Journal of the American Medical Informatics Association, 29(5), 2022, 864–872
https://doi.org/10.1093/jamia/ocac012
Advance Access Publication Date: 22 February 2022
Research and Applications

Research and Applications

Dynamic modeling of hospitalized COVID-19 patients reveals disease state–dependent risk factors

Braden C. Soper[1], Jose Cadena[2], Sam Nguyen[2], Kwan Ho Ryan Chan[2], Paul Kiszka[3], Lucas Womack[3], Mark Work[3], Joan M. Duggan[4], Steven T. Haller[4], Jennifer A. Hanrahan[4], David J. Kennedy[4], Deepa Mukundan[5], and Priyadip Ray[2]

[1]Computing Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA, [2]Engineering Directorate, Lawrence Livermore National Laboratory, Livermore, California, USA, [3]Information Technology Services, ProMedica Health System, Inc, Toledo, Ohio, USA, [4]Department of Medicine, University of Toledo College of Medicine and Life Sciences, Toledo, Ohio, USA, and [5]Department of Pediatrics, University of Toledo College of Medicine and Life Sciences, Toledo, Ohio, USA

Corresponding Author: Braden C. Soper, PhD, Computing Directorate, Lawrence Livermore National Laboratory, 7000 East Ave, Livermore, CA 94550, USA; soper3@llnl.gov

Received 19 October 2021; Revised 15 December 2021; Editorial Decision 3 January 2022; Accepted 28 January 2022

ABSTRACT

Objective: The study sought to investigate the disease state–dependent risk profiles of patient demographics and medical comorbidities associated with adverse outcomes of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections.

Materials and Methods: A covariate-dependent, continuous-time hidden Markov model with 4 states (*moderate*, *severe*, *discharged*, and *deceased*) was used to model the dynamic progression of COVID-19 during the course of hospitalization. All model parameters were estimated using the electronic health records of 1362 patients from ProMedica Health System admitted between March 20, 2020 and December 29, 2020 with a positive nasopharyngeal PCR test for SARS-CoV-2. Demographic characteristics, comorbidities, vital signs, and laboratory test results were retrospectively evaluated to infer a patient's clinical progression.

Results: The association between patient-level covariates and risk of progression was found to be disease state dependent. Specifically, while being male, being Black or having a medical comorbidity were all associated with an increased risk of progressing from the *moderate* disease state to the *severe* disease state, these same factors were associated with a decreased risk of progressing from the *severe* disease state to the *deceased* state.

Discussion: Recent studies have not included analyses of the temporal progression of COVID-19, making the current study a unique modeling-based approach to understand the dynamics of COVID-19 in hospitalized patients.

Conclusion: Dynamic risk stratification models have the potential to improve clinical outcomes not only in COVID-19, but also in a myriad of other acute and chronic diseases that, to date, have largely been assessed only by static modeling techniques.

Key words: COVID-19, disease progression, risk factors, hidden Markov model, patient trajectory

INTRODUCTION

Since its emergence in late 2019, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a global pandemic with more than 5.5 million estimated deaths worldwide.[1] An understanding of risk factors influencing disease severity is critical for the efficient clinical management of COVID-19 patients. Studies have shown that risk factors, such as obesity, sex, and age, are highly correlated with adverse outcomes in COVID-19 patients.[2–7] Furthermore, recent studies suggest such risk factors also may affect certain aspects of COVID-19 progression, specifically disease onset,[8] hospital utilization,[9] and time-to-death.[10] However, the effects of individual patient characteristics on the entire course of COVID-19 progression during a patient's hospitalization is still not well characterized. A better understanding of *how individual characteristics influence not just the final outcome, but the full patient trajectory*, could lead to better care, improved patient outcomes, and improved utilization of scarce resources.

Various approaches to disease progression modeling have been proposed in the literature. These approaches range from deterministic approaches based on differential equations,[11] statistical approaches such as autoregressive models,[12] hidden Markov models,[13] and Gaussian processes,[14,15] deep learning methods such as recurrent neural networks,[16] and computational simulation methods such as discrete event simulations (DESs).[17,18] The choice of modeling approach depends on the degree of knowledge of the underlying disease mechanism, the stochasticity and heterogeneity of disease symptoms, the number of samples available for parameter estimation, and the need for model interpretability. In this article, we focus on hidden Markov models for characterizing the disease trajectories of hospitalized SARS-CoV-2 positive patients. This particular choice is motivated by several factors: (1) a general lack of understanding of the disease mechanism, (2) significant heterogeneity of disease presentation and outcomes, and (3) a modest cohort size of 1362 hospitalized patients. In addition, HMMs are fairly easy to interpret, compared to other statistical approaches such as Gaussian processes. While DES-based methods are also easy to interpret and are capable of modeling highly complex interactions, they require either informed inputs for parameter values or extensive data for model calibration.[19] Because COVID-19 is a novel and still evolving disease, neither of these requirements is currently met.

To better understand the impact of demographics and comorbidities on the disease progression of hospitalized SARS-CoV-2 positive patients, we propose a covariate-dependent, continuous-time Markov model with 4 states (*moderate*, *severe*, *discharged*, and *deceased*) to capture the dynamic progression and regression of COVID-19 during the course of hospitalization. We assume that the underlying disease states are not directly observed; rather, these states must be inferred from observational data collected throughout the course of hospitalization. Using electronic health records (EHRs) from patients in the ProMedica healthcare system in northwestern Ohio and southeastern Michigan, we propose a hidden Markov model that allows us to infer the effects of individual patient covariates on the progression and regression of COVID-19. Demographic information (e.g., age, race, sex) along with the history of 5 vital signs and 10 laboratory test results collected during hospitalization were used to train the covariate-dependent, continuous-time hidden Markov model.

Instead of only analyzing the association between patient-level covariates and a single adverse outcome, as is done in static risk-factor analysis, we seek to uncover associations between patient-level covariates and multiple adverse disease-related events. It is hypothesized that these dynamic associations will depend on the current disease state. To the best of our knowledge, this is the first comprehensive model of disease trajectory for hospitalized COVID-19 patients which integrates demographic information, comorbidities, as well as important vitals and laboratory test results. In contrast to previously published work that simply identifies static risk factors associated with adverse outcomes, we take disease severity into account which allows us to identify *when in the course of the disease progression certain patient-level covariates are associated with adverse outcomes*, such as progressing to a more severe state. We also demonstrate for the first time that the *nature of association of certain demographic variables (such as age, sex, and race) and comorbidities (such as asthma, diabetes, hypertension and kidney disease) with adverse patient outcomes can depend on the underlying disease state of the patient*.

MATERIALS AND METHODS

Several retrospective studies have analyzed the associations between various risk factors and adverse outcomes of COVID-19 patients, a number of which have been data-driven predictive modeling approaches.[2–10] The vast majority of these studies have ignored the dynamic progression and regression of COVID-19, instead relying on static data and methods. However, many infectious diseases have a natural interpretation in terms of a finite number of progressively severe disease states.[20,21] Our objective is to investigate the disease state–dependent association between patient-level covariates and risk of progression. To this end, we model hospitalized COVID-19 patient *trajectories* given standard EHR data collected throughout the course of hospitalization. This constrains our modeling choices to discrete state space models. Multi-state Markov models (MMs) and hidden Markov models (HMMs) are 2 well-known discrete state space models[22] with a long history in disease modeling. Their generality and flexibility make them attractive models for biomedical panel data, with both MMs and HMMs having been applied to a wide variety of disease progression modeling tasks. A non-exhaustive list of such works includes applications to HIV,[20] cancer progression and diagnosis,[23,24] cancer screening,[21,25] vascular disease,[26,27] pulmonary disease,[28] neurodegenerative disease,[29,30] sepsis,[31,32] and diabetes.[33]

One benefit of MMs and HMMs is that biologically plausible models can be proposed for the various disease states and the transitions between them. This is done through the use of a Markov chain or Markov jump process. HMMs bring in the additional benefit of being able to account for stochasticity in the observation process. Finally, we note that Markov jump processes are continuous-time models, as opposed to Markov chains, which are discrete time models. As such, Markov jump processes are more appropriate when dealing with irregularly sampled data with a large amount a variability in the sampling rates. For these reasons, we focused on Markov jump processes for modeling the underlying disease progression of a COVID-19 patient.

Data

The data used in this study are composed of EHRs from patients of ProMedica, the largest healthcare system in northwestern Ohio and southeastern Michigan. The patient data used in this study corresponds to patients who (1) had a positive nasopharyngeal PCR test for SARS-CoV-2 between March 20, 2020 and December 29, 2020,

and (2) were admitted to the hospital within 3 days of a positive result. This inclusion criteria ensured that hospitalization was primarily due to COVID-19-related complications. We further dropped patients with unknown discharge status and those without any recorded vital measurements or laboratory test results. A total of 1362 patients met these criteria. A detailed flow diagram showing inclusion criteria for data preprocessing is provided in Figure 1. There are 3 main sources of data available in this dataset, all of which were collected throughout the course of the patients' hospitalizations. Patient demographic information includes age, sex, body mass index (BMI), and race. Patient comorbidities considered include asthma, hypertension, diabetes, and kidney disease. Patient vital measurements (vitals from here on) used in this study are systolic blood pressure (SBP), diastolic blood pressure (DBP), respirations (Resp), temperature (Temp), and urine output (UO). Patient laboratory test results (labs from here on) used in this study are C-reactive protein (CRP), blood urea nitrogen (BUN), lactate dehydrogenase (LDH), procalcitonin, ferritin, anion-gap, D-dimer, oxygen saturation (%O_2 Sat), hemoglobin, and platelets. More details on the ProMedica dataset are summarized in the Supplementary Material.

A continuous-time hidden Markov model for COVID-19 patient data

A finite state Markov jump process S_t with state-space $\mathcal{S}$ is fully characterized by an initial state probability distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_{|\mathcal{S}|})$ over $\mathcal{S}$ and a transition intensity matrix Q that governs the rates of transitions between the states of $\mathcal{S}$. The ij entry of Q is denoted q_{ij}. The off-diagonal elements of Q are non-negative while the diagonal elements satisfy $q_{ii} = -\sum_{j \neq i} q_{ij}$. For homogeneous continuous-time Markov jump processes, the time spent in state $i \in \mathcal{S}$ is exponentially distributed with mean $\lambda_i = -1/q_{ii}$. If the process is in state i and transitions to a different state, the process enters state $j \neq i$ with probability $p_{ij} = -q_{ij}/q_{ii}$. We then say a sequence of random variables S_t is a Markov jump process if $S_0 = i$ with probability π_i for $i = 1, 2, \ldots, |\mathcal{S}|$ and the stochastic transitions are governed by the matrix Q as described above.

To capture both COVID-19 disease progression and regression, we consider a 4-state Markov jump process. Two states correspond to the underlying disease state of a patient. We distinguish *moderate* disease burden from *severe* disease burden. We do not consider mild disease burden because of our focus on hospitalized patients. Mild cases of COVID-19 were typically treated as out-patient visits. The other 2 states correspond to the 2 possible terminal states of a patient's hospitalization: *discharged* and *deceased*. To simplify notation we sometimes label the states *discharged*, *moderate*, *severe*, and *deceased* as 0, 1, 2, and 3, respectively, giving us the state space $\mathcal{S} = \{0, 1, 2, 3\}$. We assume that patients in the *moderate* disease state can transition into the *severe* disease state or into the *discharged* state, while patients in the *severe* disease state can transition into the *moderate* disease state or the *deceased* state. A graphical representation of this model is shown in Figure 2. More mathematical details of the model can be found in the Supplementary Material.

To capture heterogeneity in disease progression, we model the intensities q_{ij} as functions of patient covariates. For this analysis we consider covariates which are static over the course of the hospitalization and have been shown to be associated with adverse outcomes in COVID-19 patients. This includes the demographic variables age, sex, BMI, and race.[34–36] Sex was recorded as either male or female. Because there were so few non-White/non-Black races represented in this dataset, we categorized race into 3 categories: White, Black, and Other.

In addition to demographic information, medical comorbidities were included as patient covariates as well. While all known comorbidities of hospitalized patients were available in the dataset, we narrowed our focus to 4 relatively common comorbidities, all of which have been shown to be associated with adverse outcomes of COVID-19: asthma, hypertension, diabetes, and kidney disease.[37] There was no missing data among these covariates.

In total there were 9 ProMedica facilities included in this study. ProMedica established care protocols that were distributed to all facilities and were used to diagnose and treat patients. A critical care telemedicine service was established to provide support to outlying hospitals and was staffed by a small group of critical care providers who established protocols for care of patients at all ProMedica facilities. The protocols and real-time management by these providers reduced any variation in care between facilities. Moreover, there were no major structural differences between facilities other than some outlying hospitals that transferred patients to Toledo Hospital when

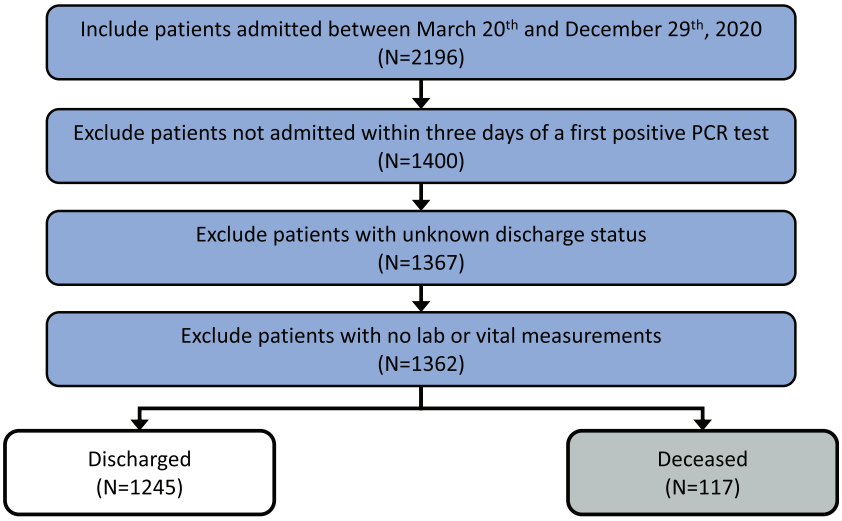

Figure 1. Flow diagram showing inclusion criteria for data preprocessing.

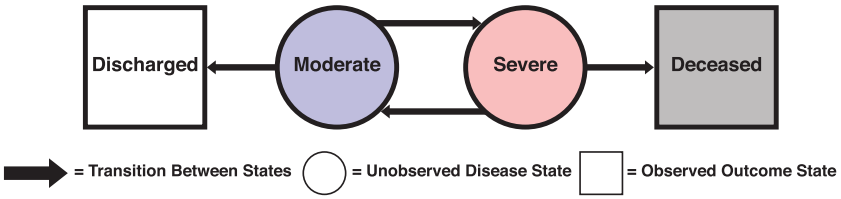

Figure 2. A 4-state Markov model for a COVID-19 positive patient: 2 hidden disease states and 2 observed outcome states.

they became critically ill. For these reasons site information was not included in the analysis.

We incorporated the above patient-level covariates into the Markov model as follows. If there are M covariates, then for patient n, let x_n be the M-dimensional vector encoding all covariates. We model the transition intensity q_{ij} for $i \neq j$ as $q_{ij}(x_n) = e^{w_{ij} \cdot x_n}$, where $w_{ij} \in \mathbb{R}^M$ is a vector of parameters to be learned, and $w_{ij} \cdot x_n$ denotes the dot product between vectors w_{ij} and x_n. Because of the restrictions placed on transitions between latent states, we must have $q_{13} = q_{20} = 0$. Because states 0 and 3 are absorbing states, we must have $q_{0i} = q_{3i} = 0$ for $i \in \{0, 1, 2, 3\}$.

We similarly assume that the initial state probabilities are functions of patient covariates. Letting $\pi = (\pi_0, \pi_1, \pi_2, \pi_3)$ denote the initial state probability distribution satisfying $\pi_i \geq 0$ and $\sum_i \pi_i = 1$, we assume $\pi_i(x_n) = \frac{e^{v_i \cdot x_n}}{\sum_j e^{v_j \cdot x_n}}$ for $i \in \{0, 1, 2, 3\}$, where $v_i \in \mathbb{R}^M$ is a vector of parameters to be learned. Note that hospitalized patients cannot be in the *discharged* state or the *deceased* state when initially admitted, so we set $\pi_0 = \pi_3 = 0$. This leaves only one set of parameters to be learned for π_1 since $\pi_2 = 1 - \pi_1$.

Given parameters v_i and w_{ij} along with covariates x_n, we can define a Markov jump process S_t^n taking values in $\{0, 1, 2, 3\}$. Conditioned on model parameters and patient covariates, we assume that individual Markov jump processes (i.e., patient disease trajectories) are independent of one another. Such an assumption requires that at no point during the study period did clinicians recommend suboptimal treatments due to the scarcity of resources, such as shortages of materials for patient care or personal protective equipment. Due to pre-pandemic planning, there were no shortages of materials for patient care or personal protective equipment for providers during the study period. Thus, such an independence assumption among disease trajectories is reasonable.

The underlying disease states *moderate* and *severe* are never directly observed. Instead, they are indirectly observed by various measurements taken throughout the course of the hospitalization. In particular, the vitals and labs can be interpreted as indirect, noisy measurements of an underlying disease state, and they can be used to infer the disease state. These measurements are sometimes referred to as *emissions* in the HMM literature. The main assumption of these emissions is that they are independent from all other model parameters when conditioned on the latent disease state. Specifically, we assume that all emissions are independent and normally distributed when conditioned on the underlying latent state. Namely, if we let y_t^k be the kth emission observed at time $t > 0$ for patient n, then conditioned on the patient being in state $i \in \{0, 1, 2, 3\}$ at time t we have

$$y_t^k | S_t^n = i \sim \mathcal{N}(\mu_i^k, \sigma_i^k),$$

where μ_i^k and σ_i^k are the mean and standard deviation, respectively, of a normal distribution and are to be learned from the data.

While not all vitals and labs were available for all patients at all times, the conditional independence assumption makes it trivial to account for missing emissions by simply integrating over the unobserved data. More details on the prevalence of observed vitals and labs among the study population can be found in the Supplementary Material.

Finally, we note that the end states *discharged* and *deceased* are observed states and thus do not have normally distributed emissions associated with them. Details on how to account for these fully observed states into the HMM framework can be found in the Supplementary Material. All model parameters were estimated via maximum likelihood estimation, while 95% confidence intervals of the maximum likelihood estimators were obtained via a naive bootstrap.[38] Statistical significance of parameter estimates at the 5% level was implied by a 95% confidence interval that excluded the null value of zero. All inference procedures were performed using Lawrence Livermore National Laboratory high-performance computing resources. Details can be found in the Supplementary Material.

Analysis of risk factors

Identification of risk factors is critical for efficient clinical management of COVID-19 patients. A significant amount of research has been published on risk factors for adverse outcomes for COVID-19 patients.[2–10] Most of these studies focus on uncovering statistically significant associations between patient covariates (risk factors) and adverse outcomes (such as the need for mechanical ventilation or death). In addition to typical terminal outcomes, our proposed model includes intermediate disease states. Specifically, our model considers the following events: (1) progression from a *moderate* disease state to a *severe* disease state; (2) regression from a *severe* disease state to a *moderate* disease state; (3) progression from a *severe* disease state to the *deceased* state; and (4) regression from a *moderate* disease state to the *discharged* state.

To investigate possible correlations of patient-level covariates with adverse disease progression indicators, we estimated statistics that characterize the underlying Markov process. First, we estimated the probability of disease progression conditioned on a state transition occurring. For example, if a patient is in the *moderate* state, then eventually the patient will transition to either the *severe* state or the *discharged* state. We estimated the probability that the patient transitions to the *severe* state (rather than the *discharged* state) when this transition occurs, and we denote this probability by p_{12}. Similarly, we estimated the probability that a patient in the *severe* state transitions to the *deceased* state (rather than the *moderate* state) conditioned on a transition occurring, and we denote this probability by p_{23}.

A given set of patient-level covariates x determines distinct transition probabilities $p_{12}(x)$ and $p_{23}(x)$. We investigate the effect of covariates on these transition probabilities by taking 2 patient covariate vectors x and x' that differ only by a single covariate of interest. We compute the relative risk of transitioning to a more severe state between the 2 different cohorts. Specifically, the relative risk

(RR) between groups x and x' of transitioning from state *moderate* to state *severe* is defined by

$$RR_{12}(x, x') = \frac{p_{12}(x')}{p_{12}(x)}.$$

Similarly, the relative risk of transitioning from the *severe* state to the *deceased* state is

$$RR_{23}(x, x') = \frac{p_{23}(x')}{p_{23}(x)}.$$

In addition to the relative risk of disease progression, we also estimated the overall relative risk of mortality as follows. Let $p^*_{i3}(x)$ be the probability that a patient with covariate vector x eventually ends in the *deceased* state starting from state $i = 1, 2$. This is known as a *hitting probability*, and details can be found in the Supplementary Materials on how to compute it. Furthermore, let $\pi_i(x)$ be the probability that patient x is in state i at the time of hospitalization (i.e., the initial state probabilities). We then define the overall relative risk of mortality as

$$RR^*(X) = \frac{\pi_1(x')p^*_{13}(x') + \pi_2(x')p^*_{23}(x')}{\pi_1(x)p^*_{13}(x) + \pi_2(x)p^*_{23}(x)}.$$

The value RR^* is not conditioned on being in either latent disease state, but is averaged over the initial state probabilities, giving us an overall relative risk of mortality from the time of hospitalization.

If we consider the covariate vector x as a baseline (or control) cohort and x' as the alternative (or treatment) cohort, then the statistics $RR_{12}(x, x')$ and $RR_{23}(x, x')$ provide information on how risk factors in vector x' and absent in vector x are associated with the likelihood of progressing to more severe disease states. In particular, if $RR_{12}(x, x') > 1$, this suggests that the risk factors present in vector x' and absent in vector x are associated with an increase in the probability of progressing from a *moderate* disease state to a *severe* disease state, whereas $RR_{12}(x, x') < 1$ suggests that the risk factors present in vector x' and absent in vector x are associated with a decrease in the probability of progressing from a *moderate* disease state to a *severe* disease state. Analogous relations hold between the statistic $RR_{23}(x, x')$ and the probability of transitioning from the *severe* state to the *discharged* state. The statistic $RR^*(x, x')$ provides information on how risk factors present in vector x' and absent in vector x are associated with the overall probability of ending in the *deceased* state. In particular, if $RR^*(x, x') > 1$, then this suggests that the risk factors present in x' and absent in vector x are associated with an increase in the overall probability of ending in the *deceased* state.

In order to isolate the association between a particular covariate and the risk of progression, we average the above statistics over the empirical distribution of patient covariates. Specifically, suppose $x_\ell = x'_\ell$ for all $\ell \neq k$ and $x_k \neq x'_k$ for some k. Then letting $x_{-k} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$ and $E_{x_{-k}}$ denote expectation with respect to the joint distribution of x_{-k}, we wish to estimate $\bar{RR}_{ij} = E_{x_{-k}}[RR_{ij}(x, x')]$. We approximate this value via Monte Carlo integration using bootstrap samples of the empirical distribution of patient covariates. Namely, if $x(b)$ is drawn with replacement from the empirical distribution of patient covariate vectors and $x'(b)$ is the same as $x(b)$ except for one covariate of interest, we approximate the mean relative risk as

$$\bar{RR}_{ij} \approx \sum_b RR_{ij}(x(b), x'(b)).$$

Bootstrapped standard errors and the bias-corrected percentile method[38,39] were used to construct 95% confidence intervals for these statistics. Statistical significance at the 5% level is implied by 95% confidence intervals that exclude the value of 1.

RESULTS

For each emission (observed lab or vital), there are 2 distinct sets of parameters: one associated with the *moderate* state and one with the *severe* state. Table 1 shows the emission distribution parameters (the mean and standard deviation of a Gaussian distribution) for both latent states. In addition, it shows the differences between the estimated *severe* and *moderate* emission distribution parameters, along with 95% confidence intervals of these differences. An $*$ denotes that the difference is significant at the 5% level. Of the 15 emission distributions, all but 4 mean parameters and 5 standard deviation parameters are significantly different at the 5% level. This indicates that our model is successfully learning to differentiate 2 distinct latent states which can be characterized by vital and lab measurements. Parameter estimates for all other model parameters can be found in the Supplementary Material.

The bootstrapped confidence intervals of the mean relative risk metrics, $\bar{RR}_{ij}(x, x')$ and $\bar{RR}^*(x, x')$, are shown in Table 2. All risk factors considered, other than BMI, were associated with a statistically significant increase in the relative risk of progressing from *moderate* to *severe* states. On the other hand, all risk factors considered, other than BMI and age, were associated with a statistically significant *decrease* in the relative risk of progressing from the *severe* to *deceased* state. Both age and BMI were associated with an increased risk of progressing from *severe* to *deceased*, but this was not significant at the 5% level. Finally, we note that only age and being Black were associated with a statistically significant increase in the relative risk of mortality.

DISCUSSION

Understanding which risk factors are associated with adverse patient-centered outcomes is critical to improving patient care. A more dynamically responsive healthcare system should also consider *when* in the course of hospitalization certain risk factors are more associated with adverse patient-centered outcomes. By modeling the entire course of disease trajectories during hospitalization with a covariate-dependent, continuous-time hidden Markov model, we found known risk factors to have different associations to disease progression depending on the disease state of the patient. The risk factors that demonstrated this pattern were being male, being Black and having a medical comorbidity.

Perhaps somewhat counter intuitive is the fact that a particular risk factor, such as being Black or being male, can be associated with an increase in the relative risk of transitioning from *moderate* to *severe*, a decrease in the relative risk of transitioning from *severe* to *deceased*, and an increase in the overall relative risk of mortality. This is best understood by remembering that the relative risks computed here are conditioned on being in a particular disease state. Taking sex as an example, females are more likely to transition to the *discharged* state from the *moderate* disease state than males, while males are more likely to transition to the *severe* state from the *moderate* disease state than females. But once we condition on being in the *severe* state and assess the risk of transitioning to the *deceased* state, we only account for those males and females that were sick enough to reach the *severe* state. Looking at only those individuals in the *severe* disease state it can happen (as is the case with sex, race,

Table 1. Maximum likelihood estimates of emission distribution parameters along with 95% CIs of the differences between *moderate* and *severe* state parameters

	moderate Mean (SD)	*severe* Mean (SD)	*severe* − *moderate* Mean (95% CI)	*severe* − *moderate* SD (95% CI)
C-reactive protein	7.62 (15.92)	13.98 (19.36)	6.37 (5.91, 7.75)*	3.44 (3.24, 4.56)*
Blood urea nitrogen	26.95 (48.90)	46.98 (66.74)	20.03 (19.11, 27.63)*	17.85 (16.59, 22.29)*
Lactate dehydrogenase	293.60 (435.70)	412.22 (633.67)	118.62 (103.83, 175.61)*	197.97 (136.99, 373.51)*
Procalcitonin	0.28 (2.96)	6.53 (30.13)	6.26 (6.14, 12.46)*	27.17 (31.41, 42.88)*
Ferritin	540.85 (1170.92)	1005.17 (1737.23)	464.33 (215.53, 719.22)*	566.31 (360.75, 857.73)*
Anion-gap	9.66 (12.76)	11.57 (14.37)	1.90 (1.68, 2.59)*	1.61 (1.53, 1.97)*
D-dimer	503.54 (2042.11)	3571.77 (9842.44)	3068.23 (2741.31, 5145.36)*	7800.33 (6535.21, 11623.34)*
% O_2 Sat	95.30 (96.22)	90.46 (104.69)	−4.85 (−7.28, −4.10)*	8.475 (7.27, 11.08)*
Hemoglobin	11.83 (14.07)	11.49 (14.18)	−0.35 (−1.69, −0.11)*	0.11 (−0.16, 0.31)
Platelets	240.79 (352.09)	252.70 (390.57)	11.91 (−4.73, 38.58)	38.48 (19.82, 88.84)*
Systolic pressure	125.51 (144.13)	121.40 (145.26)	−4.10 (−24.92, 1.45)	1.13 (−0.80, 5.45)
Diastolic pressure	70.55 (83.49)	68.95 (84.36)	−1.60 (−13.32, 0.88)	0.87 (−0.51, 4.00)
Respirations	19.02 (24.44)	25.80 (26.66)	6.78 (6.56, 7.43)*	2.23 (−2.41, 2.82)
Temperature	98.19 (99.01)	98.93 (100.93)	0.74 (−0.56, 0.92)	1.92 (−1.50, 2.48)
Urine output	329.54 (567.03)	516.24 (2839.69)	186.70 (116.68, 615.94)*	2272.67 (1044.50, 6726.36)*

*Denotes that the CI does not contain zero, indicating significance at the 5% level.

Table 2. Maximum likelihood estimates and 95% CIs for the relative risk of disease progression between 2 cohorts

Covariate[a]: x'/x	$\bar{RR}_{12}(x, x')$ *moderate* → *severe*	$\bar{RR}_{23}(x, x')$ *severe* → *deceased*	$\bar{RR}^{*}(x, x')$ *entry* → *deceased*
Age: high/low[b]	1.08 (1.01, 1.27)*	1.21 (0.94, 1.73)	1.26 (1.07, 1.58)*
Sex: male/female	1.91 (1.25, 2.40)*	0.32 (0.20, 0.46)*	1.26 (0.61, 1.53)
Race: Black/White	1.62 (1.19, 1.90)*	0.29 (0.20, 0.46)*	1.27 (1.11, 1.61)*
BMI: high/low	0.97 (0.88, 1.03)	1.04 (0.67, 1.27)	0.96 (0.78, 1.10)
Asthma: yes/no	1.33 (1.16, 1.66)*	0.41 (0.27, 0.88)*	0.97 (0.69, 1.40)
Diabetes: yes/no	1.49 (1.20, 1.72)*	0.33 (0.23, 0.49)*	1.10 (0.75, 1.33)
Hypertension: yes/no	1.55 (1.27, 1.83)*	0.35 (0.27, 0.51)*	1.09 (0.81, 1.36)
Kidney disease: yes/no	1.47 (1.22, 1.69)*	0.31 (0.22, 0.45)*	1.15 (0.92, 1.50)

*Denotes that the CI does not contain one, indicating significance at the 5% level.

[a]In each row, vector x' differs from vector x by a single covariate.

[b]For age and BMI, high/low are defined as one standard deviation above/below the population mean.

and some comorbidities) that the risk of entering the *deceased* state is higher for the individuals that were less likely to enter the *severe* state in the first place. Figure 3 shows a graphical representation of this situation.

Note that if a fixed-time (static) model had been used to discern any associations between disease severity and patient outcome (death or discharge), one may introduce an immortal time bias, leading to incorrect inferences about relative risk.[40] This is because in a fixed-time model periods of follow-up may be inappropriately assigned to a particular disease state. An HMM, on the the other hand, is not susceptible to such biases because it can infer the disease severity throughout the course of follow-up in a time-varying manner, greatly reducing such errors. We believe that the counter-intuitive results presented in this article have not been reported earlier due to the significant prevalence of static data and models used for risk stratification of COVID-19 patients.

The clinical implications revealed by the dynamic modeling in the current study are important. By only focusing on the static risk factors, a care-provider may mistakenly assign risks that do not reflect the true underlying risk *conditioned on current disease state*. For example, if a male and female patient are both assessed to be in a severe disease state, and the prevailing static risk factors are used to evaluate the risk of progression, one may mistakenly infer that the male is at higher risk than the female because being male is considered a risk factor for adverse outcomes of COVID-19. But in fact, based on our findings the female patient is at higher risk of death once we condition on the current disease states of the patients. Such a conditional risk assessment may lead to improved patient outcomes as at-risk patients can be appropriately identified for intervention. Similarly this strategy may allow more focused allocation of hospital resources, especially during a global pandemic such as COVID-19, which has repeatedly strained hospital resources during multiple waves of mass infections.

Higher age was the only risk factor shown to be associated with an increase in risk of disease progression from both the *moderate* disease state and the *severe* disease state. While the relative risk of transitioning from the *severe* state to the *deceased* state was not found to be significant at the 5% level, the CI for this value was (0.94, 1.73), which is still rather strong evidence that higher age is associated with a higher relative risk of death.

BMI, on the other hand, did not appear to be a risk factor in either state. One explanation for this is the fact that the population under consideration is biased toward high BMI. The self-reported prevalence of obesity (BMI > 30) in the state of Ohio is 34.8%,[41] whereas the prevalence of obesity in the current dataset is 60%. Thus even though we find that BMI is not associated with elevated

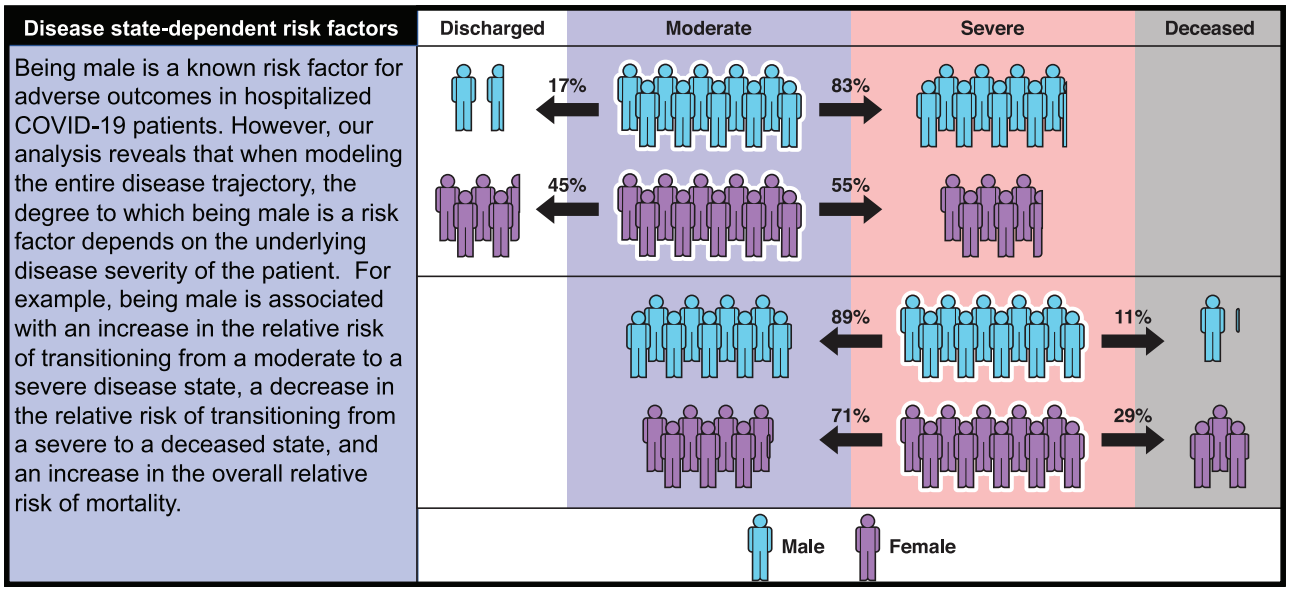

Figure 3. Dynamic disease progression modeling allows us to determine when in the course of a disease a specific patient covariate can be considered a risk factor. In COVID-19, the association between sex and the risk of disease progression is different depending on the underlying disease state.

risks of disease progression among hospitalized COVID-19 patients, it does appear to be associated with higher rates of hospitalization among the general population. Another explanation for this finding is that BMI itself is not associated with changes in the relative risk of disease progression, but various medical comorbidities that are correlated with high BMI are associated with such elevations in risk. This is supported by the fact that when we performed a similar analysis but without comorbidity information, higher BMI was found to be associated with an increased relative risk of disease progression.

Importantly, while this model was trained only on EHR data from ProMedica health system in northwestern Ohio and southeastern Michigan, there is evidence that the results may be transferable to other cohorts. We found a high degree of correlation with the Epic deterioration index,[42] a proprietary risk metric validated on much larger patient cohorts, suggesting that our approach is transferable to other cohorts. Similarly, we found a high degree of concordance with our inferred emission distribution parameters and NIH clinical guidelines on identifying COVID-19 disease severity.[43] Details of both the Epic deterioration index and NIH validations can be found in the Supplementary Material. While such validations are reassuring, future work should validate the proposed methods on a larger, independent dataset to see if similar results are observed. Extensions to the work herein involves relaxing the Markov assumption on disease dynamics, considering nonlinear effects of the covariates, and explicitly modeling interventions such as ventilation.

CONCLUSION

Compared to many reported studies that ignore the temporal progression of disease in their analysis, this study provides a unique modeling-based approach to understanding how patient demographics and medical comorbidities can present differential risk profiles depending on the underlying disease state. The proposed approach performs risk forecasting and stratification based on the full patient trajectory and serves as an exploratory tool for generating novel clinical hypotheses. We estimated the parameters of our proposed HMM based on a cohort of 1362 hospitalized SARS-CoV-2 positive patients via maximum likelihood estimation. By modeling the entire trajectory of hospitalized COVID-19 patients, we were able to show statistically significant differences in the relative risk of disease progression conditioned on current disease state. These differences should be taken into consideration when performing risk assessment among hospitalized patients. Such information is potentially more actionable throughout the course of care, possibly leading to better patient outcomes. Moreover, disease state–dependent risk assessments can be applied not only to COVID-19, but also to many other acute and chronic diseases that, to date, have largely been assessed only with static data and modeling techniques.

FUNDING

The funding was provided by the Lawrence Livermore National Laboratory (LLNL) Laboratory Directed Research and Development (LDRD) Program under Project Number 19-ERD-009, The University of Toledo Women and Philanthropy Genetic Analysis Instrumentation Center, The University of Toledo Medical Research Society, and also by David and Helen Boone Foundation Research Fund.

AUTHOR CONTRIBUTIONS

BS performed all theoretical analysis and mathematical derivations. BS and JC contributed equally to data ingestion, curation, software development, and study design. BS, JC, and PR contributed to experiment and study design. All authors contributed to the analysis of the results and the manuscript preparation. BS, JC, RC, SN, JMD, STH, JH, DJK, DM, PR: conceptualization; BS, JC, RC, SN, PK, LW, MW, PR: data curation; BS, JC, RC, SN, PK, LW, MW, PR: formal analysis; STH, DJK, PR: funding acquisition; BS, JC, RC, SN, PK, LW, MW, JMD, STH, JH, DJK, DM, PR: investigation and methodology; PK, LW, MW, STH, DJK, DM, PR: project administration; PK, LW, MW, JMD, STH, JH, DJK, DM, PR: resources; BS, JC, RC, SN, PK, LW, MW, PR: software; JMD, STH, JH, DJK, DM, PR: supervision; BS, JC, RC, SN, PK, LW, MW, PR: validation; BS, JC, RC, SN, JMD, STH, JH, DJK, DM, PR: writing-original draft, BS, JC, RC, SN, PK, LW, MW, JMD, STH, JH, DJK, DM, PR: writing, review and editing. Final version was approved by all authors.

INFORMED CONSENT

The study protocol involving analysis of fully de-identified data was reviewed and approved with Full Waiver of informed consent granted (Expedited, Category #5 research) by the respective Institutional Review Board's of ProMedica and Lawrence Livermore National Laboratory. The study was performed in compliance with all regulations and guidelines from the United State Department of Health and Human Services.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL LDRD Program under Project No. 19-ERD-009. LLNL-JRNL-826855. An abstract based on this work was accepted for presentation at the 2021 Midwest Clinical and Translational Research Meeting. We thank Dr. Amy Gryshuk for her support.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article cannot be shared publicly due to the privacy of individuals that participated in the study.

REFERENCES

1. Geneva: World Health Organization. WHO COVID-19 dashboard, 2022. https://covid19.who.int/. Accessed 17 January, 2022
2. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics* 2020; 8 (6): 890.
3. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med* 2021; 4 (1): 3–5.
4. Vaid A, Jaladanki SK, Xu J, *et al.* Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med Inform* 2021; 9 (1): e24207.
5. Li S, Lin Y, Zhu T, *et al.* Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method. *Neural Comput Appl* 2021; 1–10. https://doi.org/10.1007/s00521-020-05592-1
6. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health (Amst)* 2021; 20: 100178.
7. Nguyen S, Chan R, Cadena J, *et al.* Budget constrained machine learning for early prediction of adverse outcomes for COVID-19 patients. *Sci Rep* 2021; 11 (1): 1–14.
8. Smarr BL, Aschbacher K, Fisher SM, *et al.* Feasibility of continuous fever monitoring using wearable devices. *Sci Rep* 2020; 10 (1): 21640.
9. Roimi M, Gutman R, Somer J, *et al.* Development and validation of a machine learning model predicting illness trajectory and hospital utilization of COVID-19 patients: a nationwide study. *J Am Med Inform Assoc* 2021; 28 (6): 1188–96.
10. Wongvibulsin S, Garibaldi BT, Antar AAR, *et al.* Development of Severe COVID-19 Adaptive Risk Predictor (SCARP), a calculator to predict severe disease or death in hospitalized patients with COVID-19. *Ann Intern Med* 2021; 174 (6): 777–85.
11. Adler FR, Liou TG. The dynamics of disease progression in cystic fibrosis. *PLoS One* 2016; 11 (6): e0156752.
12. Alzakerin HM, Halkiadakis Y, Morgan KD. Autoregressive modeling to assess stride time pattern stability in individuals with Huntington's disease. *BMC Neurol* 2019; 19 (1): 1–6.
13. Liu YY, Li S, Li F, Song L, Rehg JM. Efficient learning of continuous-time hidden Markov models for disease progression. *Adv Neural Inf Process Syst* 2015; 28: 3599–607.
14. Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). PMLR; 2017. p. 1174–1182.
15. Meng R, Soper B, Lee HKH, Liu VX, Greene JD, Ray P. Nonstationary multivariate Gaussian processes for electronic health records. *J Biomed Inform* 2021; 117: 103698.
16. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017; 24 (2): 361–70.
17. Zhang X. Application of discrete event simulation in health care: a systematic review. *BMC Health Serv Res* 2018; 18 (1): 1–11.
18. Pan F, Reifsnider O, Zheng Y, *et al.* Modeling clinical outcomes in prostate cancer: application and validation of the discrete event simulation approach. *Value Health* 2018; 21 (4): 416–22.
19. Karnon J, Stahl J, Brennan A, Caro JJ, Mar J, Möller J; ISPOR-SMDM Modeling Good Research Practices Task Force. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. *Value Health* 2012; 15 (6): 821–7.
20. Satten GA, Longini IM. Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *J R Stat Soc Ser C (Appl Stat)* 1996; 45 (3): 275–309.
21. Soper BC, Nygård M, Abdulla G, Meng R, Nygård JF. A hidden Markov model for population-level cervical cancer screening data. *Stat Med* 2020; 39 (25): 3569–90.
22. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques – Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press; 2009.
23. Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Stat Med* 2003; 22 (3): 441–62.
24. Kay R. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics* 1986; 42 (4): 855–65.
25. Kirby AJ, Spiegelhalter DJ. Modeling the precursors of cervical cancer. In: Lange N, ed. *Case Studies in Biometry*. Wiley-Interscience; 1994: 359–83.
26. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (the Statistician)* 2003 Jul; 52 (2): 193–209.
27. Martino A, Guatteri G, Paganoni AM. Multivariate hidden Markov models for disease progression. *Stat Anal Data Min ASA Data Sci J* 2020; 13 (5): 499–507.
28. Antonio PG, Aman V, Yu L, David S, David B. Modeling chronic obstructive pulmonary disease progression using continuous-time hidden Markov models. *Stud Health Technol Inform* 2019; 264: 920–4.

29. Sun Z, Ghosh S, Li Y, *et al*. A probabilistic disease progression modeling approach and its application to integrated Huntington's disease observational data. *JAMIA Open* 2019; 2 (1): 123–30.
30. Williams JP, Storlie CB, Therneau TM, Jack Clifford R. Jr, Hannig J. A Bayesian approach to multistate hidden Markov models: application to dementia progression. *J Am Stat Assoc* 2020; 115 (529): 16–31.
31. Petersen BK, Mayhew MB, Ogbuefi KOE, Greene JD, Liu VX, Ray P. Modeling sepsis progression using hidden Markov models. NIPS Machine Learning for Health (ML4H). *arXiv Preprint* 2017; arXiv:1801.02736.
32. Gupta A, Liu T, Crick C. Utilizing time series data embedded in electronic health records to develop continuous mortality risk prediction models using hidden Markov models: a sepsis case study. *Stat Meth Med Res* 2020; 29 (11): 3409–23.
33. Perveen S, Shahbaz M, Ansari MS, Keshavjee K, Guergachi A. A hybrid approach for modeling type 2 diabetes mellitus progression. *Front Genet* 2019; 10: 1076.
34. Rosenthal N, Cao Z, Gundrum J, Sianis J, Safo S. Risk factors associated with in-hospital mortality in a US national sample of patients with COVID-19. *JAMA Netw Open* 2020; 3 (12): e2029058.
35. Rossen LM, Branum AM, Ahmad FB, Sutton P, Anderson RN. Excess deaths associated with COVID-19, by age and race and ethnicity—United States, January 26–October 3, 2020. *MMWR Morb Mortal Wkly Rep* 2020; 69 (42): 1522–7.
36. The Covid Tracking Project. The Atlantic. https://covidtracking.com/. Accessed February 2, 2021.
37. Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip GYH. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: a federated electronic medical record analysis. *PLOS Med* 2020; 17 (9): e1003321.
38. Efron B. *The Jackknife, the Bootstrap, and Other Resampling Plans*. BIO 63. Stanford, CA: Department of Statistics, Stanford University; 1980.
39. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000; 19 (9): 1141–64.
40. Yadav K, Lewis RJ. Immortal time bias in observational studies. *JAMA* 2021; 325 (7): 686–7.
41. Centers for Disease Control. Adult Obesity Prevalence Maps. https://www.cdc.gov/obesity/data/prevalence-maps.html. Accessed May 13, 2021.
42. Singh K, Valley TS, Tang S, *et al*. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Ann Am Thorac Soc* 2021; 18 (7): 1129–37.
43. National Institutes of Health. COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19) Treatment Guidelines. https://www.covid19treatmentguidelines.nih.gov. Accessed March 23, 2021.