Bayesian regression for group testing data

Joshua M. Tebbs

Professor
University of South Carolina
Department of Statistics

September 22, 2023

This work was funded by the National Institutes of Health (R01-AI121351).

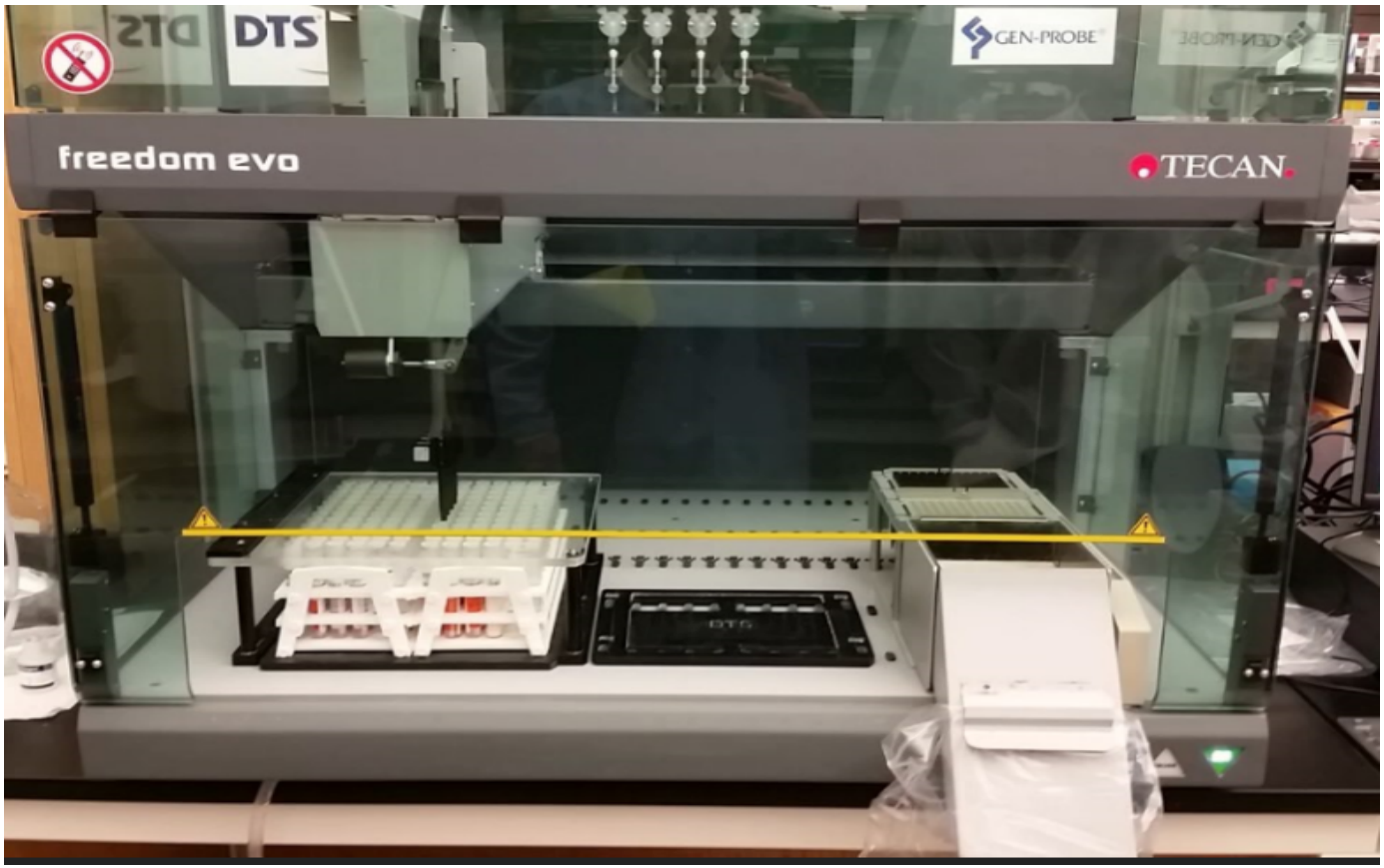
Iowa chlamydia data

- The State Hygienic Laboratory at the University of Iowa tests thousands of Iowa residents each year for chlamydia
- 2014: $N = 13862$ female subjects
  - endocervical swab (about 70 percent)
  - urine
- Swab specimens are combined and tested in pools
  - usually of size $c = 4$
  - positive pools resolved by testing each specimen individually
- Urine specimens are tested individually

Group testing

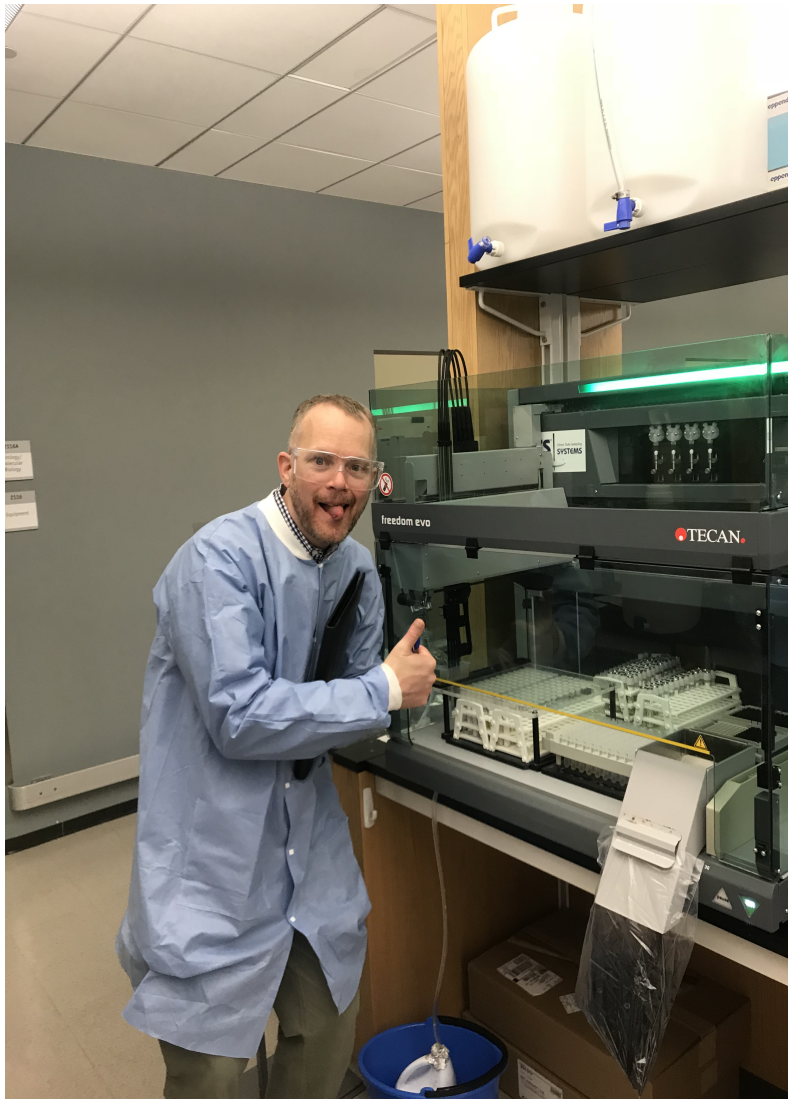
- General premise: tests are performed on "pools" of individual specimens (e.g., swabs, urine, blood, etc.)
  - positive pool: at least one individual in the pool is positive
  - negative pool: all individuals in the pool are negative
- Used to test/screen for a variety of infections
  - syphilis (Dorfman, 1943)
  - HIV, HCV, HBV
  - chlamydia, gonorrhea
  - influenza
- Cost-efficient alternative to testing subjects individually
  - SHL: saves approximately $600,000 each year by pooling specimens
- Case identification versus estimation

Iowa State Hygienic Laboratory



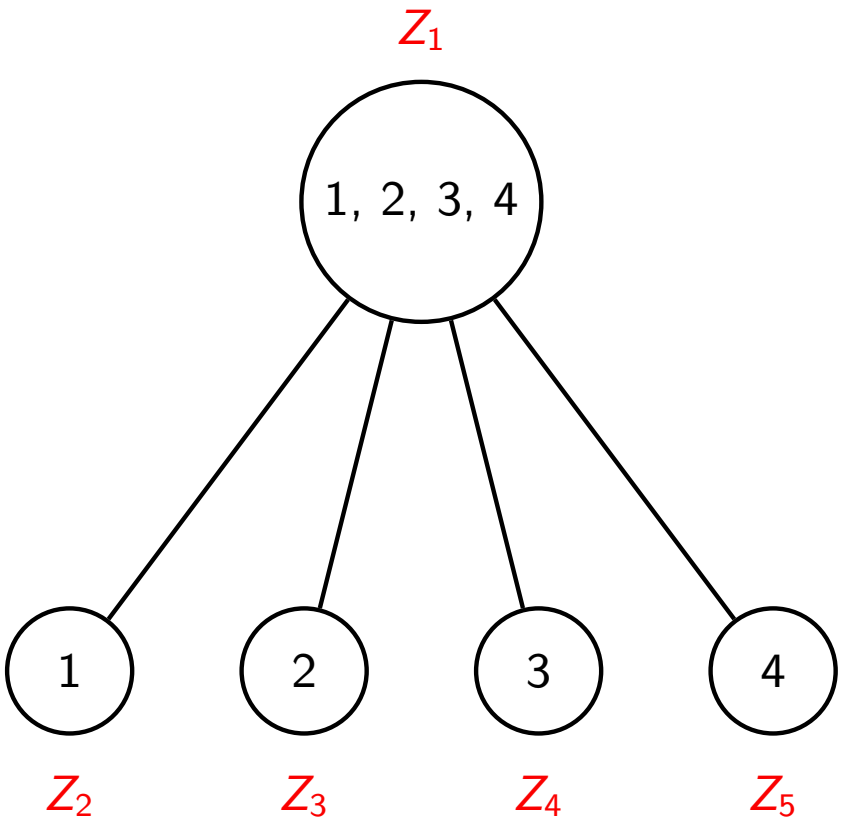
- CT testing
- Tecan pipettes 4 specimens at a time
- Entire process automated

Iowa State Hygienic Laboratory



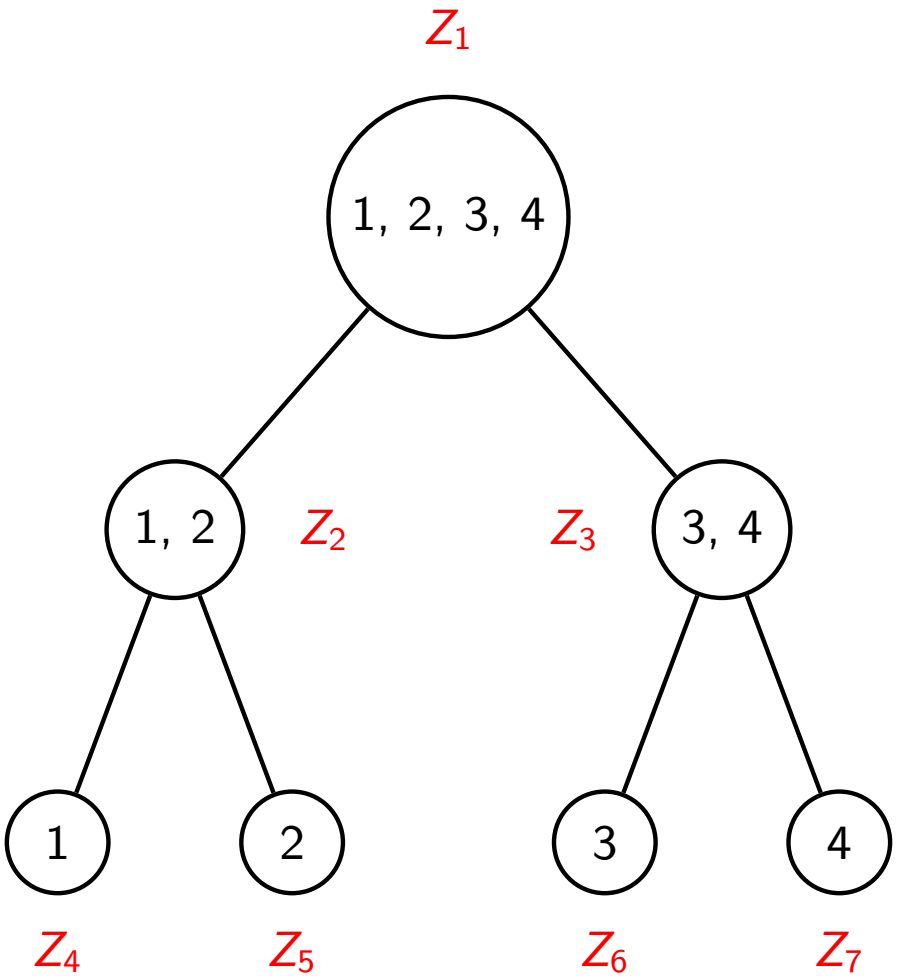
- Great colleagues at SHL!
- Wade Aldous (Associate Lab Director)
- Kris Eveland (Lab Technician)

Dorfman (two-stage hierarchical) testing



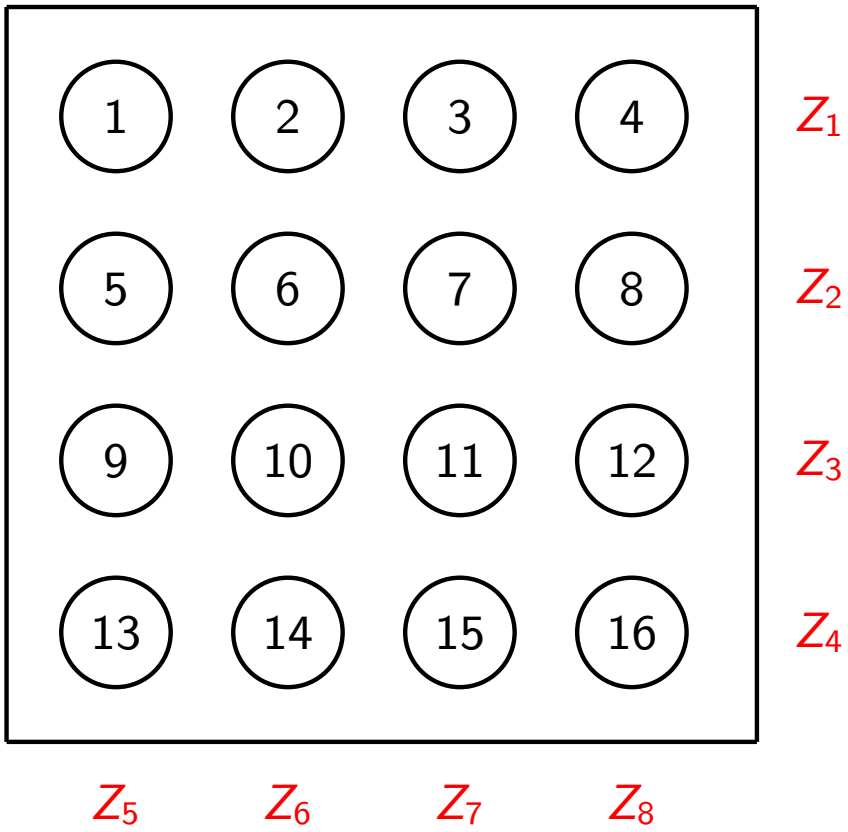
- Master pool tested in first stage
- Individual testing in second stage (if necessary)
- Most common case identification protocol
  - Iowa SHL; other labs

Three-stage hierarchical testing



- Master pool tested in first stage
- Subpools tested in second stage (if necessary)
- Individual testing in third stage (if necessary)
- Why make more complicated?
  - reduces number of tests when prevalence is small

Array testing (in two dimensions)



- Individual specimens placed in the cells of an array
- First stage: Test row and column master pools
  - rows give Z_1, Z_2, Z_3, Z_4
  - columns give Z_5, Z_6, Z_7, Z_8
- Individual retests (if necessary) in second stage give $Z_9, Z_{10}, Z_{11} \ldots,$

Objectives

- Develop a general regression framework to relate an individual's true status $\widetilde{Y}_i$ to covariates in a regression model
  - covariates measured on each individual
  - don't get to observe $\widetilde{Y}_i$, $i = 1, 2, ..., N$
- We get to observe the testing responses $\mathbf{Z} = (Z_1, Z_2, ..., Z_J)'$
  - could arise from master pools, subsets of master pools, and/or individuals
- Also want to estimate assay sensitivity and specificity
  - allow sensitivity and specificity to change with pool size
  - even allow for multiple assays to be used during the screening process

Notation and assumptions

- $\widetilde{Y}_i$ = disease status (1/0); $\mathbf{x}_i$ covariate vector; $i = 1, 2, ..., N$
- $\mathcal{P}_j \subset \{1, 2, ..., N\}$ = set of indices identifying which individuals belong to the jth pool, $j = 1, 2, ..., J$.
  - Example: $\mathcal{P}_1 = \{1, 2, 3, 4\}$, $\mathcal{P}_2 = \{1, 2\}$, $\mathcal{P}_3 = \{3, 4\}$, $\mathcal{P}_4 = \{1\}$, $\mathcal{P}_5 = \{2\}$
- $\widetilde{Z}_j = 1$ if $\mathcal{P}_j$ is truly positive; $Z_j = 1$ if $\mathcal{P}_j$ tests positively
  - $S_{e_j} = \text{pr}(Z_j = 1 | \widetilde{Z}_j = 1)$
  - $S_{p_j} = \text{pr}(Z_j = 0 | \widetilde{Z}_j = 0)$
- GLM: $\text{pr}(\widetilde{Y}_i = 1 | \mathbf{x}_i, \boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$
  - if we had the $\widetilde{Y}_i$'s, we could estimate the model directly
  - we only have the Z_j's (testing responses)

Notation and assumptions

- Assume the $\widetilde{Y}_i$'s are conditionally independent given the covariates
- The conditional distribution of $\mathbf{Z}$ can be written as

$$\pi(\mathbf{Z}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \boldsymbol{\beta}) = \sum_{\widetilde{\mathbf{Y}} \in \{0,1\}^N} \left\{ \prod_{j=1}^{J} \{S_{e_j}^{Z_j}(1-S_{e_j})^{1-Z_j}\}^{\widetilde{Z}_j} \{(1-S_{p_j})^{Z_j} S_{p_j}^{1-Z_j}\}^{1-\widetilde{Z}_j} \right.$$
$$\left. \times \prod_{i=1}^{N} \{g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\}^{\widetilde{Y}_i} \{1-g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\}^{1-\widetilde{Y}_i} \right\}$$

  - Inside the brackets: $\pi(\mathbf{Z}, \widetilde{\mathbf{Y}}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \boldsymbol{\beta})$
  - First product: Contribution of observed testing responses
  - Second product: Contribution of individual (latent) statuses
  - Observed data likelihood of $\boldsymbol{\beta}$ if $\mathbf{S}_e$ and $\mathbf{S}_p$ are known

Bayesian estimation

- Assume $\mathbf{S}_e$ and $\mathbf{S}_p$ are known (relax later)
- $\pi(\boldsymbol{\beta})$ = prior distribution for $\boldsymbol{\beta}$; e.g., $\boldsymbol{\beta} \sim \mathcal{N}_{r+1}(\mathbf{a}, \mathbf{R})$
- $\pi(\widetilde{\mathbf{Y}}, \boldsymbol{\beta} | \mathbf{Z}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}) \propto \underbrace{\pi(\mathbf{Z}, \widetilde{\mathbf{Y}} | \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \boldsymbol{\beta}) \pi(\boldsymbol{\beta})}_{(*)}$

$$(*) \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{a})' \mathbf{R}^{-1} (\boldsymbol{\beta} - \mathbf{a}) + \sum_{i=1}^{N} \widetilde{Y}_i \theta_i - b(\theta_i) \right\}$$
$$\times \prod_{j=1}^{J} \{ S_{e_j}^{Z_j} (1 - S_{e_j})^{1 - Z_j} \}^{\widetilde{Z}_j} \{ (1 - S_{p_j})^{Z_j} S_{p_j}^{1 - Z_j} \}^{1 - \widetilde{Z}_j},$$

where $\theta_i = \log[g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}) / \{1 - g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})\}]$, $b(x) = \log\{1 + \exp(x)\}$

- If the $\widetilde{Y}_i$'s were observed, sampling $\boldsymbol{\beta}$ could be done by using any Bayesian method for binary regression

Bayesian estimation

- Because we are estimating a GLM, we used Gamerman's (1997) MH algorithm because it is easier; can just work with

$$\pi(\boldsymbol{\beta}|\widetilde{\mathbf{Y}}, \mathbf{X}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{a})'\mathbf{R}^{-1}(\boldsymbol{\beta}-\mathbf{a}) + \sum_{i=1}^{N}\widetilde{Y}_i\theta_i - b(\theta_i)\right\}$$

- From $\pi(\widetilde{\mathbf{Y}}, \boldsymbol{\beta}|\mathbf{Z}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X})$, we can work out

$$\widetilde{Y}_i|\mathbf{Z}, \widetilde{\mathbf{Y}}_{-i}, \mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \boldsymbol{\beta} \sim \text{Bernoulli}\{p_{i1}^*/(p_{i0}^* + p_{i1}^*)\}, \text{ where}$$

$$p_{i1}^* = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) \prod_{j \in \mathcal{A}_i} s_{e_j}^{z_j}(1-s_{e_j})^{1-z_j}$$

$$p_{i0}^* = \{1-g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})\} \prod_{j \in \mathcal{A}_i} \{s_{e_j}^{z_j}(1-s_{e_j})^{1-z_j}\}^{I(\sum_{i' \in \mathcal{P}_{ij}} \widetilde{Y}_{i'} > 0)} \{(1-s_{p_j})^{z_j} s_{p_j}^{1-z_j}\}^{I(\sum_{i' \in \mathcal{P}_{ij}} \widetilde{Y}_{i'} = 0)},$$

where the sets $\mathcal{A}_i = \{j : i \in \mathcal{P}_j\}$ and $\mathcal{P}_{ij} = \{i' \in \mathcal{P}_j : i' \neq i\}$

Bayesian estimation

- Key point: All one needs to do is simply keep track of which individuals are in which pools

POSTERIOR SAMPLING ALGORITHM

1. Initialize $\boldsymbol{\beta}^{(0)}$ and $\widetilde{Y}_i^{(0)} = 0$, $i = 1, 2, ..., N$; set $t = 1$
2. Sample $\widetilde{Y}_i^{(t)} \sim \text{Bernoulli}\{p_{i1}^*/(p_{i0}^* + p_{i1}^*)\}$, $i = 1, 2, ..., N$
3. Sample $\boldsymbol{\beta}^{(t)}$ from $\pi(\boldsymbol{\beta}|\widetilde{\mathbf{Y}}^{(t)}, \mathbf{X})$
4. Set $t = t + 1$; repeat steps 2 and 3

- This posterior sampling algorithm is extremely fast
  - all unknown quantities are updated using standard distributions
  - invariant to group testing protocol

Unknown assay accuracy probabilities

- Allow $\mathbf{S}_e$ and $\mathbf{S}_p$ to be unknown
- Previous regression methods largely assume $S_{e_j} = S_e$ and $S_{p_j} = S_p$, $j = 1, 2, ..., J$
- S_e and S_p are usually estimated using pilot studies performed by assay manufacturers
  - these estimates are determined from testing individuals—not pools
- Goal: We want a flexible framework:
  - allow sensitivity and specificity to change with pool size
  - allow for multiple assays to be used during the testing process
    - e.g., screening tests for pools; confirmatory tests for individuals

Unknown assay accuracy probabilities

- Let $S_{e(l)}$ and $S_{p(l)}$ denote the sensitivity and specificity associated with the lth assay, $l = 1, 2, ..., L$
- Define

$$\mathcal{M}(l) = \{j : \text{the } l\text{th assay was used to test the } j\text{th pool}\}$$

- Introduce independent prior distributions $\pi(S_{e(l)})$ and $\pi(S_{p(l)})$, $l = 1, 2, ..., L$

$$\pi(\widetilde{\mathbf{Y}}, \mathbf{S}_e, \mathbf{S}_p, \beta|\mathbf{Z}, \mathbf{X}) \propto \pi(\mathbf{Z}, \widetilde{\mathbf{Y}}|\mathbf{S}_e, \mathbf{S}_p, \mathbf{X}, \beta)\pi(\beta)\prod_{l=1}^{L}\pi(S_{e(l)})\pi(S_{p(l)})$$

  - $S_{e(l)} \sim \text{beta}(a_{S_{e(l)}}, b_{S_{e(l)}})$
  - $S_{p(l)} \sim \text{beta}(a_{S_{p(l)}}, b_{S_{p(l)}})$
  - $S_{e(l)}|\mathbf{Z}, \widetilde{\mathbf{Y}}$ and $S_{p(l)}|\mathbf{Z}, \widetilde{\mathbf{Y}}$ also beta

- Augment posterior sampling algorithm: Add 1 extra step

Simulation study

- Model: $\text{logit}\{\text{pr}(\widetilde{Y}_i = 1|x_{i1}, x_{i2})\} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
  - $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)' = (-3, 2, -1)'$
  - $x_{i1} \sim \mathcal{N}(0, 1)$ and $x_{i2} \sim \text{Bernoulli}(0.5)$
  - Population prevalence: about 10 percent
- Three group testing procedures: MPT, DT, and AT
- $N = 5000$ individuals (assign to master pools at random)
  - common master pool size $c = 5$
- Three configurations of the assay accuracies
  1. common S_e and S_p for all tests (known)
  2. common S_e and S_p for all tests (unknown)
  3. different S_e and S_p for pools and individuals (unknown)
- $\pi(\boldsymbol{\beta}) \propto 1$ and $S_{e(l)}, S_{p(l)} \sim \text{beta}(1, 1)$ independently

$S_e = 0.95$ and $S_p = 0.98$ (common, known)

Parameter		Individual	MPT	DT	AT
$\beta_0 = -3$	Bias (CP95)	-0.03 (0.95)	-0.04 (0.95)	-0.02 (0.95)	-0.01 (0.95)
	SSD (ESE)	0.13 (0.13)	0.17 (0.17)	0.11 (0.12)	0.11 (0.11)
$\beta_1 = 2$	Bias (CP95)	0.02 (0.94)	0.04 (0.95)	0.02 (0.95)	0.01 (0.95)
	SSD (ESE)	0.10 (0.11)	0.15 (0.15)	0.09 (0.10)	0.09 (0.09)
$\beta_2 = -1$	Bias (CP95)	-0.01 (0.95)	-0.03 (0.96)	-0.01 (0.96)	-0.01 (0.94)
	SSD (ESE)	0.14 (0.14)	0.21 (0.22)	0.13 (0.13)	0.13 (0.13)
Average number of tests		5000	1000	2679 (46%)	2787 (44%)

- Small bias, good agreement between SSD and ESE, and CP95 within MOE
- Interesting: DT and AT give (as good or) better precision than individual testing!
  - occurs despite requiring far fewer tests
  - common theme in group testing: "Get more for less"

Iowa chlamydia data

- $N = 13862$ female specimens during 2014 (swab/urine)
- Data:
  - 2273 swab master pools of size $c = 4$
  - 12 swab master pools of size $c = 3$
  - one swab master pool of size $c = 2$
  - 416 individual swab specimens
  - 4316 individual urine specimens
- Recall: Positive swab master pools resolved using DT
- All testing performed using AC2A
  - pilot data available from product insert; see also Gaydos (2003); can be used to set informative priors

Iowa chlamydia data

- Six covariates measured on each individual:
  1. $x_1 =$ age
  2. $x_2 = 1$ if the individual is Caucasian (0, otherwise)
  3. $x_3 = 1$ if a new sexual partner was reported in the last 90 days
  4. $x_4 = 1$ if multiple partners were reported in the last 90 days
  5. $x_5 = 1$ if the individual had contact with a partner having any STD reported in the previous year
  6. $x_6 = 1$ if the individual showed symptoms of infection
- Population model:

$$\text{logit}\{\text{pr}(\widetilde{Y}_i = 1|\mathbf{x}_i)\} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6},$$

for $i = 1, 2, ..., 13862$

Iowa chlamydia data

- We envision three sets of assay accuracy probabilities:
  1. $S_{e(1)}$ and $S_{p(1)}$ for swab specimens tested in pools
  2. $S_{e(2)}$ and $S_{p(2)}$ for swab specimens tested individually
  3. $S_{e(3)}$ and $S_{p(3)}$ for urine specimens tested individually
- 13 parameters to estimate all together
  - $\pi(\boldsymbol{\beta}) \propto 1$ and $S_{e(l)}, S_{p(l)} \sim \text{beta}(1,1)$, for $l = 1, 2, 3$
  - 40000 posterior draws sampled after a burn-in of 1000 draws
  - fitting the model took about 7 minutes
- **Q:** Use informative priors for $S_{e(2)}$, $S_{p(2)}$, $S_{e(3)}$, and $S_{p(3)}$?
  - **A:** We did and got the same results

Iowa chlamydia data

Parameter	Description	Estimate	ESE	95% CI
β_0		-0.759	0.194	$(-1.126, -0.368)$
β_1	Age	-0.071	0.007	$(-0.085, -0.058)$
β_2	Race	-0.348	0.081	$(-0.505, -0.190)$
β_3	New partner	0.276	0.070	(0.139, 0.414)
β_4	Multiple partners	0.330	0.094	(0.144, 0.513)
β_5	Contact with STD	1.408	0.112	(1.189, 1.628)
β_6	Symptoms	0.290	0.077	(0.138, 0.439)
$S_{e(1)}$	Swab pool	0.891	0.069	(0.742, 0.995)
$S_{e(2)}$	Swab individual	0.998	0.002	(0.993, 1.000)
$S_{e(3)}$	Urine individual	0.836	0.091	(0.646, 0.987)
$S_{p(1)}$	Swab pool	0.999	0.001	(0.997, 1.000)
$S_{p(2)}$	Swab individual	0.978	0.007	(0.964, 0.993)
$S_{p(3)}$	Urine individual	0.989	0.007	(0.974, 0.999)

Discussion

- We have developed a general regression framework for group testing data with individually measured covariates
  - can incorporate historical information
  - estimate assay accuracy probabilities
  - invariant to how the data were collected
- Modeling extensions:
  - random effects + variable selection (*Biometrics*, 2020)
  - generalized additive regression (*Biostatistics*, 2021)
  - multivariate binary response (*Biometrics*, in revision)
  - time-to-event response (*Biometrika*, in revision)
- Can re-estimate model as new testing results arrive
  - useful to implement informative group testing case identification protocols
  - perhaps even to detect misdiagnosed individuals
    - "back-end screening"

Bayesian regression for group testing data

Joshua M. Tebbs

Professor
University of South Carolina
Department of Statistics

September 22, 2023

This work was funded by the National Institutes of Health (R01-AI121351).