Forecasting individual survival in irregularly sampled patient trajectories

Daniel Sobotka[1,2,3], **Nino Bogveradze**[2], **Lucian Beer**[2], **Philipp Seeböck**[1,2,3], **Helmut Prosch**[2] **and Georg Langs**[1,2,3]

[1] Computational Imaging Research Lab,
Department of Biomedical Imaging and Image-guided Therapy,
Medical University of Vienna, Vienna, Austria
[2] Christian Doppler Laboratory for Machine Learning Driven Precision Imaging,
Department of Biomedical Imaging and Image-guided Therapy,
Medical University of Vienna, Vienna, Austria
[3] Comprehensive Center for Artificial Intelligence in Medicine,
Medical University of Vienna, Vienna, Austria

Abstract

Time series forecasting of patient trajectories plays a critical role in the clinical environment by enabling the prediction of possibly treatment relevant patient events . Clinical data such as imaging studies, surgical records, laboratory measurements or tumor staging provide rich longitudinal information reflecting the progression of disease or treatment response. Modeling these data involves several challenges such as integrating multi-modal data, handling irregularly sampling over time, or managing missing values. Many existing forecasting approaches rely on regularly sampled data and perform poorly when facing irregularly sampled clinical data. Here, we evaluate three different deep learning models for predicting individual six month survival from irregularly sampled lung cancer patient trajectories. Results show that state-of-the-art models can integrate sparse clinical data and benefit from multi-modality, improving forecasting of clinical outcomes despite irregular sampling patterns.

1 Introduction

The prediction of future time series from observed partial time series is a widely studied problem across many academic fields, such as climate modeling [6] or biological sciences [8]. In medical research, the prediction of future patient trajectories is critical for improving clinical decision-making and patient outcomes. Time-series models are increasingly being employed to predict key health events such as the onset of heart failure [2] or patient mortality [10]. By modeling these temporal health trajectories, clinicians can anticipate adverse events, optimize treatment plans, and provide more personalized care. Incorporating clinical relevance in time series forecasting not only enhances predictive accuracy of events but also directly impacts patient treatment planing. Time-series forecasting in the medical domain presents unique challenges that distinguish it from other fields. Clinical data are often *multi-modal*, containing a variety of information from imaging data, laboratory results to surgery information. Furthermore, clinical events are typically *irregularly sampled*, with visits occurring at non-uniform intervals based on patient conditions. Finally, prediction models need to cope with *missing data* due to clinical decisions determining the modalities acquired during an examination. Compared to conventional statistical models such as auto-regressive [11] models, machine- and deep learning models offer promising results in addressing time series forecasting challenges, such as multi-objective or multi-modality forecasting [5]. Deep learning models can

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

be grouped into 1) encoder-decoder-based, 2) transformer-based and 3) GAN-based architectures, where recent research focused on transformer based deep learning time forecasting approaches [4]. Transformer-based models leverage self-attention and multi-head attention mechanisms. They have emerged as one of the most effective architectures for capturing semantic dependencies in long input sequences [9]. Beside that classification other deep learning models are focusing on forecasting, such as *DeepSurv*, a multi-layer perceptron implementing the Cox proportional hazards model [3]. *Recurrent neural networks (RNNs)* are also considered suitable for sequence modeling and can solve time series related tasks. [4]. [1] introduced *GRU-D*, an extension of the gated recurrent unit (GRU) that incorporates decay mechanisms to model multivariate time series data with missing values.

2 Methods

We evaluate a patch-based time series transformer (PatchTST) [7], GRU-D [1] and DeepSurv [3] for using multi-modal patient trajectories as basis of forecasting for survival prediction within the 6 months (24 weeks) following the last observed time point. The networks take look-back windows $L_n \mapsto (v_1, \dots v_n)$ as input and predict survival probabilities at 24 future weekly time points $t = 1, \dots, 24$. PatchTST consists of an transformer encoder that maps input patches into a latent space using a trainable linear projection and positional encoding, which encodes the temporal position of each patch in the input sequence, followed by multi-head self-attention layers. The resulting latent representations are flattened and passed through a linear layer with ReLu activation to obtain predictions $\hat{y}_{b,t}$ $(t = 1, \dots, 24)$ for each sample b in the batch. During training we ignore empty patches and use a value mask for missing values inside each patch. GRU-D is a recurrent neural network based on gated recurrent units and decay mechanisms to model past observation fading over time and how missing values should be imputed dynamically. Similar to the transformer, missing input values are masked. To account for imbalances in the survival data, a binary cross-entropy (BCE) with logits loss weighted by a sample-wise weight based on the number of non-survival weeks per sample is proposed. The sample-wise weight for batch b is defined as

$$w_b = 1 + \sum_{t=1}^{T} (1 - y_{b,t}), \tag{1}$$

where T denotes the number of 24 forecasting weeks and $y_{i,t}$ the corresponding ground truth target. The final loss is computed as

$$\mathcal{L} = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=1}^{T} w_b \cdot BCEWithLogits(\hat{y}_{b,t}, y_{b,t}) \tag{2}$$

where B denotes the batch size. In contrast to PatchTST and GRU-D, which are designed for time-series representation learning and prediction, DeepSurv optimizes the negative log partial likelihood of the Cox proportional hazards model to learn a risk score that reflects the relative hazard of an event while explicitly handling censored observations.

3 Data

In this study, a fixed temporal window of 3072 days was defined for each of the 4642 patients, corresponding to the longest observation period recorded within the clinical cohort. Each patient's temporal sequence was aligned such that the final day of the window (day 3072) represented the patient's most recent clinical encounter, which could include one or more modalities such as computed tomography (CT) imaging extracted lung tumor volume, surgical interventions, laboratory assessments or ICD diagnosis. For patients with shorter observation periods or sparse visit histories (e.g., visits limited to a single month), the earlier portion of the window was zero-padded, while the final segment contained the available clinical data. This alignment strategy ensured temporal consistency across patients and facilitated comparability independent of follow-up duration. Across the cohort, patients underwent a total of 10747 CT scans (average 2.32 per patient) and 1981 surgical procedures (0.43 per patient). Laboratory measurements included 126572 CRP tests (27.27 per patient), 110880 albumin tests (23.89 per patient), 20359 leukocyte counts (4.39 per patient), 42019 hemoglobin measurements (9.05 per patient), 42027 hematocrit measurements (9.05 per patient), and 112892 creatinine measurements (24.32 per patient). Diagnoses included 31241 C34* ICD-10 codes (6.73 per patient), and tumor

staging was recorded 237 times (0.05 per patient). From each CT scan we extracted with a pretrained 3D U-Net lung tumor annotations and used that information as tumor load input channel. We used a train/val/test ratio of $0.7/0.15/0.15$ resulting in 3249 patients for training, 696 for validation and 697 for testing. A total of 2584 patients (55.67%) are non survival during the next 24 week prediction horizon, where 2058 patients (44.33%) survive the 24 week prediction horizon. To prevent the model from overfitting to full observation windows, we employ a random endpoint strategy, where for each patient up to five additional endpoints are randomly selected from the whole patient time series. The time series is then truncated at this randomly chosen endpoint and left-padded with zeros to obtain a fixed-length input window. Survival time is recalculated relative to the new endpoint, and a discrete survival target is constructed over the prediction horizon. This strategy reduces bias toward specific temporal positions and encourages the model to learn time-robust representations. After augmentation, the dataset distribution shifted, with the proportion of non-survivors decreasing from 55.67% to 31.03%.

4 Experiments

We evaluated if multi-modal time series models can be used for future survival forecasting with irregularly sampled and missing values input time series data. We reported standard classification metrices, precision-recall (PR) curves, as well as receiver operating characteristic (ROC) analysis for the 24 weeks forecasting. We assessed prediction of survival after 6 and 24 weeks.

5 Results

All models were trained on the same combined dataset comprising both original and randomly augmented samples, while evaluation is reported separately for full windows (FW) and random windows (RW). Quantitative results for prediction week 6 and 24 are shown in Table 1.

Table 1: Per-week performance metrics for full lookback windows (FW) and random created windows (RW) for weeks 6 and 24.

Week	Metric	PatchTST		GRU-D		DeepSurv	
		FW	RW	FW	RW	FW	RW
6	Sensitivity	0.714	0.519	0.862	0.726	0.000	0.000
	Specificity	0.789	0.842	0.765	0.833	0.998	1.000
	PPV	0.689	0.272	0.706	0.331	0.000	0.000
	NPV	0.808	0.939	0.894	0.964	0.603	0.898
	MAE	0.800	0.842	0.319	0.294	0.425	0.230
	AP (PR-AUC)	0.862	0.966	0.926	0.980	0.763	0.927
	ROC-AUC	0.810	0.778	0.888	0.861	0.692	0.593
24	Sensitivity	0.963	0.891	0.990	0.955	0.016	0.004
	Specificity	0.300	0.390	0.174	0.278	0.994	0.996
	PPV	0.622	0.328	0.589	0.307	0.750	0.214
	NPV	0.872	0.914	0.932	0.948	0.457	0.749
	MAE	1.982	1.575	0.357	0.515	0.508	0.407
	AP (PR-AUC)	0.778	0.883	0.846	0.918	0.612	0.803
	ROC-AUC	0.805	0.744	0.869	0.807	0.672	0.587

5.1 Full lookback windows

This result represents the prediction at the last time point available for each patient in the data set. PatchTST and GRU-D showed highest sensitivity and NPV across week 6 and week 24, indicating strong performance in identifying patients who experience non survival within the prediction horizon. GRU-D achieved the highest sensitivity (0.862 at week 6, 0.990 at week 24), while PatchTST had slightly higher specificity at week 6 (0.789) and week 24 (0.300). DeepSurv generally failed to correctly predicted non survival patients (sensitivity of < 0.02) and therefore achieved only high

specificity in predicting survival patients. MAE values indicated that GRU-D consistently produced more accurate survival predictions than PatchTST or DeepSurv. For the PR curves similar results are visible. At week 6, all models demonstrated strong precision–recall performance when evaluated separately on the original subsets, despite being trained on the combined dataset. GRU-D achieved the highest average precision on the original data (AP = 0.926), followed by PatchTST (AP = 0.862) and DeepSurv (AP = 0.763). ROC analysis showed that GRU-D consistently achieved the highest discriminative performance across both prediction horizons (AUC of 0.888 at week 6), outperforming PatchTST (AUC = 0.810) and DeepSurv (AUC = 0.692).

5.2 Random generated windows

This result represents the prediction at randomly chosen time points in the patient trajectory data. For the random generated windows at week 6 GRU-D achieved the highest NPV (0.964), sensitivity (0.726) and specificity slightly lower (0.833) than PatchTST. At week 24 GRU-D reached a higher sensitivity (0.990), but a lower specificity (0.174) compared to week 6. Overall, FW results showed higher sensitivity and lower specificity compared to RW results. For the PR curves with GRU-D reaching an AP of 0.980, PatchTST 0.966, and DeepSurv 0.927. A similar pattern was observed at week 24, although overall performance decreased compared to week 6, reflecting the increased difficulty of longer-term prediction. For ROC analysis, GRU-D reached an AUC of 0.861, PatchTST 0.778, and DeepSurv 0.593 at week 6. A similar trend was evident at week 24, with GRU-D achieving 0.807, PatchTST 0.744, and DeepSurv 0.587.

6 Discussion

We evaluated three different deep learning models for predicting individual six month survival from irregularly sampled lung patient trajectories. We propose a balanced loss function not only for multiple forecasting time points. Using more than one forecasting time point enhances model training, since more information can be learned from each prediction sample. We enhanced our training with random generated time windows and for training we used 10 different input channels. Results showed that the recurrent neural network yields best accuracy in predicting survival at 6 and 24 weeks, compared to the transformer model. State-of-the-art deep learning models offer a viable route towards handling highly irregularly sampled time series data, and integrate multiple modalities.

Acknowledgement

This work has been partially funded by the Vienna Science and Technology Fund (WWTF, PREDICTOME) [10.47379/LS20065], European Union's Horizon Europe research and innovation programme under grant agreement No.101100633—EUCAIM and No.101080302 AI-POD, and the Austrian Science Fund (FWF, P35189 ONSET). It has been carried out within an Inter-University Cluster Project jointly funded by the University of Vienna and the Medical University of Vienna (AICARD). The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

References

[1] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[2] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[3] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[4] Xiangjie Kong, Zhenghao Chen, Weiyao Liu, Kaili Ning, Lechao Zhang, Syauqie Muhammad Marier, Yichen Liu, Yuhao Chen, and Feng Xia. Deep learning for time series forecasting: a survey. *International Journal of Machine Learning and Cybernetics*, 16(7):5079–5112, 2025.

[5] Wenxiang Li and KL Eddie Law. Deep learning models for time series forecasting: A review. *IEEE Access*, 12:92306–92327, 2024.

[6] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

[7] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[8] David S Stoffer and Hernando Ombao. Special issue on time series analysis in the biological sciences, 2012.

[9] Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58(3):80, 2025.

[10] Ruoxi Yu, Yali Zheng, Ruikai Zhang, Yuqi Jiang, and Carmen CY Poon. Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients. *IEEE journal of biomedical and health informatics*, 24(2):486–492, 2019.

[11] George Udny Yule. Vii. on a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226(636-646):267–298, 1927.