Baysian latent class model with partially informative priors for etiology estimation

Detian Deng

December 2, 2016

1 Introduction

1.1 Background

1.1.1 Child Pneumonia

Pneumonia is a form of acute respiratory infection of the lungs[?]. The infection can be caused by a variety of pathogens, indcluding bacteria, viruses, mycobacteria and fungi [?]. 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 [? ?]. Severe cases may be unable to feed or drink and may also experience unconsciousness, hypothermia and convulsions [?]. Although the majority of child pneumonia cases are nonsevere and can be managed in local primary health care facilities [?], 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 [? ?]. Under the pressure of such a severe public health burden, UNICEF and WHO declared pneumonia to be the "forgotten killer of children" in 2006 [?] and engaged the Global Action Plan for Prevention and Control of Pneumonia (GAPP) [?] 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 [? ?], 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 [?]: the wide use of pneumococcoal and haemophilus influenzae-B conjugate vaccines; the wide spread of HIV infection [?]; 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 [?] 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.[?]. 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 [?].

Specimen Measurements and Data Description

In order to maximize the detection power and accuracy of pathogen infection, the PERCH investors used multiple specimen types [**?**] 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) [**?**], 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.

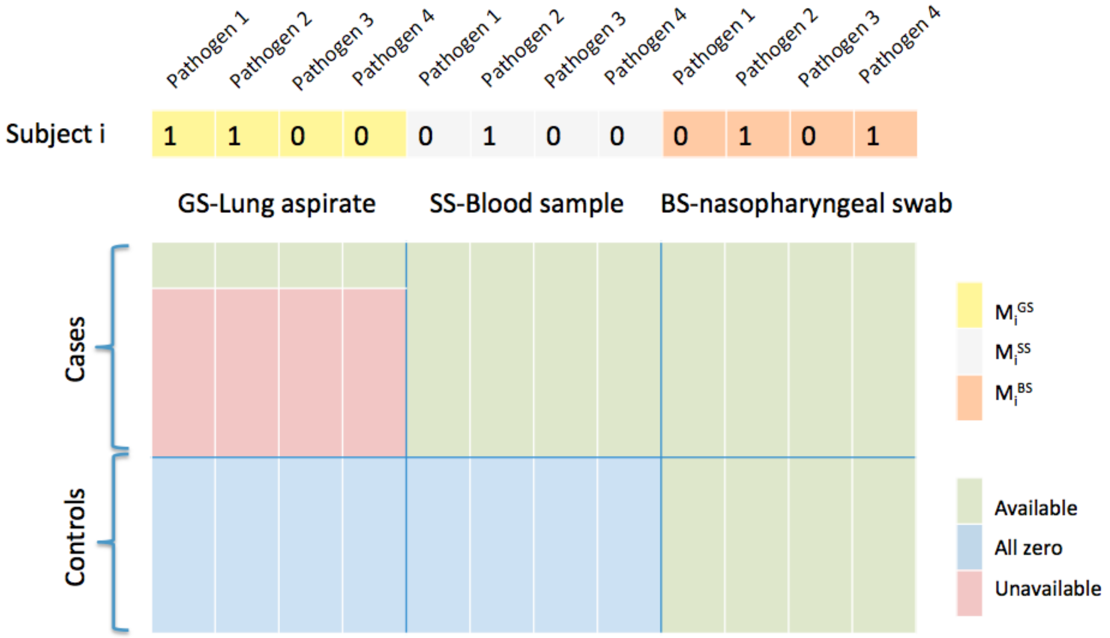

Figure 1: **Data Description**: Suppose we have 4 candidate pathogens in this demonstrative example. In the upper part of this figure, the three 4 dimensional vectors M_i^{GS}, M_i^{SS}, and M_i^{BS} are concatenated together. The GS measurement is available and it tells us pathogen 1 and 2 infect the 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 [**? ?**]. 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 [**?**].

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 [**?**]. 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)[**?**] and the nested-pLCM (npLCM) [**?**], introduced in the next section, as extensions to the classic Laten Class Model [**?**], 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[**?**]. 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) [**?**] 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[**? ?**], 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 [**?**] when the number of latent categories are large.

Therefore, Zhenke Wu, et al.(2015) developed the Bayesian partially-Latent Class Model (pLCM) [**?**] 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 [**?**] 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 [**?**] 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 [**?** **?**] suggested possible bacterial and viral coinfections for pneumonia cases and bacterial coinfections in lung tissue specimens were confirmed in the study of 2009 pandemic influenza A (H1N1) [**?**]. 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 [**?** **?** **?** **?**] and the non-likelihood based approach [**?** **?** **?**]. 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 [?], 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) [?] and later by Cox (1972) [?], 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) [?]. 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[?] was also developed as an extension to Liang and Zeger's (1986) generalized estimating equations (GEE)[?].

- **Log-linear Models**: The general log-linear model, first described by Cox (1972) [?] and discussed in depth by Haberman (1973) [?], 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) [**?**], 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) [**?**]. 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) [**?**] 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) [**?**], 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) [**?**] 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) [**?**] 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 [**?**] and the autologistic model, which models the dependence directly [**?**] 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

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

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

In order to allow the possibility of multiple pathogens infecting the lung, we propose to use a multivariate binary vector $L \in \{0,1\}^K$ as the latent variable describing the true status of the lung. Furthermore, a parsimonious quadratic exponential representation is developed to better incorporate prior knowledge. In the PERCH study, the prior knowledge scientists would like to incorporate in the model is the existence of strong competition among the majority of pathogens when they infect the lung while other few pathogens might infect independently.

In this section, we will define a novel representation of L that (1) supports stratified estimation and conditional regression modeling on the etiology parameter $\mu = \mathbb{E}(L|M, Y = 1)$, (2) reflects the competing nature among various pathogens, and (3) has good scalability, i.e. the number of parameters grows at most polynomially with K.

The method we propose is derived from the log-linear model representation, thus we will start with re-introducing the log-linear model in the context of the above four requirements and form the connections to our method.

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. $S_k = \sum_{k=0}^{K} L_i k$, 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

[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.

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

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

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

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

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

$$\sum_{k=1}^{K} \mu_k = \sum_{k=1}^{K} k\pi_k \qquad (6)$$

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

An important special case of the log-linear model is the "quadratic expoential family" (QE) described by Zhao and Prentice (1990) [**?**], 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). However, in the case where GS data are completely absent, such QE representation still over-parameterizes the latent variable, thus the parameter of interests are hardly identifiable. Therefore, further parsimonious parameterization need to be imposed.

We propose to use an exchangeable association structure with shrinkage estimation to incorporate prior knowledge as well as to improve the model 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 correlation between pathogens and I_k 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 method to estimate the probability of $I_k = 1$.

3.3 Pseudo-Quadratic Exponential Model

Given the above definition, we can decompose the $A(\Theta)$ into two components:

$$\begin{aligned} A(\Theta) =& \sum_{l^* \in \{0,1\}^K : \sum_{l^*} \leq s} \exp\{\Theta^T l^*\} + \sum_{l^* \in \{0,1\}^K : \sum l^* > s} \exp\{\Theta^T l^*\} \\ =& A_1(\Theta, s) + A_2(\Theta, s) \end{aligned}$$

Then the probability mass function can be written as:

$$\begin{aligned}P(L=l;\Theta)=&P(L=l|S\leq s;\Theta)P(S\leq s)+P(L=l|S>s;\Theta)P(S>s)\\=&\frac{\exp\{\Theta_1^T l+\Theta_2^T u_2\}\cdot 1(\sum l\leq s)}{A_1(\Theta,s)}\cdot\frac{A_1(\Theta,s)}{A(\Theta)}+\\&\frac{\exp\{\Theta_1^T l+\Theta_2^T u_2\}\cdot 1(\sum l> s)}{A_2(\Theta,s)}\cdot\frac{A_2(\Theta,s)}{A(\Theta)}\end{aligned}$$

If we allow $P(S>s)=\frac{A_2(\Theta,s)}{A(\Theta)}$ to be exactly 0, then the model violates the log linear model framework, but gains useful model sparsity for our interests, because in the etiology study, scientists have strong beliefs that there cannot be too many pathogens that jointly cause the disease. In other words, for some fixed integer S_{max} s.t. $1\leq S_{max}\leq K$, set P($S>S_{max}$) = 0. Therefore, the probability mass function of the latent variable can be simplified to:

$$P(L=l;\Theta)=\frac{\exp\{\Theta_1^T l+\Theta_2^T u_2\}\cdot 1(\sum l\leq S_{max})}{A_1(\Theta,S_{max})} \tag{8}$$

We call this representation as the Pseudo-Quadratic Exponential (PQE) Model since it adopts the same set of parameters as QE symbolically but it deviates from the log-linear family and the interpretation to the parameters are all conditioned on the assumption that $S\leq S_{max}$. This representation is useful since it imposes stronger competition among pathogens and it shrinks the size of the parameter expansion from $O(2^K)$ to $O(K^{S_{max}})$ comparing to regular QE model.

In both QE and PQE models, the parameter of interests μ can be recovered from the posterior samples of Θ using (2). And it is straightforward to extend them with conditional regression functionality by reparameterizing $\theta_{ik}^{(1)}$ as $X_i^T\beta_k$, where X_i is the vector of covariates.

4 Full Model Specification

By using a Metropolis-Hastings 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 $P(l_j|\Theta)$ can be either the QE model (7) or PQE model (8) at the user's choice, and

$$\begin{aligned}P(M_i^{SS}|l_j, \eta) =& \prod_{k=1}^{K} P(M_{ik}|l_{jk}, \eta_k) \\ =& \prod_{k=1}^{K} (\eta_k^{l_{jk}} l_{jk})^{M_{ik}} (1-\eta_k)^{l_{jk}(1-M_{ik})} \qquad (9) \\ P(M_i^{BS}|l_j, \gamma, \delta) =& \prod_{k=1}^{K} P(M_{ik}|l_{jk}, \gamma_k, \delta_k) \\ =& \prod_{k=1}^{K} (\gamma_k^{l_{jk}} \delta_k^{1-l_{jk}})^{M_{ik}} [(1-\gamma_k)^{l_{jk}} (1-\delta_k)^{1-l_{jk}}]^{1-M_{ik}} \qquad (10)\end{aligned}$$

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 (9), $P(M_i^{BS}|M_i^{GS}, \gamma, \delta)$ is defined using (10), 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:

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

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.

The prior distribution for the parameters in the likelihood of case data is:

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

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

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

The prior on Θ is defined as

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.

$$\begin{aligned} I_1 =& \Big\{ i \in \{1, 2, \ldots, n\} : Y_i = 1 \text{ and } G_i = 1 \Big\} \\ I_2 =& \Big\{ i \in \{1, 2, \ldots, n\} : Y_i = 1 \text{ and } G_i = 0 \Big\} \\ I_3 =& \Big\{ i \in \{1, 2, \ldots, n\} : Y_i = 0 \Big\} \end{aligned}$$

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

$$\begin{aligned} & P(M, \Theta, \eta, \gamma, \delta) \\ =& \prod_{i \in I_1} P(M_i^{GS}, M_i^{SS}, M_i^{BS} | \Theta, \eta, \gamma, \delta) \prod_{i \in I_2} P(M_i^{SS}, M_i^{BS} | \Theta, \eta, \gamma, \delta) \prod_{i \in I_3} P(M_i^{BS} | \gamma, \delta) P(\Theta, \eta, \gamma, \delta) \end{aligned}$$

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. With the joint density fully specified, we can build up a MCMC algorithm to simulate from the posterior distribution. The details fo the algorithm can be found in the appendix.

5 Simulation Study

A simulation study is done to explore the effectiveness of our proposed method. Based on the parameter values shown in table 2, a set of 3-dimensional multivariate binary measurement data are simulated. In total, 1000 controls and 1000 cases are simulated. For each control, only BS measurements are sampled. For cases, they all have SS and BS measurements and 3% of them have additional GS measurements simulated.

A sequence of experiments are carried out based on the simulated data set. First, with fixed 3% of GS measurements and 1000 controls, we sample the posterior distribution with case sample size 200, 400, 600, and 1000 respectively. Then, following the same sequence of sample sizes, we sample the posterior distribution without using GS measurements. Figure **??** summarizes the prior and posterior distributions of (μ, π) from these experiments. Prior distributions used for other parameters are plotted in figure **??**, as we can see, aach prior is tuned to be weakly informative, that is, to have relatively large variance and to center at wrong values that are not too far away from the truth.

The results of this simulation study suggest that only with a small amount of direct (GS) measurements, the posterior distribution is converging to the true value as the number of observation grows. Also, the posterior credible region has a high coverage probability over the true value even when direct measurements are completely absent. Since the priors for sensitivities and specificities were set to be weakly informative, the robustness of this model is empirically testified.

Table 2: Summary of the True Parameter Values

μ_1	μ_2	μ_3	π_0	π_1	π_2	π_3	$\text{TPR}_1^{(SS)}$	$\text{TPR}_2^{(SS)}$	$\text{TPR}_3^{(SS)}$	$\text{TPR}_1^{(BS)}$	$\text{TPR}_2^{(BS)}$	$\text{TPR}_3^{(BS)}$	$\text{FPR}_1^{(BS)}$	$\text{FPR}_2^{(BS)}$	$\text{FPR}_3^{(BS)}$
0.584	0.289	0.378	0.10	0.60	0.25	0.05	0.05	0.10	0.10	0.75	0.80	0.70	0.50	0.50	0.40