Supplementary Materials for "A mixed-effects Bayesian regression model for multivariate group testing data"

Christopher S. McMahan[1,*], **Chase N. Joyner**[1], **Joshua M. Tebbs**[2], **and Christopher R. Bilder**[3]

[1]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, U.S.A.

[2]Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.

[3]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, U.S.A.

[*]*email*: mcmaha2@clemson.edu

Web Appendix A: *Full conditional distributions, derivations, and expressions.* We derive full conditional distributions below and give expressions for the parameters in these distributions:

$$\begin{aligned}
\widetilde{Y}_{id} \mid \widetilde{\mathbf{Y}}_{i(-d)}, \mathbf{Z}, \boldsymbol{\Theta} &\sim \text{Bernoulli}(p^*_{id}) \\
\boldsymbol{\omega}_i \mid \widetilde{\mathbf{Y}}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R} &\sim \text{TMN}(\boldsymbol{\eta}_i, \mathbf{R}, \mathbf{L}_i, \mathbf{U}_i) \\
\boldsymbol{\beta}_{\boldsymbol{v}} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v} &\sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\
\lambda_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \mathbf{R}, w_{ld} &\sim \text{TN}(\mu_{\lambda_{ld}} w_{ld}, \sigma^2_{\lambda_{ld}} w_{ld}, 0, \infty) \\
\mathbf{a} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{b}, \mathbf{R} &\sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}) \\
\mathbf{b}_k \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{R} &\sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k}) \\
v_{rd} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, \tau_{v_{rd}} &\sim \text{Bernoulli}(p_{v_{rd}}) \\
w_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \tau_{w_{ld}} &\sim \text{Bernoulli}(p_{w_{ld}}) \\
\tau_{v_{rd}} \mid v_{rd} &\sim \text{beta}(a_v + v_{rd}, b_v + 1 - v_{rd}) \\
\tau_{w_{ld}} \mid w_{ld} &\sim \text{beta}(a_w + w_{rd}, b_w + 1 - w_{rd}) \\
S_{e(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} &\sim \text{beta}(a^\star_{e(m):d}, b^\star_{e(m):d}) \\
S_{p(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} &\sim \text{beta}(a^\star_{p(m):d}, b^\star_{p(m):d}).
\end{aligned}$$

We henceforth make use of the following notation: $\mathbf{X}_i = \oplus_{d=1}^D \mathbf{x}'_{id}$, $\mathbf{T}_i = \oplus_{d=1}^D \mathbf{t}'_{id}$, $\boldsymbol{\Lambda} = \oplus_{d=1}^D \boldsymbol{\Lambda}_d$, $\mathbf{A} = \oplus_{d=1}^D \mathbf{A}_d$, $\boldsymbol{v} = (\boldsymbol{v}'_1, ..., \boldsymbol{v}'_D)'$, and $\boldsymbol{v}_d = (v_{1d}, ..., v_{p_d d})'$.

Full conditional of $\widetilde{Y}_{id}$: From the joint distribution of the observed testing outcomes and the individuals' latent statuses, given by

$$\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\Theta}) = \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{j \in \mathcal{I}_m} \left\{ S^{Z_{jd}}_{e(m):d} (1 - S_{e(m):d})^{1 - Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S^{1 - Z_{jd}}_{p(m):d} (1 - S_{p(m):d})^{Z_{jd}} \right\}^{1 - \widetilde{Z}_{jd}} \times \prod_{i=1}^{N} \pi(\widetilde{\mathbf{Y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}),$$

it is easy to see that the full conditional distribution of $\widetilde{Y}_{id}$ is Bernoulli. In particular, $\widetilde{Y}_{id} \mid \widetilde{\mathbf{Y}}_{i(-d)}, \mathbf{Z}, \boldsymbol{\Theta} \sim \text{Bernoulli}(p^*_{id})$, where $\widetilde{\mathbf{Y}}_{i(-d)}$ is the vector $\widetilde{\mathbf{Y}}_i$ with the dth element removed,

$p^*_{id} = p^\star_{id1}/(p^\star_{id0} + p^\star_{id1})$, and

$$\begin{aligned} p^\star_{id1} &= p_{id} \prod_{j \in \mathcal{A}_i} S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1 - Z_{jd}} \\ p^\star_{id0} &= (1 - p_{id}) \prod_{j \in \mathcal{A}_i} \left\{ S_{e_j:d}^{Z_{jd}} (1 - S_{e_j:d})^{1 - Z_{jd}} \right\}^{I\left(s_{ijd} > 0\right)} \left\{ (1 - S_{p_j:d})^{Z_{jd}} S_{p_j:d}^{1 - Z_{jd}} \right\}^{I(s_{ijd} = 0)}. \end{aligned}$$

In the expressions above, $p_{id} = \pi(\widetilde{\mathbf{Y}}_{i(d)} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R})$, $\widetilde{\mathbf{Y}}_{i(d)} = (\widetilde{Y}_{i1}, ..., \widetilde{Y}_{id} = 1, ..., \widetilde{Y}_{iD})'$, the index set $\mathcal{A}_i = \{j \colon i \in \mathcal{P}_j\}$ keeps track of which pools the ith individual belongs to, and $s_{ijd} = \sum_{i' \in \mathcal{P}_j \colon i' \neq i} \widetilde{Y}_{i'd}$. If $j \in \mathcal{I}_m$, then $S_{e_j:d} = S_{e(m):d}$ and $S_{p_j:d} = S_{p(m):d}$.

Full conditional of $\boldsymbol{\omega}_i$: From the joint distribution

$$\begin{aligned} \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\Theta}) \propto \prod_{d=1}^{D} \prod_{m=1}^{M} \prod_{j \in \mathcal{I}_m} & \left\{ S_{e(m):d}^{Z_{jd}} (1 - S_{e(m):d})^{1 - Z_{jd}} \right\}^{\widetilde{Z}_{jd}} \left\{ S_{p(m):d}^{1 - Z_{jd}} (1 - S_{p(m):d})^{Z_{jd}} \right\}^{1 - \widetilde{Z}_{jd}} \\ & \times \prod_{i=1}^{N} |\mathbf{R}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{\omega}_i - \boldsymbol{\eta}_i)' \mathbf{R}^{-1} (\boldsymbol{\omega}_i - \boldsymbol{\eta}_i) \right\} \prod_{i=1}^{N} f(\boldsymbol{\omega}_i), \end{aligned}$$

one can see the full conditional distribution of $\boldsymbol{\omega}_i$ is multivariate truncated normal with mean $\boldsymbol{\eta}_i$, correlation matrix $\mathbf{R}$, lower truncation limits $\mathbf{L}_i = (L_{i1}, ..., L_{iD})'$, and upper truncation limits $\mathbf{U}_i = (U_{i1}, ..., U_{iD})'$. The truncation region for the dth dimension is $L_{id} = 0$ and $U_{id} = \infty$ if $\widetilde{Y}_{id} = 1$ and $L_{id} = -\infty$ and $U_{id} = 0$ if $\widetilde{Y}_{id} = 0$; i.e.,

$$\boldsymbol{\omega}_i \mid \widetilde{\mathbf{Y}}_i, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R} \sim \mathrm{TMN}(\boldsymbol{\eta}_i, \mathbf{R}, \mathbf{L}_i, \mathbf{U}_i).$$

Full conditional of $\boldsymbol{\beta}$: The full conditional distribution of β_{rd} is degenerate at 0 if $v_{rd} = 0$, while the nonzero elements of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_{\boldsymbol{v}}$, have the normal full conditional distribution

$$\boldsymbol{\beta}_{\boldsymbol{v}} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}, \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}).$$

The mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\beta}} &= \left\{ \boldsymbol{\Phi}(\boldsymbol{v})^{-1} + \sum_{i=1}^{N} \mathbf{X}_i(\boldsymbol{v})' \mathbf{R}^{-1} \mathbf{X}_i(\boldsymbol{v}) \right\}^{-1} \sum_{i=1}^{N} \mathbf{X}_i(\boldsymbol{v})' \mathbf{R}^{-1} \boldsymbol{\omega}^\star_{\boldsymbol{\beta} i} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \left\{ \boldsymbol{\Phi}(\boldsymbol{v})^{-1} + \sum_{i=1}^{N} \mathbf{X}_i(\boldsymbol{v})' \mathbf{R}^{-1} \mathbf{X}_i(\boldsymbol{v}) \right\}^{-1}, \end{aligned}$$

where $\boldsymbol{\Phi}(\boldsymbol{v})$ is the matrix formed by retaining the rows and columns of $\boldsymbol{\Phi} = \mathrm{diag}(\phi^2_{rd}; r = 1, ..., p_d, d = 1, ..., D)$ that correspond to the non-zero elements of $\boldsymbol{v}$. Also, $\mathbf{X}_i(\boldsymbol{v})$ is the matrix formed by retaining the columns of $\mathbf{X}_i$ corresponding to the non-zero elements of $\boldsymbol{v}$, and $\boldsymbol{\omega}^\star_{\boldsymbol{\beta} i} = \boldsymbol{\omega}_i - \mathbf{T}_i \boldsymbol{\Lambda} \mathbf{A} \mathbf{b}_{(i)}$.

Full conditional of λ_{ld}: We introduce new notation. For the ith individual, define the $q_d \times 1$ vector $\mathbf{e}_{id}$ whose lth element is $t_{idl} b_{(i)dl} + t_{idl} \sum_{m=1}^{l-1} b_{(i)dm} a_{dlm}$, where t_{idl} is the lth element of $\mathbf{t}_{id}$,

$b_{(i)dl}$ is the lth element of $\mathbf{b}_{(i)d}$, and a_{dlm} is the (l,m)th entry of $\mathbf{A}_d$. Construct $\mathbf{E}_i = \oplus_{d=1}^{D} \mathbf{e}'_{id}$. Based on this notation, we can succinctly express the full conditional distribution of λ_{ld}, the ℓth element of $\boldsymbol{\lambda}$. In particular, the full conditional of λ_{ld} is degenerate at 0 if $w_{ld} = 0$. When $w_{ld} = 1$, the full conditional is

$$\lambda_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \mathbf{R}, w_{ld} \sim \text{TN}(\mu_{\lambda_{ld}}, \sigma^2_{\lambda_{ld}}, 0, \infty),$$

where

$$\begin{aligned} \mu_{\lambda_{ld}} &= \left(\frac{1}{\boldsymbol{\Psi}_{\ell\ell}} + \sum_{i=1}^{N} \mathbf{E}_i^{\ell\prime} \mathbf{R}^{-1} \mathbf{E}_i^{\ell} \right)^{-1} \sum_{i=1}^{N} \mathbf{E}_i^{\ell\prime} \mathbf{R}^{-1} \boldsymbol{\omega}^{\star}_{\lambda_\ell i} \\ \sigma^2_{\lambda_{ld}} &= \left(\frac{1}{\boldsymbol{\Psi}_{\ell\ell}} + \sum_{i=1}^{N} \mathbf{E}_i^{\ell\prime} \mathbf{R}^{-1} \mathbf{E}_i^{\ell} \right)^{-1}. \end{aligned}$$

In the expressions above, $\mathbf{E}_i^{\ell}$ is the ℓth column of $\mathbf{E}_i$, $\boldsymbol{\Psi}_{\ell\ell}$ is the ℓth diagonal element of $\boldsymbol{\Psi} = \text{diag}(\psi^2_{ld}; l = 1, ..., q_d, d = 1, ..., D)$, $\boldsymbol{\omega}^{\star}_{\lambda_\ell i} = \boldsymbol{\omega}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{E}_i^{(-\ell)}\boldsymbol{\lambda}_{(-\ell)}$, $\mathbf{E}_i^{(-\ell)}$ is the matrix that remains after removing the ℓth column of $\mathbf{E}_i$, and $\boldsymbol{\lambda}_{(-\ell)}$ is the vector that remains after removing λ_{ld} from $\boldsymbol{\lambda}$.

Full conditional of $\mathbf{a}$: We introduce new notation. Define the $q_d \times (q_d - 1)/2$ vector $\mathbf{u}_{id} = (b_{(i)dl}\lambda_{dm}t_{idm}; l = 1, ..., q_d - 1, m = l + 1, ..., q_d)'$ and construct $\mathbf{U}_i = \oplus_{d=1}^{D} \mathbf{u}'_{id}$, where $b_{(i)dl}$ is the lth element of $\mathbf{b}_{(i)d}$, λ_{dm} is the mth element of $\boldsymbol{\lambda}_d$, and t_{idm} is the mth element of $\mathbf{t}_{id}$. The linear predictor can be re-expressed as $\eta_{id} = \mathbf{x}'_{id}\boldsymbol{\beta} + \mathbf{t}'_{id}\boldsymbol{\Lambda}_d\mathbf{b}_{(i)d} + \mathbf{u}'_{id}\mathbf{a}_d$, and it is easy to see the full conditional distribution

$$\mathbf{a} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{b}, \mathbf{R} \sim N(\boldsymbol{\mu}_{\mathbf{a}}, \boldsymbol{\Sigma}_{\mathbf{a}}).$$

The mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{a}} &= \left(\mathbf{C}^{-1} + \sum_{i=1}^{N} \mathbf{U}'_i \mathbf{R}^{-1} \mathbf{U}_i \right)^{-1} \left(\mathbf{C}^{-1}\mathbf{m} + \sum_{i=1}^{N} \mathbf{U}'_i \mathbf{R}^{-1} \boldsymbol{\omega}^{\star}_{\mathbf{a}i} \right) \\ \boldsymbol{\Sigma}_{\mathbf{a}} &= \left(\mathbf{C}^{-1} + \sum_{i=1}^{N} \mathbf{U}'_i \mathbf{R}^{-1} \mathbf{U}_i \right)^{-1}, \end{aligned}$$

where $\boldsymbol{\omega}^{\star}_{\mathbf{a}i} = \boldsymbol{\omega}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{T}_i\boldsymbol{\Lambda}\mathbf{b}_{(i)}$, $\mathbf{C} = \text{diag}(\mathbf{C}_1, ..., \mathbf{C}_D)$, and $\mathbf{m} = (\mathbf{m}'_1, ..., \mathbf{m}'_D)'$. Recall $\mathbf{m}_d$ and $\mathbf{C}_d$ are hyperparameters defined in Section 2 of the manuscript.

Full conditional of $\mathbf{b}_k$: Define $\mathcal{S}_k = \{i : \mathbf{b}_{(i)} = \mathbf{b}_k\}$ to be the index set of individuals who visited site k. The full conditional distribution of $\mathbf{b}_k$ is

$$\mathbf{b}_k \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{R} \sim N(\boldsymbol{\mu}_{\mathbf{b}_k}, \boldsymbol{\Sigma}_{\mathbf{b}_k}),$$

where the mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{b}_k} &= \left(\mathbf{I} + \sum_{i \in \mathcal{S}_k} \mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}'_i\mathbf{R}^{-1}\mathbf{T}_i\boldsymbol{\Lambda}\mathbf{A} \right)^{-1} \sum_{i \in \mathcal{S}_k} \mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}'_i\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\mathbf{b}_k i} \\ \boldsymbol{\Sigma}_{\mathbf{b}_k} &= \left(\mathbf{I} + \sum_{i \in \mathcal{S}_k} \mathbf{A}'\boldsymbol{\Lambda}\mathbf{T}'_i\mathbf{R}^{-1}\mathbf{T}_i\boldsymbol{\Lambda}\mathbf{A} \right)^{-1} \end{aligned}$$

and $\boldsymbol{\omega}^{\star}_{\mathbf{b}_k i} = \boldsymbol{\omega}_i - \mathbf{X}_i\boldsymbol{\beta}$.

Full conditional of v_{rd}: Under the Dirac spike, $\boldsymbol{v}$ should be sampled from its marginal posterior, which is obtained after integrating over $\boldsymbol{\beta}$; i.e.,

$$\begin{aligned}\pi(\boldsymbol{v} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{\tau}_v) &\propto \pi(\boldsymbol{v}|\boldsymbol{\tau}_v) \int \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\Theta})\pi(\boldsymbol{\beta} \mid \boldsymbol{v})d\boldsymbol{\beta} \\ &\propto \pi(\boldsymbol{v}|\boldsymbol{\tau}_v)\pi(\boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}),\end{aligned}$$

where $\boldsymbol{\tau}_v = (\tau_{v_{rd}}; r = 1, ...p_d, , d = 1, ..., D)'$ and

$$\pi(\boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}) \propto |\boldsymbol{\Phi}(\boldsymbol{v})|^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{1/2} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N} \boldsymbol{\omega}^{\star'}_{\boldsymbol{\beta} i}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta} i} - \boldsymbol{\mu}'_{\boldsymbol{\beta}}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\right\},$$

where $\boldsymbol{\Phi}(\boldsymbol{v})$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, and $\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta} i}$ are defined in the full conditional derivation of $\boldsymbol{\beta}$ above. If $\boldsymbol{v} = \mathbf{0}$, then this marginalized likelihood reduces to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{\omega}^{\star'}_{\boldsymbol{\beta} i}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\boldsymbol{\beta} i}\right).$$

Thus, the full conditional distribution of v_{rd}, after marginalizing over $\boldsymbol{\beta}$, is Bernoulli with success probability $p_{v_{rd}}$; i.e.,

$$v_{rd} \mid \boldsymbol{\omega}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, \tau_{v_{rd}} \sim \text{Bernoulli}(p_{v_{rd}}),$$

where $\boldsymbol{v}_{(-rd)}$ is the vector $\boldsymbol{v}$ after removing the rth element of $\boldsymbol{v}_d$ and

$$p_{v_{rd}} = \frac{\pi(\boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, v_{rd} = 1)\tau_{v_{rd}}}{\pi(\boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, v_{rd} = 0)(1 - \tau_{v_{rd}}) + \pi(\boldsymbol{\omega} \mid \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}, \mathbf{R}, \boldsymbol{v}_{(-rd)}, v_{rd} = 1)\tau_{v_{rd}}}.$$

Full conditional of w_{ld}: Under the Dirac spike, w_{ld} should be sampled from its marginal posterior, which is obtained after integrating over λ_{ld}, the ℓth element of $\boldsymbol{\lambda}$; that is, sample from

$$\begin{aligned}\pi(w_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \tau_{w_{ld}}) &\propto \pi(w_{ld}|\tau_{w_{ld}}) \int \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}, \boldsymbol{\omega} \mid \boldsymbol{\Theta})\pi(\lambda_{ld} \mid w_{ld})d\lambda_{ld} \\ &\propto \pi(w_{ld}|\tau_{w_{ld}})\pi(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, w_{ld}),\end{aligned}$$

where $\boldsymbol{\lambda}_{(-\ell)}$ is the vector $\boldsymbol{\lambda}$ with λ_{ld} removed and

$$\pi(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, w_{ld}) \propto \frac{\sigma_{\lambda_{ld}}\{1 - \Phi(-\mu_{\lambda_{ld}}/\sigma_{\lambda_{ld}})\}}{\psi_{ld}/2} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N} \boldsymbol{\omega}^{\star'}_{\lambda_\ell i}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\lambda_\ell i} - \mu^2_{\lambda_{ld}}/\sigma^2_{\lambda_{ld}}\right)\right\}.$$

All notational conventions developed to express the full conditional distribution of $\boldsymbol{\lambda}$ are adopted. When $w_{ld} = 0$, this marginalized likelihood reduces to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{\omega}^{\star'}_{\lambda_\ell i}\mathbf{R}^{-1}\boldsymbol{\omega}^{\star}_{\lambda_\ell i}\right).$$

Thus, the full conditional distribution of w_{ld}, after marginalizing over λ_{ld}, is Bernoulli with mean $p_{w_{ld}}$; i.e.,

$$w_{ld} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, \tau_{w_{ld}} \sim \text{Bernoulli}(p_{w_{ld}}),$$

where

$$p_{w_{ld}} = \frac{\pi(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, w_{ld} = 1)\tau_{w_{ld}}}{\pi(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, w_{ld} = 0)(1 - \tau_{w_{ld}}) + \pi(\boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}_{(-\ell)}, \mathbf{a}, \mathbf{b}, w_{ld} = 1)\tau_{w_{ld}}}.$$

Full conditionals of $S_{e(m):d}$ and $S_{p(m):d}$: Based on the form of $\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} \mid \boldsymbol{\Theta})$ in Section 3 of the manuscript, it is easy to establish the full conditionals

$$\begin{aligned} S_{e(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} &\sim \text{beta}(a^\star_{e(m):d}, b^\star_{e(m):d}) \\ S_{p(m):d} \mid \mathbf{Z}, \widetilde{\mathbf{Y}} &\sim \text{beta}(a^\star_{p(m):d}, b^\star_{p(m):d}), \end{aligned}$$

where

$$\begin{aligned} a^\star_{e(m):d} &= a_{e(m):d} + \sum_{j \in \mathcal{I}_m} Z_{jd}\widetilde{Z}_{jd}, \\ b^\star_{e(m):d} &= b_{e(m):d} + \sum_{j \in \mathcal{I}_m} (1 - Z_{jd})\widetilde{Z}_{jd}, \\ a^\star_{p(m):d} &= a_{p(m):d} + \sum_{j \in \mathcal{I}_m} (1 - Z_{jd})(1 - \widetilde{Z}_{jd}), \\ b^\star_{p(m):d} &= b_{p(m):d} + \sum_{j \in \mathcal{I}_m} Z_{jd}(1 - \widetilde{Z}_{jd}). \end{aligned}$$

Web Appendix B: *Additional simulation results in Section 4.* We performed four additional simulation studies to demonstrate the generality of our methods, to compare with alternative approaches, and to examine in what ways our methods are robust to model violations. In the order described in Section 4 (last paragraph), these studies examine

B.1. Single-stage group testing protocol. This study illustrates how our regression and model selection methods perform for a single-stage group testing protocol where specimens are placed in arrays.

B.2. Comparison with Joyner et al. (2020). This study compares our multivariate modeling approach with the marginal modeling methods in Joyner et al. (2020).

B.3. Homogeneous pooling. This study summarizes estimation and model selection when pools are formed homogeneously in terms of site and individual covariates (within site).

B.4. Robustness to model misspecification. We perform two simulation studies to examine the performance of our methods in the presence of model violations, namely, when (a) the linear predictor in the multivariate probit model is misspecified and (b) the probit link function is misspecified.

We describe each study, present the results, and provide summary discussions. All studies assume $D = 2$ diseases. All references are cited in the manuscript.

B.1: *Single-stage group testing protocol.* This study illustrates the performance of our methods when using a non-adaptive group testing protocol; i.e., a protocol where positive pools are not resolved adaptively. At the request of an anonymous reviewer, we consider single-stage array testing; see Hou et al. (2020). This protocol first assigns individuals to an array and then proceeds to test pools formed by combining individuals who share a common row or column of the array. No further testing is performed regardless of the outcome of the row and column pool tests. Therefore, from an estimation perspective, this protocol presents a more challenging scenario than the two-stage Dorfman algorithm used at the SHL. For the two-stage protocol, additional testing results are available when positive pools are resolved. This is not the case with single-stage protocols.

We randomly assign individuals to 5×5 arrays and consider one stratum for the assay accuracy probabilities; i.e., the testing stratum ($m = 1$) applies to all row and column pools. We set $S_{e(1):d} = 0.95$ and $S_{p(1):d} = 0.98$, for $d = 1, 2$. We simulate the execution of this single-stage protocol to produce 500 group testing data sets analogously to the study in Section 4 of the manuscript. All prior distributions and model fitting specifications are the same as those described in Section 4.

Web Table B.1 provides the average bias and the sample standard deviation of the 500 posterior mean estimates. Also provided are the average estimated posterior probabilities of inclusion for the fixed and random effects in $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, respectively. The results from this study convey the same findings we reached in Section 4 for the two-stage Dorfman protocol. Estimation is accurate and we identify nonzero fixed and random effects in this more challenging situation.

B.2: *Comparison with Joyner et al.* (2020). We seek to benchmark our multivariate modeling methods against the corresponding marginal modeling approach in Joyner et al. (2020), which also adopts a probit link. We simulate the execution of the two-stage Dorfman protocol as described in Section 4 of the manuscript. However, marginal models are used to estimate the relationship between disease statuses and covariates instead. The results from this study are shown in Web Table B.2. Therefore, the reader should compare Table 1 in the manuscript with Web Table B.2 to compare the modeling approaches.

Web Table B.2 provides the same quantities as Table 1 in the manuscript, except for the correlation matrix $\mathbf{R}_{12}$, which cannot be estimated using marginal methods. Overall, the approach in Joyner et al. (2020) does fairly well, but there are clear gains from joint modeling. For example, intercepts for fixed effects and random effects are 2-5 times more variable when estimating with marginal models and suffer from much larger bias. Similarly, estimates for the non-zero covariate effects (both fixed and random) have larger bias and are less precise when modeling marginally. Finally, although marginal models perform satisfactorily in model selection (as judged by the posterior probabilities of inclusion), the selection of nonzero effects and the exclusion of null effects is noisier than when using a joint model.

B.3: *Homogeneous pooling.* An anonymous reviewer has raised a question about pooling individuals homogeneously and whether there would be any benefit of doing so in terms of model estimation and variable selection. Our primary simulation experiment in Section 4 constructs pools randomly, which emulates how pools are formed at the SHL.

To evaluate the potential benefit of forming initial pools homogeneously, we repeated our primary experiment in Section 4 with all of the same prior choices and model fitting specifications. Simulation settings are identical except

1. initial pools (of size 4) are formed within each clinic site.
2. individuals within each pool have identical covariates.

Initial pools testing positively are resolved using Dorfman's two-stage protocol as in the primary experiment.

Web Table B.3 provides the same quantities as Table 1 in the manuscript. The results are nearly identical in the two tables, showing it is not necessarily advantageous to form initial pools within site nor homogeneously in terms of their covariates. Using Dorfman's two-stage protocol is the primary reason the results are so similar. When individuals in positive pools are retested, one ultimately obtains a substantial amount of information about each individual's true disease status. This overrides any benefit that homogeneous pooling might provide initially.

B.4: *Robustness to model misspecification.* Although the multivariate probit model is a common choice for correlated binary data, it is not immune from criticism due to potential misspecification. We therefore assess the impact of misspecifying the model when using group testing data from multiplex assays. Specifically, we misspecify the model in two ways and examine the impact of doing so separately. Both studies below use the same prior choices and simulation configruations as in Section 4 of the manuscript except where noted.

Study 1: Linear predictor misspecification
We first focus on misspecifying the form of the linear predictor $\eta_{id} = \mathbf{x}_{id}'\boldsymbol{\beta}_d + \mathbf{t}_{id}'\boldsymbol{\Lambda}_d\mathbf{A}_d\mathbf{b}_{(i)d}$, $d = 1, 2$. For each individual, we generate the covariate vector

$$\mathbf{x}_i^* = (1, x_1^*, x_2^*, x_3^*, x_4^*, \phi(x_1^* x_3^*))',$$

where $x_1^*, ..., x_4^*$ have the same covariate distributions specified in Section 4 of the manuscript and $\phi(\cdot)$ is the standard normal density. Note that the inclusion of the $\phi(x_1^* x_3^*)$ covariate creates a nonlinear relationship and hence misspecifies the model. In the linear predictor above, we set $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_i^*$ and $\mathbf{t}_{i1} = \mathbf{t}_{i2} = (1, x_1^*, x_2^*, x_3^*, x_4^*)'$. However, when we estimate the multivariate probit model

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \int_{I_{i2}} \phi(\boldsymbol{\omega}|\boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega},$$

we ignore the nonlinear covariate and assess the resulting impact of estimating a misspecified model. In the model above, we set $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where

$$\begin{aligned} \boldsymbol{\beta}_1 &= (-2.0, -0.75, 0.5, 0, 0, \beta_{61})' \\ \boldsymbol{\beta}_2 &= (-2.5, 0, 0, 0.5, -0.25, \beta_{62})', \end{aligned}$$

so that β_{6d}, the regression parameter associated with the nonlinear term for the dth disease, $d = 1, 2$, controls the amount of misspecification. Web Tables B.4 and B.5 give the results when $\beta_{61} = \beta_{62} = 2.5$ (moderate misspecifcation) and $\beta_{61} = \beta_{62} = 5$ (severe misspecifcation), respectively.

Web Tables B.4 and B.5 show that ignoring the nonlinear relationship can negatively impact performance in terms of bias in the fixed effects (most notably the intercepts). However, the variability in the fixed effects estimates is about the same as it is under no misspecification (Table 1, manuscript), and estimation performance of the random effects is also similar. Interestingly, even under severe misspecification, our approach continues to reliably identify the nonzero fixed and random effects.

Study 2: Link misspecification

We now assess the impact of misspecifying the link function in our model. To do this, we simulate the true disease status for the ith individual $\widetilde{\mathbf{Y}}_i$ according to

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \text{logit}^{-1}(\eta_{i1} + \epsilon_1) \text{logit}^{-1}(\eta_{i2} + \epsilon_2) \phi(\boldsymbol{\epsilon} | \mathbf{0}, \mathbf{R}) d\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)'$ and $\text{logit}(\cdot)$ denotes the logistic function. The use of the random vector $\boldsymbol{\epsilon}$ induces correlation between the two disease statuses $\widetilde{Y}_{i1}$ and $\widetilde{Y}_{i2}$, where we assume $\boldsymbol{\epsilon}$ is bivariate normal with mean $\mathbf{0}$ and correlation matrix $\mathbf{R}$. True disease statuses are generated using the (misspecified) model above. Using these statuses, we simulate the execution of the two-stage Dorfman protocol in Section 4, but we estimate the probit model

$$P(\widetilde{\mathbf{Y}}_i = \widetilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{b}_{(i)}, \mathbf{R}) = \int_{I_{i1}} \int_{I_{i2}} \phi(\boldsymbol{\omega} | \boldsymbol{\eta}_i, \mathbf{R}) d\boldsymbol{\omega}$$

with the resulting group testing data instead. The results from this study are shown in Web Table B.6.

Web Table B.6 reveals the same findings as Web Tables B.4 and B.5, namely, nonzero fixed and random effects can be biased, but our approach continues to perform well in terms of identifying these effects as being important (as judged by posterior probabilities of inclusion). An interesting finding in Web Table B.6 is that the bias is consistently close to $-1/2$ times the fixed or random effect. For example, the intercept parameter for disease 1 is $\beta_{11} = -2$, and the average bias is 1.00. We believe this occurs because there is a mathematical relationship between our multivariate probit model and the misspecified logistic-type model above. Such relationships have been previously documented in the literature for binary regression; the one we have identified is applicable for our more complex modeling problem with multivariate group testing data.

Table B.1: Simulation results from one-stage array testing. Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$, the assay accuracy probabilities, and the correlation matrix element $\mathbf{R}_{12}$ are also shown.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	-0.01	0.18	1.00	$\beta_{12} = -2.5$	-0.01	0.18	1.00
$\beta_{21} = -0.75$	0.01	0.15	0.99	$\beta_{22} = 0$	0.00	0.04	0.03
$\beta_{31} = 0.5$	0.01	0.08	1.00	$\beta_{32} = 0$	0.00	0.02	0.02
$\beta_{41} = 0$	0.00	<0.01	0.01	$\beta_{42} = 0.5$	0.00	0.04	1.00
$\beta_{51} = 0$	0.00	<0.01	0.01	$\beta_{52} = -0.25$	0.00	0.04	1.00
$\lambda_{11} = 1$	0.04	0.16	1.00	$\lambda_{12} = 1$	0.06	0.16	1.00
$\lambda_{21} = 0.75$	0.02	0.09	1.00	$\lambda_{22} = 0.75$	0.02	0.10	1.00
$\lambda_{31} = 0.25$	-0.02	0.09	0.91	$\lambda_{32} = 0.25$	-0.01	0.07	0.95
$\lambda_{41} = 0$	0.00	<0.01	0.01	$\lambda_{42} = 0$	0.00	<0.01	0.01
$\lambda_{51} = 0$	0.00	<0.01	0.01	$\lambda_{52} = 0$	0.00	<0.01	0.01
$a_{211} = 0.5$	-0.02	0.20	–	$a_{212} = 0.5$	-0.02	0.20	–
$a_{311} = 0.2$	0.02	0.28	–	$a_{312} = 0.2$	-0.02	0.24	–
$a_{321} = 0.5$	-0.03	0.28	–	$a_{322} = 0.5$	-0.01	0.26	–
$a_{411} = 0.1$	-0.10	0.02	–	$a_{412} = 0.1$	-0.10	0.03	–
$a_{511} = 0.0$	0.00	0.02	–	$a_{512} = 0.0$	0.00	0.02	–
$a_{421} = 0.2$	-0.20	0.02	–	$a_{422} = 0.2$	-0.20	0.03	–
$a_{521} = 0.1$	-0.10	0.02	–	$a_{522} = 0.1$	-0.10	0.02	–
$a_{431} = 0.5$	-0.50	0.02	–	$a_{432} = 0.5$	-0.50	0.02	–
$a_{531} = 0.2$	-0.20	0.02	–	$a_{532} = 0.2$	-0.20	0.02	–
$a_{541} = 0.5$	-0.50	0.02	–	$a_{542} = 0.5$	-0.50	0.02	–
$S_{e(1):1} = 0.95$	0.00	0.01	–	$S_{e(1):2} = 0.95$	0.00	0.01	–
$S_{p(1):1} = 0.98$	0.00	0.01	–	$S_{p(1):2} = 0.98$	0.00	0.01	–
$\mathbf{R}_{12} = 0.6$	-0.41	0.05					

Table B.2: Simulation results from marginal modeling in Joyner et al. (2020). Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$ and the assay accuracy probabilities are shown. The correlation matrix element $\mathbf{R}_{12}$ cannot be estimated using a marginal approach.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	-0.11	0.49	1.00	$\beta_{12} = -2.5$	-0.12	0.71	0.99
$\beta_{21} = -0.75$	-0.02	0.27	0.98	$\beta_{22} = 0$	-0.01	0.05	0.08
$\beta_{31} = 0.5$	0.07	0.13	1.00	$\beta_{32} = 0$	0.00	0.01	0.06
$\beta_{41} = 0$	0.00	<0.01	0.01	$\beta_{42} = 0.5$	0.03	0.11	1.00
$\beta_{51} = 0$	0.00	<0.01	0.02	$\beta_{52} = -0.25$	-0.03	0.07	1.00
$\lambda_{11} = 1$	0.12	0.32	1.00	$\lambda_{12} = 1$	0.17	0.42	1.00
$\lambda_{21} = 0.75$	0.04	0.16	1.00	$\lambda_{22} = 0.75$	0.04	0.23	1.00
$\lambda_{31} = 0.25$	-0.02	0.09	0.89	$\lambda_{32} = 0.25$	-0.03	0.10	0.87
$\lambda_{41} = 0$	0.00	<0.01	0.02	$\lambda_{42} = 0$	0.00	0.01	0.03
$\lambda_{51} = 0$	0.00	0.01	0.02	$\lambda_{52} = 0$	0.00	0.01	0.03
$a_{211} = 0.5$	0.03	0.16	–	$a_{212} = 0.5$	0.07	0.30	–
$a_{311} = 0.2$	-0.08	0.19	–	$a_{312} = 0.2$	-0.06	0.24	–
$a_{321} = 0.5$	-0.06	0.20	–	$a_{322} = 0.5$	0.00	0.30	–
$a_{411} = 0.1$	-0.10	0.02	–	$a_{412} = 0.1$	-0.10	0.01	–
$a_{511} = 0.0$	0.00	0.01	–	$a_{512} = 0.0$	0.01	0.05	–
$a_{421} = 0.2$	-0.20	0.02	–	$a_{422} = 0.2$	-0.20	0.02	–
$a_{521} = 0.1$	-0.10	0.02	–	$a_{522} = 0.1$	-0.10	0.05	–
$a_{431} = 0.5$	-0.50	0.01	–	$a_{432} = 0.5$	-0.50	0.02	–
$a_{531} = 0.2$	-0.20	0.02	–	$a_{532} = 0.2$	-0.20	0.02	–
$a_{541} = 0.5$	-0.50	0.01	–	$a_{542} = 0.5$	-0.50	0.01	–
$S_{e(1):1} = 0.95$	-0.01	0.02	–	$S_{e(1):2} = 0.95$	-0.01	0.01	–
$S_{e(2):1} = 0.98$	-0.01	0.01	–	$S_{e(2):2} = 0.98$	-0.01	0.01	–
$S_{p(1):1} = 0.98$	0.00	0.01	–	$S_{p(1):2} = 0.98$	0.00	<0.01	–
$S_{p(2):1} = 0.99$	0.00	<0.01	–	$S_{p(2):2} = 0.99$	0.00	<0.01	–

Table B.3: Simulation results with homogeneous pooling. Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$, the assay accuracy probabilities, and the correlation matrix element $\mathbf{R}_{12}$ are also shown.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	-0.02	0.17	1.00	$\beta_{12} = -2.5$	0.00	0.17	1.00
$\beta_{21} = -0.75$	-0.02	0.14	1.00	$\beta_{22} = 0$	0.00	0.02	0.02
$\beta_{31} = 0.5$	0.00	0.06	1.00	$\beta_{32} = 0$	0.00	<0.01	0.01
$\beta_{41} = 0$	0.00	<0.01	0.01	$\beta_{42} = 0.5$	0.00	0.03	1.00
$\beta_{51} = 0$	0.00	<0.01	0.01	$\beta_{52} = -0.25$	0.00	0.03	1.00
$\lambda_{11} = 1$	0.05	0.14	1.00	$\lambda_{12} = 1$	0.06	0.15	1.00
$\lambda_{21} = 0.75$	0.02	0.09	1.00	$\lambda_{22} = 0.75$	0.02	0.09	1.00
$\lambda_{31} = 0.25$	0.00	0.05	0.99	$\lambda_{32} = 0.25$	0.00	0.05	0.99
$\lambda_{41} = 0$	0.00	<0.01	0.01	$\lambda_{42} = 0$	0.00	0.01	0.01
$\lambda_{51} = 0$	0.00	<0.01	0.01	$\lambda_{52} = 0$	0.00	<0.01	0.01
$a_{211} = 0.5$	-0.02	0.17	–	$a_{212} = 0.5$	-0.03	0.20	–
$a_{311} = 0.2$	-0.02	0.24	–	$a_{312} = 0.2$	-0.04	0.23	–
$a_{321} = 0.5$	0.01	0.23	–	$a_{322} = 0.5$	0.00	0.23	–
$a_{411} = 0.1$	-0.10	0.02	–	$a_{412} = 0.1$	-0.10	0.03	–
$a_{511} = 0.0$	0.00	0.03	–	$a_{512} = 0.0$	0.00	0.03	–
$a_{421} = 0.2$	-0.20	0.02	–	$a_{422} = 0.2$	-0.20	0.03	–
$a_{521} = 0.1$	-0.10	0.02	–	$a_{522} = 0.1$	-0.10	0.02	–
$a_{431} = 0.5$	-0.50	0.02	–	$a_{432} = 0.5$	-0.50	0.03	–
$a_{531} = 0.2$	-0.20	0.02	–	$a_{532} = 0.2$	-0.20	0.03	–
$a_{541} = 0.5$	-0.50	0.02	–	$a_{542} = 0.5$	-0.50	0.02	–
$S_{e(1):1} = 0.95$	0.00	0.01	–	$S_{e(1):2} = 0.95$	0.00	<0.01	–
$S_{e(2):1} = 0.98$	0.00	0.01	–	$S_{e(2):2} = 0.98$	0.00	<0.01	–
$S_{p(1):1} = 0.98$	-0.01	0.02	–	$S_{p(1):2} = 0.98$	0.00	<0.01	–
$S_{p(2):1} = 0.99$	0.00	0.01	–	$S_{p(2):2} = 0.99$	0.00	<0.01	–
$\mathbf{R}_{12} = 0.6$	-0.19	0.05					

Table B.4: Robustness study with moderate misspecification ($\beta_{6d} = 2.5$, $d = 1, 2$) of the linear predictor. Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$ the assay accuracy probabilities, and the correlation matrix element $\mathbf{R}_{12}$ are also shown.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	0.26	0.15	1.00	$\beta_{12} = -2.5$	0.24	0.15	1.00
$\beta_{21} = -0.75$	0.07	0.14	0.99	$\beta_{22} = 0$	0.00	0.03	0.03
$\beta_{31} = 0.5$	0.00	0.05	1.00	$\beta_{32} = 0$	0.00	0.01	0.02
$\beta_{41} = 0$	0.00	<0.01	0.01	$\beta_{42} = 0.5$	−0.04	0.03	1.00
$\beta_{51} = 0$	0.00	<0.01	0.01	$\beta_{52} = -0.25$	0.00	0.03	1.00
$\lambda_{11} = 1$	0.02	0.12	1.00	$\lambda_{12} = 1$	0.06	0.14	1.00
$\lambda_{21} = 0.75$	−0.02	0.09	1.00	$\lambda_{22} = 0.75$	−0.05	0.09	1.00
$\lambda_{31} = 0.25$	0.00	0.05	1.00	$\lambda_{32} = 0.25$	0.00	0.05	0.99
$\lambda_{41} = 0$	0.00	<0.01	<0.01	$\lambda_{42} = 0$	0.00	0.01	0.01
$\lambda_{51} = 0$	0.00	<0.01	0.01	$\lambda_{52} = 0$	0.00	<0.01	0.01
$a_{211} = 0.5$	−0.06	0.18	–	$a_{212} = 0.5$	−0.02	0.20	–
$a_{311} = 0.2$	−0.02	0.25	–	$a_{312} = 0.2$	−0.02	0.24	–
$a_{321} = 0.5$	−0.02	0.23	–	$a_{322} = 0.5$	−0.01	0.24	–
$a_{411} = 0.1$	−0.10	0.02	–	$a_{412} = 0.1$	−0.10	0.07	–
$a_{511} = 0.0$	0.00	0.02	–	$a_{512} = 0.0$	0.00	0.02	–
$a_{421} = 0.2$	−0.20	0.02	–	$a_{422} = 0.2$	−0.20	0.02	–
$a_{521} = 0.1$	−0.10	0.02	–	$a_{522} = 0.1$	−0.10	0.02	–
$a_{431} = 0.5$	−0.50	0.02	–	$a_{432} = 0.5$	−0.50	0.02	–
$a_{531} = 0.2$	−0.20	0.02	–	$a_{532} = 0.2$	−0.20	0.02	–
$a_{541} = 0.5$	−0.50	0.02	–	$a_{542} = 0.5$	−0.50	0.02	–
$S_{e(1):1} = 0.95$	0.00	0.01	–	$S_{e(1):2} = 0.95$	−0.01	0.01	–
$S_{e(2):1} = 0.98$	0.00	0.01	–	$S_{e(2):2} = 0.98$	0.00	0.01	–
$S_{p(1):1} = 0.98$	0.00	0.01	–	$S_{p(1):2} = 0.98$	0.00	<0.01	–
$S_{p(2):1} = 0.99$	0.00	<0.01	–	$S_{p(2):2} = 0.99$	0.00	<0.01	–
$\mathbf{R}_{12} = 0.6$	−0.17	0.04					

Table B.5: Robustness study with severe misspecification ($\beta_{6d} = 5$, $d = 1, 2$) of the linear predictor. Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$ the assay accuracy probabilities, and the correlation matrix element $\mathbf{R}_{12}$ are also shown.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	0.52	0.15	1.00	$\beta_{12} = -2.5$	0.50	0.15	1.00
$\beta_{21} = -0.75$	0.12	0.13	0.98	$\beta_{22} = 0$	0.00	0.03	0.03
$\beta_{31} = 0.5$	−0.01	0.05	1.00	$\beta_{32} = 0$	0.00	<0.01	0.01
$\beta_{41} = 0$	0.00	<0.01	<0.01	$\beta_{42} = 0.5$	−0.08	0.03	1.00
$\beta_{51} = 0$	0.00	<0.01	0.01	$\beta_{52} = -0.25$	0.01	0.02	1.00
$\lambda_{11} = 1$	−0.03	0.12	1.00	$\lambda_{12} = 1$	0.02	0.13	1.00
$\lambda_{21} = 0.75$	−0.05	0.08	1.00	$\lambda_{22} = 0.75$	−0.10	0.08	1.00
$\lambda_{31} = 0.25$	−0.01	0.04	1.00	$\lambda_{32} = 0.25$	−0.01	0.04	1.00
$\lambda_{41} = 0$	0.00	<0.01	<0.01	$\lambda_{42} = 0$	0.00	<0.01	0.01
$\lambda_{51} = 0$	0.00	<0.01	0.01	$\lambda_{52} = 0$	0.00	<0.01	0.01
$a_{211} = 0.5$	−0.11	0.18	–	$a_{212} = 0.5$	0.00	0.19	–
$a_{311} = 0.2$	−0.03	0.22	–	$a_{312} = 0.2$	−0.04	0.22	–
$a_{321} = 0.5$	−0.01	0.22	–	$a_{322} = 0.5$	0.01	0.23	–
$a_{411} = 0.1$	−0.10	0.02	–	$a_{412} = 0.1$	−0.09	0.07	–
$a_{511} = 0.0$	0.00	0.02	–	$a_{512} = 0.0$	0.00	0.02	–
$a_{421} = 0.2$	−0.20	0.02	–	$a_{422} = 0.2$	−0.20	0.04	–
$a_{521} = 0.1$	−0.10	0.02	–	$a_{522} = 0.1$	−0.10	0.02	–
$a_{431} = 0.5$	−0.50	0.02	–	$a_{432} = 0.5$	−0.50	0.03	–
$a_{531} = 0.2$	−0.20	0.02	–	$a_{532} = 0.2$	−0.20	0.03	–
$a_{541} = 0.5$	−0.50	0.02	–	$a_{542} = 0.5$	−0.50	0.02	–
$S_{e(1):1} = 0.95$	0.00	0.01	–	$S_{e(1):2} = 0.95$	0.00	0.01	–
$S_{e(2):1} = 0.98$	0.00	0.01	–	$S_{e(2):2} = 0.98$	0.00	0.01	–
$S_{p(1):1} = 0.98$	0.00	0.01	–	$S_{p(1):2} = 0.98$	0.00	<0.01	–
$S_{p(2):1} = 0.99$	0.00	0.00	–	$S_{p(2):2} = 0.99$	0.00	<0.01	–
$\mathbf{R}_{12} = 0.6$	−0.14	0.03					

Table B.6: Robustness study with misspecified link function. Average bias (Bias) of the posterior mean estimates, sample standard deviation (SSD) of the estimates, and average estimated posterior probability of inclusion (PI) for the associated fixed and random effects. Averaged posterior mean estimates of the elements of $\mathbf{a}_d$, $d = 1, 2$ the assay accuracy probabilities, and the correlation matrix element $\mathbf{R}_{12}$ are also shown.

Disease 1				Disease 2			
Parameter	Bias	SSD	PI	Parameter	Bias	SSD	PI
$\beta_{11} = -2$	1.00	0.07	1.00	$\beta_{12} = -2.5$	1.27	0.06	1.00
$\beta_{21} = -0.75$	0.39	0.07	0.98	$\beta_{22} = 0$	0.00	<0.01	0.01
$\beta_{31} = 0.5$	−0.25	0.03	1.00	$\beta_{32} = 0$	0.00	<0.01	0.01
$\beta_{41} = 0$	0.00	<0.01	0.00	$\beta_{42} = 0.5$	−0.26	0.02	1.00
$\beta_{51} = 0$	0.00	<0.01	0.00	$\beta_{52} = -0.25$	0.13	0.02	1.00
$\lambda_{11} = 1$	−0.50	0.06	1.00	$\lambda_{12} = 1$	−0.51	0.06	1.00
$\lambda_{21} = 0.75$	−0.37	0.04	1.00	$\lambda_{22} = 0.75$	−0.38	0.05	1.00
$\lambda_{31} = 0.25$	−0.14	0.04	0.91	$\lambda_{32} = 0.25$	−0.15	0.04	0.86
$\lambda_{41} = 0$	0.00	<0.01	0.01	$\lambda_{42} = 0$	0.00	<0.01	0.01
$\lambda_{51} = 0$	0.00	<0.01	0.01	$\lambda_{52} = 0$	0.00	<0.01	0.01
$a_{211} = 0.5$	−0.04	0.17	–	$a_{212} = 0.5$	−0.01	0.18	–
$a_{311} = 0.2$	0.03	0.24	–	$a_{312} = 0.2$	−0.02	0.24	–
$a_{321} = 0.5$	−0.04	0.26	–	$a_{322} = 0.5$	−0.03	0.28	–
$a_{411} = 0.1$	−0.10	0.02	–	$a_{412} = 0.1$	−0.10	0.02	–
$a_{511} = 0.0$	0.00	0.02	–	$a_{512} = 0.0$	0.00	0.03	–
$a_{421} = 0.2$	−0.20	0.02	–	$a_{422} = 0.2$	−0.20	0.02	–
$a_{521} = 0.1$	−0.10	0.02	–	$a_{522} = 0.1$	−0.10	0.02	–
$a_{431} = 0.5$	−0.50	0.02	–	$a_{432} = 0.5$	−0.50	0.02	–
$a_{531} = 0.2$	−0.20	0.02	–	$a_{532} = 0.2$	−0.20	0.02	–
$a_{541} = 0.5$	−0.50	0.02	–	$a_{542} = 0.5$	−0.50	0.02	–
$S_{e(1):1} = 0.95$	0.00	0.01	–	$S_{e(1):2} = 0.95$	0.00	0.01	–
$S_{e(2):1} = 0.98$	0.00	0.01	–	$S_{e(2):2} = 0.98$	0.00	<0.01	–
$S_{p(1):1} = 0.98$	0.00	0.01	–	$S_{p(1):2} = 0.98$	0.00	0.01	–
$S_{p(2):1} = 0.99$	0.00	0.01	–	$S_{p(2):2} = 0.99$	0.00	<0.01	–
$\mathbf{R}_{12} = 0.6$	−0.48	0.02					

Web Appendix C: *Additional details for Section 5.* We provide additional details on the Iowa data analysis in Section 5 of the manuscript. This includes

C.1. AC2A informative prior construction. We summarize pilot data which were collected on female specimens to validate the performance of the AC2A and construct informative priors for the AC2A accuracy probabilities.

C.2. Sensitivity analysis. We redo our analysis in Section 5 under noninformative priors for the AC2A accuracy probabilities. These were the only parameters that were modeled informatively in Section 5.

C.3. Global goodness-of-fit assessment. We develop a simulation-based strategy to assess overall model fit for the Iowa data. We first describe our procedure and then apply it to the Iowa analysis.

Our R programs, available on GitHub at `https://github.com/mcmaha2/probit_gt`, allow the user to perform sensitivity analyses and goodness-of-fit assessments with group testing data from Dorfman's two-stage protocol.

C.1: *AC2A informative prior construction.* The Aptima Combo 2 Assay (AC2A, Hologic, Inc.) possesses different levels of sensitivity and specificity depending on the specimen type and the disease. Web Table C.1 summarizes pilot data which were collected on female specimens to validate the performance of the AC2A. These data are available from the AC2A product literature (see `www.hologic.com`) and also from Gaydos et al. (2003).

Web Table C.1 combines information from Table 5a (chlamydia, CT) and Table 9a (gonorrhea, NG) in the AC2A product literature. The number of true positives (TP), the number of false negatives (FN), the number of true negatives (TN), and the number of false positives (FP) are shown.

Table C.1: AC2A pilot data.

Disease	Stratum	TP	FN	TN	FP
CT	Swab	195	12	1154	28
	Urine	197	11	1170	13
NG	Swab	126	1	1335	17
	Urine	116	11	1347	10

In Section 5 in the manuscript, we build informative prior distributions for $S_{e(m):d}$ and $S_{p(m):d}$, $m = 1, 2, 3$, $d = 1, 2$, using the pilot data above. Informative prior distributions are

$$\begin{aligned} S_{e(m):d} &\sim \text{beta}(\text{TP} + 1, \text{FN} + 1) \\ S_{p(m):d} &\sim \text{beta}(\text{TN} + 1, \text{FP} + 1). \end{aligned}$$

These can be viewed as the posterior distribution estimates of $S_{e(m):d}$ and $S_{p(m):d}$ that would arise from analyzing the pilot data under uniform priors. For example, for individual swab specimens tested for chlamydia ($m = 1$, $d = 1$), we use $S_{e(1):1} \sim \text{beta}(196, 13)$ and $S_{p(1):1} \sim \text{beta}(1155, 29)$. Other prior distributions are formed similarly.

C.2: *Sensitivity analysis.* All parameters in our data analysis in Section 5 were modeled with diffuse prior distributions except for the AC2A accuracy probabilities (see last page). Recall

- we set $\phi^2_{rd} = \psi^2_{ld} = 100$ in the slab components to provide diffuse prior information
- we used uniform priors for all mixing weights; i.e., $a_v = b_v = a_w = b_w = 1$
- we set $\mathbf{m}_d = \mathbf{0}$, $\mathbf{C}_d = 0.5\mathbf{I}$, $d = 1, 2$, to avoid specifying a strong prior correlation between any two random effects (Chen and Dunson, 2003)
- we set $c_0 = D + 1 = 3$ and $\mathbf{S} = \mathbf{I}$, where $\mathbf{I}$ is a 2×2 identity matrix, to provide a diffuse prior in Equation (4) in the manuscript.

We have redone our Iowa data analysis using uniform$(0, 1)$ priors for $S_{e(m):d}$ and $S_{p(m):d}$, for $d = 1, 2$ and $m = 1, 2, 3$. This allows one to assess the impact of assigning informative priors for the AC2A accuracy probabilities and to see an analysis that injects little or no prior information about any of the model parameters.

The results are shown in Web Table C.2 (for chlamydia) and Web Table C.3 (for gonorrhea). Comparing these tables to Tables 2 and 3 in the manuscript, one notices at most minor changes in the estimates and posterior probabilities of inclusion for the fixed and random effects. The largest difference between the two analyses is seen in the estimates for $S_{e(2):1}$ and $S_{e(2):2}$, the AC2A sensitivities for chlamydia and gonorrhea with urine specimens. These differences are not surprising because all female urine specimens are tested individually (and only once). Therefore, there are no confirmatory or counterfactual test outcomes available to estimate these parameters. Even so, our analysis with uniform priors shows that substantial learning is still possible.

C.3: *Global goodness-of-fit assessment.* In Section 6 of the manuscript (last paragraph), we describe three types of specific model violations that could occur with our approach: a violation of linearity in the fixed and/or random effects, using an incorrect link function, and a violation of normality assumptions for the random effects. A recent set of literature provides a framework on how these assumptions can be checked individually when estimating a Bayesian model such as ours, including Hanson (2006), Jara et al. (2009), and Zhou and Hanson (2018). The general idea is to embed a parametric model within nonparametric extensions and then testing point null hypotheses with the Savage-Dickey ratio (Verdinelli and Wasserman, 1995).

To make this achievable with group testing data, one would first need to develop the methodology needed to accommodate nonparametric extensions, such as the use of Polya trees mixtures (Hanson, 2006) and Bernstein polynomials (Zhou and Hanson, 2018). No methodology currently exists within the group testing literature to estimate models with these nonparametric components. We view this to be a meaningful future research direction−generalizing existing methods to estimate Bayesian nonparametric models with group testing data.

In lieu of testing for specific departures, a global approach to model assessment emerges as an alternative. This type of assessment is also nontrivial with group testing data because the

true individual disease statuses $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, ..., \widetilde{Y}_{iD})'$ are never observed due to the combination of pooling and inherent assay error. A common strategy in assessing a Bayesian model fit is to simulate outcomes from posterior distributions and compare these to the observed outcomes. Unfortunately, in group testing, the quantities being modeled (from which the posteriors arise) are not observed.

What one *does* observe in a group testing protocol are the diagnosed testing outcomes, so we build a global goodness-of-fit strategy around these. These quantities do not appear in the model, but we can simulate them from the posteriors under the assumption that the model is correct. We can then check for agreement between the observed diagnosed statuses from group testing and those simulated under the estimated model. Here is an outline of our procedure:

SIMULATION-BASED GOODNESS-OF-FIT PROCEDURE

1. Draw B_1 realizations of $\boldsymbol{\Theta}$ from the posterior distribution using the sampling algorithm outlined in Section 3 in the manuscript. Based on these realizations, compute posterior mean estimates of all model parameters.

2. Generate B_2 realizations of the true disease statuses in $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1', ..., \widetilde{\mathbf{Y}}_N')'$ based on the model and the posterior estimates of the regression parameters. Denote these samples by $\widetilde{\mathbf{Y}}^{(b_2)}$, for $b_2 = 1, ..., B_2$.

3. For each $\widetilde{\mathbf{Y}}^{(b_2)}$, simulate the testing process based on the group testing protocol of the original data using posterior estimates of the assay accuracy probabilities. Denote the diagnosed testing outcomes from this by $\mathbf{Y}^{(b_2)}$.

4. Compute $\widehat{\mathbf{p}}^{(b_1)} = B_2^{-1} \sum_{b_2=1}^{B_2} \mathbf{Y}^{(b_2)}$. This quantity represents the model-based probability of a positive diagnosis for each individual.

5. Use the model-based probabilities and the observed diagnosed testing outcomes in the data to perform a goodness-of-fit test for each disease separately.

We implement the standard Hosmer-Lemeshow (HL) test in Step 5, although other goodness-of-fit tests or grouping strategies could be used. Here, the term "grouping" refers to how individual diagnosed outcomes and estimated probabilities are stratified to form the HL test statistic. Our code at `https://github.com/mcmaha2/probit_gt` uses R's default number of groupings associated with the HL test. We have applied this procedure to the Iowa group testing data and our analysis in Section 5 using $B_1 = 500$ and $B_2 = 5000$. The output below shows there is insignificant evidence to conclude overall lack of fit.

```
Hosmer and Lemeshow goodness of fit (GOF) test
data:  Y1, p1   # chlamydia
X-squared = 16.323, df = 12, p-value = 0.1769

Hosmer and Lemeshow goodness of fit (GOF) test
data:  Y2, p2   # gonorrhea
X-squared = 16.672, df = 12, p-value = 0.1624
```

Table C.2: Iowa data analysis with uniform priors for the AC2A accuracy probabilities. Fixed and random effects results for chlamydia. The posterior mean estimate, the estimated posterior standard deviation (ESD), and the posterior probability of inclusion (PI) are shown.

Parameter	Description	Estimate	ESD	PI
β_{11}	Intercept	-1.42	0.04	1.00
β_{12}	Age	-0.23	0.02	1.00
β_{13}	Race	-0.04	0.03	0.68
β_{14}	New partner	0.03	0.04	0.39
β_{15}	Multiple partners	0.02	0.03	0.38
β_{16}	Contact with STD	0.15	0.01	1.00
β_{17}	Symptoms	0.00	0.01	0.06
λ_{11}	Intercept	0.17	0.03	1.00
λ_{12}	Age	0.00	0.01	0.02
λ_{13}	Race	0.00	<0.01	<0.01
λ_{14}	New partner	0.05	0.05	0.59
λ_{15}	Multiple partners	0.00	0.01	0.03
λ_{16}	Contact with STD	0.00	<0.01	0.01
λ_{17}	Symptoms	0.00	<0.01	<0.01
$S_{e(1):1}$	Swab individual	0.99	<0.01	–
$S_{e(2):1}$	Urine individual	0.87	0.07	–
$S_{e(3):1}$	Swab pool	0.97	<0.01	–
$S_{p(1):1}$	Swab individual	0.98	<0.01	–
$S_{p(2):1}$	Urine individual	0.99	<0.01	–
$S_{p(3):1}$	Swab pool	0.99	<0.01	–

Table C.3: Iowa data analysis with uniform priors for the AC2A accuracy probabilities. Fixed and random effects results for gonorrhea. The posterior mean estimate, the estimated posterior standard deviation (ESD), and the posterior probability of inclusion (PI) are shown.

Parameter	Description	Estimate	ESD	PI
β_{21}	Intercept	-2.50	0.08	1.00
β_{22}	Age	0.00	<0.01	0.01
β_{23}	Race	-0.06	0.06	0.54
β_{24}	New partner	0.00	<0.01	0.01
β_{25}	Multiple partners	0.00	0.01	0.02
β_{26}	Contact with STD	0.18	0.02	1.00
β_{27}	Symptoms	0.00	0.01	0.01
λ_{21}	Intercept	0.33	0.07	1.00
λ_{22}	Age	<0.01	0.02	0.03
λ_{23}	Race	0.04	0.07	0.31
λ_{24}	New partner	0.00	0.01	0.01
λ_{25}	Multiple partners	0.00	0.01	0.01
λ_{26}	Contact with STD	0.00	0.01	0.01
λ_{27}	Symptoms	0.00	0.02	0.04
$S_{e(1):2}$	Swab individual	0.99	0.01	–
$S_{e(2):2}$	Urine individual	0.87	0.09	–
$S_{e(3):2}$	Swab pool	0.98	0.02	–
$S_{p(1):2}$	Swab individual	1.00	<0.01	–
$S_{p(2):2}$	Urine individual	1.00	<0.01	–
$S_{p(3):2}$	Swab pool	1.00	<0.01	–