Gene Expression Risk Scores for COVID-19 Illness Severity

Derick R. Peterson,1 Andrea M. Baran,1 Soumyaroop Bhattacharya,2,6 Angela R. Branche,3 Daniel P. Croft,4 Anthony M. Corbett,1 Edward E. Walsh,1,5 Ann R. Falsey,1,5 and Thomas J. Mariani2
1Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, USA, 2Division of Neonatology and Pediatric Molecular and Personalized Medicine Program, Department of Pediatrics, University of Rochester, Rochester, New York, USA, 3Division of Infectious Diseases, Department of Medicine, University of Rochester, Rochester, New York, USA, 4Division of Pulmonary and Critical Care, Department of Medicine, University of Rochester, Rochester, New York, USA, and 5Department of Medicine, Rochester General Hospital, Rochester, New York, USA

**Background.** The correlates of coronavirus disease 2019 (COVID-19) illness severity following infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) are incompletely understood.

**Methods.** We assessed peripheral blood gene expression in 53 adults with confirmed SARS-CoV-2 infection clinically adjudicated as having mild, moderate, or severe disease. Supervised principal components analysis was used to build a weighted gene expression risk score (WGERS) to discriminate between severe and nonsevere COVID-19.

**Results.** Gene expression patterns in participants with mild and moderate illness were similar, but significantly different from severe illness. When comparing severe versus nonsevere illness, we identified >4000 genes differentially expressed (false discovery rate < 0.05). Biological pathways increased in severe COVID-19 were associated with platelet activation and coagulation, and those significantly decreased with T-cell signaling and differentiation. A WGERS based on 18 genes distinguished severe illness in our training cohort (cross-validated receiver operating characteristic-area under the curve [ROC-AUC] = 0.98), and need for intensive care in an independent cohort (ROC-AUC = 0.85). Dichotomizing the WGERS yielded 100% sensitivity and 85% specificity for classifying severe illness in our training cohort, and 84% sensitivity and 74% specificity for defining the need for intensive care in the validation cohort.

**Conclusions.** These data suggest that gene expression classifiers may provide clinical utility as predictors of COVID-19 illness severity.

**Keywords.** COVID-19; severity; RNA sequencing; ICU; prediction; molecular markers.

In December 2019 a novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was identified in China as a cause of severe pneumonia with explosive human-to-human transmission [1]. Illness due to SARS-CoV-2 has been designated coronavirus disease 2019 (COVID-19), and on 11 March 2020, the World Health Organization officially declared SARS-CoV-2 a pandemic. To date there have been over 240 million infections and over 5 million deaths globally due to COVID-19 [2]. Although most patients experience mild to moderate disease, 5%–10% progress to severe or critical illness with severe pneumonia or respiratory failure [3, 4]. Early in the pandemic it became clear that certain underlying chronic medical conditions, and principally age, were key risk factors for severe disease [5, 6]. While severe disease can occur early in illness, a distinct progression to severe illness occurs in some individuals 7–12 days after symptom onset, suggesting transition from a viral phase to an inflammatory phase [7]. In addition, some young individuals without comorbidities have developed severe illness, highlighting the incomplete understanding of disease pathogenesis due to SARS-CoV-2 infection [8].

Gene expression provides an unbiased measure of the host response to a pathogen on a cellular level. We and others have previously demonstrated the potential for peripheral blood gene expression patterns to classify the ontogeny and severity of viral respiratory illness [9, 10]. We hypothesized that analysis of gene expression in the blood of patients with SARS-CoV-2–related COVID-19 might help identify those at greatest risk for severe symptoms and in need of intensive care. Gene expression analysis might also identify pathways underlying disease pathogenesis and suggest new targets amenable to potential therapeutic interventions.

**METHODS**

**Acute Illness Evaluation**

Adults ≥18 years of age, either hospitalized or community recruited, exhibiting COVID-19 symptoms and documented to
have SAR-CoV-2 by polymerase chain reaction (PCR) were eligible for the study. Participants with immunosuppression or symptoms onset greater than 28 days prior to admission were excluded. Hospitalized participants were assessed within 24 hours of admission and outpatients were brought to the clinic within 1–2 days of being identified as SARS-CoV-2 positive. Demographic, clinical, radiographic, and laboratory information, date of symptom onset, and signs and symptoms of the illness were collected. Medication use was recorded with attention to drugs that may affect transcriptional profiling.

Clinical Severity Assessment
Severity of disease in COVID-19 participants at enrollment and throughout the illness was assessed using a combination of clinical variables as well as the National Early Warning Score (NEWS) of 7 graded physiological measurements (respiratory rate, oxygen saturation, oxygen supplementation, temperature, blood pressure, heart rate, and level of consciousness) [11]. Severe illness was defined as requiring any of the following: intensive care unit (ICU) care, high-flow oxygen, ventilator support, presser support, or evidence of new end organ failure. Nonsevere illness was defined as illnesses not meeting severe criteria. In addition, a panel of 4 physicians (3 infectious disease and 1 pulmonary critical care) adjudicated all nonsevere illnesses and categorized them as mild or moderate using the NEWS as well as symptoms and physiologic parameters in the context of underlying diseases and baseline oxygen requirements. Participants were followed for the duration of hospitalization and illness, and outcomes were recorded as the highest level of care required or death.

Sample Collection and Processing
Approximately 3 mL of whole blood was collected in a Tempus Blood RNA Tube at the time of enrollment and stored at −80°C until the time of processing. The median time from symptom onset to blood collection ranged from 4 to 9 days, as shown in Table 1. Following centrifugation, RNA was isolated from the pellet using the Tempus Spin RNA Isolation Kit using the manufacturer recommended protocol. Total RNA was processed for globin reduction using GLOBINclear Human Kit as described previously [10].

RNA Sequencing
cDNA libraries were generated using 200 ng of globin-reduced total RNA. Library construction was performed using the TruSeq Stranded mRNA library kit (Illumina). cDNA quantity was determined with the Qubit Fluorometer (Life Technologies) and quality was assessed using the Agilent Bioanalyzer 2100 (Agilent). Libraries were sequenced on the Illumina NovaSeq6000 at a target read depth of approximately 20 million 1 × 100-bp single-end reads per sample. Sequences were aligned against the human genome version hg38 using the splice transcript alignment to a reference (STAR) algorithm [12], and counts were generated using HTSeq [13]. Raw counts were divided by participant-specific library size (in millions) to yield counts per million (CPM)-normalized expression, borrowing no information across participants, and gene- and sample-level filtering was performed to remove outlier samples and low-expressing genes. Normalized and filtered analytical data sets were log, transformed (after adding a pseudocount of 1 CPM) prior to analysis. We excluded data from 19 861 genes with uniformly zero reads, leaving a data set comprising 39 225 genes from 53 participants. Finally, we retained genes that had normalized counts exceeding 1 CPM in greater than 14 participants (the smallest class size). This resulted in an analytical dataset of 14 228 CPM-normalized genes.

Statistical Methods
Continuous clinical variables were compared by COVID-19 severity levels using the nonparametric Kruskal-Wallis test, and binary variables by Fisher exact test. WDifferential expression by COVID-19 severity was assessed using the nonparametric Wilcoxon rank sum test. To allow adjustment for important clinical covariates, we fit semiparametric Cox proportional hazards models for normalized gene expression as a function of severe versus nonsevere COVID-19, adjusted for race, sex, body mass index (BMI), days since symptom onset, and library size. The Benjamini-Hochberg procedure was used to control the false discovery rate (FDR). Pathway analysis of significantly differentially expressed genes was performed using ENRICHR [14]. While negative binomial regression is a useful generalized linear model (GLM) for the log of the mean of raw count data, assuming the variance of the count is a quadratic function of its mean, we used a more flexible semiparametric Cox model, which can be viewed as a GLM with complementary log-log link. The Cox model makes no assumption about the shape of the distribution of (normalized) counts, and it is invariant to monotonic transformations of the response because it only depends on the ranks. The familiar nonparametric rank-based Wilcoxon and Kruskal-Wallis tests we used are closely related to the logrank test, a special case of the Cox model.

Classifier Development and Testing
We used a version of supervised principal components analysis to build a weighted gene expression risk score (WGERS) to discriminate between severe and nonsevere COVID-19. Genes were selected based on their univariate area under the curve (AUC) and fold-change, as estimated by Hodges-Lehmann median of all pairwise shifts in log expression, with thresholds selected within the inner loop of nested 20-fold cross-validation. The cross-validation sampling strategy was designed to efficiently approximate leaving out 1 subject from each of the 3 sublevels of severity. The selected genes were standardized to Z-scores with mean 0 and unit variance, and their first PC score was used as the sole predictor for logistic
regression. The outer loop of nested cross-validation was used to estimate the receiver operating characteristic (ROC) and AUC of the adaptive procedure. The nested pooled AUC corresponds with the ROC curve, and compares samples across and within models. The nested stratified AUC only compares samples from the same model, and thus generally is preferable, but it does not correspond with any single ROC curve. A subsequent run of nonnested cross-validation produced the thresholds used to define the gene set for the final risk score. The WGERS is calculated as the linear combination of the standardized, log$_2$-transformed genes that meet the chosen thresholds, with coefficients based on the first principal component loading of the genes, scaled by the coefficient from the univariate logistic regression model. The outer loop of nested cross-validation was used to estimate the ROC and area under the ROC curve (AUC) or the adaptive procedure.

| Variable                                | Mild $(n = 14)$ | Moderate $(n = 19)$ | Severe $(n = 20)$ | $P$ Value |
|-----------------------------------------|----------------|-------------------|------------------|-----------|
| Demographics                            |                |                   |                  |           |
| Age, y, median (IQR)                    | 63.0 (41.0)    | 59.0 (29.0)       | 63.5 (19.5)      | .71       |
| Male sex, No. (%)                       | 5 (35.7)       | 11 (57.9)         | 12 (60.0)        | .37       |
| White non-Hispanic race, No. (%)        | 13 (92.9)      | 11 (57.9)         | 10 (50.0)        | .02       |
| BMI, median (IQR)                       | 28.7 (9.2)     | 30.4 (11.7)       | 26.4 (9.5)       | .29       |
| Days from symptom onset, median (IQR)   | 4.0 (5.0)      | 9.0 (6.0)         | 6.5 (4.5)        | .05       |
| Underlying conditions, No. (%)          |                |                   |                  |           |
| COPD                                    | 3 (21.4)       | 6 (31.6)          | 2 (10.0)         | .25       |
| CHF                                     | 4 (28.6)       | 1 (5.3)           | 2 (10.0)         | .17       |
| Diabetes                                | 2 (14.3)       | 8 (42.1)          | 5 (25.0)         | .20       |
| Hypertension                            | 7 (50.0)       | 11 (57.9)         | 12 (60.0)        | .88       |
| Asthma                                  | 0 (0.0)        | 1 (5.3)           | 2 (10.0)         | .77       |
| Symptoms, No. (%)                       |                |                   |                  |           |
| Cough                                   | 6 (42.9)       | 15 (78.9)         | 11 (55.0)        | .10       |
| Fever                                   | 6 (42.9)       | 15 (78.9)         | 15 (75.0)        | .08       |
| Dyspnea                                 | 4 (28.6)       | 12 (63.2)         | 13 (65.0)        | .08       |
| Rigors                                  | 0 (0.0)        | 3 (15.8)          | 0 (0.0)          | .06       |
| Physical findings                       |                |                   |                  |           |
| Systolic blood pressure, median (IQR)   | 127 (27)       | 119 (34)          | 105 (33)         | .07       |
| Oxygen saturation, median (IQR)         | 96.5 (4.5)     | 90.0 (5.0)        | 85.5 (11.5)      | .0005     |
| Laboratory data                         |                |                   |                  |           |
| Infiltrate on chest radiograph, %, (N)  | 42.8$^{a}$     | 88.8$^{b}$        | 100              | .0008     |
| C-reactive protein, mg/L, median (IQR)  | 29.6 (45.5)$^{a}$ | 107.0 (126.9)$^{d}$ | 155.4 (43.7)      | .0004     |
| D-dimer, ng/mL, median (IQR)            | 2551 (2891)$^{a}$ | 724 (420)$^{a}$   | 1209 (1023)        | .002      |
| Absolute lymphocyte count, /μL × 1000, median (IQR) | 0.9 (0.8)$^{a}$ | 1.2 (0.7)$^{d}$   | 0.5 (0.5)         | .002      |
| Level of care, No. (%)                  |                |                   |                  |           |
| Intensive care                          | 0 (0)          | 0 (0)             | 20 (100.0)       | <.0001    |
| Pressors                                | 0 (0)          | 0 (0)             | 13 (65.0)        | <.0001    |
| Worst outcome, No. (%)                  |                |                   |                  |           |
| Low-flow oxygen supplementation (~0–10 L) | 4 (28.6)   | 12 (63.2)         | 0 (0)            |           |
| Intermediate-flow oxygen supplementation (~10–20 L) | 0 (0) | 2 (10.5) | 1 (5.0) | |
| High-flow oxygen supplementation (~20–60 L) | 0 (0) | 0 (0) | 5 (25.0) | |
| Noninvasive positive pressure ventilation | 0 (0) | 0 (0) | 0 (0) | |
| Mechanical ventilation                  | 0 (0)          | 0 (0)             | 12 (60.0)        |           |
| Extracorporeal membrane oxygenation     | 0 (0)          | 0 (0)             | 1 (5)            |           |
| Death                                   | 0 (0)          | 0 (0)             | 1 (5)            |           |
| None of the above                       | 10 (71.4)      | 5 (26.3)          | 0 (0)            |           |

Continuous variables were compared using the nonparametric Kruskal-Wallis test, while categorical variables were compared using Fisher exact test. Data for laboratory variables include data from fewer than the entire cohort, as indicated.

Abbreviations: BMI, body mass index; CHF, chronic heart failure; COPD, chronic obstructive pulmonary disease; IQR, interquartile range.

*a* $n = 13$.

*b* $n = 12$.

*c* $n = 7$.

*d* $n = 18$.

*e* $n = 6$. 
AUC represents the probability that a severe subject has a higher WGERS than a nonsevere subject. The nested pooled AUC corresponds with the ROC curve, and compares samples across and within models.

To perform an independent validation of our risk score, we use a dataset from Overmyer et al [15], which had a different definition of severity in the outcome (ICU vs non-ICU), and used a different normalization for the gene expression data (TPM). Of the 18 genes used in our risk score, 2 were missing in the validation data. We imputed data for these 2 genes via multiple linear regression with coefficients estimated by regressing each on the 16 nonmissing CPM-normalized log gene expression values in the training data. We standardized the TMP-normalized validation gene expression data using means and SDs estimated from the training data, and then applied the risk score coefficients from the training data to construct a risk score for each validation subject. Apparent miscalibration required choosing a different WGERS threshold for the validation data due to gene expression measures being generally lower in the validation data compared to the training data. An ROC curve with associated AUC was used to assess the performance of the risk score in the validation data.

RESULTS

Between 30 April and 29 June 2020, 58 participants with PCR-documented COVID-19 illnesses were enrolled from inpatient and outpatient settings. Of these, 3 participants did not have blood samples collected and 2 did not meet inclusion criteria, leaving 53 participants for RNA sequencing analysis. Illnesses were adjudicated as 20 severe and 33 nonsevere (14 mild and 19 moderate). This categorization was consistent with the severity separation in the NEWS (Supplementary Figure 1). Two severely ill participants received 1 dose of remdesivir prior to blood collection. No subject received steroids or any other experimental COVID-19 treatment prior to enrollment. Five hospitalized participants had rapidly progressive hypoxemia and hemodynamic instability after enrollment and required transfer to intensive care, and 3 subsequently were mechanically ventilated. No mildly ill outpatient illnesses progressed in severity to require medical attention. There was insufficient evidence of any difference in demographic characteristics or underlying conditions by disease severity, except for race and time from disease onset (Table 1): white non-Hispanic comprised 93% mild versus 50%–58% moderate-severe (P = .02), and median time from symptom onset to enrollment was 4 days among mild, 9 days among moderate, and 6.5 days among severe disease cases (P = .047, due to heterogeneity among nonsevere). The median age of participants was 62 years with 53% of them being male. As expected, dyspnea, hypoxemia, the presence of infiltrates, and use of supplemental oxygen were more common in moderate and severe, compared to mild illness. All severely ill patients required intensive care; 15 were enrolled in the ICU and 5 were moved to ICU within 48 hours of blood sampling. All severely ill participants required supplemental oxygen; 12 (60%) were mechanically ventilated, 1 was supported with extracorporeal membrane oxygenation and survived, 13 (65%) required vasopressor support, and 1 subject died. Median NEWS were different between the 3 groups (Supplementary Figure 1). Inflammatory markers were not available for most outpatients but were notably elevated in those hospitalized with moderate to severe disease (Table 1).

Blood gene expression profiling from SARS-CoV-2–positive cases (n = 53) was completed by standard mRNA sequencing of globin mRNA-reduced RNA isolated from whole blood at the time of recruitment. On average 58 (SD 6) million reads were generated from each of the cDNA libraries, with a mapping rate of 94.2% (SD 0.6%) and transcriptome coverage of 41.3% (SD 1.3%) (Supplementary Figure 2). Exploratory principal components analysis suggested similar patterns of gene expression might be shared by participants with mild and moderate illness, but appeared distinct from those with severe illness (Figure 1A). Statistical analysis for differential gene expression confirmed significant differences when comparing mild versus severe, and moderate versus severe, but not mild versus moderate COVID-19 (Figure 1B).

We next tested for differences in gene expression when comparing participants with severe (n = 20) versus nonsevere illness (n = 33), pooling the 14 mild and 19 moderate cases. We tested for differential gene expression without (univariate) and with adjustment for variables potentially associated with severe outcome (race, sex, BMI), the number of days since onset of symptoms, and library size (Figure 1C and Supplementary Table 1). These analyses identified 6483 (46% of tested) and 8435 (59% of tested) differentially expressed genes, with and without multivariate adjustment, respectively.

We performed ontology analysis for the 6483 genes identified as differentially expressed in severe COVID-19 illness, focusing on the fully adjusted analysis (Figure 2). This analysis identified 74 pathways overrepresented by genes (n = 936) significantly upregulated in severe COVID-19, and 25 pathways overrepresented by genes (n = 5547) significantly downregulated (Figure 2 and Supplementary Table 2). Activated pathways included a number associated with infectious diseases as well as tumor necrosis factor-α (TNF-α) and nuclear factor-κB (NF-κB) signaling. Notably, there was also evidence for significant upregulation of genes associated with platelet activation and coagulation. Among pathways associated with downregulated genes in severe COVID-19 were multiple pathways involved in general host RNA metabolism as well as multiple pathways specifically associated with T-cell regulation, including T helper 2 (Th2) and Th17 differentiation. The most significantly downregulated pathway was associated with herpes simplex virus 1 (HSV1) infection.
Given the substantial number of differentially expressed genes when comparing severe versus nonsevere COVID-19, we investigated the ability of gene expression patterns to discriminate severe illness. Gene-specific thresholds for univariate AUC and magnitude change were chosen via the cross-validation procedure and used to produce an 18-gene WGERS for severe illness. Nested cross-validation was used to estimate performance via the stratified AUC (CV-AUC = 0.98). The pooled CV-AUC of 0.93 corresponds with a cross-validated ROC curve to graphically summarize performance (Figure 3A). The pooled CV-ROC curve also was used to select a risk score threshold (~1.04) with 95% sensitivity and 88% specificity, which corresponded with apparent (non–cross-validated) sensitivity of 100%, specificity of 85%, and error rate of 9% (5/53), represented via the WGERS distributions for the training data (Figure 4A). All 5 misclassified participants had moderate illness (Figure 4B).

We next identified an independent validation data set describing peripheral blood-based gene expression profiling of COVID-19 subjects who were either admitted (n = 50) to the ICU due to the severity of their acute illness [15]. Our 18-gene WGERS discriminated between ICU and non-ICU patients with an AUC of 0.85, and thresholding at 1.77 yielded 84% sensitivity and 74% specificity (Figures 3B and Figure 4C). Furthermore, all 18 genes selected in the training data were differentially expressed (FDR < 0.01) in the validation data (Supplementary Table 3).

**DISCUSSION**

SARS-CoV-2 infection causes a wide spectrum of disease ranging from minimal, often asymptomatic, respiratory illness to severe pneumonia with multisystem failure and death. Although measurements of inflammatory markers such as C-reactive protein and serum interleukin-6 (IL-6) levels are often associated with worse disease, their use to predict poor outcomes is imperfect [16–18]. Viral characteristics, such as shedding kinetics or gene sequence variation, are not reliable predictors of clinical outcome [19, 20]. Genome-wide expression profiling, a powerful and unbiased tool, can be used for multiple purposes such as relating activation or suppression.
Figure 2. Biological interpretation of gene expression patterns in severe versus nonsevere coronavirus disease 2019 (COVID-19). A and B, Pathway analysis of genes differentially expressed between severe and nonsevere COVID-19 participants. Genes identified as overexpressed (n = 936) and underexpressed (n = 5547) in cases of severe compared to nonsevere COVID-19 were used for pathway analysis using ENRICHR. With 936 genes overexpressed in severe COVID-19 cases, ENRICHR (through KEGG Pathways database) identified 74 pathways associated with COVID-19 severity, while with 5547 genes underexpressed in severe COVID-19, it identified 25 pathways. Shown here are the top 25 significant pathways (P < 0.05) associated with upregulated (A) and downregulated genes (B). The bar size represents the frequency of the pathway genes differentially expressed in severe COVID-19; red indicates upregulation and green indicates downregulation. The size of the dots is proportional to −log(P value). Larger dots represent lower P values. C, Differential expression analysis of severe and nonsevere COVID-19 participants identified 6483 genes as significantly different. Shown here is a heatmap of the top 425 differentially expressed genes, where the rows indicate genes and columns indicate participants. High expression is shown in red and low expression in green.
of molecular pathways to clinical manifestations of disease, identification of biomarkers that may allow individual prediction of disease severity, and identification of novel gene targets for therapeutic intervention. Early predictors to identify patients that will decompensate following SARS-CoV-2 infection would be highly impactful.

The goal of our study is to use gene expression analysis to identify peripheral markers pathways associated with COVID-19 severity, which may serve as predictors of disease severity potential therapeutic targets. In this study of 53 SARS-CoV-2–infected adults with illness ranging from very mild upper respiratory infection to acute respiratory failure, we identified >6000 differentially expressed genes (FDR < 0.05) between severe and nonsevere illness. The vast majority (85%) of differentially expressed genes were underexpressed, most notably with a marked effect on lymphocytes and altered function [21, 22]. Pathway analysis revealed inhibition of Th1, Th2, and Th17 cell differentiation, as well as inhibition of the T-cell receptor signaling pathway. These effects are likely related to the marked lymphopenia and poor adaptive immune response in persons with severe SARS-CoV-2 infection [23]. Also notable in severe illness was the inhibition of the mRNA surveillance pathways that include the nonsense-mediated mRNA decay pathway, which can degrade viral mRNA. Using a model coronavirus, murine hepatitis virus, Wada and colleagues showed viral transcription is enhanced by blocking this host cell pathway, demonstrated to be mediated by the viral nucleocapsid protein [24].

Several activated pathways we identified in our studies are worth comment, given what is already known about SARS-CoV-2 and COVID-19. Activation of the NF-κB and TNF signaling pathways in a setting of heightened inflammatory process is not surprising. Activation of the platelet, complement, and coagulation cascade pathways is also expected, given the characteristic hypercoagulable state that has been observed in severe illness [25]. Thrombocytopenia and activated platelets are associated with the high incidence of venous and arterial clotting, while elevated levels of serum D-dimer, a fibrinogen degradation product, and increased international normalized ratio (INR) are all features of severe COVID-19 [26]. It is interesting that the infection-related pathways most significantly activated include those principally associated with intracellular bacterial (Legionella, mycobacterial) and parasitic (Toxoplasma, Leishmania, and Trypanosoma (Chagas)) infections. These infections are associated with marked activation of macrophages, and thus may be consistent with activation of the osteoclast differentiation pathway, as osteoclasts and macrophages have many similarities [27, 28].

Our findings are generally consistent with the limited data currently available in the literature on COVID-19 and gene
Specifically, Overmyer et al reported gene expression and metabolomic data from 128 persons with and without COVID-19, where 219 molecular features with high significance to COVID-19 status and severity were discovered [15]. A number of these involved complement activation, dysregulated lipid transport, and neutrophil activation. Additionally, our data are supported by the findings noted by Ouyang et al in which Th17 and T-cell activation and differentiation were markedly downregulated in severe disease [33].

There were some novel pathways that demonstrated upregulation in our studies. A number of malignancy-related pathways were upregulated (acute myelogenous leukemia [AML], colorectal, pancreatic, and non–small cell cancer) in the severe COVID-19 patients. Similarly, in a study by Kwan et al comparing gene expression in 45 COVID-19 cases to healthy controls, 135 genes were found to be differentially expressed, with enrichment for several cancer pathways, including viral carcinogenesis and AML [30]. Identification of mutations in cell signaling, proliferation genes, and kinases such as AP2-associated protein kinase 1 (AAK1) have led to targeted treatment options for cancer patients. Baricitinib, a repurposed rheumatoid arthritis drug that interferes with the Janus kinase (JAK) pathway, was demonstrated to have efficacy when combined with remdesivir in the treatment of severe COVID-19 [34]. Baricitinib shows high affinity for AAK1 binding, potentially demonstrating some overlap in perturbations of cell signaling pathways in malignancy and COVID-19. Further investigation of gene expression pathways differentially expressed in severely ill patients may provide clues to new therapeutic targets.

Although our study was not designed to identify and validate early predictors of severe disease, the data do offer a first step. Using gene expression data, we were able develop and validate an 18-gene signature for severe disease—fully concordant with requiring ICU—with 85% AUC, 84% sensitivity, and 74% specificity in an independent validation data set. In a recent paper, Juan Guardela et al assessed the utility of blood transcript levels of 50 genes known to predict mortality in idiopathic pulmonary fibrosis patients to classify illness severity in COVID-19 [32]. A discovery cohort of 8 subjects was used, and then validated using a publicly available data set of 128 subjects [15]. The gene expression risk profile discriminated ICU admission, need for mechanical ventilation, and in-hospital mortality with an AUC of 77%, 75%, and 74%, respectively (P < .001) in a COVID-19 validation cohort.

Our current study has several limitations that are worth noting, including its relatively small sample size, the nonstandardized interval between symptom onset and sample collection, and blood collection at 1 timepoint. The complexity of the clinical data among hospitalized participants (ie, admissions only for isolation, persons with chronic oxygen requirements, COVID-19 testing for procedures) made objective criteria to distinguish mild from moderate disease difficult, necessitating the need for clinical adjudication. Lastly, certain laboratory studies were not available for all subjects.

**CONCLUSIONS**

In summary, we found a large number of differentially expressed genes in the peripheral blood that distinguished those with severe COVID-19 illness from those with mild or moderate disease. These data could be used to identify potential...
targets for interventions, as well as to develop predictors of disease severity. Future prospective studies are needed to follow mild to moderately ill patients over time and evaluate whether any of the discriminatory genes identified are affected at early stages and can serve as predictors of severity. If so, individuals with high-risk gene profiles might be hospitalized for observation, moved to a more closely monitored setting while hospitalized, or targeted for early interventions such as monoclonal antibody treatments.

Supplementary Data
Supplementary materials are available at The Journal of Infectious Diseases online. Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes
Acknowledgments. We acknowledge the University of Rochester Genomics Research Center for completing RNA isolation and sequencing. Christopher Slaunwhite, Mary Anne Formica, and Michael Peasley provided technical support. Jeanne Holden-Whiltse and Jeffrey Williams assisted with data management.

Author contributions. D. R. P., E. E. W., A. R. F., and T. J. M. conceptualized the study. D. P. C., A. R. B., E. E. W., and A. R. F. recruited patients, adjudicated clinical severity, and collected samples. D. R. P., A. M. B., S. B., A. M. C., and T. J. M. analyzed the data. D. R. P., A. M. B., S. B., A. R. B., D. P. C., E. E. W., T. J. M., and A. R. F. interpreted the results. All authors contributed to writing of the manuscript. All authors reviewed, edited, and approved the manuscript for submission.

Data availability: The raw sequence and normalized data are currently being deposited to the database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/). The accession number for this data series will be provided as soon as the submission is approved. In the meantime, summary counts data will be available upon request to the corresponding author.

Financial support. This work was supported by the National Institute of Environmental Health Sciences (grant numbers AI137364 to T. J. M. and A. R. F., and K23 ES032459 to D. C.).

Potential conflicts of interest. A. R. F. has received grants from Janssen, Merck, Sharpe and Dohme, Pfizer, and BioFire Diagnostics; and personal fees for serving on DSMB for Novavax. E. E. W. has consulted for Janssen Pharmaceuticals; and received research funding from Janssen, Gilead, Medimmune, Sanofi Pasteur, and ADMA biologics. A. R. B. consults for GSK; and has received grants from Pfizer, Merck, Janssen, and Cyanvac. All other authors report no potential conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References
1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020; 382:727–33.
2. World Health Organization. WHO coronavirus (COVID-19) dashboard. https://covid19.who.int/.
3. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020; 382:1708–20.
4. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020; 395:507–13.
5. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. JAMA 2020; 323:1239–42.
6. Booth A, Reed AB, Ponzo S, et al. Population risk factors for severe disease and mortality in COVID-19: a global systematic review and meta-analysis. PLoS One 2021; 16:e0247461.
7. Siddiqi HK, Mehra MR. COVID-19 illness in native and immunosuppressed states: a clinical-therapeutic staging proposal. J Heart Lung Transplant 2020; 39:405–7.
8. Owusu D, Kim L, O’Halloran A, et al. Characteristics of adults aged 18–49 years without underlying conditions hospitalized with laboratory-confirmed coronavirus disease 2019 in the United States: COVID-NET March-August 2020. Clin Infect Dis 2021; 72:e162–e6.
9. Wang L, Chu CY, McCall MN, et al. Airway gene-expression classifiers for respiratory syncytial virus (RSV) disease severity in infants. BMC Med Genomics 2021; 14:57.
10. Bhattacharya S, Rosenberg AF, Peterson DR, et al. Transcriptomic biomarkers to discriminate bacterial from nonbacterial infection in adults hospitalized with respiratory illness. Sci Rep 2017; 7:6548.
11. Richardson D, Faisal M, Fiori M, Beatson K, Mohammed M. Use of the first national early warning score recorded within 24 hours of admission to estimate the risk of in-hospital mortality in unplanned COVID-19 patients: a retrospective cohort study. BMJ Open 2021; 11:e043721.
12. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29:15–21.
13. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 2015; 31:166–9.
14. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 2013; 14:128.

15. Overmyer KA, Shishkova E, Miller JJ, et al. Large-scale multi-omic analysis of COVID-19 severity. Cell Syst 2021; 12:23–40.e7.

16. Bennett TD, Moffitt RA, Hajagos JG, et al; National COVID Cohort Collaborative (N3C) Consortium. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID cohort collaborative. JAMA Netw Open 2021; 4:e2116901.

17. Bentivegna M, Hulme C, Ebell MH. Primary care relevant risk factors for adverse outcomes in patients with COVID-19 infection: a systematic review. J Am Board Fam Med 2021; 34:113–26.

18. Kalil AC, Patterson TF, Mehta AK, et al; ACTT-2 Study Group Members. Baricitinib plus remdesivir for hospitalized adults with Covid-19. N Engl J Med 2021; 384:795–807.

19. Abdulrahman A, Mallah SI, Alqahtani M. COVID-19 viral load not associated with disease severity: findings from a retrospective cohort study. BMC Infect Dis 2021; 21:688.

20. Liu Y, Yan LM, Wan L, et al. Viral dynamics in mild and severe cases of COVID-19. Lancet Infect Dis 2020; 20:656–7.

21. Vigón L, Rodríguez-Mora S, Luna A, et al. Cytotoxic cell populations developed during treatment with tyrosine kinase inhibitors protect autologous CD4+ T cells from HIV-1 infection. Biochem Pharmacol 2020; 182:114203.

22. DiPiazza AT, Graham BS, Ruckwardt TJ. T cell immunity to SARS-CoV-2 following natural infection and vaccination. Biochem Biophys Res Commun 2021; 538:211–7.

23. Rydzynski Moderbacher C, Ramirez SI, Dan JM, et al. Antigen-specific adaptive immunity to SARS-CoV-2 in acute COVID-19 and associations with age and disease severity. Cell 2020; 183:996–1012.e19.

24. Wada M, Lokugamage KG, Nakagawa K, Narayanan K, Makino S. Interplay between coronavirus, a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway. Proc Natl Acad Sci U S A 2018; 115:E10157–66.

25. Ksander GA. Collagen coatings reduce the incidence of capsule contracture around soft silicone rubber implants in animals. Ann Plast Surg 1988; 20:215–24.

26. Levy JH, Iba T, Olson LB, Corey KM, Ghadimi K, Connors JM. COVID-19: thrombosis, thromboinflammation, and anticoagulation considerations. Int J Lab Hematol 2021; 43(Suppl 1):29–35.

27. Upadhyay S, Mittal E, Philips JA. Tuberculosis and the art of macrophage manipulation. Pathog Dis 2018; 76:fty037.

28. Bogdan C. Macrophages as host, effector and immunoregulatory cells in leishmaniasis: impact of tissue micro-environment and metabolism. Cytokine X 2020; 2:100041.

29. Jain R, Ramaswamy S, Harilal D, et al. Host transcriptomic profiling of COVID-19 patients with mild, moderate, and severe clinical outcomes. Comput Struct Biotechnol J 2021; 19:153–60.

30. Kwan PKW, Cross GB, Naftalin CM, et al. A blood RNA transcriptome signature for COVID-19. BMC Med Genomics 2021; 14:155.

31. Li CX, Chen J, Lv SK, Li JH, Li LL, Hu X. Whole-transcriptome RNA sequencing reveals significant differentially expressed mRNAs, miRNAs, and lncRNAs and related regulating biological pathways in the peripheral blood of COVID-19 patients. Mediators Inflamm 2021; 2021:6635925.

32. Juan Guardela BM, Sun J, Zhang T, et al. 50-gene risk profiles in peripheral blood predict COVID-19 outcomes: a retrospective, multicenter cohort study. EBioMedicine 2021; 69:103439.

33. Ouyang Y, Yin J, Wang W, et al. Downregulated gene expression spectrum and immune responses changed during the disease progression in patients with COVID-19. Clin Infect Dis 2020; 71:2052–60.

34. Richardson P, Griffin I, Tucker C, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet 2020; 395:e30–1.