Leveraging Pretrained Large Language Model for Prognosis of Type 2 Diabetes Using Longitudinal Medical Records

Phong B.H. Nguyen[1,2,3], Andreas Hungele[3,4], Reinhard W. Holl[3,4+], Michael P. Menden[1,3,5+]

[1]Computational Health Center, Institute of Computational Biology, Helmholtz Munich, Munich, Germany
[2]Faculty of Biology, Ludwig-Maximilian University Munich, Munich, Germany
[3]German Center for Diabetes Research (DZD), Munich-Neuherberg, Germany
[4]Institute for Epidemiology and Medical Biometry, Ulm University, Ulm, Germany
[5]Department of Biochemistry and Pharmacology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Melbourne, Australia

[+]Corresponding authors: reinhard.holl@uni-ulm.de & michael.menden@unimelb.edu.au

Abstract
Timely prognosis of type 2 diabetes (T2D) complications is critical for effective interventions and reducing economic burden. AI-driven large language models (LLMs) offer potential for extracting clinical insights but face challenges due to the sparse, high-dimensional nature of longitudinal medical records. This study demonstrates the utility of LLMs in medical time series prediction by preprocessing data with a missing mask, adding an embedding layer to a pretrained LLM, and fine-tuning both components. The fine-tuned model outperformed baselines in predicting both HbA1c and LDL levels using the DPV registry dataset of 449,185 T2D patients, achieving Pearson's correlations of 0.749 and 0.754, with a delta improvement of 0.253 and 0.259, respectively. The model also demonstrated robust long-term prediction for HbA1c over 554.3 days (95% CI: [547.0, 561.5]), with a 9% improvement in MSE over last-observation-based methods. Integrated gradient analysis identified significant clinical features and visits, revealing potential biomarkers for early intervention. Overall, the results showed the possibility to leverage the prediction power of LLM in T2D prognosis using sparse medical time series, assisting clinical prognosis and biomarker discovery, ultimately advancing precision medicine.

Introduction
Type 2 diabetes (T2D) is a chronic metabolic disorder that leads to elevated blood glucose levels[1] and is associated with micro- and macrovascular complications, including cardiovascular diseases, neuropathy, nephropathy, and retinopathy, contributing to significant morbidity and mortality worldwide[2]. The rising global prevalence of T2D underscores the urgent need for effective management strategies. Early prognosis of T2D complications is crucial for enabling timely interventions, improving patient outcomes, and reducing healthcare costs.

Longitudinal medical records are a vital resource for advancing precision medicine in T2D. These records provide comprehensive patient data over time[3], including demographics, medical history, test results, medications, and lifestyle factors, making them invaluable for early prognosis. Studies like the Framingham Heart Study and DCCT have demonstrated the utility of such data in predicting cardiovascular risk [4] and diabetic complications[5]. These datasets enable the identification of patterns for more accurate and personalized prognostic assessments, but their large volume and complexity pose significant analytical challenges.

Machine and deep learning has revolutionized medical prognosis by enabling the analysis of complex longitudinal medical data. Various deep learning architectures, including feed-forward neural networks,

convolutional neural networks (CNNs), and recurrent neural networks (RNNs), have been employed to model these datasets. For example, Deepr[6] used word embeddings and CNNs to predict hospital readmissions, while RETAIN[7] combined RNNs with attention mechanisms and learning interpretable representation to model temporal data and improve interpretability.

Transformer-based deep learning has significantly advanced the modeling of longitudinal medical records. Med-BERT leverages transformers to encode electronic health records (EHRs), capturing intricate patterns and temporal dependencies to enhance predictions of disease progression and hospital readmission[8]. Similarly, BEHRT extends the BERT architecture with positional embeddings for time intervals and pretraining through a masked language model[9], enabling detailed patient history representations and improved clinical predictions.

Data sparsity remains a major challenge in healthcare datasets, where missing values can contain important predictive information. Several methods have been proposed to specifically handle missingness in medical time series. Gated Recurrent Unit Decay (GRU-D)[10] uses a decay mechanism to improve imputation accuracy, while BRITS[11] applies bidirectional RNNs to manage irregular time series. NAOMI[12], a non-autoregressive model, refines imputed values with a multi-resolution strategy. However, these models are complex, difficult to scale, and still face issues such as memory loss, gradient problems, and irregular time intervals.

State-of-the-art pretrained large language models (LLMs) like GPT-3 and BERT have shown promise in biomedical tasks, including analyzing longitudinal health records for prognosis. For example, BERT was fine-tuned on clinical notes to predict outcomes like hospital readmission[13], and GatorTron[14] analyzed both structured and unstructured data to predict disease progression. MedGPT[15], fine-tuned on EHR data, outperformed traditional models in predicting diabetic complications. However, most studies trained LLMs from scratch using clinical data, neglecting the potential of fine-tuning pretrained models. Additionally, these studies focused on unstructured data and overlooked model interpretability.

This study investigated the application of foundation models to prognosis tasks using a large longitudinal T2D medical dataset. A novel approach was developed to adapt sparse, structured data for input into a pretrained large language model (LLM), leveraging minimal preprocessing alongside a learnable embedding layer. Both the embedding layer and the LLM were jointly trained to predict prognosis endpoints. Furthermore, the model's interpretability was evaluated, highlighting its potential to support personalized healthcare decision-making.

Results

This study utilized the German Diabetes Prospective Follow-up Registry (DPV) Initiative [16], which contains 649,331 electronic health records from over 400 facilities across four European countries. The analysis targeted 449,185 type 2 diabetes (T2D) patients, selecting 50,183 individuals with at least 10 clinical visits to ensure sufficient historical data (**Methods**; **Fig. 1a**). After preprocessing, the final datasets comprised 36,733 samples for HbA1c prediction and 15,139 samples for LDL prediction, both exhibiting approximately normal distributions (**Fig. 1a**).

The data preparation process structured patient records for prognosis prediction (**Fig. 1b**; **Methods**). Each timeline spanned 10 to 50 visits, with the feature window, the timeframe from first visit to the prediction point, capturing up to 321 features (e.g., diagnoses, demographics, measurements, medications, and hospitalizations), though not all were consistently recorded. The prediction window, extending from the prediction point to the target point, defined the forecasting horizon (**Fig. 1b**). The mean record lengths

for HbA1c and LDL samples were 2,651 days (95% CI [2,633; 2,669]) and 2,661 days (95% CI [2,642; 2,679]), respectively (**Fig. 1c,d**).

Adaptation of pretrained large language models to sparse multivariate medical time series

This machine learning framework was developed to process sparse medical time series data and adapt it for use with pretrained LLMs, harnessing their predictive capabilities in the medical domain. It addresses challenges associated with missing data, a common issue in medical records, and transforms the data into a format compatible with pretrained language models.

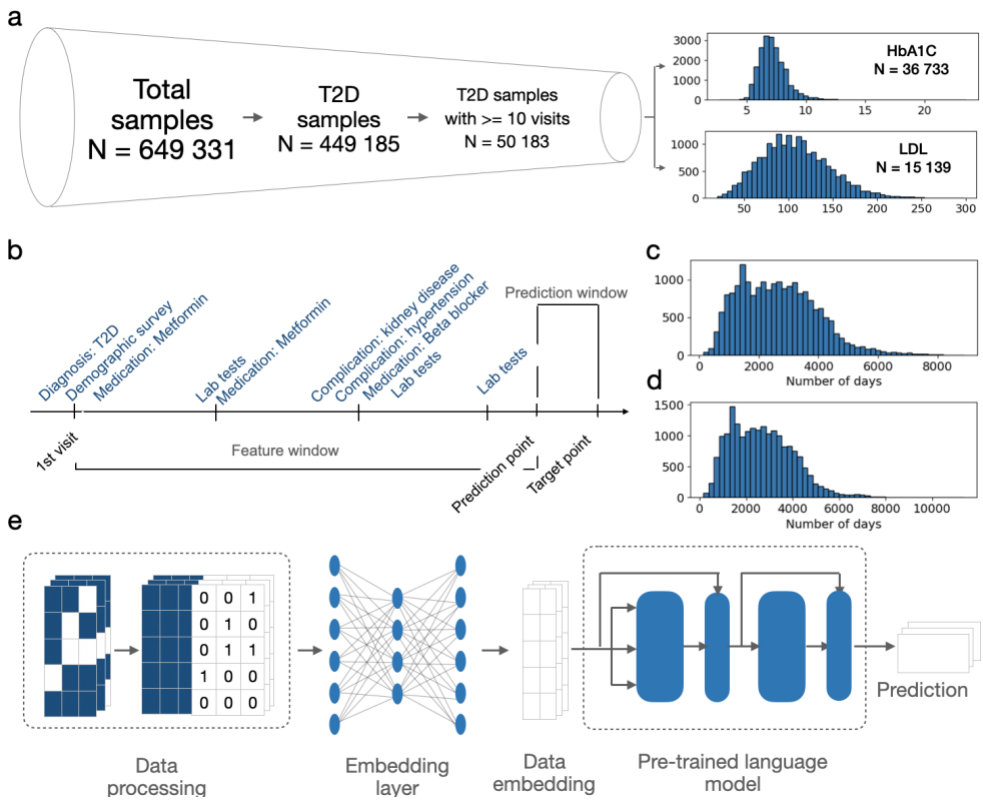

Figure 1. Overview of the data processing and machine learning framework for adapting sparse medical time series to pretrained large language models. (**a**) A filtering process was applied to select relevant samples for each prediction task, with histograms showing the distribution of selected samples for HbA1c and LDL predictions. (**b**) Example of a patient medical record, illustrating the timeline from the first visit to the target point, including the feature and prediction windows. (**c**,**d**) Distribution of time series lengths for HbA1c and LDL, respectively. (**e**) The machine learning framework processes sparse medical data, which is passed through an embedding layer to generate a reduced representation, then fed into the pretrained language model for prediction output.

The framework consists of a multi-step process designed to generate predictions from raw medical data (**Methods**, **Fig. 1e**). The first step is preprocessing, which identifies missing values for each variable using a missing mask. This mask is appended to the original data matrix, and missing values are imputed with the mean of corresponding variables across time points, ensuring the data remains complete for analysis.

Once preprocessing is complete, the data is passed through an embedding layer (**Methods**, **Fig. 1e**). This layer alters the dimensionality of the data while preserving its essential features, ensuring the data is compatible with the pretrained LLM that will be used for prediction.

In the final step, the embedded data is input into a pretrained LLM, which has been trained on vast amounts of textual data. The LLM is adapted to predict health outcomes based on the provided structured medical data (**Methods**, **Fig. 1e**). In this study, the framework was used to predict HbA1c and LDL levels for either the next visit or a future visit, demonstrating its ability to predict complex health outcomes from longitudinal medical records.

LLM-based models excel in long-term medical time series forecasting

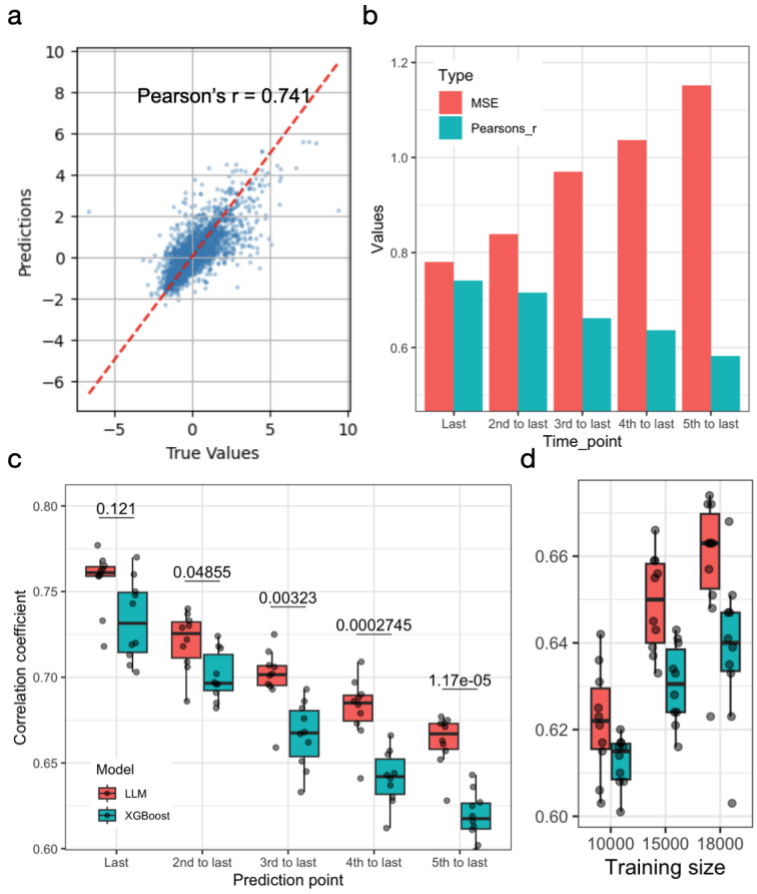

Figure 2. Comparison of LLM-based and XGBoost models in medical time series forecasting of HbA1c level. (**a**) Correlation between observed and predicted HbA1c levels using XGBoost trained on snapshot data from 17,500 samples at the prediction point. Results are shown for a test set of 2,500 samples. (**b**) XGBoost performance at varying prediction points ("last" denotes the final time point before the target point), evaluated using mean squared error (MSE) and Pearson's correlation coefficient. (**c**) Performance comparison of XGBoost and LLM-based models trained on time series across prediction points, with results from 10 resamples. P-values were calculated using the Wilcoxon Rank Sum test. (**d**) Model performance as a function of training size, within a subset of 20,000 samples, at the prediction point "4th to last," with results averaged across 10 resamples.

We first evaluated an XGBoost baseline model trained on snapshot data using a random training sample of 18,000 samples. The model predicted HbA1c levels for the next doctor visit with a Pearson's correlation coefficient of 0.741 (**Fig. 2a**), demonstrating the sufficiency of a shallow learning approach for this task. However, extending the prediction window by shifting the prediction point further from the target caused significant performance decline, with the MSE increasing from 0.780 to 1.152 and the correlation coefficient dropping from 0.741 to 0.582 (**Fig. 2b**). These results underscore the limitations of shallow snapshot-based models for long-term predictions and the necessity of capturing complex temporal patterns in historical data.

To evaluate the clinical utility of deep learning models for time series data, we benchmarked the XGBoost model against our LLM-based model trained on medical time series data, using the same dataset across multiple prediction points (**Fig. 2c**). The comparison focused on predicting HbA1c levels at varying prediction window lengths, ranging from a mean of 124.5 days (95% CI: [121.0, 126.8]) for the next visit to 554.3 days (95% CI: [547.0, 561.5]) for the fifth future visit (**Supp. Fig. S1**).

The LLM-based model outperformed the XGBoost model, especially for longer prediction windows. Although the LLM model showed a slightly higher correlation coefficient for predicting HbA1c at the next visit, this difference was not statistically significant (adjusted p-value = 0.121). As the prediction window expanded, the LLM model's advantage became more pronounced, particularly at the longest prediction window (adjusted p-value = 1.17e-05; **Fig. 2c**). Furthermore, the LLM model consistently outperformed XGBoost in predicting HbA1c at the fifth visit (**Fig. 2d**). These results underscore the LLM model's ability to capture long-term trends in time series data.

Scalability analysis reveals the robustness of LLM-based models across varying sample sizes.
To assess the LLM-based model's performance across varying sample sizes, we predicted HbA1c and LDL levels at the next doctor visit. Comparisons were made with an LSTM model trained from scratch (as opposed to our pretrained foundation model) and a random forest model relying on the last observed target value and prediction window length (LastObs). Target-related variables were excluded during training for both LLM- and LSTM-based models to prioritize the identification of key biomarkers (**Methods**). Training and testing were conducted on subsets of 20,000 samples at varying train-test ratios, providing a thorough evaluation of model performance.

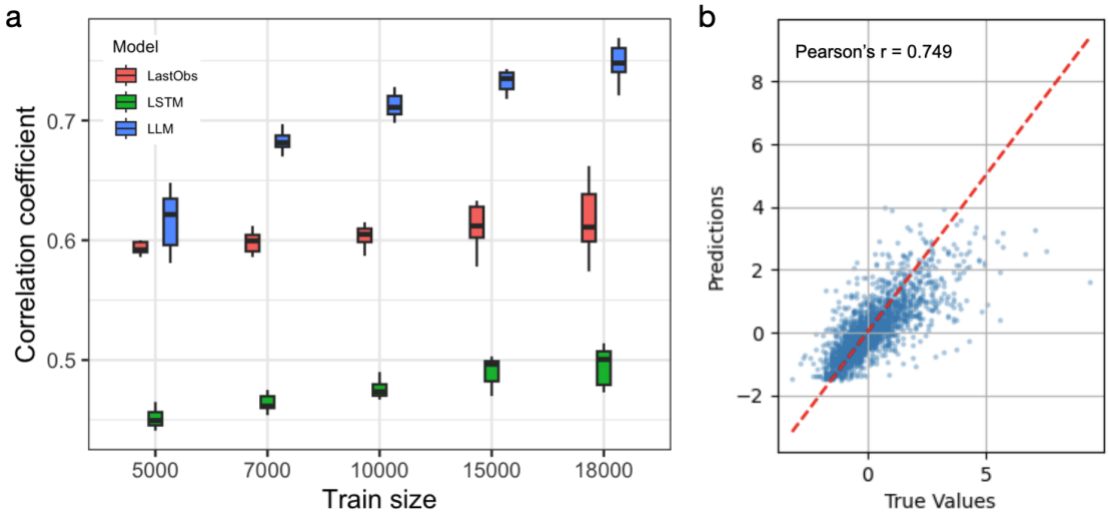

Figure 3. The LLM-based framework effectively predicts HbA1c levels for the next visit. (**a**) Comparison of the LLM-based framework, LSTM-based method, and random forest model trained on the last observed HbA1c value (LastObs) in terms of Pearson's correlation coefficient. (**b**) Predicted HbA1c values plotted against true values in the test dataset, using 18,000 samples for training and 2,000 samples for testing.

Using Pearson's correlation coefficient as the performance metric, the LLM-based model consistently outperformed both the LSTM and LastObs models across all training sizes, ranging from 5,000 to 18,000 samples, for both HbA1c and LDL predictions. This demonstrates the robustness of the LLM-based model in capturing complex patterns in the data.

For HbA1c prediction, the LLM-based model showed significant improvements in correlation coefficients as training size increased, highlighting its superior ability to leverage larger datasets. The correlation coefficients improved from 0.618 (95% CI [0.601, 0.653]) at a training size of 5,000 to 0.749 (95% CI [0.738, 0.759]) at 18,000. In contrast, the LastObs model demonstrated minimal improvement with increasing sample size, achieving correlation coefficients that ranged from 0.593 (95% CI [0.590, 0.597]) to 0.619 (95% CI [0.598, 0.640]).

The LSTM-based models, trained without target-related variables, failed to capture the complex patterns inherent in time-series data, resulting in significantly lower performance compared to the LLM-based approach. These models achieved mean correlation coefficients ranging from 0.451 (95% CI [0.445, 0.457]) to 0.495 (95% CI [0.484, 0.507]), underscoring their limited ability to learn from the available data.

Increasing the sample size had a particularly pronounced effect on the LLM-based model, further emphasizing its scalability and adaptability. Compared to the other methods, the LLM-based model showed a steeper improvement in performance, especially at smaller training sizes. At the largest sample size of 18,000, the predicted HbA1c values demonstrated a strong correlation with the true values, highlighting the model's accuracy and capacity to leverage large datasets effectively (Fig. 3b).

To further evaluate the clinical relevance of the LLM-based model, we conducted a post hoc analysis on HbA1c prediction. The raw predicted values were transformed into their original range using the known mean and standard deviation of HbA1c in our dataset. Subsequently, the samples were categorized into healthy (<7.5%) and unhealthy (≥7.5%) groups, and classification performance was assessed using the observed HbA1c values. Fisher's Exact Test revealed a strong enrichment between the true and predicted classifications (p-value = 4.90e-170; **Supp. Fig. S2**), reinforcing the model's superior performance and clinical applicability.

The LDL prediction benchmarking followed similar patterns to HbA1c, with LLM-based models achieving correlation coefficients from 0.650 (95% CI [0.639, 0.660]) at 5,000 samples to 0.749 (95% CI [0.738, 0.759]) at 18,000 samples, consistently outperforming LSTM models (**Supp. Fig. S3**). While the LastObs model outperformed the LLM-based model at 5,000 samples (correlation coefficient = 0.692), the LLM-based models significantly surpassed the LastObs models at larger sample sizes (10,000+ samples; Wilcoxon Rank Sum test, $p < 0.05$; **Supp. Fig. S3a**).

Investigation of LLM-based Model Interpretability for Prognosis of T2D

To better understand the clinical relevance of the LLM-based model in predicting T2D progression, we performed an interpretability analysis. Specifically, we focused on identifying the features most influential for predictions of HbA1c and LDL levels (**Methods**). Using integrated gradients, a robust method for evaluating feature importance, we analyzed a subset of 500 samples across 50 doctor visits. The results allowed us to determine which biomarkers and patient characteristics contributed most to model predictions (**Fig. 4**; **Supp. Fig. S4**), providing insights into the key drivers of T2D prognosis and offering a pathway for more targeted interventions and personalized care.

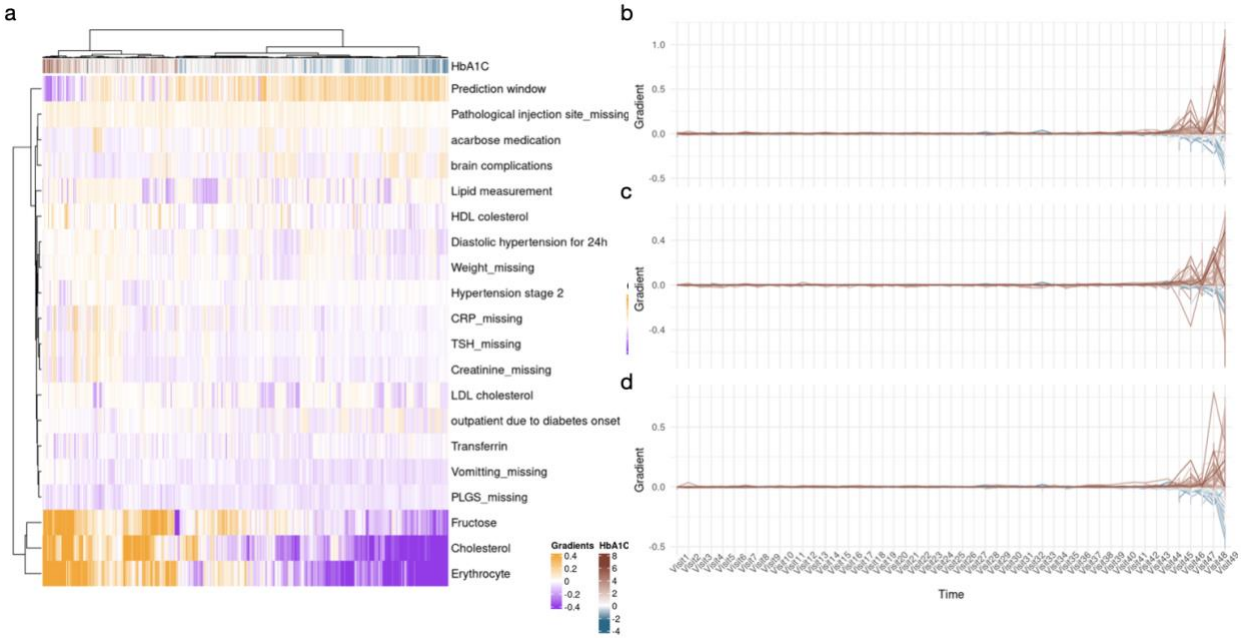

Figure 4. Most important features for HbA1c prediction. (**a**) Heatmap displaying the feature importance scores, represented by the integrated gradients of the top 20 most important features. The values represent the mean absolute gradients of the features across samples and time points.(**b-d)** Trajectories of erythrocyte count, cholesterol, and fructose gradients across visits within the feature window, respectively. Each line corresponds to one of 500 samples, with colors indicating the observed HbA1c level at the target visit.

The integrated gradients analysis for HbA1c prediction provided valuable insights into the temporal and feature-specific aspects influencing the model's decision-making. Most of the feature importance gradients were concentrated on the last five visits before the target point, indicating that the information from these visits is crucial for accurate prediction (**Supp. Fig. S4a**). This suggests that recent medical history plays a significant role in predicting future HbA1c levels.

In terms of specific features, the analysis revealed that erythrocyte levels, cholesterol levels, fructose levels, and the length of the prediction window were among the most influential factors, as indicated by the highest absolute gradients across samples and time points (**Supp. Fig. S4b**). Erythrocyte, cholesterol, and fructose levels exhibited a positive correlation with the predicted HbA1c levels (**Fig. 4a**), with their influence being most prominent during the final visits in the feature window (**Fig. 4b-d**). This finding highlights the importance of these biomarkers in shaping predictions for HbA1c.

Additionally, missing data patterns also emerged as significant predictors. Specifically, the absence of thyroid stimulating hormone and C-reactive protein levels were identified as important factors, suggesting that missing data on these features could impact the prediction of HbA1c. These findings emphasize the need for careful consideration of both observed and missing feature values in clinical predictive modeling.

The integrated gradients also gave insights into important visits and features for the LDL prediction. Specifically, most of the gradients concentrated in the last ten visits before the target point, emphasizing the critical role of these visits in predicting LDL levels (**Supp. Fig. S5a**). When analyzing the most important features, the top features were thyroid stimulating hormone level, triglyceride level and missing number of prandial insulin injections (**Supp. Fig. S5b**). All of these features are positively associated with LDL level, especially in the last ten visits of the feature window (**Supp. Fig. S6**). These findings highlight the relevance

of specific biomarkers and the impact of missing data on LDL prediction, reinforcing the importance of these factors in the model's decision-making process.

Discussion

In this work, we introduced a novel approach to adapt sparse numerical data from medical time series to a pretrained large language model (LLM). Instead of using traditional text prompt engineering, we imputed missing data and added a learnable embedding layer between the input and pretrained LLM layers. This approach allowed direct input of numerical data into the transformer architecture, effectively leveraging the LLM's predictive power. Our results demonstrate strong performance in both prediction and clinical utility.

The LLM-based approach demonstrated strong predictive performance for HbA1c and LDL, outperforming both LSTM and LastObs models across various training sizes, with robust results even at smaller sample sizes. This advantage is likely due to the LLM's pretraining on large datasets, reducing the need for extensive fine-tuning. However, performance improved significantly with larger sample sizes, indicating the model's benefit from increased data to handle clinical variability. Additionally, the LLM-based models showed high correlation in predicting HbA1c levels up to 18 months ahead, highlighting the importance of historical data in prediction. Overall, the approach shows strong clinical utility for type 2 diabetes prognosis.

An important feature of our model is its ability to provide mechanistic insights into the patient's pathological trajectory. To enhance accessibility for patients and healthcare practitioners, interpretability is essential. In this study, we used integrated gradients to interpret the model’s predictions, focusing on important visits and influential features. Our analysis revealed that only the most recent visits—five for HbA1c and ten for LDL—were critical for predictions. Clinically, this suggests that while historical data is important, the most recent visits hold the most relevance for prognosis. This highlights the model's clinical utility in guiding patient care decisions.

For HbA1c prediction, the most important features were erythrocyte, cholesterol, and fructose levels. Erythrocyte count, which is influenced by blood glucose levels through glucose binding to hemoglobin A1c[17], showed an inverse association with HbA1c levels, although this relationship is complex[18] and warrants further investigation. Cholesterol and fructose are established markers linked to insulin resistance, with high LDL cholesterol and fructose consumption contributing to poor glycemic control and elevated HbA1c levels[19,20]. These findings highlight the potential of these features as prognostic and preventive biomarkers for T2D.

For LDL prediction, the key features were thyroid stimulating hormone (TSH), triglycerides, and missing prandial insulin injections, all positively associated with LDL. Elevated TSH levels, often seen in hypothyroidism, reduce LDL clearance, leading to higher LDL levels[21]. Triglycerides, linked to various metabolic disorders, also correlate with LDL[22]. Missing prandial insulin injections, crucial for glucose control in diabetes, contribute to dyslipidemia and increased LDL[23]. These findings suggest that these features may serve as diagnostic and prognostic biomarkers for LDL and related conditions like dyslipidemia.

While valuable insights were gained, future research is needed to validate the identified biomarkers using external datasets. Integrating molecular data could further elucidate mechanisms, and increasing patient recruitment may allow for prediction of additional conditions. Despite limitations, the superior performance in predicting HbA1c and LDL supports the approach as a proof of concept.

Deep learning models demonstrate strong clinical performance but face limitations, such as reliance on large amounts of high-quality datasets, which is often not available in real-world clinical settings. These models require extensive preprocessing and significant computational resources, hindering their ability to capture complex relationships. Additionally, the use of generative AI in healthcare introduces ethical challenges, including data privacy concerns, algorithmic bias, and the potential for misuse of AI-generated predictions. Solutions may involve more efficient architectures, data augmentation, and addressing ethical considerations to ensure responsible use in clinical settings.

In conclusion, this study introduced a novel approach leveraging pretrained LLMs for the prognosis of T2D patients using time series electronic health records. The LLM-based method demonstrated superior performance in forecasting future HbA1c and LDL levels, identifying critical past visits, and highlighting influential features as potential biomarkers for these conditions. These results underscore the clinical utility of the approach in enhancing diagnosis, prognosis, and prevention strategies, thereby advancing the application of precision medicine in healthcare.

Methods

Study dataset, data processing and experimental design

The study was conducted under a data sharing agreement with the DPV Initiative, a project initiated by the German Diabetes Center (DZD) and executed and maintained by Ulm University, Germany, from which we obtained the data. The DPV Initiative has been collecting patient medical records since the 1990s from over 500 treatment facilities across Germany, Austria, Switzerland and Luxembourg. At the time of acquisition, the data consisted of 649,331 patients, of which 449,185 were of T2D. The number of doctor visits in T2D patients ranged from 1 to 241.

To ensure sufficient information for model training, we selected patient records consisting of *n* visits, where *n* ranged from 11 to 51 ($n \in [10, 51]$). The timeline was divided into three components: the feature window, the prediction point, and the prediction window.

1. **Feature Window (t_1 to t_{n-1})**:
   The feature window spans visits from the start of the timeline (t_1) to the prediction point (t_{n-1}). During this period, up to 321 features were collected per visit, including diagnoses, demographics, physical and laboratory measurements, medications, and hospitalizations. Some features were not recorded for every visit. The feature window is represented as:
   Feature Window = $\{F_1, F_2, ..., F_{n-1}\}$, where F_i is the feature set for visit t_i.
2. **Prediction Point (t_{n-1})**:
   The prediction point (t_{n-1}) is the last visit used for making predictions. Data up to this point is used to predict the target outcome.
3. **Prediction Window (t_{n-1} to t_n)**:
   The prediction window spans from the prediction point (t_{n-1}) to the target point (t_n), defining the forecasting horizon. The target point (t_n) contains the outcome to be predicted.

The final numbers of samples were 52 122 for HbA1c and 18 960 for LDL. The dataset initially comprised 321 variables, encompassing demographics, lifestyle factors, physiological measurements, diagnoses, treatments, medications, and administrative records. These variables underwent extensive preprocessing to ensure data quality and compatibility with the model. Uniform variables, those entirely missing, and variables with less than 5% data availability were excluded. Inconsistent encodings were

corrected, and highly correlated or redundant variables were removed. Categorical variables were transformed using one-hot encoding.

To address missing data, numerical variables were imputed using the global mean of the respective variable, while missing values in categorical variables were encoded as "missing" prior to one-hot encoding. Additionally, a missing mask was generated, capturing the missingness pattern by encoding '1' for missing and '0' for present data points. This mask, designed to retain potentially informative missingness patterns, was appended to the processed data table.

After preprocessing, the final datasets contained 760 variables for HbA1c prediction and 749 variables for LDL prediction. All variables were standardized to follow a normal distribution to facilitate model training.

Benchmarking with different training sizes

To optimize computational efficiency and reduce training time, a random subset of 20,000 samples was selected for HbA1c benchmarking experiments, while all available samples were used for LDL. Within the selected samples, prediction performance was assessed across various training set sizes, including 5,000, 7,000, 10,000, 15,000, and 18,000 samples (HbA1c only). The remaining samples were reserved for testing.

For deep learning models, 2,000 samples from the training set were allocated for validation. Early stopping was implemented using a patience parameter of 30 epochs to mitigate overfitting during training. To ensure robustness and reproducibility, models were trained on 10 resampled datasets for each train-test ratio.

Prediction Window Creation and Evaluation

Prediction windows were generated by shifting the prediction point backward along the patient's timeline. Specifically, the prediction point was moved from the visit immediately preceding the target point to five visits prior, resulting in corresponding prediction windows. For each window, the same training and evaluation procedure as described for benchmarking training set sizes was applied.

Since the target point was shifted up to five visits into the past, the maximum sequence length was reduced by five. Consequently, the sequence length for these prediction windows ranged from 5 to 45.

Adaptation of the pretrained LLM

To enable numerical data integration with a pretrained LLM without requiring text prompt engineering, we implemented a learnable embedding layer that transformed raw numerical inputs into embeddings suitable for the LLM while maintaining flexibility and efficiency. The embedding layer, designed as a two-layer multilayer perceptron (MLP) with 50 nodes per hidden layer, allowed customization of input and output dimensions to align with the dataset and LLM architecture. Specifically, the input dimensions were 760 for HbA1c and 749 for LDL, while the output dimension matched BERT's embedding size of 768. The embedding layer was directly connected to BERT's input encoder, bypassing its tokenizing and embedding steps, and both the embedding layer and BERT were jointly trained during model training. Although BERT was used for simplicity in this study, the approach is versatile and can be extended to other pretrained LLMs, demonstrating its broad applicability to diverse tasks.

The LLM was initialized with pretrained parameters, and dropout (70%) was applied to both the embedding layer and BERT to prevent overfitting. Models were trained for up to 150 epochs, with early stopping triggered after 30 epochs without improvement on the validation set. Training used the Adam

optimizer[24] with a learning rate of 2e-5, weight decay of 5e-3, and mean square error as the loss function. Prediction performance on the test set was assessed using Pearson's correlation coefficient.

Implementation of long short term memory (LSTM)

The LSTM model consisted of two hidden layers with 100 dimensions and was trained similarly to the LLM-based method, with the processed input data fed directly into the LSTM. Unlike the LLM, the LSTM was initialized from scratch. Input and output dimensions, training hyperparameters, and regularization mechanisms mirrored those of the LLM-based approach.

Implementation of XGBoost

The XGBoost models were trained using 100 boosted trees with mean squared error as the loss function, while all other parameters were set to the defaults of the XGBRegressor function in the Python xgboost library[25]. Only variables from the prediction time point were used as features, excluding time series data. The training and testing scheme followed the same protocol as the LLM- and LSTM-based methods.

Integrated gradients and analysis of feature importance

We used integrated gradients to interpret the model predictions. This feature attribution method quantifies the contribution of each input feature by calculating the path integral of gradients from a baseline input (typically zero) to the actual input. It computes the gradients of the model's output with respect to the input at various points along this path and averages them. Integrated gradients offer a more stable and accurate attribution of feature importance compared to simpler gradient methods. Further details can be found in the original publication[26].

To interpret the model predictions, we computed integrated gradients for a subset of 500 samples with 50 visits using the trained model. Both total gradients and average absolute gradients across all samples were calculated. The average absolute gradients were then used to identify influential visits, and the most important individual features were determined by summing the gradients across time points.

Statistical and computational analysis of results

The analysis framework was developed using PyTorch, with data processing, model training, and evaluation conducted using various Python libraries. Hypothesis testing for two continuous variables was performed using the Wilcoxon Rank Sum test, with p-values adjusted for FDR. For two categorical variables, Fisher's Exact test was used. Data visualization was carried out using ggplot2 in R.

Code availability

The code for the analysis and computational framework is located here:https://github.com/phngbh/DPV.

Data availability

The data is not publicly accessible due to the patient's confidentiality. Researchers who would like to reproduce the results could contact the data access committee at the DPV Initiative in Ulm University. Detailed information can be found here: https://buster.zibmt.uni-ulm.de/.

References

1. Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes.
2. Farmaki, P. *et al.* Complications of the type 2 diabetes mellitus. *Curr. Cardiol. Rev.* **16**, 249–251 (2020).

3. Lee, S. *et al.* Unlocking the Potential of Electronic Health Records for Health Research. *Int J Popul Data Sci* **5**, 1123 (2020).
4. Kannel, W. B. Some lessons in cardiovascular epidemiology from Framingham. *Am. J. Cardiol.* **37**, 269–282 (1976).
5. Nathan, D. M. & for the DCCT/EDIC Research Group. The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and complications study at 30 years: Overview. *Diabetes Care* **37**, 9–16 (2014).
6. Nguyen, P., Tran, T., Wickramasinghe, N. & Venkatesh, S. Deepr: A Convolutional Net for Medical Records. *arXiv [stat.ML]* (2016).
7. Choi, E. *et al.* RETAIN: An interpretable predictive model for healthcare using REverse Time AttentIoN mechanism. *arXiv [cs.LG]* (2016).
8. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, (2021).
9. Li, Y. *et al.* BEHRT: Transformer for electronic health records. *Sci. Rep.* **10**, 7155 (2020).
10. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**, (2018).
11. Cao, W. *et al.* BRITS: Bidirectional recurrent imputation for time series. *arXiv [cs.LG]* (2018).
12. Liu, Y., Yu, R., Zheng, S., Zhan, E. & Yue, Y. NAOMI: Non-Autoregressive Multiresolution Sequence Imputation. *arXiv [cs.LG]* (2019).
13. Huang, K., Altosaar, J. & Ranganath, R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv [cs.CL]* (2019).
14. Yang, X. *et al.* GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. *bioRxiv* (2022) doi:10.1101/2022.02.27.22271257.
15. Kraljevic, Z. *et al.* MedGPT: Medical concept prediction from clinical narratives. *arXiv [cs.CL]* (2021).
16. Diabetes-Patienten-Verlaufsdokumentation (DPV). https://buster.zibmt.uni-ulm.de/.
17. Wang, Y. *et al.* The relationship between erythrocytes and diabetes mellitus. *J. Diabetes Res.* **2021**, 6656062 (2021).
18. English, E. *et al.* The effect of anaemia and abnormalities of erythrocyte indices on HbA1c analysis: a systematic review. *Diabetologia* **58**, 1409–1421 (2015).
19. American Diabetes Association. Dyslipidemia management in adults with diabetes. *Diabetes Care* **27**, s68–s71 (2004).
20. Stanhope, K. L. Role of fructose-containing sugars in the epidemics of obesity and metabolic syndrome. *Annu. Rev. Med.* **63**, 329–343 (2012).
21. Duntas, L. H. & Brenta, G. A renewed focus on the association between thyroid hormones and lipid metabolism. *Front. Endocrinol. (Lausanne)* **9**, 511 (2018).
22. Parhofer, K. G. & Laufs, U. The diagnosis and treatment of hypertriglyceridemia. *Dtsch. Arztebl. Int.* **116**, 825–832 (2019).
23. Slattery, D., Amiel, S. A. & Choudhary, P. Optimal prandial timing of bolus insulin in diabetes management: a review. *Diabet. Med.* **35**, 306–316 (2018).
24. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv [cs.LG]* (2014) doi:10.48550/ARXIV.1412.6980.
25. XGBoost Python Package — xgboost 2.1.3 documentation. https://xgboost.readthedocs.io/en/stable/python/.
26. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv [cs.LG]* (2017).