Pathogen Etiology Inference for Pneumonia Using Partially Latent Class Models

An Exploratory Analysis Using Bronze-Standard Measurements

Mackenzie Mueller
Applied Bayesian Inference
Dr. Jian Kang

December 14, 2025

1 Introduction

1.1 Background

Pneumonia is an infection of the lungs that can be caused by bacteria, viruses, and fungi, and is an especially serious disease for infants and young children [4]. Because of the nature of the disease, it is difficult to diagnose in a non-invasive way. The data for this project is simulated based on the Pneumonia Etiology Research for Child Health (PERCH) case-control study, which lasted from 2011-2014 and was conducted at 9 sites in 7 countries [3]. While there are 3 ways to diagnose this disease, the data from the Bronze Standard (BrS) measurement (a nasopharyngeal PCR test) is used for the purpose of this project. The goal of the PERCH study, and the goal of my project, is to identify the percentage of specific pathogens that contribute to a pneumonia diagnosis. To address the limitations of imperfect diagnostic tests such as the nasopharyngeal PCR, this analysis applies a partially latent class modeling framework to infer pathogen-specific etiologic fractions from observed BrS data.

1.2 Data

The simulated data consist of 972 observations from children who received a nasopharyngeal test to determine pneumonia status. There are 486 cases (children with pneumonia) and 486 controls (children without pneumonia). The data provides information on the presence of 6 pathogens (unknown but labeled A through F), the age of the child (> 1 or < 1), and HIV status of the child (negative or positive). For the purpose of this project, there are

4 stratum: children < 1 who are HIV negative, children < 1 who are HIV positive, children > 1 who are HIV negative, and children > 1 who are HIV positive. The data was perfectly simulated so that the distributions of cases and controls with respect to age and HIV status are the same. See table 1 for a summary of the data.

2 Model and Methods

To specify a model, the three things to consider are the data, the model options, and the MCMC options. For this project, JAGS software was used to run Markov Chain Monte Carlo methods in RStudio version 4.5.2, with use of the baker package (version 1.0.4) [2].

To specify the data, a csv file was downloaded that contained subject ID, pathogen information, age of the child, and HIV status of the child. This data was then converted to a format of lists that made it suitable for use with the baker package in R. For model specifications, we need to consider the measurements, the likelihood, and the priors for our dataset. The measurements are Bronze Standard, abbreviated as "BrS".

For the likelihood, I followed the suggested structure from Wu et al [1]. The PLCM model takes three sets of parameters: π, ψ, θ. π is a vector of compositional probabilities for each of J pathogen causes. ψ is the false positive rate (FPR) for measurement j at site S. θ is the true positive rate (TPR) for measurement j at site S for a person whose lung is infected by pathogen j. With these parameters, the first likelihood model assumes a Bernoulli dis-

tribution with conditional independence, used to describe the model for BrS measurement for a control or a case:

$$P_i^{0,BS} = \Pr\left(M_i^{BS} = m \mid \boldsymbol{\psi}^{BS}\right) = \prod_{j=1}^{J} \left(\psi_j^{BS}\right)^{m_j} \left(1 - \psi_j^{BS}\right)^{1-m_j}$$

Then,

$$P_i^{1,BS} = \Pr\left(M_i^{BS} = m \mid \boldsymbol{\pi}, \boldsymbol{\theta}^{BS}, \boldsymbol{\psi}^{BS}\right) =$$

$$\sum_{j=1}^{J} \pi_j \left(\theta_j^{BS}\right)^{m_j} \left(1 - \theta_j^{BS}\right)^{1-m_j} \prod_{l \neq j} \left(\psi_l^{BS}\right)^{m_l} \left(1 - \psi_l^{BS}\right)^{1-m_l}$$

is the likelihood contributed by BS measurements from case i. The likelihood is fit for each of 4 stratum made up of combinations of the 2 discrete variables: age (> 1, < 1), and HIV (positive, negative). The final posterior predictive probability is a weighted average of the 4 stratum.

For the priors, we consider an etiology prior and a TPR prior. The etiology prior is specified via a vector of length equal to the number of causes. Each row is a vector of Dirichlet hyperparameters for the population etiologies in the stratum defined by the discrete covariates (age and HIV). For the TPR prior, I chose to use an informative prior to be able to include prior existing knowledge about the true positive rates of each pathogen causing pneumonia. I chose to specify the input for this prior as a Beta(6,2) distribution directly. The means that before seeing any data, the BrS test correctly detects a true cause about 75% of the time. This is a moderately informative prior, chosen because of the nature of the BrS nasopharyngeal test, which is able to detect cases most of the time.

The third consideration for my model is the MCMC options. While initial basic modeling utilized only 2 chains and 1000 iterations, this model was lacking in effective convergence. My final selected model utilizes 3 chains, 5000 iterations, with 1000 burn-in. I checked convergence for this final model using trace plots and R hat values, using a cutoff of 1.05 to show effective model convergence. All values are close to 1.0, shown in Figure 1.

3 Data Analysis

3.1 Exploratory Data Analysis

To begin exploratory data analysis (EDA), I started by modeling the probability of case/control status using all available covariates. This allowed me to estimate odds ratios of a child being a case based on the presence of a specific pathogen. Logistic regression via a glm in R was used to fit the following model:

$$\text{logit}(P(Y=1)) \sim \beta_1 X_{\text{pathogenA}} + \beta_2 X_{\text{pathogenB}} + ... + \beta_6 X_{\text{pathogenF}} + \beta_7 X_{\text{age}} + \beta_8 X_{\text{HIV}}$$

Results from this glm show that Pathogen A and Pathogen B are both strongly associated with increased odds of pneumonia status (p-values < 0.001).

Secondary EDA involved looking at co-occurrences between pathogens. These models predict the probability of each pathogen being present given the covariate data. Six logistic regression models were built with each of the 6 pathogens as the outcome variable. Based on the summary outputs, there

are strong associations between pathogen A and children older than 1, and between pathogen B and children who are HIV positive.

Additional EDA looked at the pairwise log-odds ratios for the dataset, see Figure 2. The log-odds ratios are all close to zero, which indicates minimal residual dependence between pathogen measurements. This motivates the choice to use a non-nested partially latent class model (PLCM), which assumes conditional independence.

3.2 Results

The final model returns a summary of the 6 pathogens and their respective posterior etiology means across all 4 stratum levels using the weights: (0.66,0.247,0.053,0.039). The interpretations for each of the posterior means show the percentage of each pathogen that make up the disease status (pneumonia). Results are as follows: it is estimated that Pathogen A and Pathogen B make up the greatest fractions of pneumonia cases across the 4 stratum investigated, at 38% and 26.6%, respectively. It is also estimated that pathogen C makes up 16% of pneumonia cases, Pathogen D makes up 10.7% of pneumonia cases, Pathogen E makes up 4.4% of pneumonia cases, and Pathogen F makes up 3.9% of pneumonia cases. See table 2 and figure 3 for these results.

To check the model, I compared the observed pairwise log-odds ratios to the posterior predictive distributions of pairwise log-odds ratios. The numbers are calculated using this formula:

$$\frac{\text{Predicted LOG - Observed LOR}}{\text{SE(Posterior Predictive Distrubtion of LOR)}}$$

The closer to zero the values are, the better the model is at predicting the correct log-odds ratios. We are able to check how well the model fits using R. The model is able to perfectly predict the log-odds ratios, which we can see in Figure 4 as all the values are equal to zero.

4 Discussion

Overall, the model performed well with the previously stated model specifications. Based on Dr. Wu's paper, I applied adequate likelihood and prior specifications based on former research. I was able to run an initial model, adjust the necessary components to improve it, and clearly see effective results from those changes. Based on the results, we can conclude that the two pathogens that make up the highest fraction of pneumonia cases are Pathogens A and B. We can say that the model performed well under the given specifications, and can say that the model converged successfully.

4.1 Future Directions

In this analysis, I fit a non-nested PLCM with a single latent subclass (K = 1). This was supported by initial pairwise log-odds ratios that were close to zero. However, future work for this project could include running a nested model with K $>$ 1 to relax the conditional independence assumption and to accommodate any local dependencies between pathogen measurements.

5 Figures and Tables

Age	HIV positive	% Controls	% Cases
< 1	0	66.0	66.0
≥ 1	0	24.7	24.7
< 1	1	5.3	5.3
≥ 1	1	3.9	3.9

Table 1: Distribution of controls and cases by age and HIV status

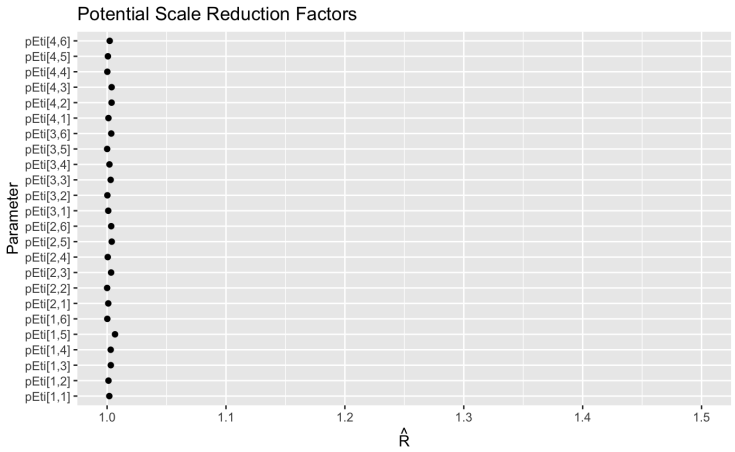

Figure 1: Rhat Values to show Model Convergence

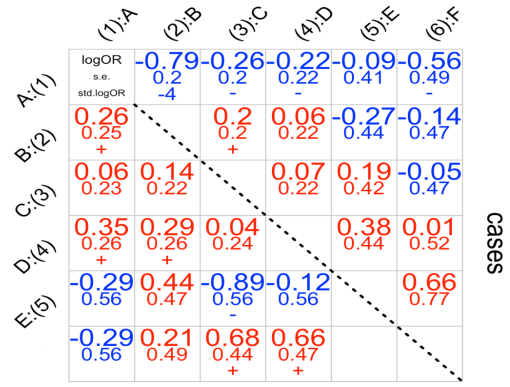

Figure 2: Pairwise log-odds ratios for Pathogens A-F

Cause	Posterior Mean	2.5% CI	97.5% CI
A	0.383	0.258	0.508
B	0.266	0.163	0.415
C	0.160	0.055	0.302
D	0.107	0.035	0.230
E	0.045	0.010	0.121
F	0.039	0.009	0.101

Table 2: Posterior means and 95% credible intervals for pathogen-specific etiology fractions.

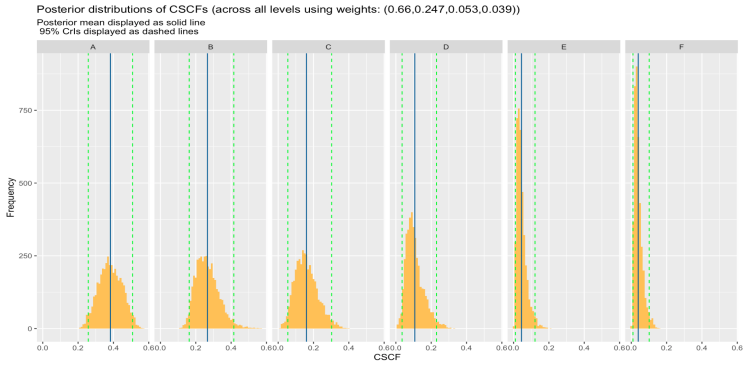

Figure 3: Posterior distributions of Pathogens

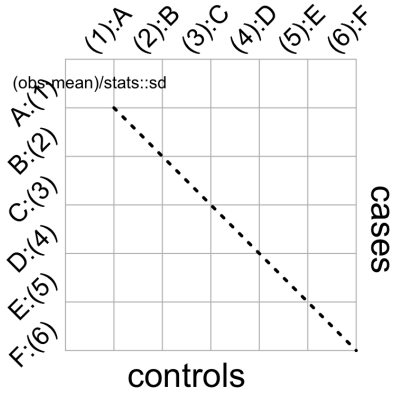

Figure 4: Pairwise LOR Comparisons

References

[1] Wu, Z., Deloria-Knoll, M., Hammitt, L. L., Zeger, S. L. and the Pneumonia Etiology Research for Child Health Core Team (2016), Partially latent class models for case–control studies of childhood pneumonia aetiology. J. R. Stat. Soc. C, 65: 97–114.

[2] Chen, I., Shi, Q., Zeger, S. L., & Wu, Z. (2022). baker: An R package for nested partially-latent class models (Version 1.0.4) [Computer software]. https://github.com/zhenkewu/baker

[3] Johns Hopkins Bloomberg School of Public Health. (n.d.). Pneumonia Etiology Research for Child Health (PERCH). International Vaccine Access Center. https://publichealth.jhu.edu/ivac/pneumonia-etiology-research-for-child-health-perch

[4] Mayo Clinic. (n.d.). Pneumonia: Symptoms and causes. https://www.mayoclinic.org/diseases-conditions/pneumonia/symptoms-causes/syc-20354204

[5] Wu, Z. (n.d.). Github Repository. GitHub. https://github.com/zhenkewu/