**Article Summary Line:** Pooled testing for SARS-CoV-2 is feasible on three different nucleic acid amplification testing platforms, but is accompanied by significant dilution-related, assay-specific, and population-dependent changes in analytical sensitivity and efficiency, which have implications for large-scale screening efforts.

**Running Title:** SARS-CoV-2 Pooled Test Performance

**Keywords:** SARS-CoV-2 Testing, Diagnostic Screening Programs, COVID-19

**Title:** Performance of Nucleic Acid Amplification Tests for Detection of SARS-CoV-2 in Prospectively Pooled Specimens

**Authors:** Hannah Wang, MD; Catherine A. Hogan, MD, MSc; Jacob A. Miller, MD; Malaya K. Sahoo, PhD; ChunHong Huang, MD; Kenji O. Mfuh, PhD; Mamdouh Sibai, BS; James Zehnder, MD; Brendan Hickey, ScM; Nasa Sinnott-Armstrong, ScM; Benjamin A. Pinsky, MD, PhD

**Author Affiliations:** Stanford University School of Medicine, Stanford, CA, USA (H. Wang, C.A. Hogan, J.A. Miller, M.K. Sahoo, C. Huang, J. Zehnder, N. Sinnott-Armstrong, B.A. Pinsky); Stanford University, Stanford (N. Sinnott-Armstrong); Stanford Health Care, Stanford (K.O. Mfuh, M. Sibai, B.A. Pinsky)

**Corresponding Author:**

Benjamin A. Pinsky, MD, PhD

3375 Hillview, Room 2913

Palo Alto, CA 94304

Phone (650) 498-5575

Fax (650) 736-1964

[bpinsky@stanford.edu](mailto:bpinsky@stanford.edu)

**Abstract Word Count:** 147

**Text Word Count:** 3495

**Abstract**

Pooled nucleic acid amplification testing (NAAT) for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has the potential to increase availability of testing at decreased cost. However, the effect of dilution on analytical sensitivity through sample pooling has not been well characterized. We tested 1,648 prospectively-pooled specimens using three different NAATs for SARS-CoV-2: a laboratory-developed real-time reverse transcription PCR (rRT-PCR) assay targeting the *E* gene, and two commercially-available Panther System assays targeting ORF1ab. Positive percent agreement (PPA) of pooled versus individual testing ranged from 71.7%-82.6% for pools of 8, and from 82.9%-100.0% for pools of 4. We developed an independent stochastic simulation model to estimate the effects of dilution on PPA and efficiency of a two-stage pooled rRT-PCR testing algorithm, and validated it using these empiric data. Our model demonstrated that PPA was dependent on the proportion of positive tests, cycle threshold distribution, and assay limit of detection**.**

**Introduction**

The ability of clinical laboratories to meet the demand for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) testing is critical to reduce coronavirus disease 2019 (COVID-19)-related morbidity, mortality, and economic impact. Pooled testing has the potential to decrease the resources required for population-level screening, and can provide valuable data to inform public health policies (*1*,*2*). Several prior experimental and modeling studies have demonstrated the feasibility of pooled SARS-CoV-2 nucleic acid amplification testing (NAAT), in pools of up to 32 individual samples (*3*–*9*). However, the potential increase in efficiency gained by pooled testing is offset by a theoretical dilution-related decrease in analytical sensitivity (*8*,*10*).

Despite this decrease in sensitivity, pooled testing of blood donors for transfusion transmitted infections such as HIV-1 and hepatitis C virus (HCV) has proven to be safe and effective (*11*). This efficacy varies depending on the performance characteristics of the assay, the prevalence of infection, viral load kinetics, and pooling size and strategy. For agents with variable seasonal or geographic prevalence like West Nile virus, many blood banks utilize adaptive risk-based pooling strategies, switching from pooled to individual testing when there is an increase in regional prevalence (*12*). Adapting a similar risk-based pooling strategy for SARS-CoV-2 has the potential to allow for more widespread testing of high-risk populations and asymptomatic critical infrastructure workers, guide aggressive contact-tracing measures, and help direct public health interventions to where they are most needed. However, there are limited prospective data on assay-specific performance characteristics of pooled testing to guide implementation of such a strategy. Furthermore, there is little evidence on parallel test performance of different assays on pooled samples to direct choice of methodology.

In this study, we aimed to evaluate the test performance characteristics of one laboratory-developed and two commercially-available SARS-CoV-2 NAATs on 1648 individual respiratory specimens prospectively grouped in pools of 8 and 4. We used these data to validate a stochastic model to estimate optimal pool size, efficiency, and expected positive percent agreement of a two-stage pooled testing algorithm that takes into account prevalence, viral load distribution, and assay analytical sensitivity.

**Methods**

Clinical Specimens

The Stanford Clinical Virology Laboratory receives samples from tertiary-care academic hospitals and affiliated outpatient facilities in the San Francisco Bay Area, California. Prospective pooling of consecutive nasopharyngeal or oropharyngeal swab specimens submitted for SARS-CoV-2 testing during the morning shift was conducted from June 10 to June 19, 2020 for evaluation of a pool size of 8, and from July 6 to July 23, 2020 for evaluation of a pool size of 4. Samples submitted for testing were collected from symptomatic and asymptomatic inpatients and outpatients, either for clinical care or in the context of COVID-related epidemiologic surveillance studies and drug trials at our institution. As samples from individuals enrolled in these studies and trials were received daily in batches, they were randomly evenly distributed among pools on a daily basis. This was done to preserve the independence between samples in the same pool; these samples had not been tested prior to receipt in our laboratory, and were otherwise treated identically to non-research samples. Non-research samples were otherwise assigned to pools consecutively. Additional laboratory-wide data on proportion of tests positive and Ct value distribution were obtained from all specimens (n=74,162) tested from March 1 through June 24, 2020. This study was conducted with Stanford institutional review board approval (protocol 48973), and individual consent was waived.

Pool size determination

In this study, an initial pool size of 8 was selected based on pilot experiments with pool sizes ranging from 4 to 10 (B. Pinsky, unpub. data), and the logistical consideration that pooling in multiples of 4 would be more efficient for the robotic liquid handlers in our laboratory. After review of the test performance characteristics of 8-sample pooling in conjunction with the results of an independent stochastic simulation model, additional testing was performed to evaluate a pool size of 4 to generate empiric data for further model validation. Subset analyses of first tests versus follow-up tests were conducted by retrospectively assigning pools to one of the two groups based on the status of the positive sample(s) in that pool. Pools containing positive samples belonging to both groups were excluded from this analysis. To validate the performance of the model for additional pool sizes, an external in silico dataset was obtained based on pool sizes of 3 and 5. The in silico analysis was performed according to FDA recommendations, which are additionally described in the Technical Appendix (*13*).

Sample Pooling, Extraction, and Nucleic Acid Amplification Testing (NAAT)

Pools were constructed prior to nucleic acid extraction by combining 500 µL from each of the individual samples. For a pool size of 8, this resulted in a total volume of 4 mL and a dilution factor of 1:8. For a pool size of 4, this resulted in a total volume of 2 mL and a dilution factor of 1:4.

Subsequently, total nucleic acids were extracted from 500 µL taken from each pool and each individual specimen on the QIAsymphony using the QIAsymphony DSP Virus/Pathogen Midi Kit (both from Qiagen, Germantown, MD), and eluted into 60 µL of buffer AVE according to manufacturer’s instructions. Real-time reverse transcription PCR (rRT-PCR) was performed using an emergency use authorization (EUA) laboratory-developed assay targeting the *E* gene on the Rotor-Gene Q instrument (Qiagen) as previously described (*14*–*16*), with pooled samples tested on the same run as the component individual samples. A Ct result between 40-45 was considered an indeterminate result, which was adjudicated by repeat testing and resulted as positive if reproducible with an acceptable amplification curve. Specimens were only reported as negative if internal control human *RNase P* was detected at a Ct of less than 35.

On the same day as QIAsymphony extraction, another 500 µL from each pool was transferred to a Hologic Panther Specimen Lysis Tube, and tested using the Panther Fusion SARS-CoV-2 assay (Hologic, Marlborough, MA) and Panther Aptima SARS-CoV-2 assay (Hologic), per manufacturer recommendations (*17*,*18*). In addition to the manufacturer-set cutoff, receiver operating characteristic (ROC) curve analysis of pooled relative light unit (RLU) values, with individual test results as the reference method, was used to determine the optimal RLU discrimination threshold. A focused electronic medical record review was conducted for all samples.

Statistical Analysis

ROC curve analysis was conducted using R package “pROC” (*19*). Positive percent agreement (PPA) and negative percent agreement (NPA) were calculated using individual testing as the reference method, and were reported with exact (Clopper-Pearson) 95% confidence intervals (95% CI) (*20*). Passing-Bablok regression was used to compare Ct values of the individuallaboratory-developed test (LDT), pooledLDT, and pooled Panther Fusion assays. The 95% CIs of slope, intercept, and bias were calculated using an ordinary non-parametric bootstrap resampling method with default parameters in R package “mcr”. Paired t-tests were used to compare the mean differences between paired Ct values among different assays. A student’s t-test was used to compare the mean difference between internal control *RNase P* Ct values in false negative and true negative pools. All comparisons were two-sided with Type I error set at 0.05. We used the laboratory-wide Ct value distribution and a separate LoD experiment to develop a stochastic simulation model to estimate PPA and efficiency for a two-stage pooled testing algorithm, which was subsequently validated using the independent empiric pools of 8 and pools of 4 data, as well as in silico pools of 5 and pools of 3 data. The methods used to develop this model are described in the Technical Appendix.

**Results**

Assay Comparisons for Pools of 8

To evaluate a pool size of 8, a total of 112 pools from 896 samples were each tested on three different NAAT platforms (Table 1). Two pools resulted as invalid, one on the Panther Fusion assay (0.9%), and one on the Panther Aptima assay (0.9%), and were excluded from subsequent analysis. All 16 individual samples in these two pools tested negative. The remaining 110 pools contained 880 individual samples. Four samples were tested in duplicate in two different pools with identical results. Among the 880 individual samples, 58 (6.6%) were positive, with a median Ct value of 31.4 (IQR 22.1-35.5). First-time diagnostic specimens had a higher median Ct value than follow-up tests (Table 2). ROC curve analysis for the Panther Aptima revealed an optimal cutoff between 343 and 393 RLUs; a cutoff of 350 was chosen as the nearest round number (Panther Aptima-350, Appendix Table 1, Appendix Figure 1).

Among the tested pools of 8, 41.8% (46/110) contained at least one positive sample. The positive pools were comprised of 36 pools with 1 positive sample, 9 pools with 2 positive samples, and 1 pool with 4 positive samples (Table 3). There were 3 false positive pools, one on each platform, in which each of the individual samples tested negative. The overall PPA of pooled testing ranged from 71.7% to 82.6%, while NPA ranged from 98.4% to 100.0% (Table 4). The 14 pools containing positive first-time diagnostic samples had higher PPA than the 28 pools containing positive follow-up tests by LDT (Appendix Table 3).

There were 16 total pools for which at least one method was falsely negative. Except for the one pool containing four positive specimens, which was not detected by Panther Aptima using the manufacturer’s cutoff (Panther Aptima-M), the remaining 15 false negative pools each contained only a single positive specimen. For all missed pools, the Ct value of the individual positive sample was greater than 34 cycles (median 36.6, IQR 35.5-37.7, Figure 1). Among individual positive specimens in the pools of 8 dataset, 22 (37.9%) had Ct>34. For the LDT, Panther Fusion, Panther Aptima-M, and Panther Aptima-350, 13/22 (59.1%), 11/22 (50.0%), 15/22 (68%), and 8/22 (36.4%) were falsely negative. Each of these false negative samples was collected from known symptomatic or convalescent patients being monitored for viral clearance; none were initial diagnostic specimens. The pooledLDT *RNase P* internal control Ct values were similar in false negative (mean 23.5, 95% CI 22.7-24.3) and true negative (mean 23.4, 95% CI 22.7-24.1, p=0.7) pools.

Linearity Studies for Pools of 8

For pools containing only a single positive sample, the pooled rRT-PCR assays showed positive systematic bias when compared to the individualLDT assay, as evidenced by the Passing-Bablok regression intercept value being greater than zero. Mean bias between pooled and individual Ct values was 3.4 cycles (95% limits of agreement 1.2-5.6, p<0.001) by LDT and 4.0 cycles (0.0-8.0, p<0.001) by Panther Fusion (Figure 2). Panther Fusion showed negative proportional bias when compared to both individual and pooledLDT assays, as evidenced by Passing-Bablok regression slopes with 95% confidence intervals that do not contain 1. This is additionally highlighted in the Bland-Altman plots, which demonstrate that at higher Ct values, Panther Fusion outperforms the LDT.

Model Estimates

The modeled PPA estimate is sensitive to the input parameters of 1) proportion of positive tests, 2) assay analytical sensitivity, and 3) viral load distribution. The analytical sensitivity of the assay is approximated in this model by the Ct value corresponding to the probability of detecting 95% of true positives, otherwise known as the 95% LoD. Specimens with Ct beyond the LoD are assigned a decreasing probability of detection based on a probit regression curve, the shape of which was determined in the initial validation of the LDT assay (Appendix Figure 5). The viral load distribution of the tested population is approximated in this model by the proportion of samples with Ct greater than the LoD. This makes the model output independent of the actual LoD Ct value itself, allowing for the model to be used across different rRT-PCR assays.

If the assay analytical sensitivity is kept constant, but the tested population changes such that a greater proportion have a Ct value beyond the 95% LoD, PPA decreases (Figure 3A). Conversely, if the patient population is kept constant, but assay analytical sensitivity increases (ie. moving from lower Ct LoD to higher Ct LoD), PPA increases (Figure 4A). However, if assay analytical sensitivity changes while the tested population shifts accordingly such that it retains the same proportion Ct > LoD, then the PPA stays constant (Appendix Figure 6). In contrast, the average expected tests per sample is almost entirely determined by pool size and prevalence, whereas analytical sensitivity (LoD Ct) and the underlying Ct distribution minimally impact efficiency due to small absolute numbers of false positive pools (Figures 3B, 4B). To achieve a 5% absolute difference in efficiency with an increase in LoD Ct from 32 to 40, a prevalence of 25% would be required.

Both PPA and tests per sample are highly dependent on pool size and prevalence of infection. As prevalence increases, PPA can counterintuitively increase with larger pool sizes as there is a greater likelihood of having more than one positive sample in a given pool, which would be expected to increase PPA. Similarly, test efficiency can decrease with larger pool sizes as the likelihood of deconvoluting a positive pool increases. Estimated PPA and average tests per sample for inputs of percentage of positive tests 0.1%-15.0% and proportion of samples with Ct value above LoD ranging from 5%-30% are available in Appendix Table 4.

Model Sensitivity Analyses and Validation

One-way deterministic and probabilistic sensitivity analyses incorporating uncertainty in the underlying model assumptions of dilutional effect and probit regression shape demonstrate a moderate (±2-7%) impact upon PPA, which is more pronounced with larger pool sizes and proportion of Ct values above the LoD (Appendix Figure 7A, 7B). In contrast, these parameters have a much smaller impact on testing efficiency (Appendix Figure 8A, 8B). The 95% confidence intervals for the empirically determined and modeled PPAs overlapped for the majority of the evaluated empiric datasets, although it overestimated PPA for the LDT follow-up tests only subset (Figure 5). For the in silico validation data, the modeled PPA was similar for pool sizes of 5 and 3, despite in silico data analysis predicting a higher PPA for pools of 3. Modeled testing efficiency was actually slightly higher for pools of 3 than pools of 5, likely due to the high prevalence of 19.1% in this dataset (Appendix Table 3).

**Discussion**

In this study, over 1600 samples were tested in pool sizes of 8 and 4 by three different SARS-CoV-2 platforms, with pooled testing demonstrating diminished positive percent agreement relative to individual samples. False negative results occurred exclusively in pools containing samples with low estimated viral load (Ct value >34). Overlapping confidence intervals in PPA and NPA at each pool size suggest that the lower test performance is inherent to the pooling process itself rather than the assay. While Panther Fusion Ct values were on average higher than those ofthe LDT, the negative proportional bias suggests that at very low estimated viral loads (Ct >36), the Panther Fusion outperformed theLDT. This may be due to the different targets of amplification (*E* gene vs. ORF1ab) or PCR reaction efficiency. These subtle differences between the two assays highlight the method-dependent nature of test performance, a variable that cannot be anticipated, and therefore is not explicitly accounted for in most statistical models of pooled testing. As such, method comparison studies should be performed prior to large-scale implementation of any pooled testing strategies, especially those that use different platforms for the pooled and individual stages of testing.

The findings in this study contrast with those of a recent study that concluded pooling in groups of 8 did not compromise test performance (*5*). This may be explained by differences in patient population, higher proportion of positive pools and rRT-PCR result interpretation. Another recent study of artificially constructed pools reported no significant decrease in sensitivity in pools of up to 32 samples (*3*). This is likely explained by the relatively low starting Ct values of their individual positive samples; none exceeded a Ct of 30. However, they, and other experimental studies have shown empirical increases in pooled Ct values directly proportional to dilution factor, a relationship that was also observed in our study (*3*,*4*,*9*).

These differences highlight the effect of viral load distribution and assay analytical sensitivity on pooled test performance, both of which should be taken into account when choosing pool size and diagnostic assay. While samples with a Ct greater than 33 have not been reported to produce cultivable virus in convalescing COVID-19 patients (*21*), more than 15% of first-time diagnostic specimens in our laboratory were detected at a Ct of 35 or above. Griesemer et al. described a similar proportion of weak positive samples with high Ct values at their public health department virology lab in New York (*10*). Assays with lower analytical sensitivity may miss specimens with late Ct values, for which the potential associated burden of onward transmission is currently unclear.

The stochastic model in this study demonstrated that expected PPA between pooled and individual rRT-PCR assays was highly dependent on assay analytical sensitivity (represented by 95% LoD), viral load distribution of test-positive patients (represented by proportion Ct >LoD), pool size, and disease prevalence (represented by proportion of tests positive). The model outputs were not always intuitive; larger pool sizes were not always less sensitive or more efficient. With increased prevalence, larger pool sizes were more sensitive as they were more likely to contain more than one positive sample per pool. They were also less efficient as a larger proportion were positive, requiring deconvolution.

The model output was largely independent of the actual LoD and viral load-to-Ct value relationship of a given assay, making it generalizable across different rRT-PCR assays. The only input parameters it requires are proportion of tests positive and proportion of samples with Ct > LoD, both of which should be readily available to any laboratories conducting clinical testing. As such, future studies on the sensitivity of pooled testing strategies should report these parameters.

Prior models of pooled testing strategies for SARS-CoV-2 have primarily examined the effect of pool size and prevalence on testing efficiency, but have not addressed the expected decrement in assay sensitivity that accompanies a putative increase in efficiency (*6*,*22*). Those that have examined sensitivity did not explicitly model the effect of variable viral load distribution of test-positive patients, a parameter that can vary based on the underlying patient population (asymptomatic vs. symptomatic, severe vs. non-severe), purpose of testing (diagnostic vs. follow-up), and specimen type (*8*,*23*–*27*). Additionally, prior modeling studies and in silico analyses have mostly used the assay’s Ct cutoff, assuming 100% detection below the cutoff, and 0% detection above it. In contrast, our model incorporates the probabilistic nature of detection at and above the LoD, which better approximates reality.

Our approach is limited by the generalizability of the probit regression shape and equation estimating dilutional effect, as demonstrated by the variability seen on probabilistic and deterministic sensitivity analysis. Furthermore, the model assumes that the PCR is 100% efficient, and that it is devoid of any proportional bias between individual and pooled tests. Additionally, the model may underestimate PPA and efficiency of pooled testing if samples in each pool are not independent; placing samples with higher pre-test probability in the same pool would decrease the total number of positive pools while increasing the likelihood of detection. Indeed, this feature could be leveraged by pooling specimens from individuals in the same household or social distancing pod, such as coworkers on the same shift, or students sharing a classroom. These factors, among others, may be the reasons for which the probabilistic sensitivity analysis confidence intervals often did not contain the empiric point estimate in our validation data. These unaccounted-for factors may limit the ability of the model to provide a reliable point estimate.

The strengths of this study include its relatively large sample size, prospective rather than experimental construction of pools, and assessment of two different pool sizes. It also compared three different SARS-CoV-2 assays, two of which are commercially available on highly automated platforms suitable for large-scale testing. The study is limited by its assessment of only a two-stage pooling strategy. An additional limitation includes selection bias, as the proportion of positive tests in the study specimens was higher due to the inclusion of follow-up samples from known COVID-19 patients enrolled in clinical research studies. Finally, test performance may also vary depending on specimen collection medium, which we did not assess in this study (*10*).

In conclusion, a two-stage pooled testing strategy for detection of SARS-CoV-2 by nucleic acid amplification is feasible and has the potential to significantly increase testing capacity. However, increased pool size and efficiency can compromise positive percent agreement. More studies examining early viral load kinetics and infectiousness are needed to fully evaluate the risks versus benefits of pooled testing. We provide a model to predict optimal pool size and associated expected positive percent agreement based on limit of detection, Ct value distribution, and proportion of positive tests. If this model can be externally validated, it may be useful in guiding SARS-CoV-2 pooled testing in other laboratories, and as part of an adaptive risk-based strategy.

**Acknowledgements**

We sincerely thank the Stanford Clinical Virology Laboratory staff for their dedication and commitment to patient care in the face of unprecedented challenges presented by the COVID-19 pandemic. We also thank Hologic, Inc. for graciously providing in silico data from their pooling evaluations.

**Disclosures**

B.H. is an employee of Google LLC.

**Financial Support**

Work done by B.H. in personal time. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of Google LLC. N.S.-A. is funded by a Stanford Graduate Fellowship.

**Author Bio**

Dr. Hannah Wang is a resident physician at Stanford Hospital in the Department of Anatomic and Clinical Pathology. Her clinical and research interests focus on infectious disease molecular diagnostics and the application of machine-learning algorithms and other computational methods to effectively leverage those diagnostics for patient care.

**References**

1. Hogan C, Sahoo M, Pinsky B. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. JAMA. 2020;323(19):1967–9.

2. Wacharapluesadee S, Kaewpom T, Ampoot W, Ghai S, Khamhang W, Worachotsueptrakun K, et al. Evaluating the efficiency of specimen pooling for PCR-based detection of COVID-19. J Med Virol 2020 Jul 1 [date cited]. Epub 2020 May 13.

3. Yelin I, Aharony N, Tamar ES, Argoetti A, Messer E, Berenbaum D, et al. Evaluation of COVID-19 RT-qPCR test in multi-sample pools. Clin Infect Dis. 2020 Jul 10 [date cited]. Epub 2020 May 2.

4. Abdalhamid B, Bilder CR, McCutchen EL, Hinrichs SH, Koepsell SA, Iwen PC. Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources. Am J Clin Pathol. 2020;153(6):715–8.

5. Ben-Ami R, Klochendler A, Seidel M, Sido T, Gurel-Gurevich O, Yassour M, et al. Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection. Clin Microbiol Infect. 2020 Jul 13 [date cited]. Epub 2020 Jun 8.

6. Eberhardt JN, Breuckmann NP, Eberhardt CS. Multi-Stage Group Testing Improves Efficiency of Large-Scale COVID-19 Screening. J Clin Virol. 2020;128(104382).

7. Lohse S, Pfuhl T, Berkó-Göttel B, Rissland J, Geibler T, Gärtner B, et al. Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. Lancet Infect Dis. 2020;3099(20):2019–20.

8. Pilcher CD, Westreich D, Hudgens MG. Group Testing for SARS-CoV-2 to Enable Rapid Scale-Up of Testing and Real-Time Surveillance of Incidence. J Infect Dis. 2020 Jul 16 [date cited]. Epub 2020 Jun 27.

9. Eis-Hübinger AM, Hönemann M, Wenzel JJ, Berger A, Widera M, Schmidt B, et al. Ad hoc laboratory-based surveillance of SARS-CoV-2 by real-time RT-PCR using minipools of RNA prepared from routine respiratory samples. J Clin Virol. 2020;127(April).

10. Griesemer SB, Slyke G Van, George K St. Assessment of sample pooling for clinical SARS-CoV-2 testing. bioRxiv. 2020 Jul 12 [date cited]. Epub 2020 May 26.

11. Busch MP, Kleinman SH. Nucleic acid amplification testing of blood donors for transfusion-transmitted infectious diseases. 2000;40(February):143–59.

12. Custer B, Tomasulo PA, Murphy EL, Caglioti S, Harpool D, McEvoy P, et al. Triggers for switching from minipool testing by nucleic acid technology to individual-donation nucleic acid testing for West Nile virus: Analysis of 2003 data to inform 2004 decision making. Transfusion. 2004;

13. Food and Drug Administration. Molecular Diagnostic Template for Laboratories. 2020 July 28 [cited 2020 Sep 18). https://www.fda.gov/media/135658/download.

14. Food and Drug Administration. EUA Summary: SARS-CoV-2 RT-PCR Assay (Stanford Health Care Clinical Virology Laboratory). 2020 April 08 [cited 2020 Jul 11]. https://www.fda.gov/media/136818/download.

15. Bulterys PL, Garamani N, Stevens B, Sahoo MK, Huang CH, Hogan CA, et al. Comparison of a laboratory-developed test targeting the envelope gene with three nucleic acid amplification tests for detection of SARS-CoV-2. J Clin Virol. 2020;129(April):10–3.

16. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020;25(3):1–8.

17. Food and Drug Administration. EUA Summary: SARS-CoV-2 Assay (Panther Fusion System). 2020 April 24 [cited 2020 Jul 11]. https://www.fda.gov/media/136156/download.

18. Food and Drug Administration. EUA Summary: Aptima SARS-CoV-2. 2020 May 14 [cited 2020 Jul 11]. https://www.fda.gov/media/138096/download.

19. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12.

20. Altman D, Machin D, Bryant T, Gardner M, Editors. Statistics With Confidence. 2nd ed. British Medical Journal. London: British Medical Journal Books; 2000.

21. La Scola B, Le Bideau M, Andreani J, Hoang VT, Grimaldier C, Colson P, et al. Viral RNA load as determined by cell culture as a management tool for discharge of SARS-CoV-2 patients from infectious disease wards. Eur J Clin Microbiol Infect Dis. 2020;39(6):1059–61.

22. Aragón-Caqueo D, Fernández-Salinas J, Laroze D. Optimization of group size in pool testing strategy for SARS-CoV-2: A simple mathematical model. J Med Virol. 2020 Jul 13 [date cited]. Epub 2020 Apr.

23. Verdun CM, Fuchs T, Harar P, Elbrächter D, Fischer DS, Berner J, et al. Group testing for SARS-CoV-2 allows for up to 10-fold efficiency increase across realistic scenarios and testing strategies. medRxiv. 2020 Jul 15 [date cited]. Epub 2020 May 5.

24. Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect Dis. 2020;20(4):411–2.

25. Xu T, Chen C, Zhu Z, Cui M, Chen C, Dai H, et al. Clinical features and dynamics of viral load in imported and non-imported patients with COVID-19. Int J Infect Dis. 2020;94(300):68–71.

26. To KKW, Tsang OTY, Leung WS, Tam AR, Wu TC, Lung DC, et al. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. Lancet Infect Dis. 2020;20(5):565–74.

27. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020;581(7809):465–9.

**Tables**

**Table 1.** Summary of nucleic acid amplification tests and pooled testing strategies evaluated, by testing platform

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Gene target(s) | Internal control | Method | Strategy | References |
| LDT | *E* | *RNase P* | rRT-PCR | Pools of 8\*, Pools of 4\* | (*1*,*14*–*16*) |
| Panther Fusion | ORF1ab | Reagent Spike-in | rRT-PCR | Pools of 8\*, Pools of 5†, Pools of 3† | (*17*) |
| Panther Aptima-M | ORF1ab | Reagent Spike-in | TMA | Pools of 8 with manufacturer-set RLU cutoff\* | (*18*) |
| Panther Aptima-350 | ORF1ab | Reagent Spike-in | TMA | Pools of 8 with RLU cutoff of >350\*‡ | (*18*) |
| \*Pooled testing strategy assessed empirically at Stanford Clinical Virology Laboratory, with individual samples evaluated by LDT.  †Pooled testing strategy assessed via *in-silico* sensitivity analysis, with individual samples evaluated by Panther Fusion.  ‡ Panther Aptima RLU cutoff of 350 selected based on receiver operating characteristic curve (Appendix Figure 1).  LDT, lab developed test; *E* gene, gene encoding SARS-CoV-2 envelope protein; *RNase P*, gene encoding ribonuclease P; rRT-PCR, real-time reverse transcription polymerase chain reaction; ORF1ab*,* gene encoding SARS-CoV-2 replicase polyprotein 1 ab;TMA, transcription-mediated amplification; RLU, relative light unit; Panther Aptima-M, Panther Aptima with manufacturer-set relative light unit cutoff; Panther Aptima-350, Panther Aptima with relative light unit cutoff of >350 considered positive. | | | | | |

**Table 2.** Proportion of positive SARS-CoV-2 tests with median Ct values in pooled testing and laboratory-wide clinical testing datasets, subset by testing indication

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | | No. positive samples/No. total samples (%) | | | Median Ct value (IQR) | | |
| All | First | Follow-up | All | First | Follow-up |
| Pools of 8\* | | 58/880 (6.6%) | 24/657 (3.7%) | 34/223 (15.2%) | 31.4 (22.1-35.5) | 24.4 (18.4-33.1) | 34.1 (29.0-36.8) |
| Pools of 4† | | 38/768 (4.9%) | 28/491 (5.7%) | 10/277 (3.6%) | 29.3 (20.3-33.9) | 27.5 (19.4-32.6) | 32.2 (24.9-34.5) |
| Hologic‡ | | 10000/52272 (19.1%) | - | - | 26.2 (20.7-32.6) | - | - |
| Lab-wide§ | | 1358/74162 (1.8%) | 1109/66070 (1.7%) | 249/8092 (3.1%) | 28.5 (23.0-34.3) | 27.2 (22.2-32.4) | 34.2 (29.0-37.4) |
|  | March | 555/8896 (6.2%) | 489/8557 (5.7%) | 66/339 (19.5%) | 26.7 (21.9-31.5) | 26.4 (21.8-31.2) | 28.6 (22.6-35.2) |
|  | April | 518/22671 (2.3%) | 404/21167 (1.9%) | 114/1504 (7.5%) | 30.6 (24.8-36.0) | 28.8 (22.7-34.6) | 35.4 (32.9-38.0) |
|  | May | 172/21833 (0.8%) | 136/19505 (0.7%) | 36/2328 (1.5%) | 27.5 (23.3-34.7) | 26.1 (22.5-31.3) | 35.4 (30.4-37.3) |
|  | June | 113/20762 (0.5%) | 80/16841 (0.5%) | 33/3921 (0.84%) | 28.2 (21.2-33.6) | 27.4 (21.3-32.7) | 30.6 (20.2-34.4) |
| \*Pools of 8 specimens were tested in our clinical laboratory from 6/10/2020-6/19/2020.  †Pools of 4 specimens were tested in our clinical laboratory from 7/6/2020-7/23/2020.  ‡Hologic data set is comprised of specimens tested clinically by Panther Fusion from 3/1/2020-7/31/2020 at two different external sites. These data were used to perform in silico sensitivity analysis to evaluate pool sizes of 3 and 5.  §Laboratory-wide samples data set is comprised of clinical specimens from 3/1/2020-6/24/2020.  No., number; Ct, cycle threshold; IQR, interquartile range. | | | | | | | |

**Table 3.** Results of 8-sample pooled testing, by testing platform and number of positive specimens per pool (n=110)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pooled Testing | | | | Individual Testing | | Total no. pools |
|  | LDT | Panther Fusion | Panther Aptima-M | Panther Aptima-350 | Positive  (no. 1PP, no. >1PP) | Negative |  |
|  | Pos | Pos | Pos | Pos | 30 (21, 9) | 0 | 30 |
|  | Pos | Pos | - | Pos | 2 (1, 1) | 0 | 2 |
|  | Pos | Pos | - | - | 0 (0, 0) | 0 | 0 |
|  | Pos | - | Pos | Pos | 0 (0, 0) | 0 | 0 |
|  | Pos | - | - | Pos | 0 (0, 0) | 0 | 0 |
|  | Pos | - | - | - | 1 (1, 0) | 1\* | 2 |
|  | - | Pos | Pos | Pos | 2 (2, 0) | 0 | 2 |
|  | - | Pos | - | Pos | 1 (1, 0) | 0 | 1 |
|  | - | Pos | - | - | 0 (0, 0) | 1† | 1 |
|  | - | - | Pos | Pos | 2 (2, 0) | 0 | 2 |
|  | - | - | - | Pos | 1 (1, 0) | 1‡ | 2 |
|  | - | - | - | - | 7 (7, 0) | 61 | 68 |
| No. positive pools | 34 | 36 | 34 | 39 | 46 (36, 10) | - | - |
| No. negative pools | 76 | 74 | 76 | 71 | - | 64 | - |
| Total no. pools | 110 | 110 | 110 | 110 | - | - | 110 |
| \*False positiveLDT Ct value was 37.5.  †False positive Panther Fusion Ct value was 38.8.  ‡False positive Panther Aptima-350 RLU value was 434.  LDT, lab developed test; Panther Aptima-M, Panther Aptima with manufacturer-set relative light unit cutoff; Panther Aptima-350, Panther Aptima with relative light unit cutoff of >350 considered positive; no., number; 1PP, 1 positive specimen in pool; >1PP, 2 or more positive specimens in pool; Pos, positive; -, negative; Ct, cycle threshold; RLU, relative light unit. | | | | | | | |

**Table 4.** Performance characteristics and efficiency of 8-sample and 4-sample pooled testing, by testing platform (n=302)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Testing platform | Pool size | PPA (95% CI) | NPA (95% CI) | Percent of pools positive | Average tests run per sample |
| LDT | 8 | 71.7% (56.5%-84.0%) | 98.4% (91.5%-100.0%) | 30.9% | 0.434 |
| Panther Fusion | 8 | 76.1% (61.2%-87.4%) | 98.4% (91.5%-100.0%) | 32.7% | 0.452 |
| Panther Aptima-M | 8 | 73.9% (58.9%-85.7%) | 100.0% (94.3%-100.0%) | 30.9% | 0.434 |
| Panther Aptima-350 | 8 | 82.6% (68.6%-92.2%) | 98.4% (91.5%-100.0%) | 34.5% | 0.470 |
| LDT | 4 | 94.3% (80.8%-99.3%)† | 100% (97.7%-100.0%) | 17.2% | 0.422 |
| Panther Fusion\* | 4 | 100.0% (85.8%-100.0%) | 100% (96.7%-100.0%) | 17.6% | 0.426 |
| Panther Aptima-M | 4 | 82.9% (66.2%-93.4%)† | 100% (97.7%-100.0%) | 15.1% | 0.401 |
| Panther Aptima-350 | 4 | 88.6% (73.3%-96.8%)† | 100% (97.7%-100.0%) | 16.2% | 0.411 |
| \*56 of the 192 pools tested on the other platforms were not tested by Panther Fusion.  †Restricting the performance characteristics comparison to only the 136 pools tested by Panther Fusion resulted in a PPA as follows: LDT 100% (95% CI 85.8%-100.0%), Aptima-M 91.7% (95% CI 73.0%-99.0%) and Aptima-350 95.8% (95% CI 78.9%-99.9%).  PPA, positive percent agreement; NPA, negative percent agreement; LDT, laboratory-developed test; Panther Aptima-M, Panther Aptima with manufacturer-set relative light unit cutoff; Panther Aptima-350, Panther Aptima with relative light unit cutoff of >350 considered positive; RLU, relative light unit. | | | | | |

**Figure Legends**

**Figure 1.** For a pool size of 8, paired individual and pooled Ct values for each individually positive sample (n=58), in order of increasing individual Ct value. The left panel contains pools comprised of only a single positive sample. The right panel contains pools comprised of two or more positive samples. The grey lines span the range of Ct values associated with a given pool. Rows without grey lines represent individually positive samples belonging to pools that tested negative by both rRT-PCR methods.

**Figure 2.** Passing-Bablok regression and Bland-Altman plots for pools of 8 containing only a single positive sample, tested by A-B) pooledLDT vs. individualLDT (n=23), C-D) pooled Panther Fusion vs. individualLDT (n=25), and E-F) pooled Panther Fusion vs. pooledLDT (n=32). For the Passing-Bablok regression plots (A, C, E), the solid line represents the line of regression, with 95% confidence interval shaded in grey. The dashed line represents the line of identity. The slope and intercept of the regression line are reported with 95% confidence intervals in parentheses. For the Bland-Altman plots (B, D, F), the solid line represents the mean difference in Ct value, with 95% limits of agreement area shaded in grey.

**Figure 3.** Model-estimated positive percent agreement (PPA) and testing efficiency, by pool size, proportion of tests positive, and proportion of samples with cycle threshold (Ct) above the 95% limit of detection (LoD). For these estimates, LoD has been held constant at the experimentally-derived Ct of 35.9, although results are independent of specific LoD value. A) Expected PPA between pooled and individual testing at pool sizes from 1-20. PPA decreases with decreasing proportion of tests positive (indicated by colored lines in each plot), and with increasing proportion of samples with Ct values beyond the 95% LoD (each panel). At >5% test positivity, expected PPA starts to increase at larger pool sizes as there is a greater likelihood of two positive samples being in the same pool. The baseline PPA (pool size of one) reflects the likelihood of obtaining the same individual result with repeat (non-pooled) testing. B) Estimated average tests per sample that would be performed at each pool size, with a lower number of average tests per sample corresponding to higher testing efficiency. Efficiency increases with decreasing proportion of tests positive, and slightly increases with increasing samples with Ct above the LoD. Each missed pool results in fewer deconvolutions, and thus fewer total tests performed.

**Figure 4.** Model-estimated positive percent agreement (PPA) and testing efficiency, by pool size, proportion of tests positive, and assay analytical sensitivity as approximated by the cycle threshold (Ct) corresponding to the 95% limit of detection (LoD). For these estimates, the population viral load distribution has been held constant at 15% of samples with Ct values above 35. A) Expected PPA between pooled and individual testing at pool sizes from 1-20. PPA decreases with decreasing proportion of tests positive (indicated by colored lines in each plot) and increases with increased analytical sensitivity (each panel). This occurs because the proportion of individuals samples with Ct value above each LoD decreases as the Ct LoD increases. B) Estimated average tests per sample that would be performed at each pool size, with a lower number of average tests per sample corresponding to higher testing efficiency. Efficiency increases with decreasing proportion of tests positive, and slightly decreases with increased analytical sensitivity, as more pools detected results in an increased number of individual tests performed at the deconvolution step.

**Figure 5.** Empiric and modeled estimates of positive percent agreement (PPA) with 95% confidence intervals (CI) for each pool size, testing platform, and sample type (all vs. first initial diagnostic vs. follow-up). The black circles represent the empiric PPA point estimates, and colored horizontal bars indicate the 95% CI. The 95% CI for the in silico data are too narrow to be visible in this plot. The grey boxplots represent the modeled estimate of PPA, with vertical black line indicating the modeled PPA point estimate, and grey box marking the 95% CI of the probabilistic sensitivity analysis. No modeled estimates are available for Panther Aptima because this is a transcription mediated amplification assay, and the model is based on dilutional effects inherent to real-time PCR only. The empiric 95% confidence intervals contain the modeled PPA point estimates for all conditions except for “pools of 8 follow-up tests only” and the “in silico” data. The data used to generate this figure is provided in Appendix Table 3.

**Technical Appendix**

Assay Comparisons for Pools of 4

To evaluate a pool size of 4, a total of 192 pools from 768 unique samples were tested on three different NAAT platforms (Table 1). Due to unforeseen logistical considerations, 56 of the 192 pools were tested only by LDT and Panther Aptima, but not tested by Panther Fusion. The remaining 136 pools were tested by all three methods. Among the 768 individual samples, 38 (4.9%) were positive, with a median Ct value of 29.3 (95% CI 20.3-33.9). First-time diagnostic samples had higher median Ct values than follow-up tests (Table 2).

Among the tested pools of 4, 18.2% (35/192) contained at least one positive sample. The positive pools were comprised of 32 pools with one positive sample, and 3 pools with two positive samples (Appendix Table 2). There were no false positive pools. The overall PPA of pooled testing ranged from 82.9% to 100% (Table 3). The 26 pools containing positive first-time diagnostic samples had higher PPA than the 9 pools containing positive follow-up tests by LDT (Appendix Table 3).

There were 6 total pools for which at least one method was falsely negative, all of which contained only a single positive specimen. Each of these 6 specimens had an individual Ct value greater than 34 cycles (median 36.4, IQR 34.6-37.5). Among individual positive specimens in the pools of 4 dataset, 10 (26.3%) had Ct>34. For the LDT, Panther Fusion, Panther Aptima-M, and Panther Aptima-350, 2/10 (20.0%), 0/6 (0.0%, 4 samples were not subjected to pooled testing), 6/10 (60.0%), and 4/10 (40.0%) were falsely negative. Four samples were first-time diagnostic specimens in individuals who were either symptomatic or had suspected exposures; the other 2 were follow-up tests in individuals with a prior diagnosis of COVID-19.

Linearity Studies for Pools of 4

For pools containing only a single positive sample, the pooled Panther Fusion assay showed positive systematic bias when compared to the individualLDT assay, as evidenced by the Passing-Bablok regression intercept value being greater than zero. ByLDT, pools resulted an average of 2.2 cycles (95% limits of agreement 0.6-3.9, p<0.001) later than the individual positive samples (Appendix Figure 3A, 3B). By Panther Fusion, pools resulted an average of 3.1 cycles (0.8-5.3, p<0.001) later than the individual positive samples (Appendix Figure 3C, 3D). Pools resulted an average of 0.73 cycles (-1.06-2.53, p<0.001) later on Panther Fusion when compared tothe LDT (Appendix Figure 3E, 3F). There was minimal proportional bias among the three assays, although the 95% confidence intervals for the Passing-Bablok regression slope for individual LDT versus pooled LDT, and for pooled LDT versus pooled Panther Fusion do not contain 1. This indicates slight positive and negative proportional biases respectively. The proportional bias is additionally highlighted in the Bland-Altman plots, which demonstrate that at higher Ct values, Panther Fusion outperforms the LDT.

External In Silico Validation Data

Data from an in silico sensitivity analysis for the Panther Fusion assay was obtained (Hologic Inc., pers. comm.) to validate our model. The data include a total of 52,272 SARS-CoV-2 tests performed from March through July 2020 at two external sites with an average prevalence of 19.1% (Table 2), and 13.0% of positive specimens with a Ct above that corresponding to the limit of detection of 35.6. The in silico sensitivity analysis was performed according to FDA recommendations by first determining the expected shift in Ct values using the Passing-Bablok regression equation generated through verification testing *(1)*. The expected positive percent agreement was calculated by dividing the number of specimens with a shifted Ct value beyond the cut-off of the assay. Pool sizes of 5 and 3 were evaluated in this manner, with expected PPA of 95.0% (94.7-95.2) and 99.9% (99.9-99.9) respectively (Appendix Table 3).

Modeling

We developed a stochastic simulation model to estimate PPA and efficiency for a two-stage pooled testing algorithm, based on characteristics of the underlying assay and patient population. To study the impact upon PPA and efficiency, we varied the proportion of positive tests (*s:* 0.1%, 1.0%, 3.0%, 5.0%, 10.0%, 15.0%), the 95% assay limit of detection Ct value (LoD; *l:* Ct corresponding to 95% detection: 32-40), the percentage of individual amplified Ct values above the LoD (*x:* 5.0%, 10.0%, 15.0%, 20.0%, 25.0%, 30.0%), and pool size (*p:* 1-20).

We fit the Ct values of samples positive for SARS-CoV-2 by rRT-PCR (n=804) received from an independent set of unique patients undergoing testing for SARS-CoV-2 from March 1 through June 24, 2020 (n=66,070) to candidate continuous probability distributions, and selected the best-fitting distribution based on the Bayesian information criterion and Kolmogorov-Smirnov statistic (0.0436). Because of slight negative skewness, a Weibull distribution best fit these data with shape and scale parameters of 4.55 and 29.86 (Appendix Figure 4). We then generated a set of random Ct values by sampling 50,000 times from this distribution. To study differing scenarios in which a variable proportion of samples had viral loads below the LoD, we generated additional sets of 50,000 Ct values with 5-30% of values above each LoD. For the base case, the expected Ct value of the pool was calculated using the following equation: *Ctpoolexpected* = -log2((∑2-*Ctsingle*)/*poolsize*). To model the probabilistic nature of detecting RNA at a given calculated pooled Ct value, we first fit a probit regression model using binary detection from an independent LoD experiment (100, 200, 500, 1000, 2000, 2500, 5000, 10,000 copies/mL in replicates of 5-20). This experiment’s LoD and confidence interval (685 copies/mL [95%CI 484-1074], Ct 35.9 [35.3-36.5]) were incorporated into the base case and sensitivity analyses.

For each pool size *p*, prevalence *s*, LoD *l*, and proportion of randomly-generated Ct values above LoD *x*, we randomly generated 10,000 pools for each possible combination of negative and positive pools (1 + *p*). Pools with zero positive samples were considered to be true negatives. For each randomly-generated pool with positive samples, we calculated the expected pooled Ct value from the individual randomly-sampled Ct values and assigned each pool as a true positive or false negative based upon the probability of detection derived from the probit regression model at a given LoD and expected pooled Ct value. Estimated PPA (true positives / [true positives + false negatives]) and average tests expected per sample ([1 + *p\**(true positive pools)]/*p*) were calculated, and results were weighted by the probability of observing a given pool combination with *i* individual positive samples (from 1 to *p*) using a binomial distribution. NPA was assumed to be 100%. Because the input datasets used to train this model were independent from the pooled datasets, we subsequently validated modeled estimates of PPA and average tests expected per sample against our empirical data for pools of 8 and 4, as well as external in silico data for pools of 5 and 3. Beyond the external factors that contributed to model estimates (individual Ct values, prevalence, proportion of samples above LoD), we also assessed model robustness via deterministic and probabilistic sensitivity analyses by varying pooling dilution (*Ctpoolexpected* = -log2((∑2-*Ctsingle*)/*poolsize ±* 1 Ct) and sampling from the 95% confidence interval of the probit regression.

The implementation code and data used to generate and validate this model can be made available upon request to the corresponding author.

**Appendix References**

1. Food and Drug Administration. Molecular Diagnostic Template for Laboratories. 2020 July 28 [cited 2020 Sep 18). https://www.fda.gov/media/135658/download.

**Appendix Table 1.** Receiver operating characteristic curve table for Panther Aptima, based on pools of 8 containing a single positive sample (n=36)

|  |  |  |
| --- | --- | --- |
| Positive percent agreement | Negative percent agreement | Relative light unit threshold |
| 1.000 | 0.047 | 301 |
| 0.972 | 0.109 | 302 |
| 0.972 | 0.156 | 304 |
| 0.972 | 0.188 | 305 |
| 0.972 | 0.203 | 306 |
| 0.972 | 0.219 | 307 |
| 0.972 | 0.234 | 308 |
| 0.944 | 0.297 | 309 |
| 0.944 | 0.359 | 310 |
| 0.944 | 0.406 | 311 |
| 0.944 | 0.438 | 312 |
| 0.944 | 0.469 | 313 |
| 0.944 | 0.500 | 314 |
| 0.944 | 0.516 | 315 |
| 0.917 | 0.578 | 316 |
| 0.889 | 0.641 | 317 |
| 0.861 | 0.688 | 318 |
| 0.833 | 0.688 | 320 |
| 0.833 | 0.703 | 321 |
| 0.833 | 0.750 | 322 |
| 0.833 | 0.797 | 323 |
| 0.833 | 0.844 | 325 |
| 0.833 | 0.891 | 326 |
| 0.833 | 0.906 | 327 |
| 0.833 | 0.922 | 328 |
| 0.806 | 0.938 | 330 |
| 0.778 | 0.953 | 332 |
| 0.778 | 0.984 | 343 |
| 0.750 | 0.984 | 393 |
| 0.750 | 1.000 | 474 |
| 0.722 | 1.000 | 531 |
| 0.694 | 1.000 | 570 |
| \*Cases defined as pools containing one or more positive samples by individual testing; controls defined as pools containing only samples negative by individual testing. Ten pools containing more than one positive sample were excluded from analysis. | | |

**Appendix Table 2.** Results of 4-sample pooled testing, by testing platform and number of positive specimens per pool (n=192)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pooled Testing | | | | Individual Testing | | Total no. pools |
|  | LDT | Panther Fusion | Panther Aptima-M | Panther Aptima-350 | Positive  (no. 1PP, no. >1PP) | Negative |  |
|  | Pos | Pos | Pos | Pos | 22 (19, 3) | 0 | 22 |
|  | Pos | Pos | - | Pos | 1 (1, 0) | 0 | 1 |
|  | Pos | Pos | - | - | 1 (1, 0) | 0 | 1 |
|  | Pos | N/A\* | Pos | Pos | 7 (7, 0) | 0 | 7 |
|  | Pos | N/A\* | - | Pos | 0 (0, 0) | 0 | 0 |
|  | Pos | N/A\* | - | - | 2 (2, 0) | 0 | 2 |
|  | Pos | - | Pos | Pos | 0 (0, 0) | 0 | 0 |
|  | Pos | - | - | Pos | 0 (0, 0) | 0 | 0 |
|  | Pos | - | - | - | 0 (0, 0) | 0 | 0 |
|  | - | Pos | Pos | Pos | 0 (0, 0) | 0 | 0 |
|  | - | Pos | - | Pos | 0 (0, 0) | 0 | 0 |
|  | - | Pos | - | - | 0 (0, 0) | 0 | 0 |
|  | - | N/A\* | Pos | Pos | 0 (0, 0) | 0 | 0 |
|  | - | N/A\* | - | Pos | 1 (1, 0) | 0 | 1 |
|  | - | N/A\* | - | - | 1 (1, 0) | 45 | 46 |
|  | - | - | Pos | Pos | 0 (0, 0) | 0 | 0 |
|  | - | - | - | Pos | 0 (0, 0) | 0 | 0 |
|  | - | - | - | - | 0 (0, 0) | 112 | 112 |
| No. positive pools | 33 | 24 | 29 | 31 | 35 (32, 3) | - | - |
| No. negative pools | 159 | 112 | 163 | 161 | - | 157 | - |
| Total no. pools | 192 | 136 | 192 | 192 | - | - | 192 |
| \*N/A, not applicable, indicates pools with no Panther Fusion testing done.  LDT, laboratory-developed test; Panther Aptima-M, Panther Aptima with manufacturer-set relative light unit cutoff; Panther Aptima-350, Panther Aptima with relative light unit cutoff of >350 considered positive; no., number; 1PP, 1 positive specimen in pool; >1PP, 2 or more positive specimens in pool; Pos, positive; -, negative; Ct, cycle threshold; RLU, relative light unit. | | | | | | | |

**Appendix Table 3.** Modeled versus empiric positive percent agreement and testing efficiency

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Testing platform | Pool size | Test type | Model input variables | | | Model estimate with PSA | | Empiric data | |
| Percent positive samples | 95% LoD | Percent Ct > LoD | PPA (95% CI) | Tests per sample (95% CI) | PPA (95% CI) | Tests per sample |
| LDT | 8 | All | 6.6% | 35.9 | 22.4% | 84.0% (78.9%-89.0%) | 0.479 (0.457-0.500) | 71.7% (56.5%-84.0%) | 0.434 |
| Panther Fusion | 8 | All | 6.6% | 35.6 | 24.1% | 82.7% (77.5%-88.1%) | 0.473 (0.451-0.496) | 76.1% (61.2%-87.4%) | 0.452 |
| Panther Aptima-M | 8 | All | - | - | - | - | - | 73.9% (58.9%-85.7%) | 0.434 |
| Panther Aptima-350 | 8 | All | - | - | - | - | - | 82.6% (68.6%-92.2%) | 0.470 |
| LDT | 8 | First† | 3.7% | 35.9 | 8.3% | 93.3% (90.2%-96.3%) | 0.368 (0.360-0.376) | 100% (76.8%-100%) | - |
| Panther Fusion | 8 | First† | 3.7% | 35.6 | 8.3% | 93.3% (90.5%-96.3%) | 0.368 (0.361-0.376) | 100% (76.8%-100%) | - |
| Panther Aptima-M | 8 | First† | - | - | - | - | - | 100% (76.8%-100%) | - |
| Panther Aptima-350 | 8 | First† | - | - | - | - | - | 100% (76.8%-100%) | - |
| LDT | 8 | Follow-up† | 15.2% | 35.9 | 32.4% | 81.9% (74.9%-88.6%) | 0.725 (0.680-0.774) | 53.6% (33.9%-72.5%) | - |
| Panther Fusion | 8 | Follow-up† | 15.2% | 35.6 | 35.3% | 79.9% (73.6%-87.1%) | 0.711 (0.664-0.763) | 60.7% (40.6%-78.5%) | - |
| Panther Aptima-M | 8 | Follow-up† | - | - | - | - | - | 60.7% (40.6%-78.5%) | - |
| Panther Aptima-350 | 8 | Follow-up† | - | - | - | - | - | 71.4% (51.3%-86.8%) | - |
| LDT | 4 | All | 4.9% | 35.9 | 15.8% | 90.0% (84.8%-94.7%) | 0.414 (0.406-0.422) | 94.3% (80.8%-99.3%) | 0.422 |
| Panther Fusion\* | 4 | All | 4.9% | 35.6 | 15.8% | 89.9% (85.8%-94.0%) | 0.414 (0.406-0.421) | 100% (85.8%-100%) | 0.426 |
| Panther Aptima-M | 4 | All | - | - | - | - | - | 82.9% (66.2%-93.4%) | 0.401 |
| Panther Aptima-350 | 4 | All | - | - | - | - | - | 88.6% (73.3%-96.8%) | 0.411 |
| LDT | 4 | First‡ | 5.7% | 35.9 | 14.3% | 91.1% (87.2%-95.2%) | 0.441 (0.432-0.449) | 96.2% (80.4%-99.9%) | - |
| Panther Fusion\* | 4 | First‡ | 5.7% | 35.6 | 14.3% | 91.0% (86.7%-94.6%) | 0.440 (0.431-0.448) | 100% (82.4%-100%) | - |
| Panther Aptima-M | 4 | First‡ | - | - | - | - | - | 88.5% (69.9%-97.6%) | - |
| Panther Aptima-350 | 4 | First‡ | - | - | - | - | - | 92.3% (74.9%-99.1%) | - |
| LDT | 4 | Follow-up‡ | 3.6% | 35.9 | 20.0% | 86.9% (81.8%-92.3%) | 0.369 (0.362-0.376) | 88.9% (51.8%-99.7%) | - |
| Panther Fusion\* | 4 | Follow-up‡ | 3.6% | 35.6 | 20.0% | 86.9% (82.1%-92.0%) | 0.369 (0.362-0.376) | 100% (47.8%-100.0%) | - |
| Panther Aptima-M | 4 | Follow-up‡ | - | - | - | - | - | 66.7% (29.9%-92.5%) | - |
| Panther Aptima-350 | 4 | Follow-up‡ | - | - | - | - | - | 77.8% (40.0%-97.2%) | - |
| Panther Fusion | 5 | In silico | 19.1% | 35.6 | 13.0% | 93.5% (90.8%-96.3%) | 0.811 (0.793-0.830) | 95.0% (94.7-95.2) | - |
| Panther Fusion | 3 | In silico | 19.1% | 35.6 | 13.0% | 93.5% (90.5%-96.5%) | 0.773 (0.759-0.787) | 99.9% (99.9-99.9) | - |
| \*56 of the 192 pools tested on the other platforms were not tested by Panther Fusion.  †For a pool size of 8, there were a total of 14 pools containing positive first tests only, and 28 pools containing positive follow-up tests only. These numbers represent the denominator for the calculation of PPA.  ‡For a pool size of 4, there were a total of 26 pools containing positive first tests only, and 9 pools containing positive follow-up tests only. These numbers represent the denominator for the calculation of PPA.  PSA, probabilistic sensitivity analysis; LoD, limit of detection; Ct, cycle threshold; PPA, positive percent agreement; CI, confidence interval; LDT, laboratory-developed test; Panther Aptima-M, Panther Aptima with manufacturer-set relative light unit cutoff; Panther Aptima-350, Panther Aptima with relative light unit cutoff of >350 considered positive; RLU, relative light unit. | | | | | | | | | |

**Appendix Table 4** (separate excel file). Lookup table for estimated positive percent agreement with deterministic sensitivity analysis upper and lower bounds for pool sizes between 1 and 20 based on model inputs of proportion tests positive (0.001 to 0.15) and proportion of tests with cycle threshold greater than the 95% limit of detection (0.05 to 0.30). Input values in cells B1 and B2 of first sheet “lookup\_table”.

**Appendix Figure 1.** A) Receiver operating characteristic curve of pools of 8 containing only a single positive sample tested by Panther Aptima, and individual samples tested by LDT, with area under the curve of 0.911 (n=36). B) Positive percent agreement (PPA) plotted against Panther Aptima relative light unit (RLU) threshold. Based on the inflection point of this curve, an RLU cutoff of >350 was chosen to maximize PPA.

**Appendix Figure 2.** For a pool size of 4, paired individual and pooled Ct values for each individually positive sample (n=38), in order of increasing individual Ct value. The left panel contains pools comprised of only a single positive sample. The right panel contains pools comprised of two or more positive samples. The grey lines span the range of Ct values associated with a given pool. Pools without a red square were false negatives by the laboratory-developed test (LDT). Pools without a blue triangle were not tested by Panther Fusion, and do not represent false negatives.

**Appendix Figure 3.** Passing-Bablok regression and Bland-Altman plots for pools of 4 containing only a single positive sample, tested by A-B) pooled LDT vs. individual LDT (n=30), C-D) pooled Panther Fusion vs. individual LDT (n=21), and E-F) pooled Panther Fusion vs. pooled LDT (n=24). For the Passing-Bablok regression plots (A, C, E), the solid line represents the line of regression, with 95% confidence interval shaded in grey. The dashed line represents the line of identity. The slope and intercept of the regression line are reported with 95% confidence intervals in parentheses. For the Bland-Altman plots (B, D, F), the solid line represents the mean difference in Ct value, with 95% limits of agreement range shaded in grey.

**Appendix Figure 4.** Continuous probability distributions fit to independent dataset of cycle threshold (Ct) values not subjected to pooled testing. Fitted theoretical weibull (red), normal (green), gamma (dark blue), and log-normal (light blue) distributions are plotted alongside empirical dataset for probability densities, quantiles (Q-Q plot), cumulative distribution functions, and probabilities (P-P plot). The fitted Weibull distribution was selected based on minimization of the Bayesian Information Criterion, Akaike information criterion, and Kolmogorov–Smirnov statistic.

**Appendix Figure 5.** Fitted probit regression (solid black line) with 95% confidence intervals (dashed black lines) derived from independent limit of detection (LoD) experiment. Probability of detection is plotted against nominal viral copies per milliliter (A, top) and corresponding cycle threshold (Ct) value (B, bottom). Solid red lines indicate 95% estimated probability of detection (horizontal) and corresponding 95% LoD (vertical, 685 cp/mL or Ct 35.9).

**Appendix Figure 6.** Model-estimated positive percent agreement (PPA) and testing efficiency, by pool size, proportion of tests positive, assay sensitivity as represented by cycle threshold (Ct) corresponding to the 95% limit of detection (LoD), and proportion of samples with Ct above the LoD. The relationship between PPA and pool size is independent of the actual Ct value corresponding to the 95% LoD due to a fixed proportion of Ct values above the LoD (5%, 15%, 25%), as demonstrated by identical plots in each vertical panel.

**Appendix Figure 7.** One-way deterministic sensitivity analysis for modeled estimates of positive percent agreement (PPA) between pooled and individual testing at pool sizes from 1-20 for variable prevalence. Solid lines indicate modeled base case estimates (presented in Figure 4A), while dashed lines indicate modeled estimates at upper and lower bounds of sensitivity analysis. A) Deterministic sensitivity analysis for deviation from pooled testing dilution effect (±1 Ct value). B) Deterministic sensitivity analysis for deviation from fitted probit regression (±2 standard deviations).

**Appendix Figure 8.** One-way deterministic sensitivity analysis for modeled estimates of tests per sample between pooled and individual testing at pool sizes from 1-20 for variable prevalence. Solid lines indicate modeled base case estimates (presented in Figure 4B), while dashed lines indicate modeled estimates at upper and lower bounds of sensitivity analysis. A) Deterministic sensitivity analysis for deviation from pooled testing dilution effect (±1 Ct value). B) Deterministic sensitivity analysis for deviation from fitted probit regression (±2 standard deviations).