Supplementary Information for:

A Deep Learning Approach to Quantitative PCR that Learns from Ground Truth

Huong Vu[1], Alexander Schubert[1,2], Saathvik Selvan[1], Petros Giannikopoulos[1,2], Ahmed Alaa[1,2], and Ziad Obermeyer[2,*]

[1]Department of Statistics, University of California, Berkeley, USA
[2]University of California, San Francisco, USA
[*]Corresponding author: zobermeyer@berkeley.edu

Contents

1 Materials

	Description and Years	**Source of Truth**		**% Positive Curves**	**Counts**	
		Positive	**Negative**		**Samples**	**Curves**
Training						
Ground truth training set	Amplification curves from samples and genes used for quality control purposes during clinical testing for SARS-CoV-2. Samples were drawn from coronavirus testing between August 2020 to March 2022.	**Control genes** RNaseP (IGI LuNER test kit) and MS2 (ThermoFisher TaqPath RT-PCR COVID-19 test kit) from Clinical and Human Normal Negative Control samples. **Target genes** N gene and E gene (LuNER) and S gene, N gene, and ORF1ab (TaqPath) from Positive Control samples.	**Control genes** RNaseP and MS2 from Positive Control and Negative Control samples. **Target genes** N gene, E gene, S gene, and ORF1ab from Human Normal Negative Control and Negative Control samples.	89.6%	N/A	18581
Evaluation						
Ground truth test set	As above.	As above.	As above.	89.9%	N/A	6175
Retested clinical samples	Clinical samples tested for SARS-CoV-2 that were considered inconclusive/invalid and retested.	**Control genes** RNaseP and MS2 from all Clinical samples. **Target genes** N gene, E gene, S gene, and ORF1ab from Clinical samples with a positive final patient results.	**Target genes** N gene, E gene, S gene, and ORF1ab from clinical samples with negative final patient results.	36.0%	277	906
SARS-CoV-2 - Known Dilution sample	A series of synthetic SARS-CoV-2 dilutions (1 copy/μL to $1e^5$ copies/μL) generated using LuNER test kits to test and calibrate analysis systems.	**Target genes** N gene and E gene from known dilution samples.	**Target genes** N gene and E gene from Negative Control samples.	88.9%	104	216
Chip60	Dilution series of human genes MLV-2v and Vimentin were developed by the Lausitz University of Applied Sciences, Senftenberg, Germany and obtained in the *chipPCR* package.	Diluted MLV-2v and Vimentin samples.	Water control curves drawn from the main samples of negative controls.	87.5%	N/A	32

Table S1. Dataset Description

2 Results

2.1 Model Performance: True Positive and True Negative Control Samples

As shown in Table S2, SPARK achieves a substantially lower Brier score on the holdout set compared to other methods. Both qPCRdeepNet and qpcR, which are trained on human-provided labels, perform similarly to threshold rules with Brier scores between 0.0329 and 0.45. In contrast, SPARK achieves a Brier score of just 0.01.

The true positive rate in the hold-out set is 89.9% which is much higher than the population prevalence of SARS-CoV-2. Hence, we re-weight the FNR and FPR calculated on the hold-out set to better approximate the models' FNR and FPR in the population. We estimate the true prevalence to be 9.0% by positive rate of expert judgment in clinical samples. Table S2 shows performance metrics after reweighting to estimated population prevalence. Between different methods, SPARK is the only model maintaining high performance in accuracy and Brier score after reweighting. SPARK's adjusted accuracy is 0.9374 and Brier score is 0.0518 while other methods' adjusted accuracies drop below 0.8 and adjusted Brier scores are above 0.1. As for the threshold rules, because we do not have predicted scores for each sample, we assume threshold rules give a probability score of 1 for positive prediction and 0 for negative prediction to approximate Brier scores.

	Hold-out Set Metrics				**Reweighted Hold-out Set Metrics**			
Model	FNR	FPR	Accuracy	Brier Score	FNR	FPR	Accuracy	Brier Score
Threshold Rules	0.025	0.222	0.955	0.045[1]	0.0023	0.2021	0.7957	0.204[2]
SPARK	0.0077	0.0681	0.9862	0.0104	0.0007	0.0610	0.9374	0.0518
qPCRdeepNet	0.0285	0.2528	0.9488	0.0406	0.0026	0.2301	0.7673	0.1267
qpcR	0.0296	0.2545	0.9476	0.0329	0.0027	0.2316	0.7658	0.1252

Table S2. Model performance comparison on hold-out and reweighted hold-out sets

We calculate the reweighted metrics with the following formulae:
With π = population prevalence; $\hat{p}$ = predicted positive fraction

$$\text{Brier score}_{\text{Threshold Rule}} = \pi \cdot FNR + (1-\pi) \cdot FPR$$

$$\text{Brier score} = \pi B_1 + (1-\pi) B_0 \text{ where } B_1 = \frac{1}{N_1} \sum_{i:y_i=1} (\hat{p}_i - 1)^2, B_0 = \frac{1}{N_0} \sum_{i:y_i=0} \hat{p}_i^2$$

$$Accuracy_{adj.} = \pi \cdot TPR + (1-\pi) \cdot TNR$$

$$FNR_{adj.} = \pi \cdot FNR$$

$$FPR_{adj.} = (1-\pi) \cdot FPR$$

We are also interested in the difference in predictions for negative curves from water samples and asymptomatic samples, LuNER and TaqPath test kit samples and different targets. The following results show the comparison in FNR, FPR, accuracy, Brier score, AUC and AUPRC from different methods between different factors.

In our hold-out sets, there are 254 negative curves from water samples and 128 negative curves from asymptomatic samples. Table S3 shows that between water and asymptomatic samples, there is not much difference in FPR, accuracy and Brier score between different methods. FPR and accuracy are calculated with the highest accuracy classification thresholds.

Model	**Water**			**Asymptomatic**		
	FPR	**Accuracy**	**Brier Score**	**FPR**	**Accuracy**	**Brier Score**
Expert Judgment	0.0	1.0	0.0	0.0	1.0	0.0
SPARK	0.012	0.988	0.011	0.0	1.0	0.002
qpcR	0.051	0.949	0.108	0.039	0.961	0.106
qPCRdeepNet	0.043	0.957	0.101	0.047	0.953	0.141

Table S3. Accuracy of water and asymptomatic samples

In our holdout set, we have 4536 curves from LuNER test kits with a true positive rate of 90.6% and 1552 curves from TaqPath test kits with a true positive rate of 87.7%. From Table S4, we notice substantial differences in FPR for SPARK and qpcR. For SPARK, the difference in FPR for curves from LuNER and TaqPath test kits is 0.042 and is statistically significant with two-proportion z-test p-value 0.037.

There is a significant difference in false positive rates (FPR) between amplification curves from the TaqPath and LuNER test kits. Additionally, the majority of true positive curves in our holdout sets are control gene curves. To understand how these factors affect model performance, we investigate the

Model	FNR	FPR	Accuracy	AUC	AUPRC	Brier Score
LuNER Test Kits						
Expert Judgment	0.044	0.0	0.961	0.978	0.996	0.039
SPARK	0.008	0.042	0.989	0.998	1.0	0.009
qpcR	0.032	0.122	0.959	0.986	0.999	0.028
qPCRdeepNet	0.035	0.054	0.963	0.982	0.998	0.040
TaqPath Test Kits						
Expert Judgment	0.043	0.0	0.962	0.978	0.995	0.038
SPARK	0.010	0.084	0.981	0.998	1.0	0.014
qpcR	0.059	0.037	0.944	0.973	0.996	0.046
qPCRdeepNet	0.053	0.047	0.948	0.977	0.997	0.041

Table S4. Accuracy comparison between LuNER and TaqPath test kit samples.

differences in evaluation metrics across target genes. Table S5 summarizes the FNR, FPR, accuracy, AUC, AUPRC, and Brier score for four methods—Expert Judgment, SPARK, qpcR, and qPCRdeepNet—across different target genes. To assess the statistical significance of performance differences, we conducted pairwise two-proportion z-tests on FNR, FPR, and accuracy for SPARK's predictions across all target gene combinations. Results are visualized as heatmaps in Figure S1, where:

- Tile color encodes the p-value of the test.
- Tile text shows the absolute difference in metric values between the column and row gene pairs.

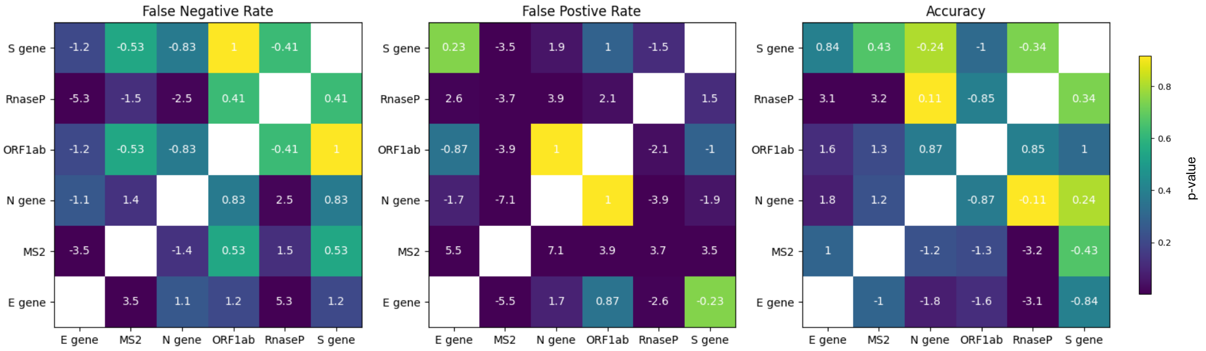

Figure S1. Pairwise two proportion z test on FNR, FPR, and accuracy for SPARK's predictions across all target gene combinations. Tile color encodes the p-value of the test. Tile text shows the absolute difference in metric values between the column and row gene pairs

2.2 Model Prediction on Retest Samples

Table S6 provides a comprehensive overview of each model's performance on the retest dataset. Among the 906 amplification curves included in our analysis, 667 were labeled as invalid, 236 as inconclusive, and 3 as negative in the initial PCR run. While we were unable to determine why these three negative-labeled samples were selected for retesting, their final patient results were reported as positive. To further investigate model performance under different retest conditions, Table S7 reports the FNR, FPR, and accuracy for each method, separately for inconclusive and invalid samples. These metrics are computed using the optimal threshold for accuracy on the full retest dataset, applied within each subset.

Model	FNR	FPR	Accuracy	AUC	AUPRC	Brier Score
RNaseP (Control)						
Expert Judgment	0.045	0	0.957	0.978	0.998	0.043
SPARK	0.006	0.088	0.990	0.997	1.0	0.008
qpcR	0.033	0.225	0.959	0.977	0.999	0.025
qPCRdeepNet	0.035	0.077	0.963	0.976	0.999	0.037
MS2 (Control)						
Expert Judgment	0.046	0	0.956	0.977	0.998	0.044
SPARK	0.001	0.283	0.979	0.992	1.0	0.015
qpcR	0.059	0	0.943	0.999	1.0	0.039
qPCRdeepNet	0.056	0.019	0.945	0.983	0.999	0.037
E gene (Target)						
Expert Judgment	0.032	0	0.986	0.984	0.982	0.014
SPARK	0.053	0.016	0.968	0.995	0.995	0.032
qpcR	0.053	0.041	0.954	0.989	0.988	0.069
qPCRdeepNet	0.074	0.066	0.931	0.983	0.982	0.011
N gene (Target)						
Expert Judgment	0	0	1.0	1.0	1.0	0
SPARK	0.025	0	0.990	1.0	1.0	0.004
qpcR	0	0.048	0.972	1.0	0.999	0.066
qPCRdeepNet	0	0.018	0.990	1.0	1.0	0.032
ORF1ab (Target)						
Expert Judgment	0	0	1.0	1.0	1.0	0
SPARK	0	0	1.0	1.0	1.0	0
qpcR	0	0	1.0	1.0	1.0	0.057
qPCRdeepNet	0	0	1.0	1.0	1.0	0.006
S gene (Target)						
Expert Judgment	0	0	1.0	1.0	1.0	0
SPARK	0	0.022	0.986	1.0	1.0	0.013
qpcR	0.148	0.109	0.877	0.853	0.907	0.132
qPCRdeepNet	0	0.130	0.918	0.998	0.996	0.174

Table S5. Performance comparison of four models across different genes and controls.

Model	FNR	FPR	Accuracy	AUC	AUPRC	Brier Score
Expert Judgment	0.368	0.160	0.765	0.736	0.568	0.235
SPARK	0.144	0.047	0.918	0.936	0.919	0.071
qpcR	0.334	0.181	0.764	0.815	0.767	0.205
qPCRdeepNet	0.340	0.214	0.741	0.774	0.765	0.280

Table S6. Brier Score of models

Model	**Invalid**			**Inconclusive**		
	FNR	**FPR**	**Accuracy**	**FNR**	**FPR**	**Accuracy**
Threshold Rules	0.854	0.005	0.699	1.0	0.0	0.619
SPARK	0.082	0.030	0.952	0.167	0.144	0.847
qpcR	0.519	0.037	0.795	0.333	0.171	0.767
qPCRdeepNet	0.472	0.081	0.783	0.144	0.274	0.775

Table S7. Accuracy of invalid and inconclusive samples

2.3 Model Prediction on Clinical Samples

Figure S2 shows amplification plots of clinical samples and the corresponding expert judgments and SPARK predictions.

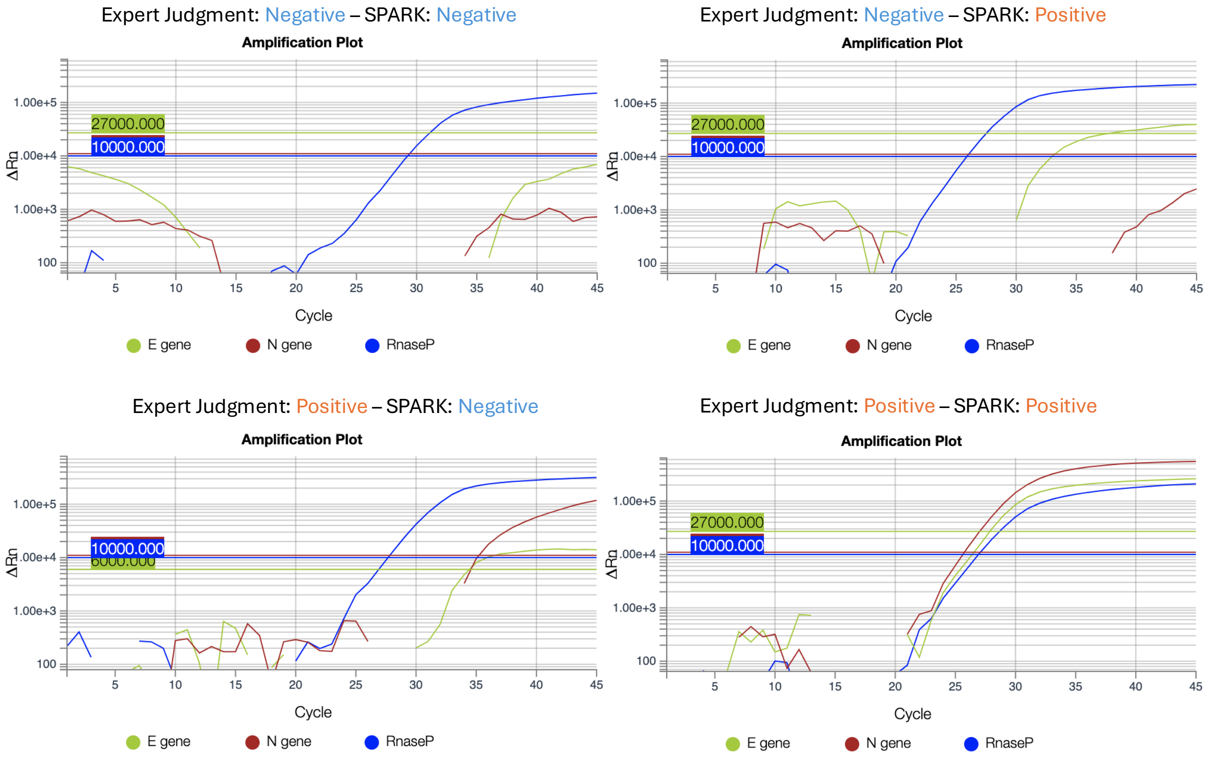

Figure S2. Amplification plots of clinical samples and corresponding expert judgments and SPARK predictions