1 Current Model

Let $S_i = (S_{i1}, \ldots, S_{iK})$, $S_{ik} = 1(\sum_{j=1}^{K} Y_{ij} = k)$, then we have $\pi_k = P(\sum_{j=1}^{K} Y_{ij} = k) = \mathbb{E}(S_i)$, and for each $i = 1, \ldots, n$

$$\mathbb{E}(\sum_{k=1}^{K} Y_{ik}) = \sum_{k=1}^{K} \mu_{ik}(\beta) = \sum_{k=1}^{K} k\pi_{ik} \tag{1}$$

Let $\gamma = (\gamma_1, \ldots, \gamma_K)$, where

$$\gamma_k = \frac{\pi_{ik}}{\sum_{k=1}^{K} \mu_{ik}} > 0 \text{ and } \sum_{k=1}^{K} k\gamma_k = 1 \tag{2}$$

Note that although π_i is dependent on μ_i, γ now can be modeled independent of μ_i, and the n equality constraints on (μ_i, π_i) specified by (1) can be reduced to only 1 equality constraint on γ in (2).

In the regression setting where $\mathrm{g}(\mu_i) = X_i^T \beta$, and g is the logit link function.

$$\begin{aligned} P(L|\mu_i, \pi_i) =& P(L|X_i; \beta, \gamma) \qquad (3) \\ =& \int P(L|\phi, \mu_i, \pi_i) P(\phi|\mu_i, \pi_i) d\phi \\ =& \int P(L|\phi) P(\phi|\mu_i, \pi_i) d\phi \\ \approx& \frac{1}{H} \sum_{h=1}^{H} P(L|\phi_i^{(h)}) \qquad (4) \end{aligned}$$

where $\phi_i^{(h)}$ is sampled from $P(\phi|\mu_i = g^{-1}(X_i^T\beta), \pi_i = \gamma\mu_i^T 1)$

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 $\mathbb{L}$.

1.1 The Likelihood for Cases

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

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

where $P(l_j|\mu,\pi)$ is defined using (10), and

$$\begin{aligned} P(M_i^{SS}|l_j,\eta) &= \prod_{k=1}^{K} P(M_{ik}|l_{jk},\eta_k) \\ &= \prod_{k=1}^{K} (\eta_k^{l_{jk}} l_{jk})^{M_{ik}} (1-\eta_k)^{l_{jk}(1-M_{ik})} \qquad (5) \\ P(M_i^{BS}|l_j,\gamma,\delta) &= \prod_{k=1}^{K} P(M_{ik}|l_{jk},\gamma_k,\delta_k) \\ &= \prod_{k=1}^{K} (\gamma_k^{l_{jk}} \delta_k^{1-l_{jk}})^{M_{ik}} [(1-\gamma_k)^{l_{jk}} (1-\delta_k)^{1-l_{jk}}]^{1-M_{ik}} \qquad (6) \end{aligned}$$

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

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

where $P(M_i^{SS}|M_i^{GS},\eta)$ is defined using (12), $P(M_i^{BS}|M_i^{GS},\gamma,\delta)$ is defined using (13), and $P(M_i^{GS}|\mu,\pi)$ is defined using (10).

1.2 The Likelihood for Controls

For controls, we only have BS measurements and we know that their lungs were not infected. Since (μ,π) are defined for case only, they are not involved in the likelihood for controls, thus the

likelihood function is:

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

1.3 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:

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

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.