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

Braden Soper (soper3@llnl.gov)
Lawrence Livermore National Laboratory https://orcid.org/0000-0002-0308-708X
Jose Cadena
Lawrence Livermore National Laboratory
Sam Nguyen
Lawrence Livermore National Laboratory
Ryan Chan
Lawrence Livermore National Laboratory
Paul Kiszka
ProMedica Health System
Lucas Womack
ProMedica Health System
Mark Work
ProMedica Health System
Joan Duggan
University of Toledo
Steven Haller
University of Toledo
Jennifer Hanrahan
University of Toledo
David Kennedy
University of Toledo
Deepa Mukundan
University of Toledo
Priyadip Ray
Lawrence Livermore National Laboratory

Article

Keywords:

Posted Date: September 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-923677/v1

Version of Record: A version of this preprint was published at Journal of the American Medical Informatics Association on January 1st, 2022. See the published version at https://doi.org/10.1093/jamia/ocac012.

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

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

[1]Lawrence Livermore National Laboratory, Livermore, CA 94550 USA
[2]Promedica Health System
[3]University of Toledo
[*]soper3@llnl.gov

ABSTRACT

The global pandemic of the SARS-CoV-2 coronavirus has significantly strained hospital resources worldwide. Improved understanding of the COVID-19 disease trajectory for patients requiring hospitalization would allow for the development of more targeted preventative, diagnostic and therapeutic strategies. A covariate-dependent, continuous-time hidden Markov model with four states (moderate-illness, severe-illness, 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 1,362 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, co-morbidities, vital signs and laboratory test results were retrospectively evaluated to predict clinical progression and outcomes. Several patient-level covariates were associated with differential impacts on the risk of progression. Specifically, while being male, being black or having a medical co-morbidity were all associated with an increased risk of progressing from the moderate to severe disease state, these factors resulted in a decreased risk of transitioning from the severe to the deceased disease state. Body mass index (BMI) alone was not found to be associated with an increased risk of disease progression, while higher age was associated with an increased risk in progressing from moderate to severe and from severe to deceased states. Regardless of the differential risk profiles, all covariates considered other than BMI and asthma were associated with an overall increased risk of transitioning to the deceased state. 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. Such dynamic risk stratification models have the potential not only to improve clinical outcomes in COVID-19, but also a myriad of other acute and chronic diseases that, to date, have largely been assessed only by static modeling techniques.

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 4.3 million estimated deaths worldwide. Understanding of risk factors influencing disease severity is critical for 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[1–6]. Furthermore, recent studies suggest such risk factors also may affect certain aspects of COVID-19 progression, specifically disease onset[7] and time-to-death[8]. 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[9], statistical approaches such as autoregressive models[10], hidden Markov models[11] and Gaussian processes[12,13], as well as deep learning tools such as recurrent neural networks[14]. The particular choice of a modeling approach depends on the amount and quality of knowledge about the underlying disease mechanism, the stochasticity and heterogeneity of the disease symptoms, sample size available for model parameter estimation as well as the need for model interpretability. In this paper, we primarily focus on hidden Markov models for characterizing the disease trajectory for hospitalized SARS-CoV-2 positive patients. The particular choice is motivated by several factors: 1) being relatively new, there is a general lack of understanding of the disease mechanism, 2) significant heterogeneity of disease presentation and outcomes and 3) a modest cohort size of $1,362$ hospitalized patients. In addition, HMMs are fairly easy to

interpret, compared to other statistical approaches such as Gaussian processes.

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-illness, severe-illness, discharged, 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 health care 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 of the patients (e.g., age, race, sex), as well as the history of five vital signs and ten laboratory test results collected throughout the hospitalization were used to train the covariate-dependent, continuous-time hidden Markov model (ct-HMM).

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, race and BMI) and comorbidities (such as asthma, diabetes, hypertension and kidney disease) with adverse patient outcomes can depend on the underlying disease state of the patient*.

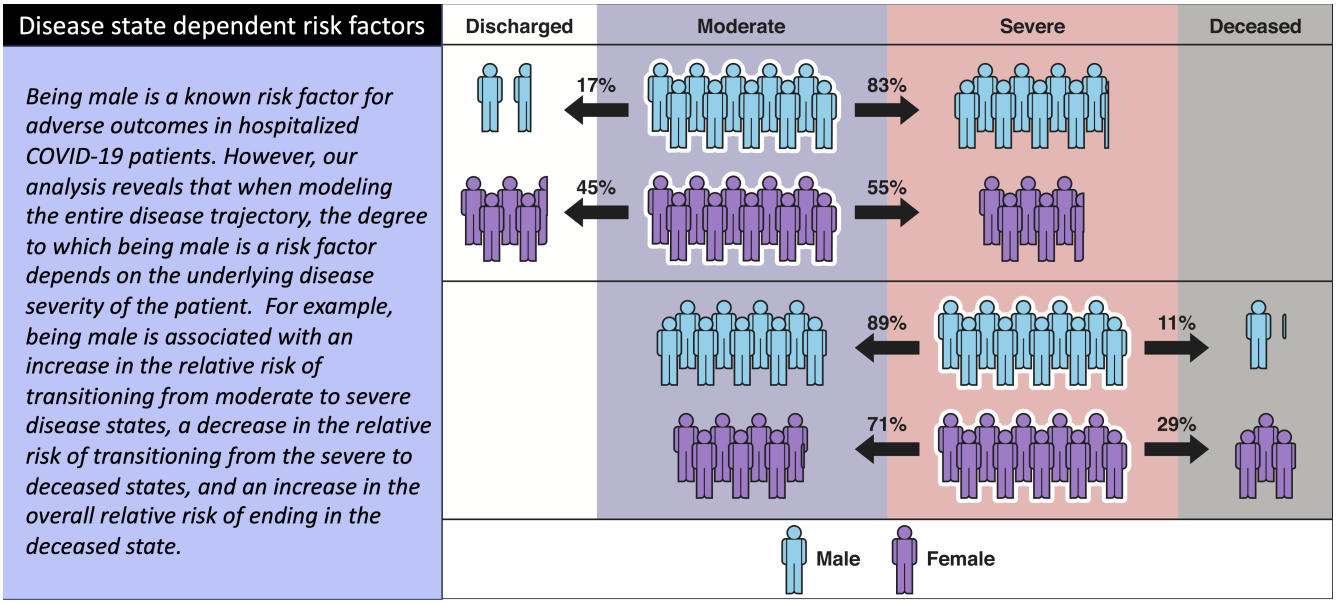


Figure 1. Dynamic disease progression modeling allows us to determine when in the course of a disease a specific patient covariate can be considered a risk factor. In COVID-19, sex has differential impacts on disease progression depending on the underlying disease state.

Data

The data used in this study is composed of electronic health records (EHRs) from patients of ProMedica, the largest health care 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 03/20/2020 and 12/29/2020, and 2) were admitted to the hospital shortly before or after the positive result. A total of 1,362 patients met these criteria. There are three main sources of data available in this dataset, all of which were collected throughout the course of the patients' hospitalizations: 1) Patient demographic information, 2) vital measurements, and 3) laboratory tests results. Details on these three data modalities are described in the following sections.

For this analysis we consider patient-level covariates which are static over the course of the hospitalization. Patient age and BMI are two of the known risk factors for adverse outcomes in COVID-19 infections[15,16]. While these are not static

covariates, on the time scale of the hospitalization they can be treated as such. Racial disparities have also been observed in COVID-19 death rates with Black Americans dying at roughly 1.5 times the rate of White Americans[16,17]. The racial composition of this data was approximately 73% White, 17% Black and 10% neither Black nor White. Because there were so few non-White/non-Black races represented in this dataset we categorized race into three categories: White, Black and Other. While all known comorbidities of hospitalized patients were available in the dataset, we narrowed our focus to four relatively common comorbidities, all of which have been shown to be associated with adverse outcomes of COVID-19[18]: Asthma, hypertension, diabetes, and kidney disease.

A total of 801 patients (58.8% of the cohort) were identified as having one or more of the four comorbidities under consideration. The number (percent) of patients in the cohort identified as having a particular condition was as follows: 629 (46.2%) had hypertension , 410 (30.1%) had diabetes, 158 (11.6%) had kidney disease, and 98 (7.2%) had asthma. Note that patients could have more than one underlying condition. Of the four considered comorbidities 411 (30.2%) had only one, 294 (21.6%) had two, 88 (6.5%) had three and 8 (0.6%) had all four.

Patient data was collected throughout the course of the hospitalization. Patient vital signs were measured and recorded as per usual nursing protocols and were recorded in the electronic medical record. The 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). Results from laboratory test that were ordered during hospitalization were also available. The 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 (%O2-sat), hemoglobin, platelets. Table 5 in the Appendix summarizes vitals, labs and patient-level covariates in the ProMedica dataset.

Methods

Several medical studies have analyzed the associations between various risk factors and adverse outcomes of COVID-19 patients. Similarly, there has been a number of data-driven predictive modeling approaches to identifying associated risk factors. However, the vast majority of these studies have ignored the dynamic progression and regression of COVID-19, instead relying on static data.

Our objective is to investigate the disease state-dependent behavior of static risk factors. To this end we model hospitalized COVID-19 patient *trajectories* given standard EHR data collected throughout the course of hospitalization and. This constrains our modeling choices to discrete state space models. Multi-state Markov models (MMs) and hidden Markov models (HMMs) are two well-known discrete state space models[19] 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[21–27], cancer screening[28–33], vascular disease[34,35], pulmonary disease[36], neurodegenerative disease[37–39], sepsis[40,41], and diabetes[42].

The benefits 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[43]. HMMs bring in the additional benefit of being able to account for measurement noise 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 will focus on Markov jump processes for modeling the underlying disease progression of a COVID-19 patient. In the remainder of this section, we lay out the details of the continuous-time hidden Markov model, which captures both the underlying disease progression and the noisy observation process.

A Continuous-Time Hidden Markov Model for COVID-19 Patient Data

To capture both COVID-19 disease progression and regression, we consider a four-state Markov model in continuous time. 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 are usually treated as out-patient visits. The other two states correspond to the two possible terminal states of a patient's hospitalization: *discharged* and *deceased*. We assume that patients in the moderate disease state can transition into the severe disease state or into the discharge state, while patients in the severe disease state can transition into the moderate disease state or the deceased state. Moreover, we assume the these transitions are stochastic and can be modeled as a continuous-time, finite-state Markov chain. A graphical representation of this model is shown in Figure 2.

The defining characteristics of a Markov jump process on a finite state space can be summarized by three probability distributions. The first is a probability distribution over the initial state of the system, which we will denote by π and refer to as the *initial state probability distirbution*. The second is a probability distribution over the amount of time spent in a state after the system enters the state. This is sometimes referred to as the *sojourn time* or *wait time* of the state. A defining characteristic

of Markov jump processes is that the wait time is exponentially distributed, and we denote the *mean wait time* by λ_i for state i. Finally, when the system transitions out of a particular state, there is a probability distribution over all possible states the system can transition to. We refer to this as the *transition probability distribution*, and we let p_{ij} denote the probability that the system transitions to state j given that it transitions out of state i. To capture heterogeneity in disease progression, we modeled these probability distributions as functions of patient-level covariates. Mathematical details of the underlying Markov jump process and our parameterization of the model can be found in the Appendix.

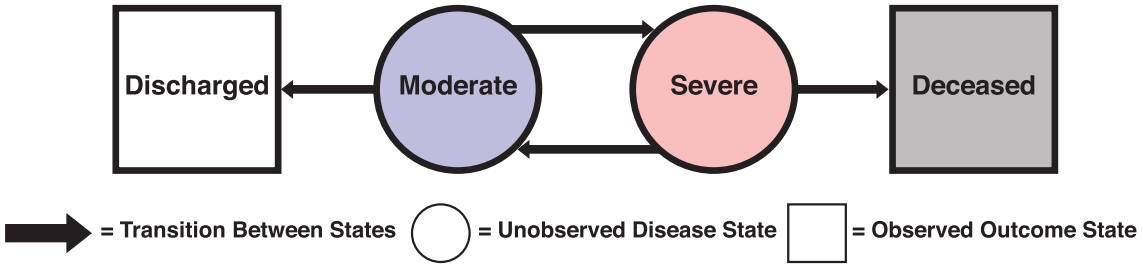


Figure 2. A four-state Markov model for a COVID-19 positive patient: two hidden disease states and two observed outcomes states.

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 vital measurements and lab results can be interpreted as indirect, noisy measurements of an underlying disease state. These measurements are taken with the express purpose of inferring the severity of the illness. These measurements are sometimes referred to as *emissions* in the HMM literature. The main assumption of these emission distributions 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. When performing inference of HMMs one objective is to learn the different parameters of the emission distributions (in our case the means and standard deviations) associated with each latent state. More precisely, for a particular patient denote the latent state at time t by S_t and the jth emission at time t by y_t^j. We assume there are parameters μ_s^j and σ_s^j such that if the jth emission is observed at time t when the patient is in state s we have

$$y_t^j | S_t = s \sim \mathcal{N}(\mu_s^j, \sigma_s^j).$$

where μ_s^j and σ_s^j are the mean and standard deviation, respectively, of a normal distribution.

Because of the conditional independence assumption made in HMMs, it is trivial to account for missing emissions by simply integrating over the unobserved data. Thus it is not necessary that all emissions be observed at each observation time, which is usually the case for EHR data. Finally, note that the end states *discharged* and *deceased* are observed states and thus do not have normally distributed emissions associated with them.

Parameter Inference

Parameter inference was performed by maximum likelihood estimation. Because the disease states are latent parameters we wish to integrate them out of the likelihood function, giving us a marginal likelihood function. Let S denote a set of sequences of disease states for all patients, let Y denote a set of observed emissions for all patients, let X denote the set of all patient covariates, and let ϕ be the set of all Markov parameters and emission distribution parameters. Then the marginal likelihood function is given by

$$L(\phi | Y, X) = \sum_S L(\phi | S, Y, X). \tag{1}$$

Note that this summation is over all possible sequences of latent states for all patients, which scales exponentially in the length of the longest observed sequence. The longest sequence in our data is on the order of 1,000 observations. Even with only two

latent states, this is prohibitively large. Recursive computations allow us to compute (1) in computational complexity that grows linearly with the sum of all sequence lengths[44]. Even with these computational savings, obtaining maximum likelihood estimates (MLEs) must be done in an iterative fashion, which increases the overall number of likelihood computations. Given these computational constraints on computing MLEs of our HMM, we still needed to parallelize the likelihood computations over multiple processors. After parallelizing the marginal likelihood computation, we applied automatic differentiation[45] and used the gradient-based optimization algorithm Adam[46] to obtain maximum likelihood estimates.

To obtain standard errors of the MLEs, we performed a naive bootstrap[47]. Specifically, we sampled $1,362$ patients with replacement from the ProMedica cohort, and from this resampled dataset we found MLEs as outlined above. We repeated this $2,500$ times to obtain $2,500$ bootstrap samples. Confidence intervals (CIs) were estimated via a bias-corrected percentile method[47,48]. Due to the computational complexity of the likelihood function, significant computational resources were required to obtain the $2,500$ bootstrap samples. We leveraged Lawrence Livermore National Laboratory (LLNL) high performance computing (HPC) resources to parallelize both individual marginal likelihood computations and batches of bootstrap samples. An HPC cluster of 100 nodes each with 36 processors was used, resulting in a total of 3600 processors used for computing all bootstrapped MLEs. Computing a single bootstrapped MLE with 100 iterations of the Adam optimization algorithm on 36 processors required between 4 and 5 hours of wall-clock time.

Finally, we note that the marginal likelihood function of HMMs is not guaranteed to be a concave function, meaning there may be many local optima in the optimization landscape. Because of this, when attempting to find an MLE for an HMM via gradient-based optimization methods, one must consider the sensitivity of the optimization procedure to the initial conditions of the parameters. To investigate the sensitivity of our inference procedure to initial conditions we first computed the MLE on the entire (non-bootstrapped) dataset with all parameters intialized at zero (assuming all parameter values have been transformed to lie on the real line). This zero-initialization acted as a least-informative reference point for the sensitivity analysis. We next ran the same MLE procedure 250 times, but for each run we initialized the parameters randomly, drawing values uniformly between -1 and 1. These job were run in parallel using 1800 processors on LLNL HPC resources. Results for these 250 runs plus the initial reference run can be found in the supplementary material. It was observed that the inference procedure was indeed sensitive to initial conditions, with each run giving different MLEs and different corresponding marginal likelihood values. However, none of the randomly initialized runs obtained a marginal likelihood value greater than that of the zero-initialized runs. For this reason we took the zero-initialized run as the global MLE, and initialized all subsequent bootstrap runs similarly.

Clinical Interpretation of Latent States

The defined latent states *moderate* and severe are purely mathematical constructs for modeling disease progression. In this section, we investigate the degree to which the defined latent states correlate with clinically defined measures of disease severity. First, we compare our disease state estimates to a proprietary risk metric used in clinical settings called Epic Deterioration Index (EDI)[49]. Second, we compare the inferred emission distributions corresponding to the defined latent states to NIH clinical guidelines on identifying disease severity in COVID-19 patients[50]. We describe these two methods in more detail below.

Epic Deterioration Index

Given the observed vital measurements and lab results collected throughout the hospitalization, estimates of the latent disease states can easily be computed via the forward recursions used to compute the marginal likelihood. These latent-state estimates come in the form of a probability distribution over the possible states *moderate*, *severe*, *discharged*, and *deceased*. More precisely, for a given patient with covariates X_i and observed emissions $Y^i_{t_1},...,Y^i_{t_n}$ from time t_1 to t_n, we can compute a probability distribution over the possible disease states conditioned on the covariates X_i and all observed emissions up until time t_n. For our purposes it is sufficient to estimate the conditional probability that the patient is in the *severe* state. We denote this probability by $r^i_{t_n}$, which is defined as

$$\alpha^i_{t_n} = P(S^i_{t_n} = severe | Y^i_{t_1},...,Y^i_{t_n}, X_i, \phi).$$

The Epic Deterioration Index (EDI) is a proprietary risk metric used in conjunction with EHRs to better inform clinical decision making[49]. The index provides a risk score on a scale of 0-100 where higher numbers indicate higher risk of deterioration. Our model's estimated probability of being in the severe state, $\alpha^i_{t_n}$, can also be seen as a risk metric that evaluates the health of a patient. Furthermore higher values indicate a worse state is more likely. Thus a high degree of correlation between the EDI and our model's severe state probability estimate would provide external validation that our model is learning to correctly identify when a patient is in poor health.

The EDI was available for most patients in our dataset, which we denote by D^i_t for patient i at time t. We validated our *severe*-state estimate α^i_t by computing the sample correlation coefficient between the series $\alpha^i = (\alpha^i_{t_1}, \alpha^i_{t_2}, ..., \alpha^i_{t_{n_i}})$ and the series $D^i = (D^i_{t_1}, D^i_{t_2}, ..., D^i_{t_{n_i}})$. The EDI was typically sampled at a higher frequency than the observed data Y^i. If the EDI was

not available at a particular observation time t_k, then a linear interpolation of the available EDIs was used to approximate the EDI at time t_k.

The sample correlation coefficients between α^i and D^i for the ith patient is given by

$$c_i = \frac{\sum_{j=1}^{n_i}(\alpha_{t_j}^i - \bar{\alpha}^i)(D_{t_j}^i - \bar{D}^i)}{\sqrt{\sum_{j=1}^{n_i}(\alpha_{t_j}^i - \bar{\alpha}^i)^2}\sqrt{\sum_{j=1}^{n_i}(D_{t_j}^i - \bar{D}^i)^2}},$$

where $\bar{\alpha}^i = \frac{1}{n_i}\sum_{j=1}^{n_i}\alpha_{t_j}^i$ and $\bar{D}^i = \frac{1}{n_i}\sum_{j=1}^{n_i}D_{t_j}^i$ are the mean risk metric and mean EDI, respectively. To discover if our risk estimates are significantly correlated to the EDI, we conducted the following hypothesis test. Define $X_i = \mathbb{1}(c_i > 0)$ and assume that $X_i \overset{iid}{\sim} \text{Bernoulli}(p)$ for $i = 1,\ldots,N$ for some $p \in [0,1]$, where N is the size of the cohort. If p is close to 0.5, this indicates a lack of correlation. If p is far from 0.5, this indicates either a positive or negative correlation depending on whether p is less than or greater than 0.5. The null and alternative hypotheses can then be stated as follows.

$$H_0 : p = 0.5\,,$$
$$H_1 : p \neq 0.5\,.$$

The z statistic for this test is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/N}},$$

where $\hat{p} = \frac{1}{N}\sum_{i=1}^{N}X_i$ is the sample proportion of positive results, $p_0 = 0.5$, and N is the number of patients in the cohort. Under the null hypothesis, the z statistic asymptotically converges to a standard Gaussian distribution by the central limit theorem, allowing us to compute approximate p-values for this hypothesis test.

NIH Guidelines

Though we have labeled the latent states *moderate* and *severe*, there is no guarantee that the learned latent states will strictly adhere to any specific clinical interpretation that is consistent with such labels. If our model successfully learns clinically meaningful moderate and severe illness states, we would expect the MLEs for the emission means to be in concordance with clinical guidelines on identifying at-risk patients. We look for such a concordance within the National Institutes of Health COVID-19 treatment guidelines[50]. The NIH guidelines categorize COVID-19 disease severeness into five levels: asymptomatic/presymptomatic, mild, moderate, severe and critical. Under the assumptions of our model, patients with asymptomatic/presymptomatic or mild levels of COVID-19 severity are not hospitalized, and thus are not represented in our training set. The critical level of disease severity is not represented in our model, so we assume it represents the most extremely ill patients in our severe disease state. Focusing on the moderate and severe illness states, the NIH guidelines state the following:

> ***Moderate Illness****: Individuals who show evidence of lower respiratory disease during clinical assessment or imaging and who have saturation of oxygen $SpO2 \geq 94\%$ on room air at sea level.*
> ***Severe Illness****: Individuals who have $SpO2 < 94\%$ on room air at sea level, a ratio of arterial partial pressure of oxygen to fraction of inspired oxygen $(PaO2/FiO2) < 300mmHg$, respiratory frequency > 30 breaths/min, or lung infiltrates $> 50\%$.*

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. Almost all of these publications focus on uncovering statistically significant associations of patient covariates (risk factors) with adverse outcomes (such as the need for mechanical ventilation of death). As discussed in the previous section, in addition to typical terminal outcomes such as death or discharge, our proposed model includes additional disease states which have statistically significant correlations with clinical measures of disease severity. Specifically, our model considers the following events that can occur during the course of hospitalization: 1) disease progression from a *moderate* disease state to a *severe* disease state; 2) disease regression from a *severe* disease state to a *moderate* disease state; 3) disease progression from a *severe* disease state to the *deceased* state; and 4) disease regression from a *moderate* disease state to the *discharged* state. We next investigate risk factors associated with the two adverse events *moderate*→*severe* and *severe*→*deceased* in addition to the overall risk of ending in the *deceased* state from the point of entry to the hospital.

When modeling disease progression it is not clear what is meant by an individual risk factor. Any particular covariate is associated with changes in initial disease state, disease progression and disease regression. Thus we must consider how a single

covariate is associated with several different processes, not just a single outcome, as is the case in static modeling approaches such as logistic regression or survival analysis. This difficulty can be seen in Table 1. Each column shows how a particular patient covariate is associated with changes in the various aspects of the disease model. In particular, we must consider how a change in a single covariate is associated with changes in the probability of entering the hospital in the moderate state, changes in transition intensities governing progression and changes in transition intensities governing regression. This makes it difficult to identify individual covariates as overall risk factors.

To address this challenge we estimated several statistics that characterize the underlying Markov process and are more directly interpretable as measures of risk. 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 at some time in the future 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_{01}. 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_{13}. Second, we estimated the average time spent in the current state after entry into it. This mean wait-time is denoted by λ_0 for the *moderate* state and λ_1 for the *severe* state.

A given set of patient-level covariates X determines distinct transition probabilities $p_{01}(X)$ and $p_{13}(X)$ and distinct mean wait-times $\lambda_0(X)$ and $\lambda_1(X)$. We investigate the effect of covariates on these transition probabilities and mean wait-times by taking two patient covariate vectors X and Y that differ only by a single covariate of interest. For transition probabilities we compute the relative risk of transitioning to a more severe state between the two different cohorts. Specifically, the relative risk (RR) between groups X and Y of transitioning from state *moderate* to state *severe* is defined by

$$RR_{01}(X,Y) = \frac{p_{01}(Y)}{p_{01}(X)}.$$

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

$$RR_{13}(X,Y) = \frac{p_{13}(Y)}{p_{13}(X)}.$$

In addition to the relative risk of disease progression, we also estimated the overall relative risk of ending in the disease state. To do this we consider the following. Let $p^*_{03}(X)$ be the probability that a patient with covariate vector X eventually ends in the deceased state starting from the *moderate* state, and let $p^*_{13}(X)$ be the probability that a patient with covariate vector X eventually ends in the deceased state starting from the *severe* state. This is known as a *hitting probability*, and details can be found in the supplementary materials on how to compute them from the Markov model parameters. We define the overall relative risk of mortality as

$$RR^*(X) = \frac{\pi_0(Y)p^*_{03}(Y) + \pi_1(Y)p^*_{13}(Y)}{\pi_0(X)p^*_{03}(X) + \pi_1(X)p^*_{13}(X)}.$$

where the hitting probabilities are averaged over the initial state probabilities $\pi(X)$. This value is not conditioned on being in either latent disease state, giving us an overall relative risk of ending up in the deceased state from the time of hospitalization. Finally, we also computed the ratio of mean wait-times in the *moderate* state $W_0(X,Y) = \frac{\lambda_0(Y)}{\lambda_0(X)}$ and the *severe* state $W_1(X,Y) = \frac{\lambda_1(Y)}{\lambda_1(X)}$.

If we consider the covariate vector X as a baseline (or control) cohort and Y as the alternative (or treatment) cohort, then the statistics $RR_{01}(X,Y)$ and $RR_{13}(X,Y)$ provide information on how risk factors in vector Y and absent in vector X are associated with the likelihood of progressing to more severe disease states. In particular, if $RR_{01}(X,Y) > 1$, this suggests that the risk factors present in vector Y 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 $RR_{01}(X,Y) < 1$ suggests that the risk factors present in vector Y and absent in vector X are associated with a decrease in the probability of progressing from a moderate disease state to a severe state. Analogous relations hold between the statistic $RR_{13}(X,Y)$ and the probability of transitioning from the *severe* state to the *discharged* state. The statistic $RR^*(X,Y)$ provides information on how risk factors present in vector Y and absent in vector X are associated with the overall probability of ending in the *deceased* state. In particular, if $RR^*(X,Y) > 1$, then this suggests that the risk factors present in Y and absent in vector X are associated with an increase in the overall probability of ending in the *deceased* state.

The statistics $W_0(X,Y)$ and $W_1(X,Y)$ provide information on how risk factors present in vector Y and absent in vector X are associated with changes to the average length-of-stay (or wait-times) in the *moderate* and *severe* states, respectively. In particular, if $W_0(X,Y) > 1$, then this suggests that the risk factors present in Y and absent in vector X are associated with an

increase in the average length-of-stay in the *moderate* state, while $W_0(X,Y) < 1$ suggests that the risk factors present in Y and absent in vector X are associated with a decrease in the average length-of-stay in the *moderate* state. Analogous relations hold between the statistic $W_1(X,Y)$ and the average length-of-stay in the *severe* state.

In order to evaluate the effect of a particular covariate on disease dynamics over the population, we average the above statistics over the empirical distribution of patient covariates. Specifically, suppose $X_\ell = Y_\ell$ for all $\ell \neq k$ and $X_k \neq Y_k$ for some k. Then letting $X_{-k} = (X_1,...,X_{k-1},X_{k+1},...,X_n)$ and $E_{X_{-k}}$ denote expectation with respect to the joint distribution of X_{-k}, we wish to estimate $\overline{RR}_{ij} = E_{X_{-k}}[RR_{ij}(X,Y)]$. 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 $Y(b)$ is the same as $X(b)$ except for one covariate of interest approximate the mean relative risk as

$$\overline{RR}_{ij} \approx \sum_b RR_{ij}(X(b),Y(b)).$$

Bootstrapped standard errors and the bias-corrected percentile method[47,48] were used to construct 95% confidence intervals for these statistics.

Results

Parameter Estimates

Table 1 gives maximum likelihood estimates (MLEs) and 95% confidence intervals (CIs) for all parameters associated with the underlying Markov model, namely the linear coefficients which determine the probability distribution over a patient's initial state and the Markov transition intensities. A $*$ denotes that the CI does not contain zero, indicating that the Markov model parameter estimates are statistically significant at the 5% level. Note that the coefficient for BMI is not significant for any of the associated Markov parameters while asthma is significant for all of them.

For each emission (observed lab or vital) in our model, there are two distinct sets of parameters: one associated with the moderate state and one with the severe state. Because we have modeled all emission distributions as Gaussian random variables, we learn a mean and variance parameter for both the moderate state and the severe state. Table 2 shows the emission distribution parameters (mean and standard deviation) for both latent states (moderate and severe). If there is a meaningful difference between the two latent states, we would expect the resulting emission distributions to differentiate the moderate and severe latent states. In Table 2 we also show the difference between the MLEs of the emission distribution parameters along with 95% confidence intervals. Again, a $*$ denotes that the CI does not contain zero, indicating that the difference between the severe and moderate state emission parameters are statistically significant at the 5% level. Of the 15 emission distributions, all mean parameters are significantly different at the 5% level other than those of platelets, systolic blood pressure and diastolic blood pressure. Similarly, all emission standard deviation parameters are significantly different at the 5% level other than platelets, systolic blood pressure, diastolic blood pressure, hemoglobin, and temperature. This indicates that our model is successfully learning to differentiate two distinct latent states which can be characterized by several vital and lab measurements.

Clinical Interpretation of Latent States

The National Institutes of Health COVID-19 treatment guidelines[50] served as our first baseline comparison to evaluate the correlation of our model's inferred latent disease states to clinically defined disease severity. Strong concordance was found between inferred emission distribution parameters of our HMM and the National Institutes of Health COVID-19 treatment guidelines. $SpO2$ is percent Oxygen saturation, which is denoted by $\%O2$Sat in our model. The mean (standard deviation) of the state-dependent distributions for this emission are $95.3\%(3.56\%)$ for the moderate state and $90.46\%(12.04\%)$ for the severe state. This is in close agreement with the NIH guidelines for determining disease severity. Note the large variance for the severe state emission distribution. This could be a result of severely ill patients having extremely low oxygen saturation levels in addition to severely ill patients having high oxygen saturation levels due to ventilation. Respiratory frequency is denoted by Respiration in our model. The mean (standard deviation) of the distributions for this emission are $19.02(3.14)$ breaths/minute for the moderate state and $25.8(5.36)$ breaths/minute for the severe state. We again have our model identifying the correct trend in mean values for the state-dependent emission distributions as specified by the NIH guidelines, namely respiratory frequency increases with more severe illness. Moreover the NIH threshold of 30 breaths/minute is within one standard deviation of the inferred severe state mean, suggesting that the model is in concordance with the NIH guidelines.

Comparing EDI[49] to our model's latent state estimates provided us another evaluation of the correlation of our model's inferred latent disease states to an externally validated disease severity metric. Figure 3 shows our model's real-time state estimations on two example patients. The left column corresponds to a patient that transitioned to the discharged state while the right column corresponds to a patient that transitioned to the deceased state. All figures are color-coded to show the latent state with the highest conditional probability at the given observation time. Blue indicates that the moderate state is more probable

($\alpha_t^i < 0.5$) while red indicates that the severe state is more probable ($\alpha_t^i \geq 0.5$). The raw data, which has been normalized to have zero mean and unit standard deviation, is given in the top row. The second row shows the model's real-time estimate of the probability of being in the severe state conditioned on all previously observed data. The bottom row shows the EDI for the two patients. The resulting z test statistic for this cohort was $z = 10.17$, giving a p-value of 2.28×10^{-20}. Repeating this experiment on ten random train/test splits with a median (min, max) sample size of 517 (511,523) resulted in a mean (std. dev.) z test statistic of 9.83 (0.73), which implies a mean (std. dev.) p-value of 2.03×10^{-15} (4.74×10^{-15}). This is strong evidence that the EDI and our model's inferred state estimation probabilities are positively correlated.

Risk Factor Analysis

Tables 3 provides bootstrap confidence intervals for the transition probabilities for various covariates. Confidence intervals for statistics associated with mean wait times are shown in Table 4. The vector Y changes a single covariate in each statistic computed. To observe the effect of age, the vector Y has the age covariate set to one standard deviation above the mean of the empirical distribution of age across the population. The same is done when testing for the effect of BMI. When testing sex, the male indicator covariate is set to one. When testing race the Black indicator covariate is set to one. When testing a particular comorbidity, the relevant indicator covariate is set to one. Confidence intervals that fall to one side of zero suggest a statistically significant effect of the covariate under consideration. Such confidence intervals are marked by a $*$.

The bootstrapped confidence intervals of the average relative risk $\overline{RR}_{ij}(X,Y)$, which gives the relative risk of transitioning from state i to state j for patients with covariate vectors Y (treatment) and X (control) are shown in Table 3. The column *moderate→severe* gives the population-averaged relative risk of transitioning from the moderate state to the severe state for the various covariate comparisons. Higher ages are associated with a modest increase in risk with a relative risk of 1.077 compared to younger patients. Being male is associated with a nearly two-fold increase in the probability of transitioning from a moderate disease state to a severe disease state relative to females. Similarly, being Black is associated with an increase in this same transition probability by a factor of 1.622. All comorbidities are associated with increased risks ranging from a low of 1.33 (asthma) to a high of 1.554 for hypertension. Higher BMI was associated with a modest decrease in risk (0.97), but this was not significant at the 5% level.

The column *severe→deceased* gives the population-averaged relative risk of transitioning from the severe state to the deceased state for the various covariate comparisons. Note that once we condition on being in the severe state, we see that several covariates are actually associated with a *decrease* in risk of transitioning to the deceased state. In particular, being male (0.324), Black (0.294) or having a comorbidity (0.305-0.408) are all associated with a decreased risk of transition from severe to deceased. Age (1.208) and BMI (1.041) are the only two covariates associated with increased risks of transitioning from severe to deceased. However note that BMI and asthma are not significant at the 5% level.

We find that the commonly cited risk factor BMI is not alone necessarily associated with adverse outcomes among hospitalized COVID-19 patients. In Table 3 we see that BMI is not significantly associated with disease progression, either from moderate to severe nor from severe to death. Instead we find that the comorbidities considered in this study (asthma, deabetes, hypertension and chronic kidney disease) are all associated with an increased risk of transitioning from moderate to severe disease states. To further investigate this we learned a similar model that did not contain comorbidities as covariates. In this model higher BMI did show a statistically significant increase in the probability of transitioning from moderate to severe illness. One interpretation of this result is that high BMI alone does not elevate the risk of adverse outcomes in COVID-19 patients, but various comorbidities that are correlated with high BMI may elevate this risk. Finally, we note that age was the only risk factor that was associated with an increase in transition from moderate to severe and from severe to death.

The third column of Table 3, labeled *entry→deceased* gives the relative risk of ending in the deceased state at the time of hospitalization. Note that despite the differential impacts of covaraites between disease states (namely both increasing and decreasing the risk depending on current disease state), the overall relative risk of ending in the deceased state is still increased for some of these covariates. In particular, being Black is associated with an increase in relative risk of mortality by a factor of 1.271, which was found to be significant at the 5% level. Higher age was also found to be associated with an increased risk of ending in the deceased state significant at the 5% level. While being male was associated with an increased risk of mortality, it was not found to be significant at the 5% level. Similarly, diabetes, hypertension and kidney disease were all associated with moderate increases in the risk of mortality, but they were not significant at the 5% level. Higher BMI and asthma were both associated with a slight decrease in the relative risk of mortality, though not at the 5% level.

Table 4 shows the population-averaged ratio of time spent in a given state before transitioning out of said state. Being male, being Black, having asthma, having diabetes and having hypertension were all associated with increased time spent in both the moderate and severe disease states, all significant at the 5% level. Male patients on average spend over 1.7 times more time in the moderate state and over 1.5 times more time in the severe state than females. Black patients on average spend over 1.2 times more time in the moderate state and over 1.5 times more time in the severe state than White patients. For patients with hypertension, time in the moderate state increased by a factor of 1.7 and a factor of 1.45 in the severe state compared to

	Intercept	Age	Male	Black	Other race
moderate	0.733 (0.595, 1.777)*	0.098 (0.004, 0.620)*	-0.310 (-0.837, 0.651)	-0.248 (-0.799, 0.300)	0.134 (-0.318, 0.692)
moderate→severe	-3.979 (-4.461, -2.062)*	-0.022 (-0.223, 0.110)	0.110 (-0.528, 0.469)	0.242 (-0.337, 0.585)	-0.174 (-0.703, 0.311)
severe→moderate	-2.375 (-2.498, -1.612)*	0.113 (-0.003, 0.343)	-0.149 (-0.638, 0.032)	-0.236 (-0.740, 0.012)	-0.558 (-1.456, -0.346)*
severe→deceased	-2.077 (-2.932, -1.856)*	0.230 (-0.046, 0.509)	-1.573 (-1.919, -1.415)*	-1.678 (-1.811, -1.451)*	-1.441 (-1.842, -1.221)*
moderate→discharged	-2.631 (-2.757, -2.315)*	-0.145 (-0.314, 0.020)	-1.740 (-1.829, -1.534)*	-1.661 (-1.799, -1.512)*	-1.441 (-1.597, -1.312)*

(a) Intercept and Demographics

	BMI	Asthma	Hypertension	Diabetes	Kidney Disease
moderate	0.069 (-0.188, 0.356)	0.404 (0.026, 1.598)*	0.206 (0.003, 0.642)*	-0.294 (-0.725, 0.200)	-0.886 (-1.619, -0.743)*
moderate→severe	0.000 (-0.076, 0.161)	-0.492 (-0.861, 0.311)	-0.102 (-0.339, 0.281)	0.093 (-0.196, 0.410)	0.255 (-0.049, 0.647)
severe→moderate	-0.049 (-0.159, 0.037)	-0.728 (-1.677, -0.387)*	-0.141 (-0.338, 0.200)	-0.001 (-0.368, 0.171)	0.194 (-0.129, 0.544)
severe→deceased	-0.024 (-0.313, 0.101)	-1.791 (-2.053, -1.397)*	-1.415 (-1.735, -1.247)*	-1.311 (-1.630, -1.157)*	-1.190 (-1.430, -1.003)*
moderate→discharged	0.050 (-0.067, 0.194)	-1.568 (-1.753, -1.351)*	-1.578 (-1.679, -1.369)*	-1.387 (-1.495, -1.247)*	-1.342 (-1.486, -1.195)*

(b) Body Mass Index and Comorbidities

Table 1. Maximum likelihood estimates and 95% CIs for Markov model parameters. Table (a) contains parameters associated with the intercept term and demographics. Table (b) contains parameters associated with the body mass index and comorbidities. The first row of each table corresponds to the parameters governing the initial state probability distribution, specifically the probability that a patient is admitted to the hospital in the moderate disease state. Subsequent rows give parameters that determine the log transition intensities between states . A $*$ denotes that the CI does not contain zero, indicating that the parameter estimates are statistically significant at the 5% level.

patients without hypertension, while patients with diabetes spent 1.36 times more time in the moderate state and 1.232 times more time in the severe state compared to patients without diabetes. Patients with asthma had the largest relative increases in time-in-state. On average they spend over more than twice as long in both the moderate state and the severe state compared to patients without asthma. BMI and chronic kidney disease were not associated with significant differences in the relative time spent in each disease state. Higher age was associated with a decrease in time spent in the severe disease state, spending 0.762 times as much time as younger patients.

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 differential impacts on disease progression depending on the disease state of the patient. The risk factors that demonstrated this pattern were being male, Black and having a medical co-morbidity.

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 the severe to deceased, and an increase in the overall relative risk of ending in the deceased state. 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, the fact that being male is associated with an increase in the relative risk of transitioning from moderate to severe means that on average males are more likely to be in the severe state during the course of hospitalization compared to females. Similarly females are more likely to transition to the discharged state from the moderate disease state than males. But once we condition on being in the severe state and assess the risk of transitioning to the deceased state, a type of internal selection bias occurs where we only account for those males and females that are sick enough to make such a transition. Looking at only those individuals in the severe disease state it can happen (as is the case with sex, race 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 1 shows a graphical representation of this situation.

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 use to assign risk stratification of the the two patients, 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.

	moderate: mean (sd)	severe: mean (sd)	severe - moderate: mean	severe - moderate: sd
CRP	7.616 (15.921)	13.981 (19.357)	6.365 (5.908, 7.752)*	3.437 (3.235, 4.558)*
BUN	26.949 (48.897)	46.979 (66.744)	20.03 (19.112, 27.625)*	17.847 (16.59, 22.293)*
LDH	293.603 (435.701)	412.222 (633.666)	118.619 (103.827, 175.606)*	197.965 (136.994, 373.51)*
Procalcitonin	0.275 (2.955)	6.531 (30.125)	6.257 (6.142, 12.462)*	27.17 (31.41, 42.88)*
Ferritin	540.846 (1170.917)	1005.174 (1737.225)	464.328 (215.532, 719.221)*	566.308 (360.748, 857.727)*
Anion gap	9.664 (12.761)	11.566 (14.369)	1.902 (1.681, 2.587)*	1.608 (1.531, 1.966)*
D-dimer	503.539 (2042.108)	3571.766 (9842.44)	3068.227 (2741.306, 5145.355)*	7800.332 (6535.207, 11623.335)*
% O2 Sat	95.303 (96.217)	90.457 (104.692)	-4.846 (-7.282, -4.104)*	8.475 (7.274, 11.083)*
Hemoglobin	11.83 (14.068)	11.485 (14.18)	-0.345 (-1.694, -0.105)*	0.112 (-0.156, 0.312)
Platelets	240.794 (352.091)	252.7 (390.568)	11.906 (-4.725, 38.576)	38.477 (19.823, 88.839)*
Systolic Pressure	125.506 (144.128)	121.404 (145.255)	-4.102 (-24.917, 1.449)	1.128 (-0.795, 5.449)
Diastolic Pressure	70.554 (83.49)	68.954 (84.357)	-1.6 (-13.316, 0.88)	0.868 (-0.507, 3.995)
Respirations	19.021 (24.438)	25.799 (26.663)	6.778 (6.562, 7.431)*	2.226 (-2.413, 2.815)
Temperature	98.188 (99.014)	98.93 (100.932)	0.741 (-0.56, 0.917)	1.919 (-1.495, 2.475)
Urine Output	329.539 (567.028)	516.239 (2839.693)	186.7 (116.68, 615.943)*	2272.665 (1044.5, 6726.363)*

Table 2. The first two columns give the maximum likelihood estimates of the emission distribution parameters (mean and standard deviation of a normal distribution) for the moderate and severe states. The last two columns give bootstrapped 95% CIs for the differences between the severe and moderate means and standard deviations. A $*$ denotes that the CI does not contain zero, indicating that the difference between the severe and moderate state emission parameters are statistically significant at the 5% level.

Relative Risk of Disease Progression (95% CI)

Covariate: Y/X	moderate $\rightarrow$ severe	severe $\rightarrow$ deceased	entry $\rightarrow$ deceased
Age: High/Low	1.077 (1.011, 1.273)*	1.208 (0.941, 1.730)	1.256 (1.066, 1.575)*
Sex: Male/Female	1.905 (1.247, 2.402)*	0.324 (0.203, 0.455)*	1.255 (0.611, 1.534)
Race: Black/White	1.622 (1.185, 1.906)*	0.294 (0.202, 0.456)*	1.271 (1.107, 1.608)*
BMI: High/Low	0.970 (0.879, 1.029)	1.041 (0.668, 1.273)	0.960 (0.775, 1.096)
Asthma: Yes/No	1.331 (1.159, 1.658)*	0.408 (0.274, 0.880)*	0.973 (0.685, 1.402)
Diabetes: Yes/No	1.487 (1.204, 1.719)*	0.332 (0.232, 0.493)*	1.098 (0.748, 1.332)
Hypertension: Yes/No	1.554 (1.216, 1.827)*	0.353 (0.270, 0.508)*	1.085 (0.807, 1.356)
Kidney Disease: Yes/No	1.474 (1.220, 1.692)*	0.305 (0.215, 0.454)*	1.147 (0.920, 1.496)

Table 3. Maximum likelihood estimate and 95% bootstrapped confidence intervals (CI) for the relative risk of disease progression between two cohorts. A * indicates that the CI does not contain the value 1, indicating a statistically significant difference at the 5% level in experiencing the event between the cohorts. For Age and BMI Low/High are defined as one standard deviation below/above the population mean.

Relative Average Time-in-State (95% CI)

Covariate: Y/X	moderate	severe
Age: High/Low	1.126 (0.852, 1.508)	0.762 (0.507, 1.055)
Sex: Male/Female	1.706 (1.351, 2.043)*	1.564 (1.311, 2.200)*
Race: Black/White	1.273 (0.945, 1.464)	1.577 (1.208, 2.436)*
BMI: High/Low	0.971 (0.748, 1.111)	1.091 (0.960, 1.444)
Asthma: Yes/No	2.177 (1.100, 2.943)*	2.448 (1.757, 5.657)*
Diabetes: Yes/No	1.355 (1.058, 1.662)*	1.232 (1.036, 1.665)*
Hypertension: Yes/No	1.720 (1.309, 2.049)*	1.452 (1.086, 1.688)*
Kidney Disease: Yes/No	1.142 (0.789, 1.561)	1.002 (0.719, 1.328)

Table 4. Maximum likelihood estimates and 95% bootstrapped confidence intervals (CI) for the relative average time spent in both latent disease states. A * indicates that the CI does not contain the value 1, indicating a statistically significant difference at the 5% level. For age and body mass index (BMI) low and high values are defined as one standard deviation below and above the population mean, respectively.

Such a conditional risk stratification 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 of $(0.959, 2.037)$, which is still rather strong evidence that higher age is associated with a higher relative risk of death. Moreover, higher age was the only factor found to decrease average time spent in the severe state, likely due to the fact that such patients are transitioning to the deceased state more quickly. These results are consistent with the well documented elevated risk that older COVID-19 patients face.

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 towards high BMI. The self-reported prevalence of obesity (BMI > 30) in the state of Ohio is 34.8%[51], whereas the prevalence of obesity in the current dataset is 60%. Thus even though we find that BMI is not associated with elevated 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 co-morbidities 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. The high degree of correlation with EDI, which was validated on much larger datasets, suggests that that our approach is capable of transferring to other cohorts of patients. Another validation of our model's state estimation results was found by comparing the inferred emission distribution parameters of our model to NIH clinical guidelines on identifying COVID-19 disease severity. Nevertheless, future work should validate the proposed model 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, the current study provides a unique modeling-based approach on the progression of disease. The proposed approach performs risk forecasting and stratification based on the full patient trajectory and serves as an exploratory tool for generating novel clinical hypothesis. We estimated the parameters of our proposed ct-HMM based on a cohort of $1,362$ hospitalized SARS-CoV-2 positive patients using Expectation-Maximization (EM) algorithm. Because the disease burden states in our model are unobserved latent variables, we validate our model in two ways: We demonstrate the model's ability to infer meaningful underlying disease states via comparisons with 1) National Institute of Health (NIH) established guidelines and 2) a widely known and proprietary risk metric called the Epic Deterioration Index. We demonstrate that our proposed model infers underlying disease states that are in general agreement with NIH guidelines on disease severity. We also show a statistically significant correlation between our model's probabilistic risk assessments of disease state and the Epic Deterioration Index. 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 conditional on current disease state. Such differences should be taken into consideration when performing risk stratification among hospitalized patients. Such information is potentially more actionable throughout the course of care, possibly leading to better patient outcomes.

Appendix

Markov Jump Process

A finite state Markov jump process with state-space $\mathscr{S}$ is fully-characterized by an initial state probability distribution $\pi = (\pi_1, \pi_2, ..., \pi_{|\mathscr{S}|})$ over $\mathscr{S}$ and a transition intensity matrix Q that governs the rates of transitions between the states of $\mathscr{S}$. 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 $i \in \mathscr{S}$ 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 goes to 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 $i = 1, 2, ..., |\mathscr{S}|$ and the stochastic transitions are governed by the matrix Q as described above.

To capture heterogeneity in disease progression, we can model the intensities as functions of patient-specific covariates. For a patient k, suppose X_k is a vector of M patient covariates, such as demographics and comorbidities. Then, we model the

	Discharged: 1245 (91.4%)	Deceased: 117 (8.6%)	Total: 1,362
Sex: Female	603 (48.4%)	55 (47.0%)	658 (48.3%)
Sex: Male	642 (51.6%)	62 (53.0%)	704 (51.7%)
Race: White	894 (71.8%)	96 (82.1%)	990 (72.7%)
Race: Black	215 (17.3%)	15 (12.8%)	230 (16.9%)
Race: Other	136 (10.9%)	6 (5.1%)	142 (10.4%)
Co-morbidity: None	533 (42.8%)	28 (23.9%)	561 (41.2%)
Co-morbidity: Asthma	92 (7.4%)	6 (5.1%)	98 (7.2%)
Co-morbidity: Hypertension	553 (44.4%)	76 (65.0%)	629 (46.2%)
Co-morbidity: Diabetes	367 (29.5%)	43 (36.8%)	410 (30.1%)
Co-morbidity: Kidney Disease	129 (10.4%)	29 (24.8%)	158 (11.6%)
Age (Years)	62.0 (49.0 - 73.0)	78.0 (69.0 - 85.0)	63.0 (50.0 - 74.0)
Body Mass Index (kg/m^2)	31.6 (27.3 - 37.3)	29.8 (25.5 - 34.4)	31.6 (27.025 - 37.275)
CRP (mg/dL)	7.3 (3.025 - 13.2)	14.85 (8.025 - 20.525)	8.1 (3.4 - 14.4)
BUN (mg/dL)	23.0 (16.0 - 35.5)	44.0 (28.0 - 72.0)	25.0 (17.0 - 40.0)
LDH (U/L)	293.0 (230.0 - 372.0)	364.0 (270.0 - 504.0)	297.0 (235.5 - 385.0)
Procalcitonin (ng/mL)	0.19 (0.1 - 0.52)	0.51 (0.22 - 2.61)	0.22 (0.103 - 0.69)
Ferritin (ng/mL)	459.0 (207.5 - 866.0)	685.0 (442.5 - 1203.5)	500.0 (231.0 - 939.75)
Anion Gap (mmol/L)	10.0 (8.0 - 12.0)	11.0 (9.0 - 13.0)	10.0 (8.0 - 12.0)
D-dimer (ng/mL)	449.0 (266.0 - 859.0)	1023.5 (510.5 - 2841.75)	494.0 (287.0 - 991.0)
% O2 Sat (%)	95.7 (92.0 - 98.0)	94.0 (91.0 - 97.0)	95.0 (92.0 - 97.8)
Hemoglobin (g/dL)	12.1 (10.2 - 13.5)	10.7 (8.8 - 12.4)	12.0 (10.0 - 13.4)
Platelets ($10^9/L$)	228.0 (167.0 - 307.0)	197.0 (145.75 - 260.0)	222.0 (164.0 - 301.0)
Systolic Pressure (mmHg)	123.0 (110.0 - 137.0)	121.0 (106.5 - 138.0)	123.0 (109.0 - 137.0)
Diastolic Pressure (mmHg)	70.5 (62.0 - 79.75)	64.0 (56.0 - 73.5)	69.0 (61.0 - 79.0)
Respirations (breaths/min)	20.0 (18.0 - 24.0)	23.0 (19.8 - 26.667)	20.0 (18.0 - 24.0)
Temperature (Fahrenheit)	98.2 (97.8 - 98.8)	98.2 (97.7 - 98.8)	98.2 (97.8 - 98.8)
Urine Output (mL)	300.0 (200.0 - 450.0)	200.0 (2.0 - 300.0)	300.0 (200.0 - 450.0)

Table 5. Data Summary: All patient-level covariates and observable vitals and labs (emissions) are summarized by discharged and deceased end-states (columns one and two, repectively) and in aggregate (column three). Numeric values are summarized by the median and interquartile range of the reference population. Binary indicator variables are summarized by the total number and proportion within the reference population. Units for vitals and labs are provided.

transition rate q_{ij} for $i \neq j$ as $q_{ij}(X_k) = e^{w_{ij} \cdot X_k}$, where $w_{ij} \in \mathbb{R}^M$ is a vector of parameters and $w_{ij} \cdot X_k$ denotes the dot product between vectors w_{ij} and X_k. Because of the restrictions placed on transitions between latent states, we must have $q_{03} = q_{12} = 0$. Because states 2 and 3 are absorbing states, we must have $q_{2i} = q_{3i} = 0$ for $i \in \{0,1,2,3\}$.

We similarly assume that the initial state probabilities are functions of patient-specific covariates. Letting $\pi = (\pi_0, \pi_1, \pi_2, \pi_3)$ we assume $\pi_i(X_k) = \frac{e^{v_i \cdot X_k}}{\sum_i e^{v_i \cdot X_k}}$ for $i \in \{0,1,2,3\}$. Note that hospitalized patients cannot be in the discharged state or the deceased state when initially admitted, so we set $\pi_2 = \pi_3 = 0$. This leaves only one set of parameters to be learned for π_0 since $\pi_1 = 1 - \pi_0$.

Transition Probability Matrix

The probability transition function is given by the solution to the Kolmogorov forward equation.

$$P'(t) = QP(t)$$

The Markov infinitesimal generator Q has ijth entry q_{ij}. Recall that we define $q_i = \sum_{j \neq i} q_{ij}$ and entry $q_{ii} = -q_i$. Our model has a simple intensity matrix Q which is as follows.

$$Q = \begin{pmatrix} -q_0 & q_{01} & q_{02} & 0 \\ q_{10} & -q_1 & 0 & q_{13} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{2}$$

In this case the Kolmogorov forwrad equations can be solved exactly, the solution of which is

$$\begin{aligned}
p_{00}(t) &= \frac{1}{2}\left(1 + \frac{q_1 - q_0}{\lambda_1 - \lambda_2}\right) e^{\lambda_1 t} + \frac{1}{2}\left(1 - \frac{q_1 - q_0}{\lambda_1 - \lambda_2}\right) e^{\lambda_2 t} \\
p_{01}(t) &= \frac{q_{01}}{\lambda_1 - \lambda_2} e^{\lambda_1 t} - \frac{q_{01}}{\lambda_1 - \lambda_2} e^{\lambda_2 t} \\
p_{02}(t) &= \frac{q_{02}}{2\lambda_1}\left(1 + \frac{q_1 - q_0}{\lambda_1 - \lambda_2}\right)\left(e^{\lambda_1 t} - 1\right) + \frac{q_{02}}{2\lambda_2}\left(1 - \frac{q_1 - q_0}{\lambda_1 - \lambda_2}\right)\left(e^{\lambda_2 t} - 1\right) \\
p_{03}(t) &= \frac{q_{13} q_{01}}{\lambda_1(\lambda_1 - \lambda_2)}\left(e^{\lambda_1 t} - 1\right) - \frac{q_{13} q_{01}}{\lambda_2(\lambda_1 - \lambda_2)}\left(e^{\lambda_2 t} - 1\right) \\
p_{10}(t) &= \frac{q_{10}}{(\lambda_1 - \lambda_2)} e^{\lambda_1 t} - \frac{q_{10}}{(\lambda_1 - \lambda_2)} e^{\lambda_2 t} \\
p_{11}(t) &= \frac{1}{2}\left(1 - \frac{q_1 - q_0}{\lambda_1 - \lambda_2}\right) e^{\lambda_1 t} + \frac{1}{2}\left(1 + \frac{q_1 - q_0}{\lambda_1 - \lambda_2}\right) e^{\lambda_2 t} \\
p_{12}(t) &= \frac{q_{02} q_{10}}{\lambda_1 - \lambda_2}\left(\frac{1}{\lambda_1}(e^{\lambda_1 t} - 1) - \frac{1}{\lambda_2}(e^{\lambda_2 t} - 1)\right) \\
p_{13}(t) &= \frac{q_{13}}{2(\lambda_1 - \lambda_2)}\left(\frac{q_0 - q_1 + \lambda_1 - \lambda_2}{\lambda_1}(e^{\lambda_1 t} - 1) - \frac{q_0 - q_1 - \lambda_1 + \lambda_2}{\lambda_2}(e^{\lambda_2 t} - 1)\right)
\end{aligned}$$

, where $\lambda_1 = \frac{-(q_0+q_1)+\sqrt{(q_0-q_1)^2+4q_{10}q_{01}}}{2}$ and $\lambda_2 = \frac{-(q_0+q_1)-\sqrt{(q_0-q_1)^2+4q_{10}q_{01}}}{2}$ are the eigenvalues of the matrix Q. This gives us a closed form representation of the probability transition matrix $P(t)$ for any set of patient covariates, which can then be used to compute the likelihood function.

Define the limiting distributions $p^*_{03} = \lim_{t\to\infty} p_{03}(t)$ and $p^*_{13} = \lim_{t\to\infty} p_{13}(t)$, which can easily be derived using the above solutions.

$$\begin{aligned}
p^*_{03} &= \lim_{t\to\infty} p_{03}(t) = \frac{q_{01} q_{13}}{q_0 q_1 - q_{01} q_{10}} \\
p^*_{13} &= \lim_{t\to\infty} p_{13}(t) = \frac{q_0 q_{13}}{q_0 q_1 - q_{01} q_{10}}
\end{aligned}$$

One interesting quantity is the ratio of these two limiting probabilities: . This gives a sense of the relative risk between the moderate disease burden and severe disease burden states for a particular patient. Then higher values $\frac{p^*_{03}}{p^*_{13}}$ indicate that the limiting distributions are similar, which means the death rates for medium-severity and high-severity patients in this cohort are similar. Then for these patients being in the high risk state is not that much worse than being in the medium risk state. On the

other hand, for lower values of $\frac{p^*_{03}}{p^*_{13}}$, the high risk state is relatively much more dangerous. This ratio has another interpretation as well. Note that

$$\frac{p^*_{03}}{p^*_{13}} = \frac{q_{01}}{q_{01}+q_{02}}.$$

The right-hand side of this equation is the probability of transitioning to the severe risk state, conditioned on transitioning out of the moderate disease state at a given time t.

Hitting Probabilities

The *hitting time* of a state in a continuous-time Markov chain is a random variable that gives the time-to-first-entry into the state given a particular starting state. More precisely, we define the hitting time of state $r \in \mathscr{S}$ of a Markov chain S_t with initial state $S_t = s$ as $T_{sr} = \inf\{t > 0 | S_t = r\}$. Expected hitting times associated with the discharged and deceased states are not defined, since there are two absorbing states. Namely, the expected hitting times of these two states is unbounded.

The hitting *probabilities* are defined however. Denote by $\mathscr{A} = \{\text{discharged}, \text{deceased}\}$ the set of end (absorbing) states of the Markov chain, and define $\tau = \inf\{t > 0 | S_t \in \mathscr{A}\}$ to be the hitting time of the set $\mathscr{A}$. Then, the hitting probability of the deceased state conditioned on starting in state s is the conditional probability $P(S_\tau = 3 | S_0 = s)$. The set of hitting probabilities over all starting states forms a vector v that is the minimal non-negative solution to $Qv[s] = 0$ for all $s \neq 3$ and $v[3] = 1$. The solution to this is easily shown to be

$$v = \left(\frac{q_{13}q_{01}}{q_0 q_1 - q_{01}q_{10}}, \frac{q_{13}q_0}{q_0 q_1 - q_{01}q_{10}}, 0, 1 \right)^T = (p^*_{03}, p^*_{13}, 0, 1)^T.$$

Notice that they correspond to the limiting distributions p^*_{03} and p^*_{13}. This is again because of the fact that deceased and discharged states are both absorbing.

The way to interpret these probabilities is that if you currently are in state s, then $v[s]$ is the probability you eventually wind up in the deceased state. Such a probability can be used as measure of risk of an adverse outcome for hospitalized patients. However, because the latent states moderate and severe are unobserved, we can't definitively know what the current state of the patient is. We can instead compute these hitting probabilities conditioned on any observations we have made of the patients up to that point. Specifically, if we have a covariate vector X_i for patient i and a sequence of observations $Y^i_{1:n} = (Y^i_{t_1}, Y^i_{t_2}, ..., Y^i_{t_n})$ made at times $t_1, t_2, ..., t_n$, then we compute the hitting probabilities conditioned on the patient covariates and all previously collected data. For a given patient i with covariate vector X_i, let τ_i denote the hitting time of the absorbing states $\mathscr{A}$, and let v_i denote the hitting probabilities. We then compute the following.

$$\begin{aligned}
P(S^i_{\tau_i} = 3 | Y^i_{1:t}, X_i) &= \sum_{s \in \mathscr{S}} P(S^i_{\tau_i} = 3, S^i_t = s | Y^i_{1:t}, X_i) \\
&= \sum_{s \in \mathscr{S}} P(S^i_{\tau_i} = 3 | S^i_t = s, Y^i_{1:t}, X_i) P(S^i_t = s | Y^i_{1:t}, X_i) \\
&= \sum_{s \in \mathscr{S}} P(S^i_{\tau_i} = 3 | S^i_t = s) P(S^i_t = s | Y^i_{1:t}, X_i) \\
&= \sum_{s \in \mathscr{S}} v^i[s] \alpha^i_t[s] \\
&= v^T_i \cdot \alpha^i_t
\end{aligned}$$

Note that given parameter estimates for the ct-HMM we can easily compute these probabilities for any new patient. Moreover these probabilities can be updated each time new observations are made, whether vital measurements or lab results. In this way the conditional hitting probability of the deceased state acts as a real-time forecast of an adverse outcome for hospitalized COVID-19 patients.

Note that we can also estimate the expected length of stay in the hospital for a given patient. In terms of the underlying Markov chain, this is the mean first-passage time to the absorbing set $\mathscr{A}$, $\bar{\tau}_i = E[\tau_i]$, where $\tau_i = \inf\{t > 0 | S_t \in \mathscr{A}, S_0 = i\}$. The set of these values, $\bar{\tau} = (\bar{\tau}_0, \bar{\tau}_1, \bar{\tau}_2, \bar{\tau}_3)$ is given by the solution to a set of linear equations.

$$\begin{aligned}
\bar{\tau}_0 &= \frac{q_{01} - q_{11}}{q_{00}q_{11} - q_{10}q_{01}} \\
\bar{\tau}_1 &= \frac{q_{10} - q_{00}}{q_{00}q_{11} - q_{10}q_{01}} \\
\bar{\tau}_2 &= 0 \\
\bar{\tau}_3 &= 0
\end{aligned}$$

Likelihood Function

Given a probabilistic model which generates random variables X given model parameters θ, the likelihood function $L(\theta|x)$ is the full joint probability distribution $P(X = x|\theta)$ taken as a function of model parameters θ given fixed observed data x. Maximizing the likelihood function $L(\theta|x)$ with respect to θ gives the maximum likelihood estimator $\hat{\theta}(x)$ of the parameters θ.

When transitions in a Markov jump processes are unobserved, which occurs when observation times do not coincide with transition times, the likelihood function depends on the Markov probability transition function $P(j,t|i,s)$, which gives the probability of the process being in state $j \in \mathscr{S}$ at time $t > 0$, given that the process was in state $x \in \mathscr{S}$ at time $s < t$. Denote by $P(t|s)$ the matrix whose i,j entry is $P(j,t|i,s)$. According to the Kolmogorov equations, the matrix $P(t|s)$ is determined by an *intensity matrix* $Q(u)$ with i,j entry $q_{ij}(u)$ defined for all $u \in [s,t]$. Specifically, given such an intensity matrix $Q(u)$, the probability transition matrix $P(t|s)$ is the unique solution to the matrix differential equation $\frac{\partial}{\partial t}P(t|s) = P(t|s) \cdot Q(t)$, with initial condition $P(s|s) = \mathbf{I}_n$, where $\mathbf{I}_n$ the identity matrix of size n. When the transition intensity matrix Q is constant we say the Markov jump process is time-homogeneous, and the solution to the Kolmogorov equation is given by $P(t|s) = e^{(t-s)Q}$ where $e^X = \sum_{k=0}^{\infty} \frac{1}{k!}X^k$ is the exponential matrix function for matrix X. In general this exponential matrix function must be approximated numerically, but in our case it is available in closed form. Details can be found in the supplementary material.

Suppose a Markov jump process with initial state probability π and transition intensity matrix Q is observed at times $\mathscr{T} = \{t_1, t_2, ..., t_n\}$, and $S_{t_j} \in \mathscr{S} = \{0, 1, 2, ..., |\mathscr{S}| - 1\}$ is the state of the Markov jump process at time $t_j \in \mathscr{T}$, which is the jth observation time. We do not assume that the observation times coincide with transition times. Suppose there are K emissions associated with each state of the Markov jump process, and denote the kth probability density function for observation y at state s by $f_k(y|\theta_s^k)$, where θ_s^k are the parameters defining the distribution. Let $Y_{t_j} = (y_{t_j}^1, y_{t_j}^2, ..., y_{t_j}^K)$ be the complete set of emissions at time t_j where $y_{t_j}^k$ is the kth emission at time t_j. Let $S = (S_{t_1}, S_{t_2}, ..., S_{t_n})$ be the vector of all of Markov states, and let $Y = (Y_{t_1}, Y_{t_2}, ..., Y_{t_n})$ be the complete set of all observed emissions. Finally, let $\theta = \theta_s^k)s,k$ be the set of all emission parameters and let $\phi = (\pi, Q, \theta)$ be the set of all model parameters. Then the full joint likelihood function can be decomposed as follows:

$$L(\phi|S,Y) = P(S,Y|\phi) = P(Y|S,\theta)P(S|\pi,Q) = P(S_{t_1}|\pi)P(Y_{t_1}|S_{t_1},\theta)\prod_{j=2}^{n} P(S_{t_j}|S_{t_{j-1}},Q)P(Y_{t_j}|S_{t_j},\theta)$$

The marginal likelihood is given by

$$L(\phi|Y) = \sum_S L(\phi|Y,S),$$

where the sum is taken over all possible sequences of S. Computing this sum is prohibitively expensive in all but the smallest problems. However, the marginal likelihood can be computed much more efficiently by recursive computations that leverage the dependence structure of the probabilistic model. See Scott (2002)[44] for details.

If M distinct Markov processes are observed according to the emission process θ, and each jump process is independent and identically distributed according to (π, Q), then we simply index all quantities by the number of processes observed and take the product over all individual likelihood functions. Specifically, we let S^i and Y^i be the sets of all Markov states and observed emissions for $i = 1, ..., M$. Then the full joint likelihood is $\prod_{i=1}^{M} L(\phi|S^i, Y^i)$. The marginal likelihood is obtained by taking the sum over all possible sequences of every Markov jump process, $\sum_{S^1} \cdots \sum_{S^M} \prod_{i=1}^{M} L(\phi|S^i, Y^i)$.

Observed States and Transition Times

Observation times typically do not coincide with transition times. It is assumed that observation times are non-informative and random. However we wish to consider cases where some states are observed directly (without emissions). We also wish to consider cases where transition times are observed. Specifically, we must account for observed transitions into a known absorbing states discharged and deceased. These are fully observed states and the exact time of transition into these states are known. Observed states can be handled by including indicator random variables as emissions in the set E_{t_i}. To account for observed transition times we must incorporate this additional information into the likelihood function.

Suppose times t_1 and t_2 are not associated with known transition times, but observation times. Suppose time t_3 is associated with a known transition time into a fully observed absorbing state $j \in \mathscr{S}$. The likelihood function for this patient is

$$P(E,S,\theta,\Lambda) = \pi(S_{t_1})P(E_{t_1}|S_{t_1})P(S_{t_2}|S_{t_1})P(E_{t_2}|S_{t_2})P(S_{t_3}|S_{t_2})P(E_{t_3}|S_{t_3}). \tag{3}$$

Because times t_1 and t_2 are associated with observation times we have

$$P(S_{t_2} = j|S_{t_1} = i) = P_{ij}(t_2 - t_1)$$

, where $P_{ij}(t_2 - t_1) = \left[e^{(t_2-t_1)Q}\right]_{ij}$. On the other hand, because t_3 is an observed transition time, there is more information that needs to be accounted for in the likelihood function. In particular, it is known that a transition occurs at time t_3. This can not be

said of time t_1 or t_2. To account for this additional information, consider the following, which follows from the properties of continuous-time homogeneous Markov chains. Because a transition occurs at time t_3, we can find a value $\varepsilon > 0$ such that with probability $1 - O(\varepsilon)$ the patient is in state $i \in N_{\text{in}}[j]$ for all $t \in (t_3 - \varepsilon, t_3)$, where $N_{\text{in}}[j] = \{i \in \mathscr{S} | q_{ij} > 0\}$ is the set of all states which can transition into state j. We can then approximate the term $P(S_{t_3}|S_{t_2})$ as

$$\begin{aligned} P(S_{t_3} = j | S_{t_2} = k) &\approx \sum_{i \in N_{\text{in}}[j]} P(S_{t_3-\varepsilon} = i | S_{t_2} = k) P(S(t) = i \; \forall t \in (t_3 - \varepsilon, t_3) | S(t_3 - \varepsilon) = i) P(S(t_3) = j | S(t) \text{ leaves } i \text{ at time } t_3) \\ &= \sum_{i \in N_{\text{in}}[j]} P_{ki}(t_3 - t_2 - \varepsilon) q_i e^{-\varepsilon q_i} \frac{q_{ij}}{q_i} \\ &= \sum_{i \in N_{\text{in}}[j]} P_{ki}(t_3 - t_2 - \varepsilon) q_{ij} e^{-\varepsilon q_i}. \end{aligned}$$

Taking the limit of the right hand side as $\varepsilon \downarrow 0$, this approximation becomes more exact, giving us the following

$$P(S_{t_3} = j | S_{t_2} = k) = \sum_{i \in N_{\text{in}}[j]} P_{ki}(t_3 - t_2) q_{ij}.$$

Parameter Initialization Sensitivity Analysis

Figure 4 provides the log likelihood values from 250 random initializations. Subfigure (a) shows the full training trajectory while subfigure (b) shows only the final values. No random initialization significantly outperformed the zero initialization, with the best initialization only resulting in a 0.1% increase in the log likelihood.

Disease State Estimation

Estimates of the latent disease states make up a sequence of probability distributions known as the forward variables in the HMM literature. Suppose patient i has N total observations at times $t_1, t_2, ..., t_N$. Let the latent disease state of patient i at time t be denoted by S_t^i, let the covariate vector of patient i be denoted by X_i, and let all of patient i's observed data up to the nth observation be $Y_{1:n}^i$. An estimate of patient i's disease state at the nth observation is given by a probability distribution over the possible states, which we denote by the vector α_n^i. For ease of notation we label the states *moderate*, *severe*, *discharged*, and *deceased* as 0, 1, 2, and 3 respectively. With this notation, the probability that patient i is in state $s = 0, 1, 2, 3$ at the nth observation is given by

$$\alpha_n^i[s] = P(S_{t_n}^i = s | Y_{1:n}^i, X_i, \phi).$$

Note that, because the end states are observed, if the patient is still hospitalized, the probability of being in one of the two end states is zero. Thus, the only non-zero values in the vector α_t^i will be the probability that the patient is in the moderate disease state $\alpha_t^i[0]$ and the probability that the patient is in the severe disease state $\alpha_t^i[1]$. Moreover, because we must have $\alpha_t^i[0] = 1 - \alpha_t^i[1]$, the single probability $\alpha_t^i[1]$ is sufficient for assessing the patient's current disease state. All forward probabilities can be computed via a forward pass of the forward-backward algorithm, a standard computational technique in HMM inference[44]. This recursive computation is used when computing the marginal likelihood for the observed data or can be computed for out-of-sample patient data.

References

1. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. & Gloaguen, R. Covid-19 pandemic prediction for hungary; a hybrid machine learning approach. *Mathematics* **8**, 890 (2020).
2. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj Digit. Medicine* **4**, 1–5 (2021).
3. Vaid, A. *et al.* Federated learning of electronic health records to improve mortality prediction in hospitalized patients with covid-19: Machine learning approach. *JMIR medical informatics* **9**, e24207 (2021).
4. Li, S. *et al.* Development and external evaluation of predictions models for mortality of covid-19 patients using machine learning method. *Neural Comput. Appl.* 1–10 (2021).
5. Pourhomayoun, M. & Shakibi, M. Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Heal.* **20**, 100178 (2021).
6. Nguyen, S. *et al.* Budget constrained machine learning for early prediction of adverse outcomes for COVID-19 patients. *Nat. Sci. Reports (to appear)* DOI: 10.21203/rs.3.rs-593801/v1 (2021).

7. Smarr, B. L. *et al.* Feasibility of continuous fever monitoring using wearable devices. *Sci. Reports* **10**, 21640, DOI: 10.1038/s41598-020-78355-6 (2020). Number: 1 Publisher: Nature Publishing Group.

8. Wongvibulsin, S. *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. *Annals Intern. Medicine* DOI: 10.7326/m20-6754 (2021).

9. Adler, F. R. & Liou, T. G. The dynamics of disease progression in cystic fibrosis. *PLoS One* **11**, e0156752 (2016).

10. Alzakerin, H. M., Halkiadakis, Y. & Morgan, K. D. Autoregressive modeling to assess stride time pattern stability in individuals with huntington's disease. *BMC neurology* **19**, 1–6 (2019).

11. Liu, Y.-Y., Li, S., Li, F., Song, L. & Rehg, J. M. Efficient learning of continuous-time hidden markov models for disease progression. *Adv. neural information processing systems* **28**, 3599 (2015).

12. Futoma, J., Hariharan, S. & Heller, K. Learning to detect sepsis with a multitask gaussian process rnn classifier. In *International Conference on Machine Learning*, 1174–1182 (PMLR, 2017).

13. Meng, R. *et al.* Nonstationary multivariate Gaussian processes for electronic health records. *J. Biomed. Informatics* **117**, 103698, DOI: https://doi.org/10.1016/j.jbi.2021.103698 (2021).

14. Choi, E., Schuetz, A., Stewart, W. F. & Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Informatics Assoc.* **24**, 361–370 (2017).

15. 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* **3**, e2029058–e2029058, DOI: 10.1001/jamanetworkopen.2020.29058 (2020). _eprint: https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2773971/rosenthal_2020_oi_200925_1610392201.69457.pdf.

16. Rossen, L. M., Branum, A. M., Ahmad, F. B., Sutton, P. & Anderson, R. N. Excess deaths associated with COVID-19, by age and race and ethnicity — united states, january 26–october 3, 2020. *MMWR. Morb. Mortal. Wkly. Rep.* **69**, 1522–1527, DOI: 10.15585/mmwr.mm6942e2 (2020).

17. The Covid Tracking Project. The Covid Tracking Project. https://covidtracking.com/. Accessed: 2021-02-02.

18. Harrison, S. L., Fazio-Eynullayeva, E., Lane, D. A., Underhill, P. & Lip, G. Y. H. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis. *PLOS Medicine* **17**, 1–11, DOI: 10.1371/journal.pmed.1003321 (2020). Publisher: Public Library of Science.

19. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning* (The MIT Press, 2009).

20. Satten, G. A. & Longini, I. M. Markov Chains With Measurement Error: Estimating the 'True' Course of a Marker of the Progression of Human Immunodeficiency Virus Disease. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **45**, 275–309 (1996). Publisher: [Wiley, Royal Statistical Society].

21. Bureau, A., Shiboski, S. & Hughes, J. P. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Stat. Medicine* **22**, 441–462 (2003).

22. Kay, R. A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies. *Biometrics* **42**, 855–865 (1986).

23. Conlon, A., Taylor, J. & Sargent, D. Multi-state models for colon cancer recurrence and death with a cured fraction. *Stat. Medicine* **33**, 1750–1766, DOI: 10.1002/sim.6056 (2014).

24. Santamaría, C., García-Mora, B., Rubio, G. & Navarro, E. A Markov model for analyzing the evolution of bladder carcinoma. *Math. Comput. Model.* **50**, 726–732, DOI: 10.1016/j.mcm.2008.12.019 (2009).

25. Dancourt, V. *et al.* Modeling recurrence in colorectal cancer. *J. Clin. Epidemiol.* **57**, 243–251, DOI: 10.1016/j.jclinepi.2003.07.012 (2004).

26. García-Mora, B., Santamaría, C., Navarro, E. & Rubio, G. Modeling bladder cancer using a Markov process with multiple absorbing states. *Math. Comput. Model.* **52**, 977–982, DOI: 10.1016/j.mcm.2010.02.036 (2010).

27. Lange, J. M., Hubbard, R. A., Inoue, L. Y. T. & Minin, V. N. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data: A Joint Model for Multistate Disease Processes and Random Informative Observation Times. *Biometrics* **71**, 90–101, DOI: 10.1111/biom.12252 (2015).

28. Chen, H. H., Duffy, S. W. & Tabar, L. A Markov Chain Method to Estimate the Tumour Progression Rate from Preclinical to Clinical Phase, Sensitivity and Positive Predictive Value for Mammography in Breast Cancer Screening. *J. Royal Stat. Soc. Ser. D (The Stat.* **45**, 307–317 (1996).

29. Duffy, S. W., Chen, H.-H. & Tabar, L. Estimation of Mean Sojourn Time in Breast Cancer Screening Using Markov Chain Model of Both Entry to and Exit from the Preclinical Detectable Phase. *Stat. Medicine* **14**, 1531–1543 (1995).

30. Uhry, Z. *et al.* Multi-state Markov models in cancer screening evaluation: a brief review and case study. *Stat. Methods Med. Res.* **19**, 463–486 (2010).

31. Yen, A. M., Chen, T. H., Duffy, S. W. & Chen, C.-D. Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. *Stat. Methods Med. Res.* **19**, 529–546, DOI: 10.1177/0962280209359862 (2010).

32. Kirby, A. J. & Spiegelhalter, D. J. Modeling the Precursors of Cervical Cancer. In Lange, N. (ed.) *Case Studies in Biometry* (Wiley-Interscience, 1994).

33. Soper, B. C., Nygård, M., Abdulla, G., Meng, R. & Nygård, J. F. A hidden Markov model for population-level cervical cancer screening data. *Stat. Medicine* **39**, 3569–3590, DOI: 10.1002/sim.8681 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8681.

34. Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W. & Couto, E. Multistate Markov models for disease progression with classification error. *J. Royal Stat. Soc. Ser. D (The Stat.* **52**, 193–209, DOI: 10.1111/1467-9884.00351 (2003).

35. Martino, A., Guatteri, G. & Paganoni, A. M. Multivariate Hidden Markov Models for disease progression. *Stat. Analysis Data Mining: The ASA Data Sci. J.* **13**, 499–507, DOI: https://doi.org/10.1002/sam.11479 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11479.

36. Powell, G. A., Verma, A., Luo, Y., Stephens, D. & Buckeridge, D. Modeling Chronic Obstructive Pulmonary Disease Progression Using Continuous-Time Hidden Markov Models. *Stud. Heal. Technol. Informatics* **264**, 920–924, DOI: 10.3233/SHTI190358 (2019).

37. Sukkar, R., Katz, E., Zhang, Y., Raunig, D. & Wyman, B. T. Disease progression modeling using hidden markov models. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, DOI: 10.1109/embc.2012.6346556 (IEEE, 2012).

38. Sun, Z. *et al.* A probabilistic disease progression modeling approach and its application to integrated Huntington's disease observational data. *JAMIA Open* **2**, 123–130, DOI: 10.1093/jamiaopen/ooy060 (2019).

39. Williams, J. P., Storlie, C. B., Therneau, T. M., Jr, C. R. J. & Hannig, J. A Bayesian Approach to Multistate Hidden Markov Models: Application to Dementia Progression. *J. Am. Stat. Assoc.* **115**, 16–31, DOI: 10.1080/01621459.2019.1594831 (2020). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2019.1594831.

40. Petersen, B. K. *et al.* Modeling sepsis progression using hidden markov models. In *NIPS Machine Learning for Health (ML4H)* (Long Beach, CA, 2017).

41. 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. Methods Med. Res.* **29**, 3409–3423, DOI: 10.1177/0962280220929045 (2020). Publisher: SAGE Publications Ltd STM.

42. Perveen, S., Shahbaz, M., Ansari, M. S., Keshavjee, K. & Guergachi, A. A Hybrid Approach for Modeling Type 2 Diabetes Mellitus Progression. *Front. Genet.* **10**, DOI: 10.3389/fgene.2019.01076 (2020). Publisher: Frontiers.

43. Fristedt, B. & Gray, L. *A Modern Approach to Probability Theory* (Birkhäuser Boston, 1997).

44. Scott, S. L. Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century. *J. Am. Stat. Assoc.* **97**, 337–351 (2002).

45. Baydin, A. G., Pearlmutter, B. A., Radul, A. A. & Siskind, J. M. Automatic Differentiation in Machine Learning: a Survey. *J. Mach. Learn. Res.* **8**, 1–43 (2018).

46. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

47. Efron, B. The Jackknife, the Bootstrap, and Other Resampling Plans | Department of Statistics. Tech. Rep. BIO 63, Stanford University (1980).

48. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Stat. Medicine* **19**, 1141–1164, DOI: https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F (2000).

49. Singh, K. *et al.* Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Annals Am. Thorac. Soc.* **18**, 1129–1137, DOI: 10.1513/annalsats.202006-698oc (2021).

50. National Institutes of Health. COVID-19 treatment guidelines panel. Coronavirus disease 2019 (covid-19) treatment guidelines. https://www.covid19treatmentguidelines.nih.gov. Accessed: 2021-03-23.

51. Center for Disease Control. Adult Obesity Prevalence Maps. https://www.cdc.gov/obesity/data/prevalence-maps.html. Accessed: 2021-05-13.

Acknowledgements

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-DRAFT. An abstract based on this work was accepted for presentation at the 2021 Midwest Clinical and Translational Research Meeting.

Author Contributions Statement

B.S. performed all theoretical analysis and mathematical derivations. B.S. and J.C. contributed equally to data ingestion, curation, software development and study design. B.S., J.C. and P.R. contributed to experiment and study design. All authors contributed to the analysis of the results and the manuscript preparation. B.S., J.C., R.C., S.N., J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: conceptualization; B.S., J.C., R.C., S.N., P.K., L.W., M.W., P.R.: data curation; B.S., J.C., R.C., S.N., P.K., L.W., M.W., P.R.: formal analysis; S.T.H., D.J.K., P.R.: funding acquisition; B.S., J.C., R.C., S.N., P.K., L.W., M.W., J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: investigation; B.S., J.C., R.C., S.N., P.K., L.W., M.W., J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: methodology; P.K., L.W., M.W., S.T.H., D.J.K., D.M., P.R.: project administration; P.K., L.W., M.W., J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: resources; B.S., J.C., R.C., S.N., P.K., L.W., M.W., P.R.: software; J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: supervision; B.S., J.C., R.C., S.N., P.K., L.W., M.W., P.R.: validation, B.S., J.C., R.C., S.N., J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: writing-original draft, B.S., J.C., R.C., S.N., P.K., L.W., M.W., J.M.D., S.T.H., J.H., D.J.K., D.M., P.R.: writing, review and editing. Final version was approved by all authors.

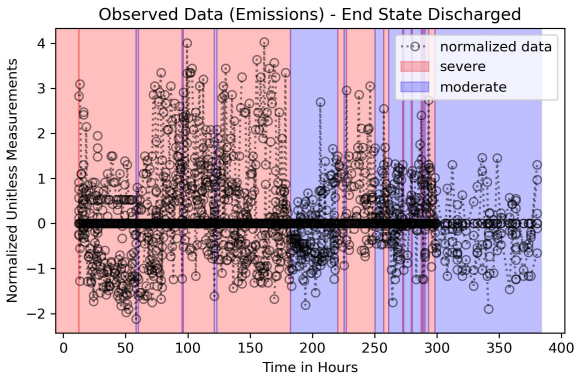
Additional Information

Competing Interests

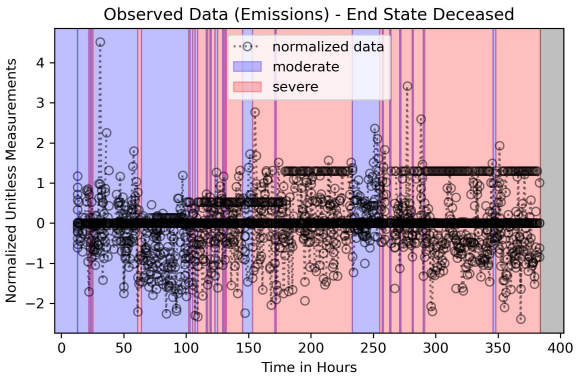
The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Informed Consent

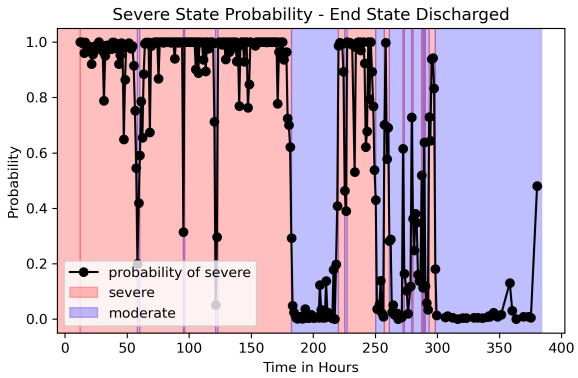
The study protocol involving analysis of fully de-identified data was reviewed and approved with Full Waiver of informed consent granted (Expedited, Category #5 research) by the respective Institutional Review Board's of ProMedica and Lawrence Livermore National Laboratory. The study was performed in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



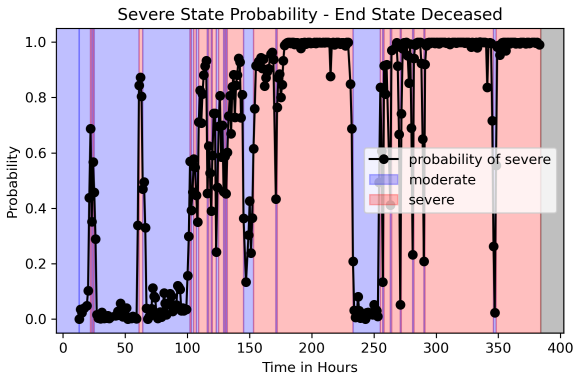
(a) Data of discharged patient



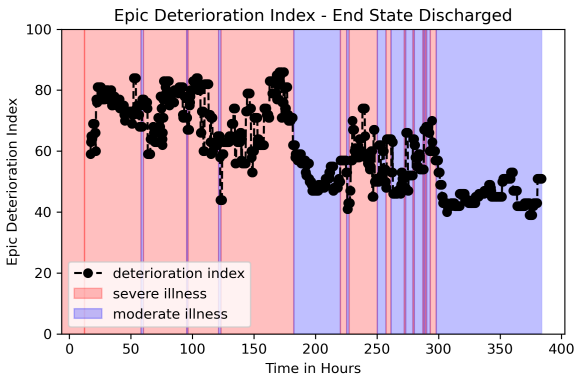
(b) Data of deceased patient



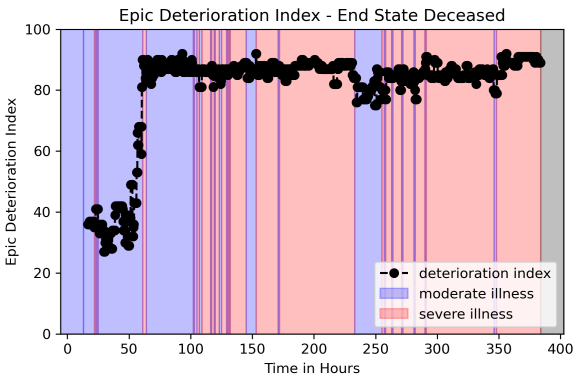
(c) PP sequence of discharged patient



(d) PP sequence of deceased patient

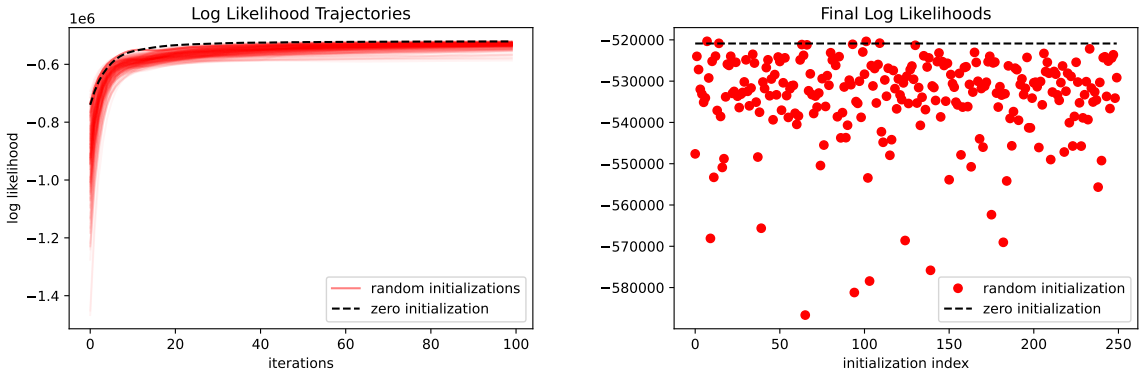


(e) EDI of discharged patient



(f) EDI of deceased patient

Figure 3. Data, posterior probability of severe illness and EDI for a patient that was discharged (left column) and a patient that died (right column). The background colors (red/blue) indicate which state the model infers as the most probable, conditioned on all previously observed data.



(a) Log likelihood trajectories

(b) Final log likelihood values

Figure 4. Results of initialization sensitivity analysis.