medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Title: Antibody response to infectious diseases and other factors accurately predict COVID-19 infection and severity risk 10-14 years later: a retrospective UK Biobank cohort study

Willette AA[1,2,†]; Willette SA[1]; Wang Q[1]; Pappas C[1]; Klinedinst BS[1]; Le S[1]; Larsen B[1]; Pollpeter A[1]; Brenner N[3]; Waterboer T[3]

(1) Department of Food Science and Human Nutrition, Iowa State University, Ames, IA, USA

(2) Department of Neurology, University of Iowa, Iowa City, IA, USA

(3) Infections and Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

Background: Several risk factors have emerged for novel 2019 coronavirus disease (COVID-19) infection and severity. Yet, it is unknown to what degree these risk factors alone or in combination can accurately predict who is most at risk. It is also worthwhile to consider serological antibody titers to non COVID-19 infectious diseases, which may influence host immunity to COVID-19.

Methods: In this retrospective study of multicenter UK Biobank participants, as of May 26th 2020, all COVID-19 testing data was collected by Public Health England for older adult in- and out-patients (69.6 ± 8.8 years). We used linear discriminant analysis with cross-validation and bootstrapping to determine the accuracy, specificity, and sensitivity of baseline data from 2006-2010 to predict COVID-19 infection and presumptive severity (i.e., testing at hospital). Receiver operating characteristic (ROC) curves were used to derive the area under the curve (AUC).

Findings: This retrospective study included 4,510 unique participants and 7,539 testing instances (i.e., test cases). Testing resulted in 5,329 negative cases and 2,210 positive cases, split into 996 mild and 1,214 severe disease outcomes. Baseline data including demographics, bioimpedance-derived body composition, vitals, serum biochemistry, self-reported illness/disability, and complete blood count. A randomized subset of 80 participants with 124 test cases also had antibody titers for 20 common to rare infectious diseases. Among all test cases, accuracy was modest for final diagnostic models of COVID-19 infection (70.2%; AUC=0.570, CI=0.556-0.584) and severity (58.3%; AUC=0.592, CI=0.568-0.615). In the serology sub-group, by contrast, final models predicted infection and severity with an accuracy of 93.5% (AUC=0.969, CI=0.934-1.000) and 74.4% (AUC=0.803, CI=0.663-0.943) respectively. Models included titers to common pathogens (e.g., human cytomegalovirus), age, blood cell counts, lipids, and other biochemical markers.

Interpretation: Risk profiles including serological titers and other risk factors could help policy makers and clinicians better identify who may get COVID-19 and require hospitalization.

†Address Correspondence to:
Auriel A. Willette, Ph.D., M.S. (Mr.)
2302 Osborn Drive
Ames, IA 50011-1078
Phone: (515) 294-3110
Email: Awillett@iastate.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Coronavirus disease 2019 (COVID-19), caused by a novel beta-coronavirus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1], has become a worldwide pandemic, severely disrupting the economic, social, and psychological well-being of countless people. Clinical presentation of COVID-19 widely varies, ranging from asymptomatic profiles to mild symptoms like high fever or cough to acute respiratory disease syndrome and death. Given this heterogeneous symptom presentation, as well as difficulties with serology testing and contact tracing, worldwide public health efforts continue to focus on containment and especially isolating adults most at risk for COVID-19 infection and severe disease.

By extension, an expanding body of research has investigated potential factors that increase COVID-19 infection and disease severity risk. It is well known, for example, that adults aged >65 years are much more likely to be hospitalized or die due to COVID-19. Obesity itself and adverse health behaviors like smoking also increase infection risk and likelihood of hospitalization[2,3]. Several age and obesity-related conditions such as cardiovascular disease, cardiometabolic diseases (e.g., type 2 diabetes), hypertension, and other disease states and syndromes are also of concern[4]. Non-white ethnicity, particularly being black regardless of country of origin, socioeconomic deprivation, and low levels of education even after adjustment for health factors point to less privilege unfortunately conferring risk[5]. Among biological markers, COVID-19 infection or severity has been related to higher C-Reactive Protein and more circulating white blood cells and lower counts of lymphocytes or granulocytes (e.g., monocytes)[6-8]. SARS-CoV-1 has a similar profile except for a relatively normal total white blood cell count[9].

These studies are invaluable for establishing or validating risk factors to guide clinical decisions and policymaker choices. However, we ultimately need to develop risk profiles derived from these factors to accurately predict who will and will not develop COVID-19, and if a COVID-19 disease course will be mild or presumptively severe (i.e., require hospitalization). Machine learning, data-driven modelling can be used to create robust, highly accurate prediction models based on routinely collected biomedical data like demographics, a complete blood count, and standard medical biochemistry data. Critically, using non COVID-19 serological data, we may gain insight into the host's ability to fight COVID-19 by examining antibody titers that detail the host response to past infectious pathogens. This "virome" may affect host innate and adaptive immunity[9,10]. For example, human cytomegalovirus vastly changes the composition of T and B cells[11], and may induce immune senescence that could account for worse SARS-CoV-2 infection outcomes.

Therefore, similar to previous work[12], our objective was to use machine learning to determine what combinations of baseline measures, collected 10-14 years ago, could best predict which older adults developed COVID-19 and if disease presentation was mild or severe. In summary, we achieved a 93.5% accuracy for predicting COVID-19 infection based on a combination of age, biochemistry and leukocyte markers, and antibody titers to common pathogens like human cytomegalovirus, human herpesvirus 6, and chlamydia trachomatis. For COVID-19 severity, due to small sample size, only antibody titers loaded for finals models that more modestly predicted severe disease (74.4%). Nonetheless, this is the first report to propose retrospective risk factor profiles to clarify and better characterize who is most at risk for COVID-19. In addition, our results suggest that past infection history and antibody response may be an invaluable, novel predictor of host immunity to COVID-19 that warrants further study.

Methods

Study design and participants

This retrospective study involved the UK Biobank cohort[13]. UK Biobank consists of approximately 500,000 people now aged 50 to 84 years (mean age=69.4 years). Baseline data was collected in 2006-2010 at 22 centers across the United Kingdom[14,15]. Summary data is listed in **Table 1**.

Our study used the May 26th, 2020 tranche of COVID-19 polymerase chain reaction (PCR) data from Public Health England. The following categories of predictors were downloaded 1) demographics; 2) health behaviors and long-term

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

disability or illness status; 3) anthropometric and bioimpedance measures of fat, muscle, and water content; 4) pulse and blood pressure; 5) a serum panel of thirty biochemistry markers commonly collected in a hospital setting; and 6) a complete blood count with a manual differential for quantitation of total white blood cells and sub-types. Among a randomized subset of 9,695 participants, as part of a separate pilot project, baseline serum was thawed and tested to determine levels of antibodies to several antigens of 20 infectious diseases. Study subjects provided electronic, signed informed consent at recruitment. Ethics approval for the UK Biobank study was obtained from the National Health Service Health Research Authority North West - Haydock Research Ethics Committee (16/NW/0274). The detailed protocol is outlined at https://www.ukbiobank.ac.uk/.

COVID-19 Testing

Through a May 26th 2020 data upload from UK Biobank, our study was based on COVID PCR test data available from March 16th to May 19th 2020. There were 4,510 unique participants that had 7,539 individual tests administered, hereafter called cases or test cases. For modeling COVID-19 infection data, each test case was coded by UK Biobank as '0' and '1', respectively representing a negative or positive PCR test. For modeling COVID-19 disease severity, each test case was coded as '0' and '1', which represented out-patient testing (i.e., mild case) or hospital in-patient testing with clinical signs of infection (i.e., presumptively severe case).

To offer insight into this frequently updated resource, roughly weekly updates by Public Health England since inception (late April/early May) have so far consisted of new participants tested for the first time (35.6%), or who have follow-up testing (64.4%). As of the May 26th, 2020 upload (see **Table 1**), a given participant had anywhere between 1 and 20 tests for COVID-19 (mean=2.5 tests), with 1.8 ± 4.1 days between each test. Based on initial modelling, there are not yet enough test cases per participant to robustly model complex changes in disease status. As of June 5th, 2020, an in- vs. outpatient identification issue had been raised for 6 out of 105 hospitals and clinics. Thankfully, we found no evidence in our prediction models that laboratory of clinical origin influenced our results.

Demographics

These factors included participant age in years at baseline, sex, education qualifications, ethnicity, and Townsend deprivation index. Sex was coded as 0 for female and 1 for male. For education, higher scores roughly correspond to progressively more skilled trade/vocational or academic training and skill need to attain the qualification. Ethnicity was coded as UK citizens who identified as White, Black/Black British, or Asian/Asian British. The Townsend index[16] is a standardized score, based on postal code (i.e., zip code) data taken from the census, indicating the relative degree of deprivation or poverty presumably experienced by the participant based on their permanent address.

Health Behaviors and Conditions

This category consisted of self-reported alcohol status, smoking status, a subjective health rating on a 1-4 Likert scale ("Excellent" to "Poor"), and whether the participant had a self-described long-term medical condition, illness, or disability. As noted in **Table 1**, 48.4% of participants indicated having such an ailment. We independently confirmed with ICD-10 based, NHS-confirmed diagnoses that this self-report data was accurate. These conditions included all-cause dementia and other neurological disorders, various cancers, major depressive disorder, cardiovascular (e.g., myocardial infarction) or cerebrovascular diseases and events (e.g., stroke), cardiometabolic diseases (e.g., type 2 diabetes), renal and pulmonary diseases (e.g., COPD), and other so called pre-existing conditions. We chose to use this single latent variable for simplicity, and because there were widely varying numbers of cases that severely underpowered our multivariable classifier analyses.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Vital Signs

The first automated reading of pulse, diastolic and systolic blood pressure at the baseline visit were used.

Body Morphometrics and Compartment Mass

Anthropometric measures of adiposity (Body Mass Index, waist circumference) were derived as described[17]. Data also included bioelectrical impedance metrics that estimate central body cavity (i.e., trunk) and whole body fat mass, fat-free muscle mass, and/or water content[18].

Blood Biochemistry

Serum biomarkers were assayed from baseline samples as previously described[19]. See **Table 1** for data summaries of the full COVID-19 sample and the sub-group with serology data. Briefly, using immunoassay or clinical chemistry devices, spectrophotometry was used to initially quantify values for 34 biochemistry analytes. UK Biobank deemed 30 of these markers to be suitably robust, after rigorous quality control to minimize systematic bias and random error in sample thawing and processing. We downloaded all fully quality-controlled data from the main showcase. We rejected a further 4 markers due data missingness >70% (estradiol, rheumatoid factor), or because there was strong overlap with other variables that had more stable distributions or trait-like qualities (glucose rejected vs. glycated hemoglobin or hba1c; direct bilirubin rejected vs. total bilirubin).

Serology Measures for Non COVID-19 Pathogenic Diseases

As described (http://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/infdisease.pdf), among 9,695 randomized UK Biobank participants selected from the full cohort, baseline serum was thawed and pathogen-specific assays run in parallel using flow cytometry on a Luminex bead platform[20].

Here, the goal of the multiplex serology panel was to measure multiple antibodies against several antigens for different pathogens, reducing noise and estimating the prevalence of prior infection and seroconversion in at least UK Biobank. All measures were initially confirmed in serum samples using gold-standard assays with median sensitivity and specificity of 97.0% and 93.7%, respectively. Antibody load for each pathogen-specific antigen was quantified using median fluorescence intensity (MFI). CagA titer load to H. pylori was excluded due to lab-based data loss. Because seropositivity is difficult to assess for several pathogens, we did not use pathogen prevalence as a predictor in models.

Table 2 shows the selected pathogens, their respective antigens, estimated prevalence of each pathogen based roughly on antibody titers, and assay values. This array ranges from delta-type retroviruses like human T-cell lymphotropic virus 1 that are rare (<1%) to human herpesviruses 6 and 7 that have an estimated prevalence of more than 90%.

Statistical Analysis

SPSS (Subscription build 1.0.0.1327) was used for all analyses. Due to differences in sample size, Mann-Whitney U and Kruskal-Wallis tests were used to compare quantitative values and categories (e.g., sex) for all 7,539 test cases and the 124 test cases with serology data (i.e., the serology sub-group). Linear discriminant analysis (LDA) was then leveraged, using individual predictors or weighted combinations of predictors, to maximally distinguish between: 1) negative or positive diagnosis for COVID-19; and 2) mild or severe COVID-19 disease status. LDA relies on a regression-like linear set of functions that can combine several data (i.e., features) and create predictive models that are straightforward to interpret. It is recognized that having a small number of test cases, such as in the serology sub-cohort, with many data types and features can lead to overfitting[21]. To guard against non-robust estimations, parametric violations, and model

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

overfitting, 1-fold cross-validation with non-parametric bootstrapping (95% Confidence Interval, 1000 iterations) was used. While logistic regression is more robust to outliers than LDA, UK Biobank data is vigorously quality-controlled to remove extreme values. Due to the small sample size of the serology sub-group, logistic regression also would be more likely to have model overfitting that inflates true accuracy.

First, LDA was used to examine how useful each baseline predictor was for correctly determining COVID-19 infection classification (negative, positive) and disease severity (mild, severe). This was done separately for all 7,539 test cases and the 124 test cases in the serology sub-group. Next, a series of forced entry models were used to see how well a set of related variables or features (e.g., demographics) predicted COVID-19 infection or disease severity. We recognize that some of these forced entry models are likely overfitted, particularly for modeling disease severity risk. Nonetheless, these models may provide a "best case scenario" for how well (or poorly) a class of predictors can perform in classification. Finally, a stepwise approach (Wilks' Lambda, F value entry=3.84) was used to combine predictors into a risk profile that best classified COVID-19 infection or separately for severity risk.

For each classification model, the accuracy (i.e., percentage of test cases that were correctly classified), sensitivity (i.e., true positives correctly identified), and specificity (i.e., true negatives correctly identified) were calculated. The area under the curve (AUC) with a 95% confidence interval (CI) was also used. Receiver operating characteristic (ROC) curves plotting sensitivity against 1-specificity were created to visualize differences in prediction accuracy among sets of similar predictors or stepwise models. For stepwise models, the Wilks' Lambda statistic and standardized coefficients were used to interpret how well and in what direction a given variable discriminated between positive vs. negative COVID-19 infection and mild vs. severe disease. A lower Wilks' Lambda corresponds to a stronger influence on the canonical classifier. Alpha was set at .05.

Role of the funding source

The funders of the study had no role in the study design, data collection, data analysis or interpretation, or writing of this report. The corresponding author (AAW) had full access to all of the data in this study and had final responsibility for the decision to submit the report for publication.

Results

As shown in **Table 1**, 7,539 total test cases for COVID-19 were conducted among 4,510 UK Biobank participants (69.6 ± 8.8 years) between March 16th to May 19rd 2020, either in outpatient or inpatient settings. There were 5,329 negative cases and 2,210 positive cases. Of the positive cases, there were 996 mild and 1,214 presumptively severe disease outcomes, defined as a test case occurring in a hospital setting. Baseline data from 10-14 years ago (Mean = 11.22 years) was available for demographic, laboratory, biochemistry, and clinical indices. A central theme of this report is the comparison of the 7,539 total test cases to a sub-group of 124 test cases with serology data (**Table 2**), in order to show that better model fit incorporating serology markers was not merely due to sample size differences or model inflation. Using non-parametric tests, then, **Table 1** indicates that the full cohort and serology sub-groups largely did not differ on most measures. A few significant differences were clinically unremarkable for the serology sub-cohort and well within the range of normal values, including lower pulse rate, several markers reflecting better kidney function, and a mean 0.6 10^9/L lower total white blood cell count due to fewer lymphocytes.

Next, each baseline variable was used to predict COVID-19 infection for a given test case. For context, 70.6% of the 7,539 test cases were negative. Consequently, any predictor achieving an accuracy of 70.6% would be performing at chance. A better measure of accuracy in this case is the AUC, where 0.5 is at-chance prediction and 1.0 is perfect accuracy. We also focused on how well true COVID-19 positive cases were identified (i.e., sensitivity). Among all participants (**Supplementary Table 1**), any given significant predictor could not correctly distinguish any true positive test cases (0% sensitivity; AUC mean and range=0.525, 0.515-0.548). For the serology sub-cohort (**Supplementary Table**

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

2), several established risk factors that loaded had better overall fit (mean AUC=0.625, AUC range=0.528-0.712), due to better sensitivity of predicting infection. Examples included ethnicity (13.2%), alcohol status (15.4%), apolipoprotein B (10.3%), and two unusually strong biochemistry analytes: urate (25.6%) and testosterone (56.4%). In order to see if biomarkers of past host response to pathogens was useful for predicting a current host response to COVID-19, we then tested each antibody titer for an antigen to a specific pathogen. As shown in **Supplementary Table 3**, antibody titers to 14 antigens across 12 pathogens each performed as well on average as other types of predictors (mean AUC=0.627, AUC range=0.505-0.707). In particular, sensitivity was notable for antibody to the pp150 Nter antigen to Human Cytomegalovirus (33.3%) and BK VP1 to Human T Lymphotropic Virus 1 (30.8%).

Lastly, as listed in **Table 3**, sets of similar predictors were forced into a classifier model to gauge how well they collectively predicted COVID-19 infection. A stepwise model was also used to create a classifier that only included predictors which each provided unique predictive utility. Among all 7,539 test cases (top row), sets of predictors including the stepwise model were only able to correctly identify COVID-19 positive test cases up to 10% of the time. **Supplementary Table 4** illustrates that predictors loading in the stepwise model included lipid and kidney health markers, white cell counts, as well as smoking status, ethnicity, and the Townsend Deprivation Index.

In the serology sub-group (bottom row), some relatively sparse predictor sets had better sensitivity (e.g., 53.3%). While the biochemistry and serology forced entry models were likely overfitted, the analyses may nonetheless provide a “best case scenario” for their usefulness as a group. Notably, the stepwise model achieved 93.5% accuracy by correctly identifying when a COVID-19 test case was negative (94.1% specificity) or positive (92.3% sensitivity). Due to potential concerns with model overfitting, the stepwise model was re-run with only predictors that had individually loaded significantly (**Supplementary Tables 2 and 3**). This model had 6 variables and still achieved 79.8% accuracy. As shown in **Supplementary Table 4**, predictors that loaded in the stepwise model included antibody titers for antigens of several common pathogens (e.g., Human Cytomegalovirus, Chlamydia Trachomatis), lipid markers, age in years, white and red cell counts, and testosterone.

Another set of analyses next determined how each baseline predictor could predict which of the 2,210 positive COVID-19 cases had a mild or severe disease course. For context, 45% and 55% of test cases were mild or severe respectively. Thus, accuracy of 50% would be considered chance prediction. Curiously, while sensitivity was the difficult metric to achieve for COVID-19 infection risk, accurately distinguishing true negatives (i.e., specificity) was problematic for disease severity. Among all 2,210 COVID-19 positive test cases (**Supplementary Table 5**), significant predictors showed a trade-off between better sensitivity or specificity and in general were only modestly useful (AUC mean and range=0.536, 0.524-0.572). Similarly, for the serology only sub-group among 39 COVID-19 positive test cases, **Supplementary Table 6** shows that only alanine aminotransferase and neutrophil count significantly predicted disease severity beyond chance. Likewise, for serology data, **Supplementary Table 7** indicates that the only significant antibodies to load were for the U14 antigen to human herpesvirus 7 (accuracy=64.1%; AUC=0.729) and JC VP1 antigen to human JC polyomavirus (accuracy=59%; AUC 0.671).

Table 4 shows the relative predictive value of groups of predictors for COVID-19 severity. First, for the full sample of 2,210 positive test cases, accuracy remained low and the proportion of true negatives (i.e., specificity) identified did not exceed 37%. This was regardless of predictor sets with a sparse or dense number of predictors. **Supplementary Table 8** illustrates that the stepwise model included only alanine aminotransferase, age in years, and monocyte count, which may explain its modest predictive utility above chance. For the serology sub-group of 39 test cases, despite strong concerns about model overfitting, the accuracy, sensitivity, and specificity were similarly modest compared to all 2,210 positive test cases for the biochemistry, immunology, and serology panels. The stepwise model was sparse and had better overall accuracy (74.4%) due to improved detection of actual mild cases (61.5%). Indeed, **Supplementary Table 8** shows that the stepwise model loaded 2 predictors, antibody titers for HTLV-1 gag antigen to the rare Human T Lymphotropic virus and JC VP1 antigen for the Human Polyomavirus that has an estimated prevalence of 57.5% in at least UK Biobank.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

The objective of this study was to determine if baseline data from 2006-2010 could predict which older adults would develop COVID-19 in 2020, and if that infection was presumptively mild or severe due to being at hospital. In summary, using machine learning, we developed separate risk profiles that accurately predicted future host immunity for COVID-19 infection (93.5%) and severity (74.4%). Such profiles only require retrospective, routine self-report and blood tests typically collected in out- and in-patient clinics and hospitals. As proof-of-principle that these profiles work, for example, we confirmed as others have noted with previous UK Biobank COVID-19 data that non-white ethnicity, low socioeconomic status, and smoking can increase infectious risk[5].

Our most novel finding was that antibody titers, reflecting pathogen exposure history and past host immunity, were strong predictors of COVID-19 infection and severity, both as a group and especially in concert with established risk factors like age, neutropenia, and dyslipidemia. This virome may consist of beneficial and detrimental pathogens that change how the immune system responds to a novel, persistent viral challenge like COVID-19[10]. For example, we found antigens to human cytomegalovirus were the strongest predictors of infection risk in our stepwise model. Older adults with prior human cytomegalovirus infection evince exhaustion of the naïve T cell pool and fewer memory versus effector cells[22]. This may explain why monocyte count was one of the few variables to predict COVID-19 severity among all test cases in this study, as innate immunity must compensate. For COVID-19 severity prediction, antibody titer to the JC polyomavirus was the only serology predictor that loaded significantly in our stepwise model and is expressed in a majority of the general population. This virus can induce hemagglutination in type O blood cells[23], which may in some way influence why this blood type may be protective for COVID-19 infection. This may also explain why higher red blood cell count appeared to be an important predictor for infection risk.

For other immunologic factors, mobilization of innate immunity was not surprisingly relevant to infection risk and severity. In particular, granulocytes (e.g., neutrophils, monocytes) loaded significantly in COVID-19 infection and severity prediction models for stepwise models, but not cytokines such as C-Reactive Protein. C-Reactive Protein has been cited as a strong risk factor for COVID-19[24]. However, this marker merely reflects signaling of the acute phase response due to systemic infection, typically being initiated by macrophages in contact with the pathogen and monocytes in the blood. Although lymphopenia and suppression of humoral immunity has been noted in COVID-19, lymphocyte cell count was in this study a modest predictor by itself and did not load in final stepwise models.

We also confirmed and extended the importance of age and biological factors related to lipids and kidney health, but curiously not obesity or comorbid conditions. Among now elderly adults in UK Biobank, age was one of the few factors to impact both infection and severity risk. Perhaps in concert, lipoprotein metabolism changes with aging along with sedentary lifestyle can induce hyperlipidemia, which is a risk factor for cardiovascular disease and may increase COVID-19 infection risk[25]. The lack of association with bioimpedance-derived fat, muscle, and water quantitation, or long-standing illness, was unexpected but may be due to complex interactions that are beyond the scope of this report. Finally, levels of testosterone by itself and in concert with other factors in the serology sub-group could strongly identify adults who would later develop COVID-19. Sex differences favoring COVID-19 infection in men have been noted, and andropause-induced reductions in testosterone occur in aging men. As testosterone normally downregulates inflammation, this loss may increase disease susceptibility[26].

Some major limitations should be noted in our study. At this time, the number of UK Biobank participants with COVID-19 and serology data is low, particularly for positive test cases. To temper this issue, we first used k-fold validation and rigorous bootstrapping to avoid model overfitting for the stepwise models. We also rigorously tested each predictor or set of predictors in the main sample or serology sub-group, where we found that model fit was not overly biased in general. Regardless, we acknowledge that sets of predictors with many variables may be overly optimistic in their prediction value. Larger sample sizes and gold standard classification schemes, such as training and testing using separate datasets, will be needed to validate that antibody titers and past pathogen history in general are relevant markers for COVID-19 infection and disease course. Another limitation was the modest predictive value of most

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

variables across all test cases, which stands in contrast to high odds ratios for some of these same factors in UK Biobank and other cohorts. Studies with smaller sample sizes will often show inflated relative risk or prediction accuracy, due simply to less heterogeneous error variance. However, our study only examined so called main effects of predictors instead of complex interactions, such as darker skin, vitamin D content, and COVID-19 infection risk. Such interactions were beyond the scope of this report, which attempted to create relatively straightforward risk profiles that could be used in a clinic or by policymakers.

In summary, this is the first study to systematically use retrospective data in a large community cohort to predict future risk for COVID-19 infection and severity. Despite baseline data having been collected 10-14 years ago, we nonetheless achieved excellent to encouraging accuracy by combining several sets of emerging risk factors together. It is especially interesting that serological data performed as well or better than any other data type. Future work should leverage past pathogen history and host immunity to inform what may happen when the host is challenged by COVID-19.

Acknowledgements

This research was conducted using the UK Biobank Resource under Application Number 25057. The study was funded by NIH AG047282 and AARGD-17-529552. No funding provider had any role in the conception, collection, execution, or publication of this work.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* 2020; **5**(4): 536-44.
2. Sattar N, McInnes IB, McMurray JJV. Obesity a Risk Factor for Severe COVID-19 Infection: Multiple Potential Mechanisms. *Circulation* 2020.
3. Simonnet A, Chetboun M, Poissy J, et al. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) requiring invasive mechanical ventilation. *Obesity (Silver Spring)* 2020.
4. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020; **395**(10229): 1054-62.
5. Patel AP, Paranjpe MD, Kathiresan NP, Rivas MA, Khera AV. Race, Socioeconomic Deprivation, and Hospitalization for COVID-19 in English participants of a National Biobank. *medRxiv* 2020: 2020.04.27.20082107.
6. Hamer M, Kivimaki M, Gale CR, David Batty G. Lifestyle risk factors, inflammatory mechanisms, and COVID-19 hospitalization: A community-based cohort study of 387,109 adults in UK. *Brain Behav Immun* 2020.
7. Liu Y, Yan LM, Wan L, et al. Viral dynamics in mild and severe cases of COVID-19. *Lancet Infect Dis* 2020; **20**(6): 656-7.
8. Qin C, Zhou L, Hu Z, et al. Dysregulation of immune response in patients with COVID-19 in Wuhan, China. *Clin Infect Dis* 2020.
9. Li T, Qiu Z, Zhang L, et al. Significant changes of peripheral T lymphocyte subsets in patients with severe acute respiratory syndrome. *J Infect Dis* 2004; **189**(4): 648-51.
10. Moss P. "The ancient and the new": is there an interaction between cytomegalovirus and SARS-CoV-2 infection? *Immun Ageing* 2020; **17**: 14.
11. Chidrawar S, Khan N, Wei W, et al. Cytomegalovirus-seropositivity has a profound influence on the magnitude of major lymphoid subsets within healthy individuals. *Clin Exp Immunol* 2009; **155**(3): 423-32.
12. Willette AA, Calhoun VD, Egan JM, Kapogiannis D, Alzheimers Disease Neuroimaging I. Prognostic classification of mild cognitive impairment and Alzheimer's disease: MRI independent component analysis. *Psychiatry Res* 2014; **224**(2): 81-8.
13. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* 2015; **12**(3): e1001779.
14. Armstrong J, Rudkin J, Allen N, Crook D, Wilson D, Wyllie D. Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank.[Google Scholar]. 2020.
15. Hilton B, Wilson D, O'Connell A-M, et al. Incidence of Microbial Infections in English UK Biobank Participants: Comparison with the General Population. *medRxiv* 2020: 2020.03.18.20038281.
16. Phillimore P, Beattie A, Townsend P. Widening inequality of health in northern England, 1981-91. *Bmj* 1994; **308**(6937): 1125-8.
17. Klinedinst BS, Pappas C, Le S, et al. Aging-related changes in fluid intelligence, muscle and adipose mass, and sex-specific immunologic mediation: A longitudinal UK Biobank study. *Brain Behav Immun* 2019; **82**: 396-405.
18. Kotler DP, Burastero S, Wang J, Pierson RN, Jr. Prediction of body cell mass, fat-free mass, and total body water with bioelectrical impedance analysis: effects of race, sex, and disease. *Am J Clin Nutr* 1996; **64**(3 Suppl): 489S-97S.
19. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 2008; **37**(2): 234-44.
20. Waterboer T, Sehr P, Pawlita M. Suppression of non-specific binding in serological Luminex assays. *J Immunol Methods* 2006; **309**(1-2): 200-4.
21. Rhenman A, Berglund L, Brodin T, et al. Which set of embryo variables is most predictive for live birth? A prospective study in 6252 single embryo transfers to construct an embryo score for the ranking and selection of embryos. *Hum Reprod* 2015; **30**(1): 28-36.
22. Weinberger B, Lazuardi L, Weiskirchner I, et al. Healthy aging and latent infection with CMV lead to distinct changes in CD8+ and CD4+ T-cell subsets in the elderly. *Hum Immunol* 2007; **68**(2): 86-90.
23. Osborn JE, Robertson SM, Padgett BL, Zu Rhein GM, Walker DL, Weisblum B. Comparison of JC and BK human papovaviruses with simian virus 40: restriction endonuclease digestion and gel electrophoresis of resultant fragments. *Journal of Virology* 1974; **13**(3): 614-22.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

24. Liu W, Tao ZW, Wang L, et al. Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease. *Chin Med J (Engl)* 2020; **133**(9): 1032-8.
25. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama* 2020; **323**(11): 1061-9.
26. Maggio M, Basaria S, Ceda GP, et al. The relationship between testosterone and molecular markers of inflammation in older men. *J Endocrinol Invest* 2005; **28**(11 Suppl Proceedings): 116-9.

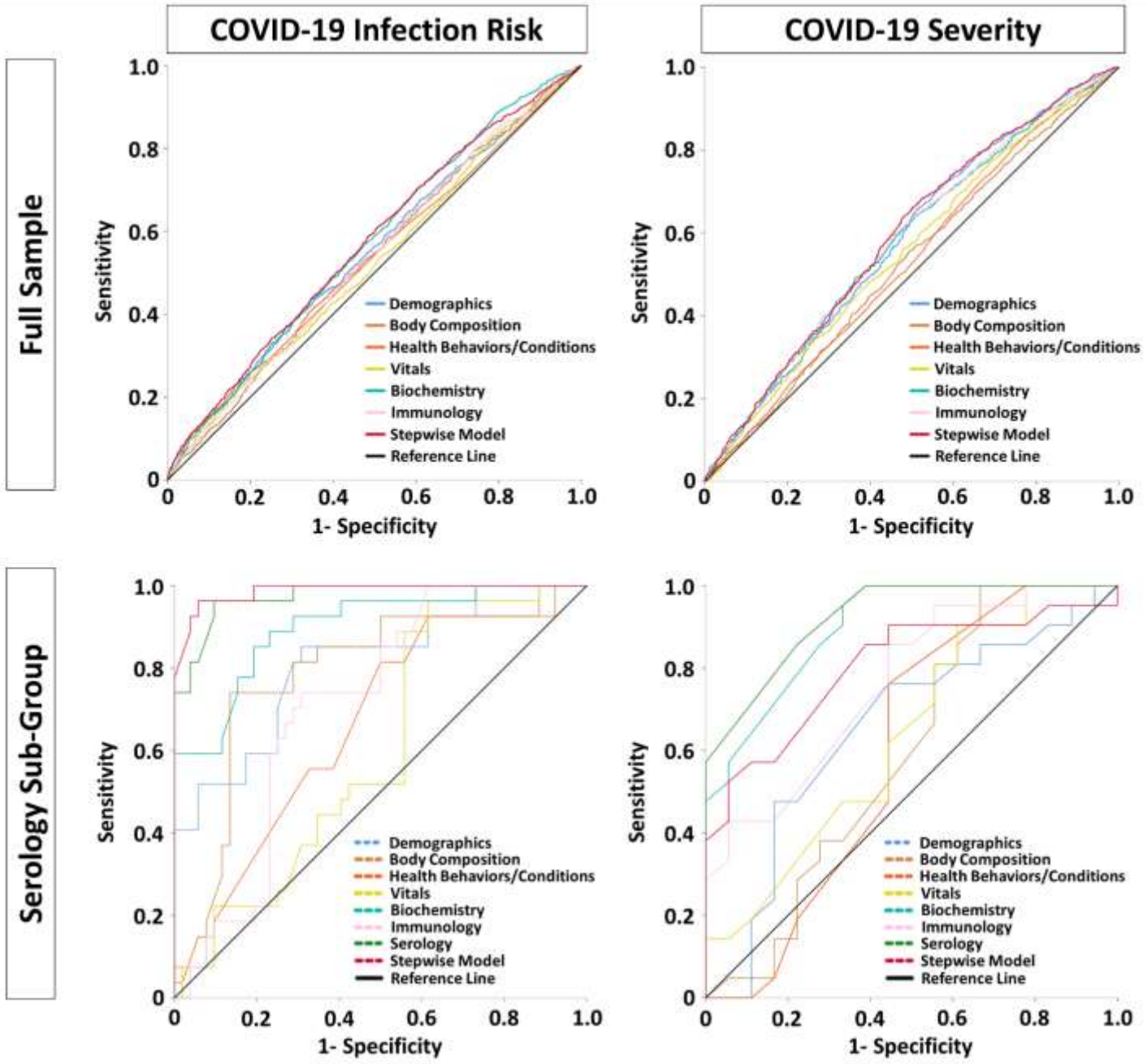
medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure Legends

Figure 1. The full sample consisted of all 7,539 test cases for COVID-19 infection risk, while there were 2,210 COVID-19 positive test cases examined for severity risk. Similarly, for the serology sub-group, there were 124 test cases and 39 test cases for COVID-19 infection and severity risk respectively. Test statistics for predictors are provided in Tables 3 and 4.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 1. Prediction accuracy of COVID-19 infection risk and severity among sets of predictors



medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Baseline Demographics and Data Characteristics

Variable	Unit	Full COVID-19 Sample	Serology Sub-Sample	P value
Total COVID-19 Test Cases	Testing Instance	7,539	124	
Total Participants		4510	80	
Test Cases per Participant		2.5 ± 1.6	2.6 ± 3.2	0.268
Mean Time between Tests	Days	2.0 ± 5.0	1.6 ± 3.2	0.951
Age at Testing	Years	69.6 ± 8.8	68.9 ± 8.4	0.474
COVID-19 Result				0.606
COVID-	Cases	5329	85	
COVID+	Cases	2210	39	
COVID-19 Severity*				0.983
Mild (i.e., outpatient)	Cases	996	18	
Severe (i.e., inpatient)	Cases	1214	21	
Age at Baseline	Years	57.5 ± 8.8	56.6 ± 8.3	0.373
Sex	% Female	48.9%	46.5%	0.692
Education Qualifications	Categories	2.59 ± 1.63	2.8 ± 1.6	0.332
Deprivation Index	Score	-0.1 ± 3.6	-1.0 ± 2.7	0.122
Ethnicity				0.353
White	%	89.4%	92.8%	
Asian or Asian British	%	3.4%	4.3%	
Black or Black British	%	4.5%	2.9%	
Other	%	2.7%	0.0%	
Smoking Status				0.091
Never	%	48.1%	56.5%	
Previous	%	38.2%	33.9%	
Current	%	13.0%	9.7%	
Alcohol Status				0.603
Never	%	6.6%	9.9%	
Previous	%	5.7%	4.2%	
Current	%	87.7%	85.9%	
Body Mass Index	kg/m^2	28.7 ± 5.7	29.8 ± 6.7	0.227
Waist Circumference	cm	95 ± 15	97 ± 17	0.693
Long-Term Medical Condition	% Present	49%	52%	0.400
Subjective Health Rating	1-4 Likert Scale	2.41 ± 0.83	2.5 ± 0.7	0.355
Pulse Rate	**Beats/Minute**	**71 ± 12**	**67 ± 10**	**0.003**
Diastolic BP	mmHg	83 ± 11	80 ± 9	0.088
Systolic BP	mmHg	140 ± 20	136 ± 17	0.768
Alanine Aminotransferase	U/L	24.4 ± 16.6	23.1 ± 10.1	0.583
Albumin	g/L	44.7 ± 2.8	44.6 ± 2.4	0.617
Alkaline Phosphatase	**U/L**	**88.0 ± 34.1**	**81.8 ± 23.3**	**0.031**
Apolipoprotein A	g/L	1.5 ± 0.3	1.5 ± 0.2	0.723
Apolipoprotein B	g/L	1.0 ± 0.2	1.0 ± 0.3	0.876
Aspartate Aminotransferase	U/L	27.0 ± 11.7	26.8 ± 12.0	0.835
Bilirubin	umol/L	9.0 ± 4.4	10.8 ± 7.3	0.667
Calcium	mmol/L	2.4 ± 0.1	2.4 ± 0.1	0.917
Cholesterol (Total)	mmol/L	5.5 ± 1.2	5.4 ± 1.2	0.493

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Creatinine	**umol/L**	**76.2 ± 30.2**	**79.2 ± 21.1**	**0.008**
Cystatin C	mg/L	1.0 ± 0.3	1.0 ± 0.2	0.162
Gamma Glutamyltransferase	U/L	45.0 ± 59.9	35.0 ± 28.2	0.901
HDL Cholesterol	mmol/L	1.4 ± 0.4	1.4 ± 0.3	0.558
Hemoglobin A1c	mmol/mol	37.6 ± 8.8	36.5 ± 4.3	0.275
Insulin-Like Growth Factor 1	nmol/L	21.0 ± 6.0	20.4 ± 4.8	0.784
LDL Cholesterol	mmol/L	3.4 ± 0.9	3.4 ± 0.9	0.687
Lipoprotein A	nmol/L	43.6 ± 48.9	43.5 ± 50.4	0.898
Phosphate	mmol/L	1.2 ± 0.2	1.1 ± 0.2	0.998
Protein (Total)	**g/L**	**72.5 ± 4.4**	**70.9 ± 4.2**	**0.003**
Sex Hormone Binding Globulin	nmol/L	50.5 ± 28.4	49.6 ± 27.0	0.728
Testosterone	nmol/L	7.1 ± 6.0	6.7 ± 5.6	0.975
Triglycerides	mmol/L	1.8 ± 1.1	1.8 ± 0.8	0.084
Urate	**umol/L**	**324.0 ± 90.5**	**353.4 ± 91.0**	**<.001**
Urea	**mmol/L**	**5.6 ± 1.9**	**5.9 ± 1.7**	**0.005**
Vitamin D	nmol/L	46.4 ± 21.4	47.1 ± 22.0	0.778
C-Reactive Protein	mg/L	3.2 ± 5.0	2.4 ± 3.3	0.212
Red Blood Cell Count	10^{12}/L	4.5 ± 0.4	4.5 ± 0.5	0.173
White Blood Cell Count	**10^9/L**	**7.2 ± 2.8**	**6.6 ± 1.4**	**0.002**
Neutrophils	10^9/L	4.4 ± 1.5	4.2 ± 1.3	0.220
Lymphocytes	**10^9/L**	**2.0 ± 2.1**	**1.8 ± 0.5**	**0.002**
Monocytes	10^9/L	0.5 ± 0.3	0.5 ± 0.1	0.389
Eosinophils + Basophils	10^9/L	0.2 ± 0.2	0.1 ± 0.1	0.162

A summary and comparison of data among either all participant test cases or a sub-group of test cases that also had non COVID-19 serology. Contemporary COVID-19 testing data has no shading. All retrospective baseline data from 2006-2010 has “blue” shading. Values are in Mean ± SD, percentages, or frequency. P values less than .05 were considered significant and applicable predictor values are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Baseline characteristics of infectious disease serology from 2006-2010

Pathogen Name	Abbreviation	UK Biobank Seroprevalence*	Antigen	Mean ± SD
Herpes Simplex Virus-1	HSV-1	69.8%	1gG	3567.9 ± 3001.3
Herpes Simplex Virus-2	HSV-2	16.2%	2mgG	382.4 ± 1180.4
Varicella Zoster Virus	VZV	92.5%	gE/gl	834.0 ± 900.0
Epstein-Barr Virus	EBV	94.7%	VCA p18	6972.0 ± 3272.9
			EBNA-1	4146.2 ± 3269.2
			ZEBRA	2246.5 ± 1658.3
			EA-D	2765.5 ± 2721.7
Human Cytomegalovirus	CMV	58.2%	pp150 Nter	1881.8 ± 2225.5
			pp 52	3284.8 ± 3296.7
			pp 28	1379.3 ± 1662.5
Human Herpesvirus-6	HHV-6	90.8%	IE1A	327.1 ± 391.9
			IE1B	575.1 ± 805.8
			p101 k	167.0 ± 416.6
Human Herpesvirus-7	HHV-7	94.7%	U14	771.8 ± 778.3
Kaposi's Sarcoma Associated Herpesvirus	KSHV	8.1%	LANA	158.1 ± 977.4
			K8.1	73.1 ± 95.0
Hepatitis B Virus	HBV	2.5%	HBc	15.6 ± 55.6
			HBe	49.6 ± 202.3
Hepatitis C Virus	HCV	0.3%	Core	6.7 ± 10.3
			NS3	37.7 ± 31.3
Toxoplasma gondii	T. gondii	28.0%	p22	51.4 ± 86.0
			sag1	121.1 ± 119.1
Human T Lymphotropic Virus 1	HTLV-1	1.6%	HTLV-1 gag	320.2 ± 357.9
			HTLV-1 env	32.8 ± 19.8
Human Immunodeficiency Virus	HIV	0.2%	HIV-1 gag	213.1 ± 452.4
			HIV-1 env	44.1 ± 24.9
Human Polyomavirus BKV	BKV	95.4%	BK VP1	3718.9 ± 2550.5
Human Polyomavirus JCV	JCV	57.5%	JC VP1	932.7 ± 1060.2
Merkel Cell Polyomavirus	MCV	66.7%	MC VP1	2454.8 ± 2366.0
Human Papillomavirus type-16	HPV 16	4.4%	L1	56.9 ± 60.2
			E6	19.3 ± 28.2
			E7	52.8 ± 104.2
Human Papillomavirus type-18	HPV 18	2.7%	L1	52.8 ± 53.1

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Chlamydia trachomatis	C. trachomatis	21.4%	momp D	103.3 ± 405.9
			momp A	42.9 ± 115.3
			tarp-D F1	96.2 ± 394.6
			tarp-D F2	171.8 ± 332.4
			PorB	23.8 ± 41.0
			pGP3	449.2 ± 1304.0
Helicobacter pylori	H. pylori	31.5%	CagA*	1725.5 ± 3135.3
			VacA	427.3 ± 1364.7
			OMP	696.7 ± 1503.0
			GroEL	779.0 ± 1799.4
			Catalase	437.2 ± 1407.8
			UreA	329.2 ± 1516.6

Antibody levels are specific to each antigen and expressed in Median Fluorescence Intensity (MFI) units. Seroprevalence of at least the main UK Biobank cohort was estimated on samples from 9,695 randomized participants, as described in white papers (see Methods). *CagA levels are based on roughly half of the original sample due to a technical lab error.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 3. Sets of predictors used to predict classification of COVID-19 test cases as negative or positive

	Sets of Predictors	Number of Predictors	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
All Test Cases	Basic Demographics	5	Enter	**<.001**	**0.577 (0.449-0.705)**	**70.0%**	**95.1%**	**9.5%**
	Body Composition*	3	Enter	**<.001**	**0.582 (0.477-0.687)**	**70.7%**	**100.0%**	**0%**
	Health Behaviors/Conditions	4	Enter	**<.001**	**0.554 (0.438-0.670)**	**69.9%**	**94.6%**	**10.2%**
	Vitals	3	Enter	**0.003**	**0.601 (0.490-0.712)**	**70.7%**	**99.8%**	**0.3%**
	Biochemistry	26	Enter	**<.001**	**0.532 (0.425-0.639)**	**70.7%**	**100.0%**	**0%**
	Immunology	8	Enter	**<.001**	**0.581 (0.479-0.683)**	**69.5%**	**96.6%**	**4.0%**
	Stepwise Model	9	Stepwise	**<.001**	**0.570 (0.556-0.584)**	**70.2%**	**95.3%**	**9.7%**
Serology Sub-Group	Basic Demographics	5	Enter	**<.001**	**0.751 (0.636-0.866)**	**62.6%**	**66.7%**	**53.3%**
	Body Composition*	3	Enter	**0.012**	**0.729 (0.618-0.840)**	**70.5%**	**98.8%**	**5.4%**
	Health Behaviors/Conditions	4	Enter	**<.001**	**0.660 (0.547-0.774)**	**71.7%**	**96.3%**	**18.4%**
	Vitals	3	Enter	0.261	0.637 (0.526-0.747)	67.2%	98.8%	0%
	Biochemistry	26	Enter	**<.001**	**0.903 (0.839-0.968)**	**74.1%**	**81.7%**	**59.5%**
	Immunology	8	Enter	0.393	0.680 (0.576-0.784)	62.0%	90.4%	0%
	Serology	44	Enter	**<.001**	**0.961 (0.927-0.994)**	**72.6%**	**80.0%**	**56.4%**
	Stepwise Model	15	Stepwise	**<.001**	**0.969 (0.934-1.000)**	**93.5%**	**94.1%**	**92.3%**

Area Under the Curve (AUC); Confidence Interval (CI). Non-parametric bootstrapping (1000 iterations, 95% CI) was used for robust estimation. *Due to several variables representing the same construct (i.e., being multicollinear), body composition consisted of: whole-body water mass; whole-body fat mass; whole-body non-fat mass (i.e., muscle, bone). P values less than .05 were considered significant and applicable sets of predictors are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 4. Sets of predictors used to predict classification of COVID-19 positive cases as mild or severe

	Predictor Classes	Number of Predictors	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
All Test Cases	**Basic Demographics**	5	Enter	**<.001**	**0.581 (0.557-0.605)**	**57.6%**	**36.9%**	**74.6%**
	Body Composition*	3	Enter	**0.025**	**0.528 (0.504-0.552)**	**55.1%**	**0.9%**	**99.5%**
	Health Behaviors/Conditions	4	Enter	**0.011**	**0.531 (0.507-0.556)**	**52.9%**	**0.0%**	**96.3%**
	Vitals	3	Enter	**<.001**	**0.554 (0.530-0.578)**	**55.1%**	**7.0%**	**94.6%**
	Biochemistry	26	Enter	**<.001**	**0.579 (0.555-0.602)**	**54.8%**	**22.1%**	**81.5%**
	Immunology	8	Enter	**<.001**	**0.581 (0.557-0.605)**	**55.1%**	**5.5%**	**95.8%**
	Stepwise Model	3	Stepwise	**<.001**	**0.592 (0.568-0.615)**	**58.2%**	**36.4%**	**76.1%**
Serology Sub-Group	**Basic Demographics**	5	Enter	0.964	0.652 (0.472-0.832)	35.9%	22.2%	47.6%
	Body Composition*	3	Enter	0.665	0.597 (0.407-0.786)	59.0%	33.3%	81.0%
	Health Behaviors/Conditions	4	Enter	0.994	0.598 (0.404-0.792)	23.7%	35.3%	14.3%
	Vitals	3	Enter	0.448	0.636 (0.459-0.814)	56.4%	44.4%	66.7%
	Biochemistry	26	Enter	**<.001**	**0.901 (0.808-0.993)**	**53.8%**	**33.3%**	**71.4%**
	Immunology	8	Enter	**<.001**	**0.763 (0.615-0.911)**	**59.0%**	**44.4%**	**71.4%**
	Serology^	36	Enter	**<.001**	**0.925 (0.847-1.000)**	**61.5%**	**38.9%**	**81.0%**
	Stepwise Model	2	Stepwise	**<.001**	**0.803 (0.663-0.943)**	**74.4%**	**61.1%**	**85.7%**

Area Under the Curve (AUC); Confidence Interval (CI). Non-parametric bootstrapping (1000 iterations, 95% CI) was used for robust estimation. *Due to several variables representing the same construct (i.e., being multicollinear), body composition consisted of: whole-body water mass; whole-body fat mass; whole-body non-fat mass (i.e., muscle, bone). ^= Due to the full serology panel of 44 antibody titers exceeding degrees of freedom, titers for 6 antigens were excluded for pathogens with the lowest estimated prevelance in the cohort (HIV, HCV, HTLV-1). P values less than .05 were considered significant and predictors and classification metrics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 1. Isolated effect of each non-serology predictor on COVID-19 risk among all test cases

Predictor	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
Basic Demographics						
Age	Enter	**<.001**	**0.520 (0.506-0.520)**	**70.7%**	**100%**	**0%**
Sex	Enter	**0.011**	**0.516 (0.502-0.530)**	**70.7%**	**100%**	**0%**
Ethnic Background	Enter	**<.001**	**0.517 (0.502-0.532)**	**70.7%**	**100%**	**0%**
Deprivation Index	Enter	**<.001**	**0.540 (0.525-0.554)**	**70.7%**	**100%**	**0%**
Education	Enter	0.265	0.504 (0.489-0.520)	71.3%	100%	0%
Body Composition						
Waist Circumference	Enter	**<.001**	**0.538 (0.524-0.552)**	**70.6%**	**100%**	**0%**
Body Mass Index	Enter	**<.001**	**0.543 (0.529-0.557)**	**70.6%**	**100%**	**0%**
Trunk Fat Mass	Enter	**<.001**	**0.530 (0.516-0.545)**	**70.6%**	**100%**	**0%**
Whole Body Fat Mass	Enter	**<.001**	**0.527 (0.513-0.542)**	**70.6%**	**100%**	**0%**
Whole Body Fat-Free Mass	Enter	**<.001**	**0.527 (0.513-0.542)**	**70.6%**	**100%**	**0%**
Whole Body Water Mass	Enter	**<.001**	**0.527 (0.512-0.541)**	**70.5%**	**100%**	**0%**
Health Behaviors and Conditions						
Smoking Status	Enter	**0.012**	**0.518 (0.504-0.533)**	**70.7%**	**100%**	**0%**
Alcohol Status	Enter	**0.018**	**0.517 (0.503-0.532)**	**70.7%**	**100%**	**0%**
Long-Term Medical Condition	Enter	0.668	0.497 (0.482-0.511)	70.5%	100%	0%
Health Rating	Enter	0.573	0.504 (0.490-0.518)	70.8%	100%	0%
Vitals						
Pulse Rate	Enter	0.648	0.503 (0.489-0.518)	70.9%	100%	0%
Diastolic BP	Enter	**0.003**	**0.518 (0.503-0.533)**	**70.9%**	**100%**	**0%**
Systolic BP	Enter	0.734	0.500 (0.485-0.514)	70.9%	100%	0%
Biochemistry						
Alanine Aminotransferase	Enter	0.663	0.505 (0.490-0.519)	70.7%	100%	0%
Albumin	Enter	0.184	0.509 (0.495-0.523)	70.7%	100%	0%
Alkaline Phosphatase	Enter	0.411	0.505 (0.490-0.519)	70.7%	100%	0%
Apolipoprotein A	Enter	**<.001**	**0.548 (0.532-0.563)**	**71.4%**	**100%**	**0%**
Apolipoprotein B	Enter	**0.016**	**0.519 (0.504-0.534)**	**70.9%**	**100%**	**0%**
Aspartate Aminotransferase	Enter	0.197	0.511 (0.496-0.525)	70.7%	100%	0%
Bilirubin (Total)	Enter	0.257	0.504 (0.490-0.519)	70.7%	100%	0%
Calcium	Enter	**0.014**	**0.517 (0.503-0.531)**	**70.7%**	**100%**	**0%**
Cholesterol (Total)	Enter	0.649	0.503 (0.488-0.519)	70.9%	100%	0%
Creatinine	Enter	**0.004**	**0.521 (0.507-0.535)**	**70.7%**	**100%**	**0%**
Cystatin C	Enter	0.066	0.513 (0.499-0.528)	70.7%	100%	0%
Gamma Glutamyltransferase	Enter	0.432	0.506 (0.492-0.520)	70.7%	100%	0%
HDL Cholesterol	Enter	**<.001**	**0.521 (0.507-0.536)**	**71.5%**	**100%**	**0%**
Hemoglobin A1c	Enter	**0.003**	**0.523 (0.508-0.538)**	**70.9%**	**100%**	**0%**
Insulin-Like Growth Factor 1	Enter	0.813	0.502 (0.487-0.517)	70.9%	100%	0%
LDL Cholesterol	Enter	0.188	0.510 (0.495-0.525)	70.9%	100%	0%
Lipoprotein A	Enter	0.209	0.511 (0.494-0.528)	71.1%	100%	0%
Phosphate	Enter	**0.049**	**0.515 (0.501-0.529)**	**70.7%**	**100%**	**0%**
Protein (Total)	Enter	0.159	0.510 (0.496-0.525)	70.7%	100%	0%
Sex Hormone Binding Globulin	Enter	**<.001**	**0.545 (0.531-0.560)**	**70.7%**	**100%**	**0%**
Testosterone	Enter	0.199	0.513 (0.499-0.527)	70.7%	100%	0%

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Triglycerides	Enter	**<.001**	**0.530 (0.515-0.545)**	**71.0%**	**100%**	**0%**
Urate	Enter	**0.003**	**0.519 (0.504-0.533)**	**70.7%**	**100%**	**0%**
Urea	Enter	0.203	0.509 (0.495-0.523)	70.7%	100%	0%
Vitamin D	Enter	0.746	0.506 (0.492-0.521)	70.7%	100%	0%
Immunology						
Red Blood Cell Count	Enter	**<.001**	**0.524 (0.509-0.539)**	**70.7%**	**100%**	**0%**
White Blood Cell Count	Enter	0.442	0.505 (0.490-0.520)	70.7%	100%	0%
C-Reactive Protein	Enter	0.171	0.511 (0.496-0.525)	70.9%	100%	0%
Neutrophils	Enter	**0.004**	**0.521 (0.507-0.536)**	**70.8%**	**100%**	**0%**
Lymphocytes	Enter	**0.009**	**0.518 (0.503-0.533)**	**70.8%**	**100%**	**0%**
Monocytes	Enter	0.152	0.505 (0.490-0.519)	70.8%	100%	0%
Eosinophils	Enter	0.968	0.514 (0.499-0.528)	70.5%	100%	0%
Basophils	Enter	0.141	0.510 (0.495-0.525)	70.6%	100%	0%

Area Under the Curve (AUC); Confidence Interval (CI). Sensitivity and specificity were the likelihood of correctly detecting when COVID-19 infection for a test case was present vs. not present. “Gray” and “white” shading are used to better visualize predictors within a set of similar variables. P values less than .05 were considered significant and applicable predictors and statistics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 2. Isolated effect of each non-serology predictor on COVID-19 risk among test cases with serology data

Predictor	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
Basic Demographics						
Age	Enter	0.418	0.545 (0.425-0.666)	68.5%	0%	0%
Sex	Enter	**0.003**	**0.665 (0.563-0.767)**	**68.5%**	**0%**	**0%**
Ethnic Background	Enter	**0.021**	**0.528 (0.415-0.642)**	**71.9%**	**98.8%**	**13.2%**
Deprivation Index	Enter	0.456	0.542 (0.430-0.654)	68.5%	0%	0%
Education	Enter	**0.016**	**0.650 (0.538-0.763)**	**69.6%**	**0%**	**0%**
Body Composition						
Waist Circumference	Enter	0.342	0.553 (0.449-0.658)	68.5%	100%	0%
Body Mass Index	Enter	0.317	0.556 (0.448-0.664)	68.5%	100%	0%
Trunk Fat Mass	Enter	0.707	0.521 (0.409-0.634)	69.7%	100%	0%
Whole Body Fat Mass	Enter	0.186	0.576 (0.469-0.683)	69.7%	100%	0%
Whole Body Fat-Free Mass	Enter	**0.008**	**0.650 (0.546-0.755)**	**67.2%**	**96.5%**	**0%**
Whole Body Water Mass	Enter	**0.012**	**0.642 (0.538-0.747)**	**68.0%**	**100%**	**0%**
Health Behaviors and Conditions						
Smoking Status	Enter	**0.018**	**0.632 (0.529-0.735)**	**68.5%**	**0%**	**0%**
Alcohol Status	Enter	**0.011**	**0.574 (0.461-0.687)**	**71.0%**	**96.5%**	**15.4%**
Long-Term Medical Condition	Enter	0.569	0.532 (0.421-0.644)	68.3%	0%	0%
Health Rating	Enter	0.773	0.516 (0.412-0.620)	68.5%	0%	0%
Vitals						
Pulse Rate	Enter	0.424	0.545 (0.438-0.652)	68.0%	0%	0%
Diastolic BP	Enter	0.058	0.393 (0.283-0.503)	67.2%	0%	0%
Systolic BP	Enter	0.515	0.537 (0.430-0.643)	68.0%	0%	0%
Biochemistry						
Alanine Aminotransferase	Enter	0.433	0.579 (0.461-0.696)	68.5%	100%	0%
Albumin	Enter	0.551	0.633 (0.529-0.738)	67.7%	98.8%	0%
Alkaline Phosphatase	Enter	0.056	0.608 (0.502-0.714)	68.5%	100%	0%
Apolipoprotein A	Enter	0.162	0.598 (0.479-0.717)	74.4%	100%	0%
Apolipoprotein B	Enter	**0.017**	**0.605 (0.501-0.708)**	**71.8%**	**100%**	**10.3%**
Aspartate Aminotransferase	Enter	0.175	0.503 (0.392-0.615)	70.2%	100%	5.1%
Bilirubin (Total)	Enter	0.080	0.556 (0.441-0.670)	62.9%	91.8%	0%
Calcium	Enter	0.611	0.558 (0.457-0.660)	68.5%	100%	0%
Cholesterol (Total)	Enter	0.156	0.579 (0.473-0.686)	68.5%	100%	0%
Creatinine	Enter	0.347	0.512 (0.403-0.621)	68.5%	100%	0%
Cystatin C	Enter	0.424	0.627 (0.514-0.740)	66.9%	97.6%	0%
Gamma Glutamyltransferase	Enter	0.092	0.598 (0.487-0.709)	68.5%	100%	0%
HDL Cholesterol	Enter	0.424	0.556 (0.433-0.679)	74.4%	100%	0%
Hemoglobin A1c	Enter	**0.029**	**0.626 (0.517-0.735)**	**62.4%**	**91.3%**	**0%**
Insulin-Like Growth Factor 1	Enter	0.619	0.528 (0.415-0.645)	68.5%	100%	0%
LDL Cholesterol	Enter	0.138	0.583 (0.478-0.689)	68.5%	100%	0%
Lipoprotein A	Enter	0.406	0.554 (0.427-0.681)	74.0%	100%	0%
Phosphate	Enter	**0.010**	**0.597 (0.492-0.702)**	**67.7%**	**97.6%**	**2.6%**
Protein (Total)	Enter	**0.010**	**0.585 (0.477-0.693)**	**65.3%**	**95.3%**	**0%**
Sex Hormone Binding Globulin	Enter	0.901	0.519 (0.413-0.624)	68.5%	100%	0%

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Testosterone	Enter	**<.001**	**0.712 (0.614-0.811)**	**66.1%**	**70.6%**	**56.4%**
Triglycerides	Enter	**0.037**	**0.617 (0.513-0.721)**	**66.9%**	**94.1%**	**7.7%**
Urate	Enter	**<.001**	**0.663 (0.553-0.773)**	**71.8%**	**92.9%**	**25.6%**
Urea	Enter	**0.042**	**0.613 (0.495-0.731)**	**67.7%**	**98.8%**	**0.0%**
Vitamin D	Enter	0.059	0.612 (0.498-0.726)	66.9%	97.6%	0%
Immunology						
C-Reactive Protein	Enter	0.850	0.561 (0.452-0.670)	67.7%	98.8%	0%
Red Blood Cell Count	Enter	0.784	0.515 (0.410-0.621)	69.1%	100%	0%
White Blood Cell Count	Enter	0.117	0.589 (0.488-0.690)	68.3%	98.8%	0%
Neutrophils	Enter	**0.016**	**0.636 (0.534-0.738)**	**67.5%**	**97.6%**	**0%**
Lymphocytes	Enter	0.434	0.544 (0.441-0.648)	69.1%	100%	0%
Monocytes	Enter	0.324	0.556 (0.449-0.663)	69.1%	100%	0%
Eosinophils	Enter	0.072	0.604 (0.500-0.707)	70.0%	100%	0%
Basophils	Enter	0.405	0.547 (0.442-0.652)	69.1%	100%	0%

Area Under the Curve (AUC); Confidence Interval (CI). Sensitivity and specificity were the likelihood of correctly detecting when COVID-19 infection for a test case was present vs. not present. “Gray” and “white” shading are used to better visualize predictors within a set of similar variables. P values less than .05 were considered significant and applicable predictors and statistics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 3. Isolated effect of each 2006-2010 antibody titer on predicting COVID-19 infection risk

Pathogen Name	Abbreviation	Antigen	Classifier Method	P value	AUC (95% CI)		Accuracy	Specificity	Sensitivity
Herpes Simplex Virus-1	HSV-1	1gG	Enter	0.426	0.529	(0.410-0.647)	68.5%	100%	0%
Herpes Simplex Virus-2	HSV-2	2mgG	Enter	0.254	0.518	(0.402-0.634)	67.7%	98.8%	0%
Varicella Zoster Virus	VZV	**gE/gI**	**Enter**	**0.040**	**0.611**	**(0.509-0.714)**	**68.5%**	**100%**	**0%**
Epstein-Barr Virus	EBV	**VCA p18**	**Enter**	**0.005**	**0.670**	**(0.569-0.771)**	**67.7%**	**92.9%**	**12.8%**
		EBNA-1	Enter	0.322	0.551	(0.442-0.659)	68.5%	100%	0%
		ZEBRA	**Enter**	**0.047**	**0.589**	**(0.491-0.688)**	**68.5%**	**100%**	**0%**
		EA-D	Enter	0.385	0.503	(0.398-0.609)	68.5%	100%	0%
Human Cytomegalovirus	CMV	**pp150 Nter**	**Enter**	**<.001**	**0.654**	**(0.539-0.769)**	**71.8%**	**89.4%**	**33.3%**
		pp 52	Enter	0.173	0.553	(0.440-0.666)	68.5%	100%	0%
		pp 28	**Enter**	**0.028**	**0.601**	**(0.485-0.717)**	**68.5%**	**94.1%**	**12.8%**
Human Herpesvirus-6	HHV-6	IE1A	Enter	0.133	0.583	(0.473-0.694)	68.5%	100%	0%
		IE1B	**Enter**	**0.018**	**0.637**	**(0.537-0.737)**	**68.5%**	**100%**	**0%**
		p101 k	Enter	0.061	0.532	(0.431-0.633)	68.5%	100%	0%
Human Herpesvirus-7	HHV-7	**U14**	**Enter**	**0.029**	**0.682**	**(0.573-0.790)**	**68.5%**	**100%**	**0%**
Kaposi's Sarcoma Associated Herpesvirus	KSHV	**LANA**	**Enter**	**0.048**	**0.546**	**(0.439-0.653)**	**70.2%**	**100%**	**5.1%**
		K8.1	Enter	0.130	0.564	(0.453-0.676)	70.2%	100%	5.1%
Hepatitus B Virus	HBV	**HBc**	**Enter**	**0.050**	**0.505**	**(0.399-0.611)**	**70.2%**	**100%**	**5.1%**
		HBe	Enter	0.682	0.600	(0.486-0.713)	68.5%	100%	0%
Hepatitus C Virus	HCV	Core	Enter	0.925	0.524	(0.415-0.633)	68.5%	100%	0%
		NS3	**Enter**	**0.008**	**0.663**	**(0.559-0.766)**	**68.5%**	**100%**	**0%**
Toxoplasma gondii	T. gondii	p22	Enter	0.684	0.617	(0.517-0.718)	68.5%	100%	0%
		sag1	Enter	0.082	0.663	(0.556-0.770)	68.5%	100%	0%
Human T Lymphotropic Virus 1	HTLV-1	**HTLV-1 gag**	**Enter**	**<.001**	**0.707**	**(0.615-0.799)**	**70.2%**	**88.2%**	**30.8%**
		HTLV-1 env	Enter	0.197	0.554	(0.451-0.658)	68.5%	100%	0%
Human Immunodeficiency Virus	HIV	**HIV-1 gag**	**Enter**	**0.009**	**0.688**	**(0.591-0.785)**	**68.5%**	**100%**	**0%**
		HIV-1 env	Enter	0.559	0.576	(0.469-0.684)	68.5%	100%	0%
Human Polyomavirus BKV	BKV	**BK VP1**	**Enter**	**0.006**	**0.648**	**(0.538-0.758)**	**68.5%**	**100%**	**0%**
Human Polyomavirus JCV	JCV	JC VP1	Enter	0.618	0.529	(0.418-0.641)	68.5%	100%	0%
Merkel Cell Polyomavirus	MCV	MC VP1	Enter	0.924	0.544	(0.441-0.647)	68.5%	100%	0%
Human Papillomavirus type-16	HPV 16	L1	Enter	0.844	0.525	(0.412-0.637)	68.5%	100%	0%
		E6	Enter	0.093	0.622	(0.517-0.728)	68.5%	100%	0%
		E7	Enter	0.188	0.646	(0.536-0.756)	66.9%	97.6%	0%
Human Papillomavirus type-18	HPV 18	L1	Enter	0.140	0.556	(0.453-0.660)	68.5%	100%	0%
Chlamydia trachomatis	C. trachomatis	momp D	Enter	0.228	0.501	(0.390-0.612)	67.7%	98.8%	0%
		momp A	Enter	0.165	0.551	(0.450-0.653)	68.5%	100%	0%
		tarp-D F1	Enter	0.099	0.549	(0.431-0.668)	69.4%	98.8%	5.1%
		tarp-D F2	Enter	0.573	0.572	(0.459-0.686)	67.7%	98.8%	0%
		PorB	Enter	0.733	0.597	(0.488-0.706)	68.5%	100%	0%
		pGP3	Enter	0.171	0.656	(0.547-0.765)	68.5%	100%	0%
Helicobacter pylori	H. pylori	CagA*	N/A	N/A	N/A	N/A	N/A	N/A	N/A
		VacA	**Enter**	**<.001**	**0.613**	**(0.506-0.719)**	**71.8%**	**97.6%**	**15.4%**
		OMP	Enter	0.258	0.510	(0.397-0.622)	66.9%	97.6%	0%
		GroEL	Enter	0.275	0.591	(0.467-0.715)	66.9%	97.6%	0%
		Catalase	Enter	0.914	0.525	(0.416-0.634)	68.5%	100%	0%
		UreA	Enter	0.891	0.567	(0.672-0.461)	68.5%	100%	0%

Area Under the Curve (AUC); Confidence Interval (CI). Sensitivity and specificity were the likelihood of correctly detecting if COVID-19 infection for a test case was present vs. not present. "Gray' and "white" shading are used to better visualize antigens specific to a given pathogen. *The CagA antigen was excluded from analysis due to roughly half of sample analyte values being lost to lab error. P values less than .05 were considered significant and applicable antigens and statistics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 4. Predictors that loaded into the stepwise models for COVID-19 infection risk

	Stepwise Predictor	Wilks' λ	Coefficient	Seroprevalence
All Test Cases	Apolipoprotein B	0.981	-0.171	
	Urea	0.981	0.193	
	Leukocyte Count	0.981	-0.264	
	Monocyte Count	0.982	0.380	
	Sex Hormone Binding Globulin	0.982	0.287	
	Smoking Status	0.982	0.275	
	Townsend Deprivation Index	0.982	-0.339	
	Ethnic Background	0.983	-0.415	
	HDL Cholesterol	0.984	0.458	
Serology Sub-Group	pp 52 antigen for Human Cytomegalovirus	0.332	0.507	0.582
	Gamma Glutamyltransferase	0.334	0.268	
	Erythrocyte Count	0.334	-0.339	
	PorB Antigen for Chlamydia trachomatis	0.336	-0.404	0.214
	Cholesterol	0.339	-0.353	
	Triglycerides	0.341	-0.369	
	pp 28 Antigen for Human Cytomegalovirus	0.345	-0.754	0.582
	IE1A Antigen for Human Herpesvirus_6	0.356	-0.490	0.908
	Monocyte Count	0.382	-0.651	
	Age in Years	0.384	0.656	
	pGP3 Antigen for Chlamydia trachomatis	0.408	0.731	0.214
	Neutrophil Count	0.420	0.782	
	NS3 Antigen for Hepatitis C Virus	0.423	0.804	0.003
	Urate	0.469	-0.961	
	Testosterone	0.608	1.441	

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 5. Isolated effect of each non-serology predictor on COVID-19 severity among all test cases

Predictor	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
Basic Demographics						
Age	Enter	**<.001**	**0.572 (0.548-0.596)**	**58.1%**	**36.7%**	**75.5%**
Sex	Enter	**0.009**	**0.528 (0.504-0.552)**	**54.9%**	**0%**	**100%**
Ethnic Background	Enter	**<.001**	**0.524 (0.500-0.548)**	**55.8%**	**12.9%**	**91.0%**
Deprivation Index	Enter	0.610	0.507 (0.483-0.531)	55.8%	12.9%	91.0%
Education	Enter	0.854	0.505 (0.478-0.532)	54.9%	0%	100%
Body Composition						
Waist Circumference	Enter	**0.003**	**0.541 (0.517-0.565)**	**54.1%**	**6.0%**	**93.6%**
Body Mass Index	Enter	0.198	0.522 (0.498-0.547)	54.8%	0%	100%
Trunk Fat Mass	Enter	0.068	0.531 (0.506-0.555)	55.0%	0.1%	99.9%
Whole Body Fat Mass	Enter	0.341	0.521 (0.497-0.546)	54.8%	0%	100%
Whole Body Fat-Free Mass	Enter	0.104	0.522 (0.498-0.547)	54.7%	0%	100%
Whole Body Water Mass	Enter	0.095	0.522 (0.497-0.546)	54.7%	0%	100%
Health Behaviors and Conditions						
Smoking Status	Enter	0.114	0.522 (0.497-0.546)	54.7%	0%	100%
Alcohol Status	Enter	0.540	0.506 (0.482-0.530)	54.8%	0%	100%
Long-Term Medical Condition	Enter	0.084	0.519 (0.494-0.543)	54.9%	0%	100%
Health Rating	Enter	0.098	0.518 (0.494-0.543)	54.9%	0%	100%
Vitals						
Pulse Rate	Enter	0.652	0.510 (0.486-0.535)	54.9%	0%	100%
Diastolic BP	Enter	0.969	0.501 (0.476-0.526)	54.9%	0%	100%
Systolic BP	Enter	**0.008**	**0.540 (0.515-0.565)**	**54.7%**	**2.0%**	**98.0%**
Biochemistry						
Alanine Aminotransferase	Enter	**0.039**	**0.540 (0.515-0.565)**	**54.8%**	**0%**	**100%**
Albumin	Enter	0.093	0.525 (0.498-0.552)	54.2%	0.4%	98.9%
Alkaline Phosphatase	Enter	0.198	0.522 (0.496-0.547)	54.8%	0%	100%
Apolipoprotein A	Enter	0.245	0.513 (0.487-0.540)	54.5%	0.1%	99.7%
Apolipoprotein B	Enter	0.862	0.502 (0.477-0.528)	54.8%	0%	100%
Aspartate Aminotransferase	Enter	**0.047**	**0.534 (0.508-0.559)**	**54.8%**	**0%**	**100%**
Bilirubin (Total)	Enter	0.965	0.509 (0.483-0.534)	54.8%	0%	100%
Calcium	Enter	0.545	0.507 (0.481-0.534)	54.5%	0%	99.8%
Cholesterol (Total)	Enter	0.352	0.510 (0.484-0.535)	54.7%	0%	100%
Creatinine	Enter	0.289	0.510 (0.485-0.535)	54.9%	0%	100%
Cystatin C	Enter	**0.006**	**0.527 (0.502-0.553)**	**54.7%**	**0.7%**	**99.5%**
Gamma Glutamyltransferase	Enter	0.181	0.529 (0.504-0.555)	54.8%	0%	100%
HDL Cholesterol	Enter	0.180	0.517 (0.490-0.544)	54.1%	0.4%	98.7%
Hemoglobin A1c	Enter	**0.002**	**0.555 (0.529-0.580)**	**54.6%**	**0.7%**	**99.3%**
Insulin-Like Growth Factor 1	Enter	**0.037**	**0.524 (0.499-0.550)**	**54.7%**	**3.8%**	**96.8%**
LDL Cholesterol	Enter	0.470	0.508 (0.483-0.533)	54.8%	0%	100%
Lipoprotein A	Enter	0.216	0.518 (0.489-0.546)	54.5%	0%	100%
Phosphate	Enter	0.357	0.513 (0.486-0.540)	54.6%	0%	99.9%
Protein (Total)	Enter	0.930	0.512 (0.486-0.539)	54.6%	0%	100%
Sex Hormone Binding Globulin	Enter	**0.036**	**0.525 (0.498-0.552)**	**54.8%**	**7.5%**	**94.2%**

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Testosterone	Enter	0.060	0.521 (0.495-0.548)	54.5%	0%	100%
Triglycerides	Enter	0.060	0.528 (0.503-0.554)	54.8%	0%	100%
Urate	Enter	**0.012**	**0.533 (0.508-0.559)**	**54.8%**	**1.7%**	**98.5%**
Urea	Enter	**0.003**	**0.530 (0.505-0.556)**	**54.9%**	**2.7%**	**97.8%**
Vitamin D	Enter	0.562	0.503 (0.477-0.529)	54.8%	0%	100%
Immunology						
Red Blood Cell Count	Enter	0.732	0.504 (0.479-0.529)	54.9%	0%	100%
White Blood Cell Count	Enter	**0.025**	**0.536 (0.511-0.561)**	**54.9%**	**0%**	**100%**
C-Reactive Protein	Enter	0.598	0.528 (0.503-0.554)	54.8%	0%	100%
Neutrophils	Enter	**0.004**	**0.535 (0.510-0.560)**	**55.6%**	**3.6%**	**98.2%**
Lymphocytes	Enter	0.212	0.504 (0.479-0.530)	55.0%	0%	100%
Monocytes	Enter	0.071	0.530 (0.505-0.555)	55.0%	0%	100%
Eosinophils	Enter	0.291	0.520 (0.495-0.545)	55.0%	0%	100%
Basophils	Enter	0.671	0.500 (0.475-0.525)	55.0%	0%	100%

Area Under the Curve (AUC); Confidence Interval (CI). Here, sensitivity and specificity are the likelihood of correctly detecting if a positive COVID-19 test case was severe or mild. "Orange" and "white" shading is used to better visualize each class of predictors for COVID-19 severity. P values less than .05 for classification accuracy were considered significant, where applicable predictors and classifier statistics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 6. Isolated effect of each non-serology predictor on COVID-19 severity for the serology sub-group

Predictor	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
Basic Demographics						
Age	Enter	0.889	0.532 (0.348-0.716)	43.6%	0%	81.0%
Sex	Enter	0.455	0.556 (0.373-0.738)	17.9%	0%	33.3%
Ethnic Background	Enter	0.889	0.520 (0.331-0.708)	52.6%	0%	95.2%
Deprivation Index	Enter	0.973	0.520 (0.331-0.708)	46.2%	0%	85.7%
Education	Enter	0.706	0.504 (0.296-0.712)	54.8%	0%	100%
Body Composition						
Waist Circumference	Enter	0.308	0.612 (0.430-0.795)	59.0%	27.8%	85.7%
Body Mass Index	Enter	0.363	0.517 (0.328-0.706)	53.8%	22.2%	81.0%
Trunk Fat Mass	Enter	0.087	0.615 (0.423-0.806)	62.2%	47.1%	75.0%
Whole Body Fat Mass	Enter	0.100	0.629 (0.437-0.822)	64.9%	35.3%	90.0%
Whole Body Fat-Free Mass	Enter	0.763	0.515 (0.325-0.704)	54.1%	0%	100%
Whole Body Water Mass	Enter	0.851	0.562 (0.372-0.752)	54.1%	0%	100%
Health Behaviors and Conditions						
Smoking Status	Enter	0.798	0.516 (0.331-0.701)	38.5%	0%	71.4%
Alcohol Status	Enter	0.845	0.505 (0.321-0.689)	53.8%	0%	100%
Long-Term Medical Condition	Enter	0.802	0.521 (0.334-0.708)	55.3%	0%	100%
Health Rating	Enter	0.999	0.501 (0.317-0.686)	55.3%	0%	100%
Vitals						
Pulse Rate	Enter	0.128	0.593 (0.412-0.773)	53.8%	33.3%	71.4%
Diastolic BP	Enter	0.984	0.520 (0.334-0.705)	48.7%	0%	90.5%
Systolic BP	Enter	0.868	0.513 (0.324-0.702)	46.2%	0%	85.7%
Biochemistry						
Alanine Aminotransferase	Enter	**0.043**	**0.690 (0.511-0.870)**	**61.5%**	**77.8%**	**47.6%**
Albumin	Enter	0.483	0.579 (0.332-0.827)	52.2%	0%	85.7%
Alkaline Phosphatase	Enter	0.311	0.538 (0.350-0.727)	53.8%	33.3%	71.4%
Apolipoprotein A	Enter	0.892	0.587 (0.351-0.824)	56.5%	0%	92.9%
Apolipoprotein B	Enter	0.587	0.542 (0.358-0.727)	46.2%	0%	85.7%
Aspartate Aminotransferase	Enter	0.688	0.642 (0.463-0.820)	51.3%	0%	95.2%
Bilirubin (Total)	Enter	0.482	0.586 (0.404-0.768)	41.0%	22.2%	57.1%
Calcium	Enter	0.190	0.635 (0.406-0.864)	60.9%	22.0%	85.7%
Cholesterol (Total)	Enter	0.945	0.517 (0.333-0.701)	43.6%	0%	81.0%
Creatinine	Enter	0.096	0.649 (0.474-0.825)	66.7%	55.6%	76.2%
Cystatin C	Enter	0.497	0.545 (0.355-0.735)	56.4%	27.8%	81.0%
Gamma Glutamyltransferase	Enter	0.992	0.577 (0.393-0.760)	53.8%	0%	100%
HDL Cholesterol	Enter	0.841	0.540 (0.286-0.794)	52.2%	0%	85.7%
Hemoglobin A1c	Enter	0.506	0.560 (0.368-0.752)	48.6%	29.4%	65.0%
Insulin-Like Growth Factor 1	Enter	0.786	0.544 (0.354-0.734)	51.3%	0%	95.2%
LDL Cholesterol	Enter	0.808	0.546 (0.362-0.731)	46.2%	0%	85.7%
Lipoprotein A	Enter	0.287	0.607 (0.385-0.829)	59.3%	71.4%	46.2%
Phosphate	Enter	0.627	0.524 (0.278-0.770)	60.9%	0%	100%
Protein (Total)	Enter	0.513	0.571 (0.335-0.808)	60.9%	0%	100%
Sex Hormone Binding Globulin	Enter	0.578	0.587 (0.347-0.828)	60.9%	0%	100%
Testosterone	Enter	0.723	0.634 (0.446-0.821)	53.8%	0%	100%

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Triglycerides	Enter	0.989	0.540 (0.351-0.728)	48.7%	0%	90.5%
Urate	Enter	0.372	0.597 (0.413-0.779)	53.8%	33.3%	71.4%
Urea	Enter	0.300	0.604 (0.419-0.790)	61.5%	44.4%	76.2%
Vitamin D	Enter	0.877	0.500 (0.315-0.685)	53.8%	0%	100%
Immunology						
Red Blood Cell Count	Enter	0.970	0.553 (0.361-0.746)	52.6%	0%	95.2%
White Blood Cell Count	Enter	0.177	0.646 (0.455-0.836)	50.0%	35.3%	61.9%
C-Reactive Protein	Enter	0.234	0.522 (0.322-0.723)	64.1%	22.2%	100%
Neutrophils	Enter	**0.049**	**0.653 (0.475-0.830)**	**57.9%**	**52.9%**	**61.9%**
Lymphocytes	Enter	0.581	0.548 (0.358-0.737)	52.6%	0%	95.2%
Monocytes	Enter	0.822	0.534 (0.343-0.724)	50.0%	0%	90.5%
Eosinophils	Enter	0.694	0.516 (0.323-0.709)	55.6%	0%	100%
Basophils	Enter	0.153	0.604 (0.423-0.785)	57.9%	35.3%	76.2%

Area Under the Curve (AUC); Confidence Interval (CI). Here, sensitivity and specificity are the likelihood of correctly detecting if a positive COVID-19 test case was severe or mild. "Orange" and "white" shading is used to better visualize each class of predictors for COVID-19 severity. P values less than .05 were considered significant, where applicable predictors and classifier statistics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 7. Isolated effect of each 2006-2010 antibody titer on predicting COVID-19 severity

Pathogen Name	Abbreviation	Antigen	Classifier Method	P value	AUC (95% CI)	Accuracy	Specificity	Sensitivity
Herpes Simplex Virus-1	HSV-1	1gG	Enter	0.185	0.626 (0.447-0.804)	59.0%	61.1%	57.1%
Herpes Simplex Virus-2	HSV-2	2mgG	Enter	0.625	0.511 (0.321-0.701)	48.7%	0%	90.5%
Varicella Zoster Virus	VZV	gE/gI	Enter	0.220	0.594 (0.412-0.776)	51.3%	38.9%	61.9%
Epstein-Barr Virus	EBV	VCA p18	Enter	0.686	0.565 (0.381-0.748)	43.6%	0%	81.0%
		EBNA-1	Enter	0.087	0.634 (0.452-0.815)	59.0%	33.3%	81.0%
		ZEBRA	Enter	0.221	0.604 (0.421-0.788)	56.4%	38.9%	71.4%
		EA-D	Enter	0.285	0.599 (0.418-0.780)	53.8%	44.4%	61.9%
Human Cytomegalovirus	CMV	pp150 Nter	Enter	0.465	0.585 (0.399-0.771)	59.0%	44.4%	71.4%
		pp 52	Enter	0.649	0.512 (0.322-0.702)	43.6%	0%	81.0%
		pp 28	Enter	0.763	0.544 (0.355-0.733)	48.7%	0%	90.5%
Human Herpesvirus-6	HHV-6	IE1A	Enter	0.592	0.538 (0.354-0.723)	53.8%	0%	85.7%
		IE1B	Enter	0.700	0.565 (0.375-0.755)	51.3%	0%	95.2%
		p101 k	Enter	0.667	0.507 (0.319-0.694)	48.7%	0%	90.5%
Human Herpesvirus-7	HHV-7	U14	Enter	**0.016**	**0.729 (0.568-0.890)**	**64.1%**	**44.4%**	**81.0%**
Kaposi's Sarcoma Associated	KSHV	LANA	Enter	1.000	0.616 (0.437-0.796)	51.3%	0%	95.2%
		K8.1	Enter	0.785	0.560 (0.371-0.748)	51.3%	0%	95.2%
Hepatitus B Virus	HBV	HBc	Enter	0.850	0.587 (0.402-0.773)	51.3%	0%	95.2%
		HBe	Enter	0.736	0.583 (0.350-0.727)	51.3%	0%	95.2%
Hepatitus C Virus	HCV	Core	Enter	0.314	0.503 (0.316-0.689)	59.0%	11.1%	100%
		NS3	Enter	0.847	0.578 (0.395-0.762)	53.8%	0%	100%
Toxoplasma gondii	T. gondii	p22	Enter	0.259	0.549 (0.357-0.741)	56.4%	5.6%	100%
		sag1	Enter	0.229	0.565 (0.379-0.751)	61.5%	27.8%	90.5%
Human T Lymphotropic Virus 1	HTLV-1	HTLV-1 gag	Enter	0.065	0.647 (0.469-0.825)	66.7%	66.7%	66.7%
		HTLV-1 env	Enter	0.570	0.595 (0.414-0.776)	51.3%	11.1%	81.0%
Human Immunodeficiency Virus	HIV	HIV-1 gag	Enter	0.364	0.538 (0.353-0.724)	53.8%	16.7%	85.7%
		HIV-1 env	Enter	0.634	0.534 (0.349-0.720)	48.7%	0%	90.5%
Human Polyomavirus BKV	BKV	BK VP1	Enter	0.782	0.562 (0.380-0.744)	51.3%	0%	81.0%
Human Polyomavirus JCV	JCV	JC VP1	Enter	**0.045**	**0.671 (0.502-0.840)**	**59.0%**	**66.7%**	**52.4%**
Merkel Cell Polyomavirus	MCV	MC VP1	Enter	0.294	0.628 (0.448-0.809)	59.0%	55.6%	61.9%
Human Papillomavirus type-16	HPV 16	L1	Enter	0.554	0.525 (0.338-0.712)	48.7%	11.1%	81.0%
		E6	Enter	0.740	0.538 (0.349-0.728)	51.3%	0%	95.2%
		E7	Enter	0.134	0.565 (0.382-0.748)	61.5%	72.2%	52.4%
Human Papillomavirus type-18	HPV 18	L1	Enter	0.828	0.511 (0.322-0.699)	46.2%	0%	85.7%
Chlamydia trachomatis	C. trachomatis	momp D	Enter	0.818	0.511 (0.322-0.699)	51.3%	0%	95.2%
		momp A	Enter	0.819	0.505 (0.315-0.695)	51.3%	0%	95.2%
		tarp-D F1	Enter	0.809	0.585 (0.403-0.766)	53.8%	0%	95.2%
		tarp-D F2	Enter	0.615	0.538 (0.343-0.734)	48.7%	0%	90.5%
		PorB	Enter	0.832	0.504 (0.319-0.689)	51.3%	0%	95.2%
		pGP3	Enter	0.464	0.603 (0.422-0.784)	56.4%	11.1%	95.2%
Helicobacter pylori	H. pylori	CagA*	N/A	N/A	NA	N/A	N/A	N/A
		VacA	Enter	0.915	0.602 (0.420-0.784)	46.2%	0%	85.7%
		OMP	Enter	0.340	0.558 (0.375-0.741)	25.6%	0%	47.6%
		GroEL	Enter	0.415	0.614 (0.433-0.795)	56.4%	22.2%	85.7%
		Catalase	Enter	0.335	0.642 (0.464-0.819)	59.0%	16.7%	95.2%
		UreA	Enter	0.300	0.606 (0.425-0.786)	53.8%	0%	100%

Area Under the Curve (AUC); Confidence Interval (CI). Here, sensitivity and specificity are the likelihood of correctly detecting if a positive COVID-19 test case was severe or mild. *The CagA antigen was excluded from analysis due to roughly half of sample analyte values being lost to lab error. "Orange" and "white" shading is used to better visualize

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

each set of antigens for a specific pathogen. P values less than .05 were considered significant, where applicable antigens and classifier metrics are bolded.

medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20127092; this version posted June 12, 2020. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Table 8. Predictors that loaded into the stepwise models for COVID-19 severity risk

	Stepwise Predictor	Wilks' λ	Coefficient	Seroprevalence
All Test Cases	Alanine Aminotransferase	0.979	0.298	
	Age in Years	0.994	0.873	
	Monocyte Count	0.980	0.351	
Serology Sub-Group	HTLV-1 gag for Human T Lymphotropic Virus 1	0.896	0.926	1.6%
	JC VP1 antigen for Human Polyomavirus JCV	0.911	0.959	57.5%