A Predictive Paradigm for COVID-19 Prognosis Based on the Longitudinal Measure of Biomarkers

Supplementary Information

Contents

Supplemental Methods

Historical regression trees

The historical regression trees (HTREE) method is an extension of the standard tree method fitting a random forest model to longitudinal data and producing a non-parametric estimate of how the response depends on all of its prior realizations as well as that of any time-varying predictor variables [1]. Data is assumed to be in the longitudinal data form:

$z_{ij} = (y_{ij}, t_{ij}, x_{ij})$ for $i = 1, 2, ..., n$ and $j = 1, 2, ..., n_i$, with y_{ij} being the response for the *i-th* subject at the *j-th* observation time t_{ij}. The tree node split of a historic regression tree is based on the concurrent and historical predictors for the response y_{ij}. The concurrent predictor (t_{ij}, x_{ij}) is a predictor value observed at the same time as the response y_{ij}. An historic predictor is one of all predictor values observed prior to the time t_{ij} of a given time point element (y_{ij}, t_{ij}, x_{ij}) for subject *i* at time t_{ij}. The node split of HTREE on a concurrent predictor follows the approach of standard classification trees. For historical predictors, the splitting is modified since, associated with each observed response y_{ij}. For these, the splitting is done by first transforming the preceding values of a predictor using a summary function, the *hrf* function in the R package "htree" [2]. The importance of a variable can be measured by comparing model prediction errors with and without the input variable under investigation in the HTREE model. The variable importance summary statistics of predictors is based on increase in the mean squared error (MSE) when the predictor *i* is replaced with permuted values in the algorithm. Specially, consider the out of bag sample corresponding to the b^{th} bootstrap sample, recalling that these $l = 1, \dots, n_b$ out of bag subjects were those not used to build the b^{th} decision tree. For out of bag subject l, data from measurement time s_{lj} is $\{S_l^{\tau}(s_{lj}), Z_l, Y_l(s_{lj})\}, l = 1, ..., n_b, j = 1, .., m_l$, and the estimated τ – year survival probability for individual l at time s_{lj} based upon traversing the b^{th} decision tree and landing in partition R_{bkl} is $\overline{S^{\tau}(s_{lj})R_{bkl}}$. For the b^{th} decision tree, model fit in the out of bag sample is characterized by

$$MSE_b(Z) = \sum_{l=1}^{n_b} 1/ml \sum_{j=1}^{ml} (S_l^{\tau}(s_{lj}) - \overline{S_l^{\tau}(s_{lj})_{R_{bkl}}})^2 ,$$

where the summand is the average mean squared error for individual l across follow-up windows, $j = 1, \ldots, m_l$. Denote $MSE_b(Z)$ the value of MSE_b when the input variable of interest has been randomly permuted as described above, altering the estimated τ-year survival probabilities for individual l at each time s_{lj} in the formula. The importance of the input variable under consideration is calculated as $\frac{\sum_{b=1}^{B}[MSE_b(Z) - MSE_b(Z)]}{\sum_{b=1}^{B} MSE_b(Z)}$, measuring the relative increase in $MSE_b(Z)$ due to permuting the input variable under consideration. The “htree” package reports the marginalized error based on this calculation, with larger values indicating greater predictor importance for assessing the impact of an input variable.

Joint model

The joint model is a dynamic prediction model often used to typify relationships between the longitudinal process and time-to-event outcome. The joint model consists of two linked sub-models: a survival sub-model and a longitudinal (mixed effect) model [3].

We let T_{i} denote the observed failure time for the $\mathrm{i}^{th}(i = 1, 2, ..., n)$ subject, which is taken as the minimum of the true event time T_i^* and the censoring time C_{i}, that is, $T_i = min\ (T_i^*,\ C_i)$. Further, we define the event indicator as $\delta_i = I(T_i^* \leq C_i)$, where I(·) is the indicator function that takes the value 1 if condition $T_i^* \leq C_i$ is satisfied, and 0 otherwise. For longitudinal responses, let $y_i(t)$ denote the value of the longitudinal biomarkers at time point t for the i^{th} subject. The actual observed longitudinal biomarkers for subject i consist of the measurements $y_{ij} = \left\{ y_i\left(t_{ij}\right), j = 1, \dots, n_i \right\}$ taken at time points t_{ij}. We will denote the true and unobserved value of the longitudinal outcome at time t as $m_i(t)$. Here, a linear mixed effects model was used to describe the subject-specific longitudinal evolutions:

$$y_i(t) = m_i(t) + \varepsilon_i(t) = x_i^T(t)\beta + z_i^T(t)b_i + \varepsilon_i(t) \quad , \quad \varepsilon_i(t) \sim N(0, \sigma^2)$$

where β denotes the vector of the unknown fixed-effects parameters, b_i the vector of random effects, $x_i(t)$ and $z_i(t)$ the row vectors of the design matrices for the fixed and random effects, respectively, and $\varepsilon_i(t)$ is the measurement error term with variance σ^2. Finally, the random effects b_i are assumed normally distributed with mean zero and covariance matrix D and independent of $\varepsilon_i(t)$. To quantify the effect of $m_i(t)$ on the risk for an event, a standard option is to use a relative risk model of the form:

$$h_i(t \mid M_i(t), \omega_i) = \lim_{dt \to 0} \Pr(t \le T_i^* t + dt \mid T_i^*, M_i(t), \omega_i) / dt = h_0(t) \exp\{\gamma^T \omega_i + \partial m_i(t)\}$$

where $M_i(t) = \{m_i(u); 0 \le u < t\}$ denotes the history of the true unobserved longitudinal process up to time point t, $h_0(\cdot)$ denotes the baseline risk function, and ω_i is a vector of baseline covariates with a corresponding vector of regression coefficients γ.

Bayesian inference was applied for parameter estimation using Markov chain Monte Carlo (MCMC) algorithms, and this can be applied to a limited class of models with the R package JMbayes [4]. Expression for the posterior distribution of model parameters is derived under the assumptions that, given the random effects, the longitudinal and event time processes are assumed independent, and the longitudinal responses of each subject are assumed independent.

Dynamic prediction

Under the Bayesian specification of the joint model, we can derive subject-specific predictions for either the survival outcome. Based on a joint model fitted to the sample $D_n = \{T_i, \delta_i, y_i; i = 1, \ldots, n\}$ from the target population, we are interested in dynamic predictions for a new subject j from the same population given the longitudinal biomarkers history: $Y_i(t) = \{y_{il}(t_{il}); 0 \le t_{il} \le t, l = 1, \ldots, n_j\}$, and has a vector of baseline covariates w_i. We supposed that the biomarker measurements have been recorded up to t, implying that subject i was event-free up to this time point. Therefore, it is more relevant to focus on conditional subject-specific predictions, given survival up to t. In particular, for any time u > t we are interested in the probability that subject j will survive at least up to time u,

$$\pi_i(\mathrm{u|t}) = \mathrm{P}(\mathrm{T}_i^* \ge u \mid \mathrm{T}_i^* > t, y_i(t), \omega_i, D_n) = \int P(\mathrm{T}_i^* \ge u \mid \mathrm{T}_i^* > t, y_i(\theta)) p(\theta \mid D_n) d\theta$$

$$P(\mathrm{T}_i^* \ge u \mid \mathrm{T}_i^* > t, y_i(t), \theta)) = \int P(\mathrm{T}_i^* \ge u \mid \mathrm{T}_i^* > t, b_i, \theta) p(b_i \mid T_i^* > t, y_i(t), \theta) db_i$$

$$= \int \frac{S_i\{u \mid H_i(u, b_i), \theta\}}{S_i\{t \mid H_i(u, b_i), \theta\}} p(b_i \mid T_i^* > t, y_i(t), \theta) db_i$$

The posterior distribution of the parameters for the original data D_n was used to obtain $\pi_i(u \mid t)$ within Monte Carlo samples by a Monte Carlo algorithm. This dynamic prediction usually is applied to predict the dynamic survival probability of subject j when new biomarker information is recorded at time t > u.

Discrimination

To measure the discriminative capability of longitudinal markers, we focused on how well the model discriminates between patients with or without the event. We used the time-dependent area under the receiver operating characteristic curve (AUC) for the occurrence of the event in a time interval. We assumed that there were longitudinal measurements: $Y_j(t) = \{y_j(t_{jl}); 0 \le t_{jl} \le t, l = 1, \ldots, n_j\}$

up to the time point t for subject j. This subject j may either experience the event, that is $\pi_j(t + \Delta t \mid t) \le c$ within a clinically relevant time interval Δt or not $\pi_j(t + \Delta t \mid t) > c$, where $0 \le c \le 1$. Thus, in this context, we define sensitivity and specificity as $P\{\pi_j(t+\triangle t) \le c \mid T_j^* \in (\mathrm{t,t+\triangle t})\}$ and $P\{\pi_j(t+\triangle t) > c \mid T_j^* > \mathrm{t+\triangle t})\}$, respectively. For a randomly chosen pair of subjects $\{i, j\}$, in which both subjects have provided measurements up to time t. Then, the discriminative capability of the assumed model can be assessed by the time-dependent AUC, which is calculated for varying values of and is given by:

$$AUC(t,\triangle t) = P[\pi_{\mathrm{i}}(\mathrm{t+\triangle t|t}) < \pi_j(\mathrm{t+\triangle t|t}) | \{\mathrm{T_i^*} \in (\mathrm{t,t+\triangle t|})\} \cap \{\mathrm{T_j^*} > \mathrm{t+\triangle t}\}$$

That is, if subject i experiences the event within the relevant time frame but subject j does not, we would expect the assumed model to assign higher probability of surviving longer than t+Δt to the subject who did not experience the event.

Time-varying effects

A time-varying joint model has been postulated to measure the relationship between survival and longitudinal biomarkers over time [5]. Specially, we have $\mathrm{h_i}(t \mid M_i(t), \omega_i) = h_0(t)\exp[\gamma^T \omega_i + \mathrm{f}\{\lambda(\mathrm{t}), \mathrm{M_i(t)}\}]$ where the function $f\{\lambda(t), M_i(t)\}$ postulates that the hazard of the event associates with the value and the slope of the longitudinal biomarkers at t or the accumulated longitudinal process up to time t. A p-splines approach based on using a high or relatively high number of equally spaced knots was adopted for λ(t). In particular, we take $\lambda(t) = \sum_{l=1}^{L} \alpha_l B_l(t)$, where *α* is a set of parameters that captures strength of the association between longitudinal biomarkers and survival outcome, and $B_l(t)$ denotes the $1-th$ basis function of a B-spline [6]. The smoothness of functions λ(t) is controlled by the following priors for the coefficient that links longitudinal and survival outcomes α : $\alpha | \tau_\alpha \sim N_L(0, \tau_\alpha M_\alpha)$ and $\tau_\alpha \sim \mathrm{Gamma}(c_1, c_2)$, where M_α are the penalty matrices. In particular, $M_\alpha = \Delta_r{}^T \Delta_r + 10^{-6} I$ and Δ_r is a r[th] order difference matrix. The scaled identity matrix I ensures a positive defined variance–covariance matrix.

Supplemental Table 1. Demographics of patients with missing laboratory test data more than 80% in the discovery dataset.

ID	Age	Gender	Outcome
187	44	Male	Survival
189	61	Male	Survival
192	34	Male	Survival
197	67	Male	Survival
200	25	Male	Survival
201	39	Male	Survival
253	51	Male	Dead
265	81	Female	Dead
268	69	Male	Dead
271	54	Male	Dead
275	65	Male	Dead
285	63	Male	Dead
289	63	Male	Dead
311	77	Female	Dead
347	80	Female	Dead
354	57	Male	Dead
359	65	Male	Dead

Supplemental Table 2. Demographics, clinical laboratory tests, and mortality outcome in the first validation dataset from Huangshi City.

	Total (n = 112)	Survived (n = 81)	Dead (n = 31)	*P-value*
Age	60.99±14.87	57.14±13.77	71.03±12.96	<0.0001[a]
Gender, n (%)				<0.0001[c]
Male	73(65.18)	54(73.97)	19(26.03)	
Female	39(34.82)	27(69.23)	12(30.77)	
Median follow-up (days)	11	15	7	<0.0001[b]
Laboratory tests (baseline)				
LDH, U/L	385.02 ± 189.09	349.12 ± 167.76	496.94 ± 212.46	0.0005[b]
WBC, 10^9/L	7.26 ± 5.18	6.18 ± 3.55	10.08 ± 7.4	0.0050[b]
NEU, 10^9/L	5.7 ± 4.13	4.94 ± 3.32	7.86 ± 5.36	0.0086[b]
Hs-CRP, mg/L	17.34 ± 43.65	19.8 ± 51.71	11.42 ± 14.36	0.6142[a]
MPV, fL	11.19 ± 1.03	11.09 ± 1.08	11.45 ± 0.83	0.1069[a]
Lymphocyte (%)	14.28 ± 8.49	15.9 ± 8.47	9.6 ± 6.74	0.0006[a]
Monocytes (%)	7.57 ± 4.6	8.2 ± 4.78	5.85 ± 3.63	0.0162[a]
Creatinine, umol/L	71.6 ± 34.79	70.48 ± 37.15	74.83 ± 27.27	0.5614[a]
PT, S	12.14 ± 1.4	12.04 ± 1.42	12.46 ± 1.31	0.2354[a]
RDW, %	13.27 ± 1.95	13.33 ± 2.22	13.10 ± 0.78	0.3969[b]
Urea, nmol/L	5.75 ± 3.26	5 ± 2.83	7.8 ± 3.54	0.0014[a]
Glucose, mmol/L	7.98 ± 5.04	7.88 ± 5.66	8.25 ± 2.97	0.1121[b]
AST, U/L	45.37 ± 27.52	39.94 ± 23.03	61.43 ± 33.54	0.0013[b]

Note: LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase;

Continuous variables were presented as mean ± standard deviation; categorical variables were presented as frequency and proportion [n (%)].

[a]*P*-value was derived from Student's *t*-test.

[b]*P*-value was derived from rank-sum test.

[c]*P*-value was derived from χ^2 test.

Supplemental Table 3. Demographics, baseline clinical laboratory tests, and mortality outcome in the second validation dataset from Wuhan Huoshenshan Hospital.

	Total (n = 1527)	**Survival (n = 1470)**	**Dead (n = 57)**	***P-Value***
Age	61.81 ± 14.13	61.44 ± 14.1	71.39 ± 11.26	< 0.0001 [a]
Gender				0.0098 [c]
Male	775 (50.75)	736(50.07)	39(68.42)	
Female	752 (49.25)	734(49.93)	18(31.58)	
Median Follow-up (days)	15	15	14	0.8084 [b]
Laboratory tests (baseline)				
LDH, U/L	213.32 ± 91.67	206.73 ± 80.13	423.08 ± 163.66	< 0.0001 [b]
WBC, 10^9/L	10.15 ± 37.47	10.14 ± 38.13	10.22 ± 8.22	< 0.0001 [b]
NEU, 10^9/L	4.4 ± 2.83	4.22 ± 2.23	9.36 ± 8.03	< 0.0001 [b]
Hs-CRP, mg/L	2.61 ± 3.07	2.58 ± 3.05	6.2 ± 3.79	0.0038 [b]
MPV, fL	10.11 ± 1.2	10.09 ± 1.19	10.78 ± 1.31	0.0001 [b]
Lymphocyte, %	24.2 ± 10.47	24.72 ± 10.13	10.27 ± 9.79	< 0.0001 [b]
Monocyte, %	7.65 ± 2.6	7.75 ± 2.52	4.98 ± 3.27	< 0.0001 [b]
Procalcitonin, ng/mL	0.11 ± 0.56	0.11 ± 0.57	0.44 ± 0.42	< 0.0001 [b]
Creatinine, umol/L	73.33 ± 64.15	72.15 ± 62.73	105.46 ± 90.31	< 0.0001 [b]
PT, S	13.13 ± 2.11	13.07 ± 2.04	15.35 ± 3.42	< 0.0001 [b]
RDW, %	13.2 ± 1.34	13.2 ± 1.32	13.19 ± 1.75	0.3587 [b]
Urea, nmol/L	5.21 ± 2.98	5.05 ± 2.68	9.6 ± 6.04	< 0.0001 [b]
Glucose, mmol/L	5.73 ± 2.33	5.67 ± 2.29	7.42 ± 2.81	< 0.0001 [b]
AST, U/L	26.07 ± 22.44	25.57 ± 21.68	41.01 ± 36.33	0.0001 [b]

Note: LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase;

Continuous variables were presented as mean ± standard deviations; categorical variables were presented as frequency and proportion [n (%)].

[a] *P*-value was derived from Student's *t*-test.

[b] *P*-value was derived from rank-sum test.

[c] *P*-value was derived from χ^2 test.

Supplemental Table 4. Comparisons between raw data from medical records and the imputed data of laboratory tests in the discovery dataset.

Laboratory test	Mean (SD)[a]	Mean (SD)[b]	*P-value*
Procalcitonin*, ng/ml	1.13 (4.7)	1.59 (5.63)	0.1034
Basophil, %	0.21 (0.22)	0.23 (0.23)	0.1651
Prothrombin time*, S	16.51 (8.76)	16.04 (7.24)	0.2547
Platelet large cell ratio	31.72 (8.57)	32.14 (8.58)	0.2826
Basophil, 10^9/L	0.02 (0.02)	0.02 (0.02)	0.2846
D-D dimer, mg/L	7.83 (9.18)	7.33 (8.97)	0.2915
Thrombocytocrit	0.21 (0.09)	0.21 (0.09)	0.3143
Mean platelet volume*, fL	10.9 (1.09)	10.95 (1.09)	0.3403
Monocytes*, %	6.16 (3.84)	6.32 (4.25)	0.3818
Glucose*, mmol/L	8.84 (5.17)	9.03 (5.42)	0.4524
Red blood cell distribution width*, %	13.06 (1.72)	13.11 (1.76)	0.5364
PLT distribution width	12.99 (2.79)	13.07 (2.83)	0.5824
White blood cell count*, 10^9/L	10.95 (20.06)	11.44 (20.56)	0.5997
Uric acid, umol/L	275.93 (151.35)	272.5 (147.51)	0.6151
Red blood cell count	6.28 (14.88)	6.6 (15.67)	0.6454
Urea*, mmol/L	9.43 (9.3)	9.24 (9)	0.6482
EGFR	81.91 (32.02)	82.53 (31.82)	0.6711
Eosinophils, %	0.64 (1.07)	0.66 (1.09)	0.6779
Eosinophil, 10^9/L	0.04 (0.06)	0.04 (0.06)	0.715
Total protein, g/L	65.38 (7.57)	65.26 (7.5)	0.7155
Hemoglobin, g/L	124.35 (19.38)	124.6 (19.44)	0.7822
RBC distribution width SD	42.35 (6.26)	42.43 (6.33)	0.7912
$HCO3^-$, mmol/L	23.2 (4.32)	23.25 (4.39)	0.7946
Creatinine*, umol/L	108.83 (133)	107.44 (128.46)	0.8162
Serum sodium	141.45 (7.09)	141.37 (6.95)	0.8187
International standard ratio	1.31 (0.82)	1.3 (0.85)	0.8235
Neutrophils, %	77.53 (16.46)	77.37 (16.18)	0.8285
Corrected calcium	2.35 (0.13)	2.35 (0.13)	0.8307
Hypersensitive C reactive protein*, mg/L	76.2 (81.16)	77.05 (82.75)	0.8329
Hematocrit	36.6 (5.26)	36.65 (5.31)	0.8353
Albumin, g/L	32.1 (6.2)	32.04 (6.15)	0.8453
Monocytes	0.49 (1.22)	0.5 (1.18)	0.8500
Indirect bilirubin, umol/L	6.89 (7.05)	6.83 (6.86)	0.8503
Lactate dehydrogenase*, U/L	470.17 (365.02)	473.11 (365.64)	0.8599
Prothrombin activity	78.9 (22.19)	78.68 (28.05)	0.8609
Platelet count	185.76 (103.87)	184.99 (103.8)	0.8694

Laboratory test	Mean (SD)[a]	Mean (SD)[b]	*P-value*
Serum potassium	4.5 (0.81)	4.49 (0.8)	0.8887
Glutamic pyruvic transaminase	37.2 (67.71)	36.8 (64.86)	0.8938
Mean corpuscular hemoglobin concentration	342.74 (17.34)	342.64 (17.2)	0.8992
Aspartate aminotransferase*, U/L	45.69 (100.36)	45.17 (96.4)	0.9081
Calcium, mmol/L	2.08 (0.16)	2.08 (0.16)	0.91
Total cholesterol, mmol/L	3.69 (0.98)	3.7 (0.99)	0.9332
Mean corpuscular hemoglobin	30.97 (2.85)	30.96 (2.85)	0.9341
γ glutamyl transpeptidase, U/L	54.93 (69.32)	54.68 (69.08)	0.9367
Alkaline phosphatase, U/L	82.2 (46.55)	82.05 (45.87)	0.941
Neutrophils*, 10^9/L	7.77 (6.06)	7.79 (6.11)	0.9442
Direct bilirubin, umol/L	9.77 (21.42)	9.71 (20.84)	0.9473
Lymphocyte*, %	15.46 (12.89)	15.43 (12.83)	0.9547
Globulin, g/L	33.25 (5.51)	33.24 (5.52)	0.9596
Serum chloride	103.03 (7.46)	103.02 (7.33)	0.9618
Total bilirubin, g/L	16.58 (26.63)	16.55 (25.82)	0.9779
Lymphocyte, 10^9/L	0.98 (1.49)	0.98 (1.46)	0.9923
Mean corpuscular volume	90.32 (6.47)	90.32 (6.56)	0.9979

Note: * prognosis biomarkers in the selected set using historical regression trees;
[a] mean (standard deviation) of raw data; [b] mean (standard deviation) of imputed data;
SD: standard deviations; *P*-value was derived from Student's *t*-test.

Supplemental Table 5. Comparisons between raw data from medical records and the imputed data of laboratory tests in the first validation dataset from Huangshi City.

Laboratory test	Mean (SD)[a]	Mean (SD)[b]	*P-value*
Prothrombin time, S	13.85 (5.35)	13.43 (4.95)	0.1535
White blood cell count, 10^9/L	8.49 (4.96)	8.12 (4.81)	0.1595
Lactate dehydrogenase, U/L	402.61 (297.8)	382.46 (274.71)	0.2446
Glucose, mmol/L	8.28 (4.22)	8.07 (4.12)	0.3577
Red blood cell distribution width, %	13.84 (1.86)	13.77 (1.74)	0.4811
Urea, nmol/L	8.92 (12.57)	8.57 (11.2)	0.5879
Hypersensitive C reactive protein, mg/L	12.57 (28.24)	11.86 (28.85)	0.7571
Creatinine, umol/L	69.78 (67.31)	70.8 (61.34)	0.7723
Monocytes, %	7.26 (4.59)	7.32 (4.58)	0.7978
Lymphocyte, %	13.28 (10.04)	13.39 (9.98)	0.8327
Aspartate aminotransferase, U/L	50.36 (99)	49.57 (89.68)	0.8781
Mean platelet volume, fL	11.1 (3.74)	11.12 (3.14)	0.9177
Neutrophils, 10^9/L	6.95 (4.8)	6.97 (4.71)	0.9270

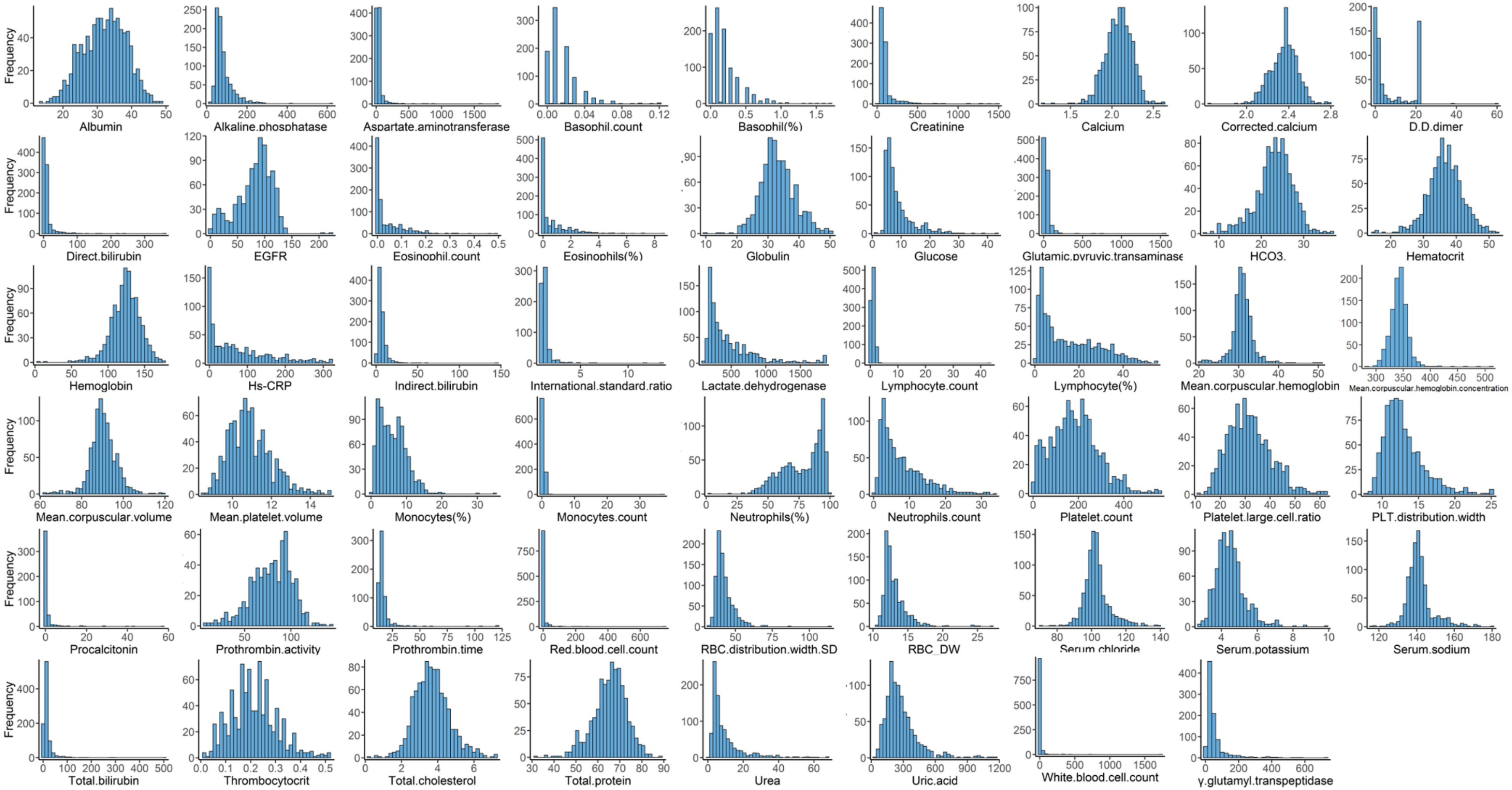
Note: [a] mean (standard deviation) of raw data; [b] mean (standard deviation) of imputed data; SD: standard deviations; *P*-value was derived from Student's *t*-test.

Supplemental Table 6. Comparisons between raw data from medical records and the imputed data of laboratory tests in the second validation dataset from Wuhan Huoshenshan Hospital.

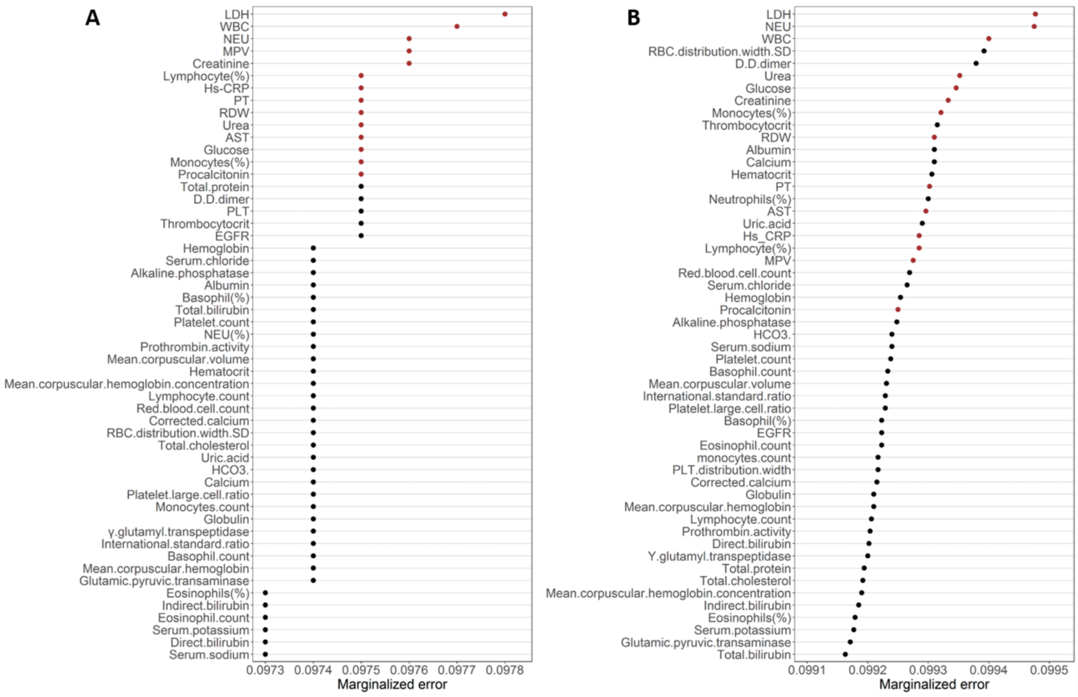
Laboratory test	**Mean (SD)[a]**	**Mean (SD)[b]**	***P-value***
Lactate dehydrogenase, U/L	218.54 (104.78)	219.25 (107.7)	0.6982
White blood cell count, 10^9/L	10.41 (58.09)	10.67 (62.28)	0.7839
Neutrophils, 10^9/L	4.8 (4.24)	4.8 (4.17)	0.9698
Hypersensitive C reactive protein, mg/L	2.61 (6.35)	2.82 (5.39)	0.0509
Mean platelet volume, fL	10.14 (1.21)	10.13 (1.2)	0.9211
Lymphocyte, %	24.16 (11.22)	24.16 (11.2)	0.9925
Monocyte, %	7.58 (2.76)	7.59 (2.78)	0.7775
Procalcitonin, ng/mL	0.44 (2.92)	0.45 (2.67)	0.8323
Creatinine, umol/L	74.74 (60.08)	74.06 (58.29)	0.4986
Prothrombin time, S	13.4 (2.59)	13.41 (2.87)	0.8118
Red blood cell distribution width, %	13.38 (1.39)	13.37 (1.37)	0.7622
Urea, nmol/L	5.8 (4.19)	5.73 (4.03)	0.3262
Glucose, mmol/L	6 (2.57)	5.96 (2.51)	0.3895
Aspartate aminotransferase, U/L	27.74 (31.1)	27.98 (31.67)	0.6418

Note: [a] mean (standard deviation) of raw data; [b] mean (standard deviation) of imputed data; SD: standard deviations; *P*-value was derived from Student's *t*-test.

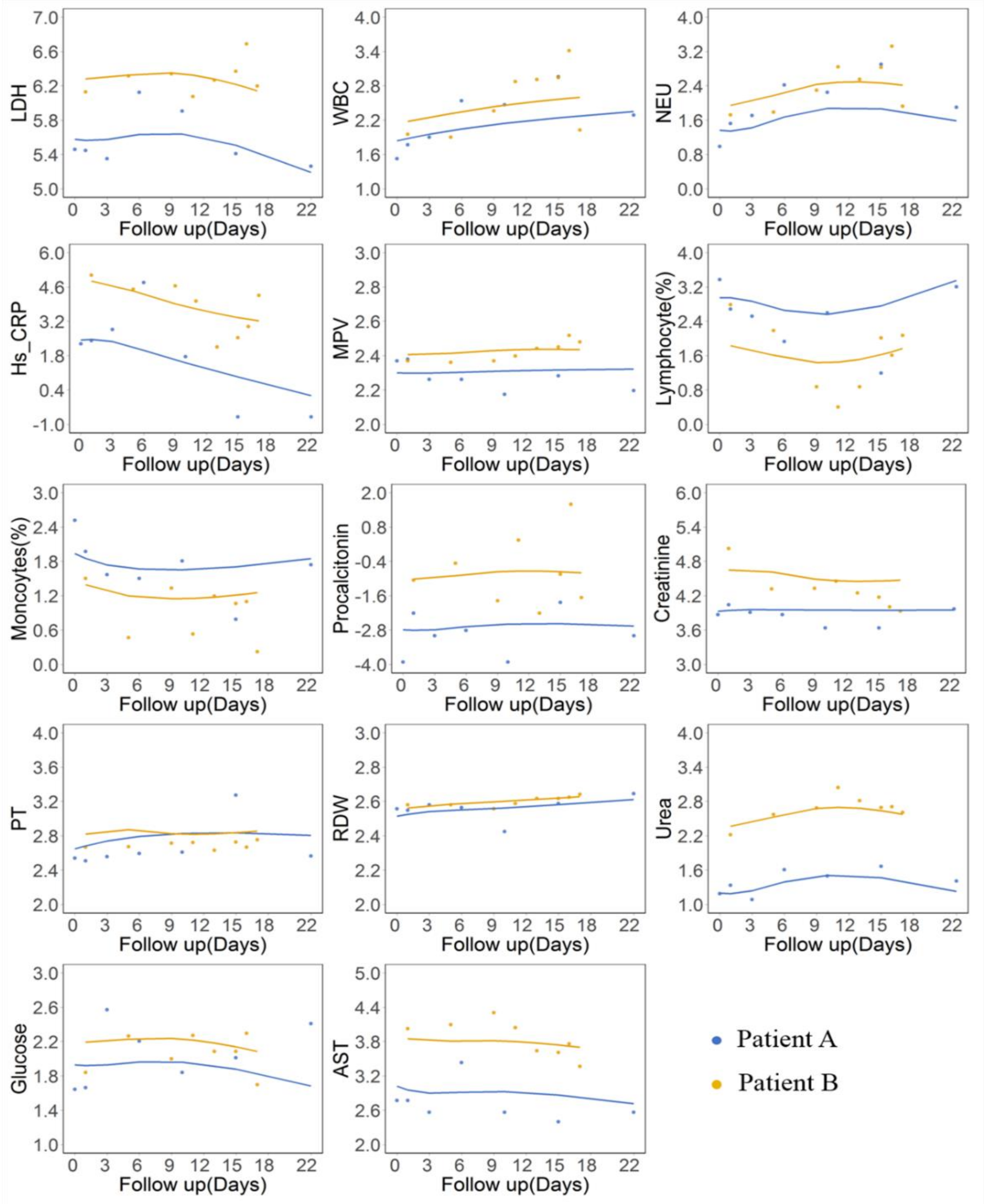
Supplemental Figure 1. Distributions of all laboratory biomarker values in the discovery dataset.



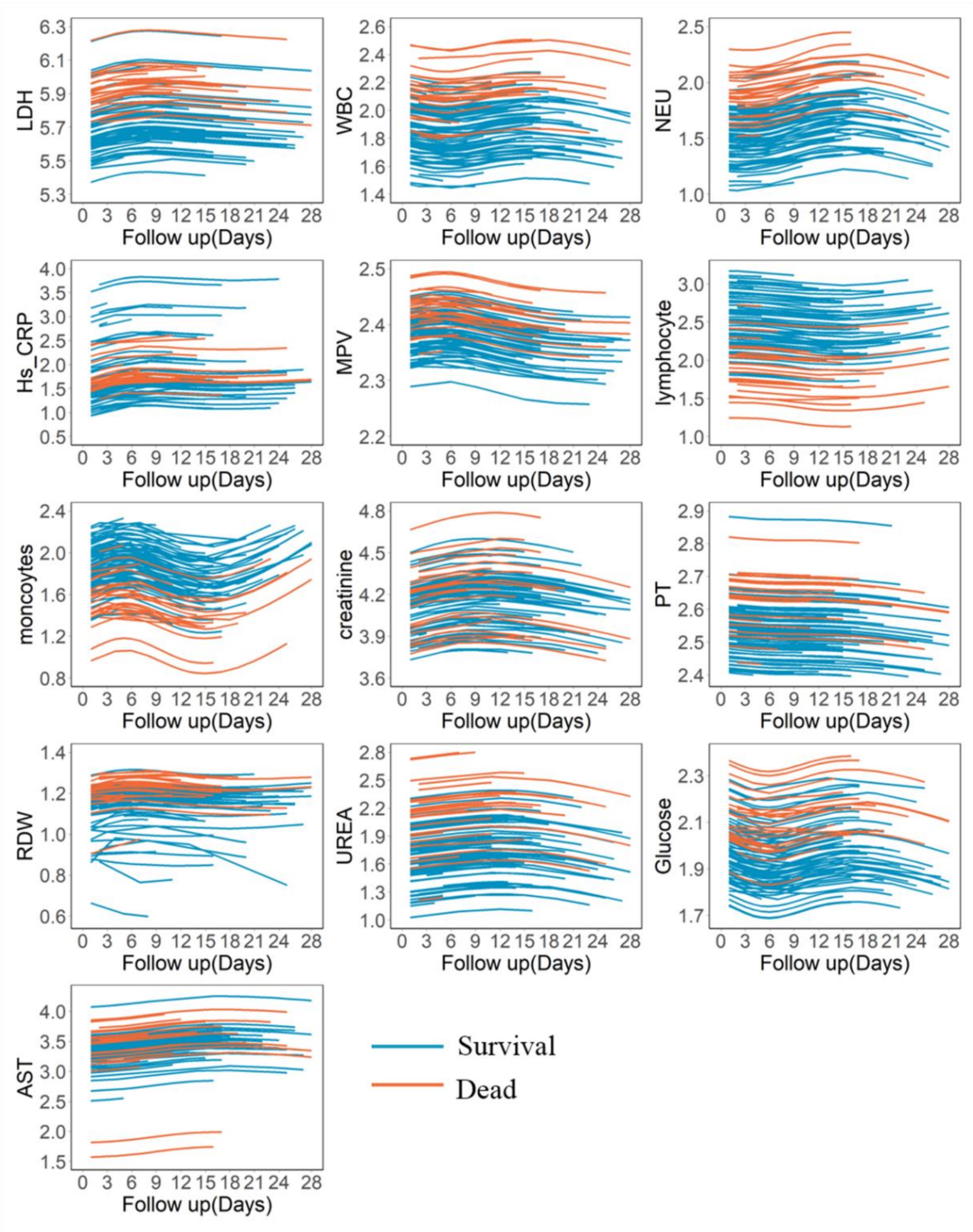
Supplemental Figure 2. Clinical biomarkers that are ranked according to the importance in HTREE model including all laboratory biomarkers in the discovery phase. A, Importance order of laboratory biomarkers in the model based on the discovery dataset. The top 14 biomarkers (red) selected using SWSFS were used for further analysis. **B**, Mean importance order of all biomarkers using three-fold cross validation. LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase.



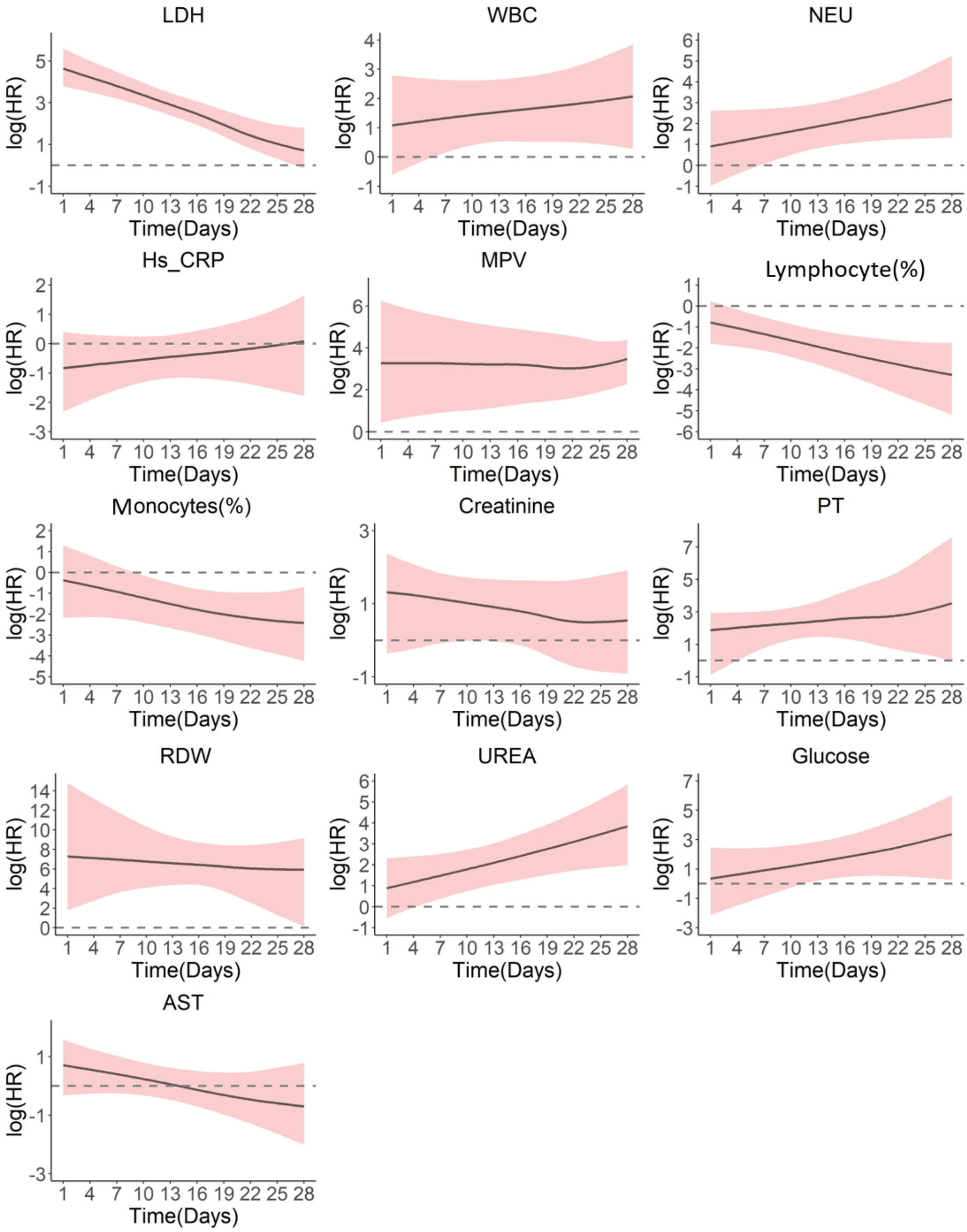
Supplemental Figure 3. The predicted longitudinal measurements of Patient A (survived) and patient B (deceased) in the discovery dataset. The predicted longitudinal measurements of joint model for patient A (blue) and patient B (yellow) based on the discovery dataset. Panels show the observed longitudinal biomarkers (points) and model-based predictions (lines) using natural cubic splines with two degrees of freedom. LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase.



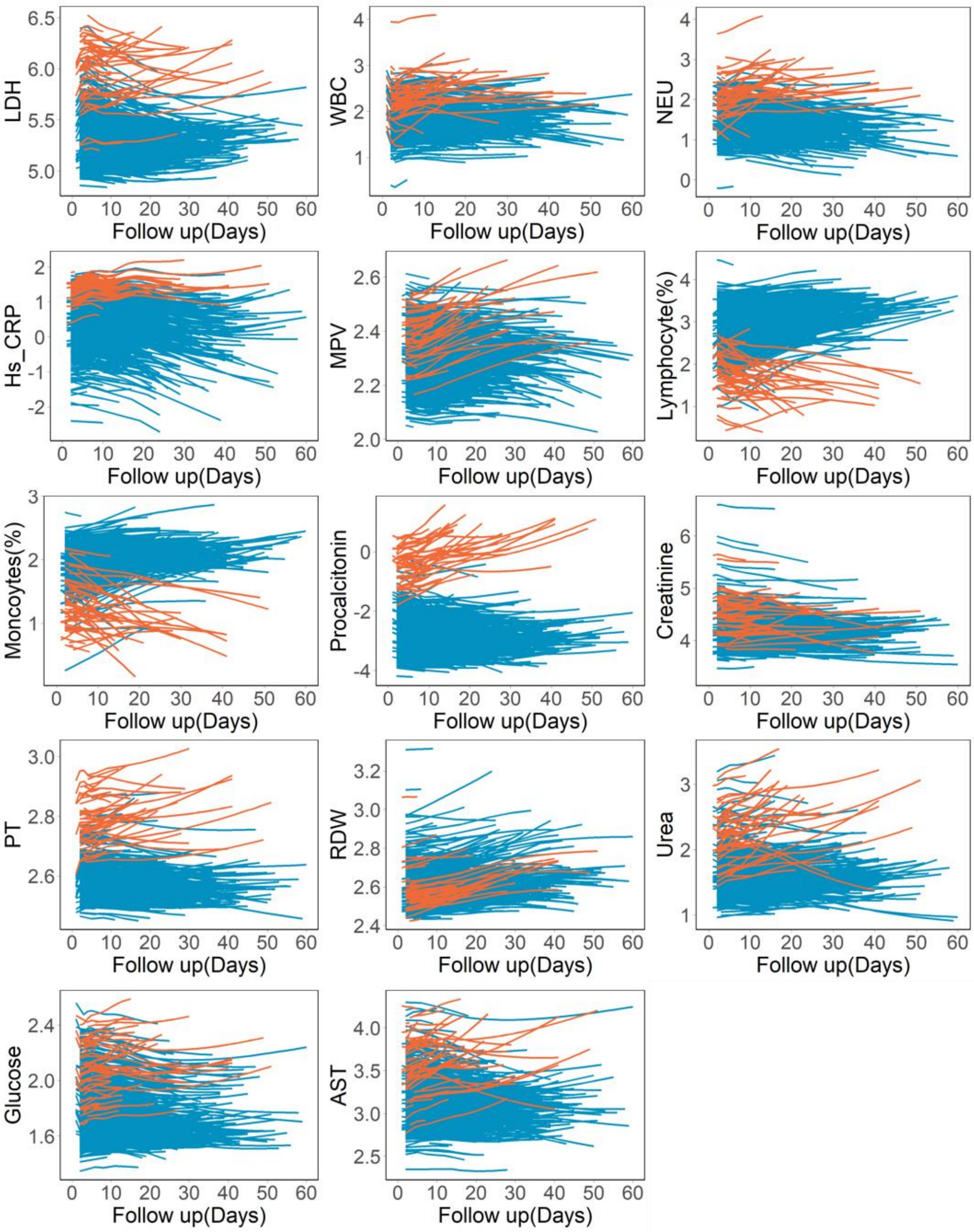
Supplemental Figure 4. Fitting trajectory patterns of biomarkers for patients who survived or deceased in the first validation dataset from Huangshi City. Lines represent averaged trajectories of patients who survived (blue) or deceased (red) during hospitalization using natural cubic splines with two degrees of freedom. LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase.



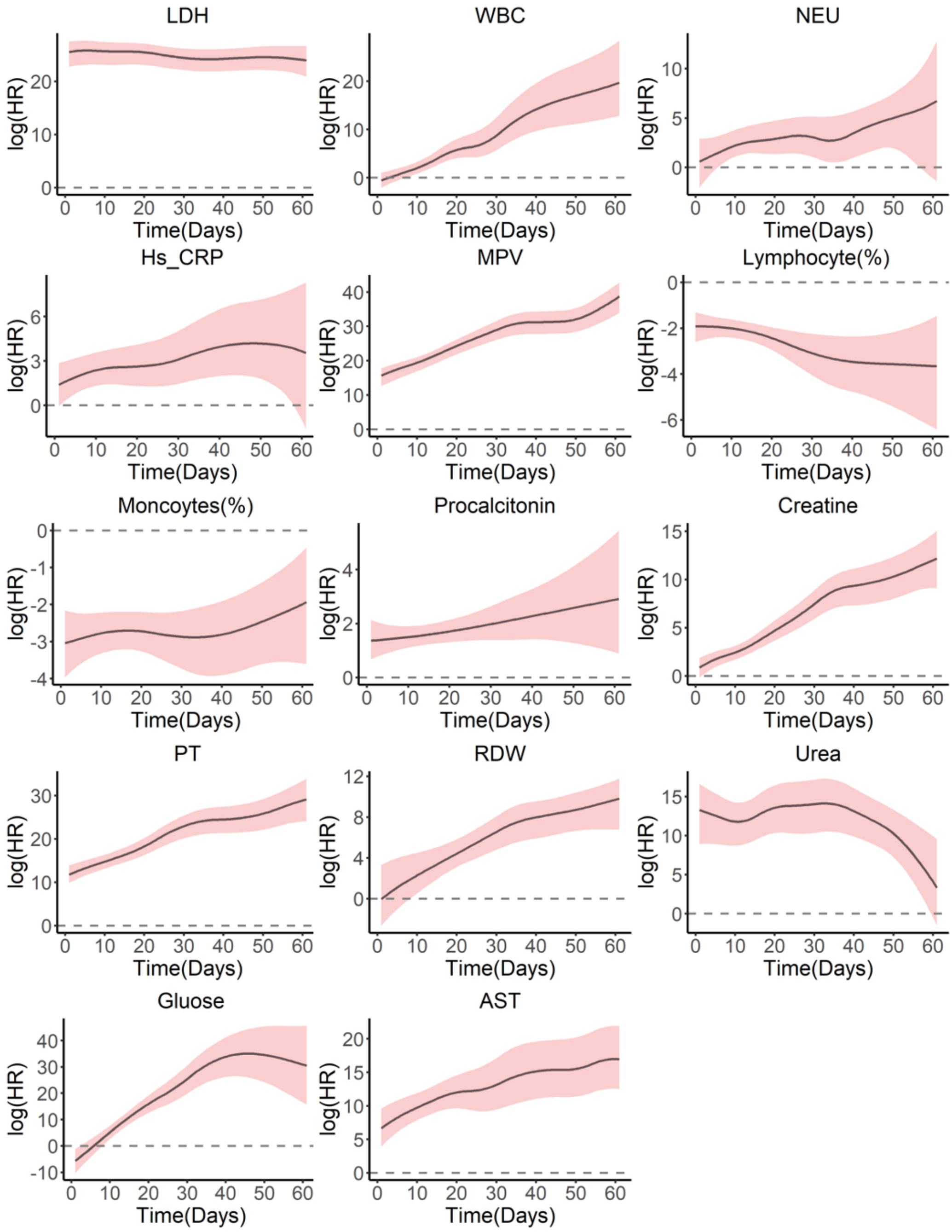
Supplemental Figure 5. Time-varying effects of biomarkers in the first validation dataset from Huangshi City. LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase.



Supplemental Figure 6. Fitting trajectory patterns of biomarkers for patients who survived or deceased in the second validation dataset from Wuhan Huoshenshan Hospital. Lines represent averaged trajectories of patients who survived (blue) or deceased (red) during hospitalization using natural cubic splines with two degrees of freedom. LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase.



Supplemental Figure 7. Time-varying effects of biomarkers in the second validation dataset from Wuhan Huoshenshan Hospital. LDH: lactate dehydrogenase; WBC: white blood cell counts; NEU: neutrophil; Hs-CRP: hypersensitive c-reactive protein; MPV: mean platelet volume; PT: prothrombin time; RDW: red blood cell distribution width; AST: aspartate aminotransferase.



References

1. Lili Zhao SM, Laura H. Mariani, Wenjun Ju. Dynamic Risk Prediction Using Longitudinal Biomakers of High Dimensions, Statistics in Medicine 2020.
2. Sexton J. Historical Tree Ensembles for Longitudinal Data, R package 2018.
3. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data, Biometrics 2011;67:819-829.
4. Rizopoulos D. TheRPackageJMbayesfor Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC, Journal of Statistical Software 2016;72.
5. Andrinopoulou ER, Eilers PHC, Takkenberg JJM et al. Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines, Biometrics 2018;74:685-693.
6. Eilers P, Marx BJSS. Flexible smoothing with B-splines and penalties 1996;11:89-121.