Marginal Regression Models for Multiple-Disease Group Testing Data

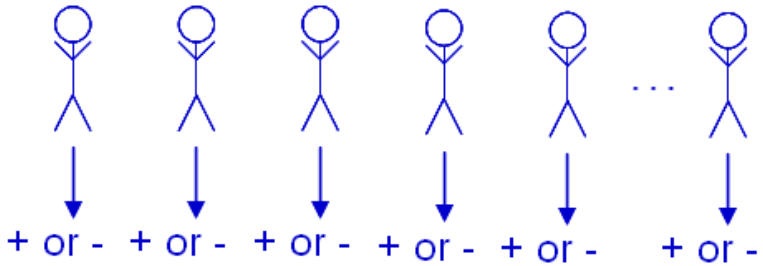
Christopher R. Bilder[1], Boan Zhang[1], and Joshua M. Tebbs[2]

[1]University of Nebraska–Lincoln, Department of Statistics
[2]University of South Carolina, Department of Statistics

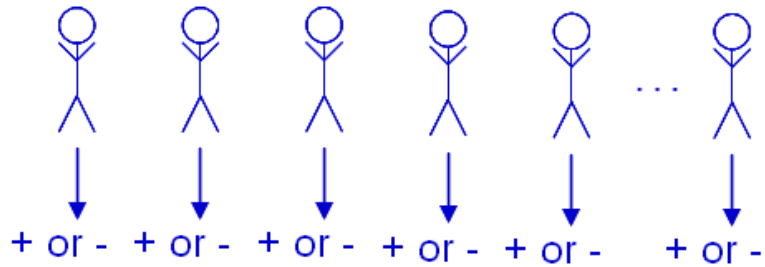
This research is supported in part by NIH grant R01AI067373

April 2, 2012

- Screen a large number of individuals for an infectious disease
- Individual testing

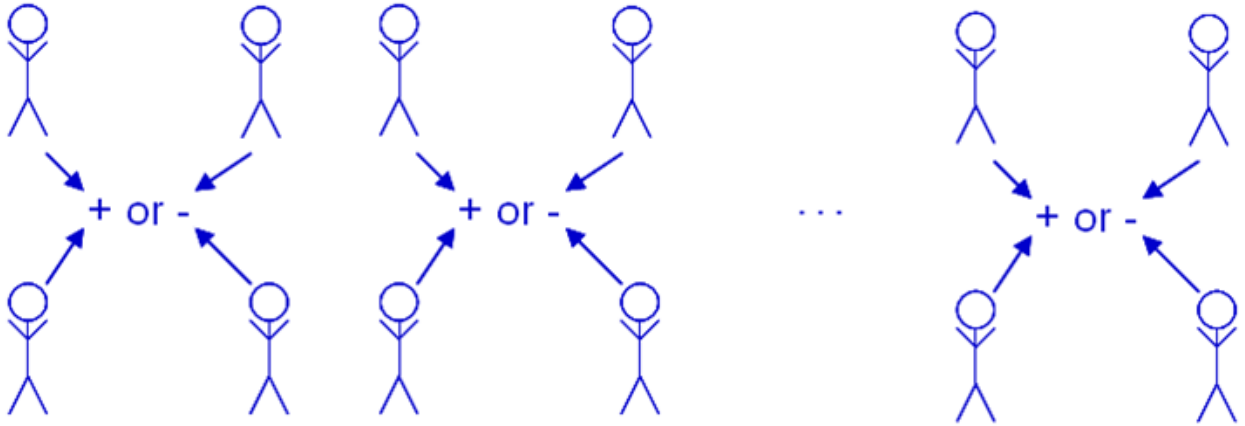


- Screen a large number of individuals for an infectious disease
- Individual testing

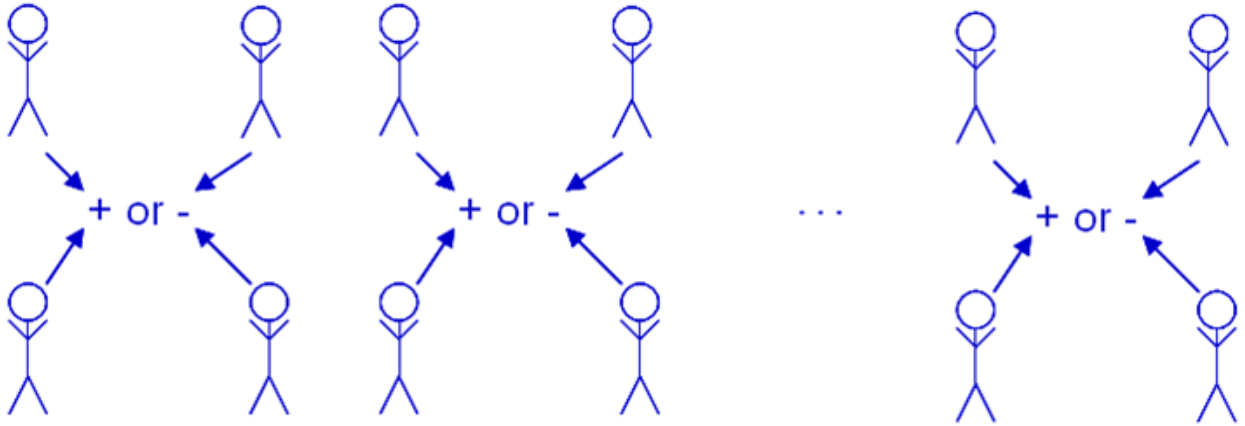


- May not feasible in high volume clinical specimen settings
  - Cost
  - Time

- Group testing (a.k.a., pooled testing)

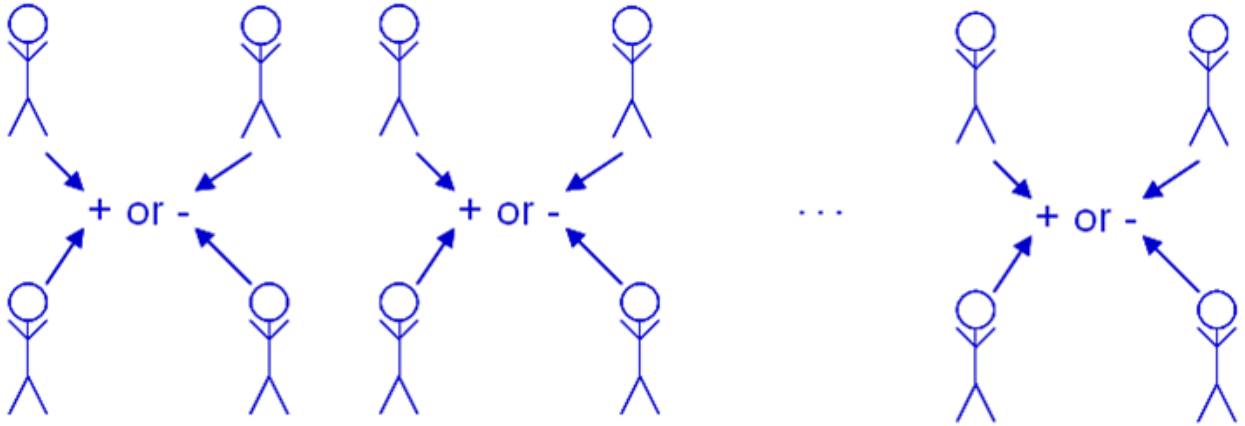


- Group testing (a.k.a., pooled testing)



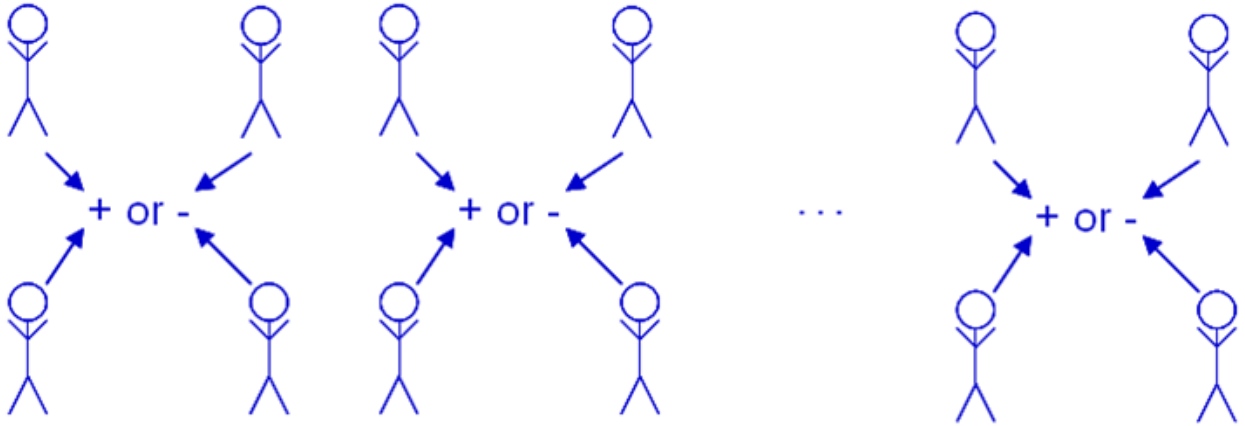
- If the GROUP is negative, then all individuals are declared negative

- Group testing (a.k.a., pooled testing)



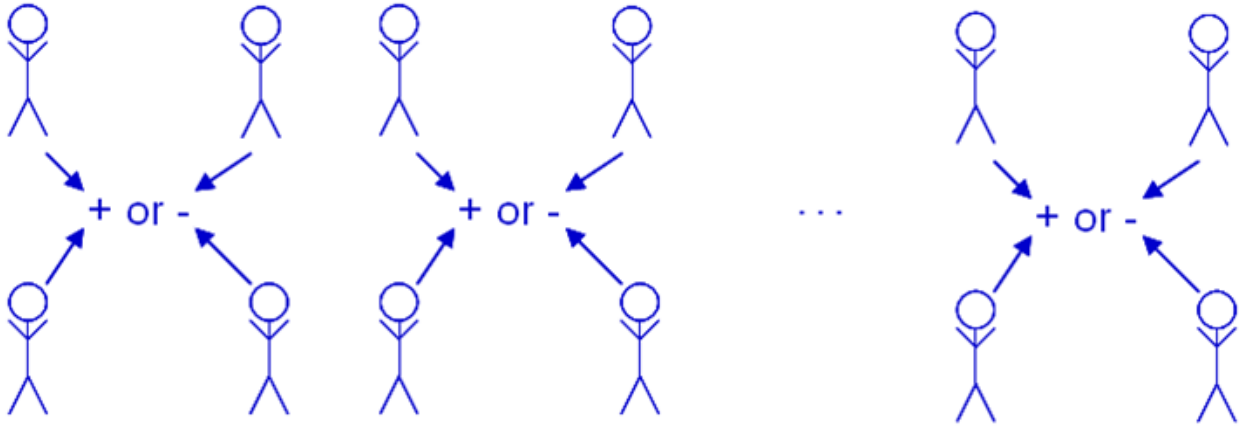
- If the GROUP is negative, then all individuals are declared negative
- If the GROUP is positive, then at least ONE individual is positive

- Group testing (a.k.a., pooled testing)



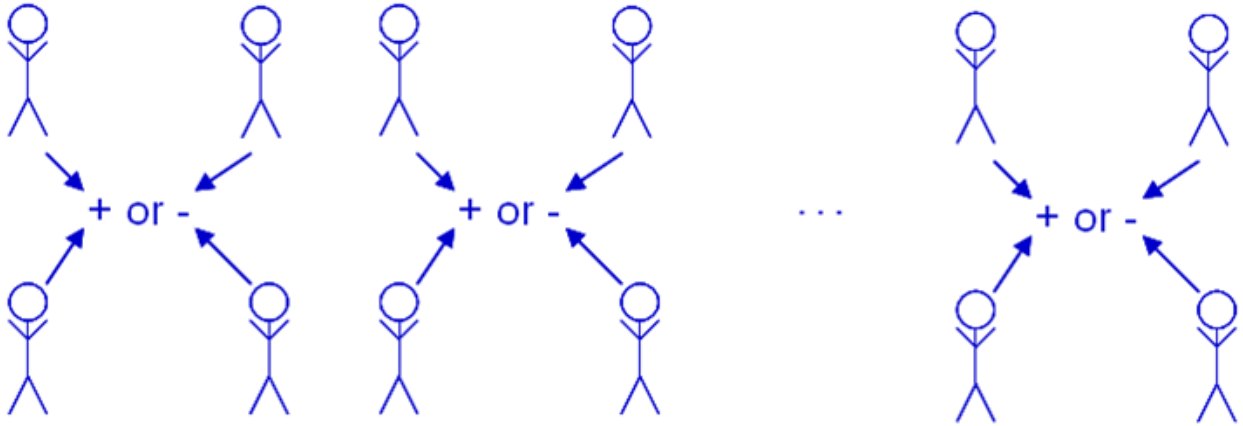
- If the GROUP is negative, then all individuals are declared negative
- If the GROUP is positive, then at least ONE individual is positive
- Benefits:
  - Reduction in tests
  - Cost savings (less assays and labor)

- Group testing (a.k.a., pooled testing)



- If the GROUP is negative, then all individuals are declared negative
- If the GROUP is positive, then at least ONE individual is positive
- Benefits:
  - Reduction in tests
  - Cost savings (less assays and labor)
- Overall disease prevalence needs to be small

- Group testing (a.k.a., pooled testing)



- If the GROUP is negative, then all individuals are declared negative
- If the GROUP is positive, then at least ONE individual is positive
- Benefits:
  - Reduction in tests
  - Cost savings (less assays and labor)
- Overall disease prevalence needs to be small
- Classification vs. estimation

- Test for multiple diseases at the same time

- Test for multiple diseases at the same time
  - American Red Cross (Stramer et al. 2004; ARC 2012)

- Test for multiple diseases at the same time
  - American Red Cross (Stramer et al. 2004; ARC 2012)
    - ≈6 million blood donations per year
    - HIV, hepatitis B, and hepatitis C
    - Groups of size 16

- Test for multiple diseases at the same time
  - American Red Cross (Stramer et al. 2004; ARC 2012)
    - ≈6 million blood donations per year
    - HIV, hepatitis B, and hepatitis C
    - Groups of size 16
- Nebraska Public Health Laboratory
  - Chlamydia and gonorrhea screening
  - ≈25,000 individuals per year
  - GenProbe Aptima Combo 2 assay

- Test for multiple diseases at the same time
  - American Red Cross (Stramer et al. 2004; ARC 2012)
    - ≈6 million blood donations per year
    - HIV, hepatitis B, and hepatitis C
    - Groups of size 16
- Nebraska Public Health Laboratory
  - Chlamydia and gonorrhea screening
  - ≈25,000 individuals per year
  - GenProbe Aptima Combo 2 assay
- Only Hughes-Oliver and Rosenberger (2000) address multiple-disease problem in group testing
  - Estimate overall prevalence
  - Perfect diagnostic tests

- Purpose

- Purpose
  - Develop group testing regression models for multiple-disease screening
  - Probability of positivity dependent on covariates
    - Imperfect diagnostic tests

- Purpose
  - Develop group testing regression models for multiple-disease screening
  - Probability of positivity dependent on covariates
    - Imperfect diagnostic tests
  - Correlated binary data problem
    - Disease responses are unobserved for each individual

- Individual responses
  - $\tilde{Y}_{ijk}$ = true unknown individual status of disease j for the ith individual in kth
    - $i = 1, \ldots, I_k$
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$

- Individual responses
  - $\tilde{Y}_{ijk}$ = true unknown individual status of disease j for the ith individual in kth
    - $i = 1, \ldots, I_k$
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $\tilde{Y}_{ijk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$
  - $\tilde{p}_{ijk} \equiv P(\tilde{Y}_{ijk} = 1 | \mathbf{x}_{ik})$ for covariates $\mathbf{x}_{ik} = (x_{1ik}, \ldots, x_{p-1,ik})'$

- Individual responses
  - $\tilde{Y}_{ijk}$ = true unknown individual status of disease j for the ith individual in kth
    - $i = 1, \ldots, I_k$
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $\tilde{Y}_{ijk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$
  - $\tilde{p}_{ijk} \equiv P(\tilde{Y}_{ijk} = 1 | \mathbf{x}_{ik})$ for covariates $\mathbf{x}_{ik} = (x_{1ik}, \ldots, x_{p-1,ik})'$
- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j} x_{1ik} + \cdots + \beta_{p-1,j} x_{p-1,ik}$

- Individual responses
  - $\tilde{Y}_{ijk}$ = true unknown individual status of disease j for the ith individual in kth
    - $i = 1, \ldots, I_k$
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $\tilde{Y}_{ijk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$
  - $\tilde{p}_{ijk} \equiv P(\tilde{Y}_{ijk} = 1 | \mathbf{x}_{ik})$ for covariates $\mathbf{x}_{ik} = (x_{1ik}, \ldots, x_{p-1,ik})'$
- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j} x_{1ik} + \cdots + \beta_{p-1,j} x_{p-1,ik}$
- Problem: $\tilde{Y}_{ijk}$ are not observed in group testing when no retests are performed

- Group responses
  - Z_{jk} = observable group responses for disease j of group k
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$

- Group responses
  - Z_{jk} = observable group responses for disease j of group k
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$

- Group responses
  - Z_{jk} = observable group responses for disease j of group k
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $Z_{jk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$

- Group responses
  - Z_{jk} = observable group responses for disease j of group k
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $Z_{jk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$
  - $\theta_{jk} \equiv P(Z_{jk} = 1)$

- Group responses
  - Z_{jk} = observable group responses for disease j of group k
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $Z_{jk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$
  - $\theta_{jk} \equiv P(Z_{jk} = 1)$
- Testing error
  - $\tilde{Z}_{jk}$ = true group response
  - $\eta_j = P(Z_{jk} = 1 | \tilde{Z}_{jk} = 1)$
  - $\delta_j = P(Z_{jk} = 0 | \tilde{Z}_{jk} = 0)$

- Group responses
  - Z_{jk} = observable group responses for disease j of group k
    - $j = 1, \ldots, J$
    - $k = 1, \ldots, K$
  - $Z_{jk} = 0\,(1)$ for negative (positive) response
  - Likely correlated across $j = 1, \ldots, J$
  - $\theta_{jk} \equiv P(Z_{jk} = 1)$
- Testing error
  - $\tilde{Z}_{jk}$ = true group response
  - $\eta_j = P(Z_{jk} = 1 | \tilde{Z}_{jk} = 1)$
  - $\delta_j = P(Z_{jk} = 0 | \tilde{Z}_{jk} = 0)$
- Relationship between individual and groups:

$$\begin{aligned} \theta_{jk} &= \eta_j P(\tilde{Z}_{jk} = 1) + (1 - \delta_j) P(\tilde{Z}_{jk} = 0) \\ &= \eta_j + (1 - \delta_j - \eta_j) \prod_{i=1}^{I_k} (1 - \tilde{p}_{ijk}) \end{aligned}$$

- Expectation-Solution algorithm (Elashoff and Ryan, 2004)
  - Generalization of GEE methodology and EM algorithm

- Expectation-Solution algorithm (Elashoff and Ryan, 2004)
  - Generalization of GEE methodology and EM algorithm
- Write generalized estimating equations in terms of $\tilde{Y}_{ijk}$
  - Usual form: THERE IS A VERY LARGE SET OF EQUATIONS HERE THAT I DECIDED NOT TO ENTER :(

- Expectation-Solution algorithm (Elashoff and Ryan, 2004)
  - Generalization of GEE methodology and EM algorithm
- Write generalized estimating equations in terms of $\tilde{Y}_{ijk}$
  - Usual form: THERE IS A VERY LARGE SET OF EQUATIONS HERE THAT I DECIDED NOT TO ENTER :(
  - $\tilde{Y}_{ijk}$ is not observed

- E-step - Replace $\tilde{Y}_{ijk}$ with $E(\tilde{Y}_{ijk}|z_{jk})$

- E-step - Replace $\tilde{Y}_{ijk}$ with $E(\tilde{Y}_{ijk}|z_{jk})$
  - Closed form expression exists:

$$E(\tilde{Y}_{ijk}|z_{jk}=1) = \frac{\eta_j \tilde{p}_{ijk}}{\theta_{jk}}$$

$$E(\tilde{Y}_{ijk}|z_{jk}=0) = \frac{(1-\eta_j)\tilde{p}_{ijk}}{1-\theta_{jk}}$$

- E-step - Replace $\tilde{Y}_{ijk}$ with $E(\tilde{Y}_{ijk}|z_{jk})$
  - Closed form expression exists:

$$E(\tilde{Y}_{ijk}|z_{jk}=1) = \frac{\eta_j \tilde{p}_{ijk}}{\theta_{jk}}$$

$$E(\tilde{Y}_{ijk}|z_{jk}=0) = \frac{(1-\eta_j)\tilde{p}_{ijk}}{1-\theta_{jk}}$$

  - Expectations can be modified for other group testing protocols

- E-step - Replace $\tilde{Y}_{ijk}$ with $E(\tilde{Y}_{ijk}|z_{jk})$
  - Closed form expression exists:

$$\begin{aligned} E(\tilde{Y}_{ijk}|z_{jk}=1) &= \frac{\eta_j \tilde{p}_{ijk}}{\theta_{jk}} \\ E(\tilde{Y}_{ijk}|z_{jk}=0) &= \frac{(1-\eta_j)\tilde{p}_{ijk}}{1-\theta_{jk}} \end{aligned}$$

  - Expectations can be modified for other group testing protocols
- S-step – solve for vector of parameters $\boldsymbol{\beta}$ in

$$\Psi^{obs}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_k \sum_i \mathbf{D}'_{ik} \mathbf{V}'_{ik} (\boldsymbol{\omega}_{ik} - \tilde{\mathbf{p}}_{ik}) = \mathbf{0}$$

where

$$\boldsymbol{\omega}_{ik} = \left(E(\tilde{Y}_{i1k}|z_{1k}), \ldots, E(\tilde{Y}_{iJk}|z_{1k}) \right)'$$

- E-step - Replace $\tilde{Y}_{ijk}$ with $E(\tilde{Y}_{ijk}|z_{jk})$
  - Closed form expression exists:

$$\begin{aligned} E(\tilde{Y}_{ijk}|z_{jk}=1) &= \frac{\eta_j \tilde{p}_{ijk}}{\theta_{jk}} \\ E(\tilde{Y}_{ijk}|z_{jk}=0) &= \frac{(1-\eta_j)\tilde{p}_{ijk}}{1-\theta_{jk}} \end{aligned}$$

  - Expectations can be modified for other group testing protocols
- S-step – solve for vector of parameters $\boldsymbol{\beta}$ in

$$\Psi^{obs}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_k \sum_i \mathbf{D}'_{ik} \mathbf{V}'_{ik} (\boldsymbol{\omega}_{ik} - \tilde{\mathbf{p}}_{ik}) = \mathbf{0}$$

  where

$$\boldsymbol{\omega}_{ik} = \left(E(\tilde{Y}_{i1k}|z_{1k}), \ldots, E(\tilde{Y}_{iJk}|z_{1k}) \right)'$$

- Iterate between E and S-steps until convergence

- Estimating the working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$
  - Want to estimate $Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) = \alpha_{jj'}$, but $\tilde{Y}_{ijk}$ is not observed

- Estimating the working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$
  - Want to estimate $Corr(\tilde{Y}_{ijk}, \tilde{Y}_{ij'k}) = \alpha_{jj'}$, but $\tilde{Y}_{ijk}$ is not observed
  - One can show that

$$
\begin{aligned}
Cov(Z_{jk}, Z_{j'k}) \quad = \quad & (\delta_j + \eta_j - 1)(\delta_{j'} + \eta_{j'} - 1) \times \\
& \left[\prod_{i=1}^{I_k} \left\{ \alpha_{jj'} \sqrt{Var(\tilde{Y}_{ijk}) Var(\tilde{Y}_{ij'k})} + (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right\} - \right. \\
& \left. \prod_{i=1}^{I_k} (1 - \tilde{p}_{ijk})(1 - \tilde{p}_{ij'k}) \right]
\end{aligned}
$$

- Provides way to estimate $\alpha_{jj'}$
  - Use method of moment estimator
  - MORE EQUATIONS HERE. DECIDED NOT TO ENTER THEM.

- Simpler form of estimator for $\alpha_{jj'}$
- MORE EQUATIONS HERE. DECIDED NOT TO ENTER THEM.

- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik}$ where
  - $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
  - $J = 2$
  - x_{1ik} $\sim$uniform(0,1), x_{2ik} $\sim$gamma(17,1.4)

- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik}$ where
  - $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
  - $J = 2$
  - x_{1ik} ~uniform(0,1), x_{2ik} ~gamma(17,1.4)
- Approximate prevalence: 3% prevalence for disease $j = 1$ and 2% for disease $j = 2$

- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik}$ where
  - $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
  - $J = 2$
  - x_{1ik} ~uniform(0,1), x_{2ik} ~gamma(17,1.4)
- Approximate prevalence: 3% prevalence for disease $j = 1$ and 2% for disease $j = 2$
- Simulate correlated binary data by adapting methods in Emrich and Piedmonte (1991)

- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik}$ where
  - $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
  - $J = 2$
  - x_{1ik} ∼uniform(0,1), x_{2ik} ∼gamma(17,1.4)
- Approximate prevalence: 3% prevalence for disease $j = 1$ and 2% for disease $j = 2$
- Simulate correlated binary data by adapting methods in Emrich and Piedmonte (1991)
- $\eta_j = \delta_j = 0.95$

- Model: $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}x_{1ik} + \beta_{2j}x_{2ik}$ where
  - $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
  - $J = 2$
  - x_{1ik} ~uniform(0,1), x_{2ik} ~gamma(17,1.4)
- Approximate prevalence: 3% prevalence for disease $j = 1$ and 2% for disease $j = 2$
- Simulate correlated binary data by adapting methods in Emrich and Piedmonte (1991)
- $\eta_j = \delta_j = 0.95$
- $B = 1000$ simulated data sets

- Mean row:
  $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
- SE/SD row: $B^{-1} \sum_{b=1}^{B} \sqrt{\widehat{Var}(\hat{\beta}_{rj})} / \sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\beta}_{rjb} - \bar{\hat{\beta}}_{rj})^2}$
- Coverage row: 95% confidence level

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
0.6	2000	5	Mean	-6.01	-7.04	0.02	1.05	0.10	0.10

- Mean row:
  $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
- SE/SD row: $B^{-1} \sum_{b=1}^{B} \sqrt{\widehat{Var}(\hat{\beta}_{rj})} / \sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\beta}_{rjb} - \bar{\hat{\beta}}_{rj})^2}$
- Coverage row: 95% confidence level

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
0.6	2000	5	Mean	-6.01	-7.04	0.02	1.05	0.10	0.10
			SE/SD	1.01	0.98	0.97	0.98	1.02	0.97

- Mean row:
  $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
- SE/SD row: $B^{-1} \sum_{b=1}^{B} \sqrt{\widehat{Var}(\hat{\beta}_{rj})} / \sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\beta}_{rjb} - \bar{\hat{\beta}}_{rj})^2}$
- Coverage row: 95% confidence level

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
0.6	2000	5	Mean	-6.01	-7.04	0.02	1.05	0.10	0.10
			SE/SD	1.01	0.98	0.97	0.98	1.02	0.97
			Coverage	0.96	0.94	0.94	0.95	0.96	0.94

- Mean row:
  $\beta_{01} = -6, \beta_{02} = -7, \beta_{11} = 0, \beta_{12} = 1, \beta_{21} = 0.1, \beta_{22} = 0.1$
- SE/SD row: $B^{-1}\sum_{b=1}^{B}\sqrt{\widehat{Var}(\hat{\beta}_{rj})}/\sqrt{(B-1)^{-1}\sum_{b=1}^{B}(\hat{\beta}_{rjb}-\bar{\hat{\beta}}_{rj})^2}$
- Coverage row: 95% confidence level

α	K	I_k	Measure	$\hat{\beta}_{01}$	$\hat{\beta}_{02}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$
0.6	2000	5	Mean	-6.01	-7.04	0.02	1.05	0.10	0.10
			SE/SD	1.01	0.98	0.97	0.98	1.02	0.97
			Coverage	0.96	0.94	0.94	0.95	0.96	0.94
0.6	1000	10	Mean						
			SE/SD						
			Coverage						
0.2	2000	5	Mean						
			SE/SD						
			Coverage						
0.2	1000	10	Mean						
			SE/SD						
			Coverage						

- Chlamydia and gonorrhea screening in Nebraska
  - "Epidemic levels" in Omaha (Zagurski, 2006)

- Chlamydia and gonorrhea screening in Nebraska
  - "Epidemic levels" in Omaha (Zagurski, 2006)
  - Part of the CDC's Infertility Prevention Project

- Chlamydia and gonorrhea screening in Nebraska
  - "Epidemic levels" in Omaha (Zagurski, 2006)
  - Part of the CDC's Infertility Prevention Project
  - Tests
    - ≈25,000 individual tests done per year at the Nebraska Public Health Laboratory (NPHL)
    - GenProbe Aptima Combo 2 assay gives test results for both diseases
    - $11 for a swab test and $16 for a urine test

- Chlamydia and gonorrhea screening in Nebraska
  - "Epidemic levels" in Omaha (Zagurski, 2006)
  - Part of the CDC's Infertility Prevention Project
  - Tests
    - $\approx$25,000 individual tests done per year at the Nebraska Public Health Laboratory (NPHL)
    - GenProbe Aptima Combo 2 assay gives test results for both diseases
    - \$11 for a swab test and \$16 for a urine test
  - NPHL is interested in using group testing
    - Other state labs that use group testing include: Idaho, Iowa, New York, Wisconsin

- Chlamydia and gonorrhea screening in Nebraska
  - "Epidemic levels" in Omaha (Zagurski, 2006)
  - Part of the CDC's Infertility Prevention Project
  - Tests
    - $\approx$25,000 individual tests done per year at the Nebraska Public Health Laboratory (NPHL)
    - GenProbe Aptima Combo 2 assay gives test results for both diseases
    - \$11 for a swab test and \$16 for a urine test
  - NPHL is interested in using group testing
    - Other state labs that use group testing include: Idaho, Iowa, New York, Wisconsin
  - Classification is primary goal

- Estimation
  - Want to understand how risk factors are related to disease status

- Estimation
  - Want to understand how risk factors are related to disease status
  - Focus: 14,530 female swab specimens screened in 2009

- Estimation
  - Want to understand how risk factors are related to disease status
  - Focus: 14,530 female swab specimens screened in 2009
  - Imperfect sensitivity and specificity
    - Gonorrhea: Sensitivity $= \eta_1 = 0.966$, Specificity $= \delta_1 = 0.980$
    - Chlamydia: Sensitivity $= \eta_2 = 0.928$, Specificity $= \delta_2 = 0.960$

- Estimation
  - Want to understand how risk factors are related to disease status
  - Focus: 14,530 female swab specimens screened in 2009
  - Imperfect sensitivity and specificity
    - Gonorrhea: Sensitivity $= \eta_1 = 0.966$, Specificity $= \delta_1 = 0.980$
    - Chlamydia: Sensitivity $= \eta_2 = 0.928$, Specificity $= \delta_2 = 0.960$
  - Artificially form groups of size 5 by testing date
    - $Z_{jk} = 1$ if any positives in group, $Z_{jk} = 0$ otherwise
    - No "best" way to obtain group responses

- Estimation (continued)
  - Covariates
    - Age
    - Race (four levels)
    - Symptoms

- Estimation (continued)
  - Covariates
    - Age
    - Race (four levels)
    - Symptoms
    - Clinical observations: Cervical friability, Pelvic inflammatory disease, Cervicitis

- Estimation (continued)
  - Covariates
    - Age
    - Race (four levels)
    - Symptoms
    - Clinical observations: Cervical friability, Pelvic inflammatory disease, Cervicitis
    - Risk history: Multiple partners, New partner in the last 90 days, Contact with someone who has a STD

- Estimation (continued)
  - Covariates
    - Age
    - Race (four levels)
    - Symptoms
    - Clinical observations: Cervical friability, Pelvic inflammatory disease, Cervicitis
    - Risk history: Multiple partners, New partner in the last 90 days, Contact with someone who has a STD
    - All covariates are binary except for age

- Estimate model with linear terms only
  - $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}\text{Age}_{ik} + \cdots + \beta_{p-1,j}(\text{Contact STD})_{ik}$

- Estimate model with linear terms only
  - $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}\text{Age}_{ik} + \cdots + \beta_{p-1,j}(\text{Contact STD})_{ik}$
  - Using a second-order approximation: $\hat{\alpha} = 0.27$

- Estimate model with linear terms only
  - $\text{logit}(\tilde{p}_{ijk}) = \beta_{0j} + \beta_{1j}\text{Age}_{ik} + \cdots + \beta_{p-1,j}(\text{Contact STD})_{ik}$
  - Using a second-order approximation: $\hat{\alpha} = 0.27$
  - Fit model to individual responses for comparison
    - Use standard GEE methodology with no testing error
    - Non-convergence when incorporating testing error via methods in Neuhaus (2002)

- OMITTED BIG TABLE

- Group testing regression models for multiple-disease data
  - Model data in the form as it arises in application
  - binGroup package (Bilder et al., *R Journal*, 2010)

- Group testing regression models for multiple-disease data
  - Model data in the form as it arises in application
  - binGroup package (Bilder et al., *R Journal*, 2010)
- Retests
  - Change $E(\tilde{Y}_{ijk}|z_{jk})$ to $E(\tilde{Y}_{ijk}|\text{observed tests and retests})$
  - Estimate working correlation structure
    - Could just use initial group tests
    - Not sure how to take into account retests

- Group testing regression models for multiple-disease data
  - Model data in the form as it arises in application
  - binGroup package (Bilder et al., *R Journal*, 2010)
- Retests
  - Change $E(\tilde{Y}_{ijk}|z_{jk})$ to $E(\tilde{Y}_{ijk}|\text{observed tests and retests})$
  - Estimate working correlation structure
    - Could just use initial group tests
    - Not sure how to take into account retests
  - Longitudinal setting
    - One disease
    - j subscript now corresponds to the *jth* time point

- Group testing regression models for multiple-disease data
  - Model data in the form as it arises in application
  - binGroup package (Bilder et al., *R Journal*, 2010)
- Retests
  - Change $E(\tilde{Y}_{ijk}|z_{jk})$ to $E(\tilde{Y}_{ijk}|\text{observed tests and retests})$
  - Estimate working correlation structure
    - Could just use initial group tests
    - Not sure how to take into account retests
  - Longitudinal setting
    - One disease
    - j subscript now corresponds to the *jth* time point
    - Restriction: Same individuals are always pooled together

Marginal Regression Models for Multiple-Disease Group Testing Data

Christopher R. Bilder[1], Boan Zhang[1], and Joshua M. Tebbs[2]

[1]University of Nebraska–Lincoln, Department of Statistics
[2]University of South Carolina, Department of Statistics

This research is supported in part by NIH grant R01AI067373

April 2, 2012