scientific reports

Check for updates

OPEN

Circulating proteins to predict COVID-19 severity

Chen-Yang Su[1,2,3,24], Sirui Zhou[1,4,24], Edgar Gonzalez-Kozlova[5,24], Guillaume Butler-Laporte[1,4], Elsa Brunet-Ratnasingham[6], Tomoko Nakanishi[1,7,8,9], Wonseok Jeon[2], David R. Morrison[1], Laetitia Laurent[1], Jonathan Afilalo[1,4], Marc Afilalo[10], Danielle Henry[1], Yiheng Chen[1,7], Julia Carrasco-Zanini[11], Yossi Farjoun[1], Maik Pietzner[11,18], Nofar Kimchi[1], Zaman Afrasiabi[1], Nardin Rezk[1], Meriem Bouab[1], Louis Petitjean[1], Charlotte Guzman[1], Xiaoqing Xue[1], Chris Tselios[1], Branka Vulesevic[1], Olumide Adeleye[1], Tala Abdullah[1], Noor Almamlouk[1], Yara Moussa[1], Chantal DeLuca[1], Naomi Duggan[1], Erwin Schurr[12], Nathalie Brassard[6], Madeleine Durand[6], Diane Marie Del Valle[13], Ryan Thompson[14], Mario A. Cedillo[15], Eric Schadt[14], Kai Nie[16], Nicole W. Simons[14], Konstantinos Mouskas[14], Nicolas Zaki[16], Manishkumar Patel[13], Hui Xie[16], Jocelyn Harris[16], Robert Marvin[16], Esther Cheng[14], Kevin Tuballes[13], Kimberly Argueta[16], Ieisha Scott[16], The Mount Sinai COVID-19 Biobank Team*, Celia M. T. Greenwood[1,4], Clare Paterson[17], Michael A. Hinterberg[17], Claudia Langenberg[11,18], Vincenzo Forgetta[1], Joelle Pineau[2], Vincent Mooser[7], Thomas Marron[19], Noam D. Beckmann[14], Seunghee Kim-schulze[16], Alexander W. Charney[20], Sacha Gnjatic[5,13,16], Daniel E. Kaufmann[6,21,22], Miriam Merad[13] & J. Brent Richards[1,4,7,23]✉

Predicting COVID-19 severity is difficult, and the biological pathways involved are not fully understood. To approach this problem, we measured 4701 circulating human protein abundances in two independent cohorts totaling 986 individuals. We then trained prediction models including protein abundances and clinical risk factors to predict COVID-19 severity in 417 subjects and tested these models in a separate cohort of 569 individuals. For severe COVID-19, a baseline model including age and sex provided an area under the receiver operator curve (AUC) of 65% in the test cohort. Selecting 92 proteins from the 4701 unique protein abundances improved the AUC to 88% in the training cohort, which remained relatively stable in the testing cohort at 86%, suggesting good generalizability. Proteins selected from different COVID-19 severity were enriched for cytokine and cytokine receptors,

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Pavilion H-413, 3755 Côte-Ste-Catherine Montréal, Montreal, QC H3T 1E2, Canada. [2]Department of Computer Science, McGill University, Montréal, QC, Canada. [3]Quantitative Life Sciences Program, McGill University, Montreal, Quebec, Canada. [4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada. [5]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Research Centre of the Centre Hospitalier de L'Université de Montréal, Montreal, QC, Canada. [7]Department of Human Genetics, McGill University, Montreal, QC, Canada. [8]Graduate School of Medicine, McGill International Collaborative School in Genomic Medicine, Kyoto University, Kyoto, Japan. [9]Japan Society for the Promotion of Science, Tokyo, Japan. [10]Department of Emergency Medicine, Jewish General Hospital, McGill University, Montreal, QC, Canada. [11]MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK. [12]Infectious Diseases and Immunity in Global Health Program, Research Institute of the McGill University Health Centre, Montréal, QC, Canada. [13]Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [14]Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [15]Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [16]Human Immune Monitoring Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [17]SomaLogic Operating Co., Inc., Boulder, CO, USA. [18]Computational Medicine, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany. [19]Immunotherapy and Phase 1 Trials, Mount Sinai Hospital, New York, NY, USA. [20]Mount Sinai Clinical Intelligence Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [21]Department of Medicine, Université de Montréal, Montreal, QC, Canada. [22]Division of Infectious Diseases, Department of Medicine, University Hospital of Lausanne and University of Lausanne, Lausanne, Switzerland. [23]Department of Twin Research, King's College London, London, UK. [24]These authors contributed equally: Chen-Yang Su, Sirui Zhou and Edgar Gonzalez-Kozlova. *A list of authors and their affiliations appears at the end of the paper. ✉email: brent.richards@mcgill.ca

but more than half of the enriched pathways were not immune-related. Taken together, these findings suggest that circulating proteins measured at early stages of disease progression are reasonably accurate predictors of COVID-19 severity. Further research is needed to understand how to incorporate protein measurement into clinical care.

A remarkable feature of COVID-19 disease is its highly variable clinical course, where some individuals manifest severe disease or death, and others remain asymptomatic. Several clinical and genetic risk factors explain a proportion of these outcomes[1–6], yet most of the host biological causes of these adverse COVID-19 outcomes remain unknown.

Recent reports have identified some of the biologic pathways influencing risk of adverse COVID-19, such as immune responses[7–11], interferon pathways[12–14], and T-cell dysfunction[15,16]. However, many such studies have focused on narrow sets of pre-selected cytokines. One way to rapidly assess thousands of potential biomarkers associated with the severity of COVID-19 is through the measurement of blood circulating proteins. Such circulating proteins may be useful because they can help to identify pathways influencing severity of disease. They may also identify individuals at high risk of a severe COVID-19 clinical course. Similarly, circulating proteomic biomarkers have recently been shown to serve as predictors of other common diseases[17–24] including cardiovascular disease. They are also relevant in drug discovery because they are generally more accessible to pharmacological manipulation than intracellular proteins[25–29]. Thus, understanding the circulating proteins associated with COVID-19 severity may be helpful to address major challenges raised by the current pandemic[15,30–41].

We undertook a large-scale study to assess the relationship of thousands of circulating proteins with COVID-19 severity. To do so, we used machine learning methods to develop a predictive model of COVID-19 severity using the circulating blood protein abundances as predictors. Proteins were measured using 4984 nucleic acid aptamers (SOMAmer® reagents)[42] targeting 4701 unique circulating human proteins in two cohorts collected from two countries, which in total included 986 individuals. The training cohort was comprised of 417 individuals from two sites of the Biobanque Québécoise de la COVID-19 (BQC19 cohort). This cohort was used to train a model to predict COVID-19 severity. This model was then tested in a separate test cohort from the Mount Sinai Hospital in New York City, which was similarly characterized for the same protein measurements and COVID-19 severity.

This large-scale study across two countries and two geographically separated cohorts identified circulating proteins associated with COVID-19 severity measured at a large-scale in well-characterized cohorts. These findings provide insights into the biological pathways influencing these outcomes and the ability of proteomics to predict these outcomes.

Results

Cohorts. To establish a proteomic-based prediction model for COVID-19 severity, we used the BQC19 cohort, which consisted of samples from two hospitals in Montreal, with proteomic measurements for training and cross-validation. The final model was tested in an independent cohort from Mount Sinai hospital in New York City. Using the same SomaScan® assay, 4984 SOMAmer reagents measured the levels of 4701 different circulating proteins in both the BQC19 and Mount Sinai cohorts. To train our models, we selected 417 individuals which included 313 nasal swab SARS-CoV-2 PCR positive patients with baseline samples collected within 14 days of symptom onset (mean and median time since symptom onset in COVID-19 patients = 7.0 days (SD = 3.96 days)). The BQC19 cohort also included an additional 104 individuals who presented to the same hospital sites with symptoms consistent with COVID-19 but had a negative SARS-CoV-2 PCR nasal swab. The Mount Sinai cohort consisted of 569 individuals with their earliest samples also collected within 14 days of symptom onset. Among them were 472 SARS-CoV-2 positive patients again confirmed by PCR, one patient confirmed by chest CT, and 96 SARS-CoV-2 negative individuals (89 with PCR confirmation). If multiple blood samples were collected from the same person, we used the samples collected at the time point closest to symptom onset. We chose to use samples close to symptom onset to reflect the proteome of acute COVID-19, rather than its recovery phase.

The demographic and clinical characteristics of the participants in the training and testing datasets are shown in Table 1. In the BQC19 cohort, the mean age across all samples was 65.3 years (SD = 18.4 years), and 52% of the cohort were men. In the Mount Sinai cohort, the mean age was 59.6 years (SD: 19.4 years), and 58.2% of the cohort were men.

For the definition of COVID-19 severity, we focused on two levels: (1) severe COVID-19 was defined as individuals who died or required any form of oxygen supplementation; and (2) critical COVID-19, defined as individuals who died or experienced severe respiratory failure (requiring non-invasive ventilation, high flow oxygen therapy, intubation, or extracorporeal membrane oxygenation). Detailed definitions of these adverse outcomes are described in Methods. The overall study design is shown in Fig. 1, which outlines the training and testing stages of the study. Consistent with recent successful large-scale genetic studies, we defined controls as all participants not meeting case criteria[1].

A visualization of the distribution of individuals across COVID-19 severity outcomes in the BQC19 and Mount Sinai cohorts is provided in Supplementary Fig. 1. In the BQC19 training cohort 175 individuals were classified as severe cases and 242 individuals were controls. The controls for severe COVID-19 were comprised of 138 SARS-CoV-2 positive individuals not meeting case definition and 104 SARS-CoV-2 negative individuals. In the case of critical disease, 93 individuals out of 313 COVID-19 positive patients were classified as critical cases and 324 individuals were controls. The controls for critical COVID-19 cases were 220 SARS-CoV-2 positive individuals not meeting case definition and 104 participants who were SARS-CoV-2 negative. In the Mount Sinai testing cohort, 392 individuals were classified as severe cases and 177 individuals were controls while for

Dataset	Training cohort (BQC19)				
		Severe		Critical	
COVID-19 severity	All samples (*n* = 417)	Cases (*n* = 175)	Controls (*n* = 242)	Cases (*n* = 93)	Controls (*n* = 324)
Age in years*	65.3 (18.4)	67.6 (17.6)	63.7 (18.8)	66.7 (16.5)	64.9 (18.9)
Female sex**	200 (48.0)	82 (46.9)	118 (48.8)	36 (38.7)	164 (50.6)
Sample processing time (hours)*	9.7 (13.4)	7.5 (6.5)	11.3 (16.5)	7.0 (6.3)	10.4 (14.8)
Hospital site**					
CHUM	105 (25.2)	42 (24.0)	63 (26.0)	32 (34.4)	73 (22.5)
JGH	312 (74.8)	133 (76.0)	179 (74.0)	61 (65.6)	251 (77.5)
Diabetes**					
No	288 (69.1)	100 (57.1)	188 (77.7)	54 (58.1)	234 (72.2)
Yes	127 (30.5)	75 (42.9)	52 (21.5)	39 (41.9)	88 (27.2)
Chronic obstructive pulmonary disease**					
No	358 (85.9)	145 (82.9)	213 (88.0)	77 (82.8)	281 (86.7)
Yes	57 (13.7)	30 (17.1)	27 (11.2)	16 (17.2)	41 (12.7)
Chronic kidney disease**					
No	363 (87.1)	151 (86.3)	212 (87.6)	77 (82.8)	286 (88.3)
Yes	52 (12.5)	24 (13.7)	28 (11.6)	16 (17.2)	36 (11.1)
Congestive heart failure**					
No	355 (85.1)	152 (86.9)	203 (83.9)	80 (86.0)	275 (84.9)
Yes	60 (14.4)	23 (13.1)	37 (15.3)	13 (14.0)	47 (14.5)
Hypertension**					
No	179 (42.9)	67 (38.3)	112 (46.3)	31 (33.3)	148 (45.7)
Yes	235 (56.4)	107 (61.1)	128 (52.9)	61 (65.6)	174 (53.7)
Liver disease**					
No	403 (96.6)	172 (98.3)	231 (95.5)	91 (97.8)	312 (96.3)
Yes	12 (2.9)	3 (1.7)	9 (3.7)	2 (2.2)	10 (3.1)
Smoking status**					
Current smoker	13 (3.1)	6 (3.4)	7 (2.9)	5 (5.4)	8 (2.5)
Ex-smoker	47 (11.3)	21 (12.0)	26 (10.7)	11 (11.8)	36 (11.1)
Never smoker	233 (55.9)	98 (56.0)	135 (55.8)	40 (43.0)	193 (59.6)
Dataset	Testing cohort (Mount Sinai)				
		Severe		Critical	
COVID-19 severity	All samples (*n* = 569)	Cases (*n* = 392)	Controls (*n* = 177)	Cases (*n* = 233)	Controls (*n* = 336)
Age in years*	59.6 (19.4)	63.0 (17.1)	52.1 (22.1)	63.7 (16.7)	56.8 (20.6)
Female sex**	238 (41.8)	154 (39.3)	84 (47.5)	87 (37.3)	151 (44.9)

Table 1. Demographic characteristics of the participating cohorts. *Mean (SD). **N (%). When the counts of each cell do not sum to the total sample size, this is due to missing data or the patient answering, "I do not know". *Severe—Cases* individuals who tested positive for COVID-19 and died or required any type of respiratory support (including oxygen delivered by nasal prongs). Controls: individuals with COVID-19 but did not meet severe case criteria or individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative. *Critical—Cases* individuals who tested positive for COVID-19 and died or required respiratory support (intubation, continuous positive airway pressure, bilevel positive airway pressure, continuous external negative pressure, or high flow positive end expiratory pressure oxygen). Controls: individuals with COVID-19 but did not meet critical case criteria or individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative. *Sample Processing Time* time in hours between sample collection and sample freezing. *CHUM* Samples recruited from the Centre Hospitalier de l'Université de Montréal. *JGH* Samples recruited from the Jewish General Hospital.

critical disease 233 individuals were cases and 336 were controls. Generally, severe, or critical COVID-19 cases were older than controls in both the training dataset and the testing dataset. Males were also more likely to have severe or critical COVID-19 as compared to females (Table 1). The age and sex distribution of the participants stratified by case/control status for the two COVID-19 severity groups are shown in Supplementary Fig. 2. The distributions suggest that males who develop severe or critical COVID-19 are generally younger than females.

Association of protein abundance with COVID-19 severity. In order to directly assess if any of the measured proteins were associated with COVID-19 severity, we used multivariable logistic regression to test the association of each of the 4984 SOMAmer reagents with the two COVID-19 outcomes while adjusting for age, sex, sample processing time, and hospital site in the BQC19 cohort. These variables were chosen because they

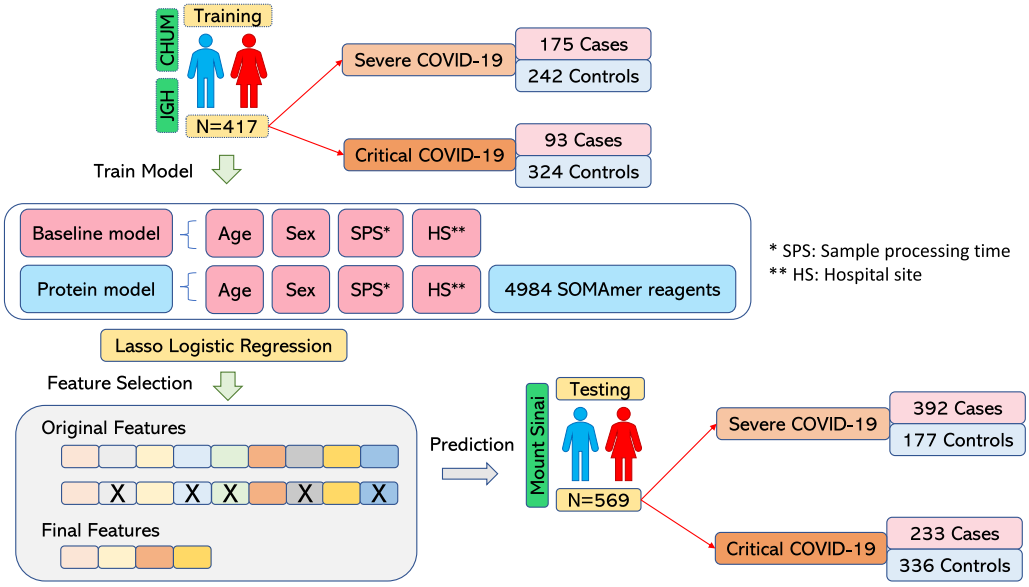


Figure 1. Overall Study design. Schematic of training and testing stages of this study. Severe COVID-19 is defined as death or use of any form of oxygen supplementation. Critical COVID-19 is defined as death or severe respiratory failure (non-invasive ventilation, high flow oxygen therapy, intubation, or extracorporeal membrane oxygenation).

are readily available in the course of clinical care, representing the minimum set of variables to predict severity. Logistic regression identified 1531 SOMAmer reagents to be associated with severe COVID-19 (Supplementary Table 1) and 1592 SOMAmer reagents (Supplementary Table 2) to be associated with critical COVID-19 when using a Benjamini–Hochberg corrected p value of 0.01 (Supplementary Fig. 3).

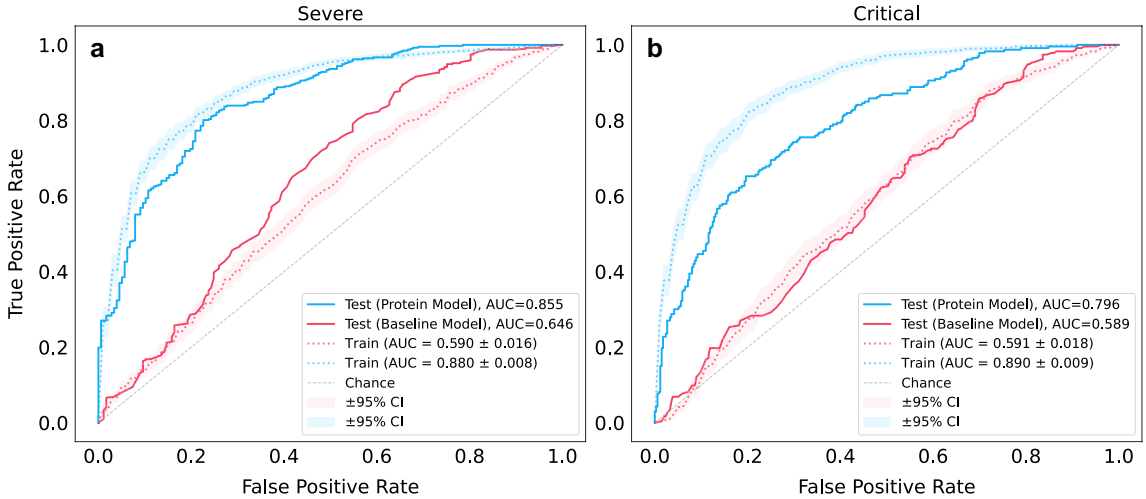
Model selection and performance using LASSO. One reason why many circulating proteins were associated with COVID-19 severity is that most of the protein levels were highly correlated with each other. Therefore, we used L1 regularized multivariable logistic regression models (LASSO)[43] to select uncorrelated proteins that best predicted COVID-19 severity in the BQC19 training cohort. We did so for three reasons: (1) LASSO performs well when the number of features is greater than the number of samples (as was the case in our experiment); (2) LASSO forces many correlated features to have a zero coefficient by randomly selecting one of the features (sometimes more than one) from a group of correlated features thereby preventing collinearity; (3) LASSO mitigates the possibility of overfitting[43].

We first defined a baseline model which included only the four covariates in the logistic regression model: age, sex, sample processing time, and hospital site to predict COVID-19 severity. We then evaluated whether the addition of proteins would aid in identifying which patients developed severe COVID-19 by adding 4984 SOMAmer reagents to the baseline model. This model, which included baseline covariates and protein levels, is termed the "protein model". To train both the baseline and the protein models, we performed 10 repeats of stratified fivefold cross-validation using LASSO logistic regression in the BQC19 cohort on both the severe and critical groups. We tuned the penalty parameter "lambda" across each of the 50 cross-validations and selected the lambda value corresponding to the model with the highest area under the receiver operator characteristic curve (AUC), which was averaged over the 50 cross-validation results. Results from the lambda parameter search are shown in Supplementary Fig. 4A, B.

For the best performing model predicting severe COVID-19, we selected a $\log_{10}$ lambda value of − 1.5 which generated an average training AUC of 59% for the baseline model. We next selected a $\log_{10}$ lambda value of 1.0, which generated an average AUC of 88% for the protein model. For the best performing model predicting critical COVID-19, we selected $\log_{10}$ lambda values of − 2.0 and 1.0 corresponding to average cross-validation training AUC scores of 59% and 89% for the baseline and protein model, respectively (Fig. 2a, b). We then used these chosen lambda hyperparameters to build baseline and protein models for severe and critical COVID-19 using the entire BQC19 cohort and evaluated their performance in the independent external test cohort from Mount Sinai. Elastic net regularized logistic regression models were also trained on the BQC19 cohort but not tested on the Mount Sinai cohort due to negligible differences in training performance (Supplementary Note).

When testing the prediction of severe COVID-19 in the independent Mount Sinai cohort, AUC performance of the baseline model improved from 59% in the BQC19 training cohort to 65% in the Mount Sinai testing cohort. The AUC of the protein model decreased slightly between training and testing (88% vs. 86%).

The AUC of the protein model for predicting critical COVID-19 also decreased from a training score of 89% to 80% in the test set. In contrast, prediction of critical COVID-19 using the baseline model was consistent



c

Mount Sinai Testing Cohort

Predicted Values

True Values

Protein model

Severe — Threshold 0.486

	Controls	Cases
Controls	True Negatives 141 24.8%	False Positives 36 6.3%
Cases	False Negatives 105 18.5%	True Positives 287 50.4%

Sensitivity = 73.2% PPV = 88.9%
Specificity = 79.7% NPV = 57.3%

Critical — Threshold 0.255

	Controls	Cases
Controls	True Negatives 234 41.1%	False Positives 102 17.9%
Cases	False Negatives 60 10.5%	True Positives 173 30.4%

Sensitivity = 74.3% PPV = 62.9%
Specificity = 69.6% NPV = 79.6%

Baseline model

Severe — Threshold 0.417

	Controls	Cases
Controls	True Negatives 107 18.8%	False Positives 70 12.3%
Cases	False Negatives 153 26.9%	True Positives 239 42.0%

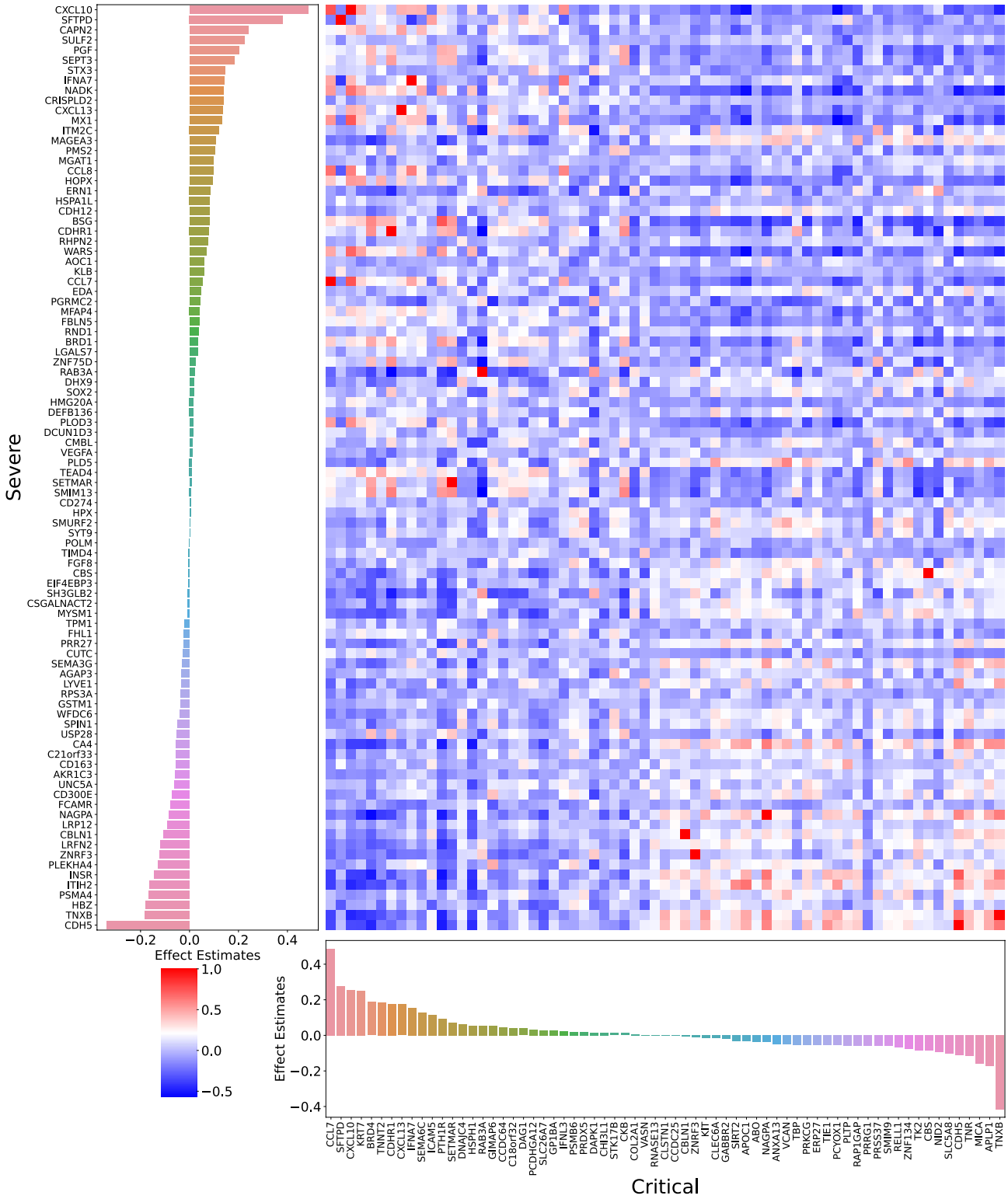
Sensitivity = 61.0% PPV = 77.3%
Specificity = 60.5% NPV = 41.2%

Critical — Threshold 0.202

	Controls	Cases
Controls	True Negatives 193 33.9%	False Positives 143 25.1%
Cases	False Negatives 116 20.4%	True Positives 117 20.6%

Sensitivity = 50.2% PPV = 45.0%
Specificity = 57.4% NPV = 62.5%

Figure 2. AUC score results. (**a**) L1 regularized logistic regression training and testing results for severe COVID-19 and (**b**) critical COVID-19. Blue and red are used to represent the protein model and baseline model, respectively, while solid and dotted lines represent the testing and training performance, respectively. Shaded areas denote the 95% confidence intervals for the training cohort. (**c**) Two-by-two contingency table results from the test set are shown for predicting severe (top left, bottom left) and critical COVID-19 (top right, bottom right) using the protein model (blue) and baseline model (red). The threshold for predicting cases was determined during training using Youden's J statistic which selects a threshold that maximizes the sum of the sensitivity and specificity score. *PPV* positive predictive value, *NPV* negative predictive value.


Severe
CXCL10
SFTPD
CAPN2
SULF2
PGF
SEPT3
STX3
IFNA7
NADK
CRISPLD2
CXCL13
MX1
ITM2C
MAGEA3
PMS2
MGAT1
CCL8
HOPX
ERN1
HSPA1L
CDH12
BSG
CDHR1
RHPN2
WARS
AOC1
KLB
CCL7
EDA
PGRMC2
MFAP4
FBLN5
RND1
BRD1
LGALS7
ZNF75D
RAB3A
DHX9
SOX2
HMG20A
DEFB136
PLOD3
DCUN1D3
CMBL
VEGFA
PLD5
TEAD4
SETMAR
SMIM13
CD274
HPX
SMURF2
SYT9
POLM
TIMD4
FGF8
CBS
EIF4EBP3
SH3GLB2
CSGALNACT2
MYSM1
TPM1
FHL1
PRR27
CUTC
SEMA3G
AGAP3
LYVE1
RPS3A
GSTM1
WFDC6
SPIN1
USP28
CA4
C21orf33
CD163
AKR1C3
UNC5A
CD300E
FCAMR
NAGPA
LRP12
CBLN1
LRFN2
ZNRF3
PLEKHA4
INSR
ITIH2
PSMA4
HBZ
TNXB
CDH5
−0.2
0.0
0.2
0.4
Effect Estimates
1.0
0.5
0.0
−0.5
Effect Estimates
0.4
0.2
0.0
−0.2
−0.4
CCL7
SFTPD
CXCL10
KRT7
BRD4
TNNT2
CDHR1
CXCL13
FBLN7
SEMA6C
ICAM5
PTH1R
SETMAR
DNAJC4
HSPH1
RAB3A
GIMAP6
CCDC64
C18orf32
DAG1
PCDHGA12
SLC26A7
GP1BA
IFNL3
PSMB6
PRDX5
DAPK1
CHI3L1
STK17B
CKB
COL2A1
VASN
RNASE13
CLSTN1
CCDC25
CBLN1
ZNRF3
KIT
CLEC6A
GABBR2
SIRT2
APOC1
ABO
NAGPA
ANXA13
VCAN
TBP
PRKCG
ERP27
TIE1
PCYOX1
PLTP
RAP1GAP
PRG4
PRSS37
SMIM9
RELL1
ZNF134
TK2
CBS
NID2
SLC5A8
CDH5
TNR
MICA
APLP1
TNXB
Critical

◂**Figure 3.** Feature importance and correlation of SOMAmers selected in the protein model to predict severe and critical COVID-19. (Left) Coefficient values of the 92 nonzero SOMAmer reagents in the final trained L1 regularized logistic regression protein model fitted to predict severe COVID-19. The original data contained 4984 SOMAmer reagents and 4 other variables: age, sex, sample processing time, and hospital site. 92 SOMAmer reagents remained within the model along with age and sample processing time which are not shown. The model was trained on the entire BQC19 cohort using lambda = 10.0 ($\log_{10}$ lambda = 1.0) which was the best lambda value found from the hyperparameter search. (Bottom) Coefficient values of the 67 nonzero SOMAmer reagents of the final trained L1 regularized logistic regression protein model fitted to predict critical COVID-19. The original data contained 4984 SOMAmer reagents and 4 variables age, sex, sample processing time, and hospital site. 67 SOMAmer reagents remained within the model along with age and sample processing time which are not shown. The model was trained on the entire BQC19 cohort using lambda = 10.0 ($\log_{10}$ lambda = 1.0) which was the best lambda value found from the hyperparameter search. (Right) Spearman's rank correlations between the 92 proteins associated with severe COVID-19 and the 67 proteins associated with critical COVID-19. These results suggest that while there were 14 overlapping proteins (SFTPD, CXCL10, RAB3A, NAGPA, CDH5, IFNA7, ZNRF3, CBS, CCL7, SETMAR, TNXB, CDHR1, CXCL13, and CBLN1), in general, the protein levels were uncorrelated with one another. Out of 6164 total correlations (92 × 67), 6150 correlations (99.8%) had a Spearman's absolute $\rho < 0.8$.

between training and test performance (AUC: 59%). The stability of these AUC estimates in the test cohort suggested that both protein models were robust.

The classification performance of the baseline and protein models in the Mount Sinai cohort is shown as two-by-two contingency tables in Fig. 2c. The baseline and protein models used thresholds of 0.417 and 0.486 to predict severe COVID-19, respectively. These thresholds were selected by computing Youden's J statistic during training and determine the threshold that maximized the sum of the sensitivity and specificity scores during training. The thresholds selected were roughly consistent with the case to control ratio in the BQC19 cohort used for training (175 cases, 242 controls). The protein model achieved a sensitivity of 73.2% compared to 61.0% for the baseline model, and a specificity of 79.7% compared to 60.5% for the baseline model, when predicting severe COVID-19 (Fig. 2c).

When predicting critical COVID-19 using the baseline and protein models, thresholds of 0.202 and 0.255 were used to predict cases, respectively, using the same method. The low threshold for predicting critical COVID-19 cases is consistent with the case to control ratio in training which was 93–324 samples. The baseline model achieved a sensitivity/specificity score of 50.2%/57.4% while the protein model achieved 74.3%/69.6%, suggesting that the protein model trained to predict critical COVID-19 had decent power to classify true positives and true negatives (Fig. 2c).

Furthermore, both the baseline and protein models demonstrated higher positive predictive values than negative predictive values when predicting severe COVID-19, compared to critical COVID-19. In contrast, these models produced higher negative predictive values than positive predictive values when predicting critical COVID-19.

Overall, these results suggest that the protein models predicting severe and critical COVID-19 both perform reasonably well in terms of the trade-off between sensitivity and specificity. The protein model is sensitive (73.2%) at identifying severe COVID-19 cases and similarly sensitive (74.3%) at identifying critical COVID-19 cases. Further, the positive predictive value for severe COVID-19 was high at 88.9%, while the negative predictive value was 57.3% (Fig. 2c). These results suggest that a protein model could predict severe COVID-19 with relatively high confidence.

Proteins prioritized by LASSO to predict COVID-19 severity. To predict severe COVID-19, the best performing protein model selected 92 proteins along with age and sample processing time (Fig. 3, Supplementary Table 3). Assessing the correlation of all 92 proteins, we found that, as expected, most of the proteins did not correlate with each other (mean absolute Spearman's $\rho = 0.17$) (Supplementary Fig. 5A). Of 8464 total correlations (92 × 92), 8372 correlations (98.9%) had Spearman's absolute $\rho < 0.8$.

Next, when predicting critical COVID-19, the best performing protein model retained age, sample processing time, and 67 proteins (Fig. 3, Supplementary Table 4). The absolute effect estimates of these proteins were generally larger than the severe COVID-19 model proteins (mean: 0.081 vs. 0.077). As expected, the 67 selected proteins also showed low levels of correlation (mean absolute Spearman's $\rho = 0.15$) (Supplementary Fig. 5B). Of 4489 total correlations (67 × 67), 4422 correlations (98.5%) had Spearman's absolute $\rho < 0.8$. The correlation between the 92 and 67 proteins selected to best predict severe and critical COVID-19 is shown in Fig. 3. Out of 6164 total correlations (92 × 67), 6150 correlations (99.8%) had Spearman's absolute $\rho < 0.8$. In general, proteins selected for predicting severe versus critical COVID-19 were not highly correlated (mean absolute Spearman's $\rho = 0.15$). A hierarchically clustered heatmap after removal of the 14 common proteins in severe and critical COVID-19 showed that the proteins selected in predicting both outcomes were also generally uncorrelated with one another where 99.2% of the correlations had Spearman's absolute $\rho < 0.8$ (Supplementary Fig. 5C).

The percent of selected proteins for the prediction of severe and critical COVID-19 that were cytokines or chemokines was only 5.4% and 4.5%, respectively. Cytokine IFNA7, as well as chemokines CXCL13, CXCL10, CCL7, and CCL8 were present in the proteins selected for predicting severe COVID-19. Three chemokines, CXCL13, CXCL10, and CCL7 were selected for predicting critical COVID-19. Importantly, among the 14 overlapping proteins, those three chemokines, CXCL13, CXCL10, and CCL7 were selected for predicting both severe and critical COVID-19. In addition, SFTPD, a surfactant protein highly specific to the lung tissue was also among the overlapped proteins in predicting both COVID-19 severity groups.

In addition, we clustered the 4984 SOMAmer reagents using uniform manifold approximation and projection (UMAP) to reduce the feature space to a 2-dimensional space and highlighted the position of the model-selected proteins predictive of either severe or critical COVID-19. Our results suggested that LASSO selected proteins were sparsely distributed across the clusters (Supplementary Fig. 6). This provides further evidence that: (1) few of the selected proteins are closely clustered with one another in UMAP space and (2) the proteins selected from the severe protein model and critical protein model were also quite distant from each other in UMAP space. Finally, we performed pathway analysis on common proteins with non-zero effect estimates that were included in the best predictive models of severe and critical COVID-19. Since LASSO is designed to pick one protein from a group of correlated proteins regardless of their biological relevance, we also included proteins highly correlated (Spearman's absolute $\rho > 0.75$) with the proteins selected by LASSO for the enrichment analysis. As a result, 171 proteins were included in the severe group (92 LASSO selected proteins and 79 correlated proteins, Supplementary Table 5); and 96 proteins were included in the critical group (67 LASSO selected proteins and 29 correlated proteins, Supplementary Table 6). Among which, 32 proteins were common between the severe and critical groups.

We found that these 32 proteins were enriched in 35 pathways (g:SCS adjusted P value < 0.05), among which 15 were directly related to immune responses (Supplementary Fig. 7, Supplementary Table 7). Prominent pathways included viral protein interaction with cytokines, cytokine and cytokine receptors (IL22RA1, TNFRSF10B, CCL7, CXCL10, CXCL13, adjusted $P = 0.0008$) and cytokine-cytokine receptor interactions (CD4, IL22RA1, TNFRSF10B, CCL7, CXCL10, CXCL13, IFNA7, adjusted $P = 0.002$).

Interestingly, more than half of enriched pathways were not related to immune response (e.g., signaling receptor binding, cell activation). This suggests that other non-immune response pathways influence COVID-19 severity. In addition, some of these pathways included protein phosphorylation (CTSG, PRKCZ, PECAM1, CD4, CLEC7A, NPPA, TNFRSF10B, PRDX4, EPHA4, TNXB, CXCL10, IFNA7, CDH5) and glycosaminoglycan binding (CTSG, CCL7, TNXB, CXCL10, CXCL13), suggesting potential avenues to explore for drug development.

Using clinical risk factors to predict COVID-19 severity. In order to contrast the prediction capabilities of protein levels with established clinical risk factors, we performed two sensitivity analyses with results shown in Table 2. In the first analysis, we added six clinical risk factors to the baseline and protein models described in the main analyses. These clinical risk factors were diabetes, chronic obstructive pulmonary disease (COPD), chronic kidney disease, congestive heart failure, hypertension, and liver disease. The prevalence of these risk factors is shown in Table 1.

Addition of these six clinical features to the baseline model improved the training AUC to 64% (from 59%) when predicting severe COVID-19 and to 61% (from 59%) when predicting critical COVID-19. However, adding these clinical risk factors to the protein model resulted in no change in the training AUC performance when predicting severe COVID-19 (AUC = 88% vs. 88%) and critical COVID-19 (AUC = 89% vs. 89%) (Table 2, Supplementary Fig. 8). 95 and 69 features with non-zero beta coefficient effect estimates were selected for the protein models predicting severe and critical COVID-19, respectively, in this sensitivity analysis (Supplementary Fig. 8). Comparing proteins selected by the protein model in this sensitivity analysis, only one protein, KIT, was added to the 94 features selected in the main analysis. For critical COVID-19, the 69 features selected remained the same.

For the second sensitivity analysis, we augmented the first sensitivity analysis with an extra covariate for smoking status. Due to missing smoking information from the CHUM hospital site in the BQC19 cohort, only 312 samples were used in model training for the second sensitivity analysis. The results suggested that the addition of smoking and 6 clinical risk factors into the original baseline model composed of age, sex, sample processing time, and hospital site also slightly improved the training performance when predicting severe COVID-19 (AUC = 66% vs. 59%) and critical COVID-19 (AUC = 61% vs. 59%) (Table 2, Supplementary Fig. 9). When adding smoking and these 6 clinical risk factors to the protein model, we found that training performance actually decreased for severe COVID-19 (AUC = 85% vs. 88%) and critical COVID-19 (AUC = 85% vs. 89%). The non-zero beta coefficients of the proteins for severe and critical COVID-19 are shown in Supplementary Fig. 9 with a total of 79 and 51 features being selected, respectively. Comparing the 79 features selected by the protein model in this sensitivity analysis to the original 94 features selected previously when predicting severe COVID-19, we observed that only 48 features overlapped. Similarly, the 51 features selected by the protein model in this sensitivity analysis only had 28 features overlapping with the 69 features selected previously. The observed decrease in AUC and

	Main analysis ($n = 417$)		Main analysis plus 6 clinical risk factors ($n = 417$)		Main analysis plus 7 clinical risk factors ($n = 312$)	
	Baseline model	Protein model	Baseline model + 6 CRFs	Protein model + 6 CRFs	Baseline model + 7 CRFs	Protein model + 7 CRFs
Severe COVID-19	0.590	0.880	0.636	0.880	0.657	0.848
Critical COVID-19	0.591	0.890	0.607	0.890	0.606	0.848

Table 2. Training AUC comparison. Main Analysis: *Baseline model* age, sex, sample processing time, and hospital site, *Protein model* baseline model and 4984 SOMAmer reagents. Main Analysis plus 6 clinical risk factors (CRFs): *Baseline model + 6 CRFs* age, sex, sample processing time, hospital site, and 6 comorbidities, *Protein model + 6 CRFs* baseline model + 6 CRFs and 4984 SOMAmer reagents. Main Analysis plus 7 clinical risk factors: *Baseline model + 7 CRFs* age, sex, sample processing time, hospital site, 6 comorbidities and smoking status, *Protein model + 7 CRFs* baseline model + 7 CRFs and 4984 SOMAmer reagents.

fewer number of overlapping proteins when comparing main analyses and sensitivity analyses may be due to the reduction in sample size used for training.

The results from these sensitivity analyses suggest that the protein measurements are likely able to act as partial proxies of the tested clinical risk factors. The addition of the clinical risk factors that we assessed may improve the predictive performance for both COVID-19 severity groups when only demographic and sample processing parameters are available. However, when protein measurements are available, adding these extra clinical risk factors may add little for improving predictions.

Discussion

In this large-scale study testing the association of 4701 circulating proteins with severe and critical COVID-19, we found that a subset of these proteins were strong predictors of COVID-19 severity. Specifically, developing a model in 417 individuals and testing its performance in 569 separate samples from an independent external cohort, we demonstrated that a proteomic model was able to predict severe COVID-19, defined as requiring the use of oxygen, with an AUC of 86% and a positive predictive value of 89%. The addition of several commonly used clinical risk factors for COVID-19 severity did not improve the performance of this model. The identified proteins were strongly enriched for cytokine signaling and immune pathways, but also highlighted non-immune pathways. Importantly, sampling was performed on presentation, but the severity outcome is the worst the patient has been at any point and not the clinical status on presentation. Hence, this is in fact a true representation of the patients' final outcomes, and our study predicts their clinical trajectory based on their protein status on presentation. Taken together, these findings demonstrate that circulating protein abundances are able to predict COVID-19 severity with reasonable accuracy.

By including an independent cohort in this study, we implemented best practices for model development and validation[44]. An important aspect of any prediction model is the testing of the model in a cohort separate from the training cohort. Therefore, a strength of this study was that our samples were recruited from three separate hospitals, across two separate health care systems in two different countries. In this study, we used the same clinical risk factors and the exact same proteomic measurement procedure to both train and test the models. This increases the probability that the results presented are generalizable and not overfitted to the training data[45]. Differences in age and sex distributions in the training and testing cohorts were handled using 10 repeats of stratified five-fold cross validation to mitigate their effects on the generalizability of the model. Indeed, for severe COVID-19, there was little change in the AUC when comparing the training and test cohorts (88% vs. 86%).

Further, most studies that have tested the association between protein levels and COVID-19 severity have focused on circulating cytokines and chemokines[15,46–52]. While this is a reasonable approach given the nature of the disease, we are unaware of any other studies that have tested the association of 4701 circulating proteins with COVID-19 severity. A recent study assessing thousands of proteins and their associations with COVID-19 severity achieved an AUC of 85%, but this was not tested in an independent cohort[41].

Interestingly only 5 of the 14 proteins selected in the final model of both severe and critical COVID-19 were cytokines or chemokines. There were also proteins selected that were not specific to immune pathway proteins, such as glycosaminoglycan binding—a favourable set of targets for drug development. The lung-specific protein SFTPD was also selected and could potentially indicate the correlation of severity with lung tissue destruction. This suggests that many of the biological pathways that influence severity of COVID-19 may act distinctly from known cytokine and chemokine proteins.

A major clinical challenge within the pandemic has been the triaging of patients to identify those most likely to require admission for hospitalization[53]. A common reason for hospitalization is the need for oxygen support. Currently, treating physicians are required to assess the need for admission using models with poor predictive performance. A model generated in China early in the pandemic to predict COVID-19 severity requires a medical history, chest X-ray, and extensive blood testing[54]. Further, the 4C Mortality Score was able to predict in-hospital mortality, but achieved an AUC of only 77%[55]. Approximately half of the patients enrolled in our study have developed severe or critical COVID-19 after their baseline proteomic measurement which were used as predictors in this study. This suggests circulating protein measurements could be considered for predicting COVID-19 severity, but this requires further study, including more sampling at the onset of symptoms.

There are differences in genetic ancestry between the training and testing cohorts. The training cohort consists of individuals from two hospital sites in Montreal and may have different allele frequencies at loci related to the identified proteins compared to individuals from the testing cohort in New York. While Ashkenazi Jews are present in both Montreal and New York, major differences include the larger French-Canadian population in Montreal and larger percentage of African genetic ancestry individuals in New York[56]. However, this difference is not expected to affect the validity of the results, as even with this genetic diversity, our model replicated well.

This study has important limitations. While the model was tested in a separate cohort, and generalized well, it should be tested in additional cohorts, especially in cohorts of diverse ancestry. The control population included individuals who were SARS-CoV-2 positive and had mild disease, in addition to individuals who were suspected to have COVID-19 but were SARS-CoV-2 negative. This means that the developed models provide insight into prediction of individuals who develop severe COVID-19 compared to mild COVID-19 and other acute diseases having symptoms consistent with COVID-19. Such control definitions reduce the potential for collider bias, but do not allow direct prediction of COVID-19 severity amongst only COVID-19 patients[57]. While critical COVID-19 case criteria allows the identification of biomarkers involved in critical illness, the specificity of the criteria limits the available cases for study, which limits statistical power, and requires consideration of the sensitivity–specificity trade-off. Hence, we defined severe COVID-19 which encompasses critical COVID-19 individuals and increased the available cases for model training. Deaths were included in both severe and critical COVID-19 case criteria to account for the effects of competing risk. The cause of death for individuals in both

severe and critical COVID-19 groups is unknown. However, cause of death is generally hard to ascertain, difficult to assess, and is an inherent limitation of the data collection process. When including clinical risk factors, smoking status was categorized as current, ex-, and never smokers instead of grouping current and ex-smokers together due to studies showing differences in risk for these two subgroups[58,59]. It may, however, be valuable to group current and ex-smokers together for sensitivity analysis in the future. We used LASSO for biomarker selection and while there are other approaches available for finding the best protein in each of the LASSO sets of proteins, this would require a different study design. For the purpose of external validation, this is not necessary. However, we recognize that other researchers may be interested in which proteins are associated with the ones selected by our models. Thus, we provide a list of highly correlated proteins with those selected by severe and critical COVID-19 in Supplementary Table 5 and 6, respectively. Last, the clinical translation of this study is hindered by the cost involved in measuring 4701 circulating proteins but could be improved by developing a specific assay to the selected proteins.

In summary, circulating protein levels are strongly associated with COVID-19 severity and able to predict the need for oxygen supplementation or death with reasonable accuracy. Measured protein levels were superior to predicting COVID-19 severity when compared to nearly all clinical risk factors tested. Further research is needed to assess whether this proteomic approach can be applied in a clinical setting to assist in triaging patients for admission to hospital.

Methods

Cohorts. The Biobanque Québécoise de la COVID-19 (BQC19) is a Québec-wide biobank which was launched to enable research into the causes and consequences of COVID-19 disease (see bqc19.ca)[60]. For this study, we used results from 417 patients (313 SARS-CoV-2 nasal swab PCR positive patients and 104 individuals who presented with symptoms consistent with COVID-19 but had negative SARS-CoV-2 PCR nasal swabs) with available proteomic data from the SomaScan SomaLogic assay. The subjects were recruited at the Jewish General Hospital (JGH) and Centre Hospitalier de l'Université de Montréal (CHUM) in Montréal, Québec, Canada, both of which are university affiliated hospitals. For each individual, blood samples drawn at the earliest time point were used for training when an individual had multiple blood draws. Selecting the blood sample at the earliest time point reflects the protein measurements during the acute phase of COVID-19 disease. The demographic characteristics of the participants in the BQC19 cohort who underwent SomaScan assays is detailed in Table 1. The demographic characteristics were obtained by medical chart review or patient interview performed by trained clinicians or trained research coordinators.

The Mount Sinai cohort used in this study was composed of results from 569 patients made up of 472 SARS-CoV-2 positive patients and 89 SARS-CoV-2 negative patients confirmed through PCR tests, one COVID-19 positive patient diagnosed by a chest CT while the remaining 7 individuals were COVID-19 negative and did not have COVID-19 symptoms during specimen collection but may have had a history of exposure. The samples donated by the patients in the Mount Sinai cohort underwent the same proteomic data collection and profiling performed as in the BQC19 cohort. The subjects were recruited at the Mount Sinai Hospital in New York City which is affiliated with the Icahn School of Medicine. Table 1 lists the demographic and sample processing parameters of participants in the Mount Sinai cohort that underwent SomaScan assays. Demographic characteristics were obtained similarly to that of the BQC19 cohort.

Demographic, sample processing, and clinical variable definitions. Age and sex from the BQC19 and Mount Sinai cohorts were collected. Sample processing time and hospital site were collected for BQC19 samples with the former being a continuous variable that quantifies the time in hours between sample collection and sample freezing.

The clinical variables were collected for the BQC19 cohort only. Clinical variables included smoking status and six different comorbidities: diabetes, COPD, chronic kidney disease, congestive heart failure, hypertension, and liver disease. All seven variables were collected as categorical values with the six comorbidities having three options (0 No, 1 Yes, and – 1 Don't know) while smoking status contained 4 categories (0 Current Smoker, 1 Ex-smoker, 2 Never smoked, and – 1 Don't know).

Proteomic measurement using the SomaScan platform. Blood samples from both the BQC19 and Mount Sinai cohorts were collected using acid citrate dextrose (ACD) tubes. Proteomic measurement was performed at SomaLogic using the SomaScan v4.0 platform. In the BQC19 cohort, a total 1038 samples collected at different time points from 503 individuals were sent to SomaLogic for proteomic profiling as previously described[2], while the Mount Sinai cohort contained 1200 samples collected at different time points from 592 individuals that were sent to SomaLogic for proteomic profiling.

SomaLogic uses the SomaScan proteomic platform which provides measurements on 4701 unique human circulating proteins using 4987 Slow Off-Rate Modified Aptamers (SOMAmer reagents) and quantifies protein levels in the form of relative fluorescence units (RFUs). Normalization and calibration steps were performed by SomaLogic to remove any systematic biases stemming from raw assays or samples. The normalization procedure involved three steps performed in a non-consecutive fashion: hybridization control normalization, intraplate median signal normalization, as well as plate scaling and calibration. More details on SomaLogic normalization can be found in their Technical Note[61].

Data preprocessing. A per-sample normalization process involved using a scale factor for a set of SOMAmer reagents to compute against a reference value generated from the median of all calibrated, unnormalized samples, and then aggregating the results within a dilution. This was done because using a normal population

reference generated from EDTA plasma tubes would have been inappropriate for normalization, since samples in this study were from ACD plasma tubes. Due to the nature of the samples that were collected from patients during acute infection, we did not apply the recommended scale [0.4–2.5] to remove samples. The raw dataset composed of 5284 SOMAmer reagents was first processed with SomaLogic package SomaDataIO v3.1.0. We removed any SOMAmer reagents that represented non-human proteins or controls (NoneX, NonHuman, Spuriomer, HybControlElution, NonBiotin, NonCleavable) and retained 4984 unique SOMAmer reagents for analysis.

Curation of samples from the longitudinal dataset. To investigate our primary study question, we focused on samples collected during the acute infection stage. Samples from the acute infection stage were defined as samples collected from SARS-CoV-2 PCR positive patients within 14 days of symptom onset. When an individual provided multiple samples collected within 14 days of symptom onset, the sample collected at the earliest timepoint was retained for analyses. Both the BQC19 and Mount Sinai samples adhered to this rule.

COVID-19 severity definitions. We defined two sets of severity groups for COVID-19: severe COVID-19 and critical COVID-19. Positive SARS-CoV-2 results were confirmed by SARS-CoV-2 viral nucleic acid amplification tests (NAAT) from relevant biologic fluids. Cases for severe COVID-19 were defined as individuals who tested positive for COVID-19 and died or required any type of respiratory support (including oxygen delivered by nasal prongs) at any timepoint. Controls for severe COVID-19 were defined as individuals who did not meet these severe case criteria; thus, controls were individuals with COVID-19 but did not meet severe case criteria or were individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative. Cases for critical COVID-19 were defined as individuals who tested positive for COVID-19 and died or required invasive respiratory support (intubation, continuous positive airway pressure, bilevel positive airway pressure, continuous external negative pressure, or high flow positive end expiratory pressure oxygen) at any timepoint. Controls for critical COVID-19 were individuals with COVID-19 but did not meet critical case criteria or were individuals who presented with symptoms of COVID-19 but were SARS-CoV-2 PCR negative.

Multivariable logistic regression. Multivariable logistic regression models were used to test the associations of either severe or critical COVID-19 on four covariates along with each SOMAmer reagent: age, sex, sample processing time, and hospital site. We used R package "glm" to perform 4984 logistic regression models in the BQC19 cohort. We first applied a false discovery rate of $P < 0.01$ (corrected P values were determined using the Benjamini–Hochberg procedure[62], p.adjust with method set to "BH" in R) to select a subset of proteins associated with severe or critical COVID-19. Volcano plots measuring the uncorrected $-\log_{10} P$ values as a function of the effect size estimates of each SOMAmer reagent were generated using the bioinfokit version 2.0.4 package in Python 3.7.

Regularized logistic regression models. We defined two model types differing in the covariates used to train the model. The first model type is a "baseline model" which was trained using age, sex, sample processing time, and hospital site. The second model type is a "protein model" which is trained using age, sex, sample processing time, hospital site, and 4984 SOMAmer reagents. We used the baseline model in our analyses as a performance benchmark to compare the results of the protein model which we expected to perform better.

To predict the two COVID-19 severity groups defined above, we used LASSO regression and elastic nets. Elastic net models were only trained in the BQC19 cohort and not tested in the Mount Sinai cohort (Supplementary Note). Specifically, for LASSO regression we used L1 Regularized Logistic Regression (Sparse Logistic Regression) as implemented in the "LogisticRegression" module from Sci-kit learn version 0.24.1, a machine learning library, with the penalty set to "L1". The L1 norm penalty adds a constraint to the effect estimates of the regression model by setting many variables to have a null effect or a coefficient of 0. This in turn allows a form of feature selection to occur and also prevents overfitting to the training dataset by forcing the model to be less complex. In addition, when multiple variables are correlated with one another such as in the case of highly correlated proteins, the penalty term from LASSO may select a single variable from the group, thus allowing a subset of uncorrelated proteins to be selected. It is important to note that although LASSO tends to select a single variable from a group of highly correlated variables, this property is not a certainty. LASSO may occasionally select more than one variable depending on the size of the dataset and the value of the penalty term. To train this model, the hyperparameter "lambda", which controls the amount of L1 regularization to add to the model, was first tuned through cross-validation (details are described in the next section). A larger value of lambda increases the amount of L1 regularization and forces more of the variables to have a null effect. On the other hand, training the model on a smaller lambda value will result in a model with more nonzero coefficients.

Cross-validation and hyperparameter tuning. Due to the relatively small size of our training dataset from BQC19, we used 10 repeats of five-fold cross-validation to tune the hyperparameter, lambda, over 17 different lambdas ($\log_{10}$ values of lambda from −2 to 2, incremented by 0.25). Each repeat of the five-fold cross-validation process involved splitting the dataset into five folds: training on four folds and validating the trained model on the final fold and performing the process five times to cover each validation fold. We used a stratified cross-validation approach, meaning that the train and validation folds maintained the same percentage of samples of each class (case/control) as the original data. This is important because of the unbalanced case/control samples for critical COVID-19 (93 cases/324 controls). A standard five-fold cross-validation split may result in train and validation folds with varying proportions of cases and controls. Since classification algorithms tend to weight each sample equally, the class that is overrepresented, such as the controls in critical COVID-19,

will receive more weight and thus bias the results. The stratified cross-validation step was performed using the RepeatedStratifiedKFold function in Sci-kit learn. Due to the relatively small sample size of our training set ($n = 417$), we performed 10 repeats of this cross-validation process to stabilize the results from training. Each repeat first shuffled the entire training set then split the data into five folds which created more variability in the data used for training.

During training on the four folds, we standardized only the protein levels using the population standard deviation (i.e., dividing by the number of samples n) and use this mean and standard deviation to standardize the protein levels in the validation fold prior to validating. This prevents information leakage which can occur if standardization of protein levels was performed on the entire dataset rather than just the training folds. Age and sample processing time were treated as continuous variables, whereas sex and hospital site were treated using dummy variables (sex [0: Female, 1: Male], and hospital site [0: CHUM, 1: JGH]).

We used the AUC to determine the model performance during cross-validation. To select the best value for the hyperparameter lambda, we compared the average AUC score (computed from 50 validation fold results) for all lambda values and selected the lambda value corresponding to the highest average AUC. Youden's J statistic was calculated for each receiver operator characteristic (ROC) curve during training. This performance metric can be calculated by subtracting the false positive rate from the true positive rate for each data point on a ROC curve and taking the maximum value. The threshold which corresponds to this maximum Youden's J statistic is the threshold that maximizes the sum of the sensitivity and specificity for that particular ROC curve. We computed the threshold corresponding to the maximum Youden's J statistic for each of the 50 ROC curves and averaged the 50 thresholds to get a single threshold value. This averaged threshold value was computed for each of the baseline and protein models predicting severe and critical COVID-19 and used to produce two-by-two contingency tables and therefore sensitivity and specificity values in the Mount Sinai cohort during model testing.

Model testing. We checked the generalizability of the baseline and protein models by testing them in an external, independent dataset from Mount Sinai. Protein measurements in the test dataset were first natural log transformed then standardized using the mean and standard deviation of the corresponding protein in the training set. Similarly, age was not standardized and kept in years. Since samples were only from a single hospital, the hospital site parameter was left as is and did not need to be dummy encoded. The variable sample processing time, however, was absent from the testing set. For this reason, we imputed the sample processing time variable in the test cohort using the mean value of the sample processing time variable in the BQC19 training cohort.

Protein correlations. Spearman's Rank Order Correlation was used to determine the correlations between individual proteins in the BQC19 cohort. Heatmaps show magnitudes of correlation coefficients between values of − 1 and 1. Correlation heatmaps showing collected clusters such as in Supplementary Fig. 5 were generated using the ggcorrplot function in R with the parameter hc.order set to TRUE to perform hierarchical clustering. Moreover, we reduced the dimension of the correlation matrix of 4984 SOMAmer reagents to a 2-dimensional space using uniform manifold approximation and projection (UMAP) from the umap-learn 0.5.1 package using default parameters. We annotated the SOMAmer reagents selected from the protein model that were associated with severe COVID-19 and critical COVID-19 as well as the proteins that overlapped between the two severity groups.

Pathway enrichment analyses. We used the web-based tool g:Profiler (https://biit.cs.ut.ee/gprofiler/gost) to investigate the possible pathways of the selected proteins as good predictors for both critical and severe COVID-19 identified by LASSO. The g:SCS algorithm was used to estimate the threshold for enrichment against all annotated genes. We selected pathways and interaction databases including Gene Ontology, KEGG[63,64], Reactome; TRANSFAC, miRTarBase, Human Protein Atlas, and CORUM.

Sensitivity analyses. We tested the effect of six established clinical risk factors which included: diabetes, COPD, chronic kidney disease, congestive heart failure, hypertension, and liver disease in the BQC19 cohort to determine whether addition of comorbidities could improve prediction of COVID-19 severity. We added these six additional covariates, with characteristics shown in Table 1, to the baseline and protein models to perform LASSO regression analysis. A total of 417 samples from the BQC19 cohort were used for training.

We performed a second sensitivity analysis by adding smoking status along with these six established clinical variables to the baseline and protein models for LASSO regression analyses. Therefore, the baseline model contained covariates age, sex, sample processing time, hospital site, and seven clinical variables while the protein model contained all the baseline variables along with 4984 SOMAmer reagents. Since smoking status was not available from the CHUM hospital site, this sensitivity analysis only involved 312 samples from the BQC19 cohort that were collected at the JGH site.

Due to missing data, we imputed the values of samples: we first converted all six comorbidity features to binary values. Any value other than a "Yes" was converted to a "No" which may include missing values being converted to a "No". For smoking status, we grouped all values into three categories: 0—Current Smoker, 1—Ex-smoker, and anything else (including missing values and − 1) was set as 2—Never smoked. Smoking status was dummy encoded and had one of the encoded variables dropped to prevent collinearity. For both sensitivity analyses, training of the L1 regularized logistic regression models used 10 repeats of stratified five-fold cross-validation as in the primary analysis.

Ethics declarations. All contributing cohorts to the present analyses received ethics approval from their respective research ethics review boards. The Biobanque Québécoise de la COVID-19 (BQC19) received ethi-

cal approval from the institutional review board (IRB) of the JGH and the CHUM. This research was reviewed and approved by the Icahn School of Medicine at Mount Sinai Program for the Protection of Human Subjects (PPHS) under study number 20-00341. The Mount Sinai PPHS is an accredited IRB. This research was considered minimal risk Human Subjects Research. All research was performed in accordance with relevant guidelines/regulations and informed consent was obtained from all subjects and/or their legal guardians from which blood samples were taken. All methods used in this study were performed in accordance with the Declaration of Helsinki.

Data availability

Code used in this analysis is available at https://github.com/chenyangsu/somalogic with additional information available upon request. The BQC19 is an Open Science Biobank. Instructions on how to access data for individuals from the BQC19 at the Jewish General Hospital site is available at https://www.mcgill.ca/genepi/mcg-covid-19-biobank. Instructions on how to access data from other sites of the BQC19 is available at https://www.bqc19.ca/en/access-data-samples. The proteomics data analyzed during the current study is available for download at https://drive.google.com/drive/folders/1xTMzgKJhkArL6XmqgcUaU3LW4kyxrBjv?usp=sharing. All datasets are available for Open Access for qualified researchers working in academia or industry through the BQC19 (bqc19.ca) and require ethical permission from a Research Ethics Board for use of the data.

Received: 2 June 2022; Accepted: 17 March 2023
Published online: 17 April 2023

References

1. The COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *medRxiv* (2021) doi:https://doi.org/10.1101/2021.03.10.21252820.
2. Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667 (2021).
3. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
4. Nakanishi, T. *et al.* Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. *medRxiv* https://doi.org/10.1101/2021.03.07.21252875 (2021).
5. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
6. Yoshiji, S. *et al.* Proteome-wide Mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity. *medRxiv* https://doi.org/10.1101/2022.06.06.22275997 (2022).
7. Jamilloux, Y. *et al.* Should we stimulate or suppress immune responses in COVID-19? Cytokine and anti-cytokine interventions. *Autoimmun. Rev.* **19**, 102567 (2020).
8. Shi, Y. *et al.* COVID-19 infection: The perspectives on immune responses. *Cell Death Differ.* **27**, 1451–1454 (2020).
9. Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C. & Klein, S. L. Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nat. Rev. Immunol.* **20**, 442–447 (2020).
10. Zhou, Z. *et al.* Heightened innate immune responses in the respiratory tract of COVID-19 patients. *Cell Host Microbe* **27**, 883-890. e2 (2020).
11. Butler-Laporte, G. *et al.* The dynamic changes and sex differences of 147 immune-related proteins during acute COVID-19 in 580 individuals. *Clin. Proteom.* **19**, 1–11 (2022).
12. Hadjadj, J. *et al.* Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science (80-.).* **369**, 718 LP – 724 (2020).
13. Acharya, D., Liu, G. & Gack, M. U. Dysregulation of type I interferon responses in COVID-19. *Nat. Rev. Immunol.* **20**, 397–398 (2020).
14. Lee, J. S. & Shin, E.-C. The type I interferon response in COVID-19: implications for treatment. *Nat. Rev. Immunol.* **20**, 585–586 (2020).
15. Del Valle, D. M. *et al.* An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat. Med.* **26**, 1636–1643 (2020).
16. Chen, Z. & Wherry, E. J. T cell responses in patients with COVID-19. *Nat. Rev. Immunol.* **20**, 529–536 (2020).
17. Williams, S. A. *et al.* Plasma protein patterns as comprehensive indicators of health. *Nat. Med.* **25**, 1851–1857 (2019).
18. Paterson, C. *et al.* Application of a 27-protein candidate cardiovascular surrogate endpoint to track risk ascendancy and resolution in COVID-19 (2020).
19. Ganz, P. *et al.* Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease. *JAMA J. Am. Med. Assoc.* **315**, 2532–2541 (2016).
20. Narasimhan, A. *et al.* Identification of potential serum protein biomarkers and pathways for pancreatic cancer cachexia using an aptamer-based discovery platform. *Cancers (Basel)* **12**, 3787 (2020).
21. Chan, M. Y. *et al.* Prioritizing candidates of post-myocardial infarction heart failure using plasma proteomics and single-cell transcriptomics. *Circulation* **142**, 1408–1421 (2020).
22. Lynch, A. M. *et al.* Plasma biomarkers of reticular pseudodrusen and the risk of progression to advanced age-related macular degeneration. *Transl. Vis. Sci. Technol.* **9**, 12 (2020).
23. Yang, J. *et al.* Impact of kidney function on the blood proteome and on protein cardiovascular risk biomarkers in patients with stable coronary heart disease. *J. Am. Heart Assoc.* **9**, e016463 (2020).
24. Paranjpe, I. *et al.* Proteomic characterization of acute kidney injury in patients hospitalized with SARS-CoV2 infection. *medRxiv* https://doi.org/10.1101/2021.12.09.21267548 (2022).
25. Folkersen, L. *et al.* Genomic evaluation of circulating proteins for drug target characterisation and precision medicine. *bioRxiv* https://doi.org/10.1101/2020.04.03.023804 (2020).
26. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **1**, 727–730 (2002).
27. Bakheet, T. M. & Doig, A. J. Properties and identification of human protein drug targets. *Bioinformatics* **25**, 451–457 (2009).
28. Lauss, M., Kriegner, A., Vierlinger, K. & Noehammer, C. Characterization of the drugged human genome. *Pharmacogenomics* **8**, 1063–1073 (2007).
29. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there?. *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
30. Wallentin, L. *et al.* Angiotensin-converting enzyme 2 (ACE2) levels in relation to risk factors for COVID-19 in two large cohorts of patients with atrial fibrillation. *Eur. Heart J.* **41**, 4037–4046 (2020).

31. Arunachalam, P. S. *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science (80-)* **369**, 1210–1220 (2020).
32. Yaşar, Ş, Çolak, C. & Yologlu, S. Artificial intelligence-based prediction of covid-19 severity on the results of protein profiling. *Comput. Methods Programs Biomed.* **202**, 105996 (2021).
33. Su, Y. *et al.* Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479-1495.e20 (2020).
34. Gisby, J. *et al.* Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death. *Elife* **10**, e64827 (2021).
35. Haljasmägi, L. *et al.* Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. *Sci. Rep.* **10**, 20533 (2020).
36. Yang, Y. *et al.* Exuberant elevation of IP-10, MCP-3 and IL-1ra during SARS-CoV-2 infection is associated with disease severity and fatal outcome. *medRxiv* https://doi.org/10.1101/2020.03.02.20029975 (2020).
37. Consiglio, C. R. *et al.* The Immunology of multisystem inflammatory syndrome in children with COVID-19. *Cell* **183**, 968-981.e7 (2020).
38. Patel, H. *et al.* Proteomic blood profiling in mild, severe and critical COVID-19 patients. *Sci. Rep.* **11**, 6357 (2021).
39. Richardson, T. G., Fang, S., Mitchell, R. E., Holmes, M. V. & Davey Smith, G. Evaluating the effects of cardiometabolic exposures on circulating proteins which may contribute to severe SARS-CoV-2. *EBioMedicine* **64**, 103228 (2021).
40. Pietzner, M. *et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 6397 (2020).
41. Filbin, M. R. *et al.* Longitudinal proteomic analysis of severe COVID-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell reports. Med.* **2**, 100287 (2021).
42. Ma, H. *et al.* Nucleic acid aptamers in cancer research, diagnosis and therapy. *Chem. Soc. Rev.* **44**, 1240–1256 (2015).
43. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
44. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
45. Wiens, J., Guttag, J. & Horvitz, E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.* **21**, 699–706 (2014).
46. Buszko, M. *et al.* Lessons learned: New insights on the role of cytokines in COVID-19. *Nat. Immunol.* **22**, 404–411 (2021).
47. Costela-Ruiz, V. J., Illescas-Montes, R., Puerta-Puerta, J. M., Ruiz, C. & Melguizo-Rodríguez, L. SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. *Cytokine Growth Factor Rev.* **54**, 62–75 (2020).
48. Leisman, D. E. *et al.* Cytokine elevation in severe and critical COVID-19: A rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. *Lancet Respir. Med.* **8**, 1233–1244 (2020).
49. Fajgenbaum, D. C. & June, C. H. Cytokine storm. *N. Engl. J. Med.* **383**, 2255–2273 (2020).
50. Liu, Y. *et al.* Elevated plasma levels of selective cytokines in COVID-19 patients reflect viral load and lung injury. *Natl. Sci. Rev.* **7**, 1003–1011 (2020).
51. De Biasi, S. *et al.* Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nat. Commun.* **11**, 3434 (2020).
52. Laing, A. G. *et al.* A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat. Med.* **26**, 1623–1635 (2020).
53. Certain Medical Conditions and Risk for Severe COVID-19 Illness | CDC. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html?CDC_AA_refVal=https:%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Fgroups-at-higher-risk.html.
54. Liang, W. *et al.* Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern. Med.* **180**, 1081–1089 (2020).
55. Knight, S. R. *et al.* Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. *BMJ* **370**, m3339 (2020).
56. Butler-Laporte, G. *et al.* Exome-wide association study to identify rare variants influencing COVID-19 outcomes: Results from the Host Genetics Initiative. *PLOS Genet.* **18**, e1010367 (2022).
57. Griffith, G. J. *et al.* Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* **11**, 1–12 (2020).
58. Simons, D., Shahab, L., Brown, J. & Perski, O. The association of smoking status with SARS-CoV-2 infection, hospitalization and mortality from COVID-19: A living rapid evidence review with Bayesian meta-analyses (version 7). *Addiction* **116**, 1319–1368 (2021).
59. Patanavanich, R., Siripoon, T., Amponnavarat, S., Glantz, S. A. & Author, C. Active smokers are at higher risk of COVID-19 Death: A systematic review and meta-analysis. *Nicotine Tob. Res.* **25**, 177–184 (2023).
60. Tremblay, K. *et al.* The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PLoS ONE* **16**, e0245031 (2021).
61. Characterization, S. R. Short Technical Note. 1–9 (2020).
62. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
63. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30 (2000).
64. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucl. Acids Res.* **44**, D457–D462 (2016).

Acknowledgements

The Richards research group is supported by the Canadian Institutes of Health Research (CIHR: 365825; 409511, 100558), the Lady Davis Institute of the Jewish General Hospital, the Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada, McGill University, Cancer Research UK [grant number C18281/A29019], and the Fonds de Recherche Québec Santé (FRQS). The Kaufmann lab's COVID-19 work is supported by the Canadian Institutes of Health Research /CITF (VR2-173203 and VS1-175561), the American Foundation for AIDS Research (AmFAR 110068-68-RGCV), the Canadian Foundation for Innovation, and FRQS. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility, and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. These funding agencies had no role in the design, implementation, or interpretation of this study. The measurement of proteomics using the SomaLogic panel was supported by the McGill Interdisciplinary Initiative in Infection and Immunity (MI4) and in whole or in part, by the Bill & Melinda Gates Foundation [INV-017895] for the Mount Sinai cohort. J.B.R. and D.E.K. are supported by FRQS Mérite Clinical Research Scholarships. C.-Y.S. is supported by a Lady Davis Institute / TD Bank Studentship Award.

S.Z. is supported by a CIHR fellowship and an FRQS postdoctoral scholarship. G.B.L. is supported by a CIHR scholarship and a joint FRQS and Québec Ministry of Health and Social Services scholarship. T.N. is supported by Research Fellowships of the Japan Society for the Promotion of Science (JSPS) for Young Scientists. M.D. is supported by a clinician-researcher salary award from the FRQS. V.M. is supported by a Canada Excellence Research Chair. S.G. was supported by grants CA224319, DK124165, and CA196521. Members of the Mount Sinai COVID-19 Biobank Team are listed in Supplementary Table 8. Other authors part of the Mount Sinai Biobank Team that did not directly contribute to the paper are listed in Supplementary Table 8.

Author contributions

Conception and design: C.Y.S., S.Z., W.J., J.P., and J.B.R.; Data analyses: C.Y.S., S.Z., and E.G.K.; Data acquisition: T.N., G.B.L., D.M., D.E.K., J.A., M.A., L.L., E.B.R., D.H., N.K., Z.A., N.R., M.B., L.P., C.G., X.X., C.T., B.V., O.A., T.A., N.A., N.B., M.D., K.N., N.W.S., K.M., D.M.D.V., N.Z., M.P., H.X., J.H., R.M., E.C., K.T., K.A., I.S., N.B., E.K., V.F., T.M., S.G., S.K.S., A.C., M.M., D.E.K., and J.B.R.; Interpretation of data: C.Y.S., S.Z., E.G.K., G.B.L., E.B.R., T.N., D.E.K., and J.B.R.; Funding acquisition: V.M., T.M., S.G., S.K.S., A.C., M.M., D.E.K., and J.B.R.; Methodology: C.Y.S., S.Z., E.G.K., C.M.T.G., C.P., M.H., J.C.Z.S., C.L., and J.B.R.; Project administration: D.M., V.F., A.C., M.M., D.E.K., J.B.R.; Validation: C.Y.S., S.Z., G.B.L., E.G.K., E.B.R., Y.C., Y.F., D.E.K., and J.B.R.; Visualization: C.Y.S. and S.Z.; Writing-original draft: C.Y.S., S.Z., E.G.K., G.B.L., E.B.R., T.N., D.E.K., and J.B.R.; Writing-review and editing: C.Y.S., S.Z., E.G.K., G.B.L., E.B.R., T.N., D.E.K., E.S., M.D., C.P., V.M., D.E.K., and J.B.R.; All authors were involved in further drafts of the manuscript and revised it critically for content. All authors gave final approval of the version to be published. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests

J.B.R. has served as an advisor to GlaxoSmithKline and Deerfield Capital and is the Founder of 5 Prime Sciences. The Lady Davis Institute has previously received funding from GlaxoSmithKline, Eli Lilly, and Biogen for research programs at Dr. Richards' laboratory unrelated to this manuscript. C.P. and M.H. are employees of SomaLogic. All other authors do not have any conflict of interest. TN has received speaking fees from Boehringer Ingelheim for talks unrelated to this research. S.G. reports other research funding from Bristol-Myers Squibb, Boehringer-Ingelheim, Celgene, Genentech, Regeneron, and Takeda. S.G. reports other research funding from Bristol-Myers Squibb, Boehringer-Ingelheim, Celgene, Genentech, Regeneron, and Takeda.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-31850-y.

Correspondence and requests for materials should be addressed to J.B.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

The Mount Sinai COVID-19 Biobank Team

Edgar Gonzalez-Kozlova[5,24], Diane Marie Del Valle[13], Ryan Thompson[14], Mario A. Cedillo[15], Eric Schadt[14], Kai Nie[16], Nicole W. Simons[14], Konstantinos Mouskas[14], Nicolas Zaki[16], Manishkumar Patel[13], Hui Xie[16], Jocelyn Harris[16], Robert Marvin[16], Esther Cheng[14], Kevin Tuballes[13], Kimberly Argueta[16], Ieisha Scott[16], Thomas Marron[19], Noam D. Beckmann[14], Seunghee Kim-schulze[16], Alexander W. Charney[20], Sacha Gnjatic[5,13,16] & Miriam Merad[13]