

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37

Performance Decay of Molecular Assays Near the Limit of Detection: Probabilistic Modeling using Real-World COVID-19 Data

Thomas J.S. Durant, MD¹; Christopher D. Koch, PhD¹; Christopher A. Kerantzas, MD, PhD¹; David R. Peaper, MD, PhD^{1*}

¹ – Yale University School of Medicine, Department of Laboratory Medicine: 55 Park Street PS345D, New Haven, CT 06511.

* To whom correspondence should be addressed: Department of Laboratory Medicine, 55 Park Street PS345D, New Haven, CT 06511.

E-mail: david.peaper@yale.edu

Abbreviations:

- COVID-19: Coronavirus disease of 2019
- Ct value: Cycle-threshold value
- EHR: Electronic health record
- EUA: Emergency use authorization
- HDO: Healthcare delivery organization
- IFU: Instructions for use
- IVD: In-vitro diagnostics
- KS_d : Kolmogorov–Smirnov distance
- LIS: Laboratory information system
- LOD: Limit of detection
- NAAT: nucleic acid amplification tests
- NPS: Nasopharyngeal swab
- RT-PCR: Real-time reverse transcription-PCR
- SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2
- UTM: universal transport medium
- VTM: viral transport medium
- WHVA: West Haven, Veteran’s Administration Hospital

Running title: SARS-CoV-2 Sample Viral Burden and Detection Rates

Support: None

Keywords: Viral burden, cycle threshold, sensitivity, detection rate, sars-cov-2, covid-19, bootstrap resampling

38 **ABSTRACT:**

39 The gold standard for diagnosis of COVID-19 is detection of SARS-CoV-2 RNA by RT-PCR. However, the
40 effect of systematic changes in specimen viral burden on the overall assay performance is not
41 quantitatively described. We observed decreased viral burdens in our testing population as the
42 pandemic progressed, with median sample Ct values increasing from 22.7 to 32.8 from weeks 14 and 20,
43 respectively. We developed a method using computer simulations to quantify the implications of
44 variable SARS-CoV-2 viral burden on observed assay performance. We found that overall decreasing
45 viral burden can have profound effects on assay detection rates. When real-world Ct values were used
46 as source data in a bootstrap resampling simulation, the sensitivity of the same hypothetical assay
47 decreased from 97.59 (95% CI 97.3-97.9) in week 12, to 74.42 (95% CI 73.9-75) in week 20. Furthermore,
48 simulated assays with a 3-fold or 10-fold reduced sensitivity would both appear to be >95% sensitive
49 early in the pandemic, but sensitivity would fall to 85.55 (95% CI 84.9-86.2) and 74.38 (95% CI 73.6-75.1)
50 later in the pandemic, respectively. Our modeling approach can be used to better quantitate the impact
51 that specimen viral burden may have on the clinical application of tests and specimens.

52

53 INTRODUCTION

54 On February 4, 2020, the U.S. Department of Health and Human Services (HHS) declared a public health
55 emergency and authorized the emergency use of in vitro diagnostics (IVD) for the detection of the
56 severe acute respiratory coronavirus 2 (SARS-CoV-2). Since this time, the number of emergency use
57 authorizations (EUA) issued by the FDA, and the number of NAATs available for clinical use, have
58 increased linearly. Currently, the required data for EUA are primarily those which describe the analytic
59 sensitivity and specificity, and among available assays, these performance metrics are generally well
60 described. The FDA only recently released (<https://bit.ly/2RUL3R2>) results of testing using a standardized
61 panel of contrived viral samples, and a group recently compared assay limits of detection (LOD) using a
62 quantitated set of viral standards.(1) Finally, there have been several method to method comparisons of
63 SARS-CoV-2 testing.(1–14) However, at the time of this writing, the clinical performance of real-time
64 reverse transcription-PCR (RT-PCR) based assays remains a difficult metric to evaluate.

65 The preferred method for a laboratory diagnosis of acute COVID-19 is an amplified RNA test,
66 and RT-PCR is the most commonly used method. Among published reports of suspected clinical
67 inaccuracies observed with SARS-CoV-2 RT-PCR, both false positive and false negative results have been
68 reported. False positive results can occur due to specimen contamination, but analytical false positives
69 have been reported to the FDA. Much more concern has been focused on falsely negative results driven
70 by variation in test sensitivity. There are many variables in the testing process which arguably influence
71 the diagnostic sensitivity of RT-PCR-based assays. Preanalytical conditions which are often cited include,
72 timing of sample collection relative to the duration of illness, severity of illness, sample collection site
73 (e.g. nasopharyngeal swab (NPS), nasal swab), and sample collection technique.(15, 16) Collectively,
74 these variables directly influence viral burden in clinical samples, and can therefore influence diagnostic
75 test performance. From an analytical standpoint, as the viral burden in clinical samples approaches or
76 passes the LOD, the probability of qualitative detection decreases. As a result, specimen viral burden is a

77 key component of overall test performance, both in the analytic and clinical aspects of assay
78 consideration.

79 The diagnostic sensitivity of a test method is thought to be fairly constant across time, but
80 publications evaluating the performance of different COVID-19 diagnostic assays have demonstrated
81 highly variable sensitivity even for commercial assays with constant procedures. Additionally, studies of
82 different specimen types with potentially different viral burdens have not been consistent in their
83 reported sensitivity.(15–18) This has profound implications for clinical decision making since different
84 levels of assay sensitivity encompassing specimen, collector, and assay may be required for different
85 clinical scenarios.

86 As of December 2020, there is no commonly available and recognized reference material for
87 quantitative viral load testing for SARS-CoV-2. Accordingly, the cycle threshold (Ct) value of real time RT-
88 PCR assays, when available, is the only widely available semi-quantitative indicator of viral burden, and
89 Ct value is inversely proportional to viral burden in a given sample. The Ct value represents the number
90 of completed PCR amplification cycles it takes for the fluorescent signal, denoting the target gene, to
91 increase beyond a threshold for positivity. Limitations of the Ct value include an inability to generalize
92 between platforms, within-platform variability between laboratories, and the absence of available Ct
93 values from some IVD instruments and analytical methods. Despite these limitations, using the Ct value
94 to evaluate viral burden in clinical samples is a common clinical practice, and is often discussed in
95 COVID-related literature as having potential as a prognostic indicator. In addition, it is often requested
96 by clinicians to help explain discrepancies between provider clinical impression and a positive test result.
97 However, there is limited evidence to support the widespread use of Ct values in these circumstances.

98 From an analytic standpoint, the Ct value remains a primary determinant in producing a
99 qualitative result. In addition, there are also emerging reports on how Ct values of the underlying

100 population may influence the diagnostic performance of an assay. In a recent study by Green et al., Ct
101 values reported from samples where SARS-CoV-2 RNA was detected and stratified by patients who
102 received repeat testing versus a single test. Testing was performed on the Cobas 6800 assay, and
103 between single and repeat testing cohorts, they observed 3.38 and 6.59% increases in Ct values for gene
104 targets 1 and 2, respectively.(10) Accordingly, the authors note that this difference in viral burden can
105 influence the clinical sensitivity of the assay. While there are emerging reports which discuss the effect
106 of variable viral burden in samples with respect to patient-level diagnosis, the degree to which shifting
107 Ct values may influence the diagnostic sensitivity has not been quantitatively described.

108 During the initial weeks of the pandemic, we observed increasing rates of false negative results
109 during validation experiments at our institution. Upon investigation we noted higher Ct values among
110 the false negative samples. While it is known that higher Ct values are less likely to be detected, there
111 was a higher proportion of low viral burden specimens in our sample cohort relative to the beginning of
112 the pandemic. Given the potential impact of viral burden on test positivity, we sought to quantitate the
113 systematic distribution of Ct values in the tested population from our institution across epidemiologic
114 weeks. Further, because it was not practical to do multiple comparisons across multiple platforms to
115 reliably describe differences in assay performance due to changes in sample viral burden, we employed
116 mathematical modeling to quantitatively describe this. To this end, we used publicly available data to
117 develop a model of test performance characteristics corresponding to different viral burdens allowing
118 probabilistic-based simulation of RNA detection for a hypothetical COVID-19 assay evaluated at different
119 timepoints during the pandemic. We then developed a computational model to simulate the
120 assessment of RT-PCR clinical performance at our institution, using a bootstrap resampling method to
121 describe variations in test sensitivity based upon different underlying assay performance assumptions.
122 We hypothesize that mathematical modeling will demonstrate a temporal decay in diagnostic sensitivity
123 as viral burden in clinical samples decreased.

124

125 **METHODS**

126 *SARS-CoV-2 RNA Molecular Testing*

127 Multiple molecular-based methods were used for detecting SARS-CoV-2 RNA during the observation
128 period of this project. The platforms which performed more than 95% of testing within our healthcare
129 delivery organization (HDO) included (1) the Panther Aptima SARS-CoV-2 Assay (Hologic; San Diego, CA)
130 (2) the TaqPath COVID-19 Combo Kit (Thermo Fisher Scientific, Inc; Carlsbad, CA), (3) the Xpert Xpress
131 SARS-CoV-2 test (Cepheid; Sunnyvale, CA), and (4) a modified version of the CDC assay. The Xpert Xpress
132 and Aptima SARS-CoV-2 testing were performed at different sites within a 5-hospital HDO in Connecticut
133 and Rhode Island starting March 28, 2020 and May 15, 2020, respectively. All testing was performed
134 within clinical laboratories consistent with the manufacturer's instructions. The modified CDC assay was
135 performed at a single laboratory within our HDO beginning on March 13, 2020. It was granted
136 Emergency Use Authorization by the US FDA, and has been previously described.(2) The preferred
137 specimen type across all platforms was a NPS collected in 3 mL of universal or viral transport medium
138 (VTM), but other specimen types and collection methods were used as needed based upon supply and
139 reagent availability.

140

141 *Data Collection and Analysis*

142 Retrospective laboratory test order and result data were extracted from our electronic health record
143 (EHR) and laboratory information system (LIS). Row-level data, indexed by container ID, included test
144 result interpretation (e.g., 'detected', 'not detected', etc.), Ct value, test method, and a collection
145 datetime stamp. Additional metadata included patient class (e.g., inpatient, outpatient, or emergency
146 department) at the time of their encounter, specimen collection type, and the resulting laboratory.

147 Custom scripts for data processing and analysis were implemented in Python (version 3.7.4), and R
148 (version 4.0.2). Linear regression of Ct values associated with each gene target was performed in
149 GraphPad Prism (version 8.4.2) (GraphPad Software; San Diego, CA, USA). Ct value analysis was limited
150 to the platforms which provide a Ct value for each gene target which included the modified CDC assay
151 and the Xpert assay. Tests which were performed at reference laboratories – i.e., ‘send-out’ tests – were
152 not included in the Ct value analysis. CDC epidemiologic week numbers are used throughout the entirety
153 of this report.

154

155 *Standardized assay performance analysis*

156 Instructions for use (IFU) available on the US FDA website were reviewed on August 13, 2020. Limit of
157 detection performance data were extracted from IFU if they met the following criteria: 1) The test was
158 commercially available, 2) the test was an extracted RT-PCR assay, 3) at least 20 samples were tested at
159 each concentration. Four assays met these criteria: Roche Cobas SARS-CoV-2, Abbott Alinity m SARS-
160 CoV-2 (Abbott Molecular Inc; Des Plaines, IL), Cepheid Xpert Xpress SARS-CoV-2, and Perkin Elmer New
161 Coronavirus Nucleic Acid Detection Kit (Perkin Elmer; Waltham, MA). The LOD, concentration tested,
162 number of samples tested, and number of samples detected were recorded for each assay. The data
163 from the Roche and Perkin Elmer assays were presented by RT-PCR target, while the data for the Abbott
164 and Cepheid assays were presented for the total assay. For the purposes of calculations and curve
165 fitting, the LOD of the Cepheid assay was changed to 0.005 PFU/mL from the claimed LOD of 0.02
166 PFU/mL since 95% of samples tested were positive at 0.005 PFU/mL. This revised LOD was consistent
167 with probit analysis determined LOD (Data Not Shown). To control for different LODs and units of
168 measure, each concentration tested was converted to a %-LOD. Individual samples with %-LOD greater
169 than 250% were excluded from further analysis. The %-Detected was calculated for each assay or gene

170 target. The %-Detected was plotted versus the %-LOD, and non-linear regression analysis with an
171 exponential plateau function (where; $Y_{\min} = 0$ and $Y_{\max} = 100$) was fit in GraphPad Prism. Additionally, the
172 fraction detected was used to determine a probit value, probit versus $\text{Log}(10)$ %-LOD was plotted, and
173 linear regression was performed (GraphPad Prism v8).

174

175 *Test Performance Simulation Model*

176 The purpose of the model simulation was to determine – for a given cohort of positive samples – what
177 fraction would be detected given the viral burden in the sample, and across the tested cohort, what is
178 the calculated sensitivity. To this end, a function to calculate the probability of detection, on a per
179 sample basis, was extrapolated from the non-linear regression model of the IFU data described above.
180 Clinical performance of a hypothetical test was first simulated using Ct values from the modified CDC
181 assay, pertaining to the N2 gene target. All N2 Ct values across the observational period, which were
182 derived from a sample that was clinically reported as ‘detected’, were collected from the LIS (Figure 1A).
183 These were subsequently collated into arrays and separated into groups by epidemiologic week (Figure
184 1B). From each dataset, corresponding to an epidemiology week, individual samples and their
185 corresponding Ct value were randomly selected using a bootstrap resampling method with replacement
186 (Figure 1C).⁽¹⁹⁾ The Ct value was then used as input for the probability function (P) (Figure 1D). The Ct
187 value at the LOD (Ct_{LOD}) for the hypothetical assay for this simulation was defined as the viral
188 concentration corresponding to N2 $Ct_{\text{LOD}} = 35.0$. Equation P yields a sum-to-one probability of ‘detected’
189 or ‘not detected’, wherein the probability was then used as input for a single binomial distribution
190 function to return N (e.g., where a high P would be likely to return $N = 1$ (i.e., ‘Detected’) and a low P
191 would be likely to return $N = 0$ (i.e., ‘Not detected’)) (Figure 1E). Output N was then compared with the
192 ground truth (i.e., the clinical test result), which in this case was always ‘detected’, to determine if the

193 simulated result was a true positive or a false negative (Figure 1F). The total number of simulated true
194 positives and false negatives were then used to calculate sensitivity for each epidemiologic week. This
195 was repeated 100 times across all weeks to calculate the average, standard deviation, median, and
196 interquartile range of the simulated sensitivity for each week.

197 A second iteration of simulated sensitivity was carried out using Ct values from the Xpert Xpress
198 SARS-CoV-2 assay, pertaining to the N2 gene target. The LoD for the hypothetical assay for this
199 simulation was defined as the viral concentration corresponding to an N2 Ct value of 40.5, 38.85, and
200 37.2, which correspond to the LoD of the Xpert assay, and a 0.5- and 1-Log reduction thereof. Ct values
201 were then grouped by epidemiologic weeks which were representative of sample viral burden
202 distributions at our institution during early and late time points during the pandemic. The simulation
203 was repeated 100 times across $Ct_{LOD} = 40.5, 38.85, \text{ and } 37.2$, and across selected weeks to calculate the
204 average, standard deviation, median, and interquartile range of the simulated sensitivity for each week.

205

206 **RESULTS**

207 *SARS-CoV-2 Testing:*

208 The observation period for this project was between March 15, 2020 and July 31, 2020 (corresponding
209 to epidemiologic weeks 12 through 31). The flagship hospital for our HDO is located in New Haven, CT,
210 and we were part of the “first wave” of COVID-19 infections in the USA. We implemented testing locally
211 on March 13, 2020, with the first full week of testing occurring on week 12. COVID-19 positivity rates by
212 PCR testing peaked at week 14 (approx. 30%). This peak has been followed by a long tail of reduced
213 positivity in the setting of dramatically increased outpatient testing. Testing of inpatients and those in
214 the ED has remained stable since early in the pandemic (Figure 2).

215

216 *Ct Value Results:*

217 While monitoring test performance for quality purposes, we observed that Ct values generated by the
218 available testing platforms, a laboratory developed variation of the CDC assay and the Cepheid
219 GeneXpert Xpress, were increasing with time as the pandemic progressed. We wished to formally
220 characterize this increase by tracking Ct values by week until mid-May 2020 (week 20), a period in which
221 nearly all local testing was performed by one of these two assays. Indeed, Ct values for all gene targets
222 increased with time (Figure 3A – 3D) such that the median Ct values for the N2 gene target increased
223 from 22.7 to 32.8 and 26.0 to 34.8 for the CDC and Xpert assays from week 14 to week 20, respectively.
224 When assessed by linear regression, the slopes for regression lines for all four gene targets were
225 significantly greater than zero. In contrast, Ct values for an influenza A gene target from the Xpert Xpress
226 Influenza A / Influenza B assay observed during the 2019 – 2020 influenza virus season, did not differ as
227 the season progressed, and the slope of the best fit line was not significantly different from zero (Figure
228 3E)(Supplemental Table 1)(Supplemental Figure 2).

229

230 *Determination of detection probability as a function of fraction of LoD*

231 We wished to determine the quantitative effect of rising Ct values on diagnostic test performance, but
232 Ct values are not transmutable among assays. Quantitative viral burden measurements would facilitate
233 comparisons among assays, but these are not widely done, and assay limits of detection can be
234 measured in a variety of units that cannot be interconverted. We hypothesized that most extracted RT-
235 PCR assays would have similar behavior at and below the LoD. Additionally, we wished to derive a
236 method wherein the probability of a single sample being reported as positive could be calculated based
237 upon its Ct value compared to the Ct_{LoD} .

238 We reviewed instructions for use available on the FDA website for commercially available,
239 extracted real-time RT-PCR assays for COVID-19. Data from assays that contained detailed LOD studies,
240 as described in the materials and methods, were extracted, and four assays (Abbott Alinity m, Perkin
241 Elmer, Roche cobas, and Xpert Xpress) met these criteria with individual gene targets from two assays (N
242 and ORF1ab from Perkin Elmer and Target 1 and Target 2 from Roche cobas) being used. Tested
243 concentrations were converted to %-LOD to normalize across testing platforms and units of measure.
244 Plotting %-detected v. %-LoD allowed us to fit an exponential plateau curve (Figure 4) to the data which
245 had an $R^2 = 0.91$ and constant ($k = 0.028737$) (Equation A):

$$246 \quad P = 100 - \left(100 \times e^{(-k \times \%LOD)} \right) \quad (A)$$

247 %-LoD can be expressed as relationship between the Ct_{sample} and Ct_{LOD} (B):

$$248 \quad \%LOD = 100 \div 2^{(Ct_{sample} - Ct_{LOD})} \quad (B)$$

249 Accordingly, by substituting the %-LoD expression in terms of Ct_{LOD} , equation (C) relates the probability
250 of viral RNA detection by RT-PCR (P) as a function of observed Ct value (Ct_{sample}) and Ct_{LOD} , where k is a
251 constant.

$$252 \quad P = 100 - \left(100 \times e^{\left(-k \times \left(100 \div 2^{(Ct_{sample} - Ct_{LOD})} \right) \right) \right) \quad (C)$$

253

254 *Test Performance Simulation Results:*

255 Using equation (C) which relates the probability (P) of detecting a sample based on the relationship of
256 Ct_{sample} to Ct_{LOD} , we simulated the performance of a single hypothetical assay as observed viral burden
257 among positive samples decreased as the pandemic progressed as summarized in Figure 1. Assuming a

258 relationship between LoD and Ct value is consistent within a given sample, samples with Ct values at the
259 assay LoD should be detected 95% of the time, and samples with Ct values above the Ct_{LoD} should have a
260 non-zero probability of detection that is inversely related to Ct_{sample} (i.e., samples with $Ct_{sample} > Ct_{LoD}$ may
261 be detected, but less than 95% of the time). Accordingly, for the simulated test performance of a
262 hypothetical assay with a Ct_{LoD} corresponding to N2 Ct = 35 in the CDC assay, the simulated diagnostic
263 sensitivity decreased from 97.59 (95% CI 97.3 to 97.9) in week 12, to 74.42 (95% CI 73.9 to 75) in week
264 20. In addition, the coefficient of variation (CV) increased from 1.6% to 3.6% from week 12 to week 20,
265 respectively. At week 16, the simulated diagnostic sensitivity fell below 95% with an average of 94.17
266 (95% CI 94 to 94.4) (Figure 5).

267 The second iteration of simulated test performance was carried out using Ct values of the N2
268 gene target, measured by the Xpert Xpress SARS-CoV-2 assay and then grouped by epidemiologic weeks
269 14 and 20. The Ct_{LoD} for the Xpert Xpress assay is available in the IFU, and we were able to simulate the
270 performance of methods with the same relative LoD ($Ct_{LoD} = 40.5$) and hypothetical methods with 3-fold
271 and 10-fold reduced LoDs, $Ct_{LoD} = 38.85$ and $Ct_{LoD} = 37.2$, respectively (Figure 6). The average simulated
272 sensitivity for $Ct_{LoD} = 40.5$, 38.85, and 37.20, during epidemiologic week 14 was 98.71 (95% CI 98.5 to
273 98.9), 97.33 (95% CI 97.1 to 97.5), and 95.79 (95% CI 95.6 to 96), respectively. During week 20, the
274 average sensitivity for $Ct_{LoD} = 40.5$, 38.85, and 37.20 was 95.41 (95% CI 95 to 95.8), 85.55 (95% CI 84.9 to
275 86.2), and 74.38 (95% CI 73.6 to 75.1), respectively (Figure 6).

276

277 **DISCUSSION**

278 In this report, we describe a novel method for quantitatively describing the effect of variations of
279 sample viral burden on clinical performance characteristics of molecular-based assays. Higher Ct values
280 correlate with lower viral burdens, and lower viral burdens can lead to variable test performance. In
281 most cases where RT-PCR testing is applied for viral diagnostics, viral burden can vary among patients,
282 but we typically operate under the assumption that the relative distribution of viral burdens at any point
283 in time should be relatively constant – i.e., average viral burden shouldn't vary for patients with
284 influenza in December, January, February or March, all things being equal. It is recognized that different
285 patient populations may have different viral burdens (e.g. adult and pediatric patients for conventional
286 respiratory viruses), but these factors are often controlled or reported in comparative studies of test
287 performance. However, if viral burden varies in a systematic manner that is not controlled for or well-
288 known, test performance characteristics may vary substantially in unexpected ways.

289 We found that Ct values systematically shifted upward (i.e. toward lower viral burdens) among
290 samples tested within our HDO as the COVID-19 pandemic progressed in Connecticut. Conversely, the
291 distribution of Ct values for influenza testing (Supplemental Table 1 and Supplemental Figure 2), did not
292 change significantly over the course of the 2019 to 2020 flu-season (slope: 0.063 (95% CI -0.02 to 0.23)).
293 We hypothesize that there are two primary reasons for the observed shift in SARS-CoV-2 gene target Ct
294 values: 1) repeat testing of known positive patients later in disease when viral shedding is lower and 2)
295 shifting of testing guidelines to test more patients with less severe symptoms. Repeat testing cannot
296 explain all these changes, as GeneXpert testing at our institution is predominantly performed on newly
297 admitted patients. Indeed, this observation likely warrants further investigation as to the underlying
298 etiology and pathogenesis of SARS-CoV-2. The primary focus of this report, however, was to develop a
299 method to quantitatively describe the effect this observation may have on the diagnostic sensitivity an
300 assay.

301 The LoD is defined as the concentration at which there is a 95% probability of amplification and
302 detection of target nucleic acid. The second edition of the Clinical and Laboratory Standards Institute
303 (CLSI) guideline EP17-A2 recommends probit analysis, a parametric statistical method for estimating the
304 LoD based on observed detection rates and concomitant nucleic acid concentration data, but probit
305 analysis would not allow us to back calculate discrete probabilities of detection each specimen.(21)
306 Nonetheless, a probit regression was used to model the IFU LoD data which provided a good fit, with an
307 R squared of 0.8767 (Supplemental Figure 1). Instead, we used an exponential method to fit publicly
308 available LoD data to model the probability of detection for Ct values in our dataset. This allowed more
309 flexibility by treating Ct values as continuous input variables for the test performance simulation which
310 could be used to calculate a sum-to-one probability of detection using equation (A). In addition, the non-
311 linear regression resulted in a slightly more optimal R squared value of 0.9128.

312 We assumed for our model that a 10-fold change in virus concentration would result in a shift of
313 3.32 cycles across the entire analytical range. While this has been demonstrated for high viral burdens,
314 samples with lower viral burdens near the LoD may not predictably behave in this manner. We also
315 assumed that our baseline tests (e.g. LDT modified CDC assay and Xpert Xpress) were highly sensitive.
316 Our internal comparisons among assays is consistent with this assumption, but more complex
317 assumptions and/or models would be needed to extrapolate the performance of a more sensitive
318 method from the data generated by a less sensitive method. Finally, we did not incorporate test
319 specificity and potential false-positive results into our models. While this could be accomplished through
320 assigning a probability of detection to negative specimens corresponding to a pre-determined false
321 positive rate, it would not have a comparable set of data as that available for test sensitivity.

322 While our analysis was performed on NP swabs, our modeling data has implications for different
323 viral burdens associated with different specimen types. For example, if viral burdens are 5- to 10-fold
324 lower in alternative specimens compared to NP swabs, then the overall clinical sensitivity of the *same*

325 assay would be expected to vary according to our model. Variations in viral burden with time would
326 compound this change in performance, as well. Indeed, the modeling data presented in this report
327 demonstrate how the rate of detection can vary significantly as the distribution of viral burden changes
328 among samples tested, and this may explain the highly variable performance seen for some sample
329 types such as nasal swabs and saliva.(17, 23)

330 In order to interpret the effects of samples with lower viral burdens on test performance
331 characteristics in publications, whether in comparisons between methods or sample types, rigorous
332 reporting of study parameters as detailed in the STARD 2015 guidelines is necessary.(24) Moreover,
333 compared to other studies of viral diagnostics, it is even more important for COVID studies to detail the
334 provenance of the specimens used including dates of sample collection, explicit case-finding criteria, and
335 local positivity rates. Currently, the literature predominantly reports on samples collected in late March
336 to early May, when testing of asymptomatic patients was not as prevalent.(2, 3, 25) As the publication
337 cycle progresses and studies from later months emerge, it will be important to observe for greater
338 variation in performance characteristics as predicted by and demonstrated in this report. The Ct values
339 of patient specimens are important data to contextualize study findings. However, Ct ranges alone are
340 not sufficient due to the stochastic nature of detection near the LoD, and the proportion of specimens
341 with a low viral burden as evidenced by higher Ct values is an important metric rather than simply the
342 maximum observed Ct.(7) Subsequently, concordance and/or percent-positive-agreement analysis
343 should be tied to ranges of Ct values from well-studied assays in order to properly assess assay
344 performance on different patient populations. Broadly, these parameters will help guide the
345 interpretation and applicability of study results for different institutions and potentially different clinical
346 use case scenarios.

347 More fundamentally, variation in assay performance over time may be partly attributable to
348 changes in case definition. Whereas early in the pandemic case definitions included clinical criteria when

349 testing was more restricted, widespread testing of asymptomatic patients has led to laboratory test-
350 based case definitions. In lieu of a definitive diagnostic gold standard, high sensitivity methods have
351 been used either individually or with other high sensitivity methods as part of a composite reference
352 standard. Importantly, a single result arising from the use of only one high-sensitivity method as a
353 reference standard may be limited by the stochastic nature of test performance near the LoD. Multi-
354 target assays are likely to be less susceptible to stochastic effects than single-target assays, but also
355 introduce the need for studies to precisely define how each target result is translated into an overall
356 positive or negative specimen result. Though marked discordance between high sensitivity methods at
357 high Ct values has not been observed in several studies, discordance may emerge in data from later
358 months in the pandemic.(2, 3, 7) These considerations again highlight the need to evaluate high
359 sensitivity methods over time, individually and together, in order to capture variations in diagnostic
360 sensitivity.

361 Laboratory testing, particularly molecular-based assays, are often thought of being deterministic
362 in nature when generating a result – i.e., if the RNA is present, it will be detected. Accordingly, if the test
363 methodology is deterministic and stationary, we often think of test performance as stationary. However,
364 as viral burden in the sample approaches the LoD, it is known that these tests exhibit more stochastic
365 characteristics. At low concentrations of pathogen nucleic acid, chance can dictate whether target
366 nucleic acid is captured in a pipet, bound during the extraction phase, efficiently eluted, or pipetted into
367 the final reaction mixture. All of this complexity is combined to determine the probability of detection at
368 and below the assay LoD. These data demonstrate how variations in characteristics of the underlying
369 test population should be considered when considering the static nature or dynamic nature of test
370 performance.

371

372 **BIBLIOGRAPHY**

- 373 1. Fung B, Gopez A, Servellita V, Arevalo S, Ho C, Deucher A, Thornborrow E, Chiu C, Miller S. 2020. Direct
374 Comparison of SARS-CoV-2 Analytical Limits of Detection across Seven Molecular Assays. *J Clin Microbiol* 58.
- 375 2. Moran A, Beavis KG, Matushek SM, Ciaglia C, Francois N, Tesic V, Love N. 2020. Detection of SARS-CoV-2 by
376 Use of the Cepheid Xpert Xpress SARS-CoV-2 and Roche cobas SARS-CoV-2 Assays. *J Clin Microbiol* 58.
- 377 3. Broder K, Babiker A, Myers C, White T, Jones H, Cardella J, Burd EM, Hill CE, Kraft CS. 2020. Test Agreement
378 between Roche Cobas 6800 and Cepheid GeneXpert Xpress SARS-CoV-2 Assays at High Cycle Threshold
379 Ranges. *J Clin Microbiol* 58.
- 380 4. Liotti FM, Menchinelli G, Marchetti S, Morandotti GA, Sanguinetti M, Posteraro B, Cattani P. 2020.
381 Evaluation of three commercial assays for SARS-CoV-2 molecular detection in upper respiratory tract
382 samples. *Eur J Clin Microbiol Infect Dis*.
- 383 5. Rhoads DD, Cherian SS, Roman K, Stempak LM, Schmotzer CL, Sadri N. 2020. Comparison of Abbott ID Now,
384 DiaSorin Simplexa, and CDC FDA Emergency Use Authorization Methods for the Detection of SARS-CoV-2
385 from Nasopharyngeal and Nasal Swabs from Individuals Diagnosed with COVID-19. *J Clin Microbiol* 58.
- 386 6. Craney AR, Velu PD, Satlin MJ, Fauntleroy KA, Callan K, Robertson A, La Spina M, Lei B, Chen A, Alston T,
387 Rozman A, Loda M, Rennert H, Cushing M, Westblade LF. 2020. Comparison of Two High-Throughput
388 Reverse Transcription-PCR Systems for the Detection of Severe Acute Respiratory Syndrome Coronavirus 2.
389 *J Clin Microbiol* 58.
- 390 7. Smithgall MC, Scherberkova I, Whittier S, Green DA. 2020. Comparison of Cepheid Xpert Xpress and Abbott
391 ID Now to Roche cobas for the Rapid Detection of SARS-CoV-2. *J Clin Virol* 128:104428.
- 392 8. Zhen W, Manji R, Smith E, Berry GJ. 2020. Comparison of Four Molecular In Vitro Diagnostic Assays for the
393 Detection of SARS-CoV-2 in Nasopharyngeal Specimens. *J Clin Microbiol* 58.
- 394 9. Tanida K, Koste L, Koenig C, Wenzel W, Fritsch A, Frickmann H. 2020. Evaluation of the automated cartridge-
395 based ARIES SARS-CoV-2 Assay (RUO) against automated Cepheid Xpert Xpress SARS-CoV-2 PCR as gold
396 standard. *Eur J Microbiol Immunol (Bp)*.
- 397 10. Green DA, Zucker J, Westblade LF, Whittier S, Rennert H, Velu P, Craney A, Cushing M, Liu D, Sobieszczyk
398 ME, Boehme AK, Sepulveda JL. 2020. Clinical Performance of SARS-CoV-2 Molecular Tests. *J Clin Microbiol*
399 58.
- 400 11. Moore NM, Li H, Schejbal D, Lindsley J, Hayden MK. 2020. Comparison of Two Commercial Molecular Tests
401 and a Laboratory-Developed Modification of the CDC 2019-nCoV Reverse Transcriptase PCR Assay for the
402 Detection of SARS-CoV-2. *J Clin Microbiol* 58.
- 403 12. Dust K, Hedley A, Nichol K, Stein D, Adam H, Karlowsky JA, Bullard J, Van Caesele P, Alexander DC. 2020.
404 Comparison of Commercial Assays and Laboratory Developed Tests for Detection of SARS-CoV-2. *J Virol*
405 *Methods* 113970.
- 406 13. Lieberman JA, Pepper G, Naccache SN, Huang M-L, Jerome KR, Greninger AL. 2020. Comparison of
407 Commercially Available and Laboratory-Developed Assays for In Vitro Detection of SARS-CoV-2 in Clinical
408 Laboratories. *J Clin Microbiol* 58.

- 409 14. Jin R, Pettengill MA, Hartnett NL, Auerbach HE, Peiper SC, Wang Z. 2020. Commercial SARS-CoV-2 Molecular
410 Assays: Superior Analytical Sensitivity of cobas SARS-CoV-2 Relative to NxTAG Cov Extended Panel and ID
411 NOW COVID-19 Test. *Arch Pathol Lab Med*.
- 412 15. Hanson KE, Caliendo AM, Arias CA, Englund JA, Lee MJ, Loeb M, Patel R, El Alayli A, Kalot MA, Falck-Ytter Y,
413 Lavergne V, Morgan RL, Murad MH, Sultan S, Bhimraj A, Mustafa RA. 2020. Infectious Diseases Society of
414 America Guidelines on the Diagnosis of COVID-19. *Clin Infect Dis*.
- 415 16. McCormick-Baw C, Morgan K, Gaffney D, Cazares Y, Jaworski K, Byrd A, Molberg K, Cavuoti D. 2020. Saliva as
416 an Alternate Specimen Source for Detection of SARS-CoV-2 in Symptomatic Patients Using Cepheid Xpert
417 Xpress SARS-CoV-2. *J Clin Microbiol* 58.
- 418 17. Wyllie AL, Fournier J, Casanovas-Massana A, Campbell M, Tokuyama M, Vijayakumar P, Warren JL, Geng B,
419 Muenker MC, Moore AJ, Vogels CBF, Petrone ME, Ott IM, Lu P, Venkataraman A, Lu-Culligan A, Klein J,
420 Earnest R, Simonov M, Datta R, Handoko R, Naushad N, Sewanan LR, Valdez J, White EB, Lapidus S, Kalinich
421 CC, Jiang X, Kim DJ, Kudo E, Linehan M, Mao T, Moriyama M, Oh JE, Park A, Silva J, Song E, Takahashi T,
422 Taura M, Weizman O-E, Wong P, Yang Y, Bermejo S, Odio CD, Omer SB, Dela Cruz CS, Farhadian S,
423 Martinello RA, Iwasaki A, Grubaugh ND, Ko AI. 2020. Saliva or Nasopharyngeal Swab Specimens for
424 Detection of SARS-CoV-2. *N Engl J Med* 383:1283–1286.
- 425 18. Hou H, Chen J, Wang Y, Lu Y, Zhu Y, Zhang B, Wang F, Mao L, Tang Y-W, Hu B, Ren Y, Sun Z. 2020.
426 Multicenter Evaluation of the Cepheid Xpert Xpress SARS-CoV-2 Assay for the Detection of SARS-CoV-2 in
427 Oropharyngeal Swab Specimens. *J Clin Microbiol* 58.
- 428 19. Bland JM, Altman DG. 2015. Statistics Notes: Bootstrap resampling methods. *BMJ* 350:h2622.
- 429 20. Massey FJ. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J Am Stat Assoc* 46:68–78.
- 430 21. Csi. 2012. Evaluation of detection capability for clinical laboratory measurement procedures; approved
431 guideline—second edition. CLSI document EP17-A2.
- 432 22. Cradic K, Lockhart M, Ozbolt P, Fatica L, Landon L, Lieber M, Yang D, Swickard J, Wongchaowart N, Fuhrman
433 S, Antonara S. 2020. Clinical Evaluation and Utilization of Multiple Molecular In Vitro Diagnostic Assays for
434 the Detection of SARS-CoV-2. *Am J Clin Pathol* 154:201–207.
- 435 23. Callahan C, Lee R, Lee G, Zulauf KE, Kirby JE, Arnaout R. 2020. Nasal-Swab Testing Misses Patients with Low
436 SARS-CoV-2 Viral Loads. *medRxiv*.
- 437 24. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet
438 HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, STARD Group. 2015. STARD
439 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem* 61:1446–1452.
- 440 25. Tu Y-P, Jennings R, Hart B, Cangelosi GA, Wood RC, Wehber K, Verma P, Vojta D, Berke EM. 2020. Swabs
441 Collected by Patients or Health Care Workers for SARS-CoV-2 Testing. *N Engl J Med* 383:494–496.

442

443

444 **FIGURE LEGENDS:**

445 **Figure 1:** Flow diagram of test performance simulation model. (A) All N2 Ct values belonging to
446 specimens which were clinically reported as ‘detected’ represent the parent distribution. (B) N2 Ct
447 values are separated by epidemiologic week. (C) N2 Ct value is randomly selected from the Week 12
448 distribution. (D) The selected N2 Ct value is used as input (Ct_{sample}) to the probability function to
449 calculate a sum-to-one probability (P). (E) P is used to randomly generate a label of ‘detected’ or ‘not
450 detected’. (F) If the randomly generated label is ‘detected’, sample is classified as True Positive. If the
451 randomly generated label is ‘not detected’, sample is classified as False Negative. (G) Sample is placed
452 back into the original epidemiology week distribution. This process repeats 100 times per week across m
453 number of epidemiologic weeks for a total of 900 calculations.

454 **Figure 2:** (Left Y-axis) Number of unique SARS-CoV-2 RT-PCR tests grouped by patient location.
455 ‘Inpatient’ represents a summation of samples collected from admitted patients or patients in the
456 emergency department. (Right Y-axis) Number of unique SARS-CoV-2 RT-PCR tests which were ‘positive’.
457 Both data are grouped by CDC epidemiologic week numbers.

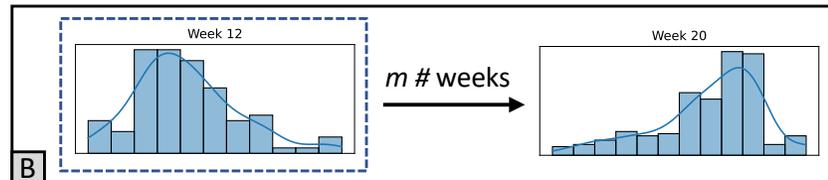
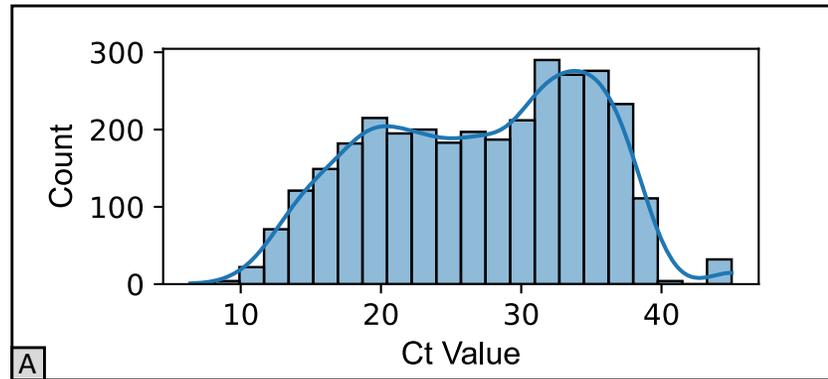
458 **Figure 3:** Boxplots representing the Ct value distributions, grouped by epidemiologic week number, test
459 method, and gene target. (A) Modified CDC assay N1 gene target. (B) Modified CDC assay N2 gene
460 target. (C) Cepheid Xpert Xpress SARS-CoV-2 E gene target. (D) Cepheid Xpert Xpress SARS-CoV-2 N2
461 gene target. (E) Cepheid Xpert Xpress Influenza A gene Target 1. Horizontal dashed lines represent the
462 assay LOD.

463 **Figure 4:** Composite test performance of commercial extracted RT-PCR assays for SARS-CoV-2. Available
464 limit of detection (LOD) data were extracted from publicly available instructions for use (IFU) and
465 normalized to %-LOD as described in methods. The %-Detected at each relative concentration was

466 calculated, plotted, and subjected to non-linear regression with an exponential plateau function with Y_0
467 = 0 and $Y_M = 100\%$. Best fit line (solid line) and 95% confidence intervals (dashed lines) are shown.

468 **Figure 5:** Boxplots representing the distributions of simulated diagnostic sensitivity using Ct values,
469 corresponding to N2 gene target, derived from samples tested by the modified CDC assay. Results are
470 grouped by epidemiologic week number. The horizontal dashed line represents the theoretical LOD
471 probability of detection: 95%.

472 **Figure 6:** Boxplots representing the distributions of simulated diagnostic sensitivity using Ct values,
473 corresponding to the N2 gene target, derived from samples tested by the Cepheid Xpert Xpress SARS-
474 CoV-2 assay. Results are grouped by epidemiologic week number and Ct_{LOD} per equation (C). The
475 horizontal dashed line represents the theoretical LOD probability of detection: 95%.



C Select random Ct value from Week 12 distribution

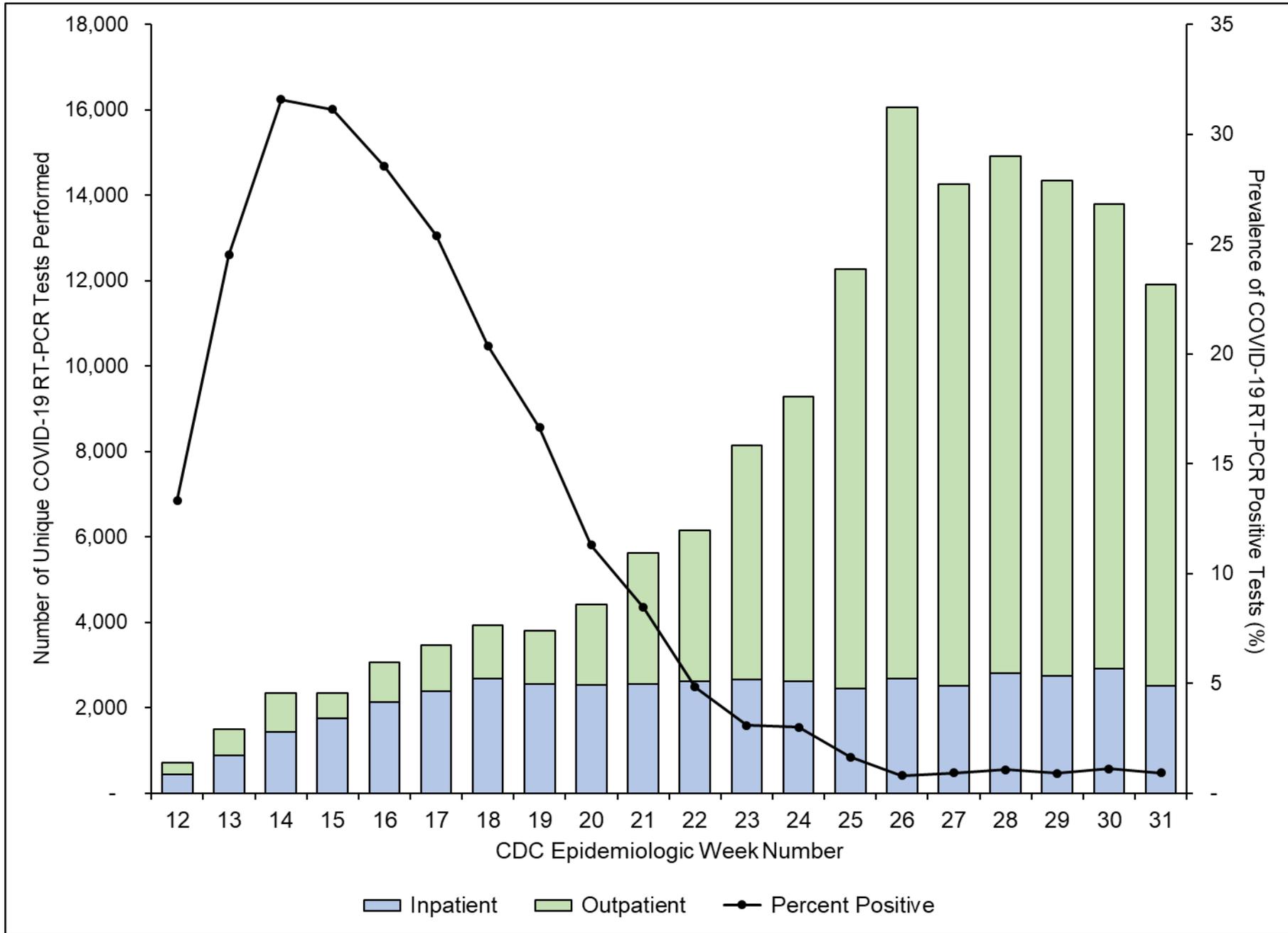
D Calculate a sum-to-one probability of detection

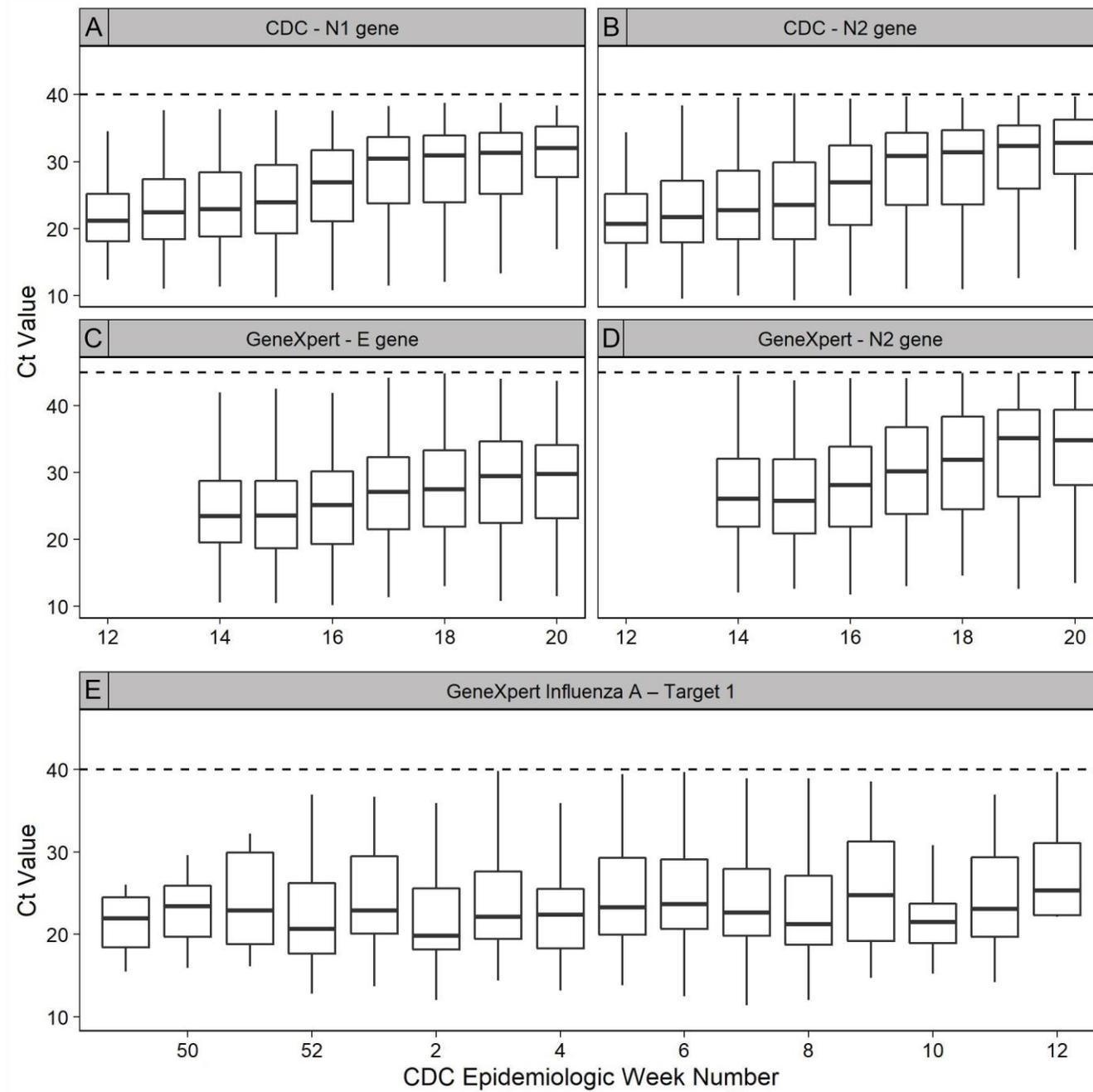
$$P = 100 - (100 \times e^{(-k \times (100 \div 2^{(Ct_{sample} - Ct_{LoD})})})}$$

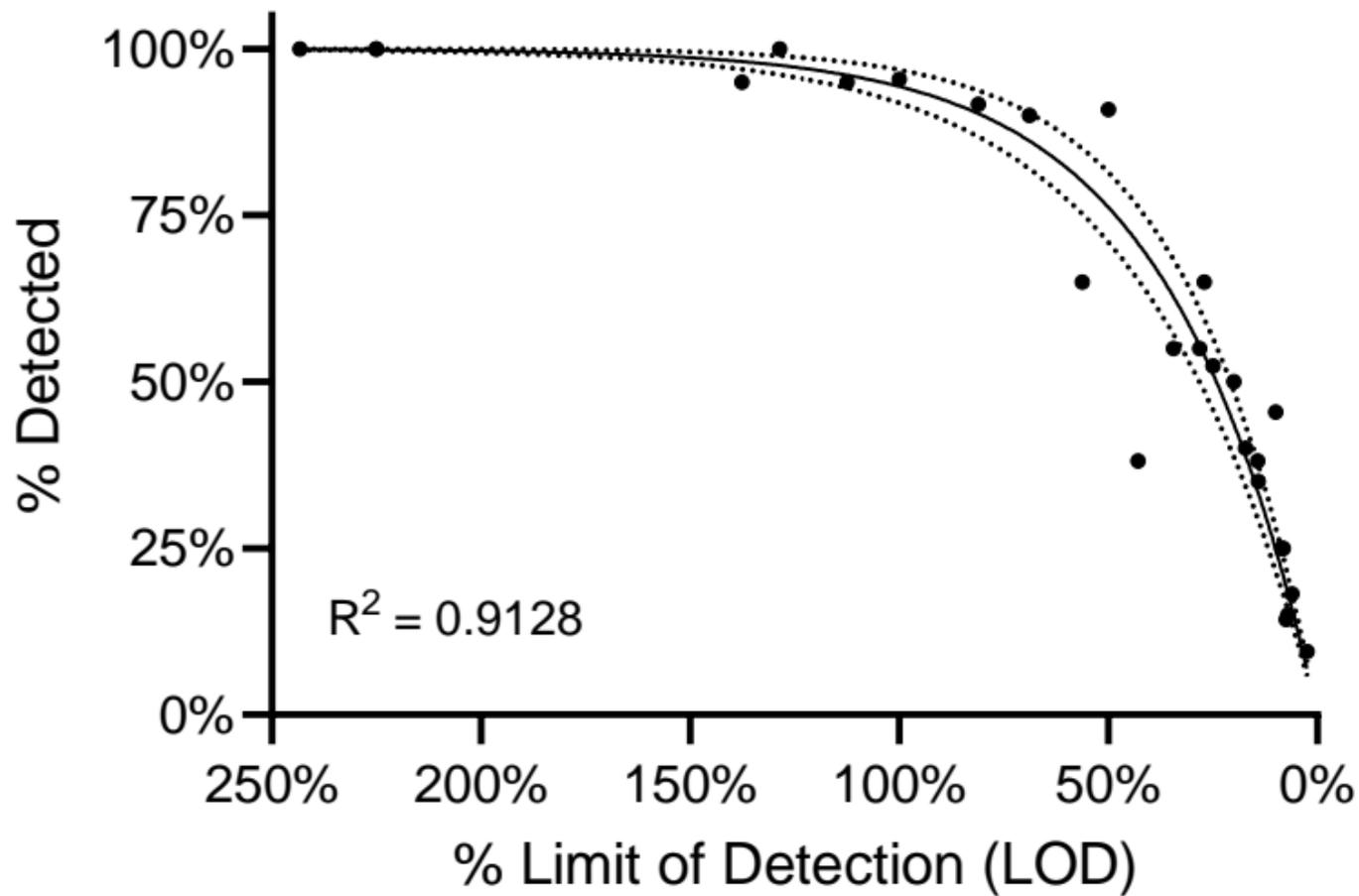
E Use P to randomly classify specimen as:
'detected' or 'not detected'

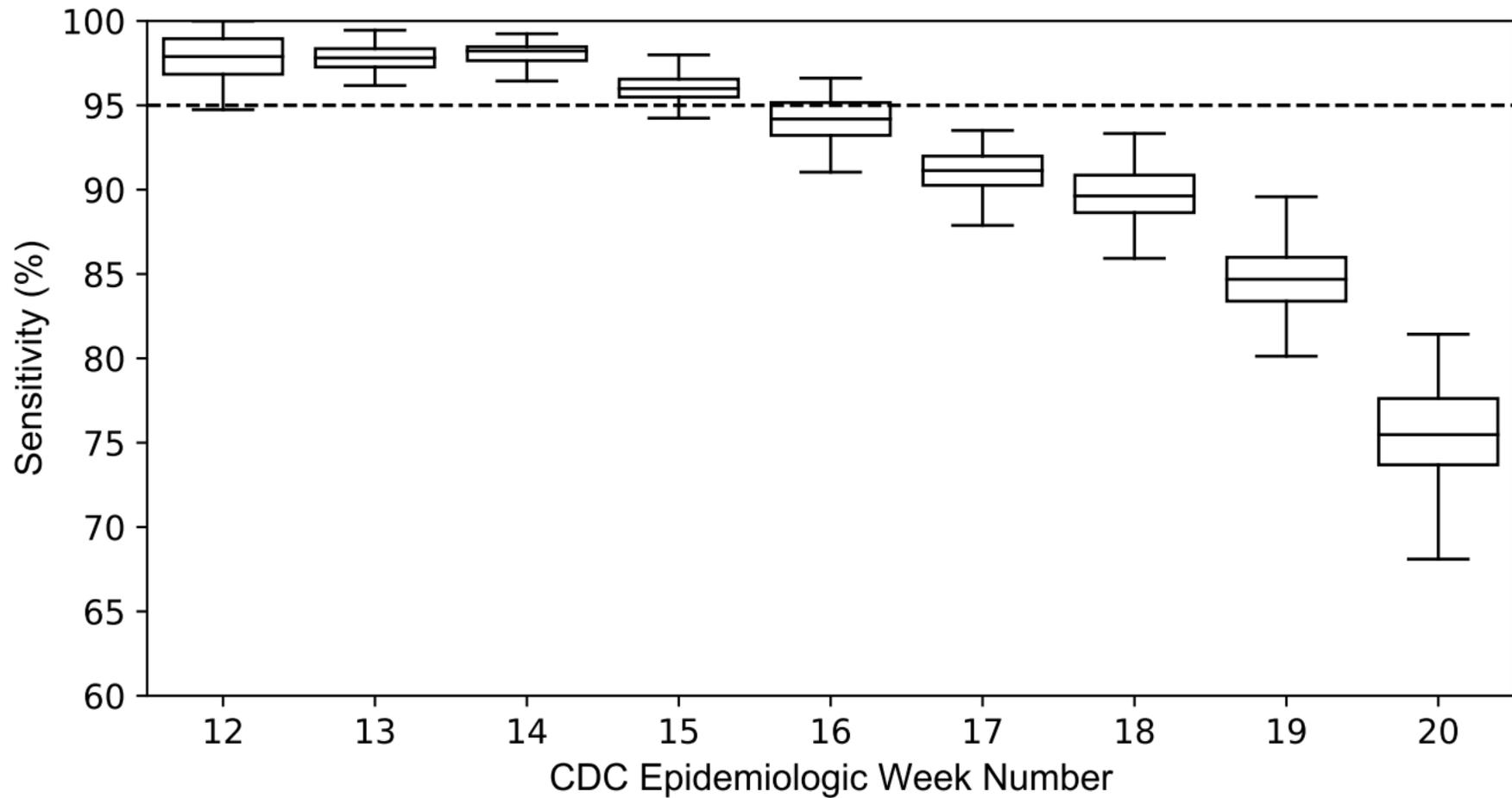
F If 'detected', classify as True Positive
If 'not detected', classify as False Negative

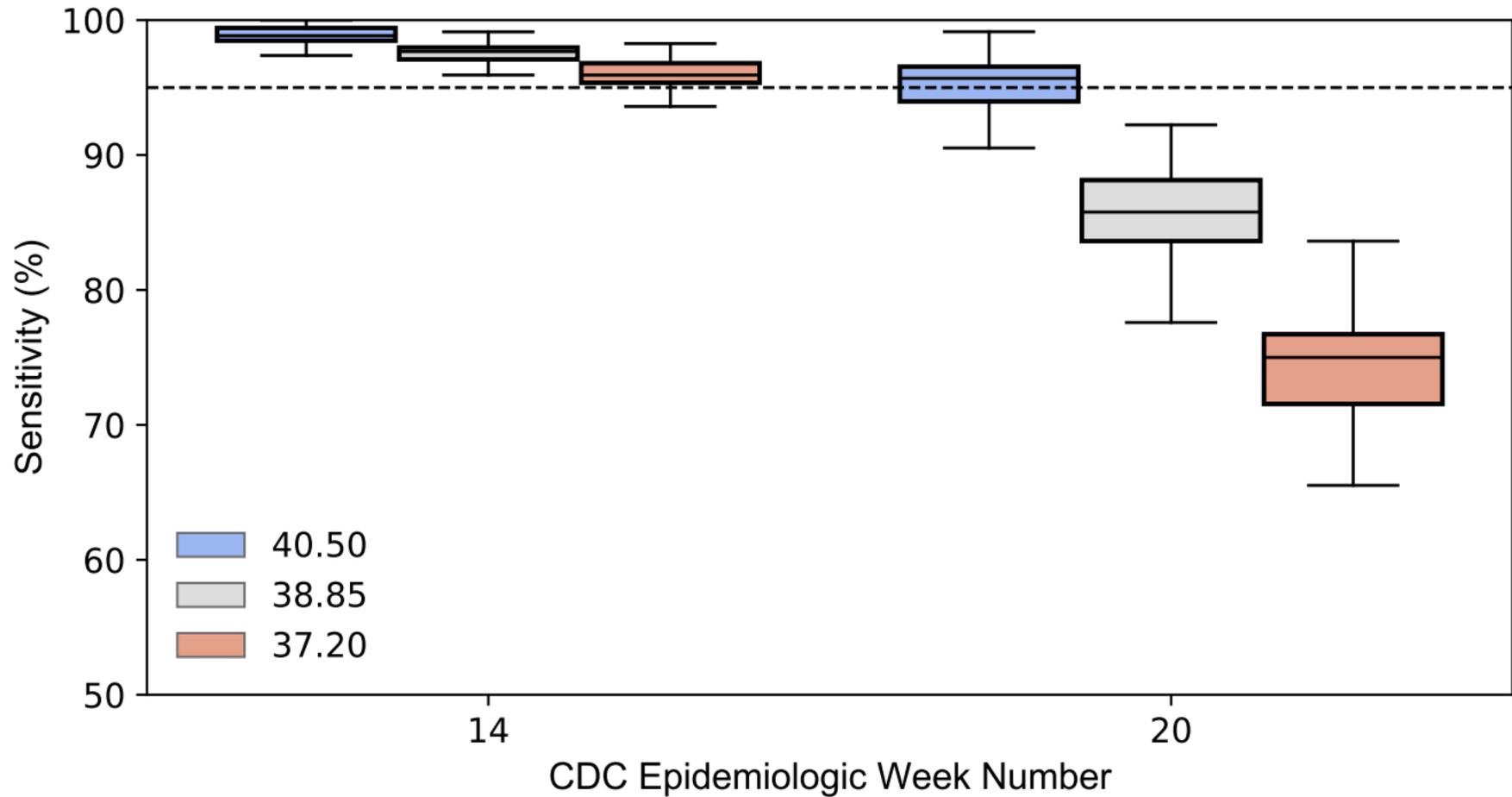
G Place Ct-value back into Week 12 distribution











Supplemental Table 1: Line of best fit equations and the associated 95% confidence intervals for the slope. Derived from a simple linear regression performed on the Ct values measured from 'positive' samples and the associated gene targets.

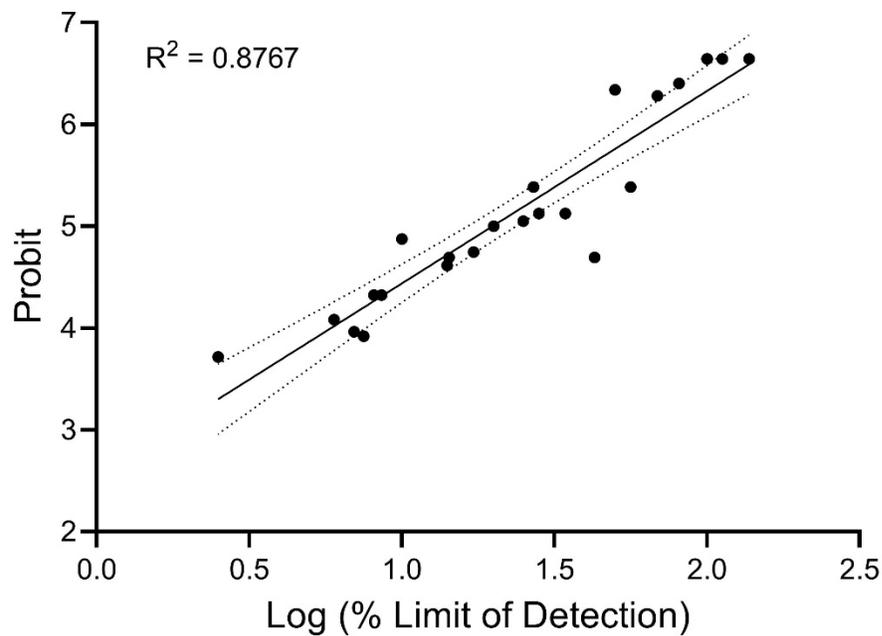
Gene Target	Equation	Slope 95% CI
<i>CDC N1</i>	$Y = 1.169 * X + 24.10$	1.064 to 1.275
<i>CDC N2</i>	$Y = 1.340 * X + 23.83$	1.227 to 1.453
<i>Cepheid E</i>	$Y = 1.046 * X + 23.23$	0.8610 to 1.231
<i>Cepheid N2</i>	$Y = 1.353 * X + 25.62$	1.169 to 1.537
<i>Flu A1</i>	$Y = 0.1036 * X + 22.91$	-0.02070 to 0.2280

Supplemental Table 2: Data extracted from commercial assay instructions for use. Roche cobas – Target 2 concentration 0.003 TCID50/mL of 0.003 TCID50/mL was excluded based on outlier analysis.

Assay	LOD	LOD Measure	Conc Tested	%LOD	Detected	Tested	% Pos
Alinity M	0.0037	TCID50/mL	0.009	243.2%	24	24	100.0%
			0.003	81.1%	22	24	91.7%
			0.001	27.0%	13	20	65.0%
			0.0003	8.1%	6	24	25.0%
Perkin Elmer (N Gene)	24.9	cp/mL	34.25	137.6%	19	20	95.0%
			17.13	68.8%	18	20	90.0%
			8.56	34.4%	11	20	55.0%
			4.28	17.2%	8	20	40.0%
			2.14	8.6%	5	20	25.0%
Perkin Elmer (ORF1ab)	9.3	cp/mL	20.93	225.1%	20	20	100.0%
			10.46	112.5%	19	20	95.0%
			5.23	56.2%	13	20	65.0%
			2.62	28.2%	11	20	55.0%
			1.31	14.1%	7	20	35.0%
			0.65	7.0%	3	20	15.0%
Roche Cobas (Target 1)	0.007	TCID50/mL	0.009	128.6%	21	21	100.0%
			0.003	42.9%	8	21	38.1%
			0.001	14.3%	8	21	38.1%
Roche Cobas (Target 2)	0.004	TCID50/mL	0.009	225.0%	21	21	100.0%
			0.001	25.0%	11	21	52.4%
			0.0003	7.5%	3	21	14.3%
			0.0001	2.5%	2	21	9.5%
Xpert Xpress	0.005	PFU/mL	0.005	100.0%	21	22	95.5%
			0.0025	50.0%	20	22	90.9%
			0.001	20.0%	11	22	50.0%
			0.0005	10.0%	10	22	45.5%
			0.0003	6.0%	4	22	18.2%

Abbreviations: cp = Copies; LOD = Limit of detection; mL = milliliter; PFU = Plaque forming units; Pos = Positive; TCID50 = Median Tissue Culture Infectious Dose;

Supplemental Figure 1: Composite test performance of commercial extracted RT-PCR assays for SARS-CoV-2. Available limit of detection (LOD) data were extracted from publicly available instructions for use (IFU) and normalized to Log %-LOD as described in methods and plotted against probit units determined using GraphPad Prism v8. Best fit line (solid line) and 95% confidence intervals (dashed lines) are shown.



Supplemental Figure 2: Boxplots representing the Ct value distributions for Ct values derived from the Cepheid Xpert Xpress Influenza A / Influenza B assay during flu-season 2019 to 2020. Results are grouped by epidemiologic week number and gene target. Horizontal dashed lines represent the assay LOD.

