Bayesian latent class model with sparse correlation for etiology estimation

Detian Deng

February 13, 2017

1 Introduction

1.1 Background

1.1.1 Child Pneumonia

Pneumonia is a form of acute respiratory infection of the lungs[1]. The infection can be caused by a variety of pathogens, indcluding bacteria, viruses, mycobacteria and fungi [24]. When a child under five gets pneumonia, the typical symptoms may include fever, cough, fast or difficult breathing, lower chest wall indrawing where the chest moves in or retracts during inhalation, and wheezing [39, 1]. Severe cases may be unable to feed or drink and may also experience unconsciousness, hypothermia and convulsions [1]. Although the majority of child pneumonia cases are nonsevere and can be managed in local primary health care facilities [27], the severe/very severe cases may result in death, especially in developing countries. In fact, pneumonia is the single largest infectious cause of death of children under 5 years of age (referred to "children" for the rest part of this article) , with an estimate of 1.6 million deaths per year accounting for 18% of the total 8.8 million childhood deaths worldwide [31, 8]. Under the pressure of such a severe public health burden, UNICEF and WHO declared pneumonia to be the "forgotten killer of children" in 2006 [43] and engaged the Global Action Plan for Prevention and Control of Pneumonia (GAPP) [35] in 2009.

1.1.2 The Need of New Etiology Information

Current prevention and treatment strategies for pneumonia were primarily developed based on the results of early pneumonia etiology studies in the 1980s [38, 34], in which two bacterial pathogens, streptococcus pneumoniae and haemophilus influenzae, were identified as the primary etiologies of pneumonia mortality. It has been 30 years since those studies conducted, and by 2015, three major changes will have taken place [27]: the wide use of pneumococcoal and

haemophilus influenzae-B conjugate vaccines; the wide spread of HIV infection [9]; the substantial improvements/changes in living conditions, nutrition, and access to health care. These changes will certainly modify the distribution of pathogens, the transmission, and the natural history of infection, which will make the understandings of pneumonia etiology based on the early studies invalid. Hence the effectiveness of the current prevention and treatment could be greatly diminished.

As a result, new information of the current etiology of severe/very severe pneumonia for children under 5 is required to ensure its prevention and treatment strategies are appropriate and effective for the epidemiologic setting of the future. In the context of such a strong need, the Pneumonia Etiology Rearch for Child Health (PERCH) project, the largest of its kind in over 20 years, was launched in 2011 and finished data collection recently.

1.1.3 Pneumonia Etiology Rearch for Child Health

The PERCH project is a case-control study that enrolled around 9500 children from 7 sites across the globe with the primary goals [27] to:

- Estimate the association between severe/very severe pneumonia and infection with confirmed and putative viral, bacterial, mycobacterial, and fungal pathogens.
- Learn the probability of severe/very severe pneumonia attributable to each of the candidate pathogens.
- Evaluate potential risk factors for infection and/or severe/very severe pneumonia due to novel or under-recognized etiologic pathogens.

A case-control design was chosen because it is more efficient than cohort studies and probe studies in terms of identifying the etiology among many different, putative etiologic pathogens. The 7 study sites are in Bangladesh, Gambia, Kenya, Mali, South Africa, Thailand and Zambia. These sites were chosen to represent the developing countries with major childhood pneumonia burdens and a range of diverse epidemiologic settings. The study enrolled about 4200 children hospitalized for severe/very severe pneumonia and approximately 5300 controls randomly selected from the corresponding communities. The inclusion-exclusion criterion are discussed in detail by Deloria-Knoll et al.[12]. For each enrolled subject, data on demographics, known and putative risk factors, and pathogen infection were collected.

More explanation on the rationale of the study can be found in the review by Adegbola, RA and Levine, OS [2].

Specimen Measurements and Data Description

In order to maximize the detection power and accuracy of pathogen infection, the PERCH investors used multiple specimen types [23] including acute blood (for cases only), nasopharyngeal(NP) swab (for both cases and controls), and lung aspirates (for only very few cases). These samples were collected and tested by a variety of conventional and novel detection techniques such as microscopy, culture, serology, antigen testing, and polymerase chain reaction (PCR) [33], targeting on more than 30 candidate pathogens. An example of a single test record is shown in table 1.

Table 1: Test record indicates that Haemophilus influenzae is detected by PCR in the lung aspirate specimen of subject 1.

subject ID	Group	Specimen Type	Detection Technique	Pathogen Name	Test Result
1	Case	Lung aspirate	PCR	Haemophilus influenzae	positive

Tests based on lung aspirates samples are considered to provide the direct observation of the lung and are assumed to have perfect sensitivity and specificity, thus they are called Gold Standard (GS) measurements. Among all peripheral measurements, we assume blood samples provide measurements with perfect specificity, but imperfect sensitivity, and NP samples provide both imperfect sensitivity and specificity, thus we call measurements from blood samples Silver Standard (SS) measurements, and those from NP samples Bronze Standard (BS) measurements.

For each child (patient) i, let Y_i indicate whether this child is a case ($Y_i = 1$) or a control ($Y_i = 0$). Suppose there are K pre-specified pneumonia causing candidates, the list of measurements can be described by three K-dimensional binary vector: M_i^{GS} (if available), M_i^{SS}, and M_i^{BS}, where $M_{ik}^{Src} = 1$ indicates that the kth pathogen is detected using the $Src \in \{GS, SS, BS\}$ measurements in subject i. The data availability and the format of measurement vector are summarized in Figure 1.

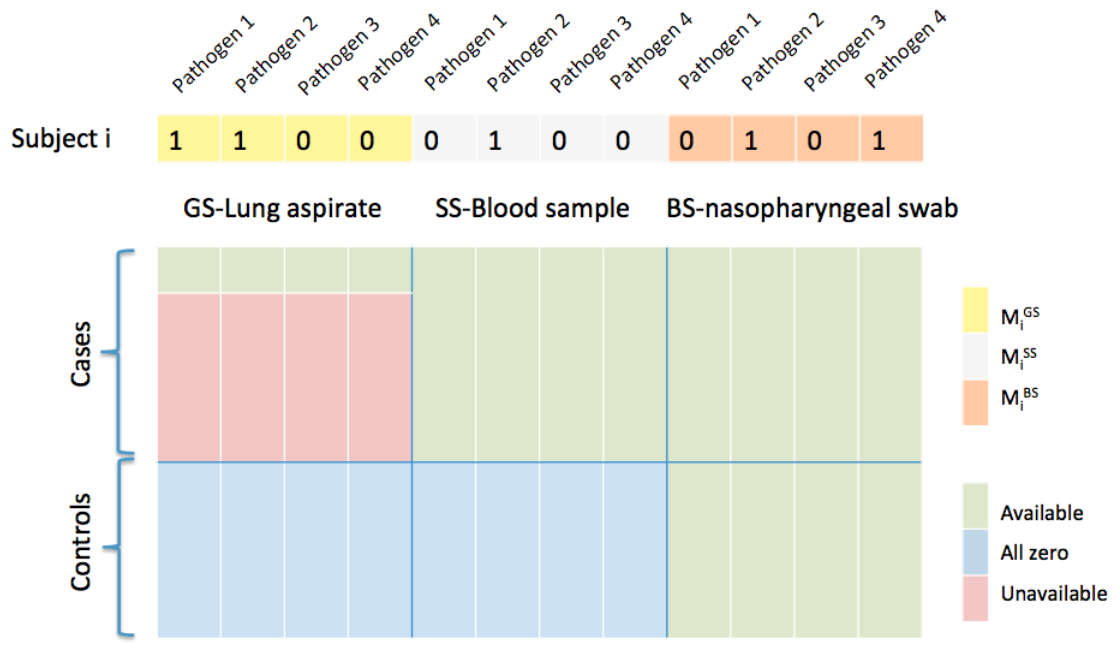

Figure 1: **Data Description**: Suppose we have 4 candidate pathogens in this demonstrative example. In the upper part of this figure, the three 4 dimensional vectors M_i^{GS}, M_i^{SS}, and M_i^{BS} are concatenated together. The GS measurement is available and it tells us pathogen 1 and 2 infect the lung of subject i. Due to the imperfect sensitivity, SS and BS measurements fail to detect pathogen 1. And because of the imperfect specificity, BS measurement detects a false positive for pathogen 4. In the lower part of this figure, the data availability is represented by different colors. As we can see, only a small fraction of cases have GS measurements.

1.2 Statistical Framework and Challenges

Due to the invasiveness of the lung aspirate procedure, GS measurements were rarely acquired [27, 23]. The actual pathogen(s) that infect the lung therefore can only be inferred from multiple peripheral measurements with imperfect sensitivity and/or specificity, i.e. the actual lung infection is a latent variable for cases. This fact poses significant statistical challenges for estimating the prevalence of the etiologic pathogens in the population of children, especially in the situation where there are multiple pathogens infecting the lung.In machine learning terms, the problem is "unsupervised" absent the GS data [17].

Let L_i be a K-dimensional binary vector describing the latent lung infection status for child i, where $L_{ik} = 1$ indicates the child's lung is infected by the kth pathogen. $L_i = (0, \ldots, 0)^T$ means the child has no infection in his/her lung, which is believed to be the true lung status for each control. We also assume, with small probability π_0, a patient identified as case has no infection in his/her lung: $L_i = (0, \ldots, 0)^T$. Using notations defined so far, our interest in this thesis project can be formulated as to estimate the population mean of true lung status of child pneumonia cases

given peripheral measurements data, that is, $\mu = \mathbb{E}[L|M^{GS}, M^{SS}, M^{BS}, Y = 1]$. We will call this parameter the etiology fraction in the following discussion.

Given that most of the measurements are imperfect in terms of sensitivity and specificity, neglecting or inappropriately adjusting (e.g. guessing the wrong value of sensitivity/specificity) for measurement error can produce significantly biased estimates [21]. Therefore, developing a statistical method for estimating μ that appropriately adjusts the measurement errors and incorporates all available sources of evidence is crucial to achieving the goal of PERCH.

Recently, Wu, et al. (2015) developed the partially-Latent Class Model (pLCM)[44] and the nested-pLCM (npLCM) [45], introduced in the next section, as extensions to the classic Laten Class Model [20], in order to deal with issues mentioned above. In this thesis project, we propose to further extend the pLCM framework in three ways. First, pLCM and npLCM assumes that a only one pathogen can be the true cause of disease for a case subject, but scientists often argue that there is a non-negligible probability that two or more pathogens can jointly cause pneumonia for a child. Our extension will allow for multiple pathogens to constitute the sufficient cause[37]. Second, in pLCM, the pneumonia case definition is assumed to be error free. In our model, we will allow for misclassification of cases and controls. Third, an novel prior specification is introduced to facilitate better incorporation of scientific knowledge. Specifically, we will place informative prior information on both the chance that each pathogen is part of the sufficient cause and on the number of pathogens that constitute the sufficient cause. We will study whether these three changes provide a better approximation to the true disease-causing mechanism and will yield more accurate estimates of the parameter of interest.

In section 2, models for multivariate binary data and the latest developments on partial latent class models are reviewed. In section 3, the redesigned representation of L is described. And in section 4, the full specification of our proposed model is explained. The future work is discussed in section 5. The model fitting algorithm is presented in the appendix.

2 Literature Review on Related Statistical Methodologies

In the previous section, we have identified the challenging statistical problem posed by the PERCH study, essential to the prevention and treatment of childhood pneumonia. In this section, we will review the current statistical methods that are relevant to solving this type of problem, introduce the foundations on which our method is based, and briefly explain what advantages our model should achieve.

2.1 Partially-Latent Class Model and Nested Partially-Latent Class Model

Latent class model (LCM) [20] is a statistical model for identifying unobserved subgroups of the population from multivariate categorical data. The model is parameterized by the prevalence of each latent class and the the conditional probabilities for the observed data given each class membership. Traditionally, given the latent class membership, the observed data are assumed to be conditionally independent. This type of model has a wide range of applications[36, 3], in which we will focus on the use in estimating the etiology fraction given multiple indirect measurements.

In this application, however, the classical LCM has its limitations. First, in the classical LCM, the number of latent classes is determined by comparing the goodness-of-fit of different models, thus the latent classes identified do not always have clear interpretations. Second, the classical LCM cannot use the partly available gold standard data and the control data. Third, it suffers from weak model identifiability [20] when the number of latent categories are large.

Therefore, Zhenke Wu, et al.(2015) developed the Bayesian partially-Latent Class Model (pLCM) [44] to estimate the population etiology distribution using PERCH data. In this model, the latent classes are specifically defined as the true infection of the lung. The conditional distributions of the measurements given the true infection are characterized by sensitivities and specificities. Then the marginal likelihood of the multivariate measurements is modeled as a function of the etiology fraction, sensitivities and specificities. This model also makes use of both control and GS, SS, and BS data. Absent GS data, the model is only partially identifiable [26] as termed in Jones et al.(2010). That is, to estimate the etiology fraction, prior information about the measurement sensitivities is needed. Specificities can be estimated from the control data. Furthermore, Wu, et al. extends pLCM to the nested-pLCM [45] by building sub-classes nested under the latent disease classes to allow dependency among measurements given the latent class.

It is important to note that the pLCM and the npLCM were both built on the assumption that each case has a single cause infecting the lung, where this cause can be a single pathogen from the list of candidates or a pre-fixed combination of candidate pathogens. This single-cause assumption enables the model to use the multinomial stochastic framework, i.e. the latent lung status is

assumed to be a multinomial variable I_i with $\omega_k = P(I_i = k)$ and $\sum_{k=1}^{K} \omega_k = 1$, where ω_k is the fraction of of disease caused by the kth pathogen or pathogen combination. This framework is conceptually straightforward and computationally efficient, because the marginal likelihood has the simple form of $\prod_i \sum_k \omega_k P(M_i|I_i = k, \lambda)$ where λ includes the sensitivities and specificities. By choosing appropriate conjugate priors for $\{\omega_k\}_{k=1}^{K}$ and λ, an efficient Gibbs Sampling algorithm can be derived.

However, the single-cause assumption is often questioned by physicians and scientists, since indirect measurements in previous studies [10, 42] suggested possible bacterial and viral coninfections for pneumonia cases and bacterial coinfections in lung tissue specimens were confirmed in the study of 2009 pandemic influenza A (H1N1) [32]. If pathogen coinfections for pneumonia were frequent, the etiology fraction ω_k would be biased and tend to be smaller since the sum of the true probability of pneumonia attributable to each of the candidate pathogens would be greater than one.

Therefore, we propose to develop a model that allows multiple pathogens to cause the disease without having to specify in advance which subset of combinations is allowed. In order to achieve this goal, we propose to model the latent status of the lung as a multivariate binary vector L_i as defined in section 1. Next, we will review the current methods for modeling multivariate binary data.

2.2 Models for Multivariate Binary Data

Multivariate binary data, or multidimensional contingency table, have been extensively studied since the 1960s. A large collections of statistical models have been proposed, and they can be generally categorized into two classes: the likelihood based approach [22, 16, 15, 47] and the non-likelihood based approach [28, 46, 30]. In this proposal, we focus on the likelihood approach in order to integrate it in the new latent class model.

We will briefly review the main methods for parameterizing the multivariate binary distribution in this section, and for each method the important parsimonious extensions and corresponding regression models will also be discussed.

- **Multinomial distribution**: Consider a multivariate binary vector of length K, denoted by $L = (L_1, \ldots, L_k)$. There are 2^K possible observations for L, termed cells. Let each cell probability be $P(L = l) = p_l$ with $\sum_l p_l = 1$, then L is a multinomial variable with $2^K - 1$ independent parameters. This is the most straightforward and flexible model but has bad scalability since the number of parameters grows exponentially as the dimension

grows. Also, it gives little insight into the structure of the data [11], thus it is hard to find a parsimonious extension of it and few regression models were built upon it.

- **Bahadur representation**: First suggested by Bahadur (1961) [4] and later by Cox (1972) [11], this representation models the joint probability of the multivariate binary data as a functions of the marginal probabilities and the second and higher-order correlation. Let $\theta_j = P(L_j = 1)$ and standardize the data as $U_j = (L_j - \theta_j)/\sqrt{\theta_j(1-\theta_j)}$. Define $\rho_{12\ldots k} = \mathbb{E}(U_1 \ldots U_k)$ as the kth order correlation between $L_1, \ldots, L_k$. Then the joint probability is defined as

$$P(L = l) = \prod_{j=1}^{K} P(L_j = l_j)\Big\{1 + \sum_{i>j} \rho_{ij} u_i u_j + \sum_{i>j>k} \rho_{ijk} u_i u_j u_k + \ldots + \rho_{12\ldots d} u_1 \ldots u_d\Big\}$$

  This representation is also a saturated model with $2^K - 1$ independent parameter. To reduce the number of parameters, one can assume parsimonious models for the correlation structure. For example, one could assume an "exchangeable" correlation structure, in which the kth-order correlations are all the same. Then the parameters would only increase linearly with the dimension. In the extreme case where all correlation parameters are set to zero, this representation becomes an independence model.

  Estimation methods for regression models using Bahadur representation were discussed by Lipsitz, et al. (1995) [29]. Since the maximum likelihood (ML) estimation with a Newton-Ralphston algorithm requires very large sample size compared to the dimension to converge to a unique solution, they proposed the "one-step" ML estimator and proved that it is asymptotically equivalent to the fully iterated ML estimator. An alternative moment-based estimation approach[29] was also developed as an extension to Liang and Zeger's (1986) generalized estimating equations (GEE)[28].

- **Log-linear Models**: The general log-linear model, first described by Cox (1972) [11] and discussed in depth by Haberman (1973) [22], is the most widely used parameterization for multivariate binary data. This representation models the joint probability in the log scale as a linear function of conditional log odds' and conditional log odds ratios. It is a member of the exponential family, thus many useful properties can be directly obtained. The general form of log-linear model can be written as:

$$P(L = l; \Theta) = \exp\Big\{\Theta_1^T l + \Theta_2^T w_2 + \ldots + \Theta_K^T w_K\Big\}/A(\Theta)$$

  where w_k is a $\binom{K}{k} \times 1$ vector of the k-way cross-products of l, $k = 1, \ldots, K$, and $\Theta = (\Theta_1, \ldots, \Theta_K)$ contains the the canonical parameters, which is a $(2^K - 1) \times 1$ vector. Θ_1 contains the k conditional log odds' and the rest contains the conditional log odds

ratios, regarded as the association parameters. Moreover, let $l^* = (l, w_2, \ldots, w_K)^T$, the normalizing term is defined as

$$A(\Theta) = \sum_{l^*: l \in \{0,1\}^K} \exp\{\Theta^T l^*\}$$

Similar to the Bahadur representation, the above model allows for varying degrees of dependence among $\{L_j\}_{j=1}^K$. Independence model is achieved when all of the tow- and higher-way association parameters are set to zero. And the other extreme is to use the full $2^K - 1$ parameters to form a saturated model.

A variety of parsimonious extensions and re-parameterizations have been developed based on the log linear model. An important special case is the "quadratic expoential family" described by Zhao and Prentice (1990) [47], which fixes the three- and higher-way association parameters at zero. In addition, they made a one-to-one transformation from (Θ_1, Θ_2) to the marginal moment parameters (μ, σ), where μ is the vector marginal mean and σ is the vector of pairwise covariances, and they derived the likelihood equation for estimating the coefficients of the regression models for μ. However, the problem of this method is that the consistency of the regressions parameters requires the correct specification of both the means and pairwise correlations.

As a method to circumvent the drawback of the above model, an important re-parameterization of the general log-linear model, the "mixed parameter" model, is proposed by Fitzmaurice and Laird (1993) [16]. Let $\Omega = (\Theta_1, \ldots, \Theta_K)$, the model is parameterized in terms of (μ, Ω), the mixture of marginal mean and conditional log odds ratios, via the one-to-one transformation from (Θ_1, Ω) to (μ, Ω). Although such transformation has no closed form, the problem can be solved using the iterative proportional fitting algorithm (Deming and Stephen, 1940) [13] within each step of the Fisher scoring algorithm. And it is shown that the regression coefficient estimator is consistent if the mean structure is correctly specified even if the correlation structure Ω is not.

- **Dependence Ratio Model**: The dependence ratio model was proposed by Ekholm (1995) [15], which models the association using dependence ratios rather than odds ratios. Let $\eta = (\eta_1, \ldots, \eta_K, \eta_{12}, \ldots, \eta_{1\ldots K}) = \mathbb{E}(l^*)$. The kth-order dependence ratio is defined as the joint success probability of k binary responses divided by the joint success probability assuming independence. For example, the 2nd order dependence ratio between L_1 and L_2 is $\lambda_{12} = \frac{\eta_{12}}{\eta_1 \eta_2}$. Therefore, dependence ratio being one indicates independence. It is shown that the joint probability can be expressed as an affine linear transformation of η

and a marginal regression model is built. Furthermore, Ekholm (2000) [14] suggested five types of parsimonious association models by constraining the structure of η based on this representation.

- **Latent Continuous Distribution**: A multivariate binary distribution can be obtained from a multivariate continuous distribution by thresholding each of the variables. For example, consider a multivariate Gaussian random vector $Z = (Z_1, \ldots, Z_K)$, the corresponding multivariate binary distribution can be constructed by letting $L_j = 1$ if and only if, say, $Z_j > 0$ and letting $L_j = 0$ other wise. This model, considered by Cox (1972), as a "historically important way" and a "useful heuristic device" but "seems unnecessary unless the Z's are of intrinsic interest".

- **Lattice Based Model** The lattice based models are extensively studied and widely used in the field of spatial analysis and statistical mechanics. The early work can date back to the Ising Model (1925) [25] and currently there are two dominant approaches for modeling binary data on a lattice: the spatial generalized linear mixed model which models the dependence by latent Gaussian Markov random field over the lattice [5] and the autologistic model, which models the dependence directly [6] thorough a linear function of the neighboring variable, termed autocovariate. The later approach is of more interest in terms of our likelihood specification, so we will focus on the autologistic model in this section.

  Suppose the multivariate binary data $L \in \{0,1\}^K$ are placed on a lattice. The conditional distribution of L_j is given by:

$$P(L_j|L_{-j}) = \text{logit}^{-1}\Big(\beta_j + \sum_{k \neq j} \alpha_{jk} L_k\Big)$$

  where β_j is the conditional log odds, $\{\alpha_{jk}\}$ are the dependence parameters, and the sum is called the autocovariate, which determines the dependence between L_j and all the other variables on the lattice L_{-j}. Let δ_{jk} be the indicator of whether L_j and L_k are neighbors, let D be a $K \times K$ adjacency matrix where $[D]_{jk} = \delta_{jk}$, and assume $\alpha_{jk} = \alpha\delta_{jk}$. By Brook's Lemma, the joint distribution of L is

$$P(L|\beta, \alpha) = \frac{\exp\left(L^T\beta + \frac{\alpha}{2}L^TDL\right)}{\sum_{Y \in \{0,1\}^K} \exp\left(Y^T\beta + \frac{\alpha}{2}Y^TDY\right)}$$

  Thus this model can be also viewed as a special case of the log-linear model.

3 Representation of the Multivariate Binary Latent Variable for Multiple Cause Etiology

To allow the possibility of multiple pathogens infecting the lung, a multivariate binary vector $L \in \{0,1\}^K$ is used as the latent variable that describes the actual status of the lung. Among all methods reviewed in the previous section, the log-linear model has the following advantages that are crucial to the etiology estimation problem. (1) It has all the properties of an exponential family distribution. (2) Its association parameters are orthogonal to the first order parameters, which makes it more convenient to do stratified estimation and conditional regression modeling on the etiology parameter $\mu = \mathbb{E}(L|M, Y = 1)$. (3) The association parameters have intuitive interpretations that reflect the complex interactions among various pathogens.

Moreover, what the etiology estimation problem also requires but the classic log-linear model does not have, are excellent scalability in terms of the sample size as well as the number of candidate pathogens, and of course good identifiability with low-quality measurement data. Also, in the PERCH study, scientists would like to incorporate their prior knowledge in the model that there exists strong competition among the majority of pathogens when they infect the lung, while other few pathogens might infect independently.

Thus, we propose a parsimonious quadratic exponential representation for the latent vector L, and extend the Bayesian partially-Latent Class Model (pLCM) accordingly, for more accurate etiology estimation. In the rest of this section, we will re-introduce the log-linear model in the context of PERCH study and derive our method from it.

3.1 Log-Linear Model

Recall that L is a K-dimensional binary random variable denoting the true state of the lung. With the same notations used in last section, the general form of the log-linear model is:

$$\begin{aligned} P(L = l; \Theta) =& \exp\{\Theta_1^T l + \Theta_2^T u_2 + \ldots + \Theta_K^T u_K\}/A(\Theta) \\ \text{where } A(\Theta) =& \sum_{l^* \in \{0,1\}^K} \exp\{\Theta^T l^*\} \end{aligned}$$

Furthermore, let $S_i \in \{0, 1, \ldots, K\}$ be the total number of pathogens infecting the lung of the ith patient, i.e. $S_k = \sum_{k=0}^{K} L_{ik}$, and define $\pi_s = P(S_i = s)$, thus $\pi = (\pi_0, \ldots, \pi_K)$ are the parameters that reflects the knowledge on how many pathogens there can be in the lung. Let $\mu = (\mu_1, \ldots, \mu_K) = \mathbb{E}(L|M, Y = 1)$ denote the parameter of our primary interests, where μ_k is

the fraction of disease potentially caused by the kth pathogen[1]. Since we allow multiple pathogen infection, the sum of all these fractions may be greater than 1. By plugging in the above joint probability function, we have

$$\begin{aligned}\pi_s :=& P(S=s)\\ =& \frac{1}{A(\Theta)} \sum_{l^*:S=s} \exp\{\Theta^T l^*\}\ , s=1,\ldots,K \qquad (1)\\ \pi_0 =& \frac{1}{A(\Theta)}\\ \mu_k =& \frac{1}{A(\Theta)} \sum_{\tilde{l}:l_k=1} \exp\{\Theta^T l^*\}\ , k=1,\ldots,K \qquad (2)\end{aligned}$$

Note that l^* has length 2^K-1, and by stacking all possible values of l^* except for the zero vector together, we get a square matrix L^* with dimension $J=2^K-1$. As we can see, it is hard to re-parameterize the log-linear model directly to a representation with parameter (μ,π), thus we bring in the un-normalized cell probabilities as the intermediate parameters, where un-normalized cell probability means the joint probability times the normalizing constant $A(\Theta)$.

Let $l^{(j)}$ be the jth possible value of L, then define the un-normalized cell probability as $\phi_j = P(L=l^{(j)})A(\Theta) = \exp(\Theta^T l_j^*), j=1,\ldots,J$. By equations (1) and (2), we know that $(\mu,\pi)^T$ is a linear combination of ϕ's, so we define the following two $K\times J$ matrices B and C to simplify the notation:

$$\begin{aligned}B[k,j] =& 1(\sum_{s=1}^{K} L^*[j,s]=k),\ k=1,\ldots,K\\ C[k,j] =& L^*[j,k]\end{aligned}$$

Recall that L^* is the matrix constructed by stacking all possible values of L together except $\{0\}^K$. $B[k,j]$ is the indicator of whether the jth possible value has k pathogens infecting the lung, and $C[k,j]$ is the indicator of whether the jth possible value has the k pathogen infecting the lung. Thus the relation defined by (1) and (2) becomes

$$\phi > 0 \qquad (3)$$

$$B\phi = \pi/\pi_0 \qquad (4)$$

$$C\phi = \mu/\pi_0 \qquad (5)$$

[1] It is not possible from observational data to determine whether an infection in the lung is part of the sufficient cause or not. That can only be determined by experimentation. Therefore, we are estimating the pathogens infecting the lung whether they constitute the cause or not.

where B and C are not independent constraints and should be compatible so that $\binom{B}{C}$ has rank $2K-1$. Explicitly, μ and π must satisfy

$$\sum_{k=1}^{K} \mu_k = \sum_{k=1}^{K} k\pi_k \tag{6}$$

Note that for any ϕ in the feasible region defined by the above linear constraints, there is a one-to-one mapping between such ϕ and Θ. In fact, Θ is the solutions to the following linear system (J equations with J unknowns):

$$L^*\Theta = \log \phi$$

3.2 Quadratic Exponential Model with Sparse Correlation

An important special case of the log-linear model is the "quadratic exponential family" (QE) described by Zhao and Prentice (1990) [47], where the three- and higher-way association parameters were fixed at zero, which shrinks the model complexity from O(2^K) to O(K^2). This parameterization is widely used in multivariate binary data analysis and performs well with fully observed GS data. However, in the case where GS data are absent, the parameter of interests are hardly identifiable with such limited information from data. Therefore, a more parsimonious and also flexible parameterization is needed.

We propose to use an exchangeable association structure with shrinkage estimation to incorporate prior knowledge as well as to improve the model scalability and identifiability.

$$\begin{aligned} P(L=l;\Theta) =& \exp\{\Theta_1^T l + \Theta_2^T u_2\}/A(\Theta) \\ \text{with } A(\Theta) =& \sum_{l^* \in \{0,1\}^K} \exp\{\Theta^T l^*\} \\ \Theta_2 =& \theta_2 \cdot (I_1, \ldots, I_{\binom{K}{2}}) \end{aligned} \tag{7}$$

where we use a single parameter θ_2 to represent the extent of negative association between pathogens and $I_{k'}$, $k' = 1, 2, \ldots, \binom{K}{2}$, is an indicator for whether a pair of pathogens can independently infect the lung. In the Bayesian framework, it is straightforward to use the stochastic search variable selection [18, 19] method to sample the indicators as latent variables and estimate the posterior distribution.

The parameter of interests μ can be recovered from the posterior samples of Θ using equation (2). When covariates are available, stratified or individual estimation for μ can be achieved by parameterizing the first-order canonical parameters for subject i and pathogen k, $\theta_{ik}^{(1)}$, as $X_i^T \beta_k$, where X_i is the vector of covariates.

4 Full Model Specification

By using a Metropolis-Hastings (within Gibbs) algorithm, we can sample from the posterior distribution as long as we have a well-defined joint distribution of the data and parameters. Throughout this section, let $M_i^{GS} \in \{0,1\}^K$ be the observed GS measurement, $M_i^{SS} \in \{0,1\}^K$ be the observed SS measurement, $M_i^{BS} \in \{0,1\}^K$ be the observed BS measurement and $L_i \in \{0,1\}^K$ be the latent status for subject i. Let $\gamma \in [0,1]^K$ and $\delta \in [0,1]^K$ represent the True Positive Rate (TPR) and False Positive Rate (FPR) for BS measurements respectively, and let $\eta \in [0,1]^K$ be the TPR for SS measurements. Also, let $\mathbb{L}$ be the set of all allowed values of L, such that $|\mathbb{L}| = J^*$ and l_j be the jth element in $\mathbb{L}$.

4.1 The Likelihood for Cases

For cases without GS measurements, and under the conditional independence assumption for measurement given the latent variables, the likelihood function is

$$\begin{aligned} P(M_i^{SS}, M_i^{BS}|\Theta,\eta,\gamma,\delta) &= \sum_{j=1}^{J^*} P(M_i^{SS}, M_i^{BS}, l_j|\mu,\pi,\eta,\gamma,\delta) \\ &= \sum_{j=1}^{J^*} \left[P(M_i^{SS}|l_j,\eta)P(M_i^{BS}|l_j,\gamma,\delta)P(l_j|\Theta)\right] \end{aligned}$$

where $P(l_j|\Theta)$ is the QE likelihood in (7) and

$$\begin{aligned} P(M_i^{SS}|l_j,\eta) &= \prod_{k=1}^{K} P(M_{ik}|l_{jk},\eta_k) \\ &= \prod_{k=1}^{K} (\eta_k^{l_{jk}} l_{jk})^{M_{ik}} (1-\eta_k)^{l_{jk}(1-M_{ik})} \qquad (8) \\ P(M_i^{BS}|l_j,\gamma,\delta) &= \prod_{k=1}^{K} P(M_{ik}|l_{jk},\gamma_k,\delta_k) \\ &= \prod_{k=1}^{K} (\gamma_k^{l_{jk}} \delta_k^{1-l_{jk}})^{M_{ik}} [(1-\gamma_k)^{l_{jk}} (1-\delta_k)^{1-l_{jk}}]^{1-M_{ik}} \qquad (9) \end{aligned}$$

For cases with GS measurements, we have $L_i = M_i^{GS}$, then the likelihood is

$$\begin{aligned} P(M_i^{GS}, M_i^{SS}, M_i^{BS}|\Theta,\eta,\gamma,\delta) =& P(M_i^{SS}, M_i^{BS}|M_i^{GS}\eta,\gamma,\delta)P(M_i^{GS}|\Theta) \\ =& P(M_i^{SS}|M_i^{GS},\eta)P(M_i^{BS}|M_i^{GS},\gamma,\delta)P(M_i^{GS}|\Theta) \end{aligned}$$

where $P(M_i^{SS}|M_i^{GS},\eta)$ is defined using (8), $P(M_i^{BS}|M_i^{GS},\gamma,\delta)$ is defined using (9), and $P(M_i^{GS}|\Theta)$ is defined using $P(l_j|\Theta)$.

4.2 The Likelihood for Controls

For controls, we only have BS measurements and we know that their lungs were not infected. Since Θ are defined for case only, they are not involved in the likelihood for controls, thus the likelihood function is:

$$P(M_i^{BS}|\gamma,\delta) = \prod_{k=1}^{K} \delta_k^{M_{ik}}(1-\delta_k)^{(1-M_{ik})} \tag{10}$$

4.3 The Hierarchical Prior Distribution

For the case likelihood, the parameters are $(\Theta, \eta, \gamma, \delta)$, and for control likelihood, the parameters are (γ, δ). Our goal is to put informative prior on $\Theta_2, \eta, \gamma, \delta$ in order to facilitate the estimation of μ. In our study, the TPR's and FPR's of the model are assumed to be mutually independent and independent from Θ; therefore we can put independent Beta priors on them.

Let (a_k, b_k) be the hyper-parameter that defines the prior of η_k, let (c_k, d_k) be the hyper-parameter for γ_k and let (e_k, f_k) be the hyper-parameter for δ_k, and $a_k, b_k, c_k, d_k, e_k, f_k$ are tuned so that the priors reflect the best knowledge of scientists on the sensitivities and specificities of each test/specimen combination. For example, scientists may believe there is 95 $\%$ of chance that the TPR of a SS measurement is between 0.01 and 0.2, then the value of (a_k, b_k) is determined by setting the 2.5th and 97.5th percentile of Beta(a_k, b_k) to 0.01 and 0.2 respectively and solving the equation.

For Θ_1, or $\boldsymbol{\beta}$ if covariates are used, independent Gaussian priors with shared hyper-parameters $N(0, \sigma_1^2)$, are used. For Θ_2, a hierarchical prior is used as suggested by the stochastic search variable selection algorithm, that is, for $\Theta_2 = \theta_2 \cdot (I_1, \ldots, I_{\binom{K}{2}})$,

$$\begin{aligned}
\theta_2 &\sim \text{N}(\alpha_2, \sigma_2^2) \\
I_{k'} &\sim \text{Bernoulli}(p_{k'}),\ k' = 1, 2, \ldots, \binom{K}{2} \\
p_{k'} &\sim \text{Beta}(\lambda_a, \lambda_b)
\end{aligned}$$

Denote the independent Gaussian prior for Θ_1 as $P(\Theta_1; \sigma_1^2)$ and let $P(\Theta_2; \alpha_2, \sigma_2^2, \lambda_a, \lambda_b)$ represent the hierarchical prior for Θ_2, then the joint prior distribution for the parameters in the likelihood of case data is:

$$P(\Theta, \eta, \gamma, \delta) = P(\Theta_1; \sigma_1^2)P(\Theta_2; \alpha_2, \sigma_2^2, \lambda_a, \lambda_b) \prod_{k=1}^{K} \Big[Beta(\eta_k; a_k, b_k) Beta(\gamma_k; c_k, d_k) Beta(\delta_k; e_k, f_k) \Big]$$

And the prior distribution for the parameters in the likelihood of control data is:

$$P(\gamma, \delta) = \prod_{k=1}^{K} \Big[Beta(\gamma_k; c_k, d_k) Beta(\delta_k; e_k, f_k) \Big]$$

4.4 The Joint Density

With the specification of likelihood and prior, we can construct the joint density needed for building the MCMC algorithm. Let G_i be an indicator of whether subject i has GS measurements, and define the following three index sets.

$$\begin{aligned} H_1 =& \Big\{ i \in \{1, 2, \ldots, n\} : Y_i = 1 \text{ and } G_i = 1 \Big\} \\ H_2 =& \Big\{ i \in \{1, 2, \ldots, n\} : Y_i = 1 \text{ and } G_i = 0 \Big\} \\ H_3 =& \Big\{ i \in \{1, 2, \ldots, n\} : Y_i = 0 \Big\} \end{aligned}$$

then we can define the joint density of data and parameters as follow by combining all building blocks together:

$$\begin{aligned} & P(M, \Theta, \eta, \gamma, \delta) \\ =& \prod_{i \in H_1} P(M_i^{GS}, M_i^{SS}, M_i^{BS} | \Theta, \eta, \gamma, \delta) \prod_{i \in H_2} P(M_i^{SS}, M_i^{BS} | \Theta, \eta, \gamma, \delta) \prod_{i \in H_3} P(M_i^{BS} | \gamma, \delta) P(\gamma, \delta) \end{aligned}$$

From the form of this model, we can see that the model we proposed incorporates data from all the sources with measurements of varying quality, and its likelihood for the latent variable adopts a quadratic exponential model with sparse correlation structure, so we will refer to this model as the Latent Sparse Correlation (LSC) model in the rest of this paper. With the joint density fully specified, the posterior distribution of our LSC model can be sampled by a Metropolis-Hastings (within Gibbs) algorithm, of which the details are described in the appendix.

5 Simulation Studies

5.1 Data Simulation

Three sets of simulation studies are carried out to empirically evaluate the effectiveness of the LSC model under different situations. For all three sets of studies, we assume there are five candidate pathogens ($A, B, \ldots, E$) and two relevant binary covariates (X_1 and X_2). In each study, 200 independent data sets are simulated, and in each simulated data set, there are 500 case subjects and 1000 control subjects.

At the data simulation stage, we first simulate the true lung infection status, then generate the BS and SS measurements. In Study I, multiple pathogens are allowed to infect the lung at the same time, and the measurements are of relatively low quality, that is, lower true positive rates and higher false positive rates. In Study II, infection is assumed to be caused by a single pathogen, and the measurement quality is the same as in Study I. In Study III, the actual lung status is generated in the same way as in Study I, but the measurements have relatively high quality. Details of the study design are described below.

I In the first set of studies, the true lung infection status L of case patients are generated by a Quadratic Exponential Model, where the first order canonical parameters are dependent on both covariates with an interaction effect, and the second order parameters are independent of the covariates. For case subject i and the kth pathogen,

$$\theta_{ik}^{(1)} = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{k2}x_{i2} + \beta_{k3}x_{i1}x_{i2}$$

Also, with $K = 5$, there are 10 second order parameters. We assume that two of them are zero, which represents that two particular pairs of pathogens ($B : C$ and $B : D$) infect lungs independently from each other. The rest eight association parameters share the same negative value: -1.5, which stands for the pairwise competition among pathogens. Then, the BS and SS measurements for case subjects are simulated based on formula (8) and (9) respectively, and the SS measurements for control subjects are simulated based on formula (10) assuming that there is no infection at all in control patients' lungs, where $\text{TPR}^{(SS)} \approx 0.1, \text{TPR}^{(BS)} \approx 0.7, \text{FPR}^{(BS)} \approx 0.45$. The actual parameter values used in the simulation process are summarized in table 2.

II In the second set of studies, the true lung infection status L of case patients are generated by a Multinomial Model, which is equivalent to the above QE model but with all the second order parameters set to negative infinity. The multinomial etiology probabilities are listed in table 3. Also, the BS and SS measurements are generated in the same way as they are in Study I with the same TPRs and FPRs.

Table 2: The model parameters used for data simulation in study I

Pathogen	A	B	C	D	E
β_0	0.21	-0.28	-0.84	-0.21	1.07
β_1	-0.1	-0.5	0.5	0.2	0.1
β_2	-0.3	0.2	-0.2	-0.1	0.3
β_3	0.4	0.3	-0.4	0.2	-0.2
TPR$^{(SS)}$	0.11	0.12	0.08	0.15	0.10
TPR$^{(BS)}$	0.80	0.60	0.70	0.70	0.65
FPR$^{(BS)}$	0.50	0.55	0.40	0.35	0.45

Table 3: The etiology probabilities used for data simulation in study II

	Other	A	B	C	D	E
strata 1	0.200	0.241	0.079	0.088	0.125	0.267
strata 2	0.171	0.224	0.113	0.106	0.191	0.196
strata 3	0.232	0.191	0.049	0.115	0.105	0.308
strata 4	0.224	0.176	0.072	0.137	0.162	0.230

III In the third set of studies, the true lung infection status of case patients are simulated from the same model as in study I, but the parameters that control the measurement quality are set differently, that is TPR$^{(SS)} \approx 0.8$, TPR$^{(BS)} \approx 0.9$, FPR$^{(BS)} \approx 0.05$. Their values are listed in table 5.

Table 4: The model parameters used for data simulation in study III that are different from study I

Pathogen	A	B	C	D	E
TPR$^{(SS)}$	0.81	0.82	0.88	0.85	0.80
TPR$^{(BS)}$	0.98	0.96	0.97	0.97	0.95
FPR$^{(BS)}$	0.050	0.055	0.040	0.035	0.045

5.2 Model Specifications

In each of the above situation, 200 independent data sets are generated. The LSC model is applied to each data set (without using GS measurements) with a series of different prior specifications on α_2, λ_a and λ_b, which represent the experts' prior knowledge on the magnitude of the competitions between pathogens. Within each study, the values for hyper-parameters a_k, b_k, c_k, d_k, e_k and f_k, $k = 1, 2, \ldots, 5$, which control the prior input on measurement quality, do not vary. Wu et

al.(2015) [44] had discussed the model sensitivity to these hyper-parameters and the partial identifiability issue, which also applies to our method. Thus we do not further study the sensitivity issue on these hyper-parameters. Their values are selected according to experts' knowledge on the quality of BS and SS measurements, and the true TPR and FPR values are set to be covered by the prior 95% credible interval. For all three studies, same values are used for σ_1, and σ_2, which are set large enough to represent non-informativeness. These hyper-parameter values are listed in table **??**.

Table 5: The common hyper-parameters used for model fitting in simulation studies

	a_k	b_k	c_k	d_k	e_k	f_k	σ_1	σ_2
Study I	7.6	59	12.7	4.8	1	1	2.2	2.2
Study II	50	10	12	1	1	1	2.2	2.2
Study III	7.6	59	12.7	4.8	1	1	2.2	2.2

For each different prior specification in each study, we have 200 sets of posterior samples produced by the LSC model. Their posterior means are collected to construct an approximate sampling distribution of the estimator. The average of these approximate sampling distributions implies empirically the values to which our parameter estimates converge. Thus the overall accuracy of the LSC model is evaluated based on these sampling distributions means. Note that a five-dimensional multivariate binary distribution can be represented by a multinomial distribution with 32 cells. Let $q_j, j = 1, \ldots, 32$ be the true multinomial cell probabilities, and let $\hat{q}_j$ be the cell probability estimations based on the sampling distribution means, then the Bhattachayya coefficient [7], $\sum_{j=1}^{32} \sqrt{q_j \hat{q}_j} \in [0, 1]$, which measures the similarity between two discrete distributions, is a good metric of the general accuracy of the LSC model.

In Study I and II, the LSC model is compared against the Bayesian partially-Latent Class Model (pLCM) [44]. The pLCM, which originally only considers single-pathogen infection, now allows multi-pathogen infection in its latest release by making it possible to manually specify candidate pathogen combinations and a Dirichlet prior with equal weights. However, the saturated model, in which all possible combinations are included, is unstable and lack of identifiability with only a few hundred case subjects. Thus, rather than the saturated model, two most commonly used pLCM specifications are applied: 1) the classic pLCM (pLCM-1) where only single pathogen infections are allowed and 2) the new pLCM (pLCM-2) where not only single pathogen but also all pairs of pathogen infections are allowed.

5.3 Results

Table 6 lists the Bhattachayya coefficients of the LSC model under different prior specifications, and of the two pLCMs in Study I. The same information is visualized in figure 2. As we can see, when the true data generating mechanism allows multi-pathogen infection, the classic pLCM (pLCM-1) performs the worst, the pLCM-2 is the second worst, and the LSC model, across all prior specifications, shows a significant amount of improvement over the two pLCM models. The variations caused by different prior specifications is relatively small comparing to the improvement over pLCM models, especially the single-pathogen model.

Table 6: Summary of the overall parameter estimation accuracy of each model fitted in study I

α_2	λ_a	λ_b	Bhattacharyya
-5	6	2	0.9838
-3.5	6	2	0.9828
-7.5	4	4	0.9828
-5	4	4	0.9815
-2	6	2	0.9790
-2	15	5	0.9790
-1.5	6	2	0.9770
-3.5	4	4	0.9765
-1	6	2	0.9748
-7.5	15	5	0.9747
-10	4	4	0.9742
-3.5	15	5	0.9728
-2	2	2	0.9712
-7.5	6	2	0.9711
-1	15	5	0.9677
-1	2	2	0.9673
-10	6	2	0.9649
-10	15	5	0.9604
pLCM-2			0.9494
pLCM-1			0.7681

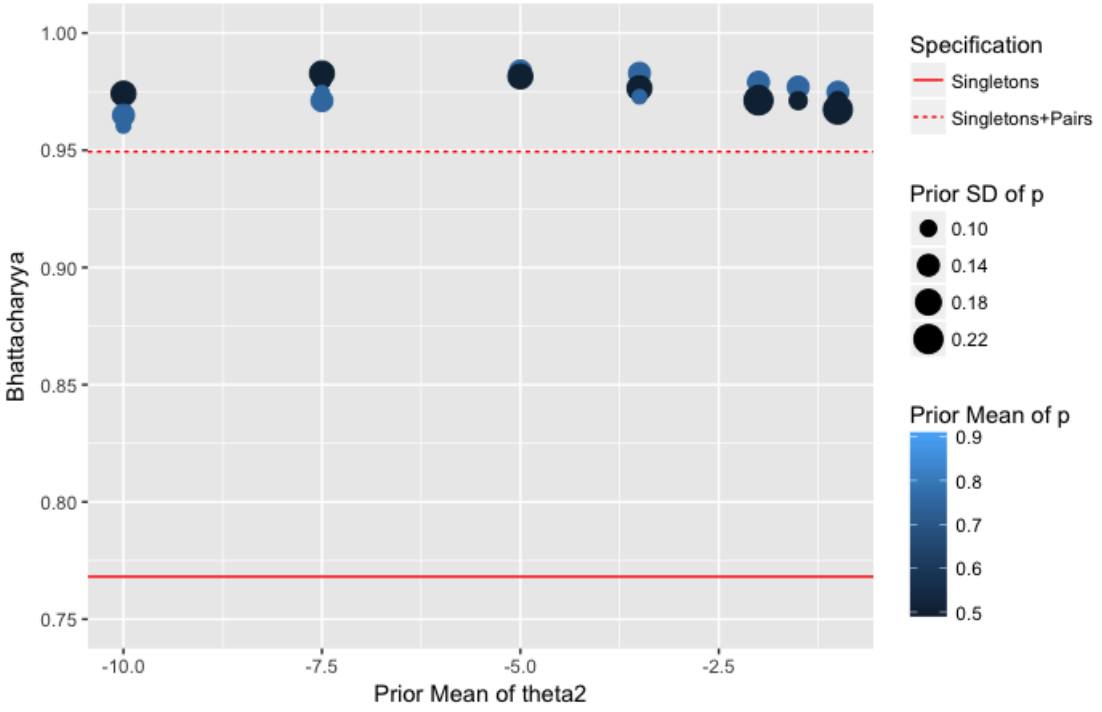

Figure 2: **Overall estimation accuracy of models in study I**: The x-axis stands for α_2, the prior mean of θ_2, and the y-axis represents the Bhattacharyya coefficient (BC). The round dots are values generated by LSC models with different priors. The color of dots indicates the mean of the prior distribution of p and the size of dots indicate the standard deviation of the prior distribution of p. These two values are calculated based on λ_a and λ_b. The two horizontal lines are benchmark values generated by pLCM models. The solid line represents the pLCM-1 model which only allows singleton infections, and the dashed line represents the pLCM-2 model which allows singleton and all the pair infections.

Model performances in Study II is summarized in table 7 and plotted in figure 3. The figure shows that when the true data generating mechanism is single pathogen infection only, the pLCM-1 has the best estimation accuracy since that is the true model, and the pLCM-2 performs the worst because it tends to attribute the cause of the disease to two-pathogen infection. The LSC model, on the other hand, shows an increasing trend in estimation accuracy as the prior mean of θ_2 gets more negative because $\theta_2 = -\infty$ makes the model equivalent to the true model. Moreover, in practice, as we can see in figure 3, if the true model is not known, the LSC model can still provide an estimate almost as accurate as the pLCM-1 with a moderately negative (e.g. between -10 and -5) prior mean of θ_2.

Table 7: Summary of the overall parameter estimation accuracy of each model fitted in study II

α_2	λ_a	λ_b	Bhattacharyya
-16	6	2	0.9925
-10	6	2	0.9828
-16	4	4	0.9761
-7.5	15	5	0.9742
-10	15	5	0.974
-7.5	6	2	0.9725
-10	4	4	0.9717
-16	15	5	0.9635
-7.5	4	4	0.9603
-4	6	2	0.9501
-4	15	5	0.9257
-1	15	5	0.9101
-1	6	2	0.9089
-1	4	4	0.9086
-4	4	4	0.9052
pLCM-1			0.9981
pLCM-2			0.8203

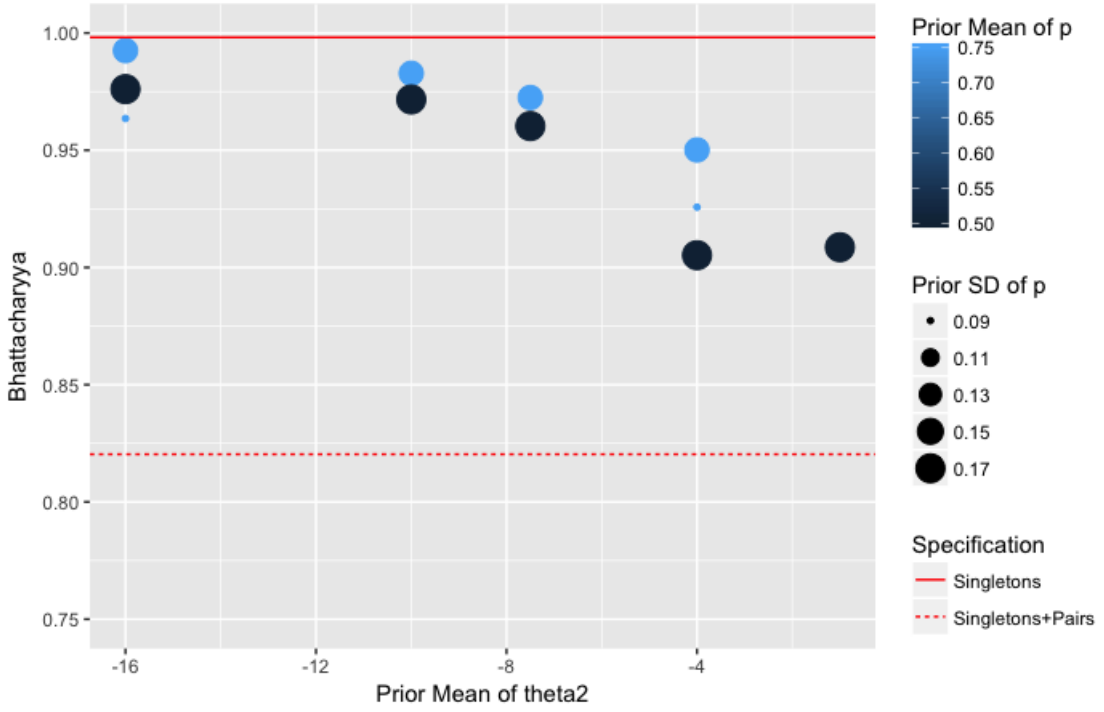

Figure 3: **Overall estimation accuracy of models in study II**: The x-axis stands for α_2, the prior mean of θ_2, and the y-axis represents the Bhattacharyya coefficient (BC). The round dots are values generated by LSC models with different priors. The color of dots indicates the mean of the prior distribution of p and the size of dots indicate the standard deviation of the prior distribution of p. These two values are calculated based on λ_a and λ_b. The two horizontal lines are benchmark values generated by pLCM models. The solid line represents the pLCM-1 model which only allows singleton infections, and the dashed line represents the pLCM-2 model which allows singleton and all the pair infections.

It is very intuitive to think that better measurement quality leads to more accurate etiology estimation. Figure 4, which compares the etiology estimations, $\hat{\mathbb{E}}(L|Y = 1, X)$, from study III and two scenarios of study I, confirms this hypothesis. In this figure, the sampling distributions of the etiology estimations for pathogen A, B, and C in the upper panel have larger standard deviations than those in the middle panel. Also, their shapes are more skewed, and the estimation bias tends to be larger. The distributions for pathogen D and E are almost the same in the upper and the middle panel. Comparing the lower panel to the middle panel, we can see that the sampling distributions estimated from high-quality measurements have much smaller standard deviations and nearly no estimation bias.

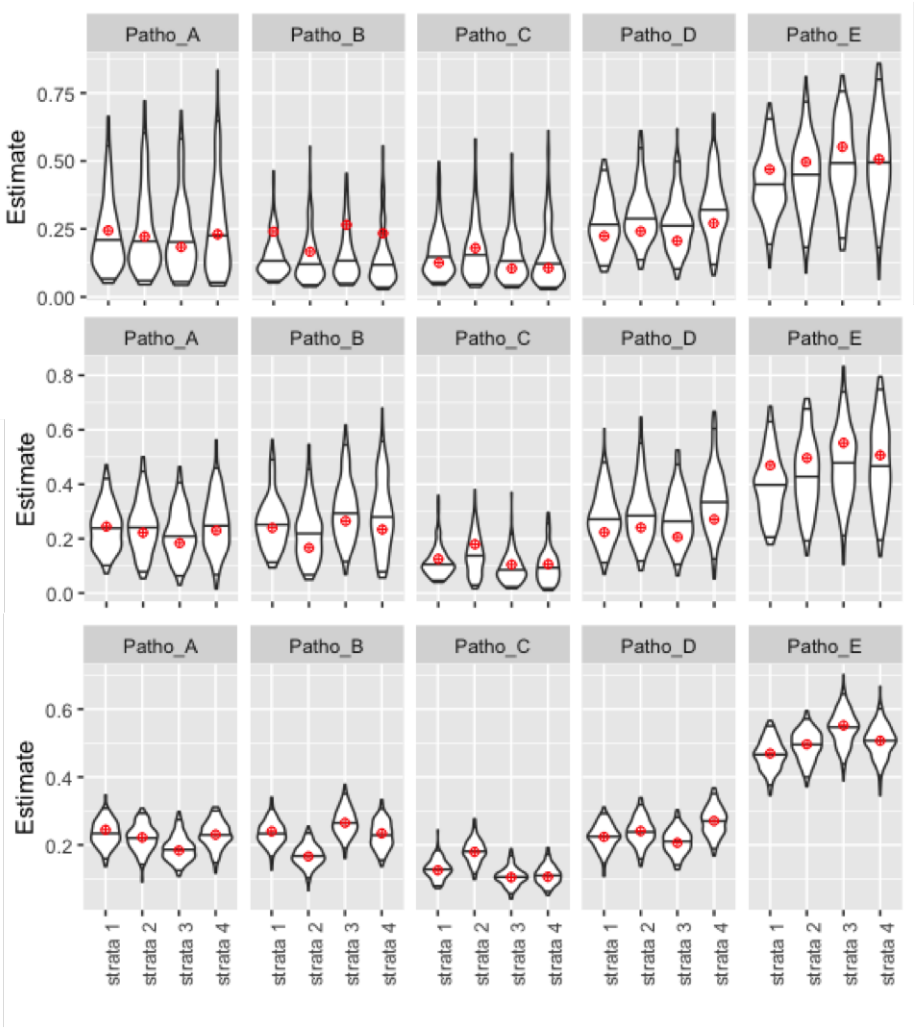

Figure 4: **The impact of data quality on the LSC model**: The results from Study I with SS measurements available for all pathogens are plotted in the middle panel, the results from Study I with SS measurements of pathogen A, B, and C removed are plotted in the upper panel, and the results from study III are plotted in the lower panel. In each plot, there are five facets, with each one corresponds to a pathogen. In each facet, the x-axis represents four strata that are determined by the two binary covariates, and the y-axis stands for the estimated value of etiology probability, $\hat{\mathbb{E}}(L|Y=1,X)$. The violin shape indicates the estimated density function of the sampling distributions of $\hat{\mathbb{E}}(L|Y=1,X)$. The three horizontal lines in each violin shape represent the $2.5th$, $50th$, and $97.5th$ percentiles respectively. The red dots show the true values of the corresponding parameters.

Besides the marginal etiology probability $\mathbb{E}(L|Y=1,X)$, we also compare the estimated probabilities of the most prevalent etiological combinations of pathogens. Figure 5 shows the sampling distributions of the singleton and doubleton etiology probabilities estimated in three scenarios. The first thing we can learn from the figure is that in the ideal circumstance with high-quality measurements, the LSC model can provide accurate etiology probability estimates for

every single combination of pathogens. While in reality, where measurement quality is relatively low and multiple pathogens do not have SS measurements, the LSC model provides more accurate (less bias and smaller variance) estimate for most etiology combinations than the pLCM-2 model. A general observation is that the pLCM-2 estimates tend to over-estimate the doubleton probabilities. Our explanation is that the multinomial likelihood in the pLCM-2 model does not take the interaction structure into account nor does it provide shrinkage on the probability estimates. Thus it attributes some of the singleton or tripleton combinations to doubleton combinations.

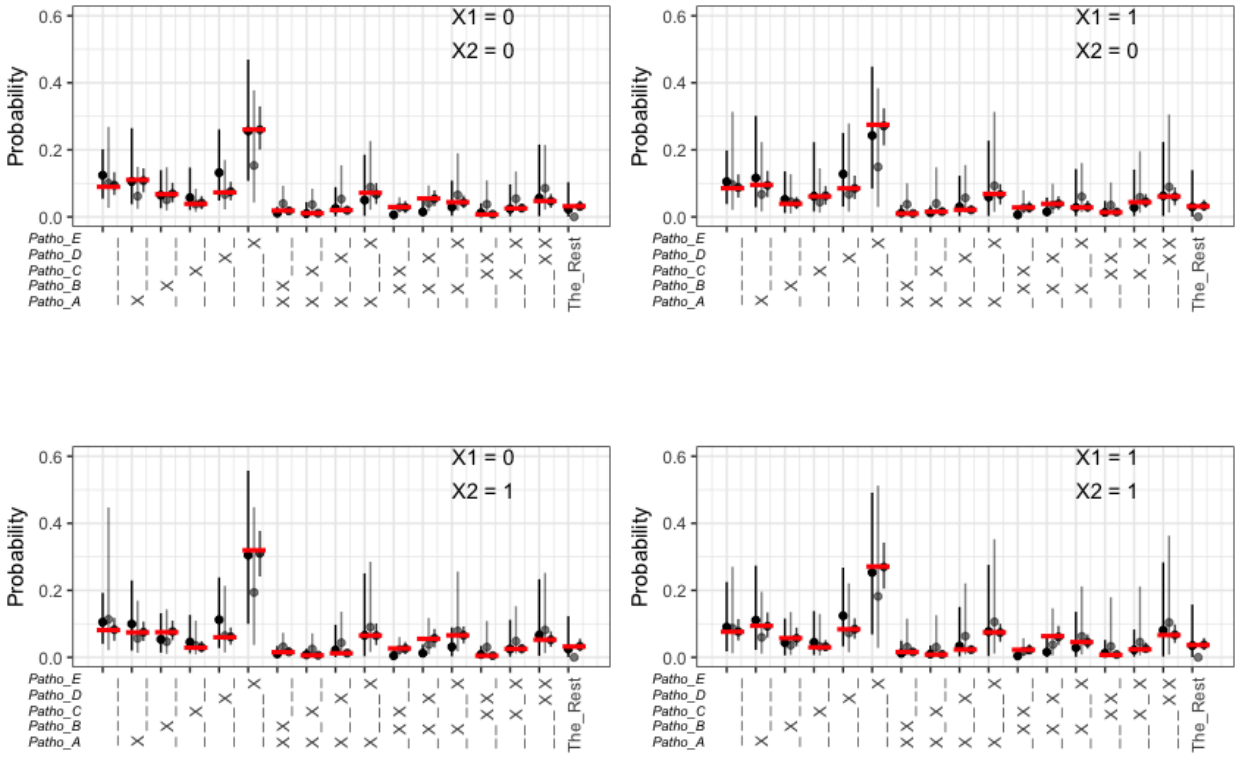

Figure 5: **Singleton and doubleton etiology probability estimation comparison**: Each of the four plots in this figure stands for a specific stratum labeled by the top-right legend. In each plot, the x-axis includes 17 different etiological combinations. Each of the first 16 combinations from the left is denoted by a unique combination of '_' and 'X', where '_' means no infection and 'X' mean infection for the corresponding pathogen listed on the very left. The last combination is labeled by 'The_Rest' indicating the sum of all the rest possible combinations, e.g. tripletons, etc. For each combination, there are three vertical lines, of which the upper and lower bounds represent the 97.5th and 2.5th percentiles of the sampling distribution. The solid dots along these lines indicate the mean of the sampling distribution. The red horizontal lines mark the true values. From left to right, the three lines correspond to three scenarios: (left) LSC estimates in Study I with SS measurements only available for pathogen D and E; (middle) pLCM-2 estimates in Study I with SS measurements only available for pathogen D and E; (right) LSC estimates in Study III.

6 PERCH Data Analysis

As what has been introduced in the first section of this paper, the PERCH study enrolled about 4200 children hospitalized for severe/very severe pneumonia and approximately 5300 controls randomly selected from communities across 7 sites around the world. To demonstrate the application of the LSC model for the analysis of PERCH study data, only the Kenya site data, where there is good availability of both BS and SS measurement data, is used so that the site-specific effect if not a concern. We picked the top 5 pathogens reported in [44] as our candidate pathogens in this analysis. These pathogens are streptococcus pneumoniae (PNEU), haemophilus influenzae (HINF), human metapneumovirus type A or B (HMPV_A_B), rhinovirus (RHINO), and respiratory syncytial virus type A or B (RSV). The BS measurements (nasopharyngeal specimen with PCR detection of pathogens - NPPCR) are available for all 281 cases and 1138 frequency-matched controls on all 5 pathogens. The SS measurements (blood culture results - BCX) are only available for all cases on the two bacteria pathogens: PNEU and HINF.

Prior scientific knowledge [33] suggests that the TPR of BS measurements (NPPCR) is in the range of 50% - 99%, and the TPR of SS measurements (BCX) is in the range of 5% - 20%. Thus we set the hyper-parameters $a_k = 7.6$, $b_k = 59.0$, $c_k = 6.0$, and $d_k = 1.3$ by range matching, and $e_k = f_k = 1$ for non-informativeness. Two categorical (binary) variables: age group and disease severity are taken into account, thus the etiology estimation is made for 4 strata. Regression coefficients including the interaction terms all have prior mean at zero and other hyper-parameter values used for PERCH data analysis are listed in table 8.

Table 8: The hyper-parameters used for the LSC model in PERCH data analysis

	α_2	λ_a	λ_b	σ_1	σ_2
LSC	-5	6	2	2.2	2.2

Figure 6 shows the singleton and doubleton etiology probability estimates for Kenya site given by three models. As we can see, the singleton etiology probability estimates made by the LSC model are quite similar to the corresponding estimates made by pLCM-1. The doubleton estimates and 'The Rest' estimates made by the LSC model are mostly nearly zero, with an exception for the PNEU-HINF pair in the Age = 0 and Severity = 1 stratum. These two pathogens are both commensals of the human nasopharynx and have long been detected together in a multispecies biofilm in infected tissue [41]. A noticeable difference between pLCM-2 results and the other two models is that pLCM-2 attributes the etiology more to RHINO and RHINO-PNEU pair and less to pathogens other than these five. Singleton et al. [40] categorize respiratory viruses into two groups based on their contribution to disease. Group 1 includes viruses with a significantly greater

contribution to respiratory symptoms, including RSV, metapneumovirus, certain para-influenza viruses, and influenza viruses. Group 2 viruses, including human rhinoviruses, adenoviruses, and coronaviruses, are less likely to be the single etiological pathogen of disease in children. Thus, it appears the pLCM-2 model overestimates the contribution of singleton RHINO in the Age = 1 strata, and the final inference should be based on the LSC model. The etiology probabilities estimated by the LSC model are listed in table 9 and 10 where 'None Above' means infection by other pathogens that are not among the listed five candidates, and 'The Rest' means infection by any other possible combinations of the listed five candidates.

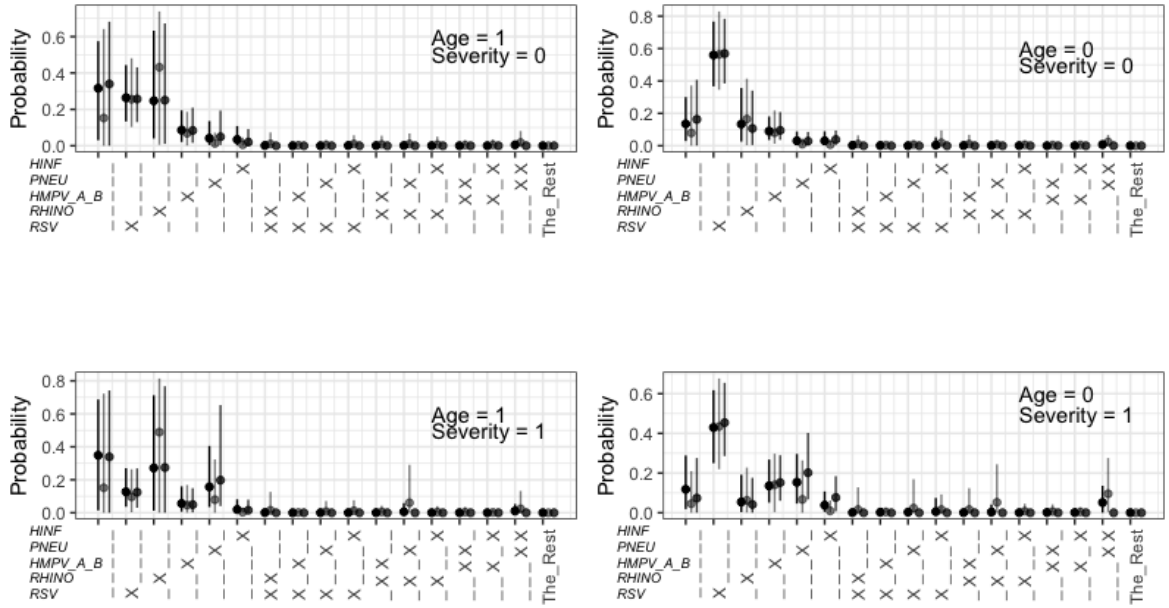

Figure 6: **Singleton and doubleton etiology probability estimation for Kenya site**: The legends and labels in this figure have the same meaning as they are in figure 5, except that the pathogen names listed in the x-axis labels in this figure are the real pathogen abbreviations. The three vertical lines for each etiological combination correspond to the three models applied to the Kenya data set: (left) the LSC model; (middle) the pLCM-2 model; (right) the pLCM-1 model.

7 Conclusion

In this paper, we propose a new method, the Latent Sparse Correlation (LSC) model, for pneumonia etiology estimation using non-Gold standard measurements. This method takes advantages of the parsimonious parameterization of the latent multivariate binary variables and the Bayesian shrinkage estimation procedure. It is shown by simulation studies that this approach can provide estimation for the latent etiology distribution reasonably well while allowing arbitrary combinations of pathogen infection. In the PERCH data analysis, the estimation results of the LSC model are consistent with published etiology research findings. A limitation of this method is its assumption that any pairs of pathogens either compete against each other or act conditionally independent. A possible extension that allows for synergic effects between pathogens is readily achievable.

Table 9: Etiology probability estimates for Kenya site

	Age = 1, Severity = 0			Age = 0, Severity = 0		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
None_Above	0.3162	0.0303	0.5748	0.1351	0.0288	0.3003
RSV	0.264	0.1335	0.4435	0.5601	0.3669	0.7666
RHINO	0.2471	0.0397	0.6324	0.1335	0.0257	0.3566
HMPV_A_B	0.086	0.0212	0.1946	0.0898	0.0355	0.1805
PNEU	0.0411	0.0041	0.136	0.0304	0.0045	0.0881
HINF	0.0341	0.0013	0.1083	0.0304	0.0058	0.0893
RSV-RHINO	0.0015	0	0.0111	0.0025	0	0.0149
RSV-HMPV_A_B	0.0004	0	0.0022	0.0016	0	0.0081
RSV-PNEU	0.0002	0	0.0011	0.0005	0	0.0033
RSV-HINF	0.0014	0	0.0188	0.0041	0	0.0523
RHINO-HMPV_A_B	0.0011	0	0.0066	0.0014	0	0.0034
RHINO-PNEU	0.0013	0	0.0124	0.0016	0	0.0194
RHINO-HINF	0.0007	0	0.0083	0.0009	0	0.0123
HMPV_A_B-PNEU	0.0001	0	0.0005	0.0002	0	0.0006
HMPV_A_B-HINF	0.0002	0	0.0028	0.0005	0	0.0075
PNEU-HINF	0.0049	0	0.0236	0.0073	0	0.0268
The_Rest	0	0	0.0001	0.0001	0	0.0006

Table 10: Etiology probability estimates for Kenya site

	Age = 1, Severity = 1			Age = 0, Severity = 1		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%
None_Above	0.3487	0.0143	0.6874	0.1181	0.0188	0.2886
RSV	0.1269	0.0365	0.2703	0.4295	0.2498	0.6179
RHINO	0.2718	0.0127	0.7124	0.0541	0.0034	0.1918
HMPV_A_B	0.0561	0.0066	0.16	0.1358	0.0512	0.268
PNEU	0.157	0.0346	0.4057	0.1535	0.0461	0.2964
HINF	0.0195	0.0003	0.0815	0.0383	0.0044	0.1073
RSV-RHINO	0.0012	0	0.0098	0.001	0	0.0067
RSV-HMPV_A_B	0.0001	0	0.0008	0.0025	0	0.0143
RSV-PNEU	0.0004	0	0.0028	0.003	0	0.0201
RSV-HINF	0.0005	0	0.0068	0.0054	0	0.0744
RHINO-HMPV_A_B	0.0009	0	0.0074	0.001	0	0.004
RHINO-PNEU	0.0055	0	0.0585	0.0031	0	0.0415
RHINO-HINF	0.0004	0	0.0053	0.0006	0	0.0073
HMPV_A_B-PNEU	0.0003	0	0.0013	0.0014	0	0.0063
HMPV_A_B-HINF	0.0001	0	0.0007	0.001	0	0.0118
PNEU-HINF	0.0105	0	0.053	0.0514	0.0001	0.1356
The_Rest	0	0	0.0002	0.0003	0	0.0026

8 References

[1] Who pneumonia fact sheets. http://www.who.int/mediacentre/factsheets/fs331/en/. Accessed: 2015-09-30.

[2] Richard A Adegbola and Orin S Levine. Rationale and expectations of the pneumonia etiology research for child health (perch) study. *Expert review of respiratory medicine*, 5(6):731, 2011.

[3] Paul S Albert, Lisa M McShane, Joanna H Shih, US National Cancer Institute Bladder Tumor Marker Network, et al. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, pages 610–619, 2001.

[4] Raghu Raj Bahadur. A representation of the joint distribution of responses to n dichotomous items. *Studies in item analysis and prediction*, 6:158–168, 1961.

[5] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014.

[6] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[7] A Bhattachayya. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35(99-109):28, 1943.

[8] Robert E Black, Simon Cousens, Hope L Johnson, Joy E Lawn, Igor Rudan, Diego G Bassani, Prabhat Jha, Harry Campbell, Christa Fischer Walker, Richard Cibulskis, et al. Global, regional, and national causes of child mortality in 2008: a systematic analysis. *The lancet*, 375(9730):1969–1987, 2010.

[9] Daniel Calder and Shamim Qazi. Evidence behind the who guidelines: hospital care for children: what is the aetiology of pneumonia in hiv-infected children in developing countries? *Journal of tropical pediatrics*, 55(4):219–224, 2009.

[10] Gustavo Cilla, Eider Onate, Eduardo G Perez-Yarza, Milagrosa Montes, Diego Vicente, and Emilio Perez-Trallero. Viruses in community-acquired pneumonia in children aged less than 3 years old: High rate of viral coinfection. *Journal of medical virology*, 80(10):1843–1849, 2008.

[11] David R Cox. The analysis of multivariate binary data. *Applied statistics*, pages 113–120, 1972.

[12] Maria Deloria-Knoll, Daniel R Feikin, J Anthony G Scott, Katherine L OBrien, Andrea N DeLuca, Amanda J Driscoll, Orin S Levine, et al. Identification and selection of cases and controls in the pneumonia etiology research for child health project. *Clinical infectious diseases*, 54(suppl 2):S117–S123, 2012.

[13] W Edwards Deming and Frederick F Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[14] Anders Ekholm, John W McDonald, and Peter WF Smith. Association models for a multivariate binary response. *Biometrics*, pages 712–718, 2000.

[15] Anders Ekholm, Peter WF Smith, and John W McDonald. Marginal regression analysis of a multivariate binary response. *Biometrika*, 82(4):847–854, 1995.

[16] Garrett M Fitzmaurice and Nan M Laird. A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):141–151, 1993.

[17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[18] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[19] Edward I George and Robert E McCulloch. Stochastic search variable selection. In *Markov chain Monte Carlo in practice*, pages 203–214. Springer, 1996.

[20] Leo A Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

[21] Paul Gustafson, Nhu D Le, and Marc Vallée. A bayesian approach to case–control studies with errors in covariables. *Biostatistics*, 3(2):229–243, 2002.

[22] Shelby J Haberman. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics*, pages 617–632, 1973.

[23] Laura L Hammitt, David R Murdoch, J Anthony G Scott, Amanda Driscoll, Ruth A Karron, Orin S Levine, Katherine L OBrien, et al. Specimen collection for the diagnosis of pediatric pneumonia. *Clinical infectious diseases*, 54(suppl 2):S132–S139, 2012.

[24] Takashi Hirama, Takefumi Yamaguchi, Hitoshi Miyazawa, Tomoaki Tanaka, Giichi Hashikita, Etsuko Kishi, Yoshimi Tachi, Shun Takahashi, Keiji Kodama, Hiroshi Egashira, et al. Prediction of the pathogens that are the cause of pneumonia by the battlefield hypothesis. *PloS one*, 6(9):e24474, 2011.

[25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[26] Geoffrey Jones, Wesley O Johnson, Timothy E Hanson, and Ronald Christensen. Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, 66(3):855–863, 2010.

[27] Orin S Levine, Katherine L OBrien, Maria Deloria-Knoll, David R Murdoch, Daniel R Feikin, Andrea N DeLuca, Amanda J Driscoll, Henry C Baggett, W Abdullah Brooks, Stephen RC Howie, et al. The pneumonia etiology research for child health project: a 21st century childhood pneumonia etiology study. *Clinical infectious diseases*, 54(suppl 2):S93–S101, 2012.

[28] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, pages 13–22, 1986.

[29] Stuart R Lipsitz, Garret M Fitzmaurice, Lynn Sleeper, and LP Zhao. Estimation methods for the joint distribution of repeated binary observations. *Biometrics*, pages 562–570, 1995.

[30] Stuart R Lipsitz, Nan M Laird, and David P Harrington. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, 78(1):153–160, 1991.

[31] Li Liu, Hope L Johnson, Simon Cousens, Jamie Perin, Susana Scott, Joy E Lawn, Igor Rudan, Harry Campbell, Richard Cibulskis, Mengying Li, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The Lancet*, 379(9832):2151–2161, 2012.

[32] J Louie, C Jean, TH Chen, S Park, R Ueki, T Harper, E Chmara, J Myers, R Stoppacher, C Catanese, et al. Bacterial coinfections in lung tissue specimens from fatal cases of 2009 pandemic influenza a (h1n1)-united states, may-august 2009. *Morbidity and Mortality Weekly Report*, 58(38):1071–1074, 2009.

[33] David R Murdoch, Katherine L OBrien, Amanda J Driscoll, Ruth A Karron, Niranjan Bhat, et al. Laboratory methods for determining pneumonia etiology in children. *Clinical infectious diseases*, 54(suppl 2):S146–S152, 2012.

[34] World Health Organization et al. Programme for the control of acute respiratory infections. *Acute respiratory infections in children: case management in small hospitals in developing countries. A manual for doctors and other senior health workers. Geneva, Switzerland: World Health Organization*, 1990.

[35] World Health Organization, UNICEF, et al. Global action plan for prevention and control of pneumonia (gapp). 2009.

[36] David Rindskopf. The use of latent class analysis in medical diagnosis. In *Papers presented*, volume 26, pages 2912–2916. World Scientific, 2002.

[37] Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

[38] FRANK SHANN. Etiology of severe pneumonia in children in developing countries. *The Pediatric Infectious Disease Journal*, 5(2):247–252, 1986.

[39] Varinder Singh and Satinder Aneja. Pneumonia–management in the developing world. *Paediatric respiratory reviews*, 12(1):52–59, 2011.

[40] Rosalyn J Singleton, Lisa R Bulkow, Karen Miernyk, Carolynn DeByle, Lori Pruitt, Kimberlee Boyd Hummel, Dana Bruden, Janet A Englund, Larry J Anderson, Lynne Lucher, et al. Viral respiratory infections in hospitalized and community control children in alaska. *Journal of medical virology*, 82(7):1282–1290, 2010.

[41] Alexandra Tikhomirova and Stephen P Kidd. Haemophilus influenzae and streptococcus pneumoniae: living together in a biofilm. *Pathogens and disease*, 69(2):114–126, 2013.

[42] P Toikka, T Juven, R Virkki, M Leinonen, J Mertsola, and O Ruuskanen. Streptococcus pneumoniae and mycoplasma pneumoniae coinfection in community acquired pneumonia. *Archives of disease in childhood*, 83(5):413, 2000.

[43] WHO UNICEF, WHO UNICEF, et al. Pneumonia: the forgotten killer of children. *UNICEF/WHO*, pages 1–40, 2006.

[44] Zhenke Wu, Maria Deloria-Knoll, Laura L Hammitt, and Scott L Zeger. Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2015.

[45] Zhenke Wu, Maria Deloria-Knoll, and Scott L Zeger. Nested partially-latent class models for dependent binary data; estimating disease etiology. 2015.

[46] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[47] Lue Ping Zhao and Ross L Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77(3):642–648, 1990.

Appendix