

Informative retesting

Joshua M. Tebbs*
Associate Professor
Department of Statistics
University of South Carolina

October 7, 2010

*This is collaborative research with Christopher R. Bilder (Department of Statistics, [University of Nebraska](#)). Our work is funded by the [National Institutes of Health](#) (R01-AI067373).

Bilder, C., Tebbs, J., and Chen, P. (2010+). Informative retesting. *Journal of the American Statistical Association*, in press.

OUTLINE

1. Nebraska IPP
2. Pooled testing
3. Informative retesting
4. Comparisons
5. Discussion

1. Nebraska IPP

- Nebraska Health and Human Services collects **chlamydia** and **gonorrhea** testing data (binary response); Region VII of IPP
- 30,000 individual tests/year
- Covariates recorded:
 - **age, gender, race**
 - **number of sexual partners, STD symptoms**
 - **clinical observations (PID, urethritis, etc.)**
 - **testing site, reason for visit**
- Researchers are interested in using **pooled testing** for estimation and identification of positive subjects

2. Pooled testing

- **General premise:** tests are performed on “pools” of individuals (e.g., urine, swabs, blood, etc.)
- The basic model assumes a binary pool response
 - **positive pool:** at least one individual in the pool is positive
 - **negative pool:** all individuals in the pool are negative
- **Intuitive:** Pooled testing only useful when dealing with low prevalence infections
- **Ubiquitous** in blood/infectious disease screening
- To **identify** who is positive, positive pools need to be **decoded**

Non-informative retesting

- Dorfman
 - individually retest all subjects in positive pools
 - currently used at American Red Cross
- Sterrett
 - individually retest until first positive is found
 - pool remaining items
 - * if pool is negative, STOP
 - * if pool is positive, repeat
 - retesting is done at random
- Both of these procedures are non-informative; covariates are ignored

Non-informative retesting

- **Non-informative** retesting (decoding) procedures essentially view the latent binary infection statuses (in a pool) as **iid** random variables with common prevalence p
- This flies in the face of reality!!
- Other non-informative procedures:
 - Halving (or more general hierarchical procedures)
 - Matrix pooling

3. Informative retesting

- Informative retesting:
 - Acknowledge that individuals in positive pools are **heterogenous**
 - Use **individual covariate information** to structure the decoding process (in pools that **test positive**)

Preliminaries

- Suppose that N individuals are drawn from a large population and that each individual is assigned initially to exactly one of K pools
- Let $Y_{ik} = 1$ if the i th individual in the k th pool is diagnosed as positive, $Y_{ik} = 0$ otherwise, for $i = 1, 2, \dots, I_k$ and $k = 1, 2, \dots, K$, so that $N = \sum_{k=1}^K I_k$
- The Y_{ik} (latent) statuses are assumed independent
- Let $Z_k = 1$ if the k th pool tests positive, $Z_k = 0$ otherwise, for $k = 1, 2, \dots, K$
- If pools are correctly classified (i.e., the diagnostic test is perfect), then

$$Z_k = 1 \iff \sum_{i=1}^{I_k} Y_{ik} > 0 \quad \text{and} \quad Z_k = 0 \iff \sum_{i=1}^{I_k} Y_{ik} = 0$$

- The goal of a retesting procedure is to determine all of the Y_{ik} diagnoses

Test misclassification

- Define $\tilde{Z}_k = 1$ if the k th pool is truly positive, $\tilde{Z}_k = 0$ otherwise, $k = 1, 2, \dots, K$
- The **true** latent individual statuses are denoted by \tilde{Y}_{ik} , where

$$\text{pr}(\tilde{Y}_{ik} = 1) = p_{ik}$$

- The **sensitivity** and the **specificity** of a diagnostic test are

$$S_e = \text{pr}(Z_k = 1 | \tilde{Z}_k = 1)$$

$$S_p = \text{pr}(Z_k = 0 | \tilde{Z}_k = 0)$$

- The probability that pool k **tests positive** is

$$\begin{aligned} \text{pr}(Z_k = 1) &= S_e + (1 - S_e - S_p)\text{pr}(\tilde{Z}_k = 0) \\ &= S_e + (1 - S_e - S_p) \prod_{i=1}^{I_k} (1 - p_{ik}) \end{aligned}$$

Assumptions

- We assume that S_e and S_p
 - are known
 - are diagnostic test dependent
 - do not depend on the covariates
 - do not depend on I_k (pool size)
- Empirical evidence supports last assumption:
 - Litvak et al. (1994), $I_k = 15$, HIV ELISA tests
 - Pilcher et al. (2005), $I_k = 90$, HIV NAT tests
 - Kacena et al. (1998a, 1998b), $I_k = 10$, C/G NAT tests

Regression models

- We model $\text{pr}(\tilde{Y}_{ik} = 1) \equiv p_{ik}$ using logistic regression
- Denote by
 - $\mathbf{x}_{ik} = (1, x_{ik1}, x_{ik2}, \dots, x_{ikp})'$ the $(p + 1) \times 1$ vector of fixed covariates for the i th individual in pool k
 - $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ the $(p + 1) \times 1$ vector of parameters
- Consider the linear **logistic regression** model

$$p_{ik} = \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{ik})}{1 + \exp(\boldsymbol{\beta}' \mathbf{x}_{ik})}$$

- In some applications, individual data are available and the model above can be fit to the individual data
 - **Nebraska IPP**: 30,000 subjects tested per year

Regression models

- The regression model can also be fit using only the **initial pooled responses** Z_k
- We can write out the likelihood function constructed from the **initial pool responses** and maximize it; that is, $\hat{\beta}$ maximizes

$$L(\beta|\mathbf{z}) = \prod_{k=1}^K \left\{ S_e + (1 - S_e - S_p) \prod_{i=1}^{I_k} (1 - p_{ik}) \right\}^{z_k} \left\{ 1 - S_e - (1 - S_e - S_p) \prod_{i=1}^{I_k} (1 - p_{ik}) \right\}^{1 - z_k},$$

where $\mathbf{z} = (z_1, z_2, \dots, z_K)'$ and

$$p_{ik} = \frac{\exp(\beta' \mathbf{x}_{ik})}{1 + \exp(\beta' \mathbf{x}_{ik})}$$

Informative Sterrett

- **Proposal:** A positive pool, say, pool k , is decoded based on $\hat{p}_{1k}, \hat{p}_{2k}, \dots, \hat{p}_{I_k k}$, ordered so that

$$\hat{p}_{(1)k} \leq \hat{p}_{(2)k} \leq \dots \leq \hat{p}_{(I_k)k},$$

where $\hat{p}_{(i)k}$ is the i th largest probability; start with the **highest-risk** subjects

- **Informative** procedures
 - One-stage Informative Sterrett (1SIS)
 - Two-stage Informative Sterrett (2SIS)
 - Full Informative Sterrett (FIS)
- **Non-informative** procedures (treat subjects as **iid** with common p)
 - Non-informative Sterrett (NIS)
 - Dorfman (D)

One-stage Informative Sterrett (1SIS)

- Order subjects according to risk

$$\hat{p}_{(1)k} \leq \hat{p}_{(2)k} \leq \cdots \leq \hat{p}_{(I_k)k}$$

- Test in order of descending probability until **1st positive** is found
- **Repool** remaining subjects
 - * If $-$, **DONE**
 - * If $+$, **Dorfman** on remaining (i.e., test **individually**)

1SIS with $I_k = 4$

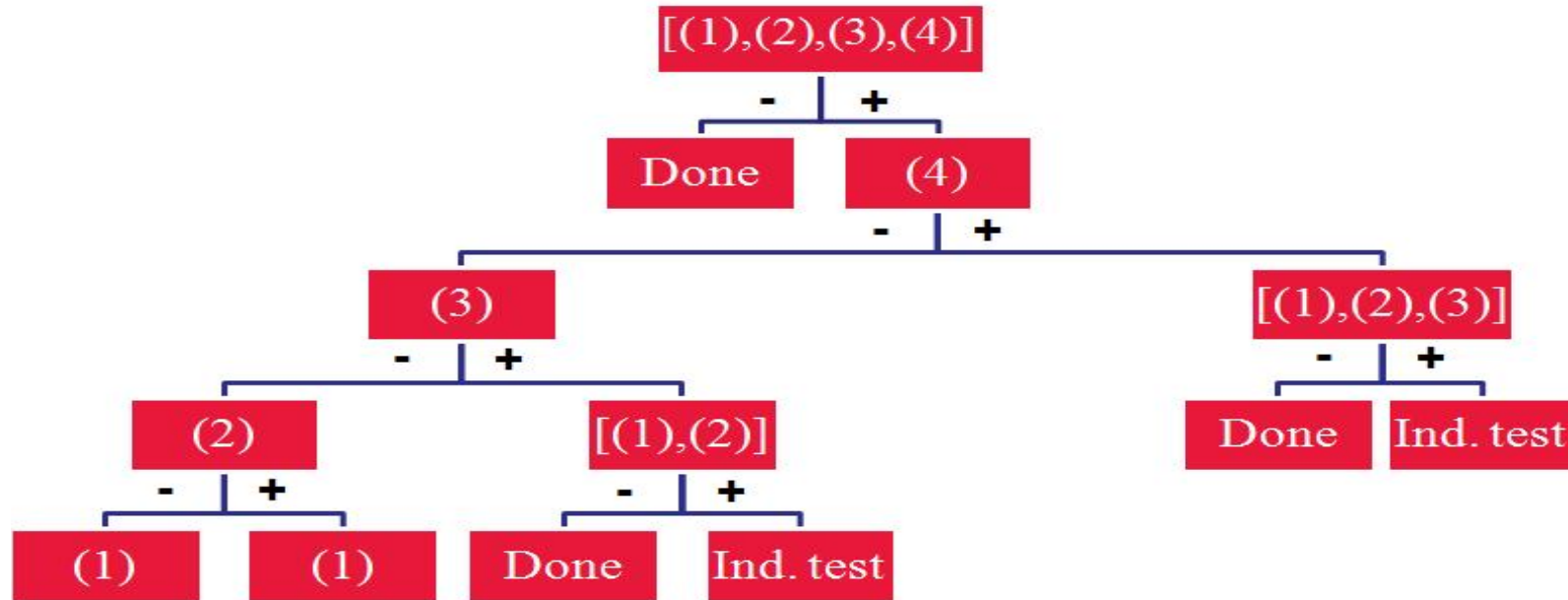


Figure 1: *One-stage Informative Sterrett (1SIS) with $I_k = 4$.*

Two-stage Informative Sterrett (2SIS)

- Order subjects according to risk

$$\hat{p}_{(1)k} \leq \hat{p}_{(2)k} \leq \cdots \leq \hat{p}_{(I_k)k}$$

- Test in order of descending probability until **1st positive** is found
- **Repool** remaining subjects
 - * If –, **DONE**
 - * If +, continue individual testing (in order of descending probability) until **2nd positive** is found; **repool** remaining subjects
 - If –, **DONE**
 - If +, **Dorfman** on remaining

2SIS with $I_k = 4$

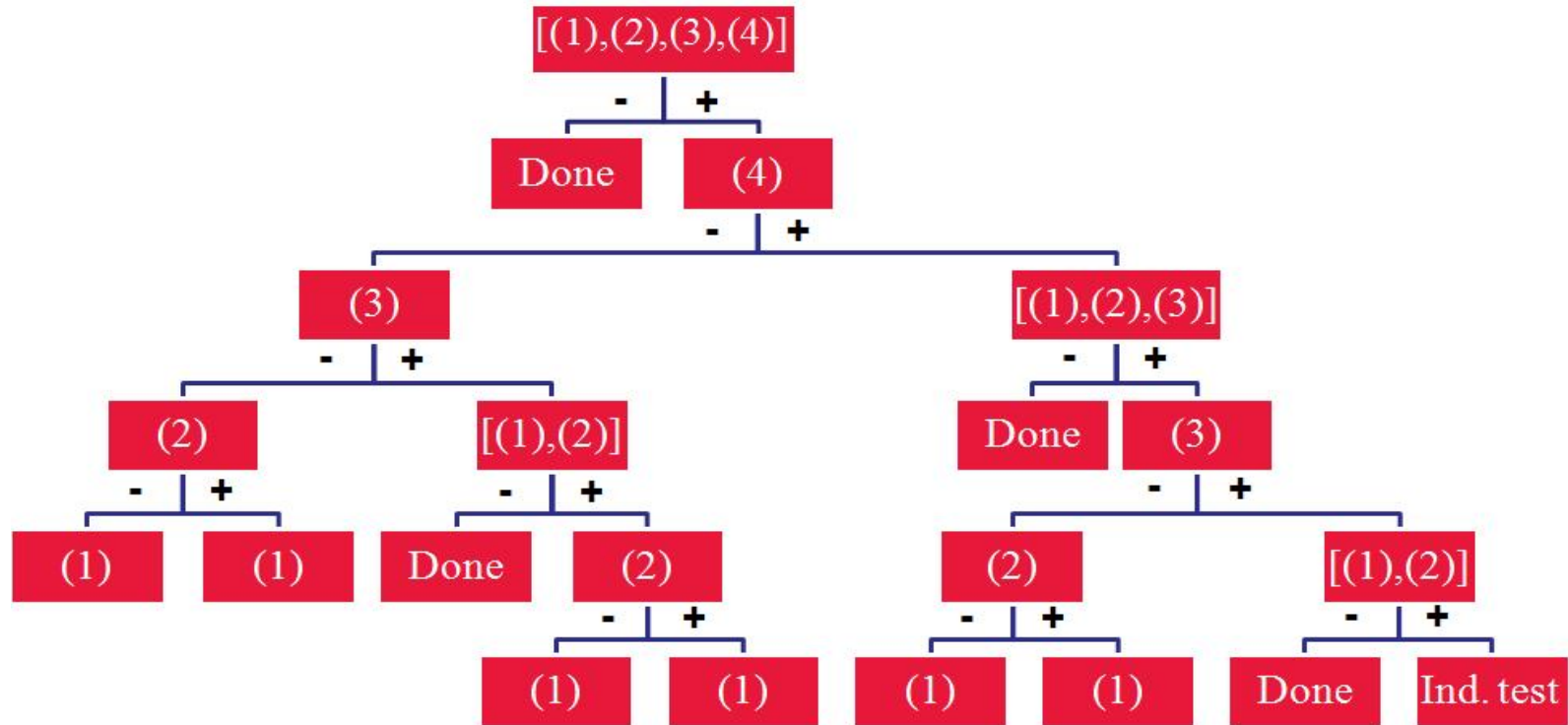


Figure 2: *Two-stage Informative Sterrett (2SIS) with $I_k = 4$.*

Full Informative Sterrett (FIS)

- Order subjects according to risk

$$\hat{p}_{(1)k} \leq \hat{p}_{(2)k} \leq \cdots \leq \hat{p}_{(I_k)k}$$

- Same as 2SIS, except that individual testing continues (in order of descending probability) until either
 - * a negative pool results
 - * the last subject is tested

FIS with $I_k = 3$

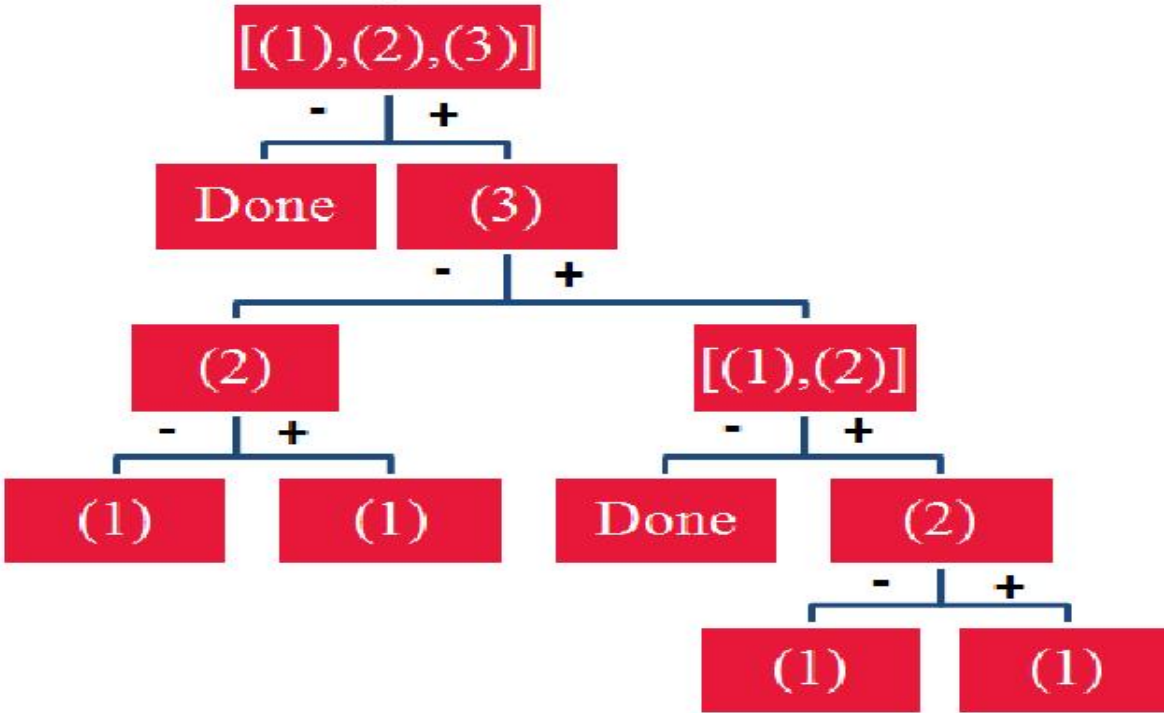


Figure 3: *Full Informative Sterrett (FIS) with $I_k = 3$.*

FIS with $I_k = 4$

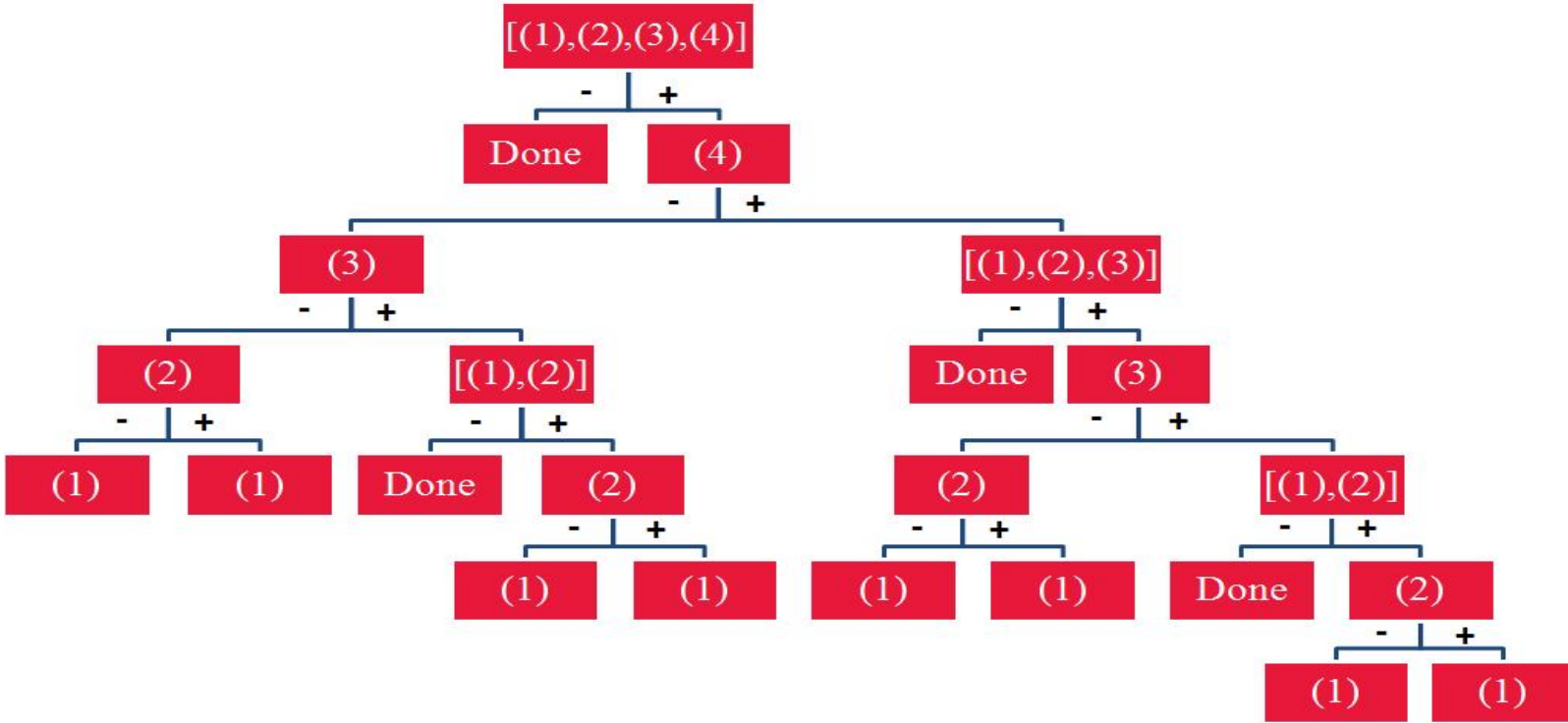


Figure 4: *Full Informative Sterrett (FIS) with $I_k = 4$.*

Probability mass functions

- For **1SIS**, **2SIS**, and **FIS**, we have derived probability mass functions for

$$T_k = \text{number of tests needed to decode pool } k$$

- The expressions (algorithms) incorporate the possibility of false positives or negatives occurring **at any stage (test)** in the decoding process
- Expressions arise by **comparing trees** of size $I_k + 1$ to I_k (for FIS) and exploiting the patterns in doing so
- PMFs depend on
 - $p_{(1)k}, p_{(2)k}, \dots, p_{(I_k)k}, S_e, S_p,$ and I_k
- Initial computational difficulty with **FIS**
 - 2^{I_k} calculations needed
 - Had to find an algorithm that would work for I_k **large** (e.g., $I_k = 90$)

1SIS PMF

$$\begin{aligned}
 \text{pr}(T = 1) &= S_p \left[\prod_{i=1}^I (1 - p_{(i)}) \right] + \bar{S}_e \left[1 - \prod_{i=1}^I (1 - p_{(i)}) \right] \\
 \text{pr}(T = t) &= \bar{S}_p S_p^{t-2} \left[\prod_{i=1}^I (1 - p_{(i)}) \right] (\bar{S}_p - S_e) + S_e \left\{ \prod_{i=I-4-t}^I [S_p(1 - p_{(i)}) + \bar{S}_e p_{(i)}] \right\} \\
 (t = 3, 4, \dots, I) &\quad \times [\bar{S}_p(1 - p_{(I-3-t)}) + S_e p_{(I-3-t)}] \left\{ S_p \left[\prod_{i=1}^{I+2-t} (1 - p_{(i)}) \right] + \bar{S}_e \left[1 - \prod_{i=1}^{I+2-t} (1 - p_{(i)}) \right] \right\} \\
 \text{pr}(T = I + 1) &= \bar{S}_p S_p^{I-2} \left[\prod_{i=1}^I (1 - p_{(i)}) \right] + S_e \left\{ \prod_{i=3}^I [S_p(1 - p_{(i)}) + \bar{S}_e p_{(i)}] + S_p^{I-2} \left[\prod_{i=1}^I (1 - p_{(i)}) \right] \right\} \\
 \text{pr}(T = I + 2) &= \sum_{m=1}^{I+4} (\bar{S}_p - S_e) \bar{S}_p^2 S_p^{I+1-m} \left[\prod_{i=1}^I (1 - p_{(i)}) \right] + S_e \left\{ \prod_{i=m}^I [S_p(1 - p_{(i)}) + \bar{S}_e p_{(i)}] \right\} \\
 &\quad \times [\bar{S}_p(1 - p_{(m-1)}) + S_e p_{(m-1)}] \left\{ \bar{S}_p \left[\prod_{i=1}^{m-2} (1 - p_{(i)}) \right] + S_e \left[1 - \prod_{i=1}^{m-2} (1 - p_{(i)}) \right] \right\}
 \end{aligned}$$

- Pool size = I

- $\bar{S}_e = 1 - S_e$, $\bar{S}_p = 1 - S_p$

4. Comparisons

- Compare 1SIS, 2SIS, FIS, NIS, and D through their
 - CDFs
 - Moments (Mean/SD)
 - Classification accuracy (see paper)

- Comparisons are made for fixed $\mathbf{p}_k = (p_{1k}, p_{2k}, \dots, p_{I_k k})'$

- Example 1. $S_e = 0.95$, $S_p = 0.99$, $I_k = 8$,

$$\mathbf{p}_k = (0.002, 0.010, 0.031, 0.053, 0.100, 0.102, 0.150, 0.172)'$$

- Example 2. $S_e = 0.99$, $S_p = 0.99$, $I_k = 10$,

$$\mathbf{p}_k = (0.01, 0.01, \dots, 0.01, 0.50)'$$

- Example 3. $S_e = 0.99$, $S_p = 0.99$, $I_k = 100$,

$$\mathbf{p}_k = (0.01, 0.01, \dots, 0.01, 0.10, 0.10)'$$

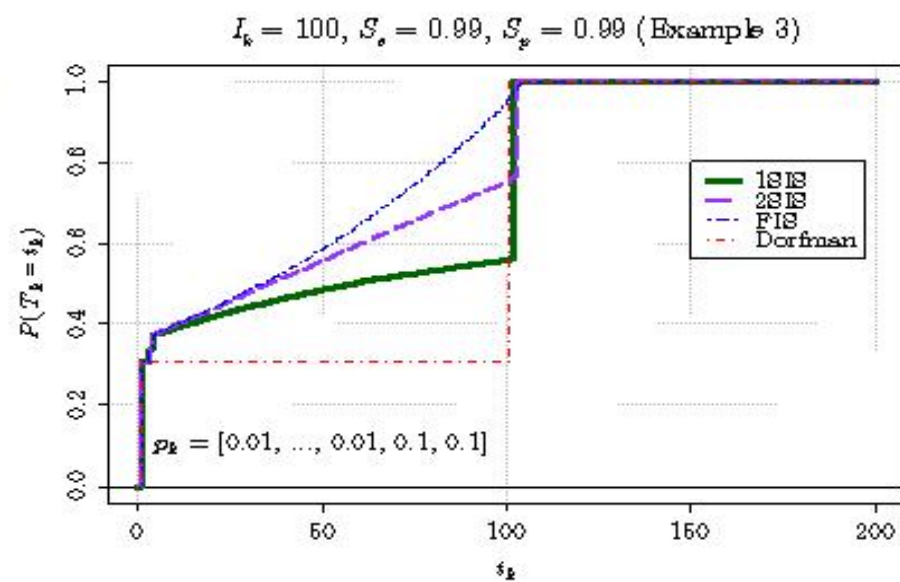
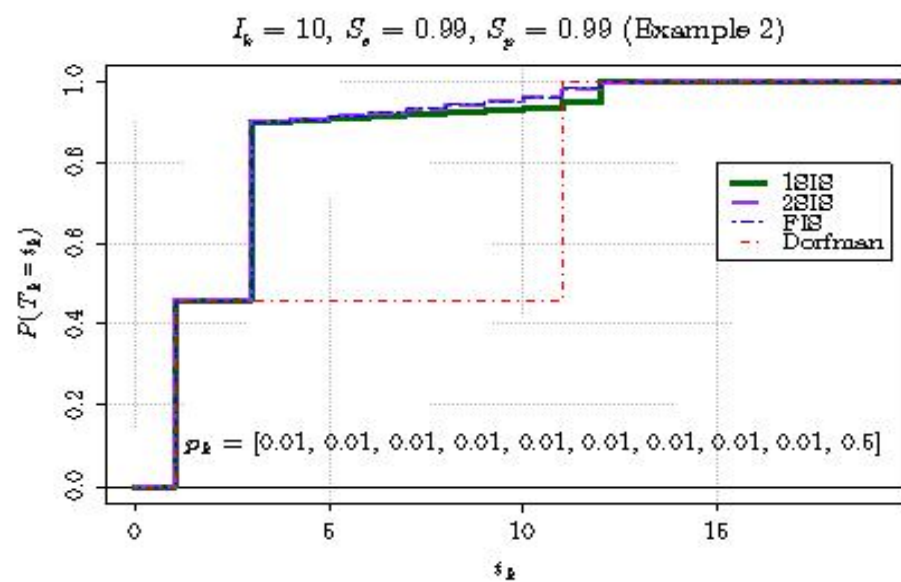
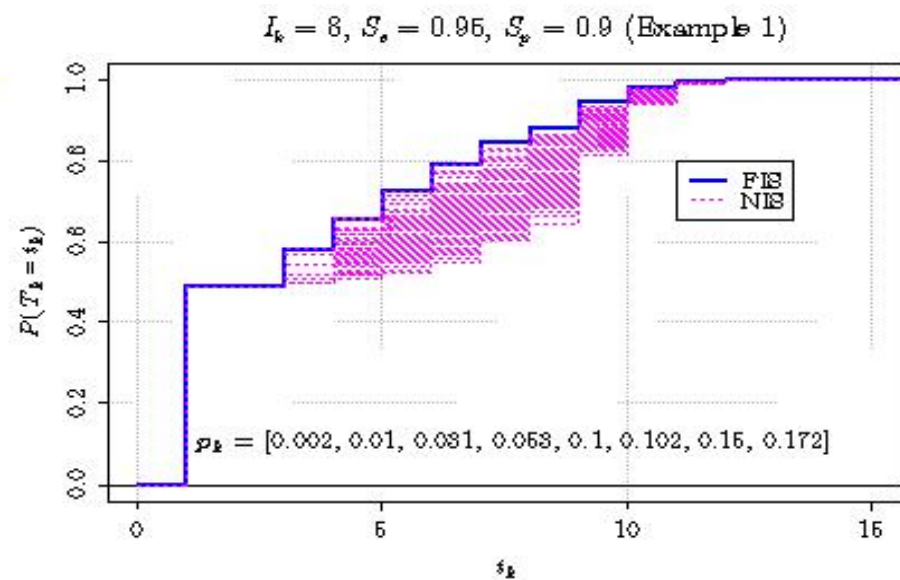
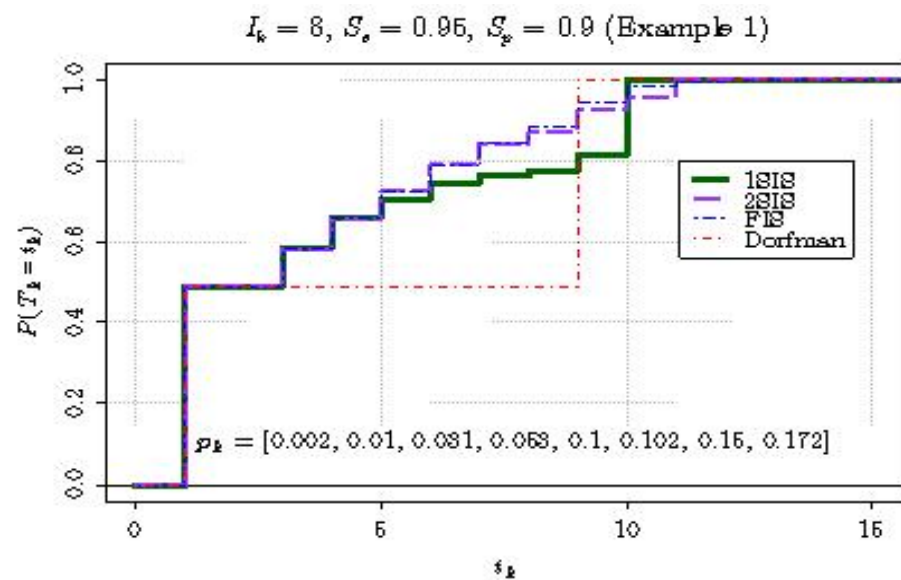


Figure 5: *CDFs for identification procedures.*

Example 1

	Unconditional		Conditional	
	Mean	SD	Mean	SD
D	5.09	4.00	9.00	0.00
1SIS	3.98	3.57	6.83	2.88
2SIS	3.67	3.20	6.22	2.60
FIS	3.61	3.10	6.11	2.45

- “Conditional” means conditional on the pool testing positive
- Example 1. $S_e = 0.95$, $S_p = 0.99$, $I_k = 8$,

$$\mathbf{p}_k = (0.002, 0.010, 0.031, 0.053, 0.100, 0.102, 0.150, 0.172)'$$

Example 2

	Unconditional		Conditional	
	Mean	SD	Mean	SD
D	6.42	4.98	11.00	0.00
1SIS	2.80	2.75	4.32	2.99
2SIS	2.67	2.43	4.09	2.56
FIS	2.67	2.42	4.08	2.54

- Example 2. $S_e = 0.99$, $S_p = 0.99$, $I_k = 10$,

$$\mathbf{p}_k = (0.01, 0.01, \dots, 0.01, 0.50)'$$

Example 3

	Unconditional		Conditional	
	Mean	SD	Mean	SD
D	70.35	46.10	101.00	0.00
1SIS	53.81	46.87	77.15	37.29
2SIS	45.58	42.90	65.27	37.24
FIS	39.80	37.32	56.94	32.40

- Example 3. $S_e = 0.99$, $S_p = 0.99$, $I_k = 100$,

$$\mathbf{p}_k = (0.01, 0.01, \dots, 0.01, 0.10, 0.10)'$$

Nebraska IPP

- Use 2004 data (30,000+ patient records) as training data
- For each infection, cross classify data by specimen type and gender
- Fit first order logit model to training data (within strata for each infection)
 - age, gender, race
 - number of sexual partners, STD symptoms
 - clinical observations (PID, urethritis, etc.)
 - type of clinic, reason for visit
- Treat year-2005 statuses as the true values
- Assign year-2005 subjects (30,067) to pools in chronological order within strata
- Apply each decoding procedure within strata to get diagnosed responses
- Can compare diagnosed responses to “true” responses

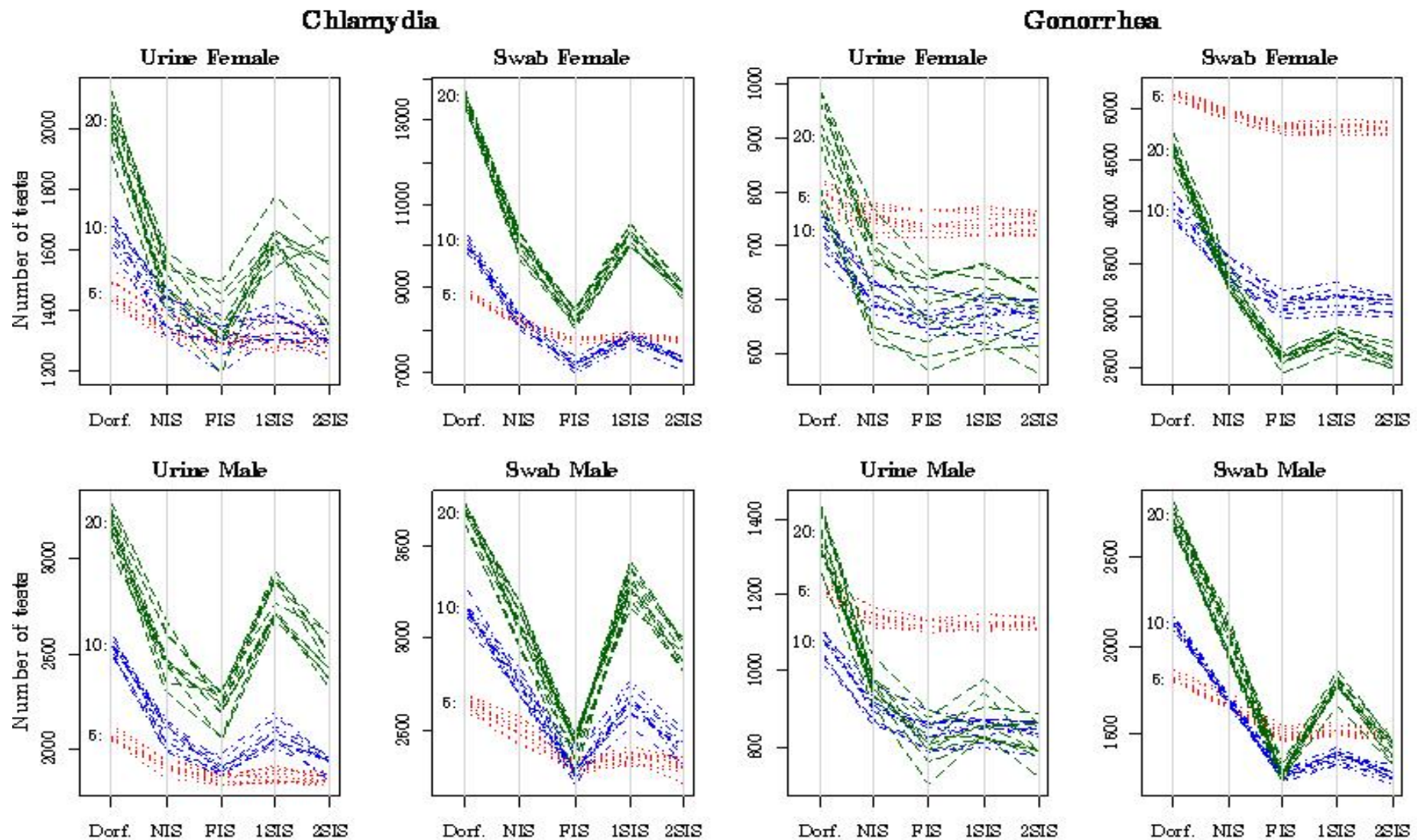


Figure 6: *Nebraska data. Number of tests for each infection, by specimen and gender.*

5. Discussion

- **Classification accuracy** analysis in paper
- Currently working on **informative versions** of
 - hierarchical decoding
 - threshold decoding
 - matrix pooling
- Design issues **not addressed**
 - Initial pool composition (in terms of covariates)
 - “Optimal” initial pool size
- **How important is the fit of the model?**
- **Multiple infection** models (with pooled responses)
- Extend informative retesting to multiple infections

Thank you!