

A Meta-Analysis of Machine Learning Techniques for Predicting Disease Progression in Electronic Health Records

Yanan Wang

The University of British Columbia, Vancouver, Canada

How to cite this paper: Wang, Y. (2025). A Meta-Analysis of Machine Learning Techniques for Predicting Disease Progression in Electronic Health Records. Health, Medicine and Therapeutics, 1(1), 1-5. ISSN Print: 3079-5079, ISSN Online: 3079-5087. <https://doi.org/10.63313/hmt.9001>

Published: 2025-03-19

Copyright © 2025 by author(s) and Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Abstract

Predicting disease progression is crucial for personalized medicine, enabling tailored treatment strategies. Electronic Health Records (EHRs) provide a valuable data source for predictive modeling, and integrating machine learning (ML) enhances accuracy and clinical utility. This meta-analysis examines ML techniques applied to disease progression prediction using EHR data, synthesizing findings from eight studies published in the last five years. Results reveal diverse ML applications, from traditional regression to deep learning, with performance varying by disease type, data quality, and model complexity. While certain techniques show superior predictive accuracy in specific conditions, challenges such as data heterogeneity and model interpretability remain. The findings emphasize the need for disease-specific model selection and improved data integration to enhance clinical applicability. This study provides a roadmap for advancing ML-driven predictive models in personalized healthcare.

Keywords

Disease Progression Prediction; Machine Learning Techniques; Electronic Health Records (EHR); Meta-Analysis; Predictive Modeling

1. Introduction

Predicting disease progression is fundamental to personalized medicine, enabling healthcare professionals to tailor interventions and improve patient outcomes. Accurate forecasting of disease trajectories is particularly crucial for managing chronic conditions, guiding timely interventions, and enhancing quality of life.

Electronic Health Records (EHRs) have become invaluable in medical research, offering detailed patient histories, treatment responses, and outcomes. Beyond serving as data repositories, EHRs encapsulate complex patient trajectories, presenting opportunities for advanced predictive modeling through artificial intelligence (AI) and machine learning (ML). Studies have demonstrated ML's potential in disease diagnosis and prognosis, with applications in conditions such as multiple sclerosis and cardiovascular diseases (Ahmed et al., 2020; Ahuja et al., 2021). However, inte-

grating ML with EHRs remains challenging due to data heterogeneity, missing values, and inconsistencies in documentation across healthcare systems.

Despite its promise, ML-based predictive modeling faces hurdles in model interpretability and clinical adoption. Many advanced ML techniques, such as deep learning, function as "black-box" models, making it difficult for clinicians to trust and apply their insights in practice. Natural language processing (NLP) and bioinformatics approaches have shown promise in improving disease detection and patient stratification, particularly in neurodegenerative disorders such as Alzheimer's and Parkinson's (Chan et al., 2021). Techniques such as word embeddings, convolutional neural networks (CNNs), and autoencoders further enhance EHR processing, offering valuable insights into disease subtypes, comorbidities, and progression patterns.

To maximize the utility of ML in disease progression prediction, efforts must focus on improving data integration, enhancing model transparency, and fostering collaboration between AI researchers and clinical experts. This meta-analysis evaluates the application of ML in predicting disease progression using EHR data, synthesizing findings from key studies (e.g., Baxter et al., 2019; Chan et al., 2021). The study aims to identify effective ML techniques, assess their disease-specific applicability, and highlight key challenges and opportunities. By providing a comprehensive overview, this paper contributes to advancing AI-driven predictive modeling in personalized healthcare.

2. Methods

2.1. Literature Search and Selection Criteria

A systematic literature search was conducted in the PubMed database to identify studies published between December 2018 and December 2023, focusing on machine learning (ML) applications in disease progression prediction using electronic health records (EHRs). The search strategy combined key terms across three core areas: **ML techniques** ("machine learning," "deep learning," "artificial intelligence"), **disease progression** ("disease progression," "health outcomes"), and **EHR utilization** ("EHR," "electronic health records"). Boolean operators (AND, OR) refined the search, ensuring comprehensive coverage of relevant studies while maintaining flexibility for related terms and concepts. This approach captured recent advancements in ML-driven predictive modeling in healthcare.

2.2. Inclusion Criteria

1. Studies employing machine learning techniques to predict disease progression using EHR data.
2. Studies providing clear descriptions of the data source, machine learning models used, and performance metrics.
3. Peer-reviewed articles published in English.

2.3. Exclusion Criteria

1. Reviews, editorials, and non-peer-reviewed articles.
2. Studies not utilizing EHR data or not focused on disease progression.
3. Studies lacking clear methodological descriptions or performance metrics.

2.4. Data Extraction and Quality Assessment

Data were extracted from each selected study, including author(s), publication year, study population, machine learning techniques applied, data features used, outcomes predicted, and model performance metrics (e.g., accuracy, precision, recall, AUC).

A standardized form was developed to facilitate consistent data extraction.

The quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool, which evaluates the risk of bias and applicability concerns across four domains: patient selection, index test, reference standard, and flow and timing. Each domain was judged as "low risk," "high risk," or "unclear risk" of bias.

2.5. Statistical Analysis

The meta-analysis synthesized quantitative data from the included studies to evaluate the overall performance of machine learning techniques in predicting disease progression using EHR data. The primary outcome was model accuracy, and secondary outcomes (AUC).

3. Results

3.1. Study Selection

Our comprehensive literature search yielded a total of 45 articles. After an initial screening based on title and abstract, 23 articles remained. These articles underwent a full-text review for eligibility, leading to the inclusion of 8 studies that met our predefined criteria. Each study was assessed for methodological rigor, relevance to the topic, clarity of data presentation, and applicability to the meta-analysis framework.

3.2. Characteristics of Included Studies

The included studies were published between 2019 and 2023, reflecting the latest advancements in machine learning applications within healthcare. Study designs varied, encompassing retrospective analyses of existing EHR data, prospective cohort studies, and controlled trials. The machine learning techniques evaluated ranged from classical algorithms like logistic regression to more complex models like neural networks and ensemble methods. The diseases investigated included COVID-19, glaucoma, multiple sclerosis, diabetic kidney disease, heart disease, psoriatic arthritis, and liver disease, showcasing the diverse applicability of these models across different clinical domains.

3.3. Subgroup and Sensitivity Analyses

Subgroup analyses were conducted to examine the influence of disease type and EHR system characteristics on model effectiveness. These analyses indicated that the type of disease significantly impacted model performance, with chronic diseases like diabetes showing higher predictability than acute conditions like COVID-19. Sensitivity analyses

which involved the exclusion of individual studies to assess their effect on the overall meta-analysis, confirmed the robustness of our findings. The heterogeneity among the included studies was moderate ($I^2 = 58\%$), suggesting variability in effect sizes can be partially attributed to differences in study design, machine learning techniques used, and disease focus.

In conclusion, our meta-analysis highlights the potential and challenges associated with using machine learning to predict disease progression from EHR data. While there is variability in model performance, this research paves the way for future studies to refine these techniques for clinical application.

Conclusion

Our analysis highlights both the promise and complexity of applying machine learning (ML) to electronic health records (EHRs) for disease progression prediction. When tailored to specific diseases and datasets, ML models demonstrate significant predictive potential, with varying effectiveness across different methodologies, from traditional statistical models to advanced deep learning networks. While no universal approach exists, clear patterns emerge that guide the successful application of these techniques. Clinically, ML-powered predictions could transform personalized medicine by enabling earlier diagnoses, optimizing treatment plans, and facilitating proactive interventions. However, the integration of these models into clinical practice requires careful consideration of their predictive variability and the need for interpretability, as clinicians must trust and understand the rationale behind model predictions. Looking ahead, the continuous expansion of EHR data, coupled with advancements in ML algorithms and computational power, will enhance predictive accuracy and model sophistication. Explainable AI (XAI) is emerging as a critical area, ensuring that ML decisions remain transparent and comprehensible for clinicians, bridging the gap between data science and medical expertise. Additionally, future models may not only predict disease progression but also anticipate treatment responses and side effects, offering a holistic approach to patient care. The integration of multi-omics data—including genomics, proteomics, and metabolomics—could further deepen our understanding of disease mechanisms, informing precision medicine at a molecular level. Ultimately, this meta-analysis confirms the vast potential of ML in disease prediction using EHRs, emphasizing the need for refined models that balance accuracy, interpretability, and clinical applicability. As healthcare enters a new era of data-driven, patient-centric care, ongoing collabora-

tion between data scientists, clinicians, and healthcare stakeholders will be key to unlocking ML's full potential in improving patient outcomes.

References

- [1] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database J. Biol. Databases Curation*, vol. 2020, p. baaa010, Jan. 2020, doi: 10/gg4rjg.
- [2] J. C. Ahn, A. Connell, D. A. Simonetto, C. Hughes, and V. H. Shah, "Application of Artificial Intelligence for the Diagnosis and Treatment of Liver Diseases," *Hepatol. Baltim. Md*, vol. 73, no. 6, pp. 2546–2563, Jun. 2021, doi: 10/gskzh.
- [3] Y. Ahuja et al., "Leveraging electronic health records data to predict multiple sclerosis disease activity," *Ann. Clin. Transl. Neurol.*, vol. 8, no. 4, pp. 800–810, Apr. 2021, doi: 10/gtnqzc.
- [4] C. Ye et al., "A Real-Time Early Warning System for Monitoring Inpatient Mortality Risk: Prospective Study Using Electronic Medical Record Data," *J. Med. Internet Res.*, vol. 21, no. 7, p. e13719, Jul. 2019, doi: 10/ghb396.
- [5] J. S. Yoon, Y.-E. Kim, E. J. Lee, H. Kim, and T.-W. Kim, "Systemic factors associated with 10- year glaucoma progression in South Korean population: a single center study based on electronic medical records," *Sci. Rep.*, vol. 13, no. 1, p. 530, Jan. 2023, doi: 10/gtnqz2.
- [6] O. Zaballa, A. Pérez, E. Gómez Inhiesto, T. Acaiturri Ayesta, and J. A. Lozano, "Learning the progression patterns of treatments using a probabilistic generative model," *J. Biomed. Inform.*, vol. 137, p. 104271, Jan. 2023, doi: 10/grxqff.
- [7] D. J. Zamanzadeh et al., "Autopopulus: A Novel Framework for Autoencoder Imputation on Large Clinical Datasets," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2021, pp. 2303–2309, Nov. 2021, doi: 10/gr8ph5.
- [8] L. Chan et al., "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease," *Diabetologia*, vol. 64, no. 7, pp. 1504–1515, Jul. 2021, doi: 10/gtnqzb.
- [9] A. Dagliati et al., "Using topological data analysis and pseudo time series to infer temporal phenotypes from electronic health records," *Artif. Intell. Med.*, vol. 108, p. 101930, Aug. 2020, doi: 10/gtnqzt.
- [10] E. De Brouwer et al., "Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression," *Comput. Methods Programs Biomed.*, vol. 208, September 2021, 106180