Strategies for pooling in array testing configurations with multiplex assays

Christopher R. Bilder, Joshua M. Tebbs, and Christopher S. McMahan
University of Nebraska-Lincoln, University of South Carolina, and Clemson University

Introduction

Abstract

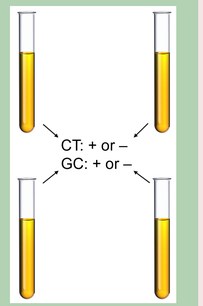
High volume screening of clinical specimens for infectious diseases is often performed by a process known as group testing. This algorithmic process involves pooling together portions of specimens from separate individuals. Each group formed is tested to detect a human body's response to infection or the pathogen that leads to disease. Follow-up retesting is performed on those groups that test positively to decode the positive individuals from the negative individuals. One of the most efficient group testing algorithms is array testing. In its simplest form, specimens are tested in a grid-like structure so that groups can be formed by row and by column. Positive-testing rows/columns present clues on which individual specimens to retest. In our presentation, we investigate how one can use multiplex assays (multiple-disease tests) together with individual risk information to increase testing efficiency. We show how particular specimen arrangements within an array can lower the number of retests needed when compared to unordered arrangements.

Corresponding author: Christopher R. Bilder, chris@chrisbilder.com, www.chrisbilder.com

This research was supported by Grant R01 AI121351 from the National Institutes of Health.

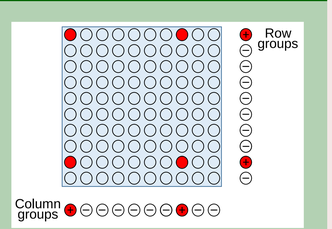
What is group testing?

- Also known as *pooled testing* and *specimen pooling*
- Used to screen a large number of individuals for infectious diseases
- Example: Chlamydia (CT) and gonorrhea (GC) testing with the Aptima Combo 2 Assay at the University of Iowa's State Hygienic Laboratory (SHL)
  - An amalgamation of specimens from 4 individuals forms a group
  - If a group tests negatively for both diseases, all individuals within it are declared disease free
  - If a group tests positively for at least one disease:
    - Need to determine who is positive and who is negative for which diseases
    - SHL retests all group members individually with the same assay; thus, a 2-stage hierarchical testing algorithm overall
- Benefits in comparison to testing each individual separately (individual testing):
  - Smaller number of tests
  - Cost savings
- Estimated savings for SHL during a recent 5-year evaluation period was approximately $3 million



Array testing

- A form of group testing that attempts to reduce the number of retests by performing more group tests in the first stage
- Specimens are arranged in a grid-like structure, like on a microplate
- Specimens are pooled by rows and columns and tests are performed upon them
- Intersections of rows and columns that test positively for at least one disease are retested individually with the same assay
- Example to the right shows a 10×10 microplate with 4 individual retests



Purpose

- A test for multiple infectious diseases is called a *multiplex assay;* examples include
  - Aptima Combo 2 Assay for chlamydia and gonorrhea
  - BD Max Assay for chlamydia, gonorrhea, and trichomoniasis
  - Procleix Ultrio Assay for HIV, hepatitis B, and hepatitis C
- Hou et al. (*Biostatistics*, 2019) is the only research article on how to use array testing with multiplex assays
  - Assumed each individual had same probability of being positive for a particular disease
- However, some individuals should be at a higher risk (probability) for being positive than others!
  - *Informative group testing* exploits these risk differences to obtain more efficient (smaller number of tests) testing algorithms
  - McMahan et al. *(Biometrics*, 2012) is the only research article on informative group testing for arrays, but this article was for single-disease assays
- Purpose: Develop informative group testing algorithms that reduce the number of tests needed when using multiplex assays on specimens arranged in arrays

Array testing algorithm

Testing configuration

- Definitions: I = number of rows and J = number of columns for the array
- Goal: Arrange specimens in the $I \times J$ array to minimize the number of retests needed
- Generalize the *gradient arrangement* proposed by McMahan et al. (2012) for single disease assays
  - Order specimens by probability of being positive for at least one disease
  - Place these ordered specimens by column (or row) into the array
  - One would expect a fewer number of columns to test positively in comparison to an unordered arrangement (equivalent to Hou et al., 2019)
- In practice, these probabilities of being positive for at least one disease are estimated using individual-specific information (personal behavior and clinical observations) as covariates in regression models; see Bilder et al. (*JASA*, 2010) and Bilder et al. (*Biometrics*, 2019)

Operating characteristics

- Definitions:
  - T = number of tests used to determine which individuals are positive/negative for K diseases in an array
  - $T_{ij} = 1$ if specimen in cell (i, j) of the array requires a retest; $T_{ij} = 0$ otherwise
  - $\tilde{\mathbf{Y}}_{ij} = (\tilde{Y}_{ij1}, \ldots, \tilde{Y}_{ijK})$ is a vector of binary random variables representing the true positive (1) or negative (0) statuses of the individual in cell (i, j) for all K diseases
  - $P(\tilde{\mathbf{Y}}_{ij} = \tilde{\mathbf{y}}) = p_{ij,\tilde{\mathbf{y}}}$
    - $1 - p_{ij,\mathbf{0}}$ represents the probability of being positive for at least one disease
    - $p_{11,\tilde{\mathbf{y}}}, \ldots, p_{IJ,\tilde{\mathbf{y}}}$ are potentially all unequal for each $\tilde{\mathbf{y}}$
- Expected number of tests for a $I \times J$ array

$$E(T) = I + J + \sum_{i=1}^{I}\sum_{j=1}^{J} P(T_{ij} = 1)$$

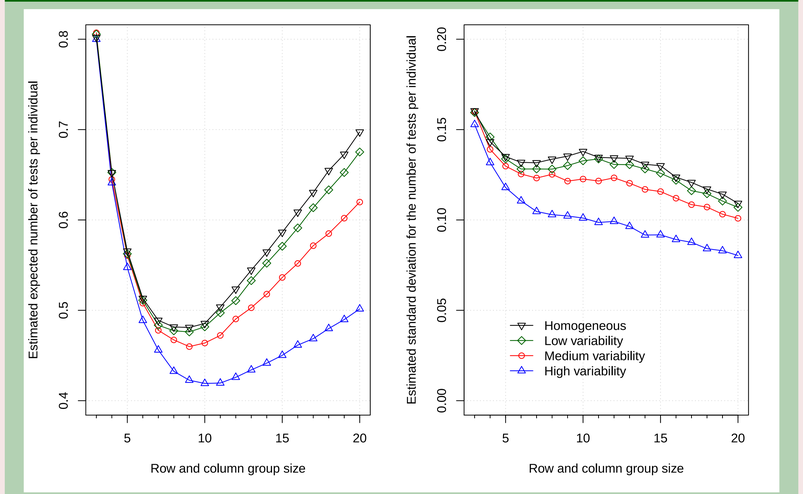
- Hou et al. (*Biostatistics*, 2019) derived $P(T_{ij} = 1)$ when $p_{11,\tilde{\mathbf{y}}} = \cdots = p_{IJ,\tilde{\mathbf{y}}}$ for all $\tilde{\mathbf{y}}$ and $K = 2$; due to its complexity, they resorted to Monte Carlo simulation for $K = 3$ as their only other K examined
- This complexity is greatly increased when $p_{11,\tilde{\mathbf{y}}}, \ldots, p_{IJ,\tilde{\mathbf{y}}}$ are unequal!
  - Therefore, we use Monte Carlo simulation alone to estimate $E(T)$
  - Summary: Simulate the number of retests needed for B arrays and record the number of retests n_b, $b = 1, \ldots, B$; average these values to estimate $\sum_{i=1}^{I}\sum_{j=1}^{J} P(T_{ij} = 1)$
  - The standard deviation for the number of tests $SD(T)$ is estimated simply as the sample standard deviation of $n_1, \ldots, n_B$

Evaluations

Process

- Focus on two diseases ($K = 2$) and square arrays ($I = J$)
- Define $\mathbf{P}_{ij} \sim$ Dirichlet($\boldsymbol{\alpha}$) as a random vector of joint probabilities of disease
  - $\mathbf{P}_{ij} = (P_{ij,00}, P_{ij,01}, P_{ij,10}, P_{ij,11})$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$
  - Simulate IJ observed values of $\mathbf{P}_{ij}$ for an array
  - Variability in the observed values emulates the heterogeneity among individuals within an array that would occur in practice
  - Purpose: Estimate $E(T)$ and $SD(T)$ using $B = 5000$ arrays
- Examine $E(T)$ and $SD(T)$ relative to the variability in the probabilities of being positive for a disease ($Var(P_{i1+})$ and $Var(P_{i+1})$)
  - Let $\boldsymbol{\alpha} = (18.25, 0.75, 0.75, 0.25)$
  - Low variability: $\mathbf{P}_{ij} \sim$ Dirichlet($4\boldsymbol{\alpha}$)
  - Medium variability: $\mathbf{P}_{ij} \sim$ Dirichlet($\boldsymbol{\alpha}$)
  - High variability: $\mathbf{P}_{ij} \sim$ Dirichlet($\boldsymbol{\alpha}/4$)
  - Homogeneous (no variability):
    - Use $E(\mathbf{P}_{ij})$ for $\mathbf{P}_{ij} \sim$ Dirichlet($\boldsymbol{\alpha}$) as the realization for each individual
    - Equivalent to Hou et al. (*Biostatistics*, 2019)
  - Each variability case has $E(P_{ij,1+}) = E(P_{ij,+1}) = 0.05$
- Assay sensitivity and specificity are set to 0.95 and 0.99, respectively, for each disease and test

Results



- Summary
  - Expected number of tests per individual, $E(T)/(IJ)$, is much less than 1, so array testing likely will lead to a significant reduction in tests in comparison to individual testing
  - Higher variability in probabilities of being positive leads to smaller values for $E(T)$ and $SD(T)$
  - The overall accuracy (not shown) had similar results regarding the variability cases
- Conclusion: Gradient arrangements of specimens can significantly reduce $E(T)$ and $SD(T)$, while also increasing overall accuracy of the algorithm!