International Journal of Population Data Science

Journal Website: www.ijpds.org

Temporal Signals of Disease: A Transformer Approach for Predicting Diabetes from Longitudinal EHRs

Jie Pan[1], Seungwon Lee[1], Adam Virani[1], Kiarash Riazi[1], Elliot Martin[1], Alexander Leung[1], Hude Quan[1], and Na Li[1]

[1]University of Calgary, Calgary, Canada

Objective

To develop and evaluate a Transformer-based temporal representation learning framework for predicting incident diabetes using longitudinal laboratory data, and to compare its performance with state-of-the-art pretrained models and traditional machine-learning baselines.

Approach

We deterministically linked a cardiac registry cohort (2015–2019) in Alberta, Canada, and retrieved three years of laboratory tests preceding each patient's diabetes diagnosis. Each laboratory event was modelled as a triplet consisting of test type, test value, and time gap with the previous test, to preserve temporal structure. Our Transformer architecture learned contextual embeddings that capture both intra-test dynamics and inter-test interactions, indicating the onset of diabetes. Performance was benchmarked against two pretrained Transformer models (Moment, PatchTST) and an XGBoost baseline trained on aggregated test summaries (mean, standard deviation, minimum, maximum).

Results

The final cohort included 30,462 patients and 19 laboratory test types relevant to diabetes. The proposed method achieved the highest predictive accuracy (AUC = 0.92), outperforming Moment (0.80), PatchTST (0.74), and XGBoost (0.88). It also demonstrated superior sensitivity (0.70) and positive predictive value (0.777) while maintaining high specificity (0.938) and negative predictive value (0.911). Conclusion: Modelling raw laboratory trajectories with a dedicated temporal Transformer substantially improves diabetes prediction compared with both pretrained sequence models and aggregation-based baselines.

Implications

This framework provides a generalizable pathway for leveraging routine laboratory data in early disease detection. It highlights the clinical value of sequence-level modelling and supports the integration of temporal representation learning into population-level surveillance and decision-support systems.

https://doi.org/10.23889/ijpds.v11i5.3567