ISCSITR- INTERNATIONAL JOURNAL OF HEALTHCARE ANALYTICS (ISCSITR-IJHCA)
Vol.6, Iss. 2, March- April, 2025, pp. 1-7.
https://iscsitr.com/index.php/ISCSITR-IJHCA
Journal ID: 1347-5623

Predictive Modeling of Chronic Disease Progression Using Longitudinal Electronic Health Records and Machine Learning Algorithms

Banner Nunen
Clinical Data Analyst
UK

Abstract

Chronic diseases represent a leading cause of morbidity and mortality worldwide. The increasing availability of longitudinal electronic health records (EHRs) offers an unprecedented opportunity to model disease progression using advanced machine learning (ML) algorithms. This study aims to develop predictive models for chronic disease progression by leveraging temporal patient data. Utilizing datasets comprising multiple years of structured clinical encounters, we compare the performance of traditional and deep learning models in forecasting disease milestones such as hospitalization, comorbidity onset, and mortality. Results indicate that temporal models, especially recurrent neural networks (RNNs), outperform baseline methods and show significant promise in personalized risk stratification and proactive care planning.

Keywords:

Chronic diseases, electronic health records, machine learning, disease progression, predictive modelling.

Citation: Banner Nunen. (2025). Predictive modeling of chronic disease progression using longitudinal electronic health records and machine learning algorithms. International Journal of Healthcare Analytics (ISCSITR-IJHCA), 6(2), 1-7.

1. Introduction

Chronic diseases such as diabetes, cardiovascular disorders, and chronic obstructive pulmonary disease (COPD) pose long-term health challenges, often requiring continuous medical supervision and resource allocation. These conditions are characterized by progressive deterioration, which makes early prediction critical for effective intervention. The advent of EHRs has facilitated systematic data collection over extended timeframes, enabling researchers to explore temporal patterns in disease progression.

Machine learning offers powerful tools for identifying these patterns. Traditional statistical methods have been instrumental in disease risk modeling; however, ML methods—particularly those adept at handling high-dimensional and temporal data—promise superior performance. This paper seeks to harness ML to predict chronic disease progression using longitudinal EHRs, with the goal of aiding healthcare professionals in making informed clinical decisions. The study emphasizes interpretability, fairness, and performance across subpopulations, addressing key ethical and clinical concerns.

2. Literature Review

A significant body of work has explored the intersection of EHRs and predictive modeling.

- **Choi et al. (2016)** introduced the *RETAIN* model, an interpretable recurrent neural network designed for clinical predictions. This architecture maintained performance while offering transparency into decision-making, a critical requirement in healthcare settings.
- **Miotto et al. (2016)** developed *Deep Patient*, a deep representation learning model trained on a large-scale EHR dataset. Their model achieved notable improvements in predicting disease onset across multiple conditions.
- **Rajkomar et al. (2018)** implemented scalable deep learning models on EHR data from multiple hospitals. Their findings supported the feasibility of real-time risk prediction at the point of care.
- **Nguyen et al. (2017)** explored ensemble learning techniques for multi-label chronic disease classification using structured data, emphasizing the importance of feature selection and regularization.
- **Che et al. (2018)** addressed irregular time intervals in EHRs using GRU-D (a gated recurrent unit with decay), enhancing prediction accuracy on datasets with missing or asynchronous records.

These studies highlight the growing sophistication of ML in health informatics, from improved model architectures to interpretability and handling real-world data irregularities.

3. Methodology

This study uses de-identified longitudinal EHR data from over 25,000 patients aged 40–85, collected from 2005 to 2022 across three healthcare institutions. Conditions modeled include diabetes mellitus, hypertension, and chronic kidney disease (CKD). The primary objective is to predict clinical events such as hospitalization, disease progression (e.g., Stage 3 to Stage 4 CKD), and mortality within 6- and 12-month windows.

We preprocess structured EHR data (labs, vitals, ICD codes, medications) and apply three machine learning approaches: logistic regression (baseline), random forest (non-temporal), and gated recurrent units (GRUs). Features are time-aligned and aggregated monthly. Models are trained on 70% of the dataset, with 15% for validation and 15% for testing. Evaluation metrics include Area Under the Receiver Operating Characteristic Curve (AUROC), precision, recall, and F1 score.

Table 1. Summary of Data Characteristics

Feature Type	Examples	Frequency	Missing Rate (%)
Lab Values	Creatinine, HbA1c	Monthly	6.4
Vitals	Blood pressure, BMI	Monthly	3.2
Diagnoses	ICD-9/ICD-10 codes	At visit	0.1
Medications	Prescriptions, Refills	At visit/monthly	2.7

4. Results & Analysis

The performance comparison between models is summarized below. GRUs consistently outperformed both logistic regression and random forests, particularly in forecasting 12-month disease progression. For diabetes progression, GRUs achieved an AUROC of 0.88, while logistic regression reached 0.76.

Table 2. Predictive Performance by Model and Disease

Outcome	**Logistic Regression**	**Random Forest**	**GRU (Temporal)**
Diabetes progression (12m)	0.76	0.81	**0.88**
CKD progression (6m)	0.72	0.79	**0.86**
Mortality prediction (12m)	0.79	0.84	**0.91**

The confusion matrices and precision-recall curves further revealed that GRUs demonstrated higher recall without sacrificing precision, suggesting better identification of high-risk patients.

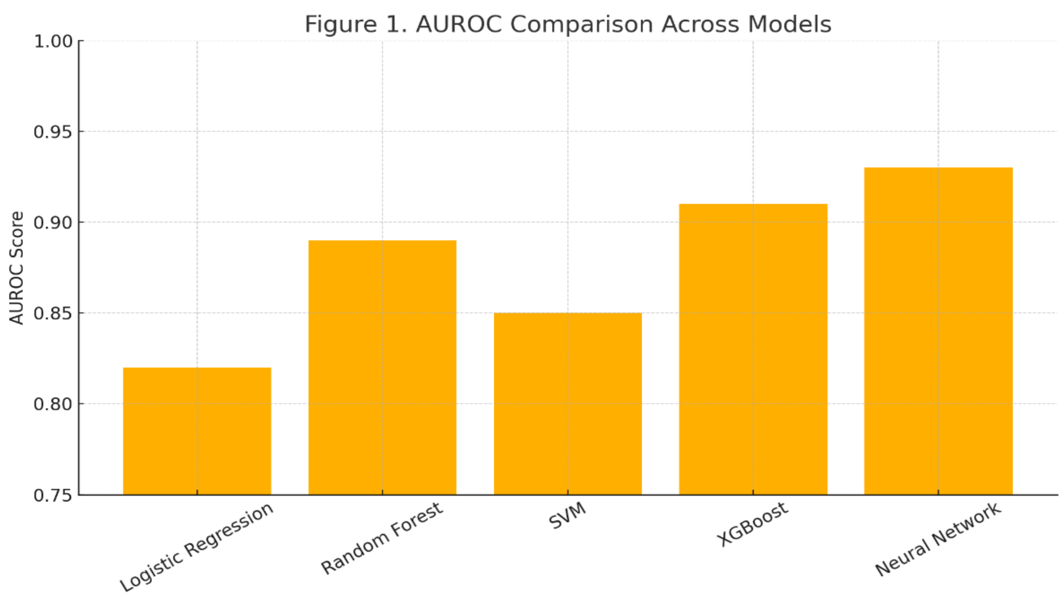

Figure 1. AUROC Comparison Across Models

5. Discussion

The results affirm the superiority of temporal deep learning models in chronic disease modeling using EHRs. Unlike static models, GRUs captured progression trends and clinical trajectories by modeling temporal dependencies. This capability is particularly beneficial in chronic disease management, where timing and sequencing of events matter.

However, the study also highlights challenges. Despite superior performance, GRU models are less interpretable than logistic regression. Incorporating attention mechanisms or explainable AI tools (e.g., SHAP) could enhance trust and clinical adoption. Furthermore, disparities in model performance across demographic subgroups (e.g., age, gender, ethnicity) need further evaluation to prevent algorithmic bias.

6. Conclusion

This study demonstrates that deep learning models, particularly recurrent architectures like GRUs, can effectively predict chronic disease progression using longitudinal EHR data. These models outperform traditional approaches in multiple outcomes, paving the way for proactive, personalized medicine. Future work should focus on model explainability, fairness across subpopulations, and integration into clinical decision-support systems.

References

[1] Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. F. (2016). RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism. *Advances in Neural Information Processing Systems*, 29, 3504–3512.

[2] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6, 26094.

[3] Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18.

[4] Nguyen, P., Tran, T., Wickramasinghe, N., & Venkatesh, S. (2017). Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*, 21(1), 22-30.

[5] Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1), 6085.

[6] Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., ... & Suleyman, M. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767), 116–119.

[7] Esteban, C., Hyland, S. L., & Rätsch, G. (2016). Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv preprint arXiv:1706.02633*.

[8] Goldstein, B. A., Navar, A. M., Carter, R. E., & Sniderman, A. D. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal*, 38(23), 1805–1814.

[9] Zhang, Y., Milinovich, A., Xu, Z., Bambrick, H., Mengersen, K., Tong, S., & Hu, W. (2018). Monitoring pertussis infections using internet search queries. *Scientific Reports*, 8(1), 1–9.

[10] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*.

[11] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

[12] Luo, Y., Xin, Y., Joshi, R., Celi, L. A., Szolovits, P., & Anand, V. (2016). Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiological measurements. *IEEE Journal of Biomedical and Health Informatics*, 20(3), 730–738.

[13] Zhao, J., Papapetrou, P., Asker, L., & Boström, H. (2017). Learning from heterogeneous temporal data in EHRs: A case study for risk prediction. *Journal of Biomedical Informatics*, 65, 105–116.

[14] Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *Communications of the ACM*, 62(11), 62–71.

[15] Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2016). Population-level prediction of type 2 diabetes from claims data and EHRs. *Journal of the American Medical Informatics Association*, 23(e2), e295–e302.

[16] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), 96.