Received: 22 March 2021 | Accepted: 25 March 2021

DOI: 10.1002/sim.8988

COMMENTARY

Statistics in Medicine WILEY

Discussion on "Is group testing ready for prime-time in disease identification"

Christopher R. Bilder[1] | **Joshua M. Tebbs**[2] | **Christopher S. McMahan**[3]

[1]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[2]Department of Statistics, University of South Carolina, Columbia, South Carolina, USA

[3]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA

Correspondence

Christopher R. Bilder, Department of Statistics, University of Nebraska-Lincoln, Hardin Hall 340, Lincoln, NE 68583, USA.
Email: chris@chrisbilder.com

Funding information

National Institute of Allergy and Infectious Diseases, Grant/Award Number: R01 AI121351; National Science Foundation, Grant/Award Number: OIA-1826715; Office of Naval Research, Grant/Award Number: N00014-19-1-2295

1 | INTRODUCTION

Is group testing ready for prime-time in disease identification? Yes! Prior to the COVID-19 pandemic, group testing (also known as pooled testing and specimen pooling) was widely used in areas including blood donation screening,[1,2] infectious disease testing in animals,[3,4] sexually transmitted infection testing,[5,6] and surveillance for pathogen contamination of food.[7,8] More areas could definitely benefit as well, as pointed out by Haber et al[9] (HMA). The use of group testing for SARS-CoV-2 detection was isolated at the beginning of the pandemic, but early accounts of its successful implementation[10-12] with little if any loss of accuracy led to its widespread use. More than 100 papers detail its implementation since April 2020. A large number of news media accounts, such as in the Washington Post,[13] ABC News,[14] and National Public Radio,[15] likely led to its adoption as well. Large laboratories, like LabCorp[16] and Quest Diagnostics,[17] received Emergency Use Authorizations (EUAs) from the Food and Drug Administration to use their assays with group testing. There is even a Wikipedia[18] web page dedicated to the use of group testing during the pandemic. Therefore, not only is group testing ready for "prime-time", but it is an established "hit" among laboratories around the world to increase testing capacity.

The focus of HMA is on the accuracy of group testing, in particular the first-stage sensitivity for Dorfman testing. We completely agree that quantifying this accuracy is extremely important for laboratories. Where we differ from HMA is how to account for the first-stage sensitivity. In our discussion, we provide comments on the methods used by HMA. Next, we discuss how laboratories choose a group size to insure first-stage sensitivity is not compromised. Relative to this discussion, we provide information about how to find the optimal group size. We conclude with reasons beyond accuracy for why some laboratories may remain hesitant toward group testing. We also conclude with what is next for laboratory implementation of group testing. Because of the worldwide importance of SARS-CoV-2 detection right now, our discussion focuses mostly on this group testing application.

2 | DILUTION

If the first stage of Dorfman testing has reduced sensitivity, the main cause is the "dilution effect." This problem can occur because a smaller portion of each individual specimen is tested in a group than if each was tested separately. Thus, less of a pathogen may be present in a group containing one positive than if the positive specimen was tested alone. Whether reduced sensitivity for the group occurs is dependent on the type, target, and implementation of the assay, in addition to the group size. Laboratories can control some of these factors to prevent the problem from occurring.

HMA advocated using a deterministic model to account for the potential dilution effect. The authors primarily focus on a model proposed by Hwang[19] ($f_H(p, k, d)$) that is a function of the overall infection prevalence (p) in a population of interest, the group size (k), and the amount of dilution (d). Section 7.1 also includes other models that are functions of group size alone. While it is correct to have these expressions as decreasing functions of group size, there is no evidence presented that these particular models describe what truly occurs. This leaves the reader uncertain if the results presented in Sections 6 and 7 occur in actual practice.

To illustrate our concern, we focus on the Hwang model. We are not aware of assays with a sensitivity that is a function of the prevalence. Hwang[19] does not provide any evidence as well. This is especially concerning because nucleic acid amplification tests (NAATs) were not available when Hwang[19] was published in 1976. In addition, examinations of assay sensitivity are always given by an assay's product insert and associated validation materials. While prevalence may be stated, this is done to summarize the specimens within a validation experiment, rather than to imply that sensitivity changes as a function of prevalence (eg, Aptima Mycoplasma genitalium assay[20] and BD Max CT/GC/TV[21]).

For a prevalence of 0.10, Figure 1 evaluates the Hwang model as a function of group size and dilution effect. Most of these dilution settings lead to unrealistic first-stage sensitivities, like for $d = 0.3$ with sensitivities dropping below 0.5. At the very least, it is doubtful that a laboratory would consider using an assay for group testing across such a large range of group sizes when first-stage sensitivities are less than 0.9.

3 | LABORATORY VALIDATION

We emphatically agree with HMA that accuracy should be a top concern whenever a laboratory implements group testing. Others in the group testing literature, such as Kim et al[22] and Hitt et al,[23] have emphasized this point as well. Drawing upon the accuracy concern, HMA proposed a way to validate the group testing process using their dilution models. Unfortunately, this validation could lead to very large sample sizes. In this section, we describe the simpler process that many laboratories use instead to ensure there is little if any loss of accuracy when compared to individual testing.

The most common NAAT approach for SARS-CoV-2 detection is through using real-time reverse transcription polymerase chain reaction (RT-qPCR). The response from the test is a cycle threshold (CT) value. This value represents the number of heating/cooling cycles completed when enough florescence is detected to declare a specimen positive. Each of these cycles results in an approximate doubling of the viral genetic material present. The maximum number of cycles used is dependent on the assay, where common values are between 35 and 40 (eg, the Centers for Disease Control and Prevention assay uses 40 cycles[24]). Thus, a CT value of 30 indicates a specimen was determined to be positive. If not enough florescence is detected prior to the maximum number of cycles, the specimen is declared negative. The amount of florescence during the cycles is tracked with an amplification curve plot, where examples are given by Yelin et al[12] and by Anderson et al[25] (see video demonstrating the testing process at the Nebraska Public Health Laboratory). While the CT is not the viral load for a specimen, it is closely related to it. A higher viral load will lead to a lower CT value (amplification threshold reached earlier due to more virus present), and a lower viral load will lead to a higher CT value (amplification threshold is reached later due to less virus present).

Rather than the process outlined in Section 7 of HMA, laboratories use the spiking procedure briefly described in Section 9. One known positive will be combined with $k - 1$ known negative specimens to evaluate how well a group size of k will work in practice. Because the known positive is diluted by the known negatives, there will likely be less virus present in the grouped specimen than if the positive specimen was tested separately (assuming the same microliters are used for a group and individual test). Thus, it is likely to take more cycles to declare the group positive than the individual positive. Fortunately, this increase in cycles is predictable due to the approximate doubling from the PCR process. Tan et al[26] explained that the increase in CT value should be approximately $\log_2 k$. For example, Abdalhamid et al[10] used a group size of 5 in their validation experiments when combining 1 known positive with 4 known negatives. This suggests CT values should increase by approximately $\log_2 5 = 2.32$. Over 25 groups in the validation experiment, the average CT increase was 2.67 and 2.24 for the assay's two targets. Yelin et al[12] provided another example of this predictable increase in CT values. They progressively doubled their group sizes up to 64, where the same positive specimen was present in each group formed, and displayed the corresponding amplification curves. Based on this process, they provided recommendations for which group sizes SARS-CoV-2 could still be detected.

Because of this predictable trend, low viral load specimens are of most concern to make sure no positive specimens are missed. What can be done to make sure the sensitivity is not reduced? An assay used with group testing should have a low limit of detection. With this type of assay, the maximum number of cycles can be increased whenever group tests are performed. For example, rather than using a maximum of 40 cycles, the Nebraska Public Health Laboratory used a maximum of 45 cycles. To reduce the possibility of an individual false positive, specimens within a positive group were retested using the original cycle threshold limit. Alternatively, because amplification curves can be examined, individual group members can be retested if their group test appears to be approaching the minimum florescence needed to be declared positive.[12] This approach is similar to the early group testing work on *Chlamydia trachomatis* detection by NAAT methods that were used at the time.[27] Other approaches are possible as well. Some laboratories will forgo these small accommodations. For example, this appears to be what LabCorp does based on their EUA. They present a histogram of approximately 150,000 CT values from individual testing and conclude 2.3% of positives may be missed using groups of size 5. It is important to note however that the EUA does not address whether some of these "missed" positives may represent a previous infection due to their high CT values.[28-31]

Laboratories will only implement a group testing algorithm if the group test has the same or similar accuracy as if each specimen were tested separately with the same assay. For this reason, it may be best to refer to first-stage "sensitivity" as first-stage "positive percent agreement" (PPA). Some product inserts, such as the Aptima Combo 2 Assay for *Chlamydia trachomatis* and *Neisseria gonorrhoeae* detection,[32] have switched to using this more precise terminology when validating their assay compared to another for individual testing. Fitzpatrick et al[33] advocated a move to this terminology as well for test validation in general. Therefore, rather than including sensitivity in an expected value of tests expression, it may be best to use PPA.

4 | OPTIMAL DESIGN

Laboratories want to use the most efficient group size possible. This group size will lead to the largest possible increase in testing capacity because the resources saved can be used to test more specimens. As discussed by McMahan et al,[34] choosing the optimal group size is similar to a power study. A power study makes assumptions regarding parameters and determines a sample size such that the power for a hypothesis test of interest is above a threshold. To find an optimal group size for group testing, one needs to make assumptions about parameter values, such as the prevalence and first-stage sensitivity. These values are determined by past testing and/or validation experiments, so that they provide the best set of information available. A number of different parameter values can be used in this process to understand how these assumptions could affect group size and other operating characteristics of interest. A group size is chosen by minimizing an objective function, most often the expected number of tests per individual, among those group sizes that either do not reduce or result in a very small reduction in the first-stage sensitivity. Group sizes after implementation can be adjusted as needed to reflect potential changing prevalences and/or laboratory resources. Even if the optimal group size is not used, it is important to note that Dorfman testing will very likely be more efficient than testing specimens separately unless a significant increase in the prevalence occurs.

As described in HMA, other objective functions are possible. Some include subjective weighting for accuracy measures. Malinosky et al[35] proposed an objective function free from these subjective weights that was the ratio of the expected number of correct classifications to the expected number of tests. Interestingly, Hitt et al[23] showed this objective function generally resulted in the same optimal group sizes as when the expected number of tests was used as an objective function. Even when there were differences, the differences were small and would not be meaningful to a laboratory. A key point made by Hitt et al[23] was the optimal design should be chosen relative to an objective function and then "examine the accuracy associated with it." If higher accuracy was needed, a "new suboptimal testing configuration would be chosen with accuracies that are acceptable."

5 | CONCLUSION

While group testing is widely used for SARS-CoV-2 detection, there remains reluctance by some laboratories to implement it. Why? One reason involves how groups are formed within a laboratory. Small portions of individual specimens need to be extracted and combined for testing. This extraction/combination process may be performed manually by a laboratory technician using a single-channel pippetting instrument. It can be monotonous and even result in repetitive-strain injuries if a considerable number of specimens need to be grouped per day. Automated liquid handling instruments can greatly ease this process and make laboratories more receptive to group testing. A second reason involves laboratory information management systems. These systems may be designed to track only individual test results, rather than group test results. In those situations, we have seen laboratory technicians record group test results by hand rather than have a computer completely track them. While this will work fine in some settings, it is not ideal for others with a high volume of clinical specimens. A third reason is specimen storage and associated logistics become more complicated. For example, a higher stage hierarchical group testing algorithm requires specimens to be capped/re-capped and moved in/out of storage multiple times. Even with these reasons, a very large number of laboratories have been able to overcome them and significantly increase their testing capacity through group testing.

An alternative to HMA's question is "Are non-Dorfman algorithms ready for prime-time?" Since Dorfman testing was conceived in the early 1940s, many algorithms have been developed that are much more efficient and/or have more desirable properties. Too often, we find laboratories think of Dorfman testing *as* group testing, rather than one way to implement it. There have been a few examples of these other algorithms in use for SARS-CoV-2 detection. These examples include Lohse et al[36] used a three-stage hierarchical algorithm and LabCorp's[16] EUA was for array testing. Kim et al[22] and Bilder et al[37] provided nice summaries for these types of algorithms in general, while the `binGroup2` package[38] in R provides computation tools. Statisticians need to get the word out to more laboratories that these other algorithms exist.

Overall infection prevalence is a large factor in determining whether any group testing algorithm is more efficient than testing each specimen separately. Some laboratories experiencing higher SARS-CoV-2 positivity rates deemed group testing to be no longer useful for fear that most groups will test positive.[39] However, this is an example of where non-Dorfman algorithms can make group testing quite useful. Informative group testing takes into account individual-specific probabilities of infection.[40-42] For example, the threshold optimal Dorfman algorithm of McMahan et al[43] provides the simplest implementation. For this algorithm, probabilities of infection for each individual are used to find a threshold for those specimens that are tested separately and those specimens that are tested in groups. Individuals displaying symptoms likely have a higher probability of infection and could be those tested separately. This type of approach has been investigated in Nebraska and implemented in Germany.[44] The end result is an algorithm that allows groups to be used only on those individuals with a low probability of infection and that resolves changing prevalence issues as described in Section 8.1 of HMA.

ACKNOWLEDGEMENTS

This research was supported by Grant OIA-1826715 from the National Science Foundation, Grant N00014-19-1-2295 from the Office of Naval Research, and Grant R01 AI121351 from the National Institute of Allergy and Infectious Diseases.

ORCID

Christopher R. Bilder https://orcid.org/0000-0002-2848-8576
Joshua M. Tebbs https://orcid.org/0000-0002-6762-7241
Christopher S. McMahan https://orcid.org/0000-0001-5056-9615

REFERENCES

1. American Red Cross Infectious disease testing; 2021. https://www.redcrossblood.org/biomedical-services/blood-diagnostic-testing/blood-testing.html. Accessed March 17, 2021.
2. Canadian Blood Services Surveillance report 2019; 2019. https://professionaleducation.blood.ca/en/transfusion/publications/surveillance-report. Accessed March 17, 2021.
3. Laurin E, Thakur K, Mohr P, et al. To pool or not to pool? guidelines for pooling samples for use in surveillance testing of infectious diseases in aquatic animals. *J Fish Dis*. 2019;42:1471-1491.
4. Nebraska Veterinary Diagnostic Center Diagnostic tests & fees; 2021. https://vbms.unl.edu/nvdc-tests-fees. Accessed March 17, 2021.
5. Boobalan J, Dinesha TR, Gomathi S, et al. Pooled nucleic acid testing strategy for monitoring HIV-1 treatment in resource limited settings. *J Clin Virol*. 2019;117:56-60.
6. De Baetselier I, Vuylsteke B, Yaya I, et al. To pool or not to pool samples for sexually transmitted infections detection in men who have sex with men? an evaluation of a new pooling method using the GeneXpert instrument in West Africa. *Sexually Transmit Dis*. 2020;47:556-561.
7. Fahey J, Ourisson P, Degnan F. Pathogen detection, testing, and control in fresh broccoli sprouts. *Nutr J*. 2006;5:1-6.
8. Mester P, Witte A, Robben C, et al. Optimization and evaluation of the qPCR-based pooling strategy DEP-pooling in dairy production for the detection of Listeria monocytogenes. *Food Control*. 2017;82:298-304.
9. Haber G, Malinovsky Y, Albert P. Is group testing ready for prime-time in disease identification? *Stat Med*. 2021.

10. Abdalhamid B, Bilder C, McCutchen E, Hinrichs S, Koepsell S, Iwen P. Assessment of specimen pooling to conserve SARS CoV-2 testing resources. *Am J Clin Pathol*. 2020;153:715-718.
11. Ben-Ami R, Klochendler A, Seidel M, et al. Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection. *Clin Microbiol Infect*. 2020;26:1248-1253.
12. Yelin I, Aharony N, Tamar E, et al. Evaluation of COVID-19 RT-qPCR test in multi sample pools. *Clin Infect Dis*. 2020;71:2073-2078.
13. Gossner O, Gollier C. A temporary coronavirus testing fix: use each kit on 50 people at a time. *The Washington Post*. March 31, 2020. https://www.washingtonpost.com/outlook/2020/03/31/coronavirus-testing-groups. Accessed March 17, 2021.
14. Abdelmalek M. With all eyes on coronavirus testing, some researchers say 'group testing' could make up the shortage. *ABC News*. May 13, 2020. https://abcnews.go.com/Health/eyes-coronavirus-testing-researchers-group-testing-make-shortage/story?id=70658896. Accessed March 17, 2021.
15. Harris R. Pooling coronavirus tests can spare scarce supplies, but there's a catch. *National Public Radio*. July 6, 2020. https://www.npr.org/sections/health-shots/2020/07/06/886886255/pooling-coronavirus-tests-can-spare-scarce-supplies-but-theres-a-catch. Accessed March 17, 2021.
16. LabCorp Emergency use authorization for COVID-19 RT-PCR test; 2020. https://www.fda.gov/media/141948/download. Accessed March 17, 2021.
17. Quest Diagnostics Emergency use authorization for SARS-CoV-2 RNA, qualitative real-time RT-PCR; 2020. https://www.fda.gov/media/136228/download. Accessed March 17, 2021.
18. Wikipedia List of countries implementing pool testing strategy against COVID-19; 2021. https://en.wikipedia.org/wiki/List_of_countries_implementing_pool_testing_strategy_against_COVID-19. Accessed March 17, 2021.
19. Hwang F. Group testing with a dilution effect. *Biometrika*. 1976;63:671-680.
20. Hologic Aptima mycoplasma genitalium assay; 2021. https://www.hologic.com/package-inserts/diagnostic-products/aptima-mycoplasma-genitalium-assay. Accessed March 17, 2021.
21. BD BD max CT/NG/TV; 2021. https://www.bd.com/resource.aspx?IDX=32632. Accessed March 17, 2021.
22. Kim H, Hudgens M, Dreyfuss J, Westreich D, Pilcher C. Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*. 2007;63:1152-1163.
23. Hitt B, Bilder C, Tebbs J, McMahan C. The objective function controversy for group testing: much ado about nothing? *Stat Med*. 2019;38:4912-4923.
24. Centers for Disease Control and Prevention CDC 2019-novel coronavirus (2019-nCoV) real-time RT-PCR diagnostic panel; 2020. https://www.fda.gov/media/134922/download. Accessed March 17, 2021.
25. Anderson J. Group testing for COVID-19, used in Nebraska, seen as 'promising way' to conserve testing supplies. *Omaha World-Herald*. 2020. https://omaha.com/livewellnebraska/group-testing-for-covid-19-used-in-nebraska-seen-as-promising-way-to-conserve-testing/article_123ac7d8-7b85-58bc-b12 d-a52f855d530d.html. Accessed March 17, 2021.
26. Tan J, Omar A, Lee W, Wong M. Considerations for group testing: a practical approach for the clinical laboratory. *Clin Biochem Rev*. 2020;41:79-92.
27. Kacena K, Quinn S, Howell M, Madico G, Quinn T, Gaydos C. Pooling urine samples for ligase chain reaction screening for genital Chlamydia trachomatis infection in asymptomatic women. *J Clin Microbiol*. 1998;36:481-485.
28. Racaniello V. This week in virology 641: COVID-19 with Dr. Anthony Fauci. 2020. https://youtu.be/a_Vy6fgaBPE?t=260. Accessed March 17, 2021.
29. Mandavilli A. Your coronavirus test is positive. maybe it shouldn't be. *The New York Times*. 2020. August 29; https://www.nytimes.com/2020/08/29/health/coronavirus-testing.html. .
30. World Health Organization Nucleic acid testing (NAT) technologies that use real-time polymerase chain reaction (RT-PCR) for detection of SARS-CoV-2; December 14, 2020. https://web.archive.org/web/20210120083427/https://www.who.int/news/item/14-12-2020-who-information-notice-for-ivd-users. Accessed March 17, 2021.
31. World Health Organization Nucleic acid testing (NAT) technologies that use real-time polymerase chain reaction (RT-PCR) for detection of SARS-CoV-2; January 20, 2021. https://www.who.int/news/item/20-01-2021-who-information-notice-for-ivd-users-2020-05. Retrieved March 17, 2021.
32. Hologic Aptima combo 2 assay for CT/NG; 2021. https://www.hologic.com/package-inserts/diagnostic-products/aptima-combo-2-assay-ctng. Accessed March 17, 2021.
33. Fitzpatrick M, Pandey A, Wells C, Sah P, Galvani A. Buyer beware: inflated claims of sensitivity for rapid COVID-19 tests. *The Lancet*. 2021;397:24-25.
34. McMahan C, Tebbs J, Bilder C. Rejoinder reaction: a note on the evaluation of group testing algorithms in the presence of misclassification. *Biometrics*. 2016;72:303-304.
35. Malinovsky Y, Albert P, Roy A. Reader reaction: a note on the evaluation of group testing algorithms in the presence of misclassification. *Biometrics*. 2016;72:299-302.
36. Lohse S, Pfuhl T, Berkó-Göttel B, et al. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. *Lancet Infect Dis*. 2020;20:1231-1232.
37. Bilder C, Iwen P, Abdalhamid B, Tebbs J, McMahan C. Tests in short supply? try group testing. *Significance*. 2020;17:15-16.
38. Hitt B, Bilder C, Schaarschmidt F, Biggerstaff B, McMahan C, Tebbs J. binGroup2: Identification and estimation using group testing; 2021. https://CRAN.R-project.org/package=binGroup2. Accessed March 17, 2021.

39. Wu K. Why pooled testing for the coronavirus isn't working in America. *The New York Times*. 2020. https://www.nytimes.com/2020/08/18/health/coronavirus-pool-testing.html. Accessed March 17, 2021.
40. Bilder C, Tebbs J, Chen P. Informative retesting. *J Am Stat Assoc*. 2010;105:942-955.
41. Black M, Bilder C, Tebbs J. Optimal retesting configurations for hierarchical group testing. *J Royal Stat Soc Ser C (Appl Stat)*. 2015;64:693-710.
42. McMahan C, Tebbs J, Bilder C. Two-dimensional informative array testing. *Biometrics*. 2012;68:793-804.
43. McMahan C, Tebbs J, Bilder C. Informative Dorfman screening. *Biometrics*. 2012;68:287-296.
44. Schneitler S, Jung P, Bub F, et al. Simple questionnaires to improve pooling strategies for SARS-CoV-2 laboratory testing. *Ann Global Health*. 2020;86.

How to cite this article: Bilder CR, Tebbs JM, McMahan CS. Discussion on "Is group testing ready for prime-time in disease identification". *Statistics in Medicine*. 2021;40:3881–3886. https://doi.org/10.1002/sim.8988