BMC Infectious Diseases

RESEARCH

Open Access

Biomarker states and risk of death among individuals hospitalized with SARS-CoV-2 infection

Tanayott Thaweethai[1,2*], Caitlin A. Selvaggi[1], Ta-Chou Ng[1], David Cheng[1,2], Tingyi Cao[1,3], Lori B. Chibnik[1,2,4], Daniel J. Shinnick[1] and Andrea S. Foulkes[1,2,3]

Abstract

Background Identifying individuals hospitalized for SARS-CoV-2 infection at increased risk of death is crucial for clinical decision making. Analyses must consider simultaneously the multitude of biomarkers across several domains and how these biomarker profiles change over time.

Methods This electronic health records-based study included individuals hospitalized at a Massachusetts General Brigham hospital for at least 24 h within 5 days prior and 30 days after diagnosis of COVID-19. K-means clustering was used to identify profiles among 20 eligible biomarkers and proportional hazards models were used to model 30-day mortality at hospitalization and 7 days after hospitalization (i.e., landmark models).

Results Twelve thousand, nine hundred forty-two individuals were included, among whom 1,198 died within 30 days. Six states were identified, characterized by the following abnormalities: (1) normal/reference, (2) hematologic, (3) inflammatory and hematological, (4) metabolic, (5) kidney, hematologic, and metabolic, and (6) cardiothrombotic, liver, and metabolic. Risk of death within 30 days was higher in States 3, 4, 5, and 6 (adjusted hazard ratios ranging from 3.6 to 7.8) compared to individuals in State 1 at hospitalization. Landmark model findings were similar.

Conclusions Distinct sub-phenotypes based on biomarker profiles were identified among patients hospitalized with SARS-CoV-2 infection, and certain phenotypes are associated with greater risk of 30-day mortality.

Keywords SARS-CoV-2, COVID-19, Biomarkers, Hospitalization, Mortality

Background

Worldwide, over 770 million individuals to date have been infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and over 7 million of these individuals have died from coronavirus disease 2019 (COVID-19) [1]. While as of May 5, 2023, the COVID-19 pandemic is no longer considered a Public Health Emergency of International Concern (PHEIC) [2], the risk of death in patients hospitalized for COVID-19 is still higher than the risk of death for seasonal influenza [3] and therefore remains a grave concern. Understanding who is at greatest risk of severe disease, particularly among hospitalized individuals, will provide the

*Correspondence:
Tanayott Thaweethai
tthaweethai@mgh.harvard.edu
[1] Massachusetts General Hospital Biostatistics, 399 Revolution Drive Ste 1068, Somerville, MA 02145, USA
[2] Department of Medicine, Harvard Medical School, 25 Shattuck St., Boston, MA 02115, USA
[3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave., Boston, MA 02115, USA
[4] Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Ave., Boston, MA 02115, USA

opportunity to focus limited clinical resources on those who are in greatest need of life saving interventions.

An extensive literature exists associating plasma biomarkers with severe outcomes, including intubation and mortality, among individuals hospitalized with SARS-CoV-2 infection [4–17]. Implicated in this literature are biomarkers of kidney injury, inflammation, cardiothrombotic dysfunction, liver abnormalities, hematologic abnormalities and metabolic dysfunction. With a few notable exceptions [18, 19], existing reports generally involve the evaluation of each plasma biomarker individually, without full consideration of the high degree of correlation and potential interactions across biomarkers. Moreover, in-hospital biomarker data are typically summarized using univariate measures, such as baseline or peak values. Consideration of multiple biomarkers simultaneously across domains, as well as evaluation of changes in biomarker profiles over time, is essential to develop clinically actionable risk groups. The primary objective of this manuscript was to evaluate the prognostic value of biomarker states, defined according to multiple laboratory measurements collected repeatedly over time, among hospitalized individuals with SARS-CoV-2 infection for predicting mortality during or following hospitalization. The secondary objective was to evaluate whether the same biomarker states were predictive of mortality when measured among individuals who had survived to seven days after hospitalization and whether having been in particular biomarker states between hospitalization and that point was also predictive of mortality. In this manuscript, electronic health records (EHR) data derived from 12,942 individuals hospitalized due to SARS-CoV-2 infection and corresponding vital records from the Massachusetts Registry of Vital Records and Statistics were evaluated to identify risk sets for 30-day survival based on biomarker profiles.

Methods

Study population and participants

Data were derived from a hospital-based cohort of $N=12{,}942$ individuals aged 18 and over at hospital admission with confirmed SARS-CoV-2 infection based on hospital record International Classification of Diseases, Tenth Edition (ICD-10) codes U07.1, B34,2, and B97.29 or positive quantitative reverse transcription polymerase chain reaction (RT-qPCR) tests between March 1, 2020 and November 30, 2021. All individuals were hospitalized at Massachusetts General Brigham (MGB), including Massachusetts General Hospital and Brigham and Women's Hospital, within 5 days prior to and 30 days after a positive SARS-CoV-2 test. Patients hospitalized for less than 24 h or with an unknown duration of hospitalization were excluded.

Exposures and outcome measures

All data were extracted from the MGB Enterprise Data Warehouse (EDW). Daily laboratory data on twenty routine biomarkers were analyzed, including blood urea nitrogen (BUN), creatinine, estimated glomerular filtration rate (eGFR), creatine phosphokinase (CPK), d-dimer, ferritin, C-reactive protein (CRP), white blood cell count (WBC), absolute lymphocyte count (ALC), hemoglobin, hematocrit (HCT), platelets, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), total bilirubin, albumin, glucose, anion gap, and lactate dehydrogenase (LDH). Laboratory measurements collected between the day of hospitalization and up to 30 days after hospitalization were included in the analysis. In the case that more than one measurement was recorded within a 24-hour period for an individual, the mean value within a day was used. Additional information on the data cleaning procedure is provided in the Supplemental Methods (Additional File 2). Death dates were obtained via hospital records and through the Massachusetts Registry of Vital Records and Statistics. In the case of an inconsistency between sources, the date listed in the Registry was used in analysis.

Statistical analysis

In the first stage of the analysis (dimension reduction), biomarker measurements on each individual-day of hospitalization were standardized by subtracting the sample mean and dividing by the sample standard deviation. K-means clustering for partially observed data (K-POD) was used to assign observations at the individual-day level to distinct groups based on minimizing the squared Euclidean distance of observed biomarker measurements within each cluster [20]. For each patient, all days for which at least one laboratory value was recorded were included in the cluster analysis. The gap statistic was used to determine the appropriate number of clusters [21], with confirmation via visual assessment of the elbow plot. The result of this dimension reduction step was a cluster assignment for each person on each day of hospitalization on which biomarker measurements were available.

In the second stage of the analysis (time-to-event analysis), Cox proportional hazards (Cox-PH) models were fitted to evaluate the association of biomarker state membership and time to death (defined as all-cause mortality). In Model 1, the association of the baseline state (i.e. at the time of hospitalization) and time from hospitalization to death was evaluated. Participants who did not die within 30 days of hospitalization were censored at 30 days. Because records for deaths that occurred outside of the hospital setting were available, it was not necessary to

censor patients who were discharged alive. Model 2 was a Cox-PH model with time-varying cluster assignments, where a patient's biomarker state membership has the potential to change when new lab assessments are performed on a given day during follow-up, but are otherwise carried forward. Model 3 was a Cox-PH landmark model in which the association between the state membership at the landmark time point and subsequent risk of mortality was evaluated among patients who survived up to the landmark and were not yet discharged at that time [22, 23]. A landmark time of 7 days was chosen as it is roughly the median length of stay for patients hospitalized with COVID-19 in the U.S. during the time period under study [24] and is therefore a reasonable time point to potentially update the information that informs survival prediction. Finally, an augmented landmark model, Model 4, was fitted in which an additional binary covariate was included for each state, indicating whether the patient was ever in a given state between hospitalization and the day prior to the landmark time. A summary of the statistical analysis approach is given in Fig. 1.

The observed transitions between clusters, hospital discharge, and death over time were calculated in unadjusted analyses. Hazard ratios and corresponding 95% confidence intervals were reported for each Cox-PH model. All Cox-PH models were adjusted for age, sex, race/ethnicity, body mass index (BMI) and presence of each of ten comorbidities (Supplemental Table 2, Additional File 1). These ten comorbidities include cancer, chronic kidney disease, chronic liver disease, chronic lower respiratory disease, dementia, heart disease, hypertension, immune disorders, stroke, and type 2 diabetes. These comorbidities were selected due to known associations with in-hospital mortality from COVID-19 in early studies from the UK and mainland China [25–27]. The predicted probability of death by 30 days from each model was estimated using the Breslow estimator of the baseline hazard function, which was then transformed into the survival function based on estimated hazard ratios. The discrimination of each Cox-PH model was assessed through the concordance statistic, or c-statistic [28–30]. Handling of missing data is summarized in the Supplemental Methods (Additional File 2).

The statistical significance level was set as 0.05. All analyses were conducted using R [31]. Survival models were fit using the *survival* package in R [28, 30].

Results

Cohort characteristics

Demographic characteristics of study participants overall and by death within 30 days, discharged and alive within 30 days, and still in hospital at 30 days after hospital admission, are provided in Table 1. As expected, the proportions of males, individuals ≥ 65

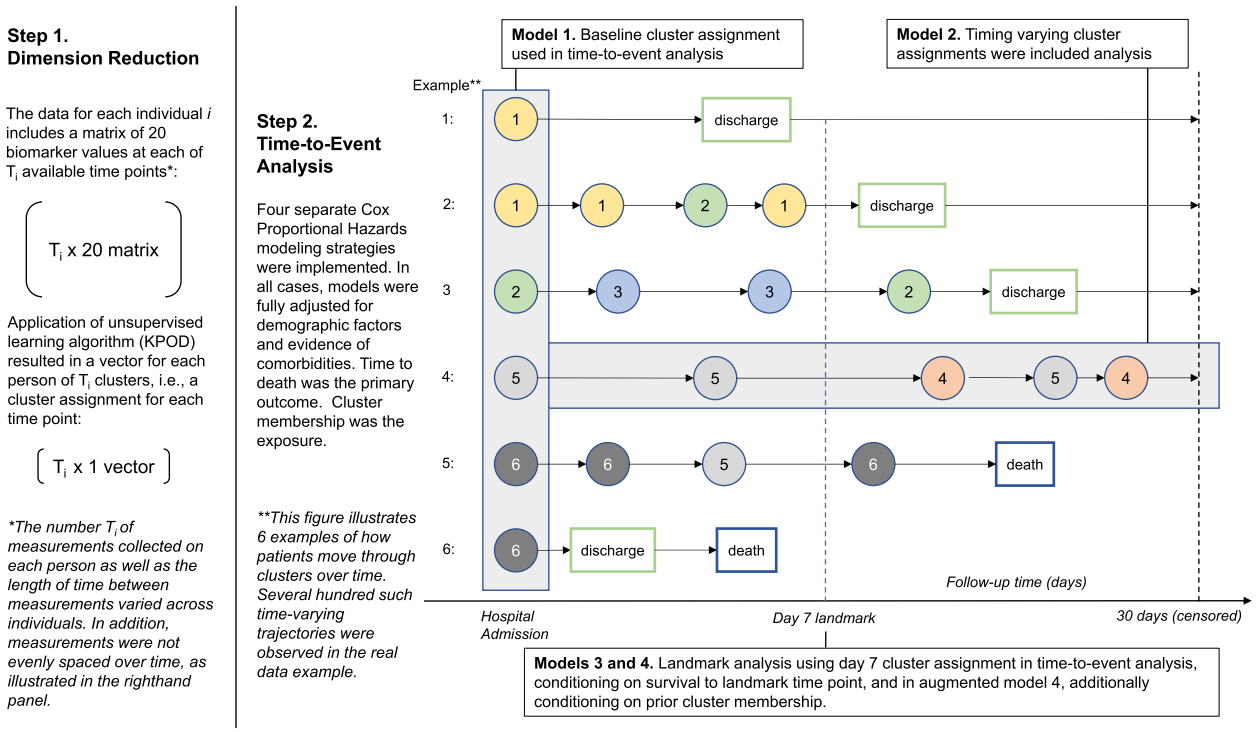

Fig. 1 Two-step analytic approach: dimension reduction followed by time-to-event analysis

Table 1 Demographic and clinical characteristics, by 30-day outcome category

	No. (%)			
		Outcome category		
Characteristic	**Total**	**Death < 30d**	**Discharge < 30d**	**In hospital ≥ 30d**
Overall	12,942 (100)	1,198 (9.3)	11,152 (86.2)	592 (4.6)
Age at admission, y				
18–49	3,944 (30.5)	61 (5.1)	3,745 (33.6)	138 (23.3)
50–64	3,512 (27.1)	191 (15.9)	3,113 (27.9)	208 (35.1)
> 64	5,486 (42.4)	946 (79.0)	4,294 (38.5)	246 (41.6)
Sex[a]				
Male/Other	6,642 (51.3)	710 (59.3)	5,568 (49.9)	364 (61.5)
Female	6,300 (48.7)	488 (40.7)	5,584 (50.1)	228 (38.5)
BMI, kg/m^2				
< 25	2,782 (21.5)	364 (30.4)	2,314 (20.7)	104 (17.6)
≥ 25 and < 30	3,572 (27.6)	309 (25.8)	3,102 (27.8)	161 (27.2)
≥ 30	4,567 (35.3)	334 (27.9)	4,004 (35.9)	229 (38.7)
Missing	2,021 (15.6)	191 (15.9)	1,732 (15.5)	98 (16.6)
Race/ethnicity[b]				
Asian, non-Hispanic	464 (3.6)	32 (2.7)	402 (3.6)	30 (5.1)
Black, non-Hispanic	1,518 (11.7)	130 (10.9)	1,309 (11.7)	79 (13.3)
Hispanic	2,658 (20.5)	136 (11.4)	2,384 (21.4)	138 (23.3)
White, non-Hispanic	7,572 (58.5)	826 (68.9)	6,448 (57.8)	298 (50.3)
Other/Unknown	730 (5.6)	74 (6.2)	609 (5.5)	47 (7.9)
Comorbidities				
0	5,516 (42.6)	321 (26.8)	4,957 (44.4)	238 (40.2)
≥ 1	7,426 (57.4)	877 (73.2)	6,195 (55.6)	354 (9.8)

[a] Sex was categorized based on administrative data in the electronic health record. Participants who were assigned neither male nor female at birth were assigned to the Male/Other category

[b] Race/ethnicity was categorized based on a combination of race and ethnicity data in the electronic health record. Individuals who reported more than one race were classified as Other/Unknown

years of age, and individuals with one or more comorbidity were greater in the group of individuals who die within 30 days of hospitalization or remain in hospital for at least 30 days than in the group of individuals who were discharged within 30 days. Summaries of the distribution of the baseline (at the time of hospitalization) and peak (maximum during hospitalization) laboratory values are provided in Supplemental Table 1 (Additional File 1). Comorbidity frequencies are given in Supplemental Table 2 (Additional File 1). The overall median length of hospital stay was 5 days [mean (standard deviation (SD)) = 8.5 (10.7), interquartile range (IQR) =(3, 10)]. Among individuals who died within 30 days, the median length of hospital stay was 8 days [mean (SD) = 9.9 (6.7), IQR = (4, 14)]. Among individuals who were discharged within 30 days, the median length of hospital stay was 5 days [mean (SD) = 6.4 (5.3), IQR = (3, 8)].

Biomarker state profiles

Six states were identified with distinct biomarker profiles, as summarized in Fig. 2. State 1 was characterized by relatively normal biomarker values; State 2 by hematological abnormalities; State 3 by inflammatory and hematological abnormalities; State 4 by metabolic abnormalities; State 5 by kidney, hematologic and metabolic abnormalities; and State 6 by metabolic, cardio-thrombotic and liver abnormalities. The transition matrix summarizing transitions in biomarker states between days in which at least one biomarker was observed or hospital discharge or death was recorded are provided in Supplemental Table 3 (Additional File 1). In total, 97,263 pairwise transitions were observed. Individuals tended to remain in the same state from one measured time point to the next, with State 5 having the highest proportion of patients remaining in the same state in a subsequent day. The proportion discharged

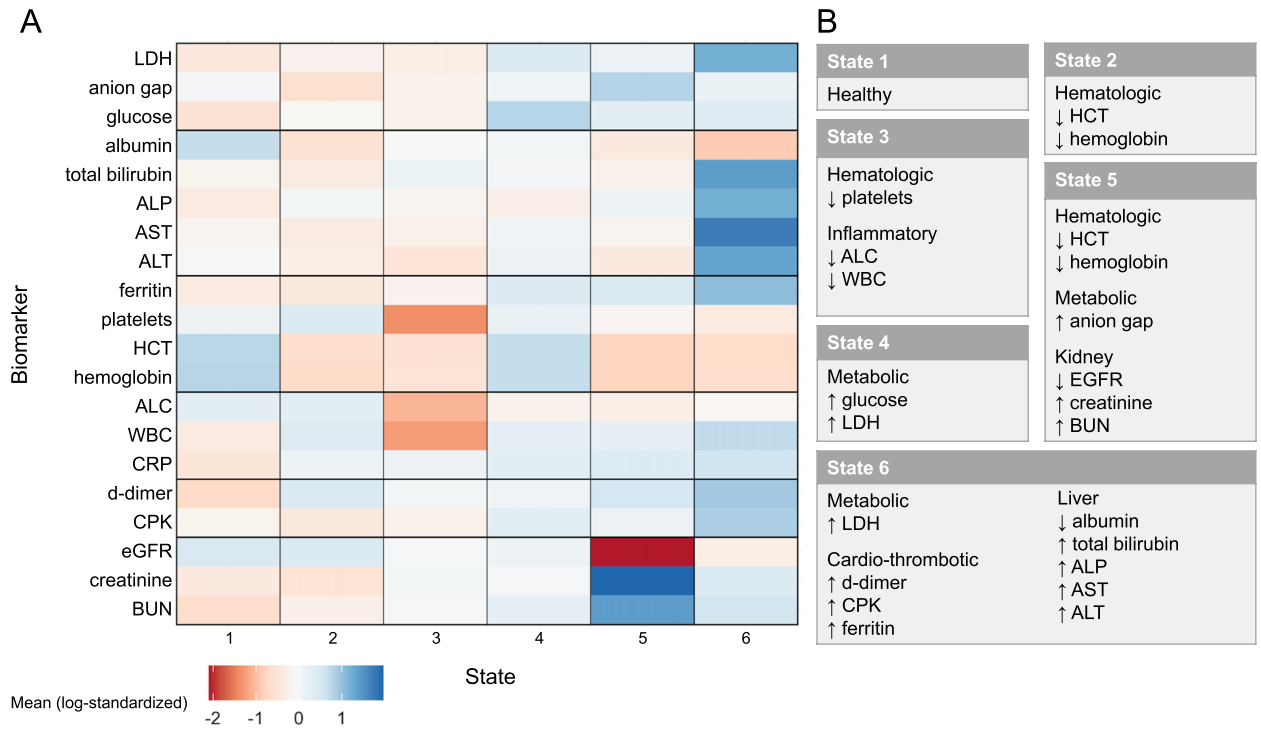

Fig. 2 States defined by plasma biomarker profiles. Biomarker states identified using unsupervised learning and all timepoints with at least one plasma biomarker measurement. A heatmap of all biomarkers by state is provided (**A**). Each state was characterized by a distinct profile of biomarker abnormalities, including hematologic, inflammatory, metabolic, kidney, liver, and cardio-thrombotic (**B**). Over time and during hospitalization, individuals transitioned between these six biomarker states, as described in Supplemental Table 3 (Additional File 1)

from State 1 was relatively high (19%) while the proportion discharged from States 5 and 6 were low (4.6% and 1.9%, respectively). Individuals in States 1 to 4 were more likely to transition between these two states than transition to States 5 and 6. The percentages transitioning from State 5 and 6 to death were 3.4% and 3.1%, respectively, while the corresponding death rates were between 0.2% and 1.3% for States 1 and 4, respectively.

Cox-PH models

The distribution of 30-day outcomes, stratified by biomarker state at baseline (i.e., time of hospitalization), are summarized in Table 2. Among individuals who were in State 1 at time of hospitalization, 5337/5636 (95%) were discharged within 30 days while 174/5636 (3%) died within 30 days. Among individuals in State 5 at hospitalization, only 610/901 (68%) were discharged while 242/901 (27%) died within 30 days. Similarly, among individuals in State 6 at hospitalization, 194/320 (61%) were discharged while 79/320 (25%) died within 30 days.

Adjusted hazard ratios for the baseline (Model 1) and time-varying (Model 2) Cox-PH models are shown in Table 3. The increased hazard of death, after adjustment for age, sex, race/ethnicity, BMI and presence of

Table 2 30-day outcomes, stratified by baseline biomarker state

	No. (%)			
		Outcome category		
Characteristic	Total	Death < 30d	Discharge < 30d	In hospital ≥ 30d
State 1	5,636	174 (3.1)	5,337 (94.7)	125 (2.2)
State 2	763	81 (10.6)	654 (85.7)	28 (3.7)
State 3	921	158 (17.2)	738 (80.1)	25 (2.7)
State 4	3,280	422 (12.9)	2,614 (79.7)	244 (7.4)
State 5	901	242 (26.9)	610 (67.7)	49 (5.4)
State 6	320	79 (24.7)	194 (60.6)	47 (14.7)

Patients who did not have at least one lab measurement on the day of hospital admissions are excluded from this table ($N = 1,121$)

Table 3 Hazard ratios for association between biomarker state and death within 30 days using 4 modeling approaches

	Hazard ratio (95% Confidence interval)				
				Augmented landmark model	
State	Baseline model	Time-varying model	Landmark model	Day 0–6	Day 7
State 1	*ref*	*ref*	*ref*	0.75 (0.63, 0.88)	*ref*
State 2	2.76 (2.12, 3.60)	3.44 (2.56, 4.62)	2.49 (1.88, 3.32)	0.96 (0.77, 1.18)	2.57 (1.92, 3.39)
State 3	3.50 (2.80, 4.37)	3.82 (2.74, 5.33)	3.93 (2.94, 5.23)	1.14 (0.93, 1.36)	4.28 (3.07, 5.79)
State 4	3.75 (3.13, 4.49)	6.83 (5.17, 9.02)	3.85 (2.97, 5.07)	1.23 (1.03, 1.46)	4.73 (3.61, 6.33)
State 5	6.40 (5.16, 7.93)	12.83 (9.49, 17.34)	7.46 (5.64, 10.04)	1.36 (1.07, 1.74)	8.09 (6.00, 10.97)
State 6	7.82 (5.96, 10.26)	16.04 (11.29, 22.79)	7.84 (5.76, 11.06)	1.24 (0.90, 1.67)	7.72 (5.60, 10.86)

Models are adjusted for age, sex, race/ethnicity, BMI, and presence of each of ten comorbidities (Supplemental Table 2, Additional File 1). The baseline model only uses biomarker data measured on the day of hospital admission, whereas the time-varying model uses biomarker data that is updated on days with new laboratory measurements. Patients who did not have at least one lab measurement on the day of hospital admission are excluded from the baseline model, but all patients are included in the time-varying model from the day of their first laboratory measurement until their death or censoring. Landmark models are restricted to patients who are alive and still hospitalized as of 7 days after hospitalization ($N = 5,193$). A patient can belong to any of the 6 states at the landmark time (day 7), but can have been in more than one of the other 5 states between hospitalization and the day before landmark (days 0–6)

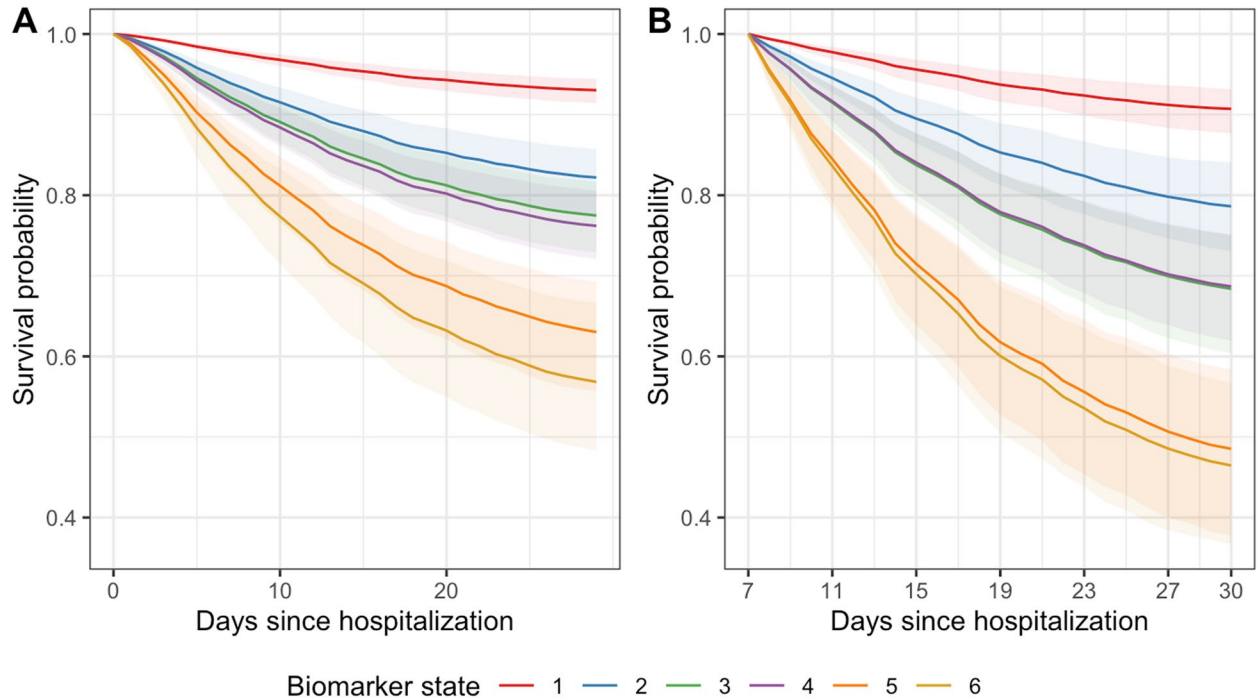

Fig. 3 Estimated 30-day survival curves, by biomarker state. Estimated survival curves are presented for a male patient who is age 65 + , White (non-Hispanic), has BMI < 25, and has none of the 10 comorbidities adjusted for in the model. The different curves correspond to the six different biomarker states the patient can belong to at the time of hospital admission (**A**) or at the day 7 landmark (**B**). Landmark model estimates are obtained from the non-augmented landmark model and additionally assume the patient is alive and not discharged by day 7 after hospital admission. Because these curves are based on Cox model estimates and are not Kaplan–Meier curves, numbers of individuals at risk at each time point are not provided on the x-axis

10 comorbidities, was statistically significantly greater for all states [State 2, adjusted hazard ratio (aHR) = 2.78, 95% confidence interval (CI) = (2.12,3.62); State 3, aHR = 3.50, 95% CI = (2.80,4.37); State 4, aHR = 3.74, 95% CI = (3.12,4.48); State 5, aHR = 6.37, 95% CI = (5.12,7.93); State 6, aHR = 7.80, 95% CI, (5.94,10.24)] as compared to the referent state (State 1). The c-statistic of the baseline model was 0.798 [95% CI: (0.786,0.807)]. The magnitude of all adjusted hazard ratios was greater in the time-varying model compared to the baseline model. The baseline model (Model 1) estimated survival curves by biomarker state at hospital admission are given in Fig. 3.

Adjusted predicted probabilities of death within 30 days by baseline biomarker state, age, sex and race/ethnicity are given in Supplemental Table 4 (Additional File 1). Predicted probabilities of death by 30 days ranged from approximately 1% for males and females in the younger age category (18-49 years) in State 1 to 43.5% in White/non-Hispanic males ≥ 65 years of age in State 6.

Results of fitting landmark models at 7 days after hospitalization are given in Table 3. In total, $N = 5{,}193$ individuals were alive and still hospitalized seven days after initial admission. The demographic and clinical characteristics of patients, stratified by whether they were included in the landmark models, are provided in Supplemental Table 5 (Additional File 1). Among patients who were excluded from the landmark models, the median time to death was 4 days (IQR: 2-5). Among patients who were included in the landmark models, 847 (16.3%) of whom died within 30 days following hospitalization and the median time to death was 14 days from admission (IQR: 10-20). Among these patients, the aHRs for mortality for each biomarker state compared to the referent state (State 1) were somewhat similar but attenuated in magnitude compared to those of the baseline model (Table 3). Augmenting the landmark model with indicator variables for whether the patient was in a given state prior to the landmark time did not substantially change the aHRs for the state a patient was in at the landmark time. However, having previously been in State 1 was associated with a reduced hazard of death [aHR = 0.75, 95% CI: (0.63,0.88)] while having previously been in States 4 or 5 was moderately associated with an increased hazard of death [State 4, aHR = 1.23, 95% CI = (1.03,1.46); State 5, aHR = 1.36, 95% CI = (1.08,1.73)]. The augmented model did not appear to greatly improve the concordance of the landmark model [Landmark c-statistic = 0.762, 95% CI = (0.749,0.776); Augmented landmark c-statistic = 0.768, 95% CI = (0.755,0.782)]. Estimated survival curves are presented for the landmark model in Fig. 3 and for the augmented landmark model in Fig. 4.

Discussion

Risk management in healthcare settings is a complex task inevitably requiring early and regular clinical evaluation and concurrent classification of patients into risk groups [32]. This manuscript evaluated the value of plasma biomarker profiles measured at hospital admission and over the course of hospitalization among individuals infected with SARS-CoV-2, as an approach to risk stratification. Leveraging the full spectrum of available data across all individuals and observed time points, six biomarker states were identified. While these states had overlapping characteristics, they were defined by unique biomarker profiles. For example, both States 2 and 5 were characterized by hematologic abnormalities (low HCT and low hemoglobin) but individuals in State 5 also exhibited metabolic (high anion gap) and kidney abnormalities (low eGFR, high creatinine and high BUN). State 1 was considered relatively healthy with normal biomarker values and was treated as the referent state for statistical modeling. All remaining states captured unique biomarker profiles, including State 6, which was characterized by a combination of metabolic, liver, and cardio-thrombotic abnormalities. Groupings of biomarkers into categories (e.g., metabolic, hematologic) was done only to enhance interpretation of the findings, and did not impact the analytic approach or results.

Compared to the referent state (State 1), each state at the time of hospitalization was predictive of an increased risk of mortality within a period of 30-days after adjustment for demographic factors and comorbidities with adjusted HRs ranging from 2.78 (State 2) to 7.80 (State 6). The estimated probability of death within 30 days of hospitalization was particularly high in White/non-Hispanic males ≥ 65 years of age who were in State 6 at baseline (estimated probability = 43.5%). The same estimated probability of death was 7.0% among White/non-Hispanic males ≥ 65 years of age who were in State 1 at baseline.

While individuals tended to remain in the same state over time, transitions between states were also frequent. Transition rates among States 1–4 ranged from 2.3% (State 1 to State 3) to 15% (State 1 to State 4). Transitions from State 6 to the other states ranged from 2% (State 6 to State 1) to 8% (State 6 to State 2). Nevertheless, landmark models that predict mortality rates conditional on survival and continued hospitalization at one week after hospital admission resulted in very similar aHRs, indicating that differences in biomarker profiles at baseline were similarly predictive of subsequent outcomes as those based on measurements at the 7-day landmark. The aHRs for the model consider biomarker states as a time-varying covariate had aHRs that were substantially greater than those of the baseline and landmark models, suggesting that an individual's most recent state is a stronger predictor of death than the state an individual was in at the time of hospitalization.

The higher degree of association between time-varying biomarker state and death may be a reflection of concurrent associations that capture changes in biomarkers resulting from the onset of severe outcomes, rather than the opposite [33]. As such, any findings from the time-varying model should be interpreted as associative, rather than causal. This is also due to the potential of time-varying confounding inherent to a time-varying Cox-PH model [34]. The analyses presented herein also

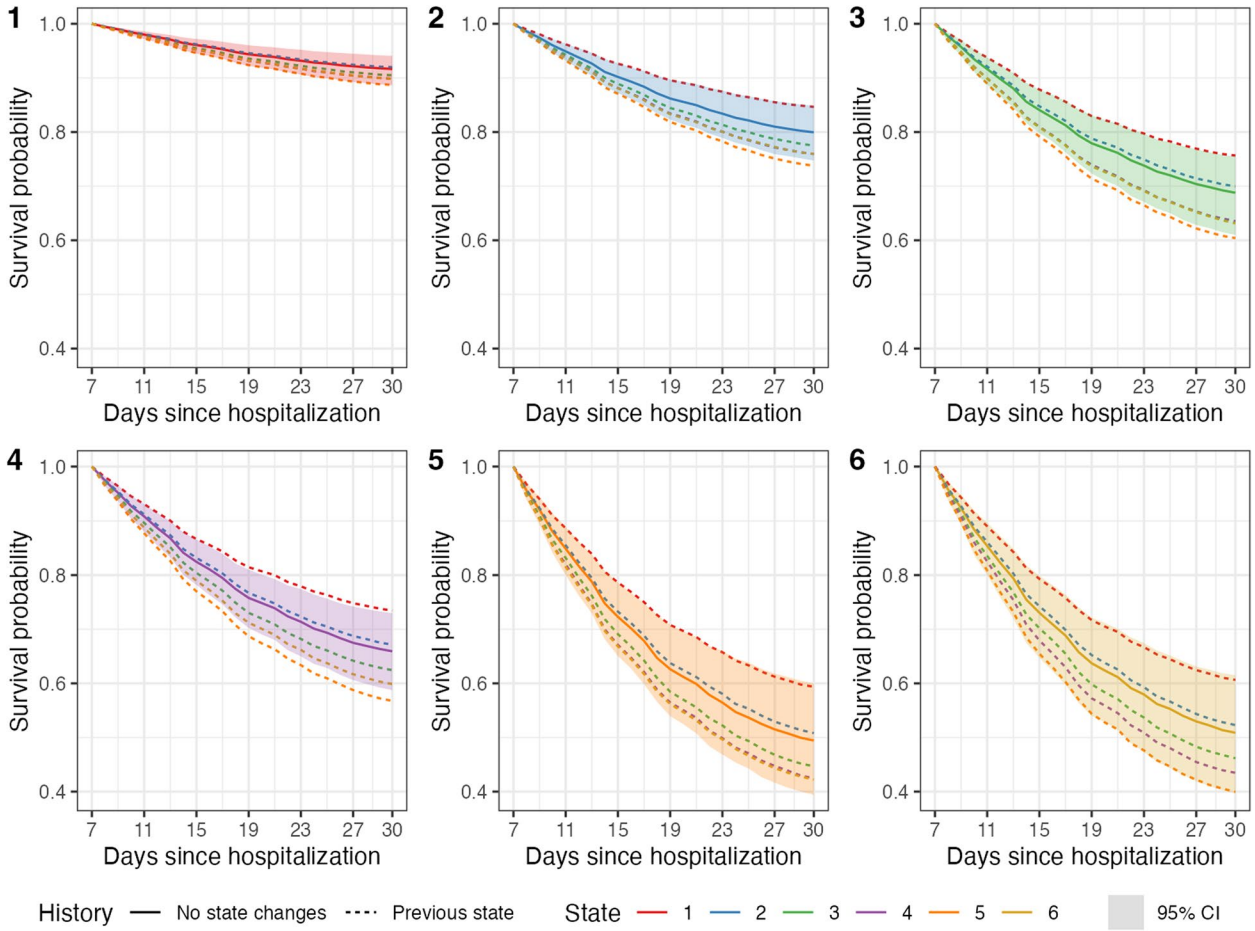

Fig. 4 Augmented landmark model survival curves, by biomarker state at 7-day landmark and by prior state membership. Estimated survival curves are presented for a patient who is age 18–49, White (non-Hispanic), has BMI < 25, has none of the 10 comorbidities adjusted for in the model, and is alive and not discharged by day 7 after hospital admission. Model estimates are obtained from the augmented landmark model. The solid curve within each panel corresponds to a patient who was in a given state at the landmark time and did not belong to any of the other states between days 0 and 6. The dashed lines within each panel correspond to a patient who was in a given state at the landmark time, but belonged to a different given state at some point between days 0 and 6 (as indicated by the color of the dashed line). For example, the solid orange line in panel (5) corresponds to a patient who was always in State 5 from hospital admission until the landmark time. The red dashed line in panel (5) corresponds to a patient who was in State 1 at some point between days 0 and 6 and was in State 5 at day 7 (but not in any other states besides those two). For visual clarity, 95% confidence intervals are only provided for the survival curve corresponding to a patient who was in the same state from hospital admission until the landmark time (i.e., the solid curve)

do not account for the impact of vaccinations or repeat infections. The time frame for data capture does cover periods in which vaccinations were available (December 2020-November 2021) and it is therefore expected that at least some of the individuals studied were vaccinated, and understanding how to best treat patients with COVID-19 who are likely to have some level of immunity against SARS-CoV-2 is essential [24].

Missing data is a ubiquitous challenge when conducting retrospective analyses of EHR that manifests in many ways [35, 36]. Because not every biomarker was measured every day, we used K-POD to account for missing biomarker measurements when defining biomarker states. Further, determining disease status based solely on the presence of a diagnostic (e.g. ICD-10) code, and assuming that patients without the diagnostic code do not have the comorbidity, can often lead to bias [37]. To address this, of patients without the diagnostic code, we sought to distinguish between patients with missing data (less frequent healthcare encounters over the prior year) and patients who truly do not have the disease (more frequent encounters), and multiply imputing only the comorbidity data that was truly missing.

The study is subject to some limitations, including challenges related to missing data. Patients who developed COVID-19 but did not have confirmed SARS-CoV-2 infection were excluded from this analysis. Death records obtained from EHR and from the Massachusetts

Registry of Vital Records and Statistics may not encompass all deaths that occurred within 30 days, especially if patients were transferred to a different hospital outside of the MGB catchment system or died out-of-state. Another limitation of the present study is that the study population is from a single hospital system, and so the findings may not be completely generalizable to other populations. The study findings may also be sensitive to the choice of comorbidities and how they were defined, but we believe the ten comorbidities selected are likely to capture most of the variability in COVID-19 mortality due to comorbidity burden. Mean lab values were used if multiple labs were reported on a single day in order to achieve dimension reduction of the data to the day-level, but it is possible that information regarding extreme values is lost in the process. Finally, cross-validation was not performed, but we do not anticipate that the estimated concordance of the various models are substantially inflated due to the large study sample size. Relatedly, the proportional hazards assumption was not checked; however, given the large sample size, hypothesis tests in this setting may detect proportional hazards assumptions that are not meaningful [38].

Conclusions

The significant associations between baseline biomarker states and mortality, after adjustment for demographic and clinical features, suggest biomarker profiles at time of hospitalization may help to define clinically relevant risk sets. In turn, these risk sets can be used to inform clinical decisions, potential interventions, and allocation of scarce resources, and ultimately could serve as stratification criteria for clinical trials.

Abbreviations

SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
COVID-19	Coronavirus disease 2019
EHR	Electronic health records
ICD-10	International Classification of Diseases, Tenth Edition
RT-qPCR	Quantitative reverse transcription polymerase chain reaction
MGB	Massachusetts General Brigham
EDW	Enterprise data warehouse
BUN	Blood urea nitrogen
eGFR	Estimated glomerular filtration rate
CPK	Creatine phosphokinase
CRP	C-reactive protein
WBC	White blood cell count
ALC	Absolute lymphocyte count
HCT	Hematocrit
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ALP	Alkaline phosphatase
LDH	Lactate dehydrogenase
K-POD	K-means clustering for partially observed data
Cox-PH	Cox proportional hazards
SD	Standard deviation
IQR	Interquartile range
aHR	Adjusted hazard ratio
CI	Confidence interval
BMI	Body mass index
c-statistic	Concordance statistic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-025-10651-w.

Additional file 1. Contains the five supplemental tables referenced in the main text: (1) Distribution of laboratory values, (2) Comorbidities by 30-day outcome, (3) Transitions between biomarker states, discharge, and death across all measured time points, (4) Predicted probabilities of death within 30 days of hospitalization, (5) Demographic and clinical characteristics, by whether they are included in the landmark model.

Additional file 2. Contains the Supplemental Methods, which provide additional detail on handling of missing data in the study.

Acknowledgements

We would like to thank Joyce Yan for her contributions to this analysis in obtaining and preparing the data for analysis. We would also like to acknowledge the Massachusetts Registry of Vital Records and Statistics for providing data regarding deaths in the Commonwealth of Massachusetts during the study period.

Clinical trial number

Not applicable.

Authors' contributions

TT planned and led the analysis, obtained the data, interpreted the findings, drafted the manuscript, and supervised the work. CAS obtained the data, performed the analysis, interpreted the findings, and drafted the manuscript. TN performed the analysis and edited the manuscript. DC, TC, LBC, and DJS edited the manuscript. ASF planned and led the analysis, interpreted the findings, drafted the manuscript, and supervised the work. All authors read and approved the final manuscript.

Authors' information

Not applicable.

Funding

All authors received support from NIH/NIGMS R01GM127862. The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data availability

The datasets used and/or analyzed during the current study are not publicly available but are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

This study was performed in line with the principles of the Declaration of Helsinki. The Partners HealthCare Institutional Review Board (IRB) (#2022P002986) approved collection of curated data based on data extractions from the EHRs on patients who receive care through the Mass General Brigham (formerly Partners) system. The IRB approved this study and a waiver of informed consent was granted.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 October 2023 Accepted: 14 February 2025
Published online: 24 February 2025

References

1. Geneva: World Health Organization. WHO COVID-19 Dashboard. 2020. Available from: https://covid19.who.int/.
2. World Health Organization. Statement on the fifteenth meeting of the IHR (2005) Emergency Committee on the COVID-19 pandemic. 2023.
3. Xie Y, Choi T, Al-Aly Z. Mortality in patients hospitalized for COVID-19 vs influenza in fall-winter 2023–2024. JAMA. 2024;331(22):1963–5.
4. Bivona G, Agnello L, Ciaccio M. Biomarkers for prognosis and treatment response in COVID-19 patients. Ann Lab Med. 2021;41(6):540–8.
5. Malik P, Patel U, Mehta D, Patel N, Kelkar R, Akrmah M, et al. Biomarkers and outcomes of COVID-19 hospitalisations: systematic review and meta-analysis. BMJ Evid Based Med. 2021;26(3):107–8.
6. Battaglini D, Lopes-Pacheco M, Castro-Faria-Neto HC, Pelosi P, Rocco PR. Laboratory biomarkers for diagnosis and prognosis in COVID-19. Front Immunol. 2022;13:857573.
7. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. JAMA. 2020;323(11):1061–9.
8. Ashktorab H, Pizuorno A, Adeleye F, Laiyemo A, Dalivand MM, Aduli F, et al. Symptomatic, clinical and biomarker associations for mortality in hospitalized COVID-19 patients enriched for African Americans. BMC Infect Dis. 2022;22(1):1–13.
9. Gupta D, Jain A, Chauhan M, Dewan S. Inflammatory markers as early predictors of disease severity in COVID-19 patients admitted to intensive care units: a retrospective observational analysis. Indian J Crit Care Med. 2022;26(4):482.
10. Silva BM, Assis LCSD, Batista Júnior MDC, Gonzalez NAP, Anjos SBD, Goes MA. Acute kidney injury outcomes in covid-19 patients: systematic review and meta-analysis. Braz J Nephrol. 2022;44(4):543-56.
11. Xu Z, Zhang Y, Zhang C, Xiong F, Zhang J, Xiong J. Clinical features and outcomes of COVID-19 patients with acute kidney injury and acute kidney injury on chronic kidney disease. Aging Dis. 2022;13(3):884.
12. Bowring MG, Wang Z, Xu Y, Betz J, Muschelli J, Garibaldi BT, et al. Outcome-stratified analysis of biomarker trajectories for patients infected with severe acute respiratory syndrome Coronavirus 2. Am J Epidemiol. 2021;190(10):2094–106.
13. La Porta E, Baiardi P, Fassina L, Faragli A, Perna S, Tovagliari F, et al. The role of kidney dysfunction in COVID-19 and the influence of age. Sci Rep. 2022;12(1):1–9.
14. Petrilli CM, Jones SA, Yang J, Rajagopalan H, O'Donnell L, Chernyak Y, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. BMJ. 2020;369:m1966.
15. Su L, Zhang J, Peng Z. The role of kidney injury biomarkers in COVID-19. Ren Fail. 2022;44(1):1280–8.
16. de Morais BF, Puga MAM, da Silva PV, Oliveira R, Dos Santos PCP, da Silva BO, et al. Serum biomarkers associated with SARS-CoV-2 severity. Sci Rep. 2022;12(1):1–9.
17. Ponti G, Maccaferri M, Ruini C, Tomasi A, Ozben T. Biomarkers associated with COVID-19 disease progression. Crit Rev Clin Lab Sci. 2020;57(6):389–99.
18. Boss AN, Banerjee A, Mamalakis M, Ray S, Swift AJ, Wilkie C, et al. Development of a mortality prediction model in hospitalised SARS-CoV-2 positive patients based on routine kidney biomarkers. Int J Mol Sci. 2022;23(13):7260.
19. Syed AH, Khan T, Alromema N. A hybrid feature selection approach to screen a novel set of blood biomarkers for early COVID-19 mortality prediction. Diagnostics. 2022;12(7):1604.
20. Chi JT, Chi EC, Baraniuk RG. k-pod: a method for k-means clustering of missing data. Am Stat. 2016;70(1):91–9.
21. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Royal Stat Soc Ser B (Statistical Methodology). 2001;63(2):411–23.
22. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. Scand J Stat. 2007;34(1):70–85.
23. van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. Boca Raton: CRC Press; 2011.
24. Meyerowitz EA, Scott J, Richterman A, Male V, Cevik M. Clinical course and management of COVID-19 in the era of widespread population immunity. Nat Rev Microbiol. 2024;22(2):75–88.
25. Russell CD, Lone NI, Baillie JK. Comorbidities, multimorbidity and COVID-19. Nat Med. 2023;29(2):334–43.
26. Guan W-J, Ni Z-Y, Hu Y, Liang W-H, Ou C-Q, He J-X, et al. Clinical characteristics of coronavirus disease 2019 in China. New Engl J Med. 2020;382(18):1708–20.
27. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ. 2020;369:m1985.
28. Therneau T. A package for survival analysis in R. R package version. 3.7-0. 2020;3(3). https://cran.r-project.org/package=survival.
29. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361–87.
30. Therneau TM. Modeling Survival Data: Extending the Cox Model. 1st edition 2000. In: Grambsch PM, editor. New York: Springer New York; Imprint: Springer. 2000.
31. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2021.
32. McGowan J, Wojahn A, Nicolini JR. Risk management event evaluation and responsibilities. StatPearls: StatPearls Publishing; 2022.
33. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health. 1999;20(1):145–57.
34. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ. 2017;359:j4587.
35. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479.
36. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013;1(3):1035.
37. Ford E, Rooney P, Hurley P, Oliver S, Bremner S, Cassell J. Can the use of bayesian analysis methods correct for incompleteness in electronic health records diagnosis data? Development of a novel method using simulated and real-life clinical data. Front Public Health. 2020;8:54.
38. Rulli E, Ghilotti F, Biagioli E, Porcu L, Marabese M, D'Incalci M, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. Br J Cancer. 2018;119(12):1456–63.

Publisher's Note