Method

To allow for the possibility of multiple pathogens infecting the lung, a multivariate binary vector $\boldsymbol{L}$ is defined to be latent variable that denotes the actual status of the lung. For every element l_k in $\boldsymbol{L}$, $l_k = 1$ suggests that pathogen k is present in the lung, otherwise $l_k = 0$. Each l_k may be correlated or independent from the other elements in $\boldsymbol{L}$, thus it allows for the multi-pathogen infection situation. Given the latent variable $\boldsymbol{L}$, the measurements for different pathogens from various peripheral sites are assumed to be conditionally independent and parameterized by their corresponding true positive rates (TPR) and false positive rates (FPR). Also, informative priors are used for the TPRs. The priors are specified as independent Beta distributions where the hyper-parameters that are determined by a credible interval matching procedure.

The full model for multivariate binary vector is too flexible to be identified and estimtated from non-Gold standard data. Some sensible structure and constraint must be imposed for identifiability. Therefore the joint distribution of the latent lung status is defined by a quadratic exponential model* with a particular association structure to incorporate prior knowledge. In the PERCH study, scientists believe that there exists strong competition among the majority of pathogens when they infect the lung, while other few pathogens, unaffected by the interspecific competition, might infect more nearly independently. Thus, the association structure is defined as a random adjacency vector**. Each element in this vector is a product of the common association parameter and an indicator that represents whether there is any competition between the corresponding pair of latent variables. A hierarchical prior*** is placed on the indicator variables to facilitate the stochastic search procedure.

Two Metropolis-within-Gibbs sampling algorithms are developed. Let K be the length of $\boldsymbol{L}$ and let n be the number of case observations. The first algorithm samples from the conditional distribution of $\boldsymbol{L}$ exactly which takes $O(K^2 2^K + n)$ in each iteration. The second algorithm samples each element in $\boldsymbol{L}$ sequentially whose time complexity is $O(K^2 n)$. The second algorithmis more scalable on K than the first, while the first algorithm is faster with small K and takes fewer samples to converge.

With the above two Monte-Carlo algorithms designed to estimate the exact posterior distribution, a variational Bayesian EM algorithm is derived for fast and scalable maximum a posteriori (MAP) estimation. This algorithm updates the parameters cyclically using either closed form or univariate optimization, which makes it extremely fast in computation.

*The quadratic exponential (QE) model is a classic model for parameterizing the joint distribution of binary variables. It has a 1-to-1 mapping to the auto-logistic model, which basically regresses each binary element on every other elements using a logistic regression. The canonical parameters in QE model are conditional log odds and conditional log odds ratios.

**The adjacency vector is a vector of the conditional log odds ratios. Each element of this vector corresponds to a pair of latent binary variables. The name 'adjacency' is borrowed from graph theory, where "= 0" means not connected, i.e. conditionally independent, and "> 0" means connected or adjacent, i.e. conditionally dependent.

***The hierarchical prior here is an analog of the spike-and-slab prior in classic Bayesian variable selection model, and leads to the sparse correlation estimates in our model.