

# The Coviral Portal: Multi-Cohort Viral Loads and Antigen-Test Virtual Trials for COVID-19

Alexandra Morgan,<sup>1</sup> Elisa Contreras,<sup>1</sup> Michie Yasuda,<sup>1</sup> Sanjucta Dutta,<sup>1</sup> Donald Hamel,<sup>3</sup> Tarini Shankar,<sup>1,4</sup> Diane Balallo,<sup>1,5</sup> Stefan Riedel,<sup>1,2</sup> James E. Kirby,<sup>1,2</sup> Phyllis J. Kanki,<sup>3</sup> and Ramy Arnaout<sup>1,2,6†\*</sup>

<sup>1</sup>Beth Israel Deaconess Medical Center, Division of Clinical Pathology, Department of Pathology, Boston, MA 02215, USA; <sup>2</sup>Harvard Medical School, Boston, MA 02215, USA; <sup>3</sup>Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA; <sup>4</sup>Northeastern University, Boston, MA 02215, USA; <sup>5</sup>Boston College, Chestnut Hill, MA 02467, USA; <sup>6</sup>Beth Israel Deaconess Medical Center, Division of Clinical Informatics, Department of Medicine, Boston, MA 02215, USA

†Corresponding author and

\*Lead Contact:

Ramy Arnaout

Beth Israel Deaconess Medical Center

330 Brookline Avenue, DA709

Boston, MA 02215

[rarnaout@bidmc.harvard.edu](mailto:rarnaout@bidmc.harvard.edu)

617-538-5681

## 19 **Abstract**

20 **Background.** Regulatory approval of new over-the-counter tests for infectious agents such as  
21 SARS-CoV-2 has historically required that clinical trials include diverse groups of specific patient  
22 populations, making the approval process slow and expensive. Showing that populations do not  
23 differ in their viral loads—the key factor determining test performance—could expedite the  
24 evaluation of new tests.

25 **Methods.** 46,726 RT-qPCR-positive SARS-CoV-2 viral loads were annotated with patient  
26 demographics and health status. Real-world performance of two commercially available  
27 antigen tests was evaluated over a wide range of viral loads. An open-access web portal was  
28 created allowing comparisons of viral-load distributions across patient groups and application  
29 of antigen-test performance characteristics to patient distributions to predict antigen-test  
30 performance on these groups.

31 **Findings.** In several cases distributions were surprisingly similar where a difference was  
32 expected (e.g. smokers vs. non-smokers); in other cases there was a difference that was the  
33 opposite direction from expectations (e.g. higher in patients who identified as White vs. Black).  
34 Sensitivity and specificity of antigen tests for detecting contagiousness were similar across most  
35 groups. The portal is at <https://arnaoutlab.org/coviral/>.

36 **Conclusions.** In silico analyses of large-scale, real-world clinical data repositories can serve as a  
37 timely evidence-based proxy for dedicated trials of antigen tests for specific populations. Free  
38 availability of richly annotated data facilitates large-scale hypothesis generation and testing.

39 **Funding.** Funded by the Reagan-Udall Foundation for the FDA (RA and JEK) and via a Novel  
40 Therapeutics Delivery Grant from the Massachusetts Life Sciences Center (JEK).

41 **Keywords:** SARS-CoV-2, COVID-19, RT-qPCR, viral load, antigen test, precision medicine,  
42 comorbidities, big data, real-world data

## 43 **Introduction**

44 Diagnosis of new infectious pathogens such as SARS-CoV-2 requires development of new  
45 diagnostic tests, which must be evaluated and approved by regulatory agencies before they can  
46 be used for patient care. Such tests include over-the-counter (OTC) antigen tests, which have  
47 been widely used for at-home testing in the context of COVID-19. In order to be approved, a  
48 new test must demonstrate a minimum level of clinical performance. Performance is typically  
49 measured as the test's sensitivity, defined as the proportion of true-positive samples that have  
50 a positive result, and its specificity, defined as the proportion of true negatives that have a  
51 negative result. Clinical performance must be demonstrated in a defined patient population or  
52 group and clinical context, for example inpatients as opposed to outpatients. However, at the  
53 start of an outbreak, epidemic, or pandemic, there may not be enough information to know  
54 whether a test can be expected to perform differently in some patient groups vs. others.  
55 Therefore at the start of development, a new diagnostic test may be approved based on its  
56 performance in the general population, not specific groups.

57 In contrast, as time goes on, evidence for clinical differences among specific groups may  
58 emerge. As this happens, it becomes reasonable to ask whether a test might perform  
59 differently in specific groups, with important implications for how and potentially even whether  
60 that test should be used in a given clinical scenario. Ideally, this question would be answered by  
61 conducting dedicated trials of the new diagnostic test in each group of interest. Unfortunately,  
62 trials are expensive and slow. Also, the number of specific groups that may be of interest is  
63 large, since specific subgroups can be defined not only based on demographics (such as age,  
64 race, and gender), comorbidities (such as diabetes, heart disease, or immunosuppression), and  
65 care settings (inpatient vs. outpatient vs. emergency room) but also by the many possible  
66 combinations of each of these characteristics, which is an essential component of precision  
67 medicine.<sup>1</sup> As a result, in practice it is prohibitively difficult to perform many separate trials on  
68 specific groups for even a single diagnostic test, much less for the many tests that are likely to  
69 be developed in response to a large-scale outbreak, such as has happened in response to the  
70 COVID-19 pandemic. This is a problem for regulators, clinicians, and patients alike.

71 One solution is to apply a new test’s various performance characteristics to real-world data  
72 collected in the course of patient care. Such characteristics include results of existing trials as  
73 well as analytical (i.e. pre-clinical) operating parameters such as the limit of detection (LOD).  
74 The LOD is defined as the lowest concentration of virus that the test can detect in 95% of  
75 replicates. It is routinely determined by manufacturers and validated by clinical laboratories  
76 before a test is put to use clinically.<sup>2</sup> The relationship between concentration and detection is  
77 usually understood to follow an S-shaped curve;<sup>3</sup> fitting it requires at least one additional  
78 datapoint besides the LOD. The concentration of the virus may be measured as the viral load,  
79 most often defined as the number of copies of viral mRNA per milliliter of testing material  
80 (copies/mL).

81 “Real-world data” means the viral-load result of a reference diagnostic test that has already  
82 been approved for the general population. Because this data is from the general population, it  
83 will presumably include results on many specific patient groups. One can apply the  
84 performance of the new test as described above to the set or “distribution” of viral loads from a  
85 group to predict what proportion of patients in the group would have tested positive with the  
86 new test. This proportion is the sensitivity of the new test for that group. In this way, one can  
87 estimate clinical sensitivity without needing a dedicated trial on that group.

88 In this study, we apply this approach to COVID-19. We use the 46,726 positive SARS-CoV-2 RT-  
89 qPCR results our institution has performed as of this writing and use our electronic health  
90 record to annotate each result according to the patient’s demographics, comorbidities, and so  
91 on. Importantly, we convert each PCR result from a Ct value to a viral load using robust (100%  
92 code-coverage) and accurate publicly available software, as previously described.<sup>2,4-6</sup> Although  
93 PCR results are typically reported simply as positive or negative, qPCR is intrinsically  
94 quantitative (the “q” in “qPCR”); we make use of this quantitative information in its natural unit  
95 of measure (viral load). This is in contrast to Ct values, which are less useful because they vary  
96 inversely with viral load and correspond to different viral loads on different PCR testing  
97 platforms.

98 We focus especially on sensitivity and specificity for infectivity or contagiousness.  
99 Contagiousness is of special interest given the public health focus on using quick, inexpensive  
100 tests to curtail community transmission in a pandemic.<sup>7,8</sup> Treating contagiousness as a function  
101 of viral load, contagiousness can be estimated using a virus culture assay we previously  
102 described in which a positive patient sample is applied to susceptible cells and monitored for  
103 virus replication. After an initial adsorption period, the cells are washed free of the initial virus  
104 to eliminate carryover. The supernatant is then sampled on a timescale of days and tested by  
105 PCR for the presence of new virus.<sup>9,10</sup> The lowest concentration of virus in a patient sample  
106 from which new virus can be recovered is the contagiousness threshold. Because cells in culture  
107 have no physical or distance barriers, no mucociliary elevator, and no protection via  
108 medications or an immune system, we consider this threshold a conservative estimate. We  
109 previously demonstrated this threshold is approximately 50,000 copies/mL and has been fairly  
110 stable even as the SARS-CoV-2 virus has evolved.<sup>9,10</sup>

## 111 **Materials and Methods**

112 **Institutional review.** Institutional Review Board approval was obtained for all described work  
113 under Beth Israel Lahey Health (BILH) IRBs 2022P000328 and 2022P000288. The Harvard T. H.  
114 Chan School of Public Health IRB20-1979 provided non-human subjects research determination  
115 for virus culture work.

116 **Defining specific patient groups.** We extracted the following information from our hospital's  
117 clinical-research data repository: demographics (age, gender, and self-reported race/ethnicity),  
118 socioeconomic status (using the median neighborhood household income for the patient's ZIP  
119 code, obtained via the 2020 U.S. census, as a proxy), care setting (inpatient, outpatient,  
120 emergency ward, or other institution), presentation/disposition (based on vital signs, which we  
121 combined into a measure of initial presentation), outcome (survived, died with COVID-19 as the  
122 cause of death, died with COVID-19 as an incidental finding), vaccination status (vaccinated,  
123 unvaccinated, or unknown), treatment (CPT-encoded procedures, remdesivir (GS-5734; Gilead  
124 Sciences, Foster City, CA) administration, steroid administration), comorbidities (according to

125 the Charlson Comorbidity Index:<sup>11</sup> body-mass index, diabetes, chronic heart disease, chronic  
126 lung disease, chronic renal disease, liver disease, dementia, chronic neurological conditions,  
127 connective-tissue disease, human immunodeficiency virus (HIV), and malignancy), and  
128 immunosuppression status<sup>12</sup> (CD4+ T-count <100 cells/ $\mu$ L, hematologic malignancy,  
129 chemo/immuno-modulating agent alone or in setting of solid malignancy, organ transplant, or  
130 rheumatologic/inflammatory condition). The rationale for extracting these data items  
131 specifically was twofold: first, this list includes the complete COVID-19 core diagnostic data at  
132 federal and state levels; second, it includes data necessary for calculating the well validated 4C  
133 Mortality Score for SARS-CoV-2.<sup>13</sup> ICD-10 codes corresponding to the listed comorbidities were  
134 determined by a physician (Dr. Arnaout) following prior methodologies<sup>14</sup> but updated for 2022-  
135 2023.

136 At presentation, patients were considered sick if any of the following were true within 1 day of  
137 the PCR test sample: systolic blood pressure <90 mmHg, diastolic blood pressure <60 mmHg,  
138 heart rate >100 beats per minute, respiratory rate >18 breaths per minute, or temperature  
139 >99.1°F. They were otherwise considered well, with the exception that if no values were  
140 recorded (NULL in the data repository) for all criteria, presentation was considered unknown  
141 and therefore not assigned.

142 Patients were designated as immunocompromised at the time of PCR testing if one of the  
143 following were true: on their most recent T-cell subset analysis report, their absolute CD4+ cell  
144 count was <100 cells/ $\mu$ l; they had a diagnosis of either lymphoma or leukemia associated with a  
145 healthcare encounter (visit, admission, or phone call) either before the PCR test or within 60  
146 days after the PCR test; they were on any of the following medications on an ongoing basis,  
147 prescribed prior to the PCR test and with enough refills to include the time up to 30 days prior  
148 to the PCR test: abatacept, adalimumab, anakinra, azathioprine, basiliximab, budesonide,  
149 certolizumab, cyclosporine, daclizumab, dexamethasone, everolimus, etanercept, golimumab,  
150 infliximab, ixekizumab, leflunomide, lenalidomide, methotrexate, mycophenolate, natalizumab,  
151 pomalidomide, prednisone, rituximab, secukinumab, serolimus, tacrolimus, tocilizumab,

152 tofacitinib, ustekinumab, or vedolizumab. Otherwise, they were designated not  
153 immunocompromised.

154 Supplementary Table 1 provides further details for the above methods.

155 **Viral load.** The SARS-CoV-2 RT-qPCR testing in this study was performed on three Abbott  
156 Molecular platforms: m2000, Alinity m, and Alinity 4-Plex (Abbott Molecular, Des Plaines, IL,  
157 U.S.A.). These detect identical SARS-CoV-2 N and RdRp gene targets. They are extremely  
158 sensitive, with LOD of ~100 copies/mL. They output a quantitative fractional cycle number  
159 (FCN), a type of Ct value described in detail elsewhere.<sup>15</sup> Together these platforms accounted  
160 for 46,726 positive tests.

161 Ct values were converted to viral loads in units of copies of viral mRNA per mL using the public  
162 Python package *ct2vl* as previously reported.<sup>4</sup> Briefly, this software was validated via calibration  
163 curves established for all platforms using an extended SeraCare panel (LGC Seracare, Milford,  
164 MA) panel based on a SARS-CoV-2 genome incorporated into replication-incompetent,  
165 enveloped Sindbis virus and calibrated based on digital PCR at US National Institutes of  
166 Standards and Technology (NIST) and LGC/Seracare.<sup>16</sup> Validation material ranged in viral load  
167 from 300 to 10<sup>6</sup> viral genome copies/mL. Results were harmonized with the cycle threshold for  
168 a spiked internal control also amplified in each SARS-CoV-2 assay to confirm lack of PCR  
169 inhibition and accurate viral load output. The standards, modeling SARS-CoV-2 virus, were run  
170 through all stages of sample preparation and extraction to allow appropriate comparison with  
171 identically processed patient samples.  $R^2$  was ~0.99 for all calibration determinations, indicating  
172 assays are robustly quantitative.

173 **Presumed SARS CoV-2 variant.** Presumed variant was inferred from the date of sample  
174 collection based on the data presented by Covariants (<https://covariants.org>) showing the  
175 frequency of sequencing particular variants in Massachusetts, the United States, and other  
176 locations.<sup>17</sup> Specimens from before June 7, 2021 were annotated as being an early variant.  
177 Specimens from between July 7, 2021 and December 6, 2021 were annotated delta variant.

178 Specimens from after January 3, 2022 were annotated as omicron variant. Results from the  
179 month between windows, when more than one major variant was common, were not  
180 annotated with a presumed variant and are omitted from by-variant comparisons.

181 **Evaluation of antigen tests vs. PCR.** Patients seeking COVID testing at a drive-through testing  
182 site near Boston affiliated with our medical center<sup>5,6</sup> between May 23 and November 4 of 2022  
183 were offered the opportunity to participate in a separate arm, providing a comparative, parallel  
184 prospective study. This represented community testing for both symptomatic and  
185 asymptomatic individuals with diverse demographics (age, race, sex, socio-economic status).  
186 Each patient who consented had both a standard-of-care PCR test and two OTC antigen tests  
187 performed (Abbott BinaxNow COVID-19 Ag Card and CareStart COVID-19 Antigen Home Test).  
188 The PCR test was performed on material collected with a nasopharyngeal swab. SARS-CoV-2 RT-  
189 qPCR testing was performed using the Abbott m2000 RealTime or Alinity m SARS-CoV-2 assays  
190 according to the manufacturer's instructions, yielding, for each positive sample, a Ct value  
191 which was converted to viral load as previously described. Specimens for the antigen tests were  
192 collected with separate nasal swabs for each test, according to the manufacturer's instructions.  
193 These were collected and the tests performed by study personnel after informed consent was  
194 obtained on-site within the time-frame constraints detailed in each test's instructions for use,  
195 as per IRB. In order to extrapolate antigen-test performance from this subset to all patients,  
196 positivity vs. viral load was modeled by logistic regression (the LogisticRegression function in  
197 Python's scikit-learn library).<sup>18</sup> LogisticRegression converges on optimal parameters in a model  
198 predicting the probability of a positive test based on viral load. Parameters were predicted  
199 separately for each test. The equation for probability was a standard sigmoid constrained to the  
200 range 0-1 (i.e., the lowest probability is zero and the highest probability is 1):  $p(\text{test success}) =$   
201  $\frac{1}{1+e^{-k(v-v_0)}}$  where  $v$ , the independent variable, is  $\log_{10}$  of the viral load. This constraint leaves  
202 two free parameters:  $v_0$  is the midpoint, i.e. the model's estimate of where the success rate  
203 passes 50%, while  $k$  controls the steepness, i.e. the change in viral load to change in probability  
204 of being positive.



205 **Contagiousness.** Samples were then stored at 4°C until contagiousness testing. This was done  
206 within a four-day time period. We previously validated that freeze-thaw does not impact viral  
207 viability and will bank remaining samples for future investigations. Quantitative viral culture  
208 was performed on a random sample of PCR-positive samples. Vero E6 cells (ATCC CRL-1586)  
209 were seeded on a 6-well flat bottom plate at  $0.3 \times 10^6$  cells per well in Eagle's minimum essential  
210 media (EMEM) containing 1% antibiotic-antimycotic, 1% HEPES and 5% fetal calf serum (FCS,  
211 Gibco) grown to confluence at approximately  $1 \times 10^6$  cells per well, inoculated with 250µL of  
212 patient sample, and incubated at 37°C for 24 hours for viral adsorption, as previously  
213 described.<sup>10,19–21</sup> Carryover of non-viable viral RNA present in samples was limited by washing  
214 cell cultures after the 24-hour viral adsorption and adding fresh EMEM composite media with  
215 reduced FCS to 2% for viral growth, meaning detectable virus represents viable replicating virus.  
216 On days 3 and 6, cell culture supernatant was removed and added to 800µL of VXL buffer  
217 (QIAGEN, German, MD) (1:1 ratio) for subsequent nucleic acid extraction and detection of virus  
218 by PCR. Viral load in culture supernatants on days 3 and 6 served as a quantitative surrogate for  
219 viable (i.e. replication-competent) virus in the patient sample and provided a measure of the  
220 magnitude of sample infectivity. SARS-CoV-2 RT-qPCR testing of Vero cell culture supernatants  
221 was performed using the Abbott m2000 Real-Time or Alinity m SARS-CoV-2 assays according to  
222 the manufacturer's instructions. The contagiousness threshold was determined by the  
223 threshold patient-sample viral load value resulting in detectable culture viral load.

224 **Whole-genome viral NGS.** Next-generation-sequencing (NGS)-based sequencing of select PCR-  
225 positive samples from the viral antigen evaluation study was performed as follows. Full-length  
226 SARS-CoV-2 viral genome sequencing was performed on the Oxford Nanopore MinION system  
227 ( $\geq R9.4$  flowcell; Oxford Nanopore Technologies-ONT, Oxford, UK) using the guppy basecaller  
228 and the downstream ARTIC network bioinformatics pipeline for genome assembly.<sup>22,23</sup> The  
229 workflow was run on a 2021 Intel Core i9-11900 Rocket Lake 3.5GHz 8-core LGA 1200 boxed  
230 processor with NVIDIA A5000 GPU. Standard coverage and quality metrics and plots were  
231 produced, single-nucleotide variants were recorded, and variants assigned using NextClade.<sup>24</sup>

232 **Web portal and privacy protection.** The portal was written using Svelte and d3 for the  
233 interactive frontend and Python run against a Postgres database for the backend. To reduce re-  
234 identification risk, ages were jittered by adjusting the patient's date of birth by a random  
235 number of days (drawn from a Gaussian distribution with a standard deviation of two years)  
236 before calculating patient's age at the time of each test. Groups smaller than 4-8 patients are  
237 suppressed and therefore not viewable. Revealing exact sizes of such small groups defined by  
238 multiple patient characteristics would pose a re-identification risk. To prevent inferring the sizes  
239 of these groups by subtraction of viewable group sizes, viewable group sizes are jittered by  
240 dropping approximately 0.5-1% of the data on any split by patient feature. To maximize  
241 consistency of the results of jittering as data are updated, jittering was performed using  
242 random number seeds based on pseudo-identifiers (which are never uploaded and thus  
243 inaccessible to/safe from the web client). For ease of visualization, plots of viral load  
244 distributions are shown as kernel-density estimates (i.e. smoothed) using a Gaussian kernel of  
245 width 0.25 log<sub>10</sub> viral load units (~1.7-fold).

246 **Statistical tests.** The geometric mean viral load for each patient group was calculated as a  
247 summary statistic. The geometric (as opposed to arithmetic) mean was chosen because viral  
248 loads vary over many orders of magnitude.<sup>2</sup> The Kolmogorov-Smirnov test (KS;  
249 `scipy.stats.kstest`) was used to compare distributions. This test was used because data were not  
250 distributed normally and KS does not require normality (unlike, for example, the t-test, which  
251 requires normal distributions). KS tests the null hypothesis that the distributions of viral loads  
252 for two patient groups are statistically indistinguishable.<sup>25</sup> The p-value gives the probability that  
253 distributions from the two groups are drawn from the same underlying distribution. A large p-  
254 value means the two groups are statistically indistinguishable; a small p-value means they are  
255 different. Interpretation of p-values as significant vs. not significant requires a significance  
256 threshold, which requires correction for multiple comparisons if multiple comparisons are  
257 performed.<sup>26,27</sup> Because the number of comparisons performed via the web portal is up to the  
258 user, un-corrected p-values are reported, with interpretation as significant or not significant left  
259 to the user.

260 **Software and hardware.** Data extraction, annotation, statistics, and analyses were performed  
261 using standard Unix tools and Python 3.9+ using the pandas, numpy, scipy, and scikit libraries  
262 and the interactive Jupyter notebook environment. Figures were created using Python graphics  
263 libraries matplotlib and seaborn, and OmniGraffle 7 (The Omni Group, Seattle, WA).

## 264 **Results**

265 **A web portal for large-scale real-world SARS-CoV-2 viral load results for different patient**  
266 **groups.** 46,726 COVID-19 PCR results representing approximately 39,180 unique individuals  
267 were converted to viral loads and annotated for patient demographics, comorbidities,  
268 presentation, treatment, and socioeconomic status and made available for interactive  
269 investigation via a public web portal at <https://arnaoutlab.org/coviral/> (Table 1). The portal<sup>28,29</sup>  
270 allows users to visualize the viral load distribution for any patient group, to compare  
271 distributions between groups, and to estimate, for each group, the sensitivity and specificity of  
272 a given OTC test for detecting contagious individuals. Users can define and compare complex  
273 subgroups by selecting multiple characteristics via checkboxes in the user interface (Fig. 1). In  
274 this work, all the figures that contain distributions are direct screenshots from the portal.

275 **Overall viral load distributions.** Viral loads varied over nearly 10 orders of magnitude, from 7  
276 copies/mL (the lowest our system will report) to 1.5 billion copies/mL (99<sup>th</sup> percentile). This  
277 extraordinary range is consistent with observations from early in the pandemic (spring-summer  
278 of 2020).<sup>2</sup> The referenced early observations suggested that viral loads were, to a good  
279 approximation, distributed fairly uniformly over the range. In contrast, the current dataset,  
280 which is ten times as large (46,726 results vs. 4,774 in the previous work<sup>2</sup>), demonstrates clear  
281 bimodality: patients' viral loads tend to be either very low, with a peak around the LOD of 100  
282 copies/mL, or else very high, with a peak around 100 million copies/mL (Fig. 2). This bimodality  
283 is apparent in retrospect (e.g., Fig. 2a of Arnaout et al. 2021<sup>2</sup>) but required a large dataset to  
284 visualize clearly. Further research is needed to understand the reason(s) for these two peaks.

285 **Viral load comparisons among patient groups: remdesivir treatment and patient**

286 **presentation.** The web portal allows statistical comparisons of many thousands of specific  
287 patient groups. Here we describe several examples that are illustrative of the questions that  
288 can be asked and answered using this resource. Remdesivir (Gilead Sciences, Foster City, CA) is  
289 an intravenously administered RNA polymerase inhibitor initially approved by the FDA for  
290 treatment of SARS-CoV-2 in hospitalized adults and adolescents.<sup>30</sup> Of the 46,726 test results in  
291 our dataset, 688 were from patients who then received remdesivir treatment. In practice, at  
292 our institution, remdesivir was used for sicker patients. We hypothesized that viral loads would  
293 be higher on average in patients who received remdesivir and in sicker patients. The data  
294 supported this hypothesis: viral loads were higher on average in both remdesivir-receiving and  
295 sicker-appearing patients, with means of  $8.6 \times 10^5$  copies/mL in patients who received remdesivir  
296 vs.  $1.2 \times 10^5$  in those who did not (Fig. 3a) and  $9.4 \times 10^4$  in sick-appearing patients vs.  $4.1 \times 10^4$  in  
297 well-appearing patients (Fig. 3b). In both cases, the difference was due to a greater fraction  
298 patients in the high-viral-load peak. This was especially clear in the remdesivir comparison (Fig.  
299 3a). In each case, the KS p-value was  $4.3 \times 10^{-16}$ , which we interpret as rejecting the null  
300 hypothesis of no difference, with high confidence. These are examples in which the data  
301 confirmed hypotheses regarding differences in viral load.

302 **Unexpected findings: pulmonary disease.** Serious cases of COVID-19 are marked by life-  
303 threatening respiratory distress. This evolution became less common with the emergence of  
304 the omicron strain and the increasing prevalence of prior immunological exposure including  
305 vaccination. We hypothesized that patients with pulmonary disease would have higher viral  
306 loads than patients without pulmonary disease, especially for early viral variants, which had a  
307 stronger tropism for lung as opposed to the upper respiratory tract. However, this hypothesis  
308 was not supported (Fig. 3c). Viral loads for the 975 patients with pulmonary disease tested  
309 during the early-variant era were statistically indistinguishable from those for the 27,308  
310 patients with no pulmonary disease who were tested during the same time period. This is an  
311 example of unexpected findings that this dataset and its web-portal interface can reveal.

312 **Comparisons among multiple groups: survivorship and causes of death.** The web portal also  
313 allows users to compare more than two groups of patients at a time. In quantifying mortality  
314 during the pandemic, one distinction of value has been between individuals who died with  
315 COVID-19 as the proximal cause of death and individuals who died with COVID-19 as an  
316 incidental finding. We compared these two groups with survivors (Fig. 3d). We found that the  
317 398 patients who died from COVID-19 in our dataset had higher viral loads than either of the  
318 other two groups, and that viral loads were statistically indistinguishable between the  
319 approximately 46,000 survivors and the 143 patients who died with COVID-19 as an incidental  
320 finding ( $p=0.07$ ). For ease of comparison, the web portal displays distributions in a ridgeline plot  
321 from lowest to highest mean, top to bottom. When there are three or more groups, p-values  
322 are displayed as a heatmap, accompanied by explanatory text. (Because KS p-values are  
323 symmetric, only the top half of the heatmap is shown.)

324 **Complex patient subgroups: race and presumed variant.** The ability to interrogate complex  
325 subgroups by checking multiple boxes in the web-portal interface allows more subtle  
326 investigations. For example, Black patients have experienced disproportionate morbidity and  
327 mortality during the pandemic.<sup>31</sup> However, the 13,806 patients who self-reported as White in  
328 our dataset on average had slightly higher viral loads than the 8,299 who self-reported as Black  
329 (KS  $p=3.3 \times 10^{-13}$ ). That the viral loads in the White group were on average *higher* suggests that  
330 differences in outcome between these groups are not explained by differences in viral load (Fig.  
331 3f-g), despite the clear relationship between viral load and survivorship described above (Fig.  
332 3d). Interestingly, the observed difference is more pronounced during the delta-variant wave  
333 (Fig. 3h). During the delta wave (July 7, 2021 to December 6, 2021), viral loads were on average  
334 three times as high for White patients ( $8.0 \times 10^5$  copies/mL,  $n=1,024$ ) as Black patients ( $2.7 \times 10^5$   
335 copies/mL,  $n=665$ ; KS test  $p=5.2 \times 10^{-8}$ ) with a distinctly sharper high-viral-load peak in White  
336 patients. This difference was greater in patients over 30 years old and was almost entirely  
337 absent in patients under 30 (30-60 y.o.: 398 Black patients and 719 White patients,  $p=2.5 \times 10^{-5}$ ;  
338 <30 y.o.: 266 Black vs. 299 White patients,  $p=0.14$ ). In contrast, viral load distributions for Black  
339 and White patients were more similar for both early in the pandemic and during the omicron  
340 variant time period ( $p=4.0 \times 10^{-5}$  for 4,393 Black and 7,042 White patients and  $p=0.02$  for 2,295

341 Black and 4,341 White patients, respectively). This example illustrates the utility and (statistical)  
342 power of the portal for investigating subgroups of interest.

343 **Antigen test performance.** In the head-to-head comparison of PCR and antigen test results, 281  
344 patients consented to participate. Of the PCR samples collected, 277 were tested; the  
345 remaining four were mishandled or leaked. Of the 277, 65 had a positive COVID-19 result by  
346 PCR (23%). The PCR-positive samples were all tested on either the Alinity m SARS-CoV-2 real  
347 time RT-PCR assay or the Alinity m Resp-4-Plex PCR assay. Viral loads in the PCR-positive  
348 patients ranged from approximately 10 to approximately  $10^9$  copies/mL, with a peak in the  
349 distribution between  $10^6$  and  $10^8$ . Of the 65 positive samples, 3 were sequenced and 20  
350 selected at random were used to assess contagiousness in viral culture.

351 **Antigen test performance.** Of the 65 patients with positive PCR tests, 43 tested positive on the  
352 Binax antigen test and 40 tested positive on the CareStart antigen test. No invalid antigen tests  
353 (lacking the control line) were observed. Only one of the patients who tested negative by PCR  
354 tested positive on the antigen tests (both Binax and CareStart), confirming the high specificity  
355 of these tests. The proportion of positive antigen tests varied with viral load. At viral loads less  
356 than  $10^3$  copies/mL, both antigen tests were always negative; at viral loads greater than  $10^7$   
357 copies/mL, both antigen tests were always positive. However, there was an overlap of antigen-  
358 test-positive and antigen-test-negative results at intermediate viral loads (Fig. 4a).  $k$  and  $v_0$   
359 values (see Methods) were comparable between the two antigen tests ( $k=1.184$ ,  $v_0=4.538$  for  
360 Binax and  $k=1.142$ ,  $v_0=4.995$  for CareStart). The resulting S-shaped curves were used to predict  
361 antigen test performance in the web portal.

362 The OTC antigen tests that have been widely available on the market since 2021 are  
363 considerably less sensitive than RT-qPCR for detecting SARS-CoV-2 infection. However, because  
364 their LODs are generally above the contagiousness threshold, they are quite sensitive for  
365 detecting contagiousness.<sup>10</sup> Based on our clinical experience, we hypothesized that antigen  
366 tests would perform similarly on different patient groups and subgroups; this hypothesis was  
367 largely supported (Fig. 4b-d). The web portal also allows users to estimate sensitivity and

368 specificity for the BinaxNow COVID-19 Ag Card and CareStart COVID-19 Antigen Home Test,  
369 based on the modelled performance curves, on any sufficiently large user-selected patient  
370 group (Fig 6). The two tests performed well: sensitivities for detecting contagiousness were  
371 roughly 0.85-0.90 across patient groups.

372 **Contagiousness for omicron-era virus.** For early-pandemic and delta-wave strains of SARS-CoV-  
373 2, the threshold viral load for contagiousness has previously been found to be approximately  
374  $10^5$  copies/mL.<sup>10</sup> For omicron variants, we found that the threshold is statistically  
375 indistinguishable from this, at  $4.5 \times 10^4$  copies/mL (confidence interval,  $1.1 \times 10^4$ - $1.9 \times 10^5$   
376 copies/mL;  $p=0.23$ , Fig. 5). The omicron threshold was based on 20 PCR-positive results from  
377 our head-to-head clinical trial of 277 total patients. We confirmed that the dominant strain  
378 circulating in the Massachusetts Bay area was omicron (BA5.2/Clade 22B) via next-generation  
379 sequencing, with only rare single-nucleotide differences relative to those already described (Fig.  
380 6a-d). Because the strain mix in Massachusetts (Fig. 6e) has been highly representative of the  
381 strain mix in the country as a whole throughout the pandemic (Fig. 6f), these results support  
382 the generalizability of these findings from a particular geographic area to the entire population.

## 383 **Discussion**

384 The COVID-19 pandemic has proven a catalyst for accelerating medical advances, including the  
385 development of more efficient methods for developing and testing critical diagnostic assays.<sup>32-</sup>  
386 <sup>37</sup> It has also drawn attention to the value of reliable public data, including large public  
387 datasets.<sup>38-42</sup> Here we describe such a dataset, to our knowledge the first large dataset of SARS-  
388 CoV-2 viral loads in patients across the history of the pandemic through to the present day. The  
389 rich clinical annotations of this dataset reveal similarities and differences in viral loads among  
390 patients by demographics, presentation, and comorbidity as well as by vaccination status,  
391 treatment, and socioeconomic status. Protected patient data was safeguarded by multiple  
392 mechanisms. The data revealed cases in which differences were expected as well as cases in  
393 which they were unexpected. The size of the dataset and extent of the annotation allow more  
394 comparisons than can reasonably be summarized in a single publication; availability of this

395 dataset via the web allows anyone—clinicians, investigators, developers, regulators, and  
396 patients, alone or with the assistance of artificial intelligence and/or machine learning (AI/ML)-  
397 based tools—to explore and conduct research, to test existing hypotheses, and to generate  
398 new research questions.

399 The clinical utility of measuring and investigating viral load in SARS-CoV-2 has been  
400 demonstrated.<sup>43–46</sup> This is consistent with both the advantage of viral loads relative to Ct  
401 values<sup>47,48</sup> and their utility in earlier viral infections such as HIV and hepatitis C (HCV). SARS-  
402 CoV-2 viral loads have already proven useful in the development and characterization of COVID-  
403 19 diagnostics in multiple contexts, including testing on nasopharyngeal secretions, nasal  
404 secretions, and saliva.<sup>5,6</sup> Here we demonstrate their potential to regulators as a tool to  
405 streamline evaluation of new OTC tests. Specifically, we demonstrate that a test’s analytical  
406 performance measure, namely the LOD, can be used to estimate that test’s clinical sensitivity  
407 for detecting contagiousness in any patient group, without having to conduct a dedicated  
408 clinical trial for that group. The alternative of innumerable dedicated trials is beyond reasonable  
409 expectations of the financial capability of developers or the bandwidth of their clinical testing  
410 partners. For this reason regulatory agencies such as the FDA have expressed interest in  
411 methods that use large-scale real-world data to streamline test evaluation.

412 Two assumptions implicit in this approach are worth mention. First, we model the success rate  
413 of antigen tests solely as a function of viral load. The assumption is that no other non-negligible  
414 factor varies systematically between patient groups. Second, we assume that the curve—  
415 success rate as a function of viral load—can be adequately predicted from the data from a fairly  
416 small study. There are two possible sources of error in this curve: sampling error, which can be  
417 reduced by increasing the number of subjects sampled; and lack-of-fit error, the error inherent  
418 in trying to fit a function of the wrong form. The function used in the web portal’s calculations,  
419 was chosen on the basis of its history of use in dose-response-type situations, but carries the  
420 (reasonable) assumptions that the probability increases smoothly and continuously with  
421 increasing viral load, and approaches 0 with sufficiently low viral load and 1 with sufficiently  
422 high viral load. These constraints leave only two free parameters, which is desirable for



423 statistical power and robustness to the noise inherent in any such study (e.g., sampling error  
424 and measurement error).

425 A user of the web portal who is accustomed to thinking of test quality solely in terms of LOD  
426 (limit of detection) might initially be surprised that the portal prefers two parameters, not one,  
427 to define an antigen test's performance. In effect, the LOD parameter sets the location of the S-  
428 shaped curve that relates viral load and performance, and the second parameter—here, the  
429 50% detection threshold—sets the S-shaped curve's steepness. Without the second parameter,  
430 one could fit a curve in which the sensitivity is 0% at all viral loads below the LoD, and 95% at  
431 any higher viral load, which is clearly quite different from the relationship observed in the head-  
432 to-head study. How different the true shape of the antigen test performance curve is from the  
433 logit function used here, and thus whether fitting a different function would work better, can  
434 be elucidated by larger head-to-head studies; however, in the midst of a public health  
435 emergency, the cost in time of sampling more subjects must be weighed against the value of  
436 complete certainty. At any rate, the fitting error was low.

437 Three important limitations to the dataset in its current form also deserve mention. The first is  
438 due to incompleteness of some of the data fields, for example presentation and vaccination  
439 status (Table 1). Presentation information was only sometimes available in structured form in  
440 our data repository; we did not attempt to extract data from notes to complement incomplete  
441 records. Vaccination status was likewise only sometimes available in a structured manner;  
442 integration with state-level records could potentially fill in missing records. Second, patient-  
443 level annotations are not yet available for download as part of the dataset. Making viral loads  
444 freely and easily available for patient groups required significant attention to avoid potential  
445 loopholes that might risk patient privacy via identifiability. Methods included suppressing data  
446 transmission for groups that were small enough to present potential “journalist risk,”<sup>49</sup> jittering  
447 counts to prevent deduction of the sizes of suppressed groups, and rounding viral loads to two  
448 log-scale decimal places.<sup>50</sup> Further work is necessary to make patient-level annotations  
449 available. Third and finally, the size of the dataset, while large, is still insufficient for the  
450 smallest groups—for example, cystic fibrosis patients, patients who recently delivered, or

451 Native Americans—to be sufficiently sizeable to draw robust conclusions. One solution is to add  
452 data from other care-giving institutions that performed substantial COVID-19 testing; another is  
453 to supplement existing large datasets, for example the 50-million-person CVD-COVID-UK  
454 initiative, with viral loads. The free availability of methods to convert from Ct values to viral  
455 loads facilitates such advances.<sup>4</sup>

456 As part of the drive toward precision medicine, clinical care benefits from personalization of  
457 diagnostic testing: the right test for the right patient, where the importance of patient  
458 heterogeneity is increasingly accepted. We have demonstrated that large-scale real-world data  
459 can assist this effort by enabling the estimation of personalized clinical sensitivities and  
460 specificities without the need for dedicated clinical trials on every patient group. This work is  
461 generalizable beyond COVID-19. Laboratory testing is an exceptionally rich source of real-world  
462 medical information. It is the highest volume medical activity, with some 20 billion tests  
463 performed each year in the United States alone. It is also the most cost-effective, costing just  
464 pennies on the healthcare dollar. It is integral to decision-making across medicine, for patients  
465 at every level of acuity, from screening to emergencies. Its results are almost always numerical  
466 or categorical, making it especially amenable to modern approaches like machine learning. And  
467 computational re-analysis is substantially less expensive than de novo trials. The present work  
468 supports the view that meaningful value can come economically from additional repurposing of  
469 the vast stores of real-world laboratory results that exist in healthcare institutions.

## 470 **Acknowledgements**

471 The authors acknowledge Elliot Hill for assistance on plugging ct2vl into the data processing  
472 workflow; to Timothy Graham, Gail Piatkowski, Baevin Feeser, and Griffin Weber for their  
473 advice and guidance on extracting data from EMR databases; to funding from the Reagan-Udall  
474 Foundation for the FDA (R.A.); and to funding via a Novel Therapeutics Delivery Grant from the  
475 Massachusetts Life Sciences Center (J.E.K.). We would also like to acknowledge Abbott  
476 (Scarborough, ME) for provision of the Binax Now antigen tests and Ginkgo Bioworks (Boston,  
477 MA) for provision of CareStart antigen tests used in this study.

## 478 **Author Contributions**

479 Conceptualization, J.K. and R.A.; Software, A.M.; Investigation, D.H., E.C., S.D., and M.Y.;  
480 Resources: S.R.; Data Curation, A.M.; Writing — Original Draft, R.A. and A.M.; Writing — Review  
481 & Editing, P.K. and J.K.; Visualization, T.S., D.B., A.M., and R.A.; Supervision, R.A; Funding  
482 Acquisition, P.K., S.R., and R.A.

## 483 **Declaration of interests**

484 The authors declare no competing interests.

## 485 **Table and Figure Legends**

486 **Table 1: Summary of patient characteristics.** Shown are counts for select high-level categories  
487 as of April 14, 2023. Counts may differ somewhat from the counts presented through the web  
488 portal as a result of jittering and as more data continues to be added through the portal. Note  
489 that counts broken down by characteristics do not add up to the total, because of the nulling  
490 out of some data to reduce re-identification risk (see Methods).

491 **Table S1: Detailed criteria for defining patient characteristics, Related to Methods.** Criteria for  
492 each of the characteristics available on the portal are described.

493 **Fig. 1: Patient characteristics available for defining patient groups and subgroups.** This  
494 screenshot from the web portal demonstrates the ability for users to define patient groups by  
495 demographic, comorbidity, treatment, vaccination status, pandemic era, and so on, using  
496 checkboxes.

497 **Fig. 2: Overall bimodal distribution of viral loads.** When no checkboxes are selected to  
498 constrain or partition the dataset, users see the distribution of all viral loads. The marked  
499 bimodal distribution is clearly apparent.

500 **Fig. 3: Viral load comparisons by remdesivir treatment, patient presentation, and outcome.**

501 Screenshot from the web portal. Viral loads are higher in (a) patients who received remdesivir  
502 vs. patients who did not receive remdesivir and in (b) patients who presented as ill-appearing  
503 vs. patients who presented as well-appearing (see Table S1 for precise definitions). Note the  
504 bimodal distributions, with a low-viral-load peak and a high-viral-load peak. (c) Viral load  
505 distributions and pairwise p-values by presence or absence of pulmonary disease for patients  
506 during the early variant era. (d) Patients who died from COVID-19 had higher viral loads than  
507 either patients who died with COVID-19 as an incidental finding or survivors. Viral loads for  
508 were statistically indistinguishable between the latter two groups. The web portal displays  
509 distributions in a ridgeline plot from lowest to highest mean, top to bottom. (e) Because there  
510 are three or more groups, p-values are displayed as a heatmap, accompanied by explanatory  
511 text. Because KS p-values are symmetric, only the top half of the heatmap is shown. Fig. 3f-h:  
512 viral load distributions by self-reported race and by race-plus-presumed variant. (f) Viral load  
513 distributions by race. (g) Statistical comparisons among these distributions. (h) Viral load  
514 distributions between Black vs. White patients in the delta-variant era.

515 **Fig. 4: Antigen test results from head-to-head trial and performance on patient subgroups.** (a)

516 Each antigen test result for each PCR-positive patient, vs.  $\log_{10}$  of viral load according to the  
517 simultaneous PCR test. (b) Antigen test performance on patients in neighborhoods stratified by  
518 median household income. (c) Performance of BinaxNow COVID-19 Ag Card on patients with  
519 median household income  $> \$130,000$ . (d) Performance of BinaxNow COVID-19 Ag Card on  
520 patients with median household income  $< \$52,000$ . The user can use the radio buttons to select  
521 any of the patient groups in the viral load distributions in the top section of the web portal. The  
522 imputed positive antigen-test results will appear as a lighter shade, and the negative results a  
523 darker shade.

524 **Fig. 5: Determination of the contagiousness threshold for omicron-era SARS-CoV-2 strains.**

525 Below a certain viral load in the patient sample (x-axis), no virus was recoverable from culture  
526 (y-axis, maximum of day 3 and day 6 supernatants). For viewing convenience, culture-negative  
527 samples are plotted at 0.1 copies/mL (dotted line, 1 copy/mL). The gray region shows the

528 confidence interval for the threshold. The red line shows the midpoint of this region (on the  
529 log<sub>10</sub> scale), 45,000 copies/mL, as the most likely threshold.

530 **Fig. 6. Sequencing late-2022 strain and generalizability of Massachusetts-level results to the**  
531 **United States as a whole.** Results of sequencing of a BA5.2/Clade 22B patient sample from  
532 Aug.-Sep. 2022 (97.6% coverage). (a) Sample relative to COVID-19 phylogeny (with clade labels).  
533 (b) First 64 of the 72 nucleotide substitutions relative to the original Wuhan strain. (c) 52 amino  
534 acid substitutions relative to the Wuhan strain. (d) The five unique (“private”) mutations  
535 relative to the phylogenetic tree. (e) Distribution of strains in Massachusetts near the time of  
536 the sample according to covariants.org. (f) Comparison by frequency of the strains circulating in  
537 Massachusetts to those circulating in the United States at the same times demonstrating  
538 generalizability of Massachusetts-state variant patterns to the country as a whole. Red line, 1:1.  
539 Gray, early strains; purple, delta strains; green, omicron strains.  $R^2$  is for least-squares linear  
540 regression of USA vs. Massachusetts data (regression slope=0.97, intercept=0.00).

541 **Tables and Figures**

542 Table 1: Summary of patient characteristics

543	<b>Sex</b>	
	Female	25,884
544	Male	20,608
	<b>Age</b>	
	<30 years old	12,446
545	30-60 years old	21,889
	> 60 years old	12,100
	<b>Self-reported Race or Ethnicity</b>	
546	Unknown/Other	14,530
	White	13,806
547	Black	8,299
	Hispanic	7,540
548	Asian/Pacific Islander	2,472
	<b>Setting</b>	
	Inpatient	2,157
549	Outpatient	11,758
	Emergency room	1,779
550	Other institutions	31,031
	<b>Variant</b>	
	Early	28,289
551	Delta	2,911
	Omicron	11,264
	<b>Vaccination status</b>	
552	Vaccinated	6,806
	Unvaccinated	6,960
553	Unknown	32,732
	<b>Outcome</b>	
	Died from COVID-19	398
554	Died with COVID-19 as an incidental finding	143
	Survived	45,938
	<b>Testing platform</b>	
555	Abbott m2000	24,243
	Abbott Alinity	20,593
556	Abbott Alinity 4-plex	1,889
	<b>Total</b>	
557		46,726



560

Table S1, continued

Checkbox name	Notes
	ICD10 codes: [G041,'G800','G801','G802','G808','G809','G8100','G8101','G8102','G8103','G8104','G8110','G8111','G8112','G8113','G8114','G8190','G8191','G8192','G8193','G8194','G8220','G8221','G8222','G8250','G8251','G8252','G8253','G8254','G830','G8310','G8311','G8312','G8313','G8314','G8320','G8321','G8322','G8323','G8324','G8330','G8331','G8332','G8333','G8334','G834','G835','G8381','G8382','G8383','G8384','G8389','G839','G89031','G89032','G89033','G89034','G89039','G89041','G89042','G89043','G89044','G89049','G89051','G89052','G89053','G89054','G89059','G89061','G89062','G89063','G89064','G89065','G89069','G89131','G89132','G89133','G89134','G89139','G89141','G89142','G89143','G89144','G89149','G89151','G89152','G89153','G89154','G89159','G89161','G89162','G89163','G89164','G89165','G89169','G89231','G89232','G89233','G89234','G89239','G89241','G89242','G89243','G89244','G89249','G89251','G89252','G89253','G89254','G89259','G89261','G89262','G89263','G89264','G89265','G89269','G89311','G89312','G89313','G89314','G89319','G89341','G89342','G89343','G89344','G89349','G89351','G89352','G89353','G89354','G89359','G89361','G89362','G89363','G89364','G89365','G89369','G89811','G89812','G89813','G89814','G89819','G89841','G89842','G89843','G89844','G89849','G89851','G89852','G89853','G89854','G89859','G89861','G89862','G89863','G89864','G89865','G89869','G89931','G89932','G89933','G89934','G89939','G89941','G89942','G89943','G89944','G89949','G89951','G89952','G89953','G89954','G89959','G89961','G89962','G89963','G89964','G89965','G89969','Q','R','Q532']
Disabilities	
	ICD10 codes:[A1884,'A3282','A3681','A381','A395','A5203','B2682','B332','B376','B5881','C452','D8685','G130','G712','G713','G720','G721','G722','G7249','G7281','G7289','G729','G737','I01','I02','I05','I06','I07','I08','I09','I111','I110','I13','I130','I20','I21','I22','I23','I24','I25','I252','I3','I4','I5','I501','I5020','I5021','I5022','I5023','I5030','I5031','I5032','I5033','I5040','I5041','I5042','I5043','I50810','I50811','I50812','I50813','I50814','I5082','I5083','I5084','I5089','I509','I5181','I70','I9713','I97130','I97131','I1082','I1182','O101','O29121','O29122','O29123','O29129','O903','O92912','Q2','R570','S26','T28','Z95','Z95811','Z95812']
Heart conditions	
	Immunosuppressed vs. immunocompetent. Immunosuppressed if most recent T-cell subset analysis report, their absolute CD4 cell count was <100 cells/μl; they had a diagnosis of either lymphoma or leukemia associated with an encounter either before the PCR test or within 60 days of the PCR test; they were on any of the following drugs on an ongoing basis, prescribed prior to the PCR test and with enough refills to include the time up to 30 days prior to the PCR test: abatacept, adalimumab, anakinra, azathioprine, basiliximab, budesonide, certolizumab, cyclosporine, daclizumab, dexamethasone, everolimus, etanercept, golimumab, infliximab, kekizumab, leflunomide, lenalidomide, methotrexate, mycophenolate, natalizumab, pomalidomide, prednisone, rituximab, secukinumab, serrolimus, tacrolimus, tocilizumab, tofacitinib, ustekinumab, or vedolizumab. Otherwise, immunocompetent
Immune status	
	ICD10 codes:[A5445,'A5274','B180','B181','B182','B188','B189','B190','B1910','B1911','B1920','B1921','B199','B251','B581','B8500','B8501','B8510','B8511','B864','K700','K702','K7030','K7031','K7040','K7041','K709','K713','K714','K7150','K7151','K716','K717','K718','K7210','K7211','K7290','K7291','K730','K731','K732','K738','K739','K740','K7400','K7401','K7402','K741','K742','K743','K744','K745','K7460','K7469','K751','K752','K753','K754','K7581','K7589','K759','K760','K761','K762','K763','K764','K765','K766','K767','K7681','K7689','K769','K77','K9182','Z944']
Liver disease	
	ICD10 codes:[F060,'F061','F062','F0630','F0631','F0632','F0633','F0634','F11150','F11151','F11159','F11250','F11251','F11259','F11950','F11951','F11959','F12150','F12151','F12159','F12250','F12251','F12259','F12950','F12951','F12959','F13150','F13151','F13159','F13250','F13251','F13259','F13951','F13959','F14150','F14151','F14159','F14250','F14251','F14500','F14501','F14599','F15150','F15151','F15159','F15250','F15251','F15259','F15950','F15951','F15959','F16150','F16151','F16159','F16250','F16251','F16259','F16950','F16951','F16959','F18150','F18151','F18159','F18250','F18251','F18259','F18950','F18951','F18959','F19150','F19151','F19159','F19250','F19251','F19259','F19950','F19951','F19959','F200','F201','F202','F203','F205','F2081','F2089','F209','F21','F22','F23','F24','F250','F251','F258','F259','F28','F29','F3010','F3011','F3012','F3013','F302','F303','F304','F308','F309','F310','F3110','F3111','F3112','F3113','F312','F3130','F3131','F3132','F314','F315','F3160','F3161','F3162','F3163','F3164','F3170','F3171','F3172','F3173','F3174','F3175','F3176','F3177','F3178','F3181','F3189','F319','F320','F321','F322','F323','F324','F325','F328','F3281','F3289','F329','F32A','F330','F331','F332','F333','F3340','F3341','F3342','F338','F339','F340','F341','F348','F3481','F3489','F349','F39','F4489','F843']
Mental health conditions	
	median household income by zip code from the 2020 Census (2020 American Community Survey). URL structure (replace <ZIPCODE> with an actual zip code): <a href="https://api.census.gov/data/2020/acs/acs5/subject?get=NAME,S1901_C01_O12E&amp;for=zip%20code%20at%20county%20area:&lt;ZIPCODE&gt;">https://api.census.gov/data/2020/acs/acs5/subject?get=NAME,S1901_C01_O12E&amp;for=zip%20code%20at%20county%20area:&lt;ZIPCODE&gt;</a> . In the text that this URL will return, S1901_C01_O12E is the column name for median household income
Neighborhood income	
	ICD10 codes:[E7500,'E7501','E7502','E7509','E7510','E7511','E7519','E7523','E7525','E7526','E7529','E754','F0150','F0151','F0280','F0281','F0390','F0391','F05','F842','G08','G10','G110','G111','G1110','G1111','G1119','G112','G113','G114','G118','G119','G120','G121','G1220','G1221','G1222','G1223','G1224','G1225','G1229','G128','G129','G131','G132','G138','G20','G210','G2111','G2119','G212','G213','G214','G218','G219','G230','G231','G232','G238','G239','G2409','G241','G242','G248','G254','G255','G2570','G2571','G2579','G2581','G2582','G2583','G2589','G259','G26','G300','G301','G308','G309','G3101','G3109','G311','G312','G3181','G3182','G3183','G3185','G3189','G319','G320','G3281','G3289','G35','G360','G368','G369','G370','G371','G372','G373','G374','G375','G378','G379','G40001','G40009','G40011','G40019','G400191','G400199','G40111','G40119','G40201','G40209','G40211','G40219','G40301','G40309','G40319','G40401','G40409','G40411','G40419','G40441','G40449','G4042','G40501','G40509','G40801','G40802','G40803','G40804','G40811','G40812','G40813','G40814','G40821','G40822','G40823','G40824','G40833','G40834','G40839','G40901','G40909','G40911','G40919','G40A01','G40A09','G40A11','G40A19','G40B01','G40B09','G40811','G40819','G47411','G47419','G47421','G47429','G803','G890','G910','G911','G912','G913','G914','G918','G919','G930','G9340','G9341','G9349','G935','G936','G937','G9381','G9382','G9389','G939','G94','O99350','O99351','O99352','O99353','O99354','O99355','P9160','P9161','P9162','P9163','R561','R569']
Neurological disorders	
	ICD10 codes:[K250,'K251','K252','K253','K254','K255','K256','K257','K259','K260','K261','K262','K263','K264','K265','K266','K267','K269','K270','K271','K272','K273','K274','K275','K276','K277','K279','K280','K281','K282','K283','K284','K285','K286','K287','K289']
Peptic ulcer disease	
	ICD10 codes:[A5200,'A5201','A5202','A5209','I700','I701','I70201','I70202','I70203','I70208','I70209','I70211','I70212','I70213','I70218','I70219','I70221','I70222','I70223','I70228','I70229','I70231','I70232','I70233','I70234','I70235','I70238','I70239','I70241','I70242','I70243','I70244','I70245','I70248','I70249','I7025','I70261','I70262','I70263','I70268','I70269','I70291','I70292','I70293','I70298','I70299','I70301','I70302','I70303','I70308','I70309','I70311','I70312','I70313','I70318','I70319','I70321','I70322','I70323','I70328','I70329','I70331','I70332','I70333','I70334','I70335','I70338','I70339','I70341','I70342','I70343','I70344','I70345','I70348','I70349','I7035','I70362','I70363','I70368','I70369','I70391','I70392','I70393','I70398','I70399','I70401','I70402','I70403','I70408','I70409','I70411','I70412','I70413','I70418','I70419','I70421','I70422','I70423','I70428','I70429','I70431','I70432','I70433','I70434','I70435','I70438','I70439','I70441','I70442','I70443','I70444','I70445','I70448','I70449','I7045','I70461','I70462','I70463','I70468','I70469','I70491','I70492','I70493','I70498','I70499','I70501','I70502','I70503','I70508','I70509','I70511','I70512','I70513','I70518','I70519','I70521','I70522','I70523','I70528','I70529','I70531','I70532','I70533','I70534','I70535','I70538','I70539','I70541','I70542','I70543','I70544','I70545','I70548','I70549','I70551','I70552','I70553','I70556','I70559','I70561','I70562','I70563','I70568','I70569','I70601','I70602','I70603','I70608','I70609','I70611','I70612','I70613','I70618','I70619','I70621','I70622','I70623','I70628','I70629','I70631','I70632','I70633','I70634','I70635','I70638','I70639','I70641','I70642','I70643','I70644','I70645','I70648','I70649','I7065','I70661','I70662','I70663','I70668','I70669','I70691','I70692','I70693','I70698','I70699','I70701','I70702','I70703','I70708','I70709','I70711','I70712','I70713','I70718','I70719','I70721','I70722','I70723','I70728','I70729','I70731','I70732','I70733','I70734','I70735','I70738','I70739','I70741','I70742','I70743','I70744','I70745','I70748','I70749','I70751','I70752','I70753','I70759','I70761','I70762','I70763','I70768','I70769','I70791','I70792','I70793','I70798','I70799','I708','I7090','I7091','I7092','I7100','I7101','I7102','I7103','I711','I712','I713','I714','I715','I716','I718','I719','I720','I721','I722','I723','I724','I725','I726','I728','I729','I7301','I731','I7381','I7389','I739','I7401','I7409','I7410','I7411','I7419','I742','I743','I744','I745','I748','I749','I75011','I75012','I75013','I75019','I75051','I75052','I75053','I75059','I75081','I7589','I770','I771','I772','I773','I774','I775','I776','I7770','I7771','I7772','I7773','I7774','I7775','I7776','I7777','I7779','I77810','I77811','I77812','I77819','I7789','I779','I780','I781','I788','I789','I790','I791','I798','K551','Z95820','Z95828']
Peripheral vascular disease	
	ICD10 codes:[Z94000,'Z94003','Z94640','Z94645']
Substance abuse	
	ICD10 codes:[F1010,'F1011','F10120','F10121','F10129','F10130','F10131','F10132','F10139','F1014','F10150','F10151','F10159','F10180','F10181','F10182','F10188','F10189','F1019','F1020','F1021','F10220','F10221','F10229','F10230','F10231','F10232','F10233','F10239','F1024','F10250','F10251','F10259','F1026','F1027','F10280','F10281','F10282','F10288','F1029','F1094','F10950','F10951','F10959','F1096','F1097','F10980','F11','F1110','F1111','F11120','F11121','F11122','F11129','F1113','F1114','F11181','F11182','F11188','F1119','F1120','F1121','F11220','F11221','F11222','F11229','F1123','F11280','F11288','F1129','F11299','F113120','F113121','F113122','F113123','F113129','F11313','F11314','F11318','F11319','F11320','F11321','F113220','F113221','F113229','F11323','F11324','F11329','F1133121','F1133122','F1133123','F1133129','F113313','F113314','F113318','F113319','F113320','F113321','F1133220','F1133221','F1133229','F113323','F113324','F11326','F11327','F113280','F113281','F113282','F113288','F11329','F14','F1410','F1411','F14120','F14121','F14122','F14129','F1413','F1414','F14180','F14181','F14182','F14188','F1419','F1420','F1421','F14220','F14221','F14229','F1423','F14239','F1424','F14280','F14281','F14282','F14288','F1429','F1510','F1511','F15120','F15121','F15122','F15129','F1513','F1514','F15180','F15181','F15182','F15188','F1519','F1520','F1521','F15220','F15221','F15222','F15229','F1523','F1524','F15280','F15281','F15282','F15288','F1529','F16','F1610','F1611','F16120','F16121','F16122','F16129','F1614','F16180','F16181','F16188','F1619','F1620','F1621','F16220','F16221','F16229','F1624','F16280','F16283','F16288','F1629','F1810','F1811','F18120','F18121','F18122','F18129','F1814','F1817','F18180','F18188','F1819','F1820','F1821','F18220','F18221','F18229','F1824','F1827','F18280','F18288','F1829','F19','F1910','F1911','F19120','F19121','F19122','F19129','F19130','F19131','F19132','F19139','F1914','F1916','F1917','F19180','F19181','F19188','F19189','F1919','F1920','F1921','F19220','F19221','F19222','F19229','F19230','F19231','F19232','F19239','F1924','F1926','F1927','F19280','F19281','F19282','F19288','F1929','G621','I426','K2920','K2921','K7010','K7011','O99310','O99311','O99312','O99314','O99315','O99320','O99321','O99322','O99323','O99324']
Tocilizumab	matches to case-insensitive searches for the string "tocilizumab"
Transplanted organ and tissue status	ICD10 codes: ["Z94"]
Ventilation assistance	CPT4 codes: ["94002","94003","94640","94645"]

561



562 Figure 1: Patient characteristics available for defining patient groups and subgroups

Explore groups ▾

**Group** Reset

**Sex**

- Male
- Female

**Age**

- <30 years old
- 30-60 years old
- >60 years old

**Patient Location**

- Inpatient
- Outpatient
- Emergency room
- Institutional

**BMI**

- Underweight
- Healthy weight
- Overweight
- Obese

**Immune Status**

- Immunosuppressed
- Immunocompetent

**Smoking Status**

- Current smokers
- Former smokers
- Never smoked

**Presentation**

- Sick-appearing
- Well-appearing

**Neighborhood Income**

- < \$52,000
- \$52,000-\$78,000
- \$78,000-\$104,000
- \$104,000-\$130,000
- >\$130,000

**Vaccination Status**

- Vaccinated
- Unvaccinated
- Unknown

**Presumed Variant**

- Early variants
- Delta
- Omicron

**Race/Ethnicity**

- White
- Black
- Asian/Pacific islander
- Hispanic
- Unknown/Other

**Pregnancy Status**

- Pregnant
- Not pregnant

**Outcome**

- Survived
- Died from COVID-19 (causal)
- Died with COVID-19 (incidental)

**Blood Products**

- Received blood products
- Did not receive blood products

**Dexamethasone**

- Received dexamethasone
- Did not receive dexamethasone

**Remdesivir**

- Received remdesivir
- Did not receive remdesivir

**Tocilizumab**

- Received tocilizumab
- Did not receive tocilizumab

**Ventilation Assist**

- Received ventilation assist
- Did not receive ventilation assist

**Heart Conditions**

- Known heart conditions
- No reported heart conditions

**Peripheral Vascular Disease**

- Known peripheral vascular disease
- No reported peripheral vascular disease

**Cerebrovascular Disease**

- Known cerebrovascular disease
- No reported cerebrovascular disease

**Neurological Disorders**

- Known neurological disorders
- No reported neurological disorders

**Pulmonary Disease**

- Known pulmonary disease
- No reported pulmonary disease

**Connective Tissue Disease**

- Known connective tissue disease
- No reported connective tissue disease

**Peptic Ulcer**

- Known peptic ulcer
- No reported peptic ulcer

**Liver Disease**

- Known liver disease
- No reported liver disease

**Diabetes**

- Known diabetes
- No reported diabetes

**Disabilities**

- Known disabilities
- No reported disabilities

**Renal Disease**

- Known renal disease
- No reported renal disease

**Cancer**

- Known cancer
- No reported cancer

**Acquired Immunodeficiency Syndrome**

- Known acquired immunodeficiency syndrome
- No reported acquired immunodeficiency syndrome

**Substance Abuse**

- Known substance abuse
- No reported substance abuse

**Mental Health Conditions**

- Known mental health conditions
- No reported mental health conditions

**Sickle Cell & Thalassemia**

- Known sickle cell & thalassemia
- No reported sickle cell & thalassemia

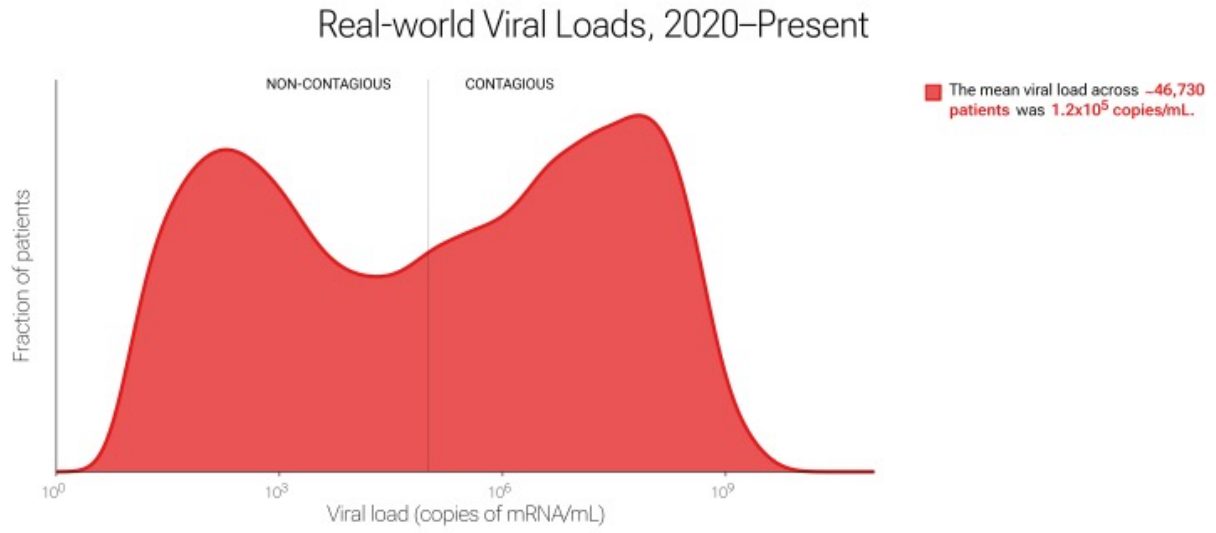
**Transplanted Organ And Tissue Status**

- Known transplanted organ and tissue status
- No reported transplanted organ and tissue status

563

564

Figure 2: Overall bimodal distribution of viral loads



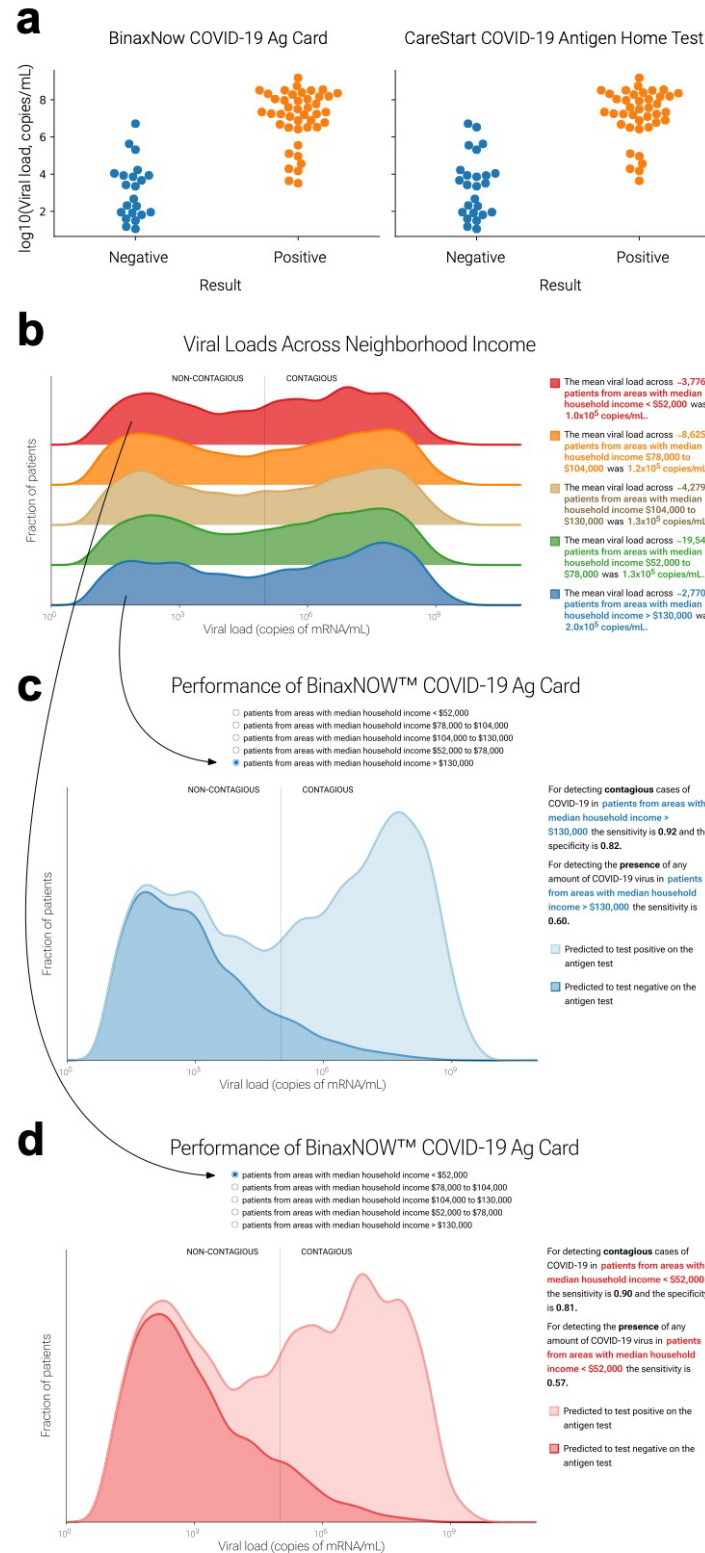
565

566 Figure 3: Viral load comparisons by remdesivir treatment, patient presentation, and outcome



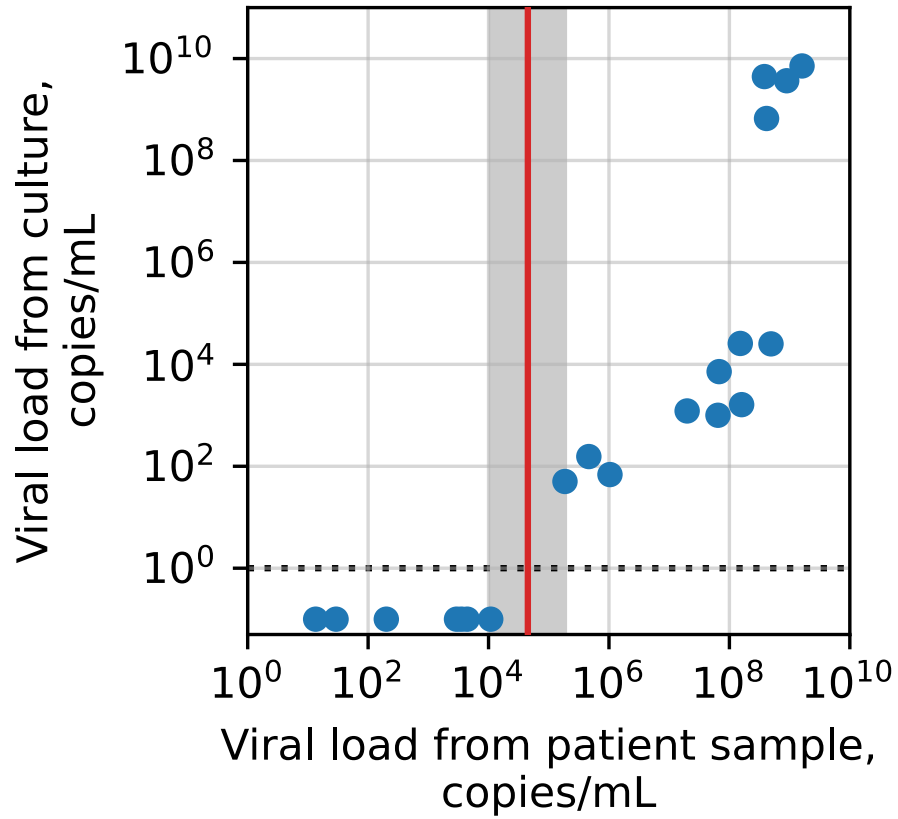
567

568 Figure 4: Antigen test results from head-to-head trial and performance on patient subgroups



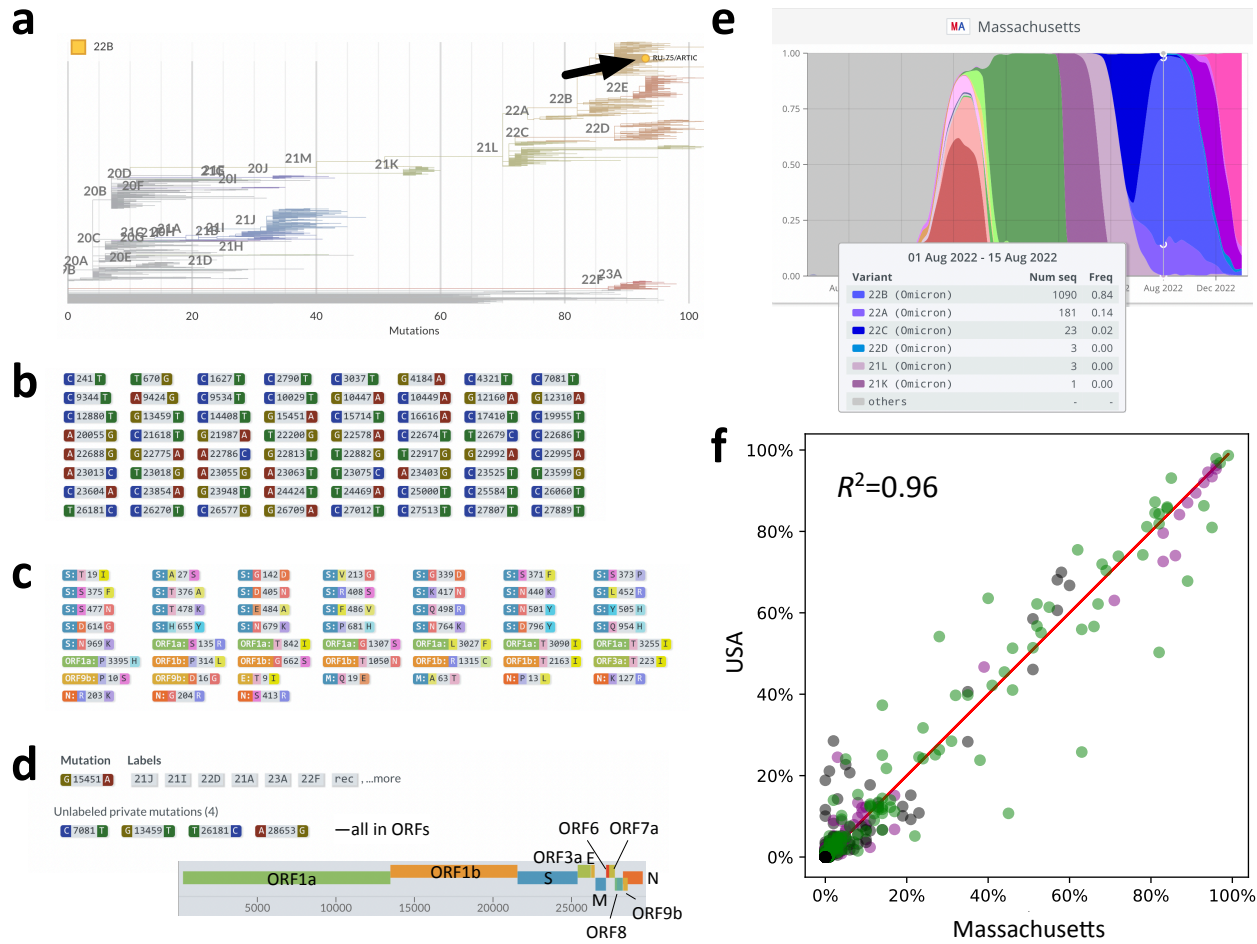
569

570 Figure 5: Determination of the contagiousness threshold for omicron-era SARS-CoV-2 strains



571

572 Figure 6: Sequencing late-2022 strain and generalizability of Massachusetts-level results to the  
 573 United States as a whole



574

## 575 **References**

- 576 1. National Research Council, Division on Earth and Life Studies, Board on Life Sciences,  
577 Committee on a Framework for Developing a New Taxonomy of Disease (2011). Toward  
578 Precision Medicine: Building a Knowledge Network for Biomedical Research and a New  
579 Taxonomy of Disease (National Academies Press).
- 580 2. Arnaout, R., Lee, R.A., Lee, G.R., Callahan, C., Cheng, A., Yen, C.F., Smith, K.P., Arora, R., and  
581 Kirby, J.E. (2021). The Limit of Detection Matters: The Case for Benchmarking Severe Acute  
582 Respiratory Syndrome Coronavirus 2 Testing. *Clin Infect Dis*. 10.1093/cid/ciaa1382.
- 583 3. Wild, D. ed. (2013). *The immunoassay handbook: theory and applications of ligand binding,*  
584 *ELISA, and related techniques* 4th ed. (Elsevier).
- 585 4. Hill, E.D., Yilmaz, F., Callahan, C., Cheng, A., Braun, J., and Arnaout, R. (2022). *ct2vl :*  
586 *Converting Ct Values to Viral Loads for SARS-CoV-2 RT-qPCR Test Results (Microbiology)*  
587 10.1101/2022.06.20.496929.
- 588 5. Callahan Cody, Lee Rose A., Lee Ghee Rye, Zulauf Kate, Kirby James E., Arnaout Ramy, and  
589 Miller Melissa B. Nasal Swab Performance by Collection Timing, Procedure, and Method of  
590 Transport for Patients with SARS-CoV-2. *Journal of Clinical Microbiology* 59, e00569-21.  
591 10.1128/JCM.00569-21.
- 592 6. Callahan, C., Ditelberg, S., Dutta, S., Littlehale, N., Cheng, A., Kupczewski, K., McVay, D.,  
593 Riedel, S., Kirby, J.E., and Arnaout, R. (2021). Saliva is Comparable to Nasopharyngeal Swabs  
594 for Molecular Detection of SARS-CoV-2. *Microbiol Spectr* 9, e0016221.  
595 10.1128/Spectrum.00162-21.
- 596 7. Pilarowski, G., Lebel, P., Sunshine, S., Liu, J., Crawford, E., Marquez, C., Rubio, L., Chamie, G.,  
597 Martinez, J., Peng, J., et al. (2021). Performance Characteristics of a Rapid Severe Acute  
598 Respiratory Syndrome Coronavirus 2 Antigen Detection Assay at a Public Plaza Testing Site  
599 in San Francisco. *J Infect Dis* 223, 1139–1144. 10.1093/infdis/jiaa802.

- 600 8. Mina, M.J., Parker, R., and Larremore, D.B. (2020). Rethinking Covid-19 Test Sensitivity - A  
601 Strategy for Containment. *N Engl J Med* 383, e120. 10.1056/NEJMp2025631.
- 602 9. Stanley, S., Hamel, D.J., Wolf, I.D., Riedel, S., Dutta, S., Contreras, E., Callahan, C.J., Cheng,  
603 A., Arnaout, R., Kirby, J.E., et al. (2022). Limit of Detection for Rapid Antigen Testing of the  
604 SARS-CoV-2 Omicron and Delta Variants of Concern Using Live-Virus Culture. *J Clin*  
605 *Microbiol* 60, e00140-22. 10.1128/jcm.00140-22.
- 606 10. Kirby, J.E., Riedel, S., Dutta, S., Arnaout, R., Cheng, A., Ditelberg, S., Hamel, D.J., Chang, C.A.,  
607 and Kanki, P.J. (2022). SARS-CoV-2 Antigen Tests Predict Infectivity based on viral culture:  
608 comparison of antigen, PCR viral load, and viral culture testing on a large Sample Cohort.  
609 *Clinical Microbiology and Infection*, S1198743X22003743. 10.1016/j.cmi.2022.07.010.
- 610 11. Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of  
611 classifying prognostic comorbidity in longitudinal studies: development and validation. *J*  
612 *Chronic Dis* 40, 373–383. 10.1016/0021-9681(87)90171-8.
- 613 12. Greenberg, J.A., Hohmann, S.F., Hall, J.B., Kress, J.P., and David, M.Z. (2016). Validation of a  
614 Method to Identify Immunocompromised Patients with Severe Sepsis in Administrative  
615 Databases. *Annals ATS* 13, 253–258. 10.1513/AnnalsATS.201507-415BC.
- 616 13. Knight, S.R., Ho, A., Pius, R., Buchan, I., Carson, G., Drake, T.M., Dunning, J., Fairfield, C.J.,  
617 Gamble, C., Green, C.A., et al. (2020). Risk stratification of patients admitted to hospital  
618 with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and  
619 validation of the 4C Mortality Score. *BMJ* 370, m3339. 10.1136/bmj.m3339.
- 620 14. Elixhauser, A., Steiner, C., Harris, D.R., and Coffey, R.M. (1998). Comorbidity measures for  
621 use with administrative data. *Med Care* 36, 8–27. 10.1097/00005650-199801000-00004.
- 622 15. Shain, E.B., and Clemens, J.M. (2008). A new method for robust quantitative and qualitative  
623 analysis of real-time PCR. *Nucleic Acids Res* 36, e91. 10.1093/nar/gkn408.



- 624 16. Kirby, J.E., Cheng, A., Cleveland, M.H., Degli-Angeli, E., DeMarco, C.T., Faron, M., Gallagher,  
625 T., Garlick, R.K., Goecker, E., Coombs, R.W., et al. (2022). A Multi-Institutional Study  
626 Benchmarking Cycle Threshold Values for Major Clinical SARS-CoV-2 RT-PCR Assays.  
627 2022.06.22.22276072. 10.1101/2022.06.22.22276072.
- 628 17. CoVariants <https://covariants.org/>.
- 629 18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,  
630 Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in  
631 Python. *Journal of Machine Learning Research* 12, 2825–2830.
- 632 19. Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer,  
633 D., Jones, T.C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized  
634 patients with COVID-2019. *Nature* 581, 465–469. 10.1038/s41586-020-2196-x.
- 635 20. La Scola, B., Le Bideau, M., Andreani, J., Hoang, V.T., Grimaldier, C., Colson, P., Gautret, P.,  
636 and Raoult, D. (2020). Viral RNA load as determined by cell culture as a management tool  
637 for discharge of SARS-CoV-2 patients from infectious disease wards. *Eur J Clin Microbiol*  
638 *Infect Dis* 39, 1059–1061. 10.1007/s10096-020-03913-9.
- 639 21. Huang, C.-G., Lee, K.-M., Hsiao, M.-J., Yang, S.-L., Huang, P.-N., Gong, Y.-N., Hsieh, T.-H.,  
640 Huang, P.-W., Lin, Y.-J., Liu, Y.-C., et al. (2020). Culture-Based Virus Isolation To Evaluate  
641 Potential Infectivity of Clinical Specimens Tested for COVID-19. *J Clin Microbiol* 58, e01068-  
642 20. 10.1128/JCM.01068-20.
- 643 22. Artic Network <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>.
- 644 23. Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., Oliveira, G.,  
645 Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A., et al. (2017). Multiplex PCR method for  
646 MinION and Illumina sequencing of Zika and other virus genomes directly from clinical  
647 samples. *Nat Protoc* 12, 1261–1276. 10.1038/nprot.2017.066.

- 648 24. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P.,  
649 Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution.  
650 *Bioinformatics* 34, 4121–4123. [10.1093/bioinformatics/bty407](https://doi.org/10.1093/bioinformatics/bty407).
- 651 25. Pratt, J.W., and Gibbons, J.D. *Concepts of nonparametric theory* (Springer International  
652 Publishing).
- 653 26. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and  
654 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*  
655 (Methodological) 57, 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- 656 27. Tukey, J. (1953). Multiple Comparisons. *J. Am. Stat. Assoc.* 48, 624–625.
- 657 28. Morgan, A. COVIRAL webapp back-end. <https://github.com/chhotii-alex/antigen-flask>.
- 658 29. Morgan, A. COVIRAL webapp front-end. <https://github.com/chhotii-alex/antigen-sensitivity>.
- 659 30. Lamb, Y.N. (2020). Remdesivir: First Approval. *Drugs* 80, 1355–1363. [10.1007/s40265-020-](https://doi.org/10.1007/s40265-020-01378-w)  
660 [01378-w](https://doi.org/10.1007/s40265-020-01378-w).
- 661 31. Ahmed, A., Song, Y., and Wadhera, R.K. (2022). Racial/Ethnic Disparities in Delaying or Not  
662 Receiving Medical Care During the COVID-19 Pandemic. *J GEN INTERN MED* 37, 1341–1343.  
663 [10.1007/s11606-022-07406-7](https://doi.org/10.1007/s11606-022-07406-7).
- 664 32. Watson, C. (2022). Rise of the preprint: how rapid data sharing during COVID-19 has  
665 changed science forever. *Nature Medicine* 28, 2–5. [10.1038/s41591-021-01654-6](https://doi.org/10.1038/s41591-021-01654-6).
- 666 33. Rando, H.M., Boca, S.M., McGowan, L.D., Himmelstein, D.S., Robson, M.P., Rubinetti, V.,  
667 Velazquez, R., Greene, C.S., and Gitter, A. (2021). An Open-Publishing Response to the  
668 COVID-19 Infodemic. *CEUR Workshop Proc* 2976, 29–38.
- 669 34. Ross, J.S. (2021). Covid-19, open science, and the CVD-COVID-UK initiative. *BMJ* 373, n898.  
670 [10.1136/bmj.n898](https://doi.org/10.1136/bmj.n898).

- 671 35. Callahan, C.J., Lee, R., Zulauf, K.E., Tamburello, L., Smith, K.P., Previtiera, J., Cheng, A., Green,  
672 A., Azim, A.A., Yano, A., et al. (2020). Open Development and Clinical Validation Of Multiple  
673 3D-Printed Nasopharyngeal Collection Swabs: Rapid Resolution of a Critical COVID-19  
674 Testing Bottleneck. *J Clin Microbiol*, JCM.00876-20, jcm;JCM.00876-20v1.  
675 10.1128/JCM.00876-20.
- 676 36. Arnaout, R.A. (2021). Cooperation under Pressure: Lessons from the COVID-19 Swab Crisis.  
677 *Journal of Clinical Microbiology* 59, e01239-21. 10.1128/JCM.01239-21.
- 678 37. Tse, E.G., Klug, D.M., and Todd, M.H. (2020). Open science approaches to COVID-19.  
679 *F1000Res* 9, 1043. 10.12688/f1000research.26084.1.
- 680 38. Perillat, L., and Baigrie, B.S. (2021). COVID-19 and the generation of novel scientific  
681 knowledge: Evidence-based decisions and data sharing. *J Eval Clin Pract* 27, 708–715.  
682 10.1111/jep.13548.
- 683 39. Li, R., von Isenburg, M., Levenstein, M., Neumann, S., Wood, J., and Sim, I. (2021). COVID-19  
684 trials: declarations of data sharing intentions at trial registration and at publication. *Trials*  
685 22, 153. 10.1186/s13063-021-05104-z.
- 686 40. Gardener, A.D., Hick, E.J., Jacklin, C., Tan, G., Cashin, A.G., Lee, H., Nunan, D., Toomey, E.C.,  
687 and Richards, G.C. (2022). Open science and conflict of interest policies of medical and  
688 health sciences journals before and during the COVID-19 pandemic: A repeat cross-  
689 sectional study: Open science policies of medical journals. *JRSM Open* 13,  
690 20542704221132139. 10.1177/20542704221132139.
- 691 41. Besançon, L., Peiffer-Smadja, N., Segalas, C., Jiang, H., Masuzzo, P., Smout, C., Billy, E.,  
692 Deforet, M., and Leyrat, C. (2021). Open science saves lives: lessons from the COVID-19  
693 pandemic. *BMC Med Res Methodol* 21, 117. 10.1186/s12874-021-01304-y.
- 694 42. Wood, A., Denholm, R., Hollings, S., Cooper, J., Ip, S., Walker, V., Denaxas, S., Akbari, A.,  
695 Banerjee, A., Whiteley, W., et al. (2021). Linked electronic health records for research on a

- 696 nationwide cohort of more than 54 million people in England: data resource. *BMJ* 373,  
697 n826. [10.1136/bmj.n826](https://doi.org/10.1136/bmj.n826).
- 698 43. Rao, S.N., Manissero, D., Steele, V.R., and Pareja, J. (2020). A Systematic Review of the  
699 Clinical Utility of Cycle Threshold Values in the Context of COVID-19. *Infect Dis Ther* 9, 573–  
700 586. [10.1007/s40121-020-00324-3](https://doi.org/10.1007/s40121-020-00324-3).
- 701 44. Satlin, M.J., Chen, L., Gomez-Simmonds, A., Marino, J., Weston, G., Bhowmick, T., Seo, S.K.,  
702 Sperber, S.J., Kim, A.C., Eilertson, B., et al. (2022). Impact of a Rapid Molecular Test for  
703 *Klebsiella pneumoniae* Carbapenemase and Ceftazidime-Avibactam Use on Outcomes After  
704 Bacteremia Caused by Carbapenem-Resistant Enterobacterales. *Clinical Infectious Diseases*  
705 75, 2066–2075. [10.1093/cid/ciac354](https://doi.org/10.1093/cid/ciac354).
- 706 45. Savela, E.S., Vilorio Winnett, A., Romano, A.E., Porter, M.K., Shelby, N., Akana, R., Ji, J.,  
707 Cooper, M.M., Schlenker, N.W., Reyes, J.A., et al. (2022). Quantitative SARS-CoV-2 Viral-  
708 Load Curves in Paired Saliva Samples and Nasal Swabs Inform Appropriate Respiratory  
709 Sampling Site and Analytical Test Sensitivity Required for Earliest Viral Detection. *J Clin*  
710 *Microbiol* 60, e0178521. [10.1128/JCM.01785-21](https://doi.org/10.1128/JCM.01785-21).
- 711 46. Vilorio Winnett, A., Porter, M.K., Romano, A.E., Savela, E.S., Akana, R., Shelby, N., Reyes,  
712 J.A., Schlenker, N.W., Cooper, M.M., Carter, A.M., et al. (2022). Morning SARS-CoV-2  
713 Testing Yields Better Detection of Infection Due to Higher Viral Loads in Saliva and Nasal  
714 Swabs upon Waking. *Microbiol Spectr* 10, e0387322. [10.1128/spectrum.03873-22](https://doi.org/10.1128/spectrum.03873-22).
- 715 47. Brümmer, L.E., Katzenschlager, S., Gaeddert, M., Erdmann, C., Schmitz, S., Bota, M., Grilli,  
716 M., Larmann, J., Weigand, M.A., Pollock, N.R., et al. (2021). Accuracy of novel antigen rapid  
717 diagnostics for SARS-CoV-2: A living systematic review and meta-analysis. *PLoS Med* 18,  
718 e1003735. [10.1371/journal.pmed.1003735](https://doi.org/10.1371/journal.pmed.1003735).
- 719 48. Lee, R.A., Herigon, J.C., Benedetti, A., Pollock, N.R., and Denkinger, C.M. (2021).  
720 Performance of Saliva, Oropharyngeal Swabs, and Nasal Swabs for SARS-CoV-2 Molecular

- 721           Detection: A Systematic Review and Meta-analysis. *Journal of Clinical Microbiology*.  
722           10.1128/JCM.02881-20.
- 723   49. El Emam, K. (2013). *Guide to the de-Identification of Personal Health Information* (Auerbach  
724           Publishers, Incorporated).
- 725   50. El Emam, K., Rodgers, S., and Malin, B. (2015). Anonymising and sharing individual patient  
726           data. *BMJ* 350, h1139. 10.1136/bmj.h1139.
- 727