A Deep Learning Model for Patient Subtyping and Survival Analysis in Kidney Disease Trajectory

Pumeng Shi

University of California, Berkeley, USA

Correspondence: pumengshi@berkeley.edu

Summary

Chronic kidney disease (CKD) is a global public health problem, affecting over 10% of the general population worldwide. CKD patients face an increased risk of progressing to end-stage kidney failure (ESKF), which is associated with higher hospitalization rates and cardiovascular mortality. The estimated glomerular filtration rate (eGFR) is a key indicator of kidney function and a critical predictor of CKD progression. However, existing predictive models often fail to incorporate the dynamic changes in eGFR over time, leading to less accurate forecasts of disease progression. In this study, we develop a novel deep learning model, the time-dependent LSTM (TdLSTM), to improve the prediction of eGFR and time-to-ESKF by effectively handling time-varying variables and irregular measurement time intervals in longitudinal study. Our method captures the evolving patterns of CKD progression, enabling more accurate predictions of patient outcomes. The specifically designed survival time estimate approach yields accurate predictions of the time of reaching the end stage of the disease. We trained and validated the TdLSTM model using data from two cohorts of CKD patients, demonstrating its superiority over existing models in predicting CKD progression and identifying distinct patient subtypes with varying progression rates. The proposed TdLSTM model enhances our ability to predict CKD outcomes by integrating temporal information into the predictive process, ultimately generating better prediction accuracy on the redefined time-varying evaluation metrics such as AUC and concordance index.

Introduction

Chronic kidney disease (CKD) is estimated to affect more than 10% of the global population, amounting to over 850 million individuals[1]. By 2040, CKD is projected to be the 5th highest cause of years of life lost (YLL) globally[2]. Persons suffering from CKD are at an increased risk of progression to end-stage kidney failure (ESKF), hospitalization, and cardiovascular mortality[3]. The sooner kidney disease is found, the better the chance of slowing or stopping it from worsening. Preventing kidney function decline is likely to be most effective in patients with mild to moderate CKD who are at high risk of progression to ESKF. Hence, accurately predicting CKD progression is of great clinical significance. Previous research has primarily focused on the overall disease progression rather than the individual patient's risk of transitioning to ESKF, often overlooking censored data and failing to capture personalized temporal differences. Identifying underlying patterns in the KD trajectory, especially in our sparse and irregular longitudinal data, is crucial for a deeper understanding of disease development.

Discoverying potential patterns while forecasting the occurrence of ESKF in a cohort of CKD patients is the interest of this paper. For this, we aim to predict the time of reaching ESKF, *i.e.*, time-to-ESKF, by analyzing comprehensive biomarkers. As a typical biomarker for CKD, the estimated glomerular filtration rate (eGFR) measures the level of kidney function and determines the stage of kidney disease[4]. The ESKF in our study is defined as dialysis initiation, preemptive transplantation, death due to kidney failure, or eGFR <15 ml/min/1.73m². It can be calculated from the results of the blood creatinine test, as well as age, body size, and gender. If the eGFR is low, the kidneys may not be working as well as they should. People with a lower eGFR are at increased risk of having CKD progress to kidney failure. A reduced eGFR is strongly associated with adverse cardiovascular and renal outcomes, and as eGFR declines these risks increase. For this reason, we estimate eGFR values over time up to a specified landmark. In our longitudinal study, eGFR is a time-varying variable within the 2-year prediction horizon, posing challenges to our prediction. If the time-varying eGFR is not properly accounted for, it can lead to an underestimating of the standard error and result in p-values that are too low when assessing the association between time-varying eGFR and CKD progression. There have been no existing models directly applicable to time-varying biomarkers in our study. Most of models fail to account for the evolution of eGFR.

This work aims to develop a deep learning model that effectively handles time-varying variables to estimate the impact of eGFR decline on prognosis and explore subtypes of patients with different progression rate. We utilize Long Short-Term Memory (LSTM) networks, which are well-suited for processing time-varying inputs by storing and updating context information. Unlike feed-forward neural networks, LSTMs – an advanced form of recurrent neural networks (RNNs) – use internal memory to process sequential data[5]. The feedback loop within LSTM units allows the network to retain information from past inputs indefinitely, enabling it to exhibit temporal dynamic behavior[6]. Moreover, LSTMs excel in managing long-term dependencies through a gated architecture that determines what information to retain or discard from previous time steps. This capability overcomes the limitations of traditional RNNs, making LSTMs ideal for analyzing the progression of CKD over time.

The duration between consecutive elements in a sequence is often irregular, especially in health data such as our CKD data (which will be discussed below), where patient records are unstructured, with follow-up intervals ranging from days to months or even years. These varying time gaps can signal different health conditions; for example, frequent admissions might indicate a deteriorating health status, while long intervals between records may reduce the relevance of past data for predicting current outcomes. Traditional LSTM architectures do not account for such time irregularities in longitudinal data. In healthcare, where the timing between hospital visits or clinical measurements is crucial for decision-making, an LSTM architecture that incorporates these irregular elapsed times is essential for accurately analyzing temporal data.

To address the limitations of LSTM in handling irregular time intervals in longitudinal data, we propose a novel architecture – time-dependent LSTM (TdLSTM), which incorporates the varying elapsed time between consecutive elements of a sequence as part of the patient's medical history, allowing the model to adjust its memory content accordingly. Our study design is illustrated in Figure 1. We trained the TdLSTM models using CKD data, preserving temporal information in the representation of patient variables through a backpropagation learning process. The TdLSTM encoder generates variable representations that are used to cluster patients into distinct subtypes with varying progression rates. The TdLSTM decoder then produces embeddings of patient variables for eGFR prediction and time-to-ESKF prediction. Additionally, we

developed a survival time estimation approach to accurately predict the exact time-to-ESKF based on the predicted survival probability.

We conducted experiments on the two cohorts of moderate CKD patients described by five variables, such as age, gender, eGFR, UACR, and diagnosis. The patient with eGFR >30 ml/min/1.73m² extracted from MASTERPLAN and NephroTest studies, respectively. We perform the eGFR and time-to-ESKD predictions by training the model on the MASTERPLAN dataset and externally validating its performance on the NephroTest dataset, for both baseline and follow-up studies. TdLSTM with a set of cross-validations on the combination of the two datasets is investigated in comparison with other predictive models. The unsupervised clustering analyzes the subtypes that are associated with KD progression. To comphensively evaluate our predictions, we specifically designate three evaluate metrics to measure the model performance. By investigating the error of estimate time-to-ESKF, we demonstrate the effectiveness of our proposed deep learning model. The advantage of our model is the capability to learn more representative embedding that includes evolutional clinical information to detect the temporal differences in disease progression between individuals by unsupervised clustering.

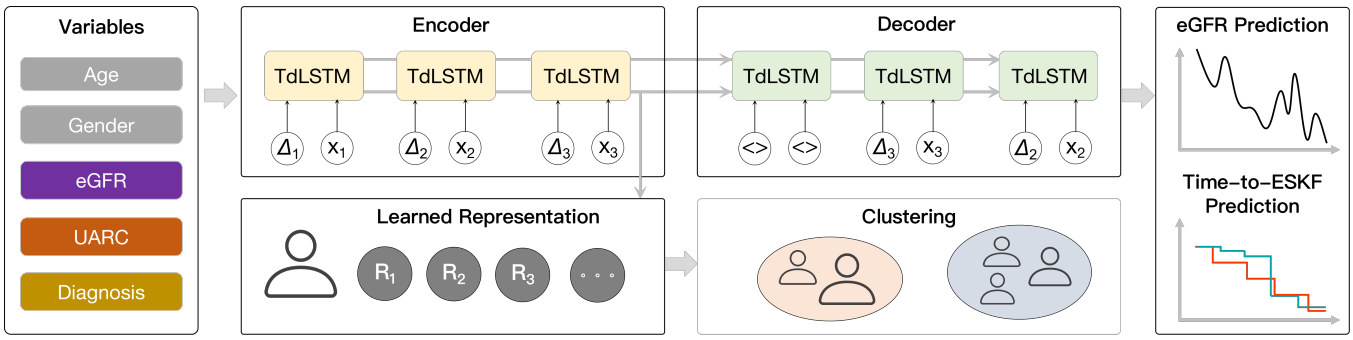

Figure 1: An overview of the study design. Patients' clinical data, including demographics (age and gender), eGFR, UARC, and diagnosis. The proposed TdLSTM model encoder generates variable representation for clustering patients and the decoder learned the embedding is used for eGFR and time-to-ESKF prediction.

Results

Statistics

We collected data from the MASTERPLAN[7] and NephroTest[8] studies, where each patient was recorded with five variables: age, gender, eGFR, urine albumin/creatinine ratio (UACR), and diagnosed nephropathy (including diabetic nephropathy, glomerulonephritis, hypertension or vascular disease, polycystic kidney disease, tubulointerstitial nephritis, and other or unknown). The MASTERPLAN study is a randomized controlled trial comparing intensified care, facilitated by a nurse practitioner, to standard care provided by a nephrologist. Serum creatinine levels were standardized to a central laboratory in the MASTERPLAN study. In the NephroTest study, plasma creatinine was initially measured using a modified kinetic Jaffe colorimetric method with a Konelab 20 analyzer (Thermo-Fisher Scientific) from the study's inception until 2008, and later with an enzymatic assay. Creatinine values obtained with the Jaffe method were standardized to an IDMS-traceable creatinine assay as previously described[9]. We calculated eGFR using the CKD-EPI

equation for serum creatinine and selected 505 patients at stage 3 with eGFR >30 ml/min/1.73m² from the MASTERPLAN study and 398 patients from the NephroTest study[10]. During the follow-up period, 11% (n = 63) of the MASTERPLAN patients reached ESKF, while in NephroTest were 4% ($n = 14$) reached ESKF. Compared to the MASTERPLAN, eGFR measurements in NephroTest were less frequent due to the annual follow-up schedule. Both studies received ethical approval from the respective committee (local IRB UMC Utrecht and CCTIRS MG/CP09.503 France), and all participants provided written informed consent. The time-to-ESKF is calculated from a 2-year landmark point up to a 4-year total follow-up period. The prediction horizon represents the maximum duration for which the models assess ESKF risk, while the landmark signifies the time since baseline when patient data is included in the model. Follow-up was censored upon reaching ESKF, death, or at the end of the 4-year period, whichever occurred first.

Table 1: Characteristics and distribution of the MASTERPLAN and NephroTest cohorts.

Variables		MASTERPLAN	NephroTest
		505 paitents	398 patients
Age, mean±SD, years	Baseline	58 ± 15	56 ± 15
Gender, Male:Female	Baseline	347:158	266:132
Diagnosed Nephropathy	Baseline		
Diabetic (%)		48 (10%)	36 (9%)
Glomerrular (%)		92 (18%)	74 (19%)
Vascular (%)		138 (27%)	100 (25%)
Congential (%)		63 (12%)	28 (7%)
Interstitial (%)		54 (11%)	47 (12%)
Other or Unknown (%)		110 (22%)	113 (28%)
eGFR, mean±SD, median (Q1-Q3)	Baseline	50 ± 18, 47 (36-63)	52 ± 19, 51 (37-66)
(ml/min/1.73m²)	Time-varying	49 ± 19, 43 (34-60)	48 ± 18, 45 (36-57)
UACR, mean, median (Q1-Q3)	Baseline	55, 40 (15-85)	52, 13 (2-62)
(mg/g)	Time-varying	284, 36 (13-284)	44, 7 (3-62)

The average age of patients in the MASTERPLAN and NephroTest cohorts was 58 and 56 years, respectively. The gender distribution was similar across both cohorts, with 69% male participants in MASTERPLAN and 67% in NephroTest. Regarding nephropathy diagnoses, diabetic nephropathy remained stable at 9-10% in both cohorts. Glomerular nephropathy was observed in 18% of MASTERPLAN participants and 19% of NephroTest participants. Vascular nephropathy was more prevalent in MASTERPLAN (27%) compared to NephroTest (25%). Congential nephropathy showed a large decrease in NephroTest (7%) compared to MASTERPLAN (12%). Interstitial nephropathy remained stable at 11% in MASTERPLAN and 12% in NephroTest. The category of other or unknown nephropathy was more common in the NephroTest cohort, while in the MASTERPLAN cohort, it remained around 22%. The estimated glomerular filtration rate (eGFR) was also comparable between the two cohorts. Baseline eGFR values were around 50 ml/min/1.73m² in MASTERPLAN and 52 ml/min in NephroTest, both declining to 48-49 ml/min/1.73m²

in longitudinal study. The urine albumin-to-creatinine ratio (UACR) showed same trends: in MASTERPLAN, UACR mostly decreased from 40mg/g (median) at baseline to 36mg/g (median) at longitudinal base, while in NephroTest, it decreased from 50 mg/g to 44mg/g. (Due to missing UACR values in longtidual MASTERPLAN study, the mean of UACR had a huge difference compared to others.) These detailed data provide insights into the stability and shifts in key health indicators and diagnoses within and between these two cohorts over time.

Subtyping

We clustered the patients into two distinct subgroups, each corresponding to a different progression pattern. In the MASTERPLAN study, 359 patients were assigned to the first subgroup (pattern 1), with 45 of these patients progressing from Stage 3 to ESKF, while 314 were censored. The second subgroup (pattern 2) included 146 patients, of whom 18 converted from Stage 3 to ESKF, and 128 were censored. In the NephroTest set, 281 patients were categorized into pattern 1, with 11 patients converting from Stage 3 to ESKF, and 270 being censored. Pattern 2 comprised 117 patients, of whom 2 progressed to ESKF, while 115 were censored. Figure 2 provides a visualization of these patterns using 2D t-SNE (t-distributed Stochastic Neighbor Embedding)[11]. Initially, we observed indistinct clusters when visualizing the baseline variables, as shown in subfigure A for MASTERPLAN and subfigure C for NephroTest. In these subfigures, patients are distributed with significant overlap in the 2-dimensional space formed by the original 5 variables. However, when we visualized the latent representation learned from the hidden layer of our model, a clear separation between the two clusters emerged. In this new 2-dimensional space, which combines baseline variables with time-to-ESKF information, the patients in both datasets formed two distinct and separable clusters, effectively distinguishing the two patterns (subtypes), as shown in subfigure B for MASTERPLAN and subfigure D for NephroTest.

Time-to-ESKF Prediction

We compared our TdLSTM model against the Kidney Failure Risk Equation (KFRE), the Cox model using baseline eGFR, and the time-dependent Cox model incorporating time-varying eGFR. Figure 3 illustrates the calibration between predicted and observed risks of progression to ESKF at a 2-year landmark with a 2-year prediction horizon, resulting in a total follow-up of 4 years in the two cohorts. The KFRE consistently overestimated the risk of ESKF in both baseline and longitudinal analyses. The Cox model[12] using only the baseline eGFR value demonstrated reasonable calibration among low-risk participants but tended to overestimate the risk in high-risk individuals. Conversely, the Cox model that included time-varying eGFR variables showed improved calibration, particularly in the longitudinal study. Our TdLSTM model achieved the best calibration overall, with predicted risks closely aligning with observed risks, especially in the longitudinal study. This demonstrates the superior ability of our model to provide accurate risk predictions over time, effectively capturing the dynamic nature of disease progression.

We further assessed the performance of time-to-ESKF prediction by incorporating the discovered patterns as an independent variable, comparing it against the five baseline variables, using 5-fold cross-validation on the combined MASTERPLAN and NephroTest datasets. Figure 4 presents the average results from 25 cross-validation experiments, highlighting the significant impact of different patterns on the C-

index, which measures the risk of progression for both baseline and follow-up studies. The time-to-ESKF prediction utilizing only the two identified patterns achieved a C-index of 0.62 (p = 0.17) for baseline data and 0.68 (p = 7.6e-16) for follow-up data. These results outperform the C-index values obtained using individual baseline variables: 0.56 for age, 0.51 for gender, 0.6 for eGFR, 0.57 for UACR, and 0.54 for diagnosis. Similarly, the pattern-based model outperformed the follow-up data results for the same vari-

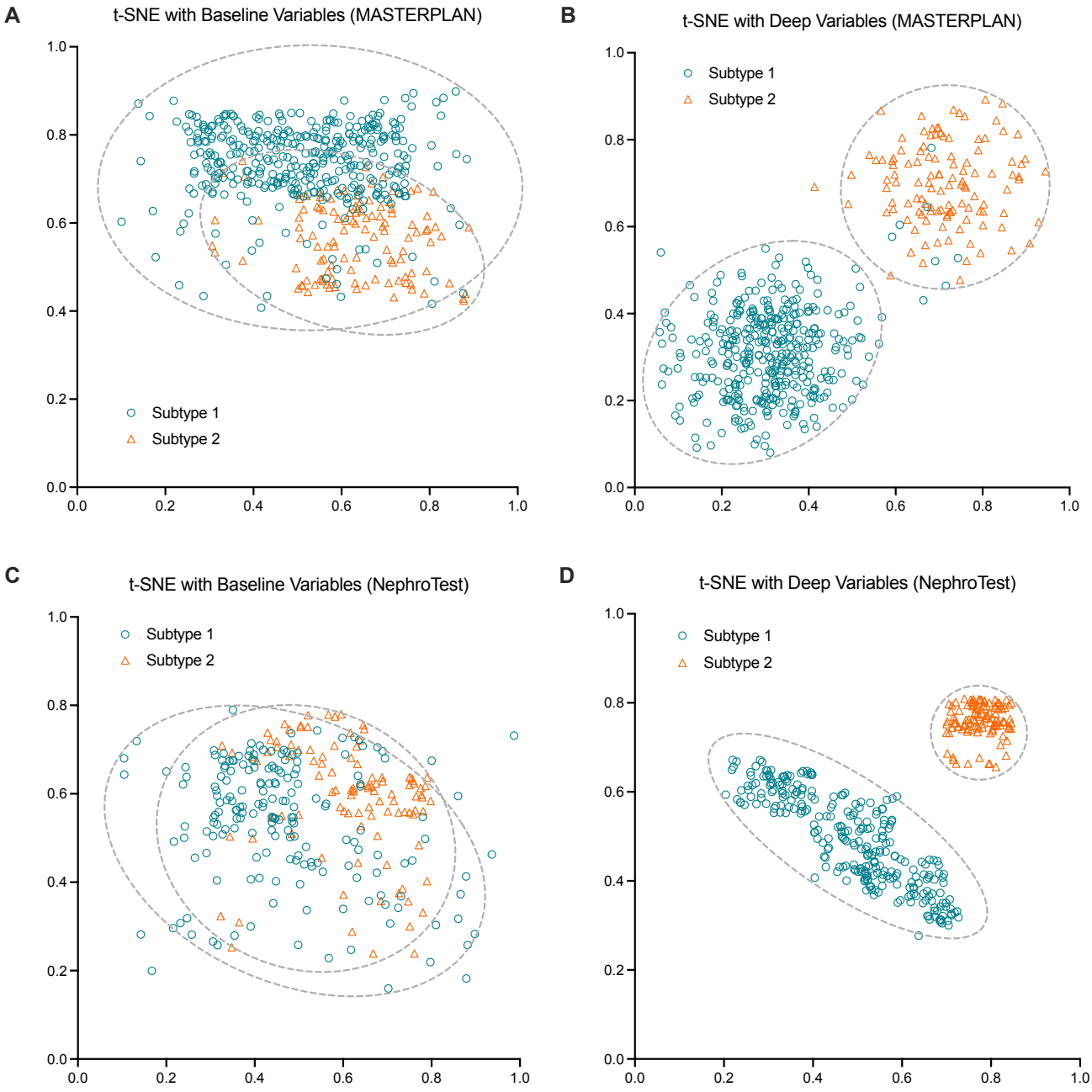

Figure 2: t-SNE for the distribution of the identified patterns on the MASTERPLAN and NephroTest datasets. Subfigures A and C are two-dimensional t-distributed stochastic neighbor embeddings (t-SNE) of the variables of the patients with two patterns on the MASTERPLAN and NephroTest follow-up datasets. Subfigures B and D are two-dimensional t-SNE by the hidden-layer activations of the model for the two patterns on the two follow-up datasets. The hidden-layer variables contain the representation of baseline variables and the temporal representation from the ESKF time prediction.

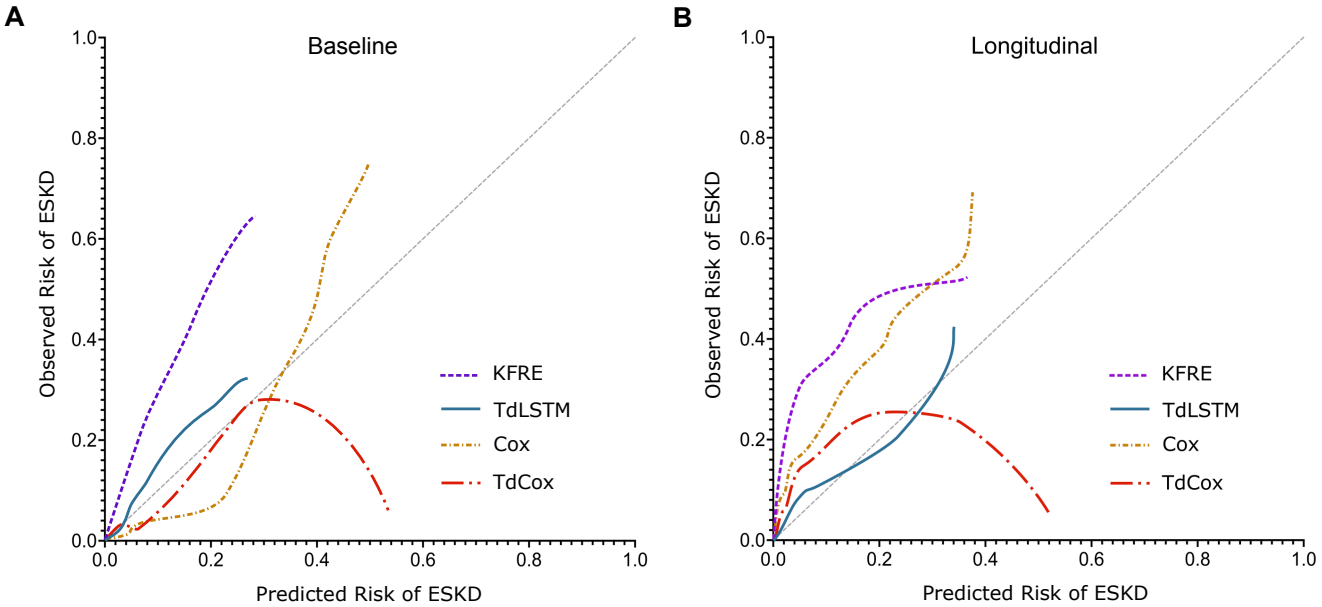

Figure 3: Calibration between predicted and observed risk of progression to ESKF at two years follow-up in the NephroTest cohort. The line indicates the average calibration line.

ables, which were 0.58, 0.52, 0.64, and 0.6, respectively. This demonstrates the robustness of the discovered patterns in enhancing prediction accuracy compared to traditional baseline variables.

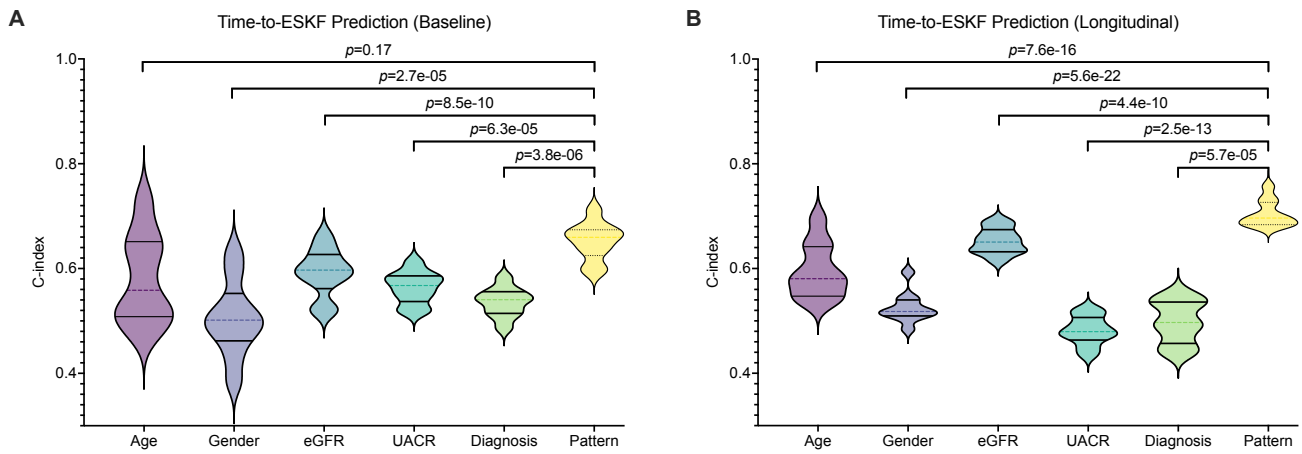

Figure 4: Predictive ability of the identified patterns. We evaluated the performance of time-to-ESKF prediction by C-index and compared the performance of the discovered patterns of the other five variables. The violin plots showed 5-fold cross-validation results on the mixed MASTERPLAN and NephroTest datasets for each variable.

To further assess the model's effectiveness in predicting the probability of being free from ESKF, we compared the area under the ROC curve (AUC) associated with predicted probabilities over the 2-year follow-up period. Figure 5 illustrates the AUC results for 1-year and 2-year progression predictions based on baseline data (left) and follow-up data (right). The findings indicate that when using baseline variables, our model achieves an AUC of 0.833 for 1-year progression prediction and 0.707 for 2-year progression. However, when using follow-up variables, the AUC improves significantly, reaching 0.955 for 1-year pro-

gression and 0.84 for 2-year progression. This demonstrates the model's enhanced predictive power when incorporating follow-up data, particularly in long-term predictions.

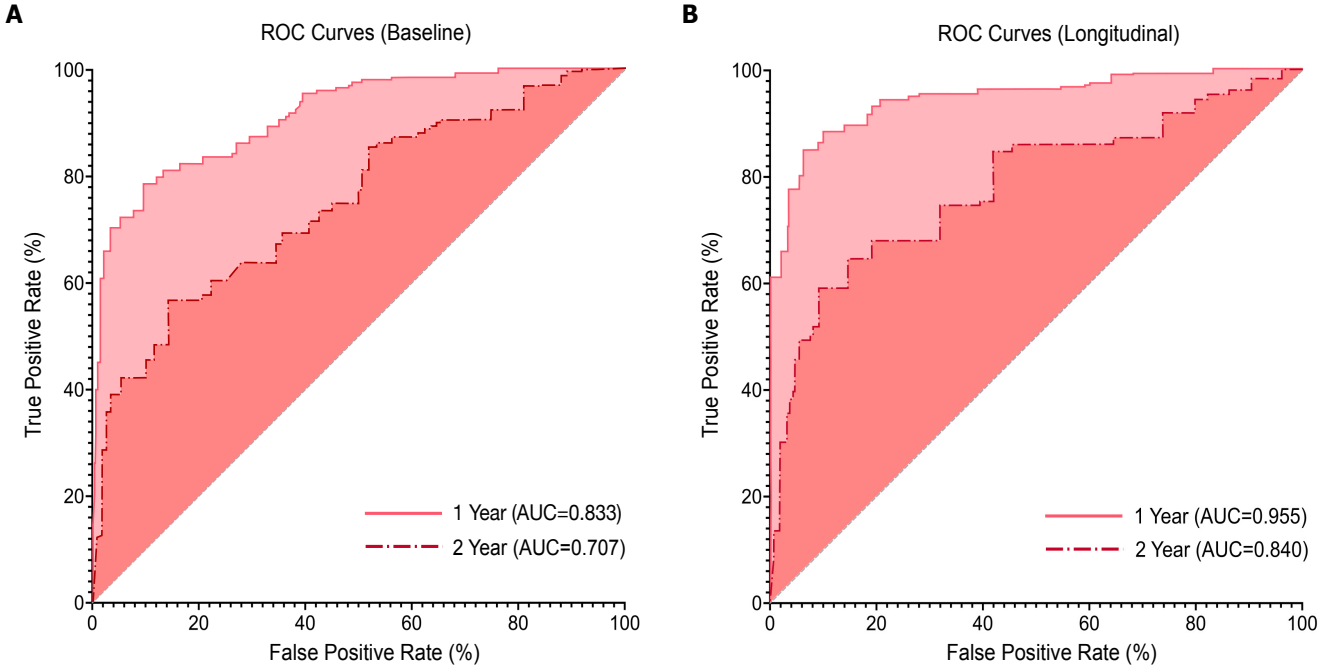

Figure 5: ROC curves for prediction of ESKF time on baseline and follow-up datasets, associated with 1 and 2 years of CKD progression.

We assessed the relative error of the estimated survival time, as summarized in Table 2. The proposed TdLSTM model significantly outperformed Cox-type models, yielding much lower relative errors[13]. Specifically, TdLSTM achieved a relative error of 0.453 on baseline data (p = 3.9e-7) and 0.377 on follow-up data (p = 0.0002). These results highlight the superior accuracy of TdLSTM in estimating survival time, particularly when follow-up data is incorporated.

Table 2: Comparison of the 5-fold cross-validation (5CV) results, and RE of the estimated survival times in the form of the mean $\pm$ standard deviation.

Dataset	Model	Relative Error	p-value
Baseline	TdLSTM	0.453 ± 0.061	3.9e-7
	Cox	0.551 ± 0.049	0.006
Time-varying	TdLSTM	0.377 ± 0.053	0.0002
	TdCox	0.483 ± 0.023	0.0044

Discussion

This work introduces a novel recurrent neural network (RNN)-based model for predicting eGFR and time-to-ESKF in patients with chronic kidney disease (CKD). The model performs nonlinear prediction and

survival analysis using longitudinal data, where each patient's trajectory is represented as a latent state derived from their baseline clinical variables and disease progression over time. Unlike traditional approaches, the model effectively handles time-varying variables and identifies distinct subtypes of CKD progression.

To address long-term dependencies, we implemented an enhanced version of Long Short-Term Memory (LSTM) networks. While standard LSTMs assume uniformly spaced time intervals, our model adjusts for irregular time intervals, a critical factor in longitudinal data. This innovation allows the model to better capture the nonlinear progression of CKD, improving predictive performance over simpler methods that rely on linear concatenation[14]. Our model maintains the temporal continuity of individual disease progression without explicitly modeling CKD stages. By aligning patients on the time axis and performing deep survival clustering, we ensure that both baseline variables and the time-to-ESKF are incorporated into the latent representations, enabling the identification of hidden patterns in patient trajectories. The result is a more nuanced understanding of disease progression, particularly in cases with heterogeneous trajectories.

In the context of CKD progression, we compared the performance of our model to a straightforward Cox proportional hazards model[12] with a linear eGFR slope. While the Cox model assumes that patients with slower progression are less likely to reach ESKF, it suffers from biases such as informative drop-out. In contrast, our model effectively handles such biases and proves to be more robust, particularly when the outcome of interest is rare (e.g., <10%) or when follow-up data is missing. Our approach was validated using data from the MASTERPLAN and NephroTest cohorts. In cases where follow-up was highly complete and the proportion of missing eGFR values was low, our model's predictive accuracy was comparable to existing benchmarks like the Kidney Failure Risk Equation (KFRE)[15]. However, when missing data or dropout occurred, our model outperformed traditional methods by offering more reliable predictions for time-to-ESKF, leveraging both recent and historical patient data. In practice, this model offers physicians a powerful tool for predicting ESKF in stage 3 CKD patients. The ability to account for time-varying factors and irregular time intervals makes it particularly useful in cases with complex or incomplete data, providing more accurate and individualized risk assessments than traditional methods.

We implemented the model using PyTorch (version 2.4) and Scikit-learn (version 1.5.1). Missing data were addressed using multivariate imputation by chained equations (MICE)[16] with the Scikit-learn "IterativeImputer" method, where a random forest algorithm was applied to perform imputation. Imputation was carried out five times, with the average of these results used as the final imputation.

Limitations of the Study

In this work, the relatively small dataset limits the potential for extensive deep model parameter tuning. The primary goal is to develop a data-driven model tailored for analyzing kidney disease data in longitudinal studies. While our model shows promise, further validation of data from multiple sources would be beneficial. A significant challenge is incorporating the uncertainty in eGFR slope estimates into the categorization of eGFR trajectories. Failing to account for this uncertainty can lead to an underestimation of standard errors and an overly optimistic assessment of the statistical significance (p-values) when analyzing the association between time-varying exposures and eGFR slope – especially when comparing periods of progression and non-progression within a single patient. Moreover, the approach by Li et al. may not fully address the issue of informative censoring or missing data that occurs when patients with more rapid

disease progression have fewer eGFR measurements available. This phenomenon, known as missingness not at random, can bias the results and will be a promising direction for future research.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Data Availability

All relevant data are public data and also have been available in the corresponding articles.

Method Details

Data Formulization

Given N patients described by V variables, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbb{R}^{N \times V}$, the eGFR prediction model outputs a series of eGFR values $Z = (z_1, z_2, \cdots)$, while the time-to-ESKF prediction model generates an outcome vector of survival probabilities $P = (p(t_1), p(t_2), \cdots)$ over the disjoint time points $t_1, t_2, \cdots$ for these patients. In our eGFR learning process, the outcomes Z can be modeled as a function of the variables $\mathbf{X}$, where the learning objective is to find the mapping f such that $\mathbf{Z} = f(\mathbf{X})$. The survival curves can be plotted by identifying a survival function S that yields the survival probability $p = S(t)$ for all $t = t_1, t_2, \cdots$. In time-to-ESKF prediction, we denote the ESKF response variables by $(\mathcal{T}, \zeta)$, where $\mathcal{T}$ is the measurement time, *i.e.*, the minimum of ESKF time T and (right-)censoring time C, *i.e.*,

$$\mathcal{T} = \min\{T, C\} = \begin{cases} T, & \text{if } \zeta = 1 \\ C, & \text{otherwise } (\zeta = 0) \end{cases}$$

The indicator $\zeta = \mathbb{1}\{T \leq C\}$ equals 1 if the patient reached ESKF and 0 otherwise. That is, $\zeta \in \{0, 1\}$ indicates either censorship or ESKF. We shall assume without loss of generality that the patients are ascendingly sorted according to observation, and T and C are conditionally independent observed.

Proposed Model

The schematic of the proposed TdLSTM includes LSTM units, each comprising an input, three gates (input, forget, and output), and an output activation function. The output of the unit is recurrently connected back to the unit input and all of the gates, *i.e.*, $\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t)$, where $\odot$ is a Hadamard (element-wise) product[14], and the output gate $\mathbf{O}_t$ modulates the amount of memory content exposure. Unlike a simple

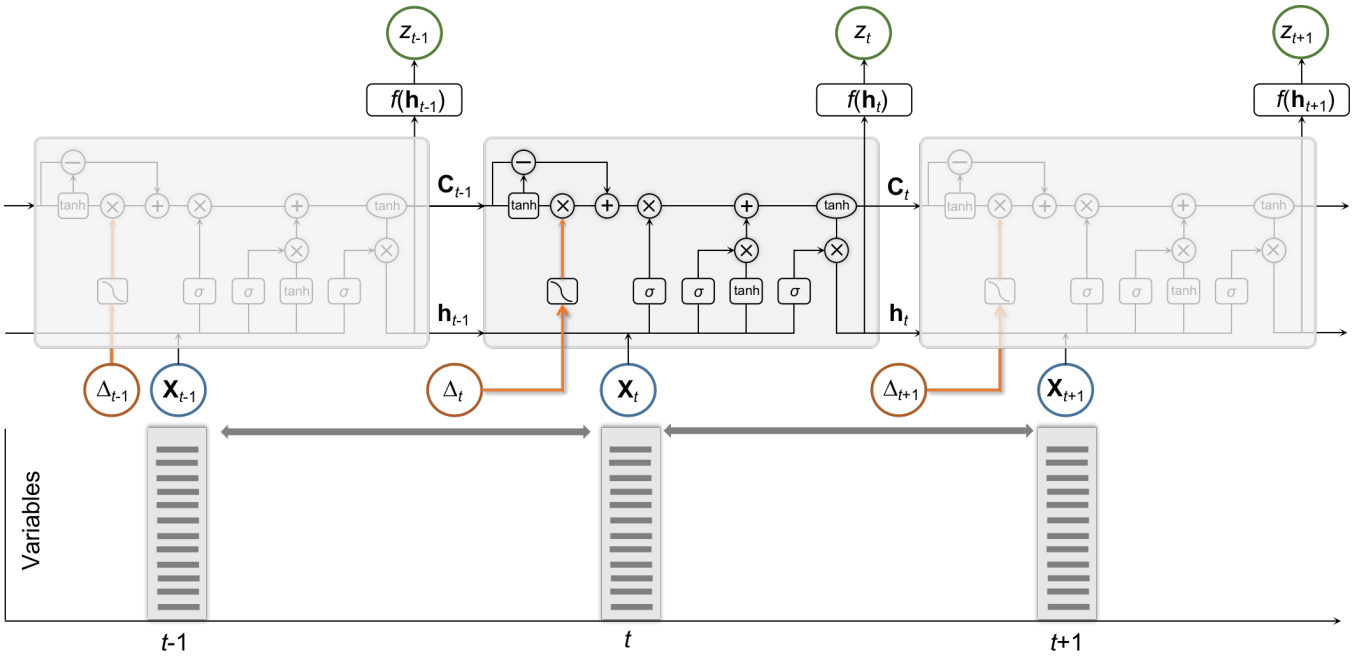

Figure 6: Illustration of the proposed time-dependent LSTM unit, and its application on analyzing our kidney patients' health records. Green boxes indicate networks and yellow circles denote point-wise operations. TD-LSTM takes two inputs, input record and the elapsed time

recurrent unit that computes a weighted sum of the input variables and applies a nonlinear function, each LSTM unit maintains a memory $\mathbf{C}_t$ at time t. This memory cell is updated by partially forgetting the adjusted previous memory $\mathbf{C}^*_{t-1}$ and adding new candidate memory content $\widetilde{\mathbf{C}}_t$ such as

$$\mathbf{C}_t = \mathbf{\Gamma}_t \odot \mathbf{C}^*_{t-1} + \mathbf{I}_t \odot \widetilde{\mathbf{C}}_t.$$

- This decomposition is data-driven, with the parameters of the decomposition network being learned simultaneously alongside the other network parameters through back-propagation. There is no strict requirement for the type of activation function used in the decomposition network, though the tanh function has been shown to perform slightly better.

  $$\mathbf{C}^{\mathrm{L}}_{t-1} = \mathbf{C}_{t-1} - \mathbf{C}^{\mathrm{S}}_{t-1}.$$

  The short-term memory $\mathbf{C}^{\mathrm{S}}_{t-1}$ is obtained by a network, *i.e.*,

  $$\mathbf{C}^{\mathrm{S}}_{t-1} = \tanh([\mathbf{w}^{\mathrm{d}}\boldsymbol{\omega}^{\mathrm{d}}])^{\top} + \mathbf{b}^{\mathrm{d}}.$$

  After obtaining the short-term memory, it is adjusted by applying an elapsed time weight, resulting in the discounted short-term memory, as follows:

  $$\widehat{\mathbf{C}}^{\mathrm{S}}_{t-1} = \phi(\Delta_t) \odot \mathbf{C}^{\mathrm{S}}_{t-1}.$$

  This adjustment accounts for the varying time intervals between consecutive elements in the sequence, ensuring that the memory reflects the temporal dynamics of the data more accurately.

  As a guideline, let Δ_t represent the elapsed time between $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$. The function $\phi(\cdot)$ is a heuristic decaying function that adjusts the influence of the short-term memory based on Δ_t. The function

$\phi(\cdot)$ should be monotonically non-increasing, meaning that as Δ_t increases, the effect of the short-term memory decreases. Different types of monotonically non-increasing functions can be chosen for $\phi(\cdot)$ depending on the measurement type of the durations specific to the application domain. In healthcare, where the elapsed time between successive records can range from days to years, we normalize the elapsed time to a common unit, such as days. For datasets with large elapsed times, we might use $\phi(\cdot) = \frac{1}{\Delta_t}$ or $\phi(\cdot) = \frac{1}{\log(e+\Delta_t)}$, ensuring that the short-term memory's influence diminishes appropriately with increasing elapsed time. Finally, to reconstruct the adjusted previous memory $\mathbf{C}^*_{t-1}$, the complement subspace of the long-term memory $\mathbf{C}^{\mathtt{L}}_{t-1}$ is combined with the discounted short-term memory $\widehat{\mathbf{C}}^{\mathtt{S}}_{t-1}$, *i.e.*,

$$\mathbf{C}^*_{t-1} = \mathbf{C}^{\mathtt{L}}_{t-1} \oplus \widehat{\mathbf{C}}^{\mathtt{S}}_{t-1}.$$

Here, $\oplus$ denotes the combination of the complement subspace of the long-term memory and the discounted short-term memory.

- The input weights in terms of input gate, forget gate, output gate and cell unit in hidden layer are denoted by $\mathbf{w}^{\mathrm{i}}, \mathbf{w}^{\mathrm{f}}, \mathbf{w}^{\mathrm{o}}, \mathbf{w}^{\mathrm{c}} \in \mathbb{R}^{D\times V}$ respectively, the recurrent weights by $\boldsymbol{\omega}^{\mathrm{i}}, \boldsymbol{\omega}^{\mathrm{f}}, \boldsymbol{\omega}^{\mathrm{o}}, \boldsymbol{\omega}^{\mathrm{c}} \in \mathbb{R}^{D\times D}$, and the bias as $\mathbf{b}^{\mathrm{i}}, \mathbf{b}^{\mathrm{f}}, \mathbf{b}^{\mathrm{o}}, \mathbf{b}^{\mathrm{c}} \in \mathbb{R}^{D}$. The input gate $\mathbf{I}^{\tau}$, forget gate $\mathbf{\Gamma}^{\tau}$, output gate $\mathbf{O}^{\tau}$, and candidate cell $\widetilde{\mathbf{C}}^{\tau}$ are computed by

$$\begin{pmatrix} \mathbf{I}_t \\ \mathbf{\Gamma}_t \\ \mathbf{O}_t \\ \widetilde{\mathbf{C}}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \begin{pmatrix} [\mathbf{w}^{\mathrm{i}}\ \boldsymbol{\omega}^{\mathrm{i}}] \\ [\mathbf{w}^{\mathrm{f}}\ \boldsymbol{\omega}^{\mathrm{f}}] \\ [\mathbf{w}^{\mathrm{o}}\ \boldsymbol{\omega}^{\mathrm{o}}] \\ [\mathbf{w}^{\mathrm{c}}\ \boldsymbol{\omega}^{\mathrm{c}}] \end{pmatrix} \Big[\mathbf{X}_t \mathbf{H}_{t-1}\Big]^{\top} + \begin{pmatrix} \mathbf{b}^{\mathrm{i}} \\ \mathbf{b}^{\mathrm{f}} \\ \mathbf{b}^{\mathrm{o}} \\ \mathbf{b}^{\mathrm{c}} \end{pmatrix},$$

where $[\mathbf{X}_t\ \mathbf{H}_{t-1}]^{\top}$ is the concatenation of the two vectors: input variables $\mathbf{X}_t$ and $\mathbf{H}_{t-1}$.

Model Learning

For patient i ($i = 1, 2, \ldots, N$) with variables $\mathbf{X}$ at time t, TdLSTM estimates the survival probability (*i.e.*, the probability of staying before the end stage) as follows:

$$S_i(t|\mathbf{X}_t, \mathbf{W}) = \sigma(\mathbf{w}^{\mathrm{out}}_t \mathbf{H}_t + b^{\mathrm{out}}_t).$$

TdLSTM cannot be an effective prediction model unless it achieves the objective that *the predicted survival probabilities approach the actual progression of ESKF*. We leverage the Kullback-Leibler (KL) divergence between the distributions of $S_i(t|\mathbf{X}, \mathbf{W})$ and survival status $\zeta_i(t) \in \{0, 1\}$ to quantify such approachability, as follows:

$$\mathcal{D}_i(t, \mathbf{W}) = \zeta_i(t) \ln \frac{\zeta_i(t)}{S_i(t|\mathbf{X}_t, \mathbf{W})} + (1 - \zeta_i(t)) \ln \frac{1 - \zeta_i(t)}{1 - S_i(t|\mathbf{X}_t, \mathbf{W})}. \tag{1}$$

The optimal weights aim to make $S_i(t|\mathbf{X}_t, \mathbf{W})$ as close as possible to 1 if patient i remains free of ESKF by time t_k, and close to 0 otherwise, with outputs of 1 and 0 representing definite true and definitely false predictions, respectively. Our learning objective is to minimize $\mathcal{D}_i(t, \mathbf{W})$ across time points $t_0, t_1, t_2, \ldots, t_K$ where survival statuses are known for all N_{tr} training patients. In this work, we set the time interval as one month, so $t_0 = 0$, $t_1 = 1$ month, continuing up to $t_K = 24$ months.

It is important to note that S naturally decreases from 1 to 0 over time, from the beginning to the onset of ESKF. Therefore, the minimization must be constrained by this monotonic behavior. The established penalty method converts the constrained optimization problem into a series of unconstrained optimization problems. Accordingly, we apply the static penalty[17] for violating the inequality constraints and minimize the average error calculated by

$$\begin{aligned} f(\mathbf{W}^*) = \underset{\mathbf{W}}{\text{minimize}} \quad & \frac{1}{N_{\text{tr}}}\sum_{i=1}^{N_{\text{tr}}}\Big(\frac{1}{K}\sum_{k\in\mathcal{K}}\mathcal{D}_i(t_k,\mathbf{W})\Big) \\ \text{subject to} \quad & S_i(t_k|\mathbf{X},\mathbf{W}) - S_i(t_{k+1}|\mathbf{X},\mathbf{W}) > 0 \;\; \forall k = 1,2\ldots,K-1,\; i = 1,2,\ldots,N_{\text{tr}}. \end{aligned}$$

The penalty method, which transforms a constrained minimization problem into a series of unconstrained optimization problems, proves to be effective. The solution to the unconstrained problem converges to that of the original problem. Consequently, by leveraging the benefits of the exterior-point method (EPM)[18], we establish the following objective function to account for violations of the survival probability constraints:

$$\mathcal{E}(\mathbf{W}^*) = \underset{\mathbf{W}}{\text{minimize}} \quad E(\mathbf{W}) + \frac{1}{N_{\text{tr}}\times(K-1)}\sum_{i=1}^{N_{\text{tr}}}\sum_{k=1}^{K-1}(\max\{0, S_i(t_k|\mathbf{X},\mathbf{W}) - S_i(t_{k+1}|\mathbf{X},\mathbf{W})\})^2.$$

Model learning is conducted using the forward-only Levenberg-Marquardt algorithm[19], which combines the speed efficiency of the Gauss-Newton algorithm[20] with the stability of the steepest descent method.

Model Evaluation

Upon obtaining the predicted survival probabilities for different time points from our model with the learned model parameters $\widehat{\mathbf{W}}$, the subsequent task entails estimating the time-to-ESKF for the test individual. To initiate this process, we establish the relative error (RE) at time t using the following definition:

$$E(t) = \sum\nolimits_{k=1}^{K}(|\log t - \log t_k|)S\left(t_k \mid \mathbf{X};\widehat{\mathbf{W}}\right),$$

thereby seeking the time at which E is the lowest, *i.e.*, the time-to-ESKF is

$$\widehat{T} = \text{argmin}_{t\in\{t_1,\ldots,t_K\}}E(t).$$

To evaluate the model's performance on the N patients, we redefine three independent metrics: the kidney failure AUC and the time-dependent concordance index (C-index), redefined as follows ($\mathbb{1}$ is the indicator function)

- AUC provides a probability measure of classification ability at a pre-specified time snapshot (*e.g.*, at t_K = 2 years in our case). It quantifies the model's ability to address the issue *"Is patient i likely to reach ESKF by time t?"*.

$$\text{AUC} = \frac{\sum_{i:\zeta_i(t_K)=0}\sum_{j:\zeta_j(t_K)=1}\mathbb{1}\{S_i(t_K|\mathbf{X},\widehat{\mathbf{W}}) < S_j(t_K|\mathbf{X},\widehat{\mathbf{W}})\}}{\big|\{i:\zeta_i(t_K)=0\}\big|\times\big|\{j:\zeta_j(t_K)=1\}\big|}.$$

- As a generalization of the AUC, C-index estimates how accurately the model can answer the question *"Which of patient i and patient j is more likely to reach the ESKF?"*.

$$\text{C-index} = \frac{\sum_{i:\zeta_i(t_K)=0}\sum_{j:T_i<T_j}\mathbb{1}\{S_i(t_k|\mathbf{X},\widehat{\mathbf{W}}) < S_j(t_k|\mathbf{X},\widehat{\mathbf{W}})\}}{\big|\{i:\zeta(t_K)=0\}\big|\times\big|\{j:T_i<T_j\}\big|}.$$

References

[1] Anna Francis, Meera N. Harhay, Albert C. M. Ong, Sri Lekha Tummalapalli, Alberto Ortiz, Agnes B. Fogo, Danilo Fliser, Prabir Roy-Chaudhury, Monica Fontana, Masaomi Nangaku, Christoph Wanner, Charu Malik, Anne Hradsky, Dwomoa Adu, Sunita Bavanandan, Ana Cusumano, Laura Sola, Ifeoma Ulasi, and Vivekanand Jha. Chronic kidney disease and the global public health agenda: an international consensus. *Nature Reviews Nephrology*, 20:473–485, 2024.

[2] K.J. Foreman, N. Marquez, A. Dolgert, K. Fukutaki, N. Fullman, M. McGaughey, M.A. Pletcher, A.E. Smith, K. Tang, C.W. Yuan, and J.C. Brown. Chronic kidney disease and the global public health agenda: an international consensus. *The Lancet*, 392(10159), 2024.

[3] Intisar Al Alawi, Issa Al Salmi, Adhra Al Mawali, Yacoub Al Maimani, and John A Sayer. End-stage kidney failure in oman: An analysis of registry data with an emphasis on congenital and inherited renal diseases. *International Journal of Nephrology*, 2017(1):6403985, 2017.

[4] Josef Coresh, Tanvir Chowdhury Turin, Kunihiro Matsushita, Yingying Sang, Shoshana H Ballew, Lawrence J Appel, Hisatomi Arima, Steven J Chadban, Massimo Cirillo, Ognjenka Djurdjev, et al. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. *JAMA*, 311(24):2518–2531, 2014.

[5] Yitan Zhu, Dehua Bi, Milda Saunders, and Yuan Ji. Prediction of chronic kidney disease progression using recurrent neural network and electronic health records. *Scientific Reports*, 13(1):22091, 2023.

[6] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pages 65–74, 2017.

[7] Arjan D van Zuilen, Michiel L Bots, Arzu Dulger, Ingeborg van der Tweel, Marjolijn van Buren, Marc A G J Ten Dam, Karin A H Kaasjager, Gerry Ligtenberg, Yvo W J Sijpkens, Henk E Sluiter, Peter J G van de Ven, Gerald Vervoort, Louis-Jean Vleming, Peter J Blankestijn, and Jack F M Wetzels. Multifactorial intervention with nurse practitioners does not change cardiovascular outcomes in patients with chronic kidney disease. *Kidney International*, 82(6):710–717, 2012.

[8] Marc Froissart, Jerome Rossert, Christian Jacquot, Michel Paillard, and Pascal Houillier. Predictive performance of the modification of diet in renal disease and Cockcroft-Gault equations for estimating renal function. *Journal of the American Society of Nephrology*, 16(3):763–773, 2005.

[9] Olivier Moranne, Marc Froissart, Jerome Rossert, Cedric Gauci, Jean-Jacques Boffa, Jean Philippe Haymann, Mona Ben M'rad, Christian Jacquot, Pascal Houillier, Benedicte Stengel, Bruno Fouqueray, and NephroTest Study Group. Timing of onset of CKD-related metabolic complications. *Journal of the American Society of Nephrology*, 20(1):164–171, 2009.

[10] Andrew S Levey, Lesley A Stevens, Christopher H Schmid, Yaping Lucy Zhang, Alejandro F Castro, 3rd, Harold I Feldman, John W Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, Josef Coresh,

and CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, 150(9):604–612, 2009.

[11] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

[12] Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 1(43):45–57, 1975.

[13] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems (NIPS)*, 24:1845–1853, 2011.

[14] Jianfei Zhang, Lifei Chen, Yanfang Ye, Gongde Guo, Rongbo Chen, Alain Vanasse, and Shengrui Wang. Survival neural networks for time-to-event prediction in longitudinal study. *Knowledge and Information Systems*, 62:3727–3751, 2020.

[15] Navdeep Tangri, Morgan E Grams, Andrew S Levey, Josef Coresh, Lawrence J Appel, Brad C Astor, Gabriel Chodick, Allan J Collins, Ognjenka Djurdjev, C Raina Elley, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. *JAMA*, 315(2):164–174, 2016.

[16] Stef Van Buuren and Catharina GM Oudshoorn. Multivariate imputation by chained equations, 2000.

[17] Zbigniew Michalewicz and Marc Schoenauer. Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary Computation*, 4(1):1–32, 1996.

[18] Hiroshi Yamashita and Takahito Tanabe. A primal-dual exterior point method for nonlinear optimization. 20(6):3335–3363, 2010.

[19] Bogdan M Wilamowski and Hao Yu. Neural network learning without backpropagation. 21(11):1793–1803, 2010.

[20] Philip E Gill and Walter Murray. Algorithms for the solution of the nonlinear least-squares problem. *SIAM Journal on Numerical Analysis*, 15(5):977–992, 1978.