A circulating proteome-informed prognostic model of COVID-19 disease activity that relies on routinely available clinical laboratories

William Ma[1,*], Antoine Soulé[1,*], Karine Tremblay[2], Simon Rousseau[3,†], and Amin Emad[1,4,†]

[*] These authors contributed equally.

[1] Department of Electrical and Computer Engineering, McGill University, Montréal, QC, Canada

[2] Pharmacology-physiology Department, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Saguenay, QC, Canada; Centre intégré universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean, Saguenay, QC, Canada; CRCHUS, Sherbrooke, Canada.

[3] The Meakins-Christie Laboratories at the Research Institute of the McGill University Heath Centre Research Institute, & Department of Medicine, Faculty of Medicine, McGill University, Montréal, QC, Canada

[4] Mila, Quebec AI Institute, Montréal, QC, Canada

† Co-corresponding Authors:
Amin Emad,
755 McConnell Engineering Building, 3480 University Street, Montreal H3A 0E9, Canada
Email: amin.emad@mcgill.ca

Simon Rousseau,
RI-MUHC, E M3.2244, 1001 Décarie, Montréal H4A 3J1, Canada,
Email: simon.rousseau@mcgill.ca

Abstract

A minority of people infected with SARS-CoV-2 will develop severe COVID-19 disease. To help physicians predict who is more likely to require admission to ICU, we conducted an unsupervised stratification of the circulating proteome that identified six endophenotypes (EPs) among 731 SARS-CoV-2 PCR-positive hospitalized participants in the Biobanque Québécoise de la COVID-19, with varying degrees of disease severity and times to intensive care unit (ICU) admission. One endophenotype, EP6, was associated with a greater proportion of ICU admission, ventilation support, acute respiratory distress syndrome (ARDS) and death. Clinical features of EP6 included increased levels of C-reactive protein, D-dimers, interleukin-6, ferritin, soluble fms-like tyrosine kinase-1, elevated neutrophils, and depleted lymphocytes, whereas another endophenotype (EP5) was associated with cardiovascular complications, congruent with elevated blood biomarkers of cardiovascular disease like N-terminal pro B-type natriuretic peptide (NT-proBNP), Growth Differentiation Factor-15 (GDF-15), and Troponin T. Importantly, a prognostic model solely based on clinical laboratory measurements was developed and validated on 903 patients that generalizes the EPs to new patients recruited across all pandemic waves (2020-2022) and create new opportunities for automated identification of high-risk groups in the clinic. Thus, this novel way to address pathogenesis that leverages detailed phenotypic information but relies on routinely available information in the clinic to favor translation may find applications in other diseases beyond COVID-19.

31 **Introduction**

32 An important challenge facing respirologists and critical care physicians is the heterogeneity in
33 outcome following SARS-CoV-2 infections. A minority of people infected with SARS-CoV-2 will
34 develop a severe form of coronavirus disease 2019 (COVID-19) requiring hospitalization and
35 respiratory support. Defining the molecular mechanisms related to specific severe outcomes is
36 important to identify treatable traits and improve the survival of critically ill patients. Successfully
37 reaching this precision medicine goal requires a more granular definition of the underlying
38 pathophysiology. A symptom-based method to discover molecular mechanisms of the disease is
39 inherently confounded by the fact that the same higher-level condition, such as severe COVID-
40 19 disease, can be produced by several different molecular mechanisms, a phenomenon termed
41 the "many-one" limitation (1). Recent advances in computing strategies, such as machine
42 learning, have enabled the development of methods that help overcome this limitation by,
43 instead of using symptoms, starting from molecular profiles to define endophenotypes, i.e.,
44 subgroups of individuals who are inapparent to traditional classification methods but share a
45 common set of molecular factors that can lead to identification of treatable traits (2-4). Current
46 investigations of endophenotypes in COVID-19 have mainly relied on supervised approaches
47 using fixed outcomes (such as disease severity) and integrating clinical variables at the onset (5).
48 We hypothesize that using an unsupervised approach and exploiting a rich molecular dataset can
49 provide novel mechanistic insights into the pathobiology of severe COVID-19 that can help
50 physicians improve diagnosis, prognosis, and clinical management.

51

52 This study identified six endophenotypes linked to diverse clinical trajectories of COVID-19 using
53 the extensive molecular phenotyping of a cohort of 731 SARS-CoV-2 positive hospitalized patients
54 from the *Biobanque Québécoise de la COVID-19* (BQC19, www.quebeccovidbiobank.ca) (6), a
55 prospective observational cohort of SARS-CoV-2-positive and negative participants recruited in
56 the province of Québec, Canada, to improve our understanding of COVID-19 pathobiology and
57 our capacity to alter disease outcomes. The molecular signature of each endophenotype was
58 used to build a prognostic model of disease severity that generalizes the EPs to new patients and
59 was validated on a separate group of 903 patients. This prognostic model solely utilizes clinical
60 laboratory measurements, creating the possibility of automated identification of high-risk groups
61 in the clinic.
62
63 **Results**
64 **Unsupervised clustering of SARS-CoV-2-positive hospitalized BQC19 participants reveal**
65 **endophenotypes associated with varying disease severity**
66 We aimed to identify endophenotypes of COVID-19, based on the circulating proteome of
67 patients, in a cohort of SARS-CoV-2 positive hospitalized participants in the BQC19 (Table 1) using
68 an unsupervised approach. Figure S1 shows the distribution of the patient hospital admission
69 dates and the corresponding waves of COVID-19 infection as defined by National Institute of
70 Public Health of Quebec (INSPQ, https://www.inspq.qc.ca/covid-19). Consensus agglomerative
71 clustering was performed on participants (n = 731, Table 1) for whom the circulating proteome
72 was measured using a multiplex SOMAmer affinity array (SomaLogic, ~5,000 aptamers) (7). The
73 optimal number of clusters (k = 6) was identified first using two criteria: Akaike's Information

74 Criteria (AIC) and Bayesian Information Criteria (BIC) (Figure 1A). Then, consensus agglomerative
75 clustering (Euclidean distance and Ward linkage) (8, 9) using 1000 bootstrap subsamples of the
76 participants was performed to obtain six robust clusters (Figure 1, Figures S2 and S3). The
77 distribution of Rand-Index, showing the concordance between each one of the 1000 subsampled
78 clusterings and the final consensus clustering, is provided in Figure S2B (mean Rand-Index =
79 0.823), reflecting a high degree of consistency and robustness.
80
81 The clinical and pathological characteristics of patients in each endophenotype is provided in
82 Table 1. To characterize the identified endophenotypes (EPs) with respect to disease severity, we
83 performed two-sided Fisher's exact tests to assess their enrichment (or depletion) in either of
84 "severe" or "deceased" outcomes. EP6 was significantly enriched in the severe/deceased
85 outcomes (Benjamini–Hochberg false discovery rate (FDR) = 1.74E-21) with either of these
86 outcomes observed in 74.6% of EP6 patients. Meanwhile, EP1 was significantly depleted in
87 severe/deceased outcomes (FDR = 1.89E-13) (Figure 2A, Table 1, Table S1) with either of these
88 outcomes observed in only 13.2% of EP1 patients. In addition, EP6 was enriched in participants
89 (a) receiving oxygen therapy (FDR = 4.23E-18), (b) receiving ventilatory support (FDR = 4.59E-18),
90 and (c) being admitted to intensive care unit (ICU) (FDR = 9.51E-28) (Figure 2A, Table 1, Table S1).
91 Kaplan–Meier analysis (10) also confirmed that the identified EPs have a distinct temporal
92 pattern of admission to ICU (multivariate logrank test P = 5.00E-30), with EP1 and EP6 having the
93 highest and lowest probability, respectively, of not being admitted to ICU or dying prior to that
94 in a 40-day span since their admission to the hospital (Figure 2C). A similar pattern was also
95 observed when patients that died before admission to ICU were excluded (Figure S4, multivariate

96 logrank test P = 5.39E-30). A two-sided Mann–Whitney U (MWU) test showed that patients in
97 EP5 were generally older than other EPs (FDR = 7.73E-5), while EP3 included younger patients
98 (FDR = 1.53E-4). Notably, EP6 (which had the most severe patients) did not show enrichment in
99 older patients or individuals with high body mass index (BMI) (two-sided MWU FDR > 0.05)
100 (Figures 2D-F, Table 1, Table S1).
101
102 These analyses revealed that the unsupervised approach using the circulating proteome of the
103 patients was able to identify endophenotypes with distinct disease characteristics and outcomes.
104 We identified EP6 as a group of participants with an increase in key measures of COVID-19 disease
105 severity, including admission to ICU and the need for ventilatory support.
106
107 **EP6 is enriched among BQC19 participants with acute respiratory distress syndrome**
108 To better characterize all EPs with regards to different complications, we performed two-sided
109 Fisher's exact tests comparing each EP to the rest. In accordance with increased COVID-19 disease
110 severity, EP6 was enriched in several medical complications including ARDS (FDR = 1.12E-11),
111 acute kidney injury (FDR = 5.73E-8), secondary bacterial pneumonia (FDR = 2.25E-5), liver
112 dysfunction (FDR = 1.37E-3), cardiovascular complications (FDR = 1.37E-3), and bacteremia (FDR
113 = 4.28E-3) (for the full list, see Figure 3 and Table S2). Notably, the frequency of ARDS was 9% in
114 EP1 compared to 50% in EP6 making this complication a key feature of this endophenotype
115 (Figure 3, Table S2).
116

117 **Clinical laboratories reveal that members of EP6 have increased levels of C-reactive protein, D-**
118 **dimers, elevated neutrophils, and depleted lymphocytes**

119 To further characterize each EP, we assessed the clinical laboratory results obtained from blood
120 draws and compared them between the groups. We focused on 21 markers that were measured
121 in at least 50% of the patients of the cohort and used the summary value reported in the BQC19
122 database corresponding to the most extreme measurement among multiple blood draws (Figure
123 4A, Table S3 also includes first blood draw characteristics). Figure 3A shows the elevation and
124 depletion of these markers in the identified EPs. EP6 is characterized by abnormal values in
125 markers of inflammation (lymphopenia, total white blood cell count, neutrophilia, C-reactive
126 protein (CRP)), liver damage (alanine aminotransferase (ALT)), coagulopathy (D-dimers, low
127 hemoglobin, International Normalized Ratio (INR), and hyperglycemia (glucose). We also used 22
128 markers from the Elecsys diagnostic panel (Roche Diagnostic) to further characterize EP6, (Figure
129 4B and Table S3). This led to additional elevated and highly significant markers: (a) alpha-1
130 antitrypsin, an acute phase reactant elevated during inflammatory conditions; (b) Interleukin 6
131 (IL-6), a pleiotropic cytokine associated with systemic inflammatory response syndrome (11),
132 shown to be elevated in severe COVID-19 (12) and linked to endothelial damage and liver injury
133 (13); (c) ferritin, an iron-storage protein and acute phase reactant, elevated in COVID-19, and like
134 other hyperferritinic syndrome, associated with coagulopathy (14, 15); and (d) soluble vascular
135 endothelial growth factor (VEGF) receptor sFLT1 (soluble fms-like tyrosine kinase-1), previously
136 shown to be associated with endothelial damage and COVID-19 severity (16). The overall
137 characteristics of each EP are summarized in Table 2.

138

139 **EP5 is associated with cardiovascular complications**

140 EP5 comes second in the order of severity established in Figure 2. Interestingly, it is molecularly
141 and clinically distinct from EP6 (Figures 2-4, Table 1). A striking feature of EP5 is the increase in
142 markers of cardiovascular diseases, such as higher levels of N-terminal pro B-type natriuretic
143 peptide (NT-proBNP), indicative of ventricular dysfunction (17), Growth Differentiation Factor-15
144 (GDF-15) associated with cardiometabolic risk (18) and Troponin T linked to cardiac damage all
145 suggestive of high risk for cardiovascular events (19) (Figure 4B). Accordingly, this group was
146 enriched for cardiovascular complications during hospitalization (FDR = 1.46E-2, Figure 3). As
147 postulated, the unsupervised clustering was able to distinguish different types of COVID-19
148 disease trajectory.

149

150 **A computational prognostic model based on blood biomarkers predicts EPs in a separate**
151 **validation cohort**

152 Since each EP showed a clear and distinct clinical laboratory result signature based on 21 blood
153 markers and 22 Elecsys diagnostic markers, we sought to develop a computational prognostic
154 model of disease severity based on these signatures. We focused on data from the first blood
155 draw (Figure 4B and S5, Table S3) and developed a nearest-centroid classifier, capable of dealing
156 with missing values, to predict EPs based on these 43 markers (see Methods for details). To test
157 the prognostic ability of this model on an independent yet similar dataset, we analyzed 903 SARS-
158 CoV-2 positive hospitalized BQC19 participants that did not have circulating proteome data and
159 had not been used to identify the endophenotypes (see Figure S6 for the distribution of the
160 patient hospital admission dates). These patients were recruited across all waves of the pandemic

161 between March 2020 and October 2022. The clinical and pathological characteristics of patients
162 in each predicted endophenotype (PEP) are provided in Figure 5 and Table S4.
163
164 Our prognostic model identified 167 of these 903 patients as belonging to predicted EP6 (PEP6).
165 Fisher's exact tests showed significant enrichments of PEP6 in severe/deceased (FDR = 5.62E-21),
166 while PEP1 and PEP2 were significantly depleted in these outcomes (FDR = 5.29E-8 and FDR =
167 1.19E-8, respectively), as shown in Table S4. Like EP6, PEP6 was also significantly enriched in
168 participants (a) receiving oxygen therapy (FDR = 2.87E-13), (b) receiving ventilatory support (FDR
169 = 1.50E-13), and (c) being admitted to ICU (FDR = 1.34E-19) (Table S4). Kaplan–Meier analysis
170 also confirmed that these PEPs have a distinct temporal pattern of admission to ICU (multivariate
171 logrank test P = 1.56E-21), with PEP6 having the highest chance of being admitted to ICU (or dying
172 prior to that) in the 40-day span following admission to hospital (Figure 5B). These results suggest
173 that our prognostic model based on 43 blood biomarkers can be used to generalize the definition
174 of endophenotypes to patients for whom proteomic data is not available. Since the 21 blood
175 markers are more commonly available, we also developed a prognostic model only based on
176 these markers, which also showed strong prognostic capabilities (Figure S7). Therefore, it is
177 possible to leverage detailed molecular information on a smaller number of participants to
178 predict clinical outcomes on a larger population using routinely available information collected
179 during hospitalization.
180
181 **Discussion**

182 In this study, we have bridged the gap between the circulating proteome and routinely available
183 blood diagnostic biomarkers, using machine learning algorithms, to prognosticate COVID-19
184 outcomes in hospitalized patients. The model performed on participants recruited across all the
185 pandemic waves from 2020 to 2022, demonstrating that it performs despite mutations in
186 infecting strains and the development of immunity. This showcases a novel analytical pipeline
187 that can support physicians in making more informed decision on potential unfavorable
188 trajectories early during hospitalization and adjust follow-ups and treatments accordingly.
189
190 The major strength of this study is the use of an unsupervised approach for analysis of a large
191 and well-phenotyped cohort. This broad-based approach led to the identification of six COVID-
192 19 disease endophenotypes in hospitalized patients that could not be captured by simply
193 classifying the population solely based on severity, with different clinical trajectories and
194 distinguishing characteristics that are summarized in Tables 1 and 2. We identified two
195 endophenotypes with more favorable outcomes (EP1 and EP2), three endophenotypes with
196 intermediate outcomes in terms of severity (EP3, EP4 and EP5) and one endophenotype which
197 led to worst outcomes compared to all others (EP6). EP6, was associated with ARDS, the worst
198 clinical manifestation of COVID-19 that was reflected by a greater proportion of ICU admission,
199 mechanical ventilation, and severe/fatal outcomes (Figures 2 and 3). Clinical features of this
200 endophenotype were consistent with published literature, including increased levels of CRP, D-
201 dimers, IL-6, ferritin, sFLT1, elevated circulating neutrophils, and reduced peripheral blood
202 lymphocytes (Figure 3, Table S3), presenting a profile associated with systemic inflammatory
203 response syndrome and abnormal coagulation. Possible molecular effectors of COVID-19 disease

204 severity in EP6 are discussed in an accompanying study. Another endophenotype (EP5), while
205 leading to unfavorable clinical trajectory during hospitalization, was instead associated with clear
206 markers of cardiovascular disease, cardiovascular complications during hospitalization, and older
207 age. The distribution of clinical laboratories in each endophenotype was sufficient to train an
208 accurate prognostic model that could readily support future clinical care, since it only requires
209 data from routine clinical laboratory results for prognosis.
210
211 The identification of endophenotypes was done systematically using robust consensus clustering
212 of aptamer expression levels in which the optimum number of clusters was determined
213 congruently using two well-established measures: AIC and BIC. The consensus clustering using
214 bootstrap sampling (1000 times) ensured identification of robust clusters that are not sensitive
215 to exclusion of some of the samples (20% randomly selected and excluded at each cycle). The
216 mean Rand-index between each of the 1000 subsampled clusterings and the final consensus
217 clustering was 0.823, reflecting a high degree of concordance and robustness. Moreover,
218 identifying the best number of clusters using AIC/BIC (both of which agreed with each other)
219 allowed us to reveal the patterns of the EPs directly from the data instead of imposing a pattern
220 onto it through human supervision. This is an important strength of the study that enabled us to
221 identify distinct molecular patterns of patients that could have remained undetected using other
222 traditional approaches.
223
224 Additionally, to improve the translational applicability of the EPs, we developed a prognostic
225 model based only on measurements of conventional clinical laboratory blood markers to test the

226 generalizability of these endophenotypes to samples without measured aptamer data.
227 Characteristics of EPs predicted solely based on their blood markers were consistent with the
228 original EPs, suggesting that clinical blood markers could be used as surrogates for assignment of
229 these EPs to new patients and potentially automating identification of high-risk groups in the
230 clinic. This approach takes into account the effect of multiple blood variables simultaneously and
231 incorporates the full distribution of each variable. This is in contrast to the clinical laboratory
232 results that are automatically flagged as within or outside normal range, one variable at a time,
233 therefore increasing the clinical applicability of our model by leveraging a wider spectrum of
234 information to prognosticate patient outcomes.
235
236 *Limitations and considerations*
237 The data presented in this study comes from individuals participating in the BQC19, a prospective
238 observational cohort built to study COVID-19 in Québec (Canada) with its specific population
239 profile as reported previously (6). A chronological bias may also be present, as most of the
240 participants used for endophenotyping in this study were recruited during the first two waves of
241 the pandemic (Figure S1), prior to widespread vaccination in Québec and the appearance of the
242 Omicron variant and sub-variants. Therefore, some of the features of the identified
243 endophenotypes may change over the course of the pandemic. It will be essential to continue
244 longitudinal assessments of the molecular profiles to better understand the dynamic nature of
245 host-pathogen interactions. It will also be interesting to compare the profiles of COVID-19 ARDS
246 to other viral-induced ARDS, to identify common as well as distinguishing features of these
247 conditions.

248

249 **Conclusion**

250 Respiratory infections represent an important challenge for respirologists and critical care
251 physicians due to the heterogeneity of outcomes. Developing better ways to prognosticate poor
252 outcomes is crucial in improving patients' care and survival. In this manuscript, we proposed a
253 novel experimental approach that leverages detailed proteomic information but relies on
254 routinely available information in the clinic for prognostication to favor translation that may find
255 applications in many other diseases beyond COVID-19.

256

257 **Methods:**

258 **Datasets and preprocessing**

259 The Biobanque Québécoise de la COVID-19 (BQC19; www.quebeccovidbiobank.ca) is aimed at
260 coordinating the collection of patients' data and samples for COVID-19 related research. Data
261 and samples were collected from ten sites across the province of Québec (Canada) (6). BQC19
262 organizes the collected data, including clinical information and multi-omics experimental data,
263 before making it available in successive releases. For this study, we used the circulating proteome
264 determined using SOMAmers. Our main corpus of analysis consisted of n = 1,634 hospitalized
265 and SARS-CoV-2 positive patients (based on qRT-PCR) of BQC19. This included n = 731 patients
266 (Figure S1) for which both clinical and proteomic data was available as well as n = 903 patients
267 (Figure S6) for whom proteomic data was not available but whose clinical data contained
268 measurements for more than half (at least 11 out of 21) of the blood markers that we used as a
269 validation set for the prognostic model developed in this study.

270

271 We also obtained data (n = 731) corresponding to the circulating proteome measured between
272 April 2, 2020 and April 20, 2021 by a multiplex SOMAmer affinity array (SomaLogic, 4,985
273 aptamers) from BQC19. When measurements of the same patients but at different time points
274 were available, we used the one corresponding to the first time point. SomaScan is a
275 biotechnological protocol commercialized by the SomaLogic company (7). It relies on a set of
276 artificial aptamers linked to a fluorophore and each designed to bind a single protein. Once added
277 to the sample, the activity of each aptamer is measured through fluorescence and used to
278 approximate the expression level of the targeted protein. SomaScan protocol comprises several
279 levels of calibration and normalization to correct technical biases. Log2 and Z-score normalization
280 were performed on each aptamer separately in addition to the manufacturer's provided
281 normalized data (hybridization control normalization, intraplate median signal normalization,
282 and median signal normalization). Since the data was analyzed by SomaLogic in two separate
283 batches, we applied the z-score transformation separately to each batch, to reduce batch effects.
284 These additional transformations ensure that the measured values of different aptamers are
285 comparable and can be used in cluster analysis.

286

287 **Consensus agglomerative clustering**

288 Patients were clustered using agglomerative clustering with Euclidean distance and Ward's
289 linkage (8, 9). To identify number of clusters k, we used the elbow method based on the AIC and
290 BIC . More specifically, we calculated the AIC and BIC for clustering using k = 2, 3, …, 20 and used
291 the Kneedle algorithm (20) to identify the value of k corresponding to the "elbow", where

292 increasing the value of k does not provide much better modeling of the data. Kneedle identified
293 k = 6 as the number of clusters based on both AIC and BIC (Figure 1A).
294
295 Given the number of clusters in the data, we then used consensus clustering with sub-sampling
296 to obtain robust endophenotypes. We randomly sampled 80% of the patients 1000 times. Each
297 time, we used the agglomerative clustering above with k = 6 to identify clusters. Given these 1000
298 clusterings, we calculated the frequency of two patients appearing in the same cluster, when
299 both were present in the randomly formed dataset. We then performed one final agglomerative
300 clustering of these frequency scores to identify the six endophenotypes (Figure S2A and Figure
301 1B).
302
303 **Nearest-centroid predictor based on blood markers**
304 In order to predict endophenotypes from blood tests, we developed a missing-value resilient
305 nearest-centroid classifier. We used the dataset of patients that were used to form the original
306 EPs (n = 731) as the training set and the dataset of patients that did not have proteome data as
307 the validation set (n = 903). First, we z-score normalized each of the 43 markers across all the
308 patients in the training set, one marker at a time. We then formed a marker signature (a vector
309 of length 43) for each EP. Each element of an EP's signature corresponds to the mean of the
310 corresponding marker across all patients of that EP.
311
312 To predict the EP label of each patient in the test set, we first z-score normalized their blood
313 marker measurements using the mean and standard deviation of the markers calculated from

314 the training set. Then, we calculated the cosine distance between each test patient's blood
315 marker profile and the centroids (excluding missing values) and identified the nearest EP as the
316 predicted EP (PEP) label of the patient.
317
318 **Statistics**
319 Several non-parametric tests, including Mann–Whitney U test, Fisher's exact test, and
320 Spearman's rank correlation, were used in this study. Benjamini–Hochberg false discovery rate
321 (FDR) was used to adjust the p-values for multiple tests.
322
323 **Study approval**
324 The study was approved by the Institutional Ethics Review Board of the "Centre intégré
325 universitaire de santé et de services sociaux du Saguenay-Lac-Saint-Jean" (CIUSSS-SLSJ) affiliated
326 to the Université de Sherbrooke [protocol #2021-369, 2021-014 CMDO – COVID19].
327
328 **Acknowledgements**
329 This work was made possible through open sharing of data and samples from the Biobanque
330 Québécoise de la COVID-19, funded by the Fonds de recherche du Québec - Santé, Génome
331 Québec, the Public Health Agency of Canada and, as of March 2022, the Ministère de la Santé et
332 des Services Sociaux du Québec. We thank all participants to BQC19 for their contribution. This
333 study was supported by the Fonds de recherche du Québec - Santé (FRQS)- Cardiometabolic
334 Health, Diabetes and Obesity Research Network (CMDO)- Initiative. This work was also supported
335 by Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2019-

04460 (AE). The authors also acknowledge the in-kind contribution of Roche Diagnostics, a division of Hoffmann–La Roche Limited, which provided the reagents for the biomarker analyses conducted on the BQC19 blood samples.

Conflict of Interests

The authors have declared that no conflict of interest exists.

Figures

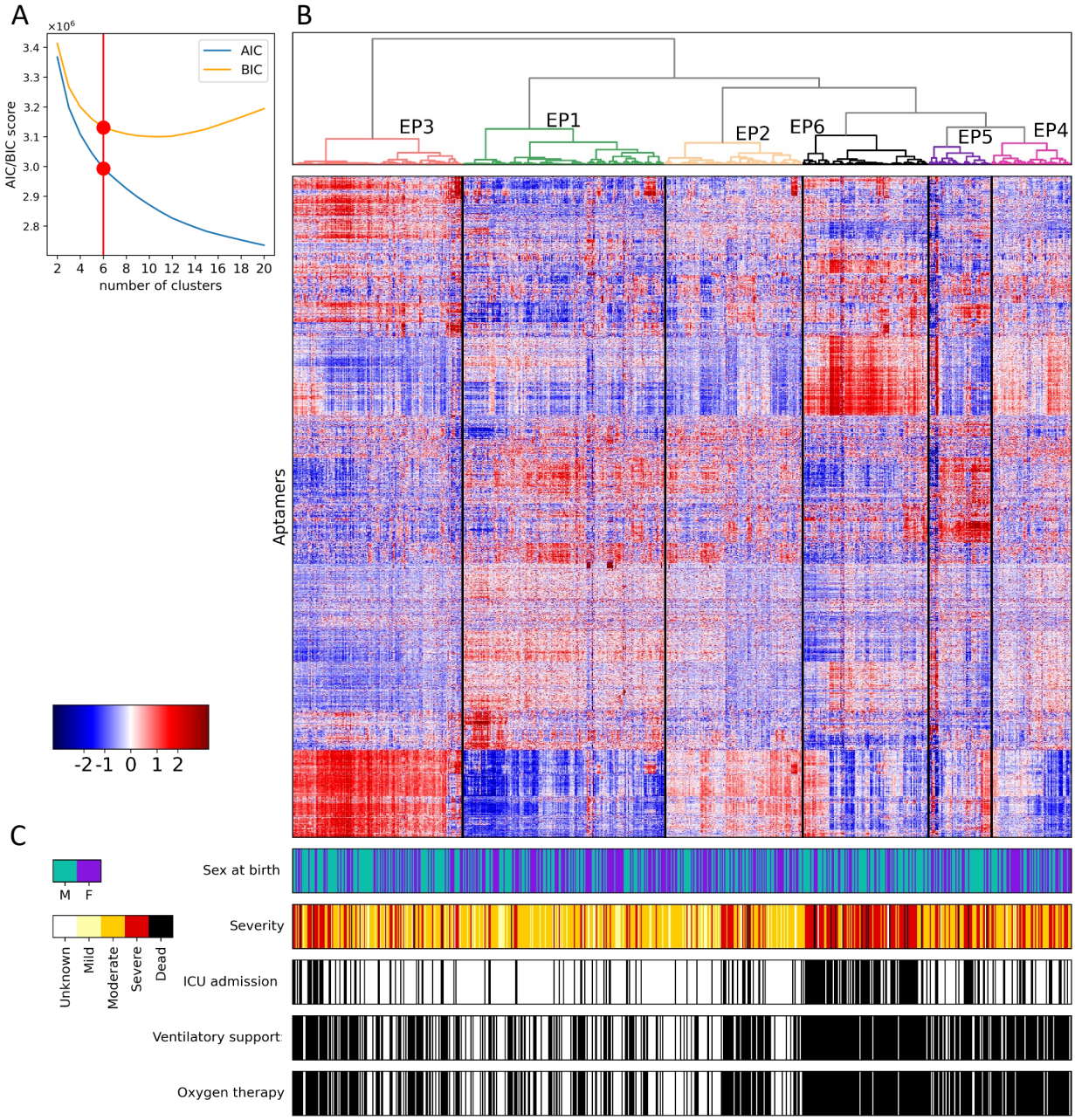

Figure 1: Unsupervised consensus clustering of SARS-CoV-2 positive patients. A) The elbow points (circles in red) of Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC) curves versus number of clusters consistently corresponded to k=6 as the optimal number of clusters. B) The heatmap shows the expression of aptamers (rows) in each sample (columns). The dendrogram shows the identified endophenotypes. C) Characterization of samples based on sex at birth, highest world health organization (WHO) severity level achieved, intensive care unit (ICU) admission, ventilatory support, and oxygen therapy. For the last three rows, a sample colored “black” reflects a label of “yes”.

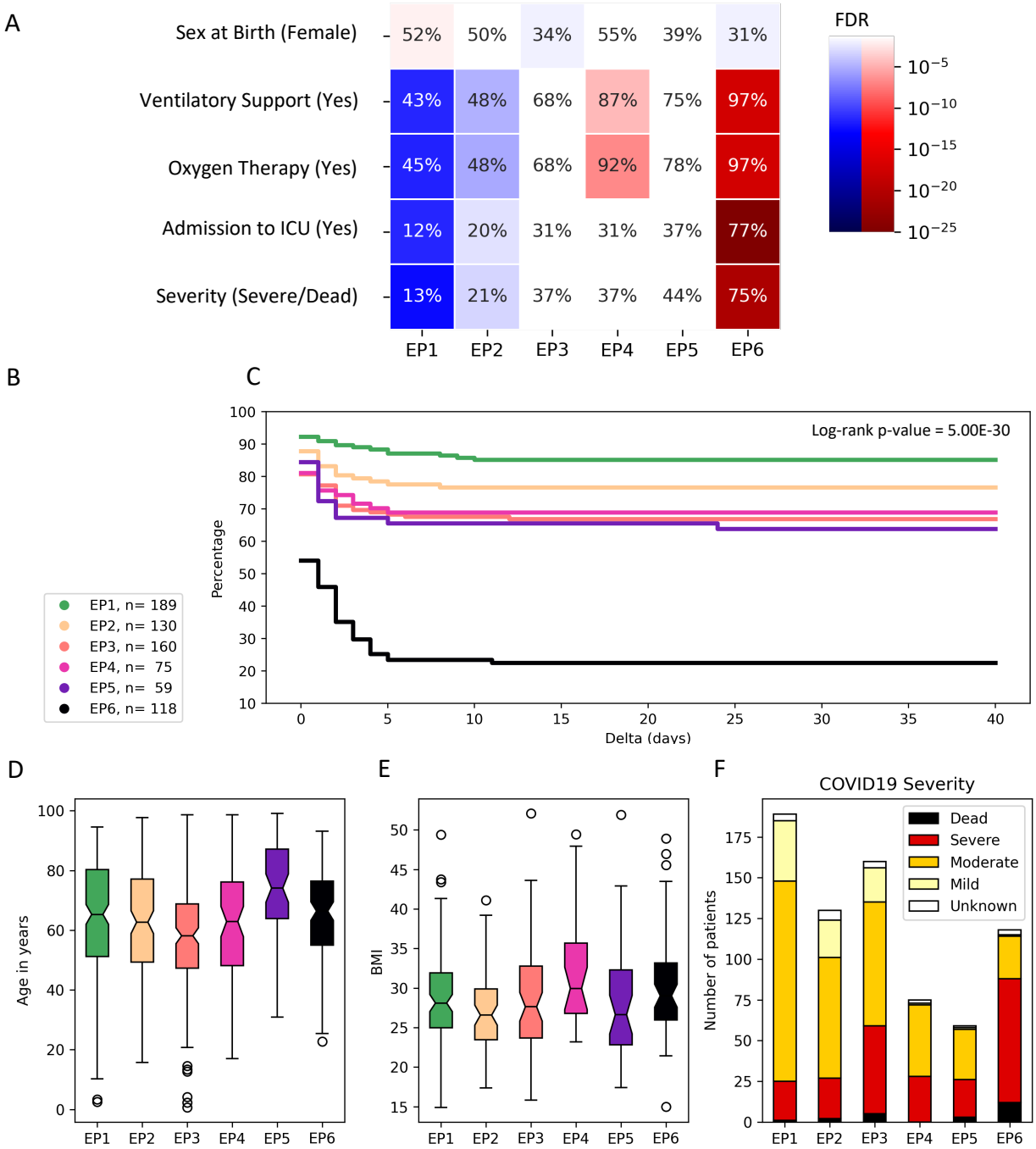

Figure 2: Characterization of endophenotypes (EPs).
A) Enrichment or depletion of each EP in clinical variables (one cluster versus rest). Two-sided Fisher's exact tests are used to calculate the p-values, which are corrected for multiple tests using Benjamini–Hochberg false discovery rate (FDR). Gradients of blue show depletion, while gradients of red show enrichment. FDR values above 0.05 are depicted as white. B) The number of patients in each EP and the colors used to represent them in panels C, D, and E. C) Kaplan–Meier analysis of the time between patients' admission to the hospital and their admission to intensive care unit (ICU) (or death if earlier) for each EP (Delta). D) Distribution of age in each EP. E) Distribution of BMI in each EP. F) COVID-19 severity in each EP.

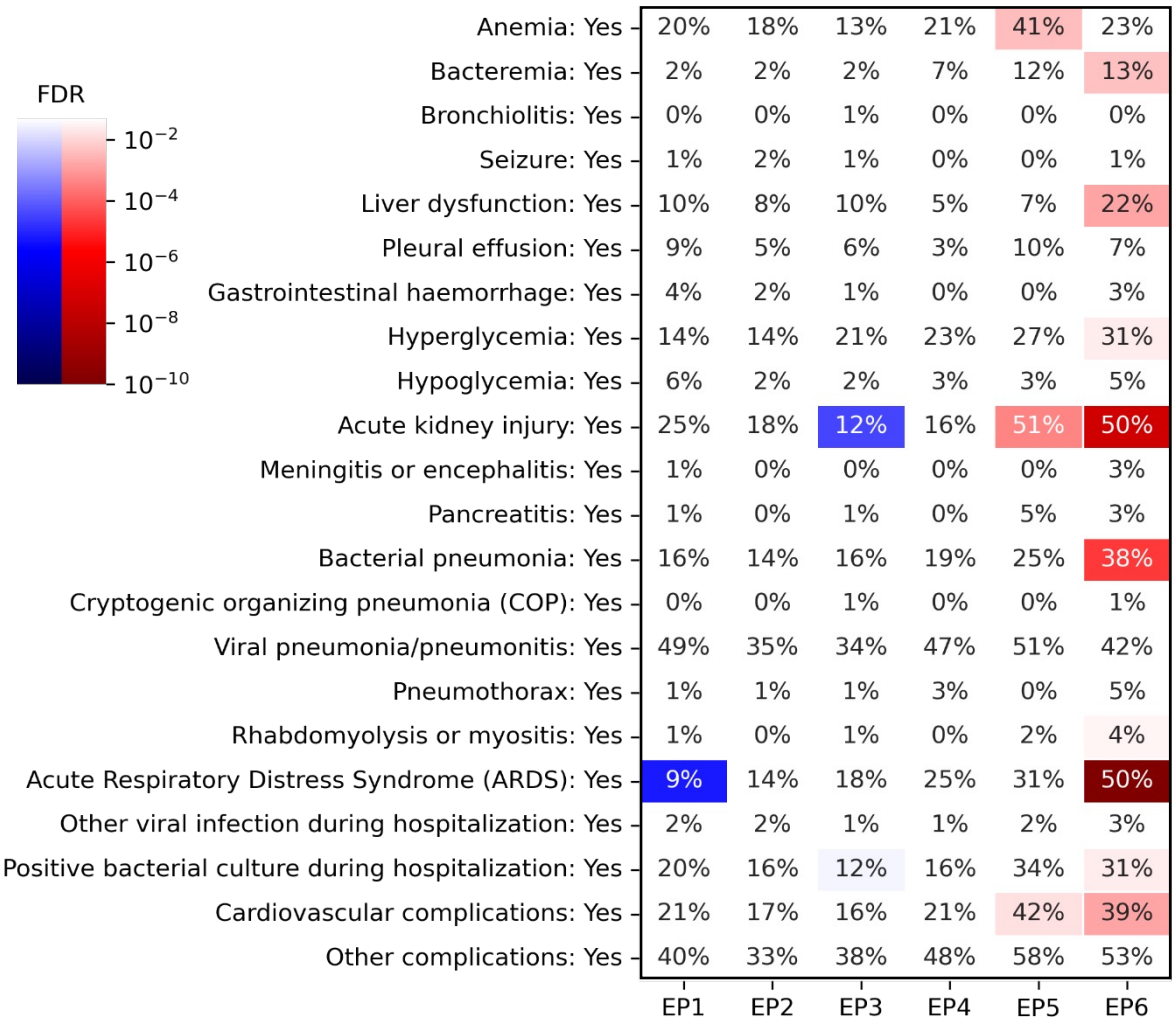

Figure 3: Frequency and significance of complications in different EPs.
The value in each cell shows the percentage of patients of that EP (column) that suffered from the complication (row). The colors represent two-sided Fisher's exact test false discovery rate (FDR, corrected for multiple tests). Red represents enrichment, while blue represents depletion. FDR values below 0.05 are shown as white.

A

	EP1	EP2	EP3	EP4	EP5	EP6
WBC (Highest) (x 10^9/L)	9.3	8.9	10.2	10.6	10.5	14.8
Neutrophil (Highest) (relative %)	77.6	79.7	81.1	83.4	82.9	89.2
Lymphocyte (Lowest) (relative %)	14.0	11.6	8.7	8.5	6.8	4.5
Monocyte (Highest) (relative %)	11.2	10.3	9.6	9.7	11.8	9.8
Eosinophil (Highest) (relative %)	1.9	1.7	1.1	1.0	2.0	2.5
Basophil (Highest) (relative %)	0.4	0.4	0.2	0.5	0.6	0.6
Platelet (Lowest) (x 10^9/L)	186	198	234	203	130	180
Haemoglobin (Lowest) (g/L)	116	118	118	116	81	100
Urea (Highest) (mmol/L)	7.3	7.0	7.8	7.7	19.9	16.4
Creatinine (Highest) (umol/L)	83	87	76	79	255	114
Sodium Na+ (Highest) (mmol/L)	142	141	142	142	142	146
Potassium K+ (Highest) (mmol/L)	4.4	4.4	4.5	4.5	4.9	4.9
C-reactive protein (Highest) (U/L)	77	87	102	124	94	160
LDH (Highest) (U/L)	333	344	394	418	372	589
Albumin (Lowest) (g/L)	31	31	30	30	27	25
ALT (Highest) (U/L)	35	40	53	65	29	100
APTT (Highest) (sec)	34	32	33	35	36	38
INR (Highest)	1.2	1.2	1.1	1.2	1.6	1.3
Bilirubin (Highest) (umol/L)	9.0	9.0	10.0	9.0	9.0	13.0
Glucose (Highest) (mmol/L)	8.0	7.5	8.4	8.6	11.6	11.5
D-Dimer (Highest) (ug/L)	974	938	1053	1002	1775	2083

FDR: 10^{-5}, 10^{-10}, 10^{-15}, 10^{-20}, 10^{-25}

B

	EP1	EP2	EP3	EP4	EP5	EP6
Alpha-1 Antitrypsin (g/L)	1.2	1.4	1.6	1.5	1.5	1.8
Anti-SARS-CoV-2 antibodies (U/mL)	4.09	2.56	10.39	21.12	0.40	30.67
Alkaline phosphatase (U/L)	50	62	55	53	72	63
Cancer antigen 125 (U/mL)	8.9	9.5	8.8	8.3	13.0	12.8
Creatine Kinase (U/L)	38	59	64	49	71	107
Creatine Kinase MB (ng/mL)	0.86	0.88	1.07	0.90	1.82	1.61
Calcium (mmol/L)	1.6	1.7	1.7	1.6	1.6	1.6
Cholesterol (mmol/L)	2.4	2.8	2.7	2.5	2.4	2.6
Cystatin C (mg/L)	0.76	0.82	0.82	0.83	2.42	1.17
Ferritin (ug/L)	213	289	566	474	634	835
Growth/diff. factor-15 (pg/mL)	2017	2176	2141	2136	7199	4454
Human Epididymis Protein 4 (pmol/L)	81	73	105	98	482	233
Interleukine-6 (pg/mL)	12	17	15	29	43	53
Brain natriuretic peptide 32 (pg/mL)	140	91	140	198	2438	339
Procalcitonin (ng/mL)	0.051	0.072	0.065	0.074	0.246	0.185
Phosphate (mmol/L)	0.73	0.92	0.83	0.83	1.15	0.89
Free T4 (pmol/L)	16	15	16	16	15	14
Triglycerides (mmol/L)	1.00	1.08	1.29	1.14	1.10	1.53
Troponin T HS (ng/L)	9.9	9.0	9.0	10.0	58.3	19.1
Vitamin B12 (pmol/L)	360	373	423	441	564	581
Vitamin D (nmol/L)	30	36	25	30	43	24
Soluble fms-like tyrosine kinase 1 (pg/mL)	76	87	98	89	114	126

FDR: 10^{-5}, 10^{-10}, 10^{-15}, 10^{-20}, 10^{-25}

369

Figure 4: Patterns of blood markers and Roche diagnostic markers in the endophenotypes (EPs). Heatmaps show the false discovery rate (FDR) values for two-sided one-vs-rest Mann–Whitney U tests for 21 blood markers (most extreme value during hospitalization) (A) and 22 Roche diagnostic markers (B) for each EP. FDR values below 0.05 are shown as white. The numerical values show the median value of the maker in each EP. Abbreviations used: WBC = white blood cells, LDH = lactate dehydrogenase, ALT = alanine aminotransferase, aPTT = activated partial thromboplastin time, INR = International Normalized Ratio.

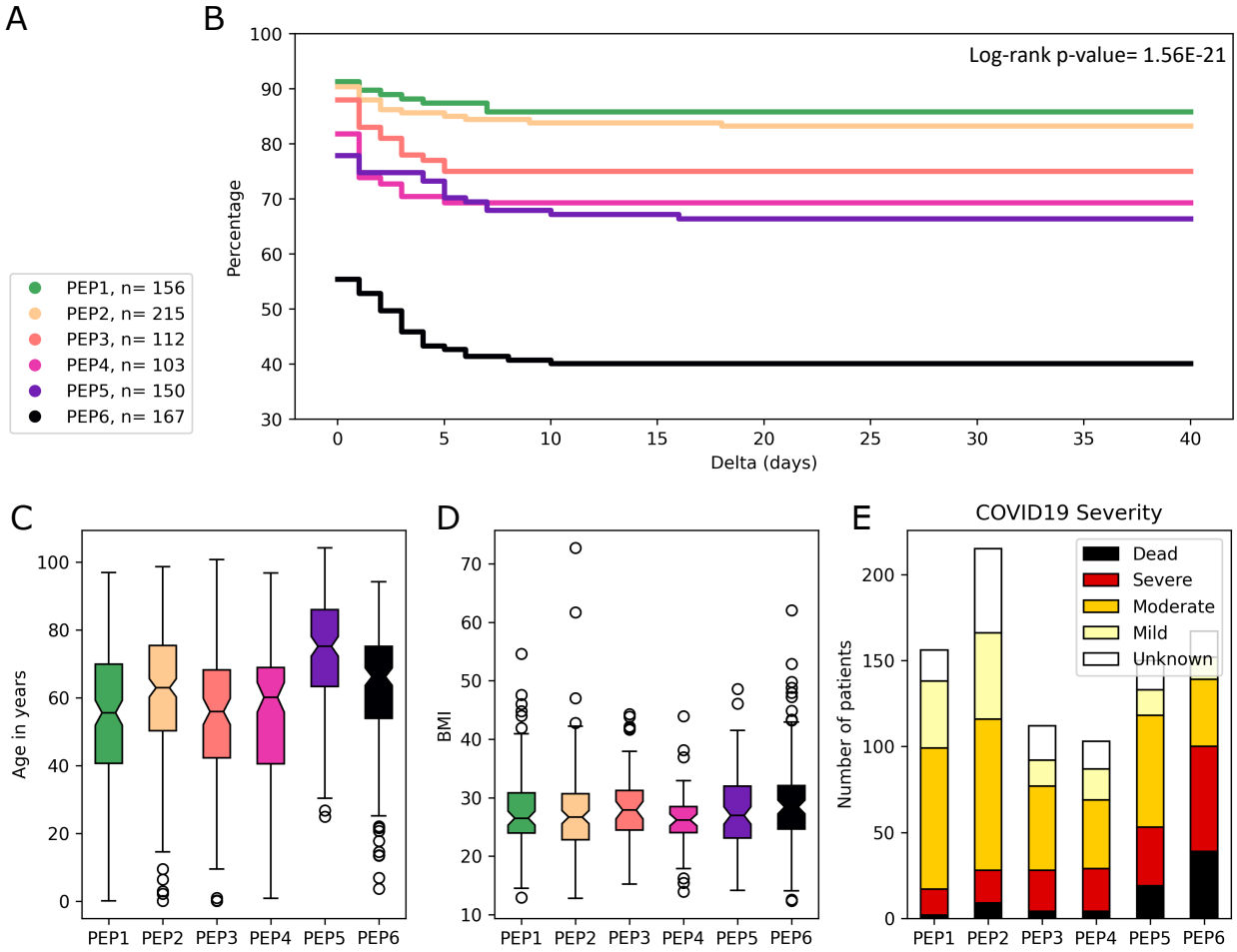

Figure 5: Characterization of predicted endophenotypes (PEPs) based on the prognostic model using 21 blood markers and 22 Roche diagnostic markers. A) The number of patients in each PEP and the colors used to represent them in panels B, C, D, and E. C) Kaplan–Meier analysis of the time between patients' admission to the hospital and their admission to intensive care unit (ICU) (or death if earlier) for each PEP (Delta). D) Distribution of age in each PEP. E) Distribution of BMI in each PEP. F) World health organization COVID-19 severity in each PEP.

388 **Tables**

389 **Table 1:** Clinical and pathological characteristics of the BQC19's participants used in to identify
390 endophenotypes (EPs) in this study.
391

		Cohort (n=731) No. (%)	EP1 (n=189) No. (%)	EP2 (n=130) No. (%)	EP3 (n=160) No. (%)	EP4 (n=75) No. (%)	EP5 (n=59) No. (%)	EP6 (n=118) No. (%)
Age (years)	**<45**	17.0	18.0	21.5	21.2	16.0	6.8	10.2
	45-65	34.9	31.7	34.6	43.1	36.0	18.6	36.4
	>65	48.2	50.3	43.8	35.6	48.0	74.6	53.4
	Unknown	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Body mass index in kg/m^2	**<20**	2.2	2.6	3.8	0.6	0.0	5.1	1.7
	20-25	11.8	11.6	14.6	12.5	4.0	18.6	9.3
	25-35	27.2	33.9	26.9	18.8	21.3	28.8	31.4
	>35	7.0	5.8	3.1	6.9	10.7	10.2	9.3
	Unknown	51.8	46.0	51.5	61.3	64.0	37.3	48.3
Sex at birth	**Female**	43.8	52.4	50.0	34.4	54.7	39.0	31.4
	Male	56.2	47.6	50.0	65.6	45.3	61.0	68.6
Severity	**Deceased**	3.1	0.5	1.5	3.1	0.0	5.1	10.2
	Severe	31.5	12.7	19.2	33.8	37.3	39.0	64.4
	Moderate	51.2	65.1	56.9	47.5	58.7	52.5	22.0
	Mild	11.5	19.6	17.7	13.1	1.3	1.7	0.8
	Unknown	2.7	2.1	4.6	2.5	2.7	1.7	2.5
Oxygen therapy	**Yes**	66.6	45.0	48.5	68.1	92.0	78.0	97.5
	No	22.8	45.0	38.5	11.2	5.3	13.6	1.7
	Unknown	10.5	10.1	13.1	20.6	2.7	8.5	0.8
Ventilatory support	**Yes**	65.1	43.4	47.7	68.1	86.7	74.6	96.6
	No	11.4	11.6	16.9	16.2	4.0	13.6	1.7
	Unknown	23.5	45.0	35.4	15.6	9.3	11.9	1.7
Admission to intensive care unit	**Yes**	32.0	12.2	20.0	30.6	30.7	37.3	77.1
	No	65.8	84.7	77.7	66.2	68.0	62.7	22.0
	Unknown	2.2	3.2	2.3	3.1	1.3	0.0	0.8

392

393

Table 2: Summary of the characteristics of each endophenotype.
In this table, High (Low), denoted as H (L) implies that the average value of the variable in the corresponding EP was significantly higher (lower) than the other EPs (considered together), while N (Nondescript) implies that it was not significantly different.

Endophenotype	Age	Sex at birth	BMI	Blood markers
EP1	H	F	N	High lymphocyte, Low neutrophil
EP2	N	N	L	High albumin, Low white blood cells
EP3	L	M	N	High platelet, Low creatinine
EP4	N	N	N	High lactate
EP5	H	N	N	High creatinine, Low haemoglobin
EP6	N	M	N	High white blood cells, Low lymphocyte

Abbreviations used: H = High, L = Low, N = Nondescript, F = Female, M = Male

401 **References**

402 1. Hull DL. Informal aspects of theory reduction. PSA: Proceedings of the biennial meeting
403 of the philosophy of science association: Cambridge University Press; 1974. p. 653-670.

404 2. Te Pas MFW, Madsen O, Calus MPL, Smits MA. The Importance of Endophenotypes to
405 Evaluate the Relationship between Genotype and External Phenotype. *International*
406 *Journal of Molecular Sciences* 2017; 18: 472.

407 3. Blatti III C, Emad A, Berry MJ, Gatzke L, Epstein M, Lanier D, et al. Knowledge-guided
408 analysis of "omics" data using the KnowEnG cloud platform. *PLoS biology* 2020; 18:
409 e3000583.

410 4. Emad A, Ray T, Jensen TW, Parat M, Natrajan R, Sinha S, et al. Superior breast cancer
411 metastasis risk stratification using an epithelial-mesenchymal-amoeboid transition gene
412 signature. *Breast Cancer Research* 2020; 22: 74.

413 5. Al-Hadrawi DS, Al-Rubaye HT, Almulla AF, Al-Hakeim HK, Maes M. Lowered oxygen
414 saturation and increased body temperature in acute COVID-19 largely predict chronic
415 fatigue syndrome and affective symptoms due to Long COVID: A precision nomothetic
416 approach. *Acta Neuropsychiatr* 2022: 1-12.

417 6. Tremblay K, Rousseau S, Zawati MnH, Auld D, Chassé M, Coderre D, et al. The Biobanque
418 québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and
419 biological determinants of COVID-19 clinical trajectories. *PloS one* 2021; 16: e0245031.

420 7. Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody E, et al. Aptamer-based multiplexed
421 proteomic technology for biomarker discovery. *Nature Precedings* 2010: 1-1.

422 8. Ward Jr JH. Hierarchical grouping to optimize an objective function. *Journal of the*
423 *American statistical association* 1963; 58: 236-244.

424 9. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which
425 Algorithms Implement Ward's Criterion? *Journal of Classification* 2014; 31: 274-295.

426 10. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of*
427 *the American Statistical Association* 1958; 53: 457-481.

428 11. Oda S, Hirasawa H, Shiga H, Nakanishi K, Matsuda K, Nakamua M. Sequential
429 measurement of IL-6 blood levels in patients with systemic inflammatory response
430 syndrome (SIRS)/sepsis. *Cytokine* 2005; 29: 169-75.

431 12. Liu F, Li L, Xu M, Wu J, Luo D, Zhu Y, et al. Prognostic value of interleukin-6, C-reactive
432 protein, and procalcitonin in patients with COVID-19. *J Clin Virol* 2020; 127: 104370.

433 13. McConnell MJ, Kawaguchi N, Kondo R, Sonzogni A, Licini L, Valle C, et al. Liver injury in
434 COVID-19 and IL-6 trans-signaling-induced endotheliopathy. *J Hepatol* 2021; 75: 647-658.

435 14. Ruscitti P, Berardicurti O, Di Benedetto P, Cipriani P, Iagnocco A, Shoenfeld Y, et al. Severe
436 COVID-19, Another Piece in the Puzzle of the Hyperferritinemic Syndrome. An
437 Immunomodulatory Perspective to Alleviate the Storm. *Front Immunol* 2020; 11: 1130.

438 15. Ding J, Hostallero DE, El Khili MR, Fonseca GJ, Milette S, Noorah N, et al. A network-
439 informed analysis of SARS-CoV-2 and hemophagocytic lymphohistiocytosis genes'
440 interactions points to Neutrophil extracellular traps as mediators of thrombosis in COVID-
441 19. *PLoS Computational Biology* 2021; 17: e1008810.

442 16. Greco M, Suppressa S, Lazzari RA, Sicuro F, Catanese C, Lobreglio G. sFlt-1 and CA 15.3 are
443 indicators of endothelial damage and pulmonary fibrosis in SARS-CoV-2 infection. *Sci Rep*
444 2021; 11: 19979.

445 17. Corteville DC, Bibbins-Domingo K, Wu AH, Ali S, Schiller NB, Whooley MA. N-terminal pro-
446 B-type natriuretic peptide as a diagnostic test for ventricular dysfunction in patients with
447 coronary disease: data from the heart and soul study. *Arch Intern Med* 2007; 167: 483-9.

448 18. Adela R, Banerjee SK. GDF-15 as a Target and Biomarker for Diabetes and Cardiovascular
449 Diseases: A Translational Prospective. *J Diabetes Res* 2015; 2015: 490842.

450 19. Pandey A, Patel KV, Vongpatanasin W, Ayers C, Berry JD, Mentz RJ, et al. Incorporation of
451 Biomarkers Into Risk Assessment for Allocation of Antihypertensive Medication According
452 to the 2017 ACC/AHA High Blood Pressure Guideline: A Pooled Cohort Analysis.
453 *Circulation* 2019; 140: 2076-2088.

454 20. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a "kneedle" in a haystack: Detecting
455 knee points in system behavior. 31st international conference on distributed computing
456 systems workshops: IEEE; 2011. p. 166-171.

457