

SUPPLEMENTARY MATERIAL FOR "DYNAMIC PREDICTION OF RESIDUAL LIFE WITH LONGITUDINAL COVARIATES USING LONG SHORT-TERM MEMORY NETWORKS"

BY GRACE RHODES^{1,*}, MARIE DAVIDIAN^{1,†} AND WENBIN LU^{1,‡}

¹Department of Statistics, North Carolina State University, *gmrhodes@ncsu.edu; †davidian@ncsu.edu; ‡wlu4@ncsu.edu

1. LSTM-GLM Automated Hyperparameter Selection. We introduce an automated selection process for two important LSTM-GLM hyperparameters: the dimension of the context vectors, s , and the number of LSTM autoencoder training epochs, ep . The process is outlined below at a given prediction time τ .

1. Specify a set of D candidate hyperparameter settings $\mathcal{D} = \{(s_1, ep_1), \dots, (s_D, ep_D)\}$.
 2. For each $d = 1, \dots, D$:
 - a) For each of the $k = 1, \dots, p$ biomarkers:
 - i. Train an LSTM autoencoder for ep_d epochs to construct s_d -dimensional, window-specific context vectors $\psi_{ik}^d(\tau)$ for all i such that $Y_i > \tau$.
 - b) Define $\psi_i^d(\tau) = \{\psi_{i1}^d(\tau), \dots, \psi_{ip}^d(\tau)\}$.
 3. Using R unique divisions of the data:
 - a) Divide the data into a training data set containing 100ω percent of at-risk patients and a testing data set containing the other $100(1 - \omega)$ percent of at-risk patients, where $\omega \in (0, 1)$. Stratify on the censoring indicator Δ_i .
 - b) For each $d = 1, \dots, D$:
 - i. Fit the LSTM-GLM on the baseline covariates \mathbf{X}_i and the window-specific context vectors $\psi_i^d(\tau)$ using patients i in the training data set.
 - ii. Compute the loss defined in Equation (1) on patients i in the testing data set.
- $$(1) \quad \frac{1}{\sum_i \frac{\Delta_i I(Y_i > \tau)}{\widehat{G}(Y_i)}} \sum_{i=1}^m \frac{\Delta_i I(Y_i > \tau)}{\widehat{G}(Y_i)} \left[(Y_i - \tau) - g\{\eta(\tau) + \gamma(\tau)^T \mathbf{X}_i + \alpha(\tau)^T \psi_i^d(\tau)\} \right]^2$$
4. For each $d = 1, \dots, D$:
 - a) Calculate the median of the R testing losses computed in Step 3b. Denote the median testing loss for hyperparameter setting d as $m(d)$.
 5. Define $d^{opt} = \operatorname{argmin}_d \{m(d)\}$. Select $s = s_{d^{opt}}$ and $ep = ep_{d^{opt}}$.

The selection of R and ω should be guided by the size of the data set under study and the computational resources available. Note, the automated hyperparameter selection process must be repeated at each prediction time $\tau \in \mathcal{T}$.

2. Comparative Methods for Performance Evaluation. We dynamically predict MRL using the LSTM-GLM, the LSTM-NN, and six variations of the dynamic transformed MRL model, and we compare the prediction performance of the competing methods. For each of the six dynamic transformed MRL models, we define a distinct function of the history of longitudinal biomarker measurements, $\zeta_i(\tau) = f\{\mathbf{Z}_i(t_{i1}), \dots, \mathbf{Z}_i(t_{i\tau_i})\}$.

First, we specify $\zeta_i^{(B)}(\tau) = \mathbf{Z}_i(t_{i1})$ to be the p -dimensional vector of baseline biomarker measurements collected on patient i . Second, we specify $\zeta_i^{(L)}(\tau) = \mathbf{Z}_i(t_{i\tau_i})$ to be the p -dimensional vector of biomarker measurements collected most recently before prediction

time τ on patient i . We refer to $\zeta_i^{(L)}(\tau)$ as the “last-value carried forward” vector. Third, let

$$Z_{ik}^{avg}(\tau) = \left\{ \sum_{j=1}^{n_i} I(t_{ij} < \tau) \right\}^{-1} \sum_{j=1}^{n_i} I(t_{ij} < \tau) Z_{ik}(t_{ij})$$

be the average value of biomarker k prior to time τ for patient i . We define the p -dimensional vector $\zeta_i^{(A)}(\tau) = \{Z_{i1}^{avg}(\tau), \dots, Z_{ip}^{avg}(\tau)\}$.

Next, we construct two formulations of $\zeta_i(\tau)$ that contain the intercept and slope of each biomarker regressed against time. To maintain the dynamic nature of prediction, at each prediction time τ , we conduct regression using only biomarker measurements taken at times $t_{ij} < \tau$ on patients with $Y_i > \tau$. Specifically, for each biomarker $k = 1, \dots, p$, at each prediction time τ , we fit the linear model

$$(2) \quad Z_{ik}(t_{ij}) = \beta_{0ik}^{(\tau)} + \beta_{1ik}^{(\tau)} t_{ij} + \epsilon_{ijk}^{(\tau)},$$

where $t_{ij} \in [0, \tau)$, $\beta_{0ik}^{(\tau)}$ and $\beta_{1ik}^{(\tau)}$ are scalar parameters, and $\epsilon_{ijk}^{(\tau)}$ is a scalar error term.

First, we fit Equation (2) independently for each patient i and each biomarker k at each prediction time τ via linear regression. We assume $\epsilon_{ijk}^{(\tau)} \stackrel{ind}{\sim} \mathcal{N}(0, (\sigma_{ik}^{(\tau)})^2)$, where $\mathcal{N}(\mu, \Sigma)$ denotes a normal distribution with mean μ and variance-covariance Σ .

Second, we frame Equation (2) as a linear mixed effects model by defining

$$\beta_{0ik}^{(\tau)} = \eta_{0k}^{(\tau)} + b_{0ik}^{(\tau)}, \quad \beta_{1ik}^{(\tau)} = \eta_{1k}^{(\tau)} + b_{1ik}^{(\tau)},$$

where $\eta_{0k}^{(\tau)}$ and $\eta_{1k}^{(\tau)}$ are scalar fixed parameters, and $b_{0ik}^{(\tau)}$ and $b_{1ik}^{(\tau)}$ are scalar random effects. Let $\mathbf{b}_{ik}^{(\tau)} = (b_{0ik}^{(\tau)}, b_{1ik}^{(\tau)})$. We assume $\mathbf{b}_{ik}^{(\tau)} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_k^{(\tau)})$ and $\epsilon_{ijk}^{(\tau)} \stackrel{ind}{\sim} \mathcal{N}(0, (\sigma_k^{(\tau)})^2)$. Additionally, we assume the random effects $\mathbf{b}_{ik}^{(\tau)}$ and the errors $\epsilon_{ijk}^{(\tau)}$ are independent. We fit the linear mixed effects model independently for each biomarker k at each prediction time τ based on the data for all patients with $Y_i > \tau$ via maximum likelihood methods.

We then specify the “linear regression” vector $\zeta_i^{(S)}(\tau)$ and the “mixed effects” vector $\zeta_i^{(M)}(\tau)$ to be the $2p$ -dimensional vector $(\hat{\beta}_{0i1}^{(\tau)}, \hat{\beta}_{1i1}^{(\tau)}, \dots, \hat{\beta}_{0ip}^{(\tau)}, \hat{\beta}_{1ip}^{(\tau)})$, where the parameter estimates are obtained via the aforementioned methods.

Lastly, motivated by the work of [Lin et al. \(2018\)](#), we construct the sixth formulation of $\zeta_i(\tau)$ using FPCA. At each prediction time τ , we approximate the measurement of biomarker k for patient i at time $t_{ij} \in [0, \tau)$ as

$$Z_{ik}(t_{ij}) \approx \mu_k^{(\tau)}(t_{ij}) + \sum_{l=1}^{L_k^{(\tau)}} \lambda_{ilk}^{(\tau)} \cdot \rho_{lk}^{(\tau)}(t_{ij}) + \epsilon_{ijk}^{(\tau)},$$

where $\mu_k^{(\tau)}(\cdot)$ is the mean function, $\lambda_{ilk}^{(\tau)}$ is the l th FPC score, $\rho_{lk}^{(\tau)}(\cdot)$ is the l th eigenfunction, and $\epsilon_{ijk}^{(\tau)} \stackrel{ind}{\sim} \mathcal{N}(0, (\sigma_k^{(\tau)})^2)$.

Define $L_k^{(\tau)}$ to be the minimum number of FPC scores required to explain 99 percent of the total variance of biomarker k with respect to prediction time τ . We then specify $\zeta_i^{(F)}(\tau) = (\lambda_{i11}^{(\tau)}, \dots, \lambda_{iL_1^{(\tau)}1}^{(\tau)}, \dots, \lambda_{i1p}^{(\tau)}, \dots, \lambda_{iL_p^{(\tau)}p}^{(\tau)})$ to be a vector of the $L_k^{(\tau)}$ FPC scores for each of the $k = 1, \dots, p$ biomarkers. Then the dimension of $\zeta_i^{(F)}(\tau)$ is $\sum_{k=1}^p L_k^{(\tau)}$.

We estimate $\mu_k^{(\tau)}(\cdot)$, $\rho_{lk}^{(\tau)}(\cdot)$, and $\lambda_{ilk}^{(\tau)}$ independently for each biomarker k at each prediction time τ using the principal components analysis through conditional estimation (PACE) algorithm ([Yao, Müller and Wang \(2005\)](#)). To uphold the dynamic nature of prediction, we conduct estimation at prediction time τ using only biomarker measurements collected at times $t_{ij} < \tau$ on patients with $Y_i > \tau$.

3. Simulated Survival Times. We conduct two simulation studies to assess the prediction performance of the LSTM-GLM and the LSTM-NN. In each study, we conduct 500 simulations. For each of the 500 simulations, we generate a new data set of survival times T_i and censoring times C_i for all patients $i = 1, 2, \dots, 5000$. In both studies, we generate the censoring times as $C_i \stackrel{iid}{\sim} \mathbb{U}(0, 100)$. Conversely, we generate the survival times T_i using a different model for each study.

In the first study, we generate T_i according to the accelerated failure time (AFT) model

$$\nu_i = \int_0^{T_i} \exp\{\beta_1 B_i(s) + \beta_2 X_i\} ds,$$

where $\beta_1 = \beta_2 = 1$. For each patient, we generate a random $\nu_i = \exp(\theta_i)$, where $\theta_i \stackrel{iid}{\sim} \mathcal{N}(3, 1)$.

In the second study, we generate T_i according to the Cox proportional hazards model

$$h_i(t | B_i(t), X_i) = \lambda \exp\{\beta_1 B_i(t) + \beta_2 X_i\},$$

where $\beta_1 = \beta_2 = 1$ and $\lambda = 0.05$. For each patient, we generate a random $\nu_i = -\log(1 - \theta_i)$, where $\theta_i \stackrel{iid}{\sim} \mathbb{U}(0, 1)$.

Define $\lambda = 1$ for the AFT model. Then for both the AFT and Cox models,

$$\nu_i = \int_0^{T_i} \lambda \exp\{\beta_1 B_i(s) + \beta_2 X_i\} ds.$$

Define $K(t) = 1 + \sum_{j=1}^8 jI\{t > j, t \leq (j+1)\} + 9I\{t > 9\}$. Then

$$\begin{aligned} \nu_i(t) &= \int_0^t \lambda \exp\{\beta_1 B_i(s) + \beta_2 X_i\} ds \\ &= \nu_i^*\{t; K(t)\} + I\{K(t) > 1\} \sum_{j=1}^{K(t)-1} \nu_i^*(j; j), \end{aligned}$$

where

$$\begin{aligned} \nu_i^*(t; K) &= \int_{K-1}^t \lambda \exp\{\beta_1 B_i(s) + \beta_2 X_i\} ds \\ &= \frac{\lambda \exp\{\beta_2 X_i + \beta_1(a + b_{i0}) - \beta_1 \sum_{j=1}^K (c_j + b_{ij})(j-1)\}}{\beta_1 \sum_{j=1}^K (c_j + b_{ij})} \times \\ &\quad \left[\exp\left\{t\beta_1 \sum_{j=1}^K (c_j + b_{ij})\right\} - \exp\left\{(K-1)\beta_1 \sum_{j=1}^K (c_j + b_{ij})\right\} \right]. \end{aligned}$$

We invert $\nu_i(t)$. Define

$$R\{\nu_i(t)\} = 1 + \sum_{j=1}^8 jI\{\nu_i(t) > \nu_i(j), \nu_i(t) \leq \nu_i(j+1)\} + 9I\{\nu_i(t) > \nu_i(9)\}.$$

To simplify notation, we suppress the dependence of R on $\nu_i(t)$. Then

$$\nu_i^{-1}(t) = \frac{\log \left[\frac{\{\nu_i(t) - I(R > 1) \sum_{j=1}^{R-1} \nu_i^*(j; j)\} \beta_1 \sum_{j=1}^R (c_j + b_{ij})}{\lambda \exp\{\beta_2 X_i + \beta_1(a + b_{i0}) - \beta_1 \sum_{j=1}^R (c_j + b_{ij})(j-1)\}} + \exp\left\{(R-1)\beta_1 \sum_{j=1}^R (c_j + b_{ij})\right\} \right]}{\beta_1 \sum_{j=1}^R (c_j + b_{ij})}.$$

For each simulation, we compute the survival time $T_i = \nu_i^{-1}(t)$ for each randomly generated ν_i , $i = 1, 2, \dots, 5000$. We administratively censor patients whose randomly generated ν_i is not in the range of $\int_0^{T_i} \lambda \exp\{\beta_1 B_i(s) + \beta_2 X_i\} ds$.

4. Supplemental Simulation Studies. We repeat the simulation studies described in Section 4 of the manuscript. To simulate a longitudinal biomarker measured using a precise instrument, we reduce both the measurement error and the variation in measurement times. Specifically, we generate the longitudinal biomarker $Z_i(t_{ij}) = B_i(t_{ij}) + \epsilon_{ij}$ at the patient-specific measurement times $t_{ij} = \min(0, \tau_j + \epsilon_{ij})$, where $\epsilon_{ij} \stackrel{iid}{\sim} \mathcal{N}(0, 0.05^2)$ and $\epsilon_{ij} \stackrel{iid}{\sim} \mathcal{N}(0, 0.01^2)$. With reduced measurement error, we are less concerned with overfitting the LSTM-NN, so we train the network for 25,000 epochs. We define all other simulation settings to be identical to those described in Section 4 of the manuscript.

We plot the distributions of the 500 testing losses and 500 testing C-indexes for each of the eight studied dynamic MRL models in Figure (1). For both the AFT and Cox simulations,

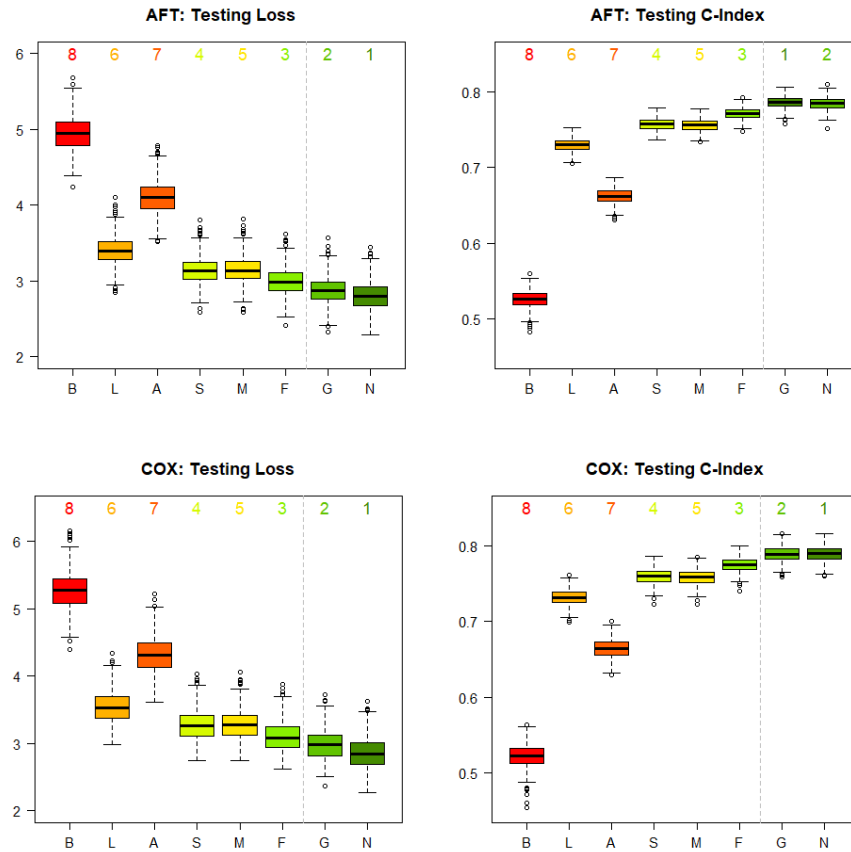


FIG 1. Distribution of the 500 testing losses and 500 testing C-indexes for each of the 8 dynamic prediction models. The models resulting in the lowest median testing loss and the highest median testing C-index are labelled 1. The models resulting in the highest median testing loss and the lowest median testing C-index are labelled 8. The six dynamic transformed MRL models are labelled according to their formulation of $\zeta_i(\tau)$. “B” represents the baseline vector. “L” represents the last-value carried forward vector. “A” represents the average vector. “S” represents the linear regression vector. “M” represents the mixed effects vector. “F” represents the FPCA vector. Furthermore, “G” represents the LSTM-GLM, and “N” represents the LSTM-NN.

the LSTM-NN results in the lowest median testing loss, followed by the LSTM-GLM. The LSTM-NN and LSTM-GLM also result in the highest median testing C-indexes. The FPCA model consistently results in the third-best median testing loss and testing C-index. Compared to the simulation studies presented in Section 4 of the manuscript, these results demonstrate a more significant improvement in calibration and discrimination for the LSTM-GLM and the LSTM-NN relative to competing methods. Thus, these results suggest that the LSTM-GLM and the LSTM-NN are especially useful for producing accurate dynamic predictions of MRL in settings where the longitudinal biomarkers are measured using precise instruments.

5. MIMIC-III Data Application Hyperparameter Selection. In the MIMIC-III data application, we must specify the hyperparameter settings of the LSTM-GLM and the LSTM-NN. The LSTM autoencoders used to construct the window-specific context vectors for the LSTM-GLM and the LSTM-NN have two important hyperparameters: the dimension of the window-specific context vectors, s , and the number of training epochs, ep_a . As s increases, the total number of parameters in the autoencoder increases, so more epochs are needed to train the autoencoder. Accordingly, at each prediction time $\tau \in \mathcal{T}$, we construct four sets of window-specific context vectors using the hyperparameter settings $(s, ep_a) \in \{(3, 150), (5, 150), (5, 300), (7, 300)\}$.

We fit four LSTM-GLMs that each regress on one of the four sets of context vectors. The distribution of 100 testing losses for each LSTM-GLM is plotted at each prediction time in Figure (2). For each $\tau \in \mathcal{T}$, we define the LSTM-GLM that results in the lowest median testing loss to be the “best” LSTM-GLM at time τ . The hyperparameter settings of the best LSTM-GLM at each $\tau \in \mathcal{T}$ can be seen in Table (1).

Additionally, we fit an “automated” LSTM-GLM. For each unique data division in the performance evaluation, we select the hyperparameter settings of the automated LSTM-GLM by conducting the automated hyperparameter selection process detailed in Section 1 on the training data. We define the set of candidate hyperparameter settings to be $\mathcal{D} = \{(3, 150), (5, 150), (5, 300), (7, 300)\}$. We set the number of iterations to $R = 50$, and we set the proportion of patients in the sub-training data set to $\omega = 0.5$.

The LSTM-NN feed-forward neural network has three additional hyperparameters that influence prediction performance: the L_2 -penalty tuning parameter, λ , the dimension of the parameter matrices, u , and the number of training epochs, ep_n . We train eight LSTM-NNs using all eight possible combinations of $\lambda \in \{0.005, 0.01\}$, $u \in \{1, 2\}$, and $ep_n \in \{2000, 3000\}$.

It would be computationally expensive to tune λ , u , and ep_n with respect to all four sets of window-specific context vectors. Consequently, we train the LSTM-NNs using only the set of context vectors constructed with hyperparameter settings $(7, 300)$. Because these context vectors have the largest dimension, they have the potential to retain the most information from the biomarker trajectories. Moreover, feed-forward neural networks like the LSTM-NN are capable of handling a large number of covariates.

TABLE 1

The hyperparameter settings of the best LSTM-GLM and the best LSTM-NN at each prediction time $\tau \in \mathcal{T}$, where the “best” model is defined to be the one resulting in the lowest median testing loss.

Prediction Time Days	LSTM-GLM		LSTM-NN				
	s	ep_a	s	ep_a	λ	u	ep_n
1	3	150	7	300	0.01	1	2000
1.5	7	300	7	300	0.005	2	3000
2	3	150	7	300	0.005	2	2000
2.5	5	300	7	300	0.01	2	2000
3	7	300	7	300	0.01	2	3000

The distribution of 100 testing losses for each LSTM-NN is plotted at each prediction time in Figure (3). For each $\tau \in \mathcal{T}$, we define the LSTM-NN that results in the lowest median testing loss to be the “best” LSTM-NN at time τ . The hyperparameter settings of the best LSTM-NN at each $\tau \in \mathcal{T}$ can be seen in Table (1).

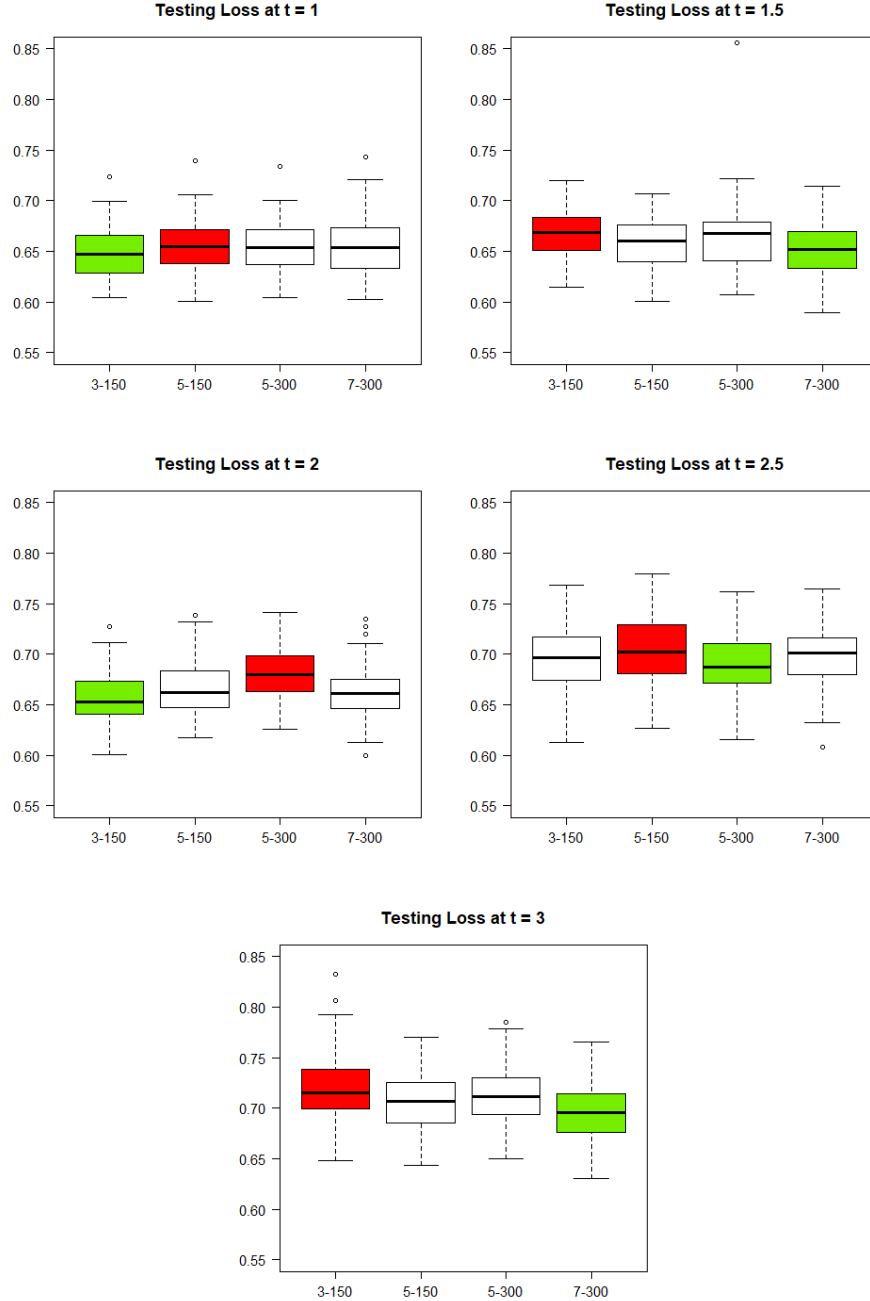


FIG 2. Distribution of the 100 testing losses for each of the four LSTM-GLMs fit using a different set of window-specific context vectors at prediction times $\tau \in \mathcal{T} = \{1, 1.5, 2, 2.5, 3\}$. Let s be the dimension of the window-specific context vectors, and let ep_a be the number of epochs used to train the LSTM autoencoders. The x-axis is labelled with the hyperparameter settings “ $s - ep_a$.”

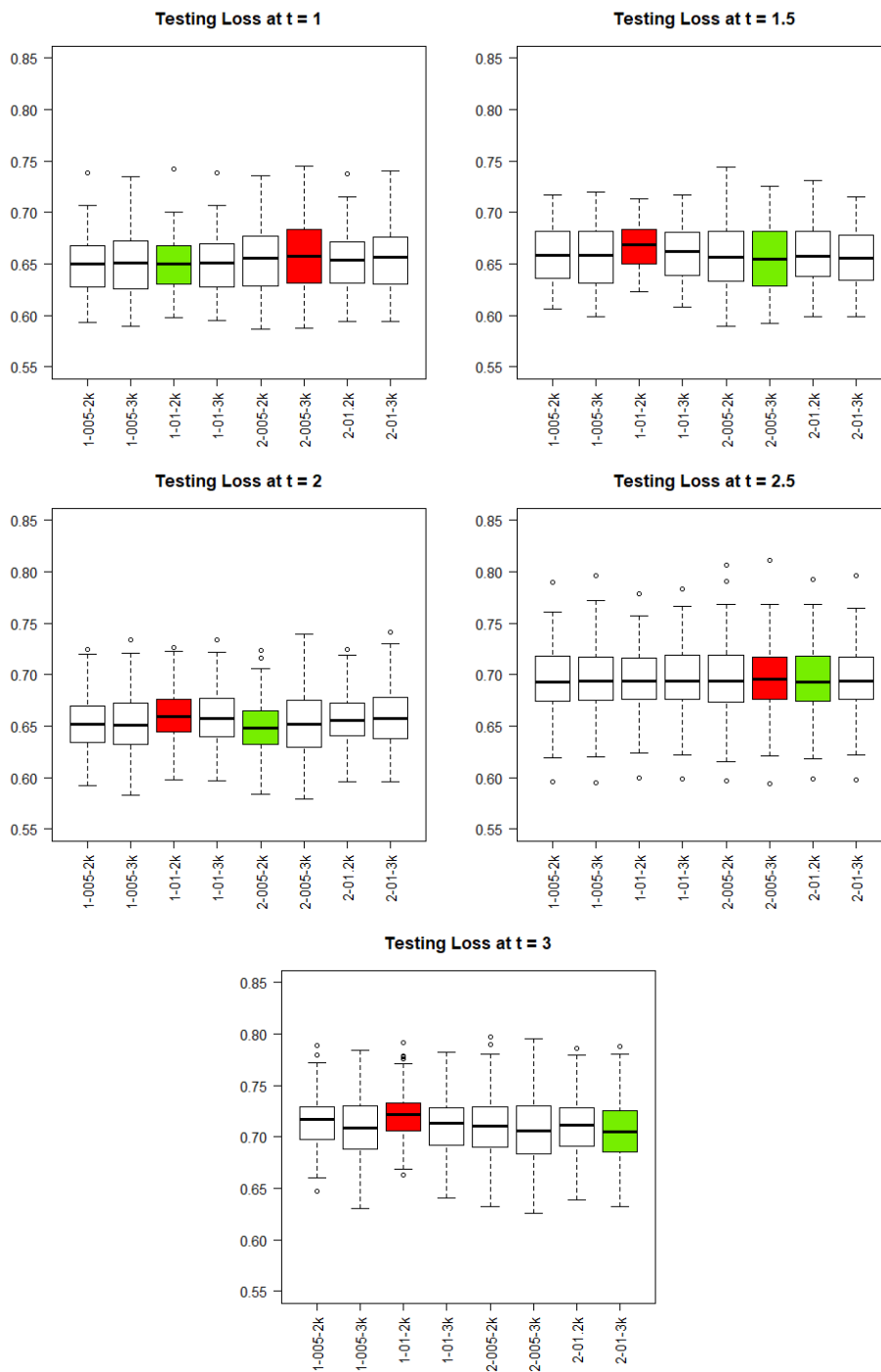


FIG 3. Distribution of the 100 testing losses for each of the eight LSTM-NNs trained using different hyperparameter settings at prediction times $\tau \in \mathcal{T} = \{1, 1.5, 2, 2.5, 3\}$. Let λ be the L_2 -penalty tuning parameter in the LSTM-NN objective function. Let u be the dimension of the LSTM-NN parameter matrices \mathbf{W}_1 and \mathbf{W}_2 . Let ep_n be the number of epochs used to train the LSTM-NN. The x-axis is labelled with the hyperparameter settings “ $u - \lambda - ep_n$.”

REFERENCES

- LIN, X., LU, T., YAN, F., LI, R. and HUANG, X. (2018). Mean residual life regression with functional principal component analysis on longitudinal data for dynamic prediction. *Biometrics* **74** 1482–1491.
- YAO, F., MÜLLER, H. G. and WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100** 577–590.