Pooled testing with penalized regression models

Christopher R. Bilder
University of Nebraska–Lincoln
Department of Statistics
chris@chrisbilder.com

Research is supported by NIH grant R01 AI121351

Joint work with
Pranta Das at University of Nebraska-Lincoln,
Joshua M. Tebbs at University of South Carolina, and
Christopher S. McMahan at Clemson University

- Infectious disease testing
  - Timely
  - Efficient

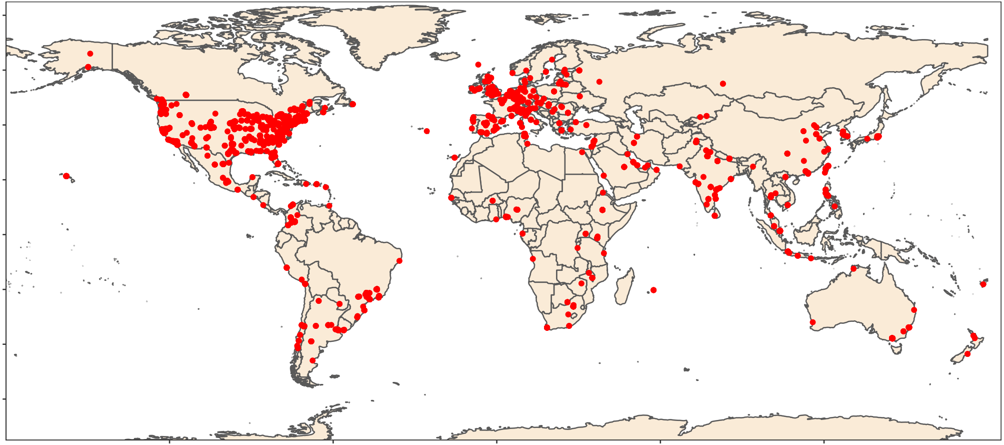
- Infectious disease testing
  - Timely
  - Efficient
- COVID-19 pandemic

- Infectious disease testing
  - Timely
  - Efficient
- COVID-19 pandemic
- Pooled testing
  - Also known as “group testing”
  - Process in Nebraska (Abdalhamid et al. 2020)

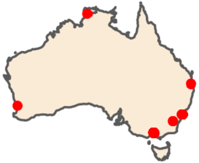
- Infectious disease testing
  - Timely
  - Efficient
- COVID-19 pandemic
- Pooled testing
  - Also known as “group testing”
  - Process in Nebraska (Abdalhamid et al. 2020)
  - Combine together portions of 5 specimens from different individuals into a “pool”
  - Test as if it were a single specimen
    - Pool tests negative: All 5 individuals are negative
    - Pool tests positive: Retest each individual separately to determine who is positive or negative

- Infectious disease testing
  - Timely
  - Efficient
- COVID-19 pandemic
- Pooled testing
  - Also known as “group testing”
  - Process in Nebraska (Abdalhamid et al. 2020)
  - Combine together portions of 5 specimens from different individuals into a “pool”
  - Test as if it were a single specimen
    - Pool tests negative: All 5 individuals are negative
    - Pool tests positive: Retest each individual separately to determine who is positive or negative
  - Decrease number of tests, increase testing capacity

- Widely used during pandemic
  - A Shiny App for Pooled Testing
  - 91 countries during 6 months of 2020



- Australia
  - App: 73 separate visits from Sydney!



- Australia
  - App: 73 separate visits from Sydney!

  - Department of Health, Disability and Ageing: "Revised testing framework for COVID-19 in Australia, March 2022"

- Australia
  - App: 73 separate visits from Sydney!

  - Department of Health, Disability and Ageing: "Revised testing framework for COVID-19 in Australia, March 2022"
  - Microbiological Diagnostic Unit Public Health Lab at U. of Melbourne (Chong et al. 2020)

- Australia
  - App: 73 separate visits from Sydney!

  - Department of Health, Disability and Ageing: "Revised testing framework for COVID-19 in Australia, March 2022"
  - Microbiological Diagnostic Unit Public Health Lab at U. of Melbourne (Chong et al. 2020)
- Widely used elsewhere: Blood donations, sexually transmitted infections, congenital infections, animal infections, food safety surveillance, computer networks assessments, flower infection levels

- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)

- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
  - Different pool sizes
  - Other algorithms exist

- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
  - Different pool sizes
  - Other algorithms exist
  - Statistician involvement: Efficiency (expected number of tests per individual) is comparison metric

- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
  - Different pool sizes
  - Other algorithms exist
  - Statistician involvement: Efficiency (expected number of tests per individual) is comparison metric
- During COVID-19 pandemic
  - Over 1 GB of papers published on pooled testing during the first two years of pandemic!

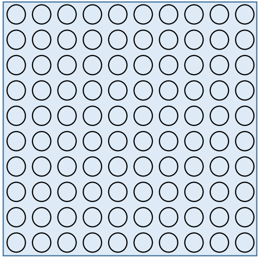
- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
  - Different pool sizes
  - Other algorithms exist
  - Statistician involvement: Efficiency (expected number of tests per individual) is comparison metric
- During COVID-19 pandemic
  - Over 1 GB of papers published on pooled testing during the first two years of pandemic!
  - Two new innovative algorithms developed
    - Shental et al. (2020), Ghosh et al. (2021), Zismanov et al. (2024)
    - Non-statistical journal papers

- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
  - Different pool sizes
  - Other algorithms exist
  - Statistician involvement: Efficiency (expected number of tests per individual) is comparison metric
- During COVID-19 pandemic
  - Over 1 GB of papers published on pooled testing during the first two years of pandemic!
  - Two new innovative algorithms developed
    - Shental et al. (2020), Ghosh et al. (2021), Zismanov et al. (2024)
    - Non-statistical journal papers
  - Use viral load responses rather than binary (positive/negative) responses
  - Use linear model to predict positive/negative

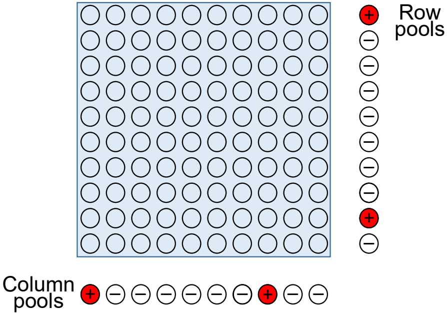
- Algorithmic process
  - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
  - Different pool sizes
  - Other algorithms exist
  - Statistician involvement: Efficiency (expected number of tests per individual) is comparison metric
- During COVID-19 pandemic
  - Over 1 GB of papers published on pooled testing during the first two years of pandemic!
  - Two new innovative algorithms developed
    - Shental et al. (2020), Ghosh et al. (2021), Zismanov et al. (2024)
    - Non-statistical journal papers
  - Use viral load responses rather than binary (positive/negative) responses
  - Use linear model to predict positive/negative
  - Want to avoid retesting in a second stage

- Algorithmic process
    - Dorfman testing for previous example: Stage 1 test in pool, Stage 2 test separately (if needed)
    - Different pool sizes
    - Other algorithms exist
    - Statistician involvement: Efficiency (expected number of tests per individual) is comparison metric
- During COVID-19 pandemic
    - Over 1 GB of papers published on pooled testing during the first two years of pandemic!
    - Two new innovative algorithms developed
        - Shental et al. (2020), Ghosh et al. (2021), Zismanov et al. (2024)
        - Non-statistical journal papers
    - Use viral load responses rather than binary (positive/negative) responses
    - Use linear model to predict positive/negative
    - Want to avoid retesting in a second stage
- Purpose: Examine use of viral load response and linear model prediction with "array testing" algorithm

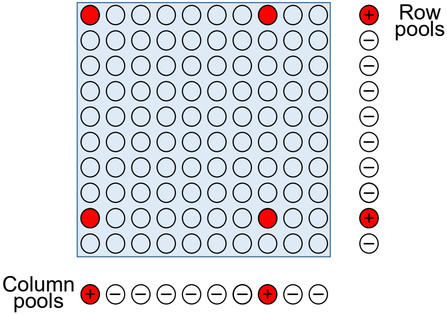
- Array testing



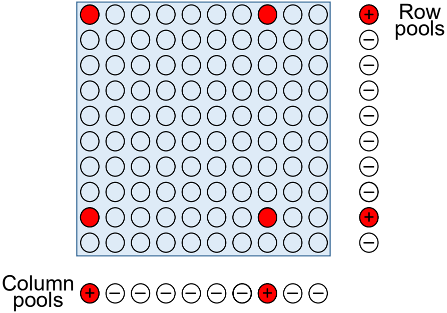
- Array testing



- Array testing



- Array testing



- Key aspect: Test in multiple pools (groups) during first stage to reduce the number of retests in a second stage

- A 3×3 array

	Column1	Column 2	Column 3
Row 1	1	2	3
Row 2	4	5	6
Row 3	7	8	9

- A 3 × 3 array

	Column1	Column 2	Column 3
Row 1	1	2	3
Row 2	4	5	6
Row 3	7	8	9

- Alternative form

	Specimens									Regular
Pools	1	2	3	4	5	6	7	8	9	Array
1	1	1	1	0	0	0	0	0	0	Row 1
2	0	0	0	1	1	1	0	0	0	Row 2
3	0	0	0	0	0	0	1	1	1	Row 3
4	1	0	0	1	0	0	1	0	0	Col. 1
5	0	1	0	0	1	0	0	1	0	Col. 2
6	0	0	1	0	0	1	0	0	1	Col. 3

- A 3 × 3 array

	Column1	Column 2	Column 3
Row 1	1	2	3
Row 2	4	5	6
Row 3	7	8	9

- Alternative form

	Specimens									Regular
Pools	1	2	3	4	5	6	7	8	9	Array
1	1	1	1	0	0	0	0	0	0	Row 1
2	0	0	0	1	1	1	0	0	0	Row 2
3	0	0	0	0	0	0	1	1	1	Row 3
4	1	0	0	1	0	0	1	0	0	Col. 1
5	0	1	0	0	1	0	0	1	0	Col. 2
6	0	0	1	0	0	1	0	0	1	Col. 3

- Pooling matrix: $X_{6\times 9}$ is a matrix of 0's and 1's

- Could we use a linear model to predict positives/negatives rather than going to a second stage?

- Could we use a linear model to predict positives/negatives rather than going to a second stage?
- Define
  - R = Number of rows of array, C = Number of columns of array
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_{RC})'$, a vector of viral loads for the pools; observable
  - $\boldsymbol{\beta} = (\beta_1, ..., \beta_{RC})'$, a vector of true individual viral loads; not observable

- Could we use a linear model to predict positives/negatives rather than going to a second stage?
- Define
  - R = Number of rows of array, C = Number of columns of array
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_{RC})'$, a vector of viral loads for the pools; observable
  - $\boldsymbol{\beta} = (\beta_1, ..., \beta_{RC})'$, a vector of true individual viral loads; not observable
- $E(\boldsymbol{Y}) = \mathrm{X}\boldsymbol{\beta}$
  - A linear model!
  - Pool viral loads are sums of individual viral loads

- Could we use a linear model to predict positives/negatives rather than going to a second stage?
- Define
  - R = Number of rows of array, C = Number of columns of array
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_{RC})'$, a vector of viral loads for the pools; observable
  - $\boldsymbol{\beta} = (\beta_1, ..., \beta_{RC})'$, a vector of true individual viral loads; not observable
- $E(\boldsymbol{Y}) = \mathrm{X}\boldsymbol{\beta}$
  - A linear model!
  - Pool viral loads are sums of individual viral loads
  - β_i
    - Equal 0: Specimen has no virus, negative individual
    - Greater than 0: Specimen has virus, positive individual

- Could we use a linear model to predict positives/negatives rather than going to a second stage?
- Define
  - R = Number of rows of array, C = Number of columns of array
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_{RC})'$, a vector of viral loads for the pools; observable
  - $\boldsymbol{\beta} = (\beta_1, ..., \beta_{RC})'$, a vector of true individual viral loads; not observable
- $E(\boldsymbol{Y}) = \mathrm{X}\boldsymbol{\beta}$
  - A linear model!
  - Pool viral loads are sums of individual viral loads
  - β_i
    - Equal 0: Specimen has no virus, negative individual
    - Greater than 0: Specimen has virus, positive individual
- Fit model to estimate $\boldsymbol{\beta}$
  - Assume $MVN(0, \sigma_y^2 \mathrm{I})$ for $\boldsymbol{Y}$

- Could we use a linear model to predict positives/negatives rather than going to a second stage?
- Define
  - R = Number of rows of array, C = Number of columns of array
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_{RC})'$, a vector of viral loads for the pools; observable
  - $\boldsymbol{\beta} = (\beta_1, ..., \beta_{RC})'$, a vector of true individual viral loads; not observable
- $E(\boldsymbol{Y}) = \mathrm{X}\boldsymbol{\beta}$
  - A linear model!
  - Pool viral loads are sums of individual viral loads
  - β_i
    - Equal 0: Specimen has no virus, negative individual
    - Greater than 0: Specimen has virus, positive individual
- Fit model to estimate $\boldsymbol{\beta}$
  - Assume $MVN(0, \sigma_y^2 \mathrm{I})$ for $\boldsymbol{Y}$
  - RC columns of X (# of specimens) $>$ $R + C$ rows of X (# of pools)

- Could we use a linear model to predict positives/negatives rather than going to a second stage?
- Define
  - R = Number of rows of array, C = Number of columns of array
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_{RC})'$, a vector of viral loads for the pools; observable
  - $\boldsymbol{\beta} = (\beta_1, ..., \beta_{RC})'$, a vector of true individual viral loads; not observable
- $E(\boldsymbol{Y}) = \mathrm{X}\boldsymbol{\beta}$
  - A linear model!
  - Pool viral loads are sums of individual viral loads
  - β_i
    - Equal 0: Specimen has no virus, negative individual
    - Greater than 0: Specimen has virus, positive individual
- Fit model to estimate $\boldsymbol{\beta}$
  - Assume $MVN(0, \sigma_y^2\mathrm{I})$ for $\boldsymbol{Y}$
  - RC columns of X (# of specimens) $>$ $R + C$ rows of X (# of pools)
  - Penalized regression model: non-negative LASSO

- Options based on threshold $c > 0$ (determined by assay manufacturer)

- Options based on threshold $c > 0$ (determined by assay manufacturer)
    - #1: Estimated viral load for specimen, $\hat{\beta}_i$
        - Equal to or larger than threshold declare positive
        - Less than a threshold declare negative

- Options based on threshold $c > 0$ (determined by assay manufacturer)
    - #1: Estimated viral load for specimen, $\hat{\beta}_i$
        - Equal to or larger than threshold declare positive
        - Less than a threshold declare negative
    - #2: Same as #1 but
        - Retest specimens in a second stage for estimates in an indeterminate range
        - Indeterminate range: $0 < \hat{\beta}_i < c$

- Algorithms investigated
  - Array testing with linear model, no retests (option #1)
  - Array testing with linear model, potential retests (option #2)
  - Array testing
  - Dorfman testing

- Algorithms investigated
  - Array testing with linear model, no retests (option #1)
  - Array testing with linear model, potential retests (option #2)
  - Array testing
  - Dorfman testing
- Comparison metric: Efficiency
  - Expected number of tests per individual
  - “Best” algorithm has the lowest value

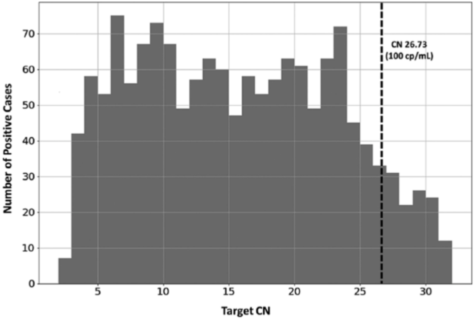
- Algorithms investigated
  - Array testing with linear model, no retests (option #1)
  - Array testing with linear model, potential retests (option #2)
  - Array testing
  - Dorfman testing
- Comparison metric: Efficiency
  - Expected number of tests per individual
  - "Best" algorithm has the lowest value
- Comparison metric: Positive percentage agreement (PPA)
  - Probability of declaring positive given the individual would test as positive
  - Like a sensitivity

- Algorithms investigated
  - Array testing with linear model, no retests (option #1)
  - Array testing with linear model, potential retests (option #2)
  - Array testing
  - Dorfman testing
- Comparison metric: Efficiency
  - Expected number of tests per individual
  - “Best” algorithm has the lowest value
- Comparison metric: Positive percentage agreement (PPA)
  - Probability of declaring positive given the individual would test as positive
  - Like a sensitivity
- Estimate efficiency and PPA
  - No closed form expressions for linear model-based algorithms
  - Use Monte Carlo simulation

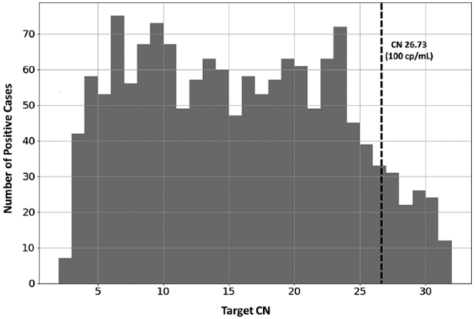
- Monte Carlo simulation summary
    - Simulate individual positive/negative status with Bernoulli(p), p is infection prevalence

- Monte Carlo simulation summary
  - Simulate individual positive/negative status with Bernoulli(p), p is infection prevalence
  - Simulate reverse transcription polymerase chain reaction (RT-PCR) assay testing process

- Monte Carlo simulation summary
  - Simulate individual positive/negative status with Bernoulli(p), p is infection prevalence
  - Simulate reverse transcription polymerase chain reaction (RT-PCR) assay testing process
    - Emulate Abbott Realtime SARS-CoV-2 Assay (Hirschhorn et al. 2021)
    - Individual test result: Sample CN (also known as CT) and convert to viral load

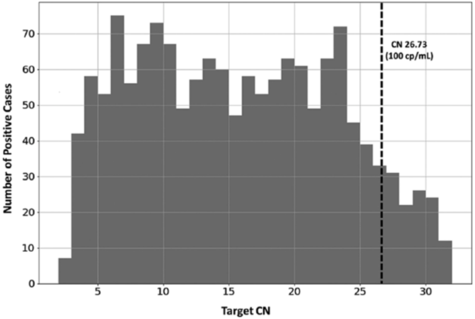


- Monte Carlo simulation summary
  - Simulate individual positive/negative status with Bernoulli(p), p is infection prevalence
  - Simulate reverse transcription polymerase chain reaction (RT-PCR) assay testing process
    - Emulate Abbott Realtime SARS-CoV-2 Assay (Hirschhorn et al. 2021)
    - Individual test result: Sample CN (also known as CT) and convert to viral load



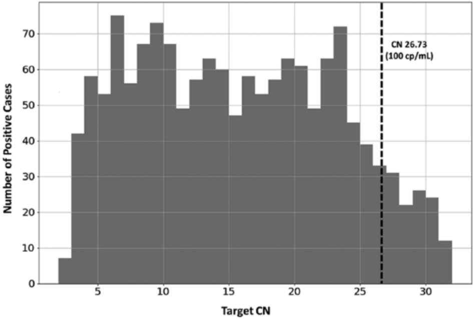
    - Emulate pool test results using Tan et al. (2020) and Arnout et al. (2021)

- Monte Carlo simulation summary
  - Simulate individual positive/negative status with Bernoulli(p), p is infection prevalence
  - Simulate reverse transcription polymerase chain reaction (RT-PCR) assay testing process
    - Emulate Abbott Realtime SARS-CoV-2 Assay (Hirschhorn et al. 2021)
    - Individual test result: Sample CN (also known as CT) and convert to viral load

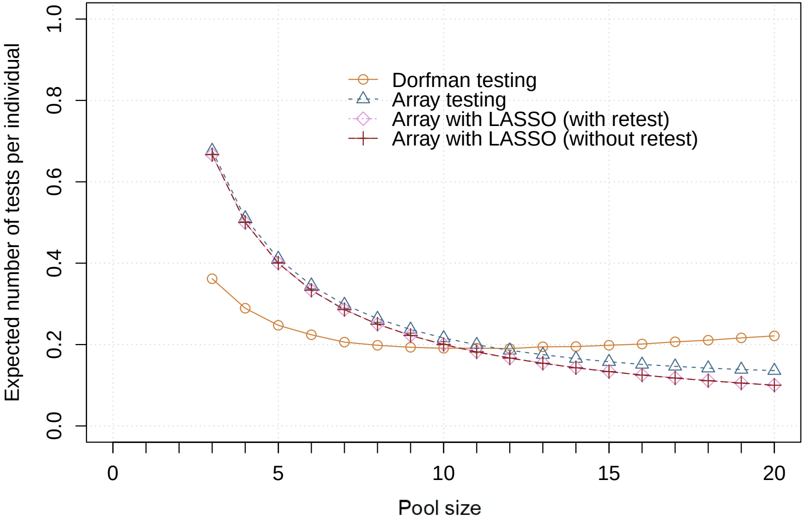


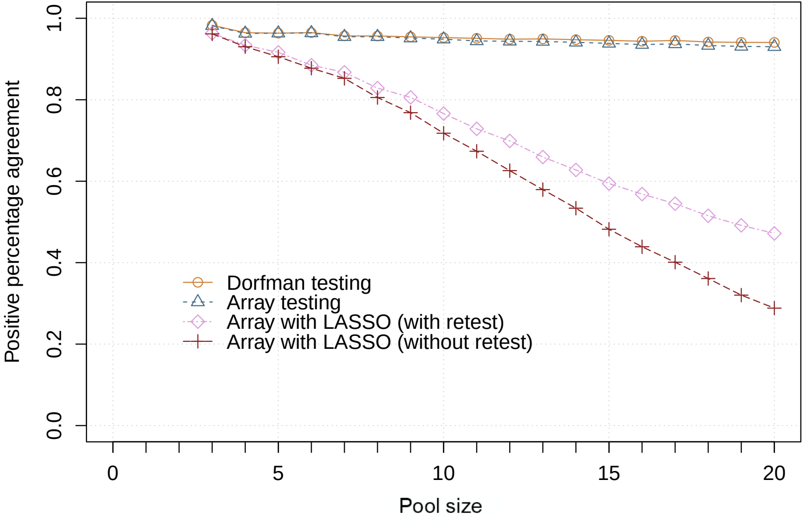
    - Emulate pool test results using Tan et al. (2020) and Arnout et al. (2021)
  - Repeat data simulation process 10,000 times

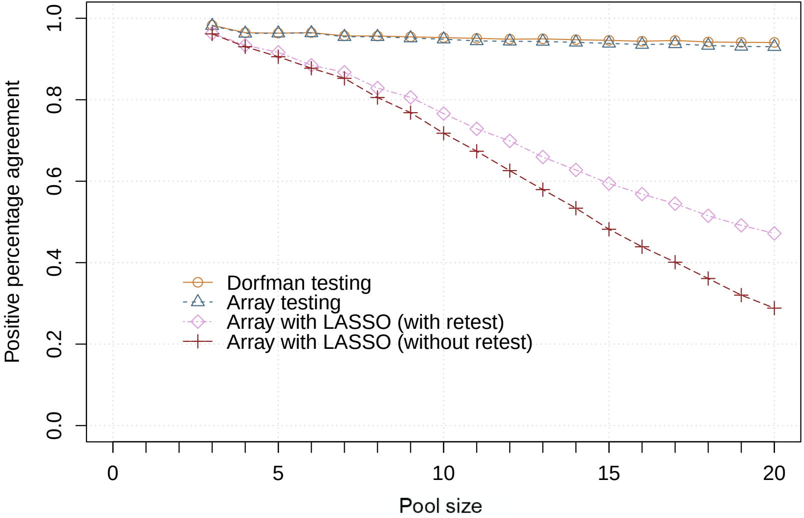
- Monte Carlo simulation summary
  - Simulate individual positive/negative status with Bernoulli(p), p is infection prevalence
  - Simulate reverse transcription polymerase chain reaction (RT-PCR) assay testing process
    - Emulate Abbott Realtime SARS-CoV-2 Assay (Hirschhorn et al. 2021)
    - Individual test result: Sample CN (also known as CT) and convert to viral load



    - Emulate pool test results using Tan et al. (2020) and Arnout et al. (2021)
  - Repeat data simulation process 10,000 times
  - One setting: $p = 0.01$, pool sizes 3 to 20


Dorfman testing
Array testing
Array with LASSO (with retest)
Array with LASSO (without retest)
Expected number of tests per individual
0.0
0.2
0.4
0.6
0.8
1.0
Pool size
0
5
10
15
20


Positive percentage agreement
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Pool size
Dorfman testing
Array testing
Array with LASSO (with retest)
Array with LASSO (without retest)



- Why do the linear model-based algorithms perform poorly?

- Linear model-based algorithms with array testing: Not ready for labs yet!

- Linear model-based algorithms with array testing: Not ready for labs yet!
- Other investigations
  - Different p
  - Each individual put into more groups; e.g., 3D array testing
  - Adaptive LASSO
  - Pooling matrices from other authors

- Linear model-based algorithms with array testing: Not ready for labs yet!
- Other investigations
  - Different p
  - Each individual put into more groups; e.g., 3D array testing
  - Adaptive LASSO
  - Pooling matrices from other authors
- Incorporate statistical inference - $H_0 : \beta_i = 0$ vs. $H_a : \beta_i > 0$

Pooled testing with penalized regression models

Christopher R. Bilder
University of Nebraska–Lincoln
Department of Statistics
chris@chrisbilder.com

Research is supported by NIH grant R01 AI121351

Joint work with
Pranta Das at University of Nebraska-Lincoln,
Joshua M. Tebbs at University of South Carolina, and
Christopher S. McMahan at Clemson University