Multi-stage group testing with heterogeneous probabilities of disease positivity

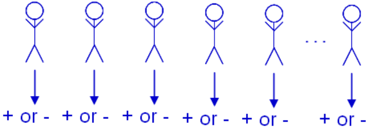
Christopher R. Bilder[1], Joshua M. Tebbs[2], and Michael S. Black[3]

[1]University of Nebraska–Lincoln, Department of Statistics
[2]University of South Carolina, Department of Statistics
[3]University of Wisconsin-Platteville, Department of Mathematics

This research is supported in part by NIH grant R01AI067373

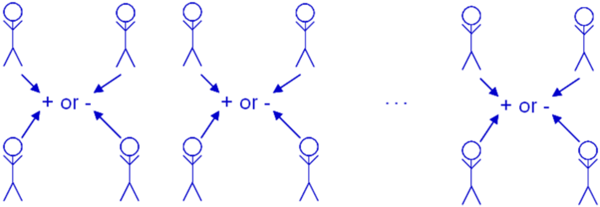
June 18, 2014

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



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

- 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
  - "Decode" the positive group
- Benefits:
  - Reduction in tests
  - Cost savings (less tests and labor)
- Overall disease prevalence needs to be small

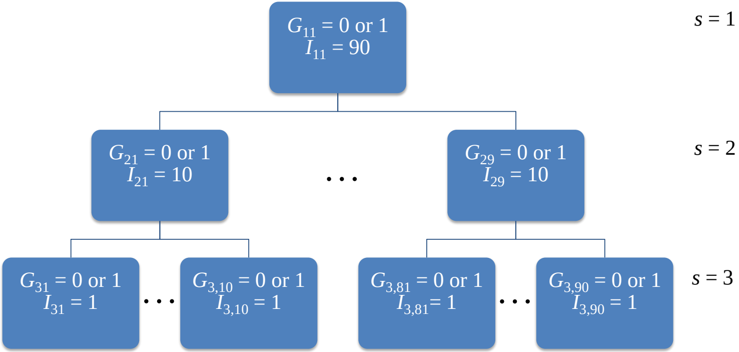
- American Red Cross (Stramer et al. 2004; ARC 2014)
  - Millions of blood donations per year
  - HIV, hepatitis B, and hepatitis C
  - 1st stage - Initial group of size 16
  - 2nd stage - Individual testing
- HIV screening by public health clinics: Los Angeles three-stage hierarchical group testing
  - 1st stage - Initial group of size 90
  - 2nd stage - Subgroups of size 10
  - 3rd stage - Individual testing
- Number of tests can be further reduced by allowing more than two stages

- Informative retesting
  - Incorporate factors that influence positive or negative disease status
  - Estimate the probability that an individual is positive
  - These probabilities are used to select
    - Number of subgroups
    - Subgroup sizes
    - Members of each subgroup

    in order to form a retesting configuration
  - Goal is to reduce the number of tests
  - Papers include: Bilder et al. (*JASA*, 2010), McMahan et al. (*Biometrics*, 2012), McMahan et al. (*Biometrics*, 2012b), Black et al. (*JRSS-C*, 2012)

- Purpose
  - Examine hierarchical group testing methods (three or more stages)
  - Incorporate informative retesting ideas
  - Determine the retesting configuration that minimizes the number of tests

- Consider a group with I individuals
- Define G_{sj} as a binary random variable denoting the test status for group j at the sth stage
    - $G_{sj} = 0$ for a negative test result
    - $G_{sj} = 1$ for a positive test result
- Define I_{sj} as the number of individuals in group j at the sth stage ($I_{11} \equiv I$)
- Los Angeles example:



$G_{11} = 0$ or 1
$I_{11} = 90$

$s = 1$

$G_{21} = 0$ or 1
$I_{21} = 10$

. . .

$G_{29} = 0$ or 1
$I_{29} = 10$

$s = 2$

$G_{31} = 0$ or 1
$I_{31} = 1$

. . .

$G_{3,10} = 0$ or 1
$I_{3,10} = 1$

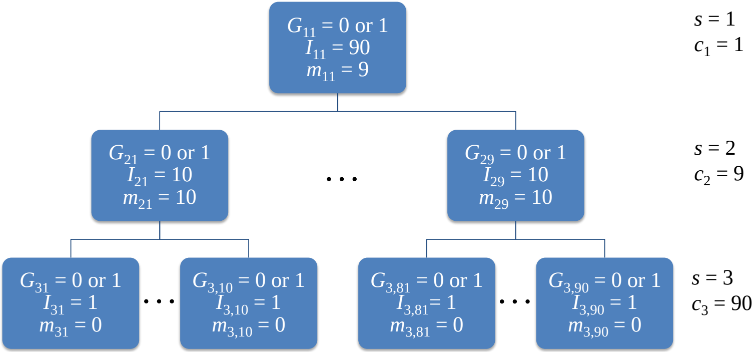
$G_{3,81} = 0$ or 1
$I_{3,81} = 1$

. . .

$G_{3,90} = 0$ or 1
$I_{3,90} = 1$

$s = 3$

- If $G_{sj} = 1$, the corresponding group is divided into m_{sj} subgroups
- Define c_s as the total number possible of subgroups at the sth stage
- Los Angeles example:



$G_{11} = 0$ or 1, $I_{11} = 90$, $m_{11} = 9$

$s = 1$, $c_1 = 1$

$G_{21} = 0$ or 1, $I_{21} = 10$, $m_{21} = 10$ … $G_{29} = 0$ or 1, $I_{29} = 10$, $m_{29} = 10$

$s = 2$, $c_2 = 9$

$G_{31} = 0$ or 1, $I_{31} = 1$, $m_{31} = 0$ … $G_{3,10} = 0$ or 1, $I_{3,10} = 1$, $m_{3,10} = 0$

$G_{3,81} = 0$ or 1, $I_{3,81} = 1$, $m_{3,81} = 0$ … $G_{3,90} = 0$ or 1, $I_{3,90} = 1$, $m_{3,90} = 0$

$s = 3$, $c_3 = 90$

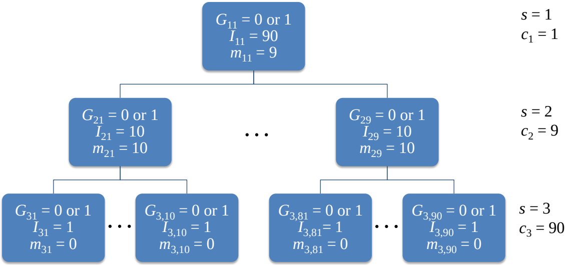
- Let T be the number of tests for one group
- The expected number of tests is

$$E(T) = 1 + \sum_{s=1}^{S-1} \sum_{j=1}^{c_s} m_{sj} P\left(\bigcap_{\{(s'j'):G_{sj}=1\}} \{G_{s'j'} = 1\} \right)$$

where S is the total number of stages

- Los Angeles example with $s = 2$, $j = 1$:

$$P\left(\bigcap_{\{(s'j'):\{G_{21}=1\}\}} \{G_{s'j'} = 1\} \right) = P(\{G_{11} = 1\} \cap \{G_{21} = 1\})$$



- Define $\tilde{G}_{sj}$ as a binary random variable denoting the TRUE status for group j at the sth stage
- Accuracy of an assay
  - $S_e = P(G_{sj} = 1 | \tilde{G}_{sj} = 1)$ is the sensitivity
  - $S_p = P(G_{sj} = 0 | \tilde{G}_{sj} = 0)$ is the specificity
- Then $P\left(\bigcap_{\{(s'j'):G_{sj}=1\}\}} \{G_{s'j'} = 1\}\right)$

$$
\begin{aligned}
= \quad & (1 - S_p)^s \left\{ \prod_{i=1}^{I_{11}} (1 - p_i) \right\} + \sum_{a=1}^{s-1} S_e^a (1 - S_p)^{s-a} \left\{ \prod_{i \in B_{a+1,j'}} (1 - p_i) \right\} \\
& + S_e^s \left\{ 1 - \prod_{i \in B_{sj}} (1 - p_i) \right\}
\end{aligned}
$$

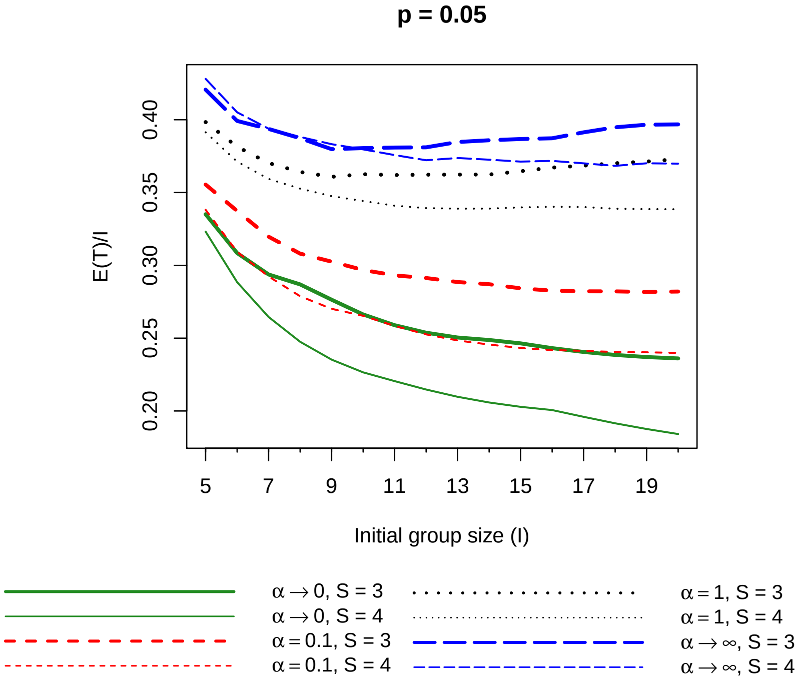
where

- p_i is the probability that individual i is truly positive
- $i \in B_{sj}$ means the set of individuals who belong to the jth ordered group at the sth stage
- $i \in \bar{B}_{sj}$ means the set of individuals who below to the parent group of B_{sj} excluding those in B_{sj} itself

- Want to minimize the number of tests
- Find the retesting configuration that essentially achieves the above by minimizing $E(T)$
  - p_i is unknown
  - In practice, estimate p_i and minimize the estimated $E(T)$
- Simplification
  - Order individuals by p_i values
  - Individuals are assigned to subgroups successively by this ordering

- Examine ALL possible retesting configurations
  - Define configuration with minimum $E(T)$ as the optimal retesting configuration (ORC)
  - $(S-1)^{I-1}$ possible configurations
- Use a search algorithm
  - Formulate as an integer program and use method of steepest descent
  - Define configuration resulting from algorithm as the candidate retesting configuration (CRC)
  - Algorithm is not guaranteed to find ORC, but we have found it to work well

- Examine $E(T)$ in specific situations
- Let $P_i \sim beta(\alpha, \alpha(1-p)/p)$ for $i = 1, \ldots, I$, $\alpha > 0$, $0 < p < 1, \text{ and } E(P_i) = p$
  - p represents the overall prevalence
  - As $\alpha \to \infty$, $Var(P_i) \to 0$; p_i's become homogeneous
  - As $\alpha \to 0$, $Var(P_i)$ increases; p_i's become more heterogeneous
- Use $E(P_{(i)})$ for p_i in $E(T)$
- $S_e = S_p = 0.95$
- CRC results in the same configurations as ORC
  - All $S = 3$ cases
  - All $S = 4$ cases where ORC was calculated ($I \leq 14$)


p = 0.05
E(T)/I
0.20
0.25
0.30
0.35
0.40
5
7
9
11
13
15
17
19
Initial group size (I)
α → 0, S = 3
α → 0, S = 4
α = 0.1, S = 3
α = 0.1, S = 4
α = 1, S = 3
α = 1, S = 4
α → ∞, S = 3
α → ∞, S = 4

- Sherlock et al. (2007)
  - Examines publicly funded HIV testing practices across United States
  - Three-stage hierarchical group testing

Location	Observed prevalence	1st stage group size	2nd stage group sizes
Los Angeles	0.0045	90	9 groups of size 10
North Carolina	0.0021	90	9 groups of size 10
San Francisco	0.0175	50	5 groups of size 10
Seattle-King County	0.0164	30	3 groups of size 10
Atlanta	0.0030	48	6 groups of size 8

- Quote from the paper:

> *... the use of pooled NAATs to detect acute HIV infection is becoming a popular strategy for the screening of large populations. However, the most efficient approach remains to be determined.*

- Can we do better?
- ORC assuming homogeneity
  - Use observed prevalence as the true prevalence p
  - Find configuration that minimizes $E(T)$
- CRC accounting for heterogeneity
  - Exact amount of heterogeneity is unknown
  - $P_i \sim beta(\alpha, \alpha(1-p)/p)$ for $i = 1, \ldots, I$, $\alpha > 0$, $0 < p < 1, \text{ and } E(P_i) = p$
- Assume $S_e = S_p = 0.99$ and only examine the same 1st stage group size as originally used

		1st stage	2nd stage
Location	Observed	group size	group sizes
Los Angeles	0.0045	90	9 groups of size 10
North Carolina	0.0021	90	9 groups of size 10
San Francisco	0.0175	50	5 groups of size 10
Seattle-King County	0.0164	30	3 groups of size 10
Atlanta	0.0030	48	6 groups of size 8

	ORC homogeneity	Reduction in $E(T)$		
	2nd stage	from CRC under heterogeneity		
Location	group sizes	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0.1$
Los Angeles	10 groups of size 9	8.6%	15.2%	36.8%
North Carolina	10 groups of size 9	7.7%	13.6%	33.1%
San Francisco	2 groups of size 7, 6 groups of size 6	8.4%	15.1%	37.4%
Seattle-King County	6 groups of size 5	7.2%	12.2%	32.0%
Atlanta	6 groups of size 7, 1 group of size 6	6.5%	11.1%	27.3%

- Limitations
  - Comparsion of $E(T)$, not the actual number of tests that may occur
  - Amount of heterogeneity is unknown
    - Levels of variability are not extreme
    - Los Angeles with $\alpha = 0.1$: 0.001 and 0.999 quantiles for beta distribution are slightly larger than 0 and approximately equal to 0.0445, respectively
- Potential for significant benefits from using ORC and CRC

Multi-stage group testing with heterogeneous probabilities of disease positivity

Christopher R. Bilder[1], Joshua M. Tebbs[2], and Michael S. Black[3]

[1]University of Nebraska–Lincoln, Department of Statistics
[2]University of South Carolina, Department of Statistics
[3]University of Wisconsin-Platteville, Department of Mathematics

This research is supported in part by NIH grant R01AI067373

June 18, 2014