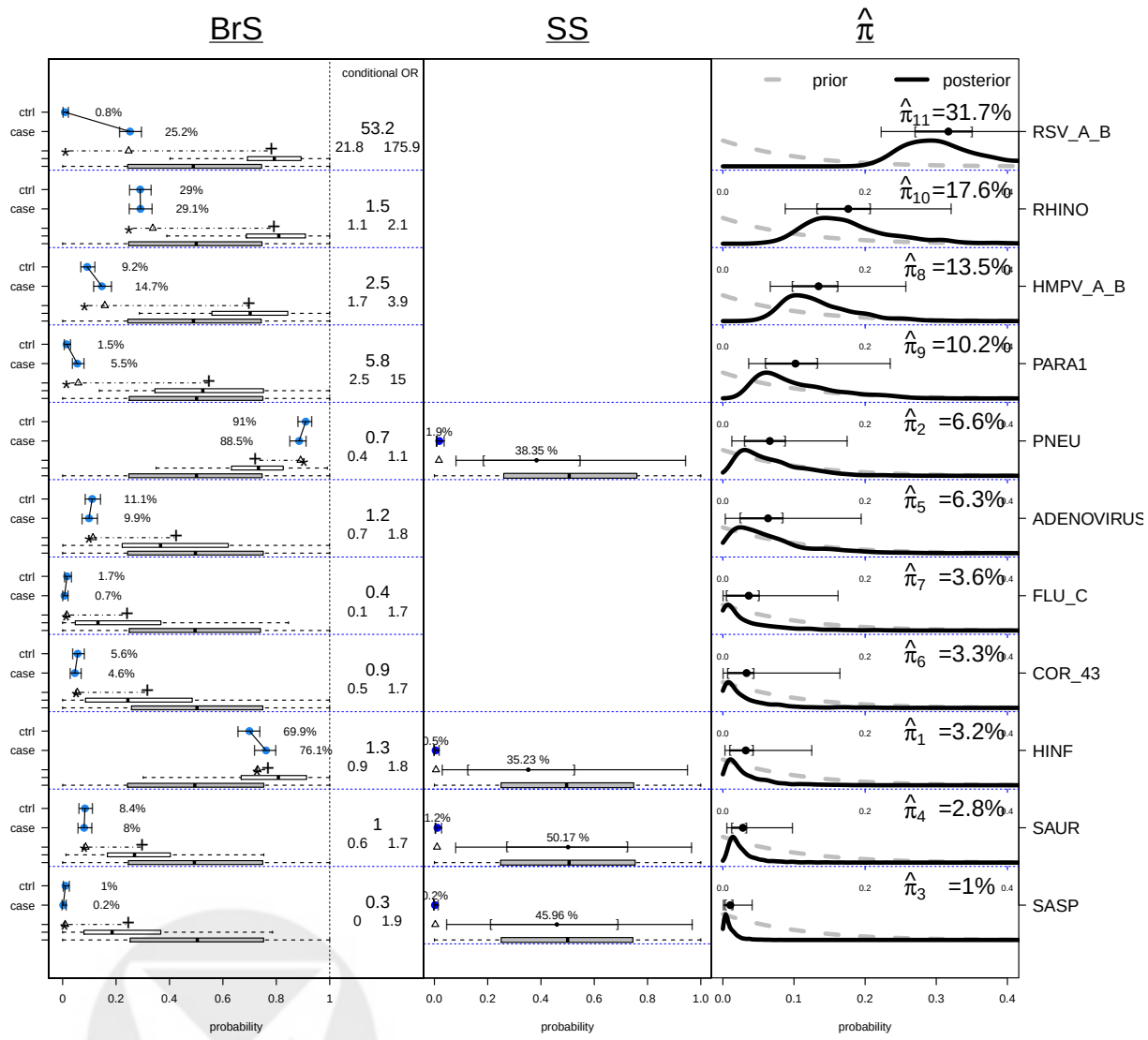


Figure 4: Results on using uniform priors on TPRs. As in Figure 3 with uniform priors on the TPRs.

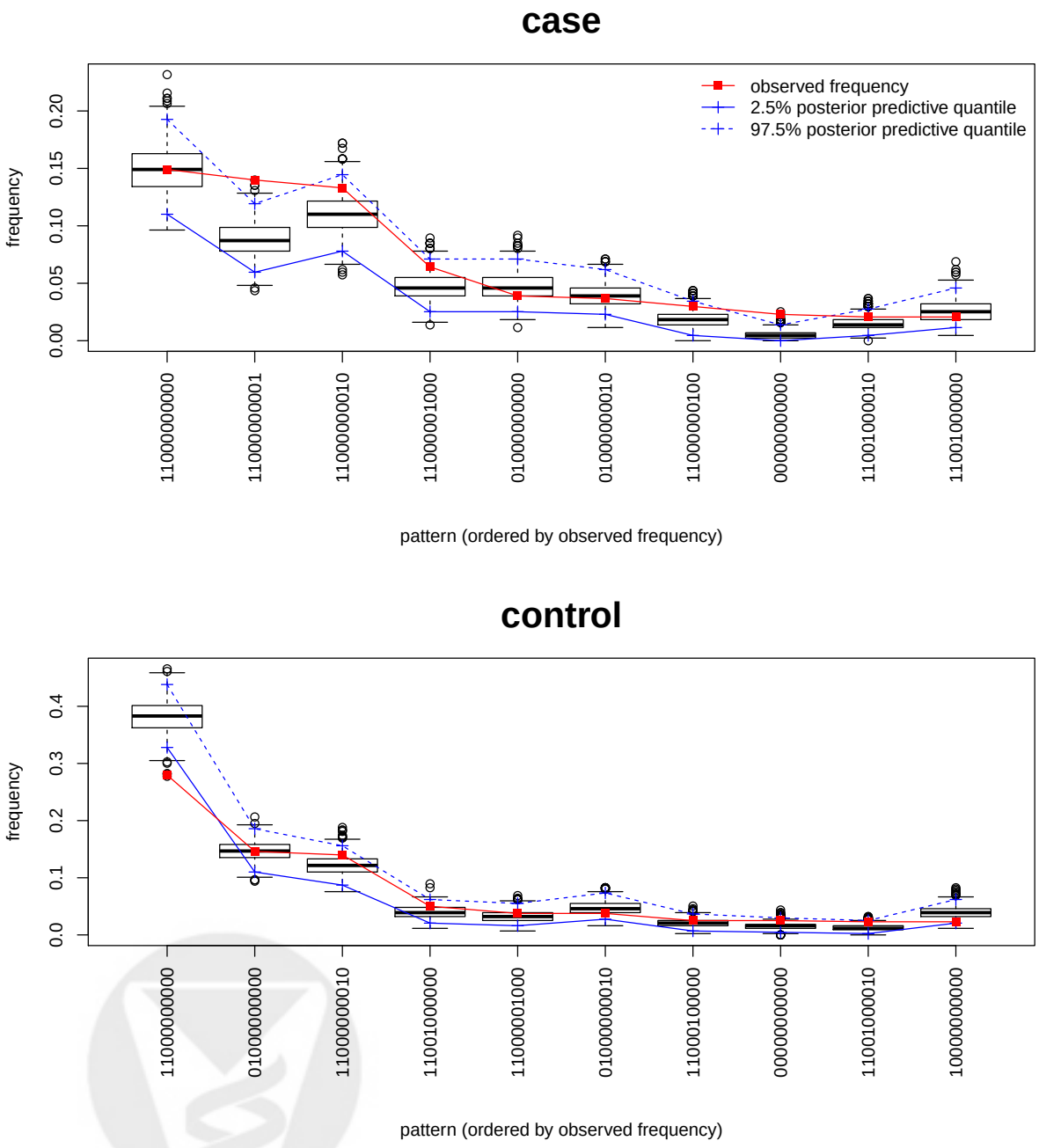


Figure 5: Posterior predictive checking for 10 most frequent BrS measurement patterns among cases and controls with expert priors on TPRs.

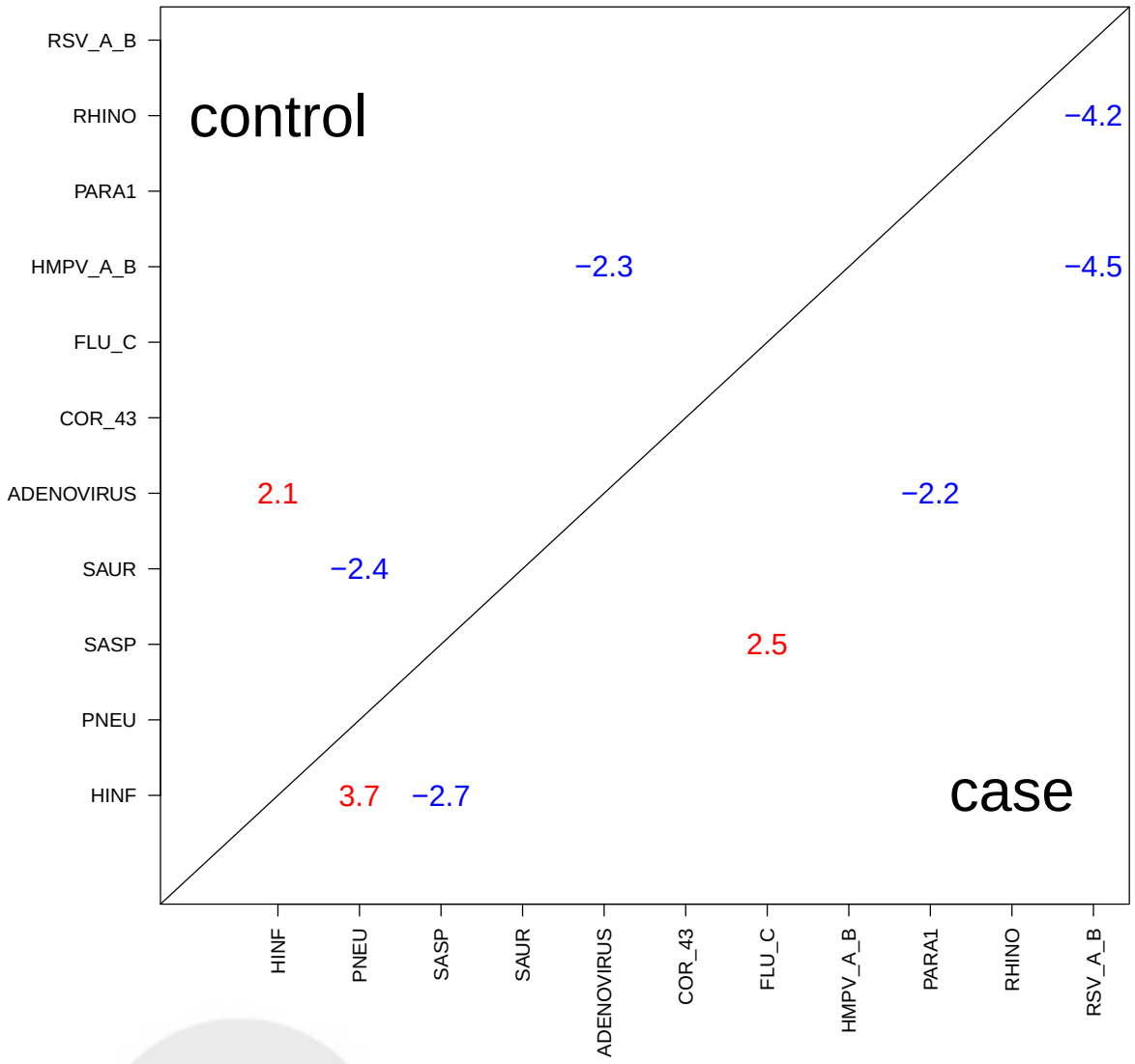


Figure 6: Posterior predictive checking for pairwise odds ratios separately for cases (lower right triangle) and controls (upper left triangle) with expert priors on TPRs. Each entry is a standardized log odds ratio (SLOR): the observed log odds ratio for a pair of BrS measurements minus the mean LOR for the posterior predictive distribution divided by the standard deviation of the posterior predictive distribution. The first significant digit of absolute SLORs are shown in red for positive and blue for negative values, and only those greater than 2 are shown.

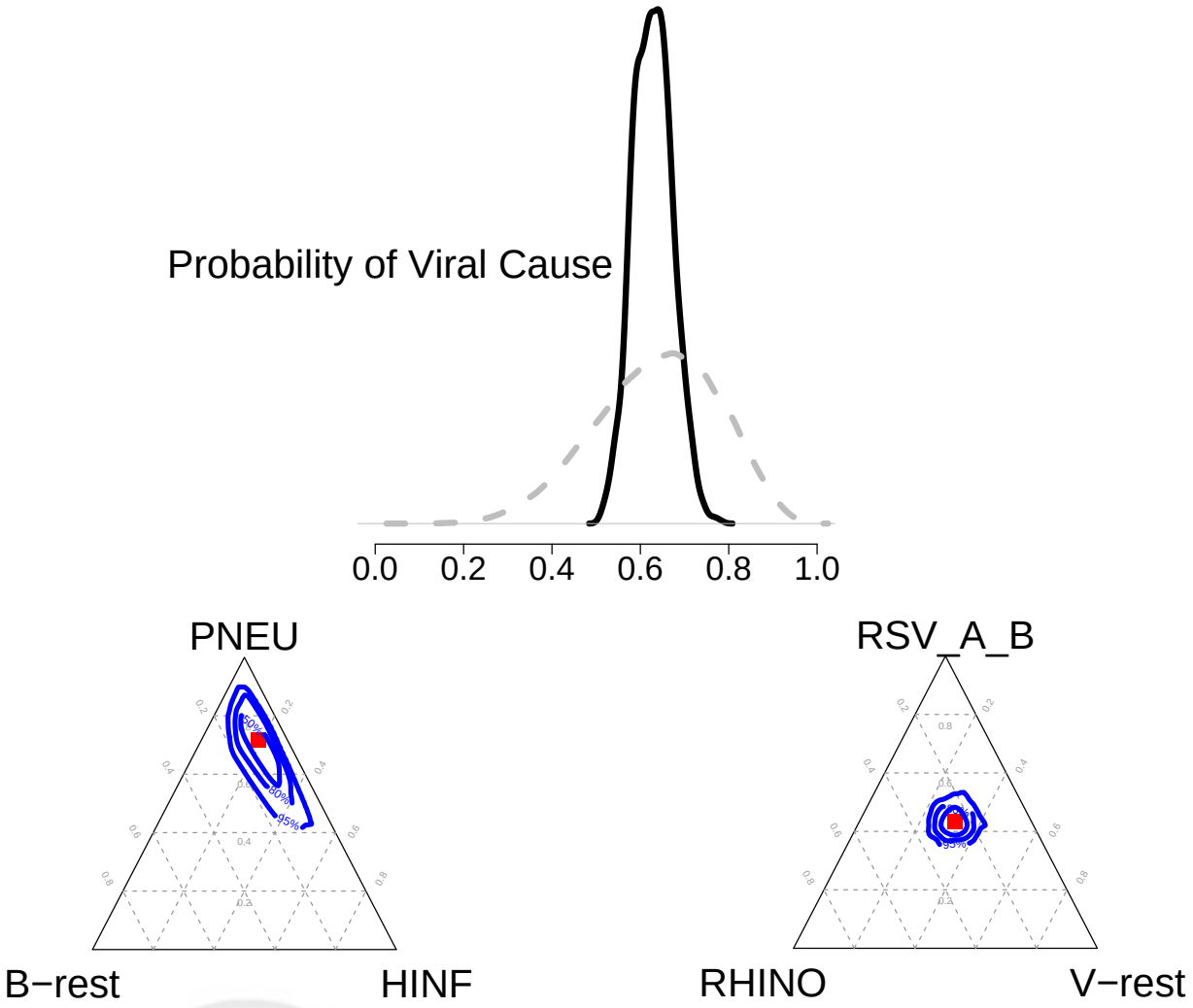


Figure 7: Summary of posterior distribution of pneumonia etiology estimates using expert (left) and uniform (right) priors on TPRs. In each subfigure, top: posterior (solid) and prior (dashed) distribution of viral etiology; bottom left: posterior etiology distribution for top two bacterial causes given bacteria is a cause; bottom right: posterior etiology distribution for top two viral causes given virus is a cause. B-rest and V-rest stand for the rest of bacteria and viruses other than the top two species, respectively. The nested blue circles are 95%, 80%, and 50% credible regions for population etiology estimates within bacterial or viral group.