Predicting Neurological Disease Risk Using Machine Learning Models: Evidence from Electronic Health Records Across the United States

Nicolla Eddy Wopara[1]

[1]Technology Management, University of California, Santa Barbara

Publication Date: 2026/09/19

Abstract: Electronic health records (EHRs) were primarily designed to document care, rather than forecast disease. Within the past five years, research experts have queried whether machine learning can be used to mine the same records, including lab values, years of diagnosis codes, clinical notes, and medication orders, to identify who is going to have a stroke, Alzheimer's disease, or other neurological conditions before the symptoms are obvious. This review scans recent evidence on the question and synthesises multiple studies to identify the gap. In analyses of several dozen United States studies, machine learning models built on routine hospital and health-system data have predicted the risk of stroke with accuracy (area under the curve, or AUC) of 0.90–0.99, consistently outperforming tools such as the CHA2DS2-VASc score. The onset of Alzheimer's disease and related dementias 1-5 years in advance is predicted with a gradient-boosted model trained on de-identified HER data, with AUCs between 0.809 and 0.833, while surfacing sleep apnea and headache as previously underappreciated risk signals. A separate model built on over 52,000 Cleveland Clinic records indicated that ordinary clinical data comprising patient-reported outcomes, sleep assessments, fall histories, and weight trends are characterised by easily detectable signs of impending neurological disease. Similarly, an increasing body of fairness research shows that these models can demonstrate worse performance, or different behaviour, especially for Black and other minority patients. The reason for this is that the training data in itself is a reflection of unequal access to care. The study, therefore, concludes that machine learning has the potential to predict neurological disease with EHR data, which are already sitting in hospital systems. Meanwhile, that power will only translate into equitable benefit if bias auditing is routinely carried out as much as accuracy testing.

Keywords: *Machine Learning, Neurological Disease Risk, Electronic Health Records, United States.*

How to Cite: Nicolla Eddy Wopara (2026) Predicting Neurological Disease Risk Using Machine Learning Models: Evidence from Electronic Health Records Across the United States. *International Journal of Innovative Science and Research Technology*,11(9), 527-533. https://doi.org/10.38124/ijisrt/26sep309

I. INTRODUCTION

A piece of data is added to the patient's electronic health record every time they visit a doctor, fill a prescription, or get a blood test (Ehrenstein et al., 2019). With many millions of Americans visiting the clinic and a record of decades of visits, this result is clearly one of the richest and largest datasets ever assembled regarding human health (Batarseh & Latif, 2016; Tanner, 2017; Tang et al., 2024). While this is not deliberately designed for research, it is increasingly being used for research all the same.

Naturally, neurological disease is targeted for this type of data mining (Chase et al., 2017; Kaswan et al., 2021). Currently, however, none of these conditions is curable, although many of them respond to treatment the earlier they are spotted [*See Figure 1*]. In the United States, the burden of neurological disease is vast and increasing, with incident cases of Alzheimer's disease (AD), Parkinson's disease (PD), other dementias, multiple sclerosis, and motor neuron diseases increasing progressively from 1990 to 2017 (GBD 2017 US Neurological Disorders, 2021 as cited in Felix et al., 2024).

Consequently, this creates an opportunity that if the record of an individual's medical care already has some early signals of a necessary diagnosis, a robust algorithm might interpret these signals quite earlier than a clinician (Haick & Tang, 2021).

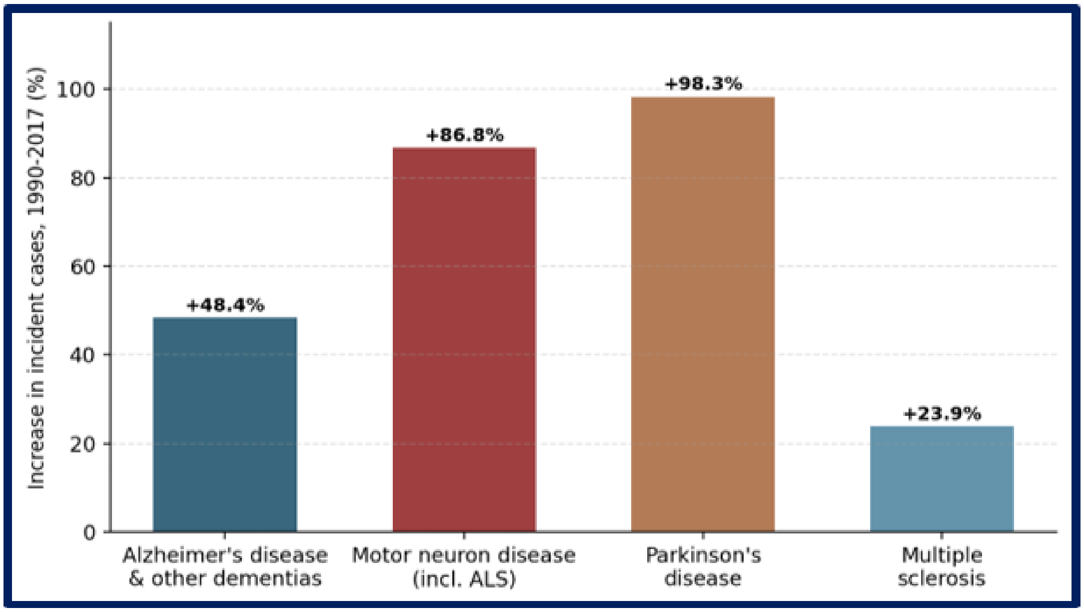

Fig 1: Increase in U.S. Incident Cases of Major Neurodegenerative Conditions [Adapted from Digital Health, 2024 – GBC-Derived U.S. Incident Trends (1990-2017)]

This review focuses on three query areas: how well machine learning (ML) models predict neurological disease risk when trained on real-world HER data, contrary to curated research cohorts; the kinds of data that are most significant; and if the models work equally well for everyone [inclusive], or whether they manifest the disparities documented in other locations in neurological-health literature.

It is, however, of significant interest to precisely understand the meaning of prediction in this context. This is because no single study reviewed in this paper claims to diagnose a neurological condition solely from data, while none is an actual replacement for a clinical evaluation. Essentially, the focus is on estimating a probability (a risk score) that a patient will receive a diagnosis within a specified period based on patterns learned from several other patients with known diagnoses (Newman & Kohn, 2020). This distinction is important for the use of these tools, as a high-risk score from an electronic health record (EHR) indicates the need to look closer (Ashburner et al., 2025; Akter et al., 2025).

II. DATA: SOURCES AND METHODS

This paper takes a narrative approach, drawing on recently published and peer-reviewed studies that have built and validated machine learning models with the United States hospital or health-system HER data for neurological risk prediction. This also includes the algorithmic-bias and fairness literature that pertain to their application, according to Heseltine-Carp et al. (2025). Thus, instead of re-running a single model, the reported performance, methods, and prediction values are compared across the studies to identify patterns common to all.

The aim of this study is underscored by three major types of studies:

- Disease-conversion prediction studies: These studies use years of pre-diagnosis records to predict the chances that a patient will later be diagnosed with a certain neurological condition (Heseltine-Carp et al., 2025). These studies include a 2024 Cleveland Clinic study using XGBoost models on approximately 50,000 records to predict conversion to Alzheimer's disease, multiple sclerosis, ALS, or Parkinson's disease (Felix et al., 2024), and a 2025 study predicting Alzheimer's disease as well as other related dementias one to five years before they are diagnosed (Li et al., 2023).
- Acute-event prediction studies: These studies comprise real-time or near-real-time HER data including notes, vital signs, and prior diagnoses, to identify patients at high risk (short-term) of a stroke or related event. A 13-hospital Pennsylvania study covering 56,452 emergency department encounters falls under this type (Abedi et al., 2024).
- Systematic reviews and meta-analyses: These studies pool results across several individual machine learning stroke-prediction studies to establish a logical and realistic range of model performance. Asadi et al. (2024) investigation of nearly 50 studies within 2013-2024 falls under this category.

Similarly, this review draws on parallel and vital literature examining algorithmic fairness to examine whether these models perform consistently across ethnic, racial, and socioeconomic groups. Some academic health systems and Mass General Brigham studies are drawn for the purpose (Yang et al., 2024; Huang et al., 2022).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep309

III. RESULTS

➢ *Predicting Stroke Risk*

Stroke is the most studied neurological application of machine learning (ML) on electronic health records (EHR) data. On the whole, the results are encouraging. A 2024 systematic review of 49 studies (2013 – 2024) found that machine learning models predicted stroke risk with AUCs between 0.64 and 0.99 across the atrial-fibrillation-based and general populations. Where the two were head-to-head tested, ML models consistently outperformed the conventional CHA2DS2-VASc risk score commonly used in clinics today across the United States (Heseltine-Carp et al., 2025).

In addition, the choice of model is important. Across these studies, tree-based ensemble methods, especially XGBoost and Random Forest, have shown a tendency to have superior performance to the simpler logistic regression models and deep neural networks. This review identified Random Forest as the top-performing algorithm in approximately 25% of all stroke ML studies (2019 – 2023) (Asadi et al., 2024) [*See Figure 2*]. A 13-hospital Pennsylvania study combining structured HER data and natural-language processing of physician notes achieved AUROCs (0.88 – 0.92) from the information available before the stroke happened, representing the more clinically meaningful test. According to Abedi et al. (2024), a model that performs excellently only when fed data extracted after the diagnosis is not exactly making any predictions.

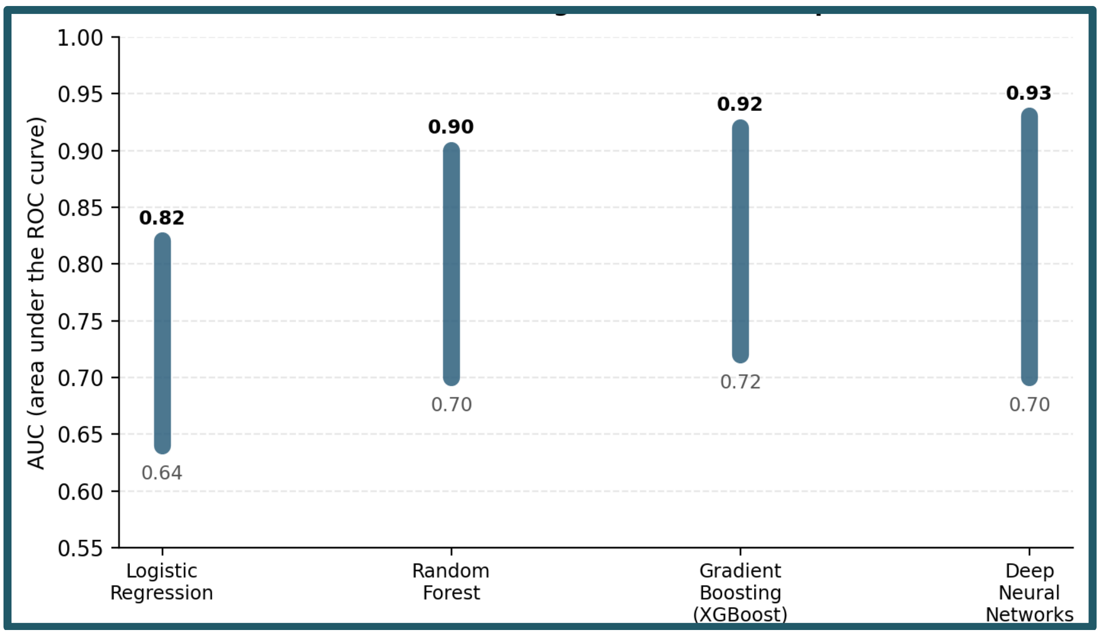

Fig 2: Reported Range of Model Performance in Stroke Risk Prediction Studies [Using EHR/Routine Hospital Data] [Adapted from Asadi et al., 2024; Systematic Review PubMed 2025; Health Science Reports, 2024]

In Figure 2 above, bars show typical reported AUC range per algorithm group across reviewed studies [2013 – 2024]. AUC of 1.0 = perfect prediction; 0.5 = no better than chance

The study is a noteworthy reminder of the clinical effectiveness of this approach – where the researchers identify that stroke misdiagnosis occurs in approximately 9% of all stroke patients, while their models were purposefully built to aid emergency department clinicians catch cases they would have missed ordinarily (Abedi et al., 2024).

➢ *Predicting Alzheimer’s Disease and Related Dementias*

Dementia prediction is largely a harder problem than stroke prediction. This is because Alzheimer's and other dementias develop silently over years and do not announce themselves in a single acute event (Akter et al., 2025). This was addressed by a 2025 study, where the researchers trained a gradient-boosted model on de-identified EHR data to predict Alzheimer’s disease and other related dementias (ADRD) at five lead times of one year apart in five years before diagnosis. The model achieved AUC-ROC scores (0.809 – 0.833) across the periods, and the accuracy declined slightly with the lengthening of the prediction horizon (Akter et al., 2025) [*See Figure 3*].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep309

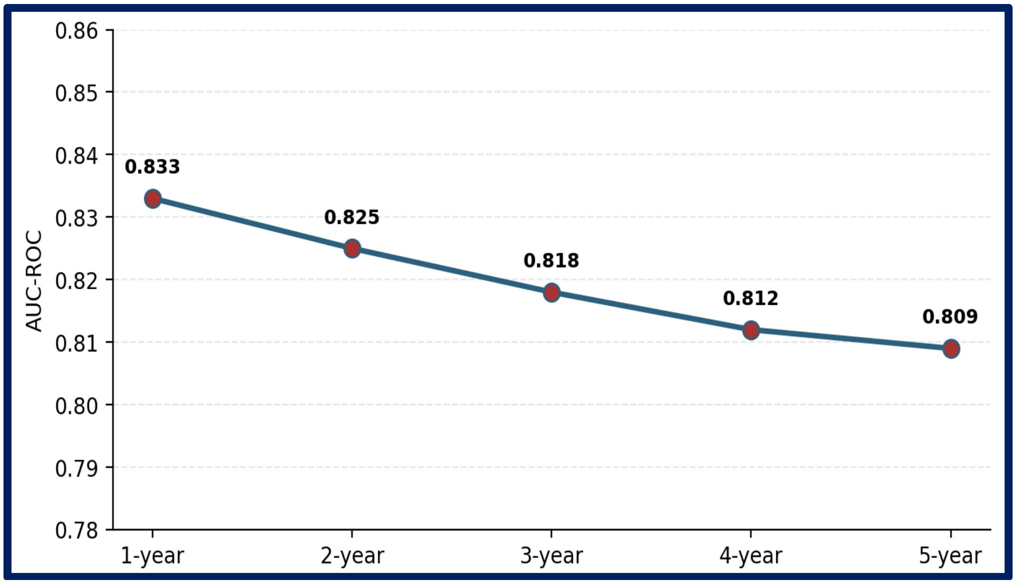

Fig 3: Model Accuracy for Predicting Alzheimer's Disease and Other Related Dementias by Years Prior to Clinical Diagnosis

Figure 3 above demonstrates the Gradient-Boosted model performance predicting ADRD from EHR data, with a prediction window before diagnosis [x-axis].

The model focused on an interesting area beyond the accuracy figure. With SHAP (Shapley Additive Explanations), which quantifies how each factor contributed to an individual prediction, the researchers found age, depression, sleep apnea, heart disease, and headache as the strongest predictors (Akter et al., 2025). Specifically, headache and sleep apnea were flagged as 'novel' risk factors, not emphasised in traditional dementia risk assessments (Kitamura et al., 2020). In a different Cleveland Clinic study which spans four different neurological diseases namely ALS, Alzheimer's, Parkinson's, and multiple sclerosis, a similar conclusion was reached from a different perspective, that patient-reported outcomes, fall histories, sleep assessments, longitudinal weight change, and additional diagnoses were important predictors, which a clinic already collects routinely, but rarely analyses together (Felix et al., 2024).

➢ *Fairness and Bias*

Fairness and bias imply the same data enabling prediction that can encode inequity and inclusivity. A body of research complicates the optimistic pictures demonstrated above. Since machine learning models learn patterns from historical data while historical healthcare data reflects historical inequities, many studies have discovered that clinical prediction models can demonstrate inconsistent performance across racial groups regardless of the strength of the overall accuracy (Ilapaka & Ghosh, 2026). In the review of a widely cited 2019 analysis, commercial algorithm for managing population health assigned Black patients the same risk score as White patients despite the Black patients being sicker; because the algorithm used for healthcare cost as a substitute for healthcare need, and Black patients in the training data had incurred lower costs for reasons linked to unequal access not just overall better health (Obermeyer et al., 2019; Siddique et al., 2024).

This finding is not just a one-off discovery: A 2024 study using Mass General Brigham EHR data built separate one-year mortality prediction models for five chronic conditions, including dementia, and a matched-counterpart method to distinguish racial disparities in model performance from the differences in patient populations (Yang et al., 2024). Similarly, a broader 2022 scoping review of 12 clinical machine learning fairness studies found that two-thirds of the studies (8 of them) concluded the presence of racial bias in the model examined. However, all the studies that attempted bias-mitigation strategies succeeded in improving fairness by the metric selected, which suggests that the problem can be addressed once measured (Huang et al., 2022).

Another dimension of this concern is seen in data completeness. Research on EHR data continuity found that patients with less continuous contact in a single health system, that is, those who are disproportionately from minority and disadvantaged backgrounds, generate noisier and sparser records, which can degrade model performance for the exact screening tool a patient needs to help (ARXIV). In this same review, a striking example is seen from dementia and outside stroke but illustrates the general problem – that video-based tools for detecting neurological signs, for example, abnormal blink rate, useful for Tourette syndrome and Parkinson's disease, have less reliable performance for Asian patients (Darkis & Cerna, 2025). This is a reminder that bias can enter through the measurement layer or sensor, rather than the statistical model (Turaev et al., 2023).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep309

➢ *Predictor Variables*

A pattern emerges across the dementia and stroke literatures, in which variables eventually become very important. Easily extractible structured data from any EHR, such as blood pressure readings, medication lists, age, and prior diagnoses, consistently form the backbone of the models, which is part of why they can be built without the need for new infrastructure. However, the strongest-performing studies are likely to add an additional layer beyond the basics: either unstructured clinical notes processed with natural language processing (NLP) as in the Pennsylvania stroke study (Abedi et al., 2024), or patient-reported-longitudinal information such as sleep complaints, fall histories, and weight trends, as in the Cleveland Clinic discoveries (Felix et al., 2024). This suggests the prediction accuracy ceiling is not particularly set by the algorithm choice, rather by how much of a full clinical picture of a patient the model is allowed to access.

IV. DISCUSSION

Collectively, the evidence supports a particular and confident claim – that electronic health records already contain a usable signal for neurological disease risk, and ML models are adequate to discover it. AUCs between 0.80 and 0.90 for stroke prediction, and low-to-mid range of 0.80s for multi-year-ahead dementia prediction, are strong by the standards of a clinical prediction task, and beat the traditional risk scores currently used in practice in head-to-head comparisons (Asadi et al., 2024; Heseltine-Carp et al., 2025).

Meanwhile, the argument regarding the inherent fairness of the predictive power, or whether it has to be made fair, persists. Based on this literature, the answer is the latter: bias does not penetrate these models via some dramatic flaw, but through the ordinary, cumulative fact that the training data was a result of a healthcare system where documentation quality, access, and diagnosis rates already differ by geography and race (Wang et al., 2024). These same disparities are more broadly documented in Alzheimer's and stroke outcomes. Therefore, by default, a model trained on the data will reproduce similar patterns unless deliberate checks and efforts to correct them are put in place. Since very few bias-mitigation efforts reviewed in this study succeeded once researchers looked for the problem, in a sense, this appears like a solvable governance and engineering problem, rather than an unavoidable cost of using AI and machine learning in medicine (Huang et al., 2022).

Meanwhile, a translation gap also exists. The majority of the high-performing models described in this study were developed and validated at a jurisdiction-based hospital network or single health system, and tested on data from the same institutions. While this is a reasonable first step, it is different from showing the performance of a model when deployed at a new hospital with a different patient mix, EHR vendor, and documentation habits. In this review, very few studies provided a genuine, multi-region validation, while fairness researchers have flagged the potential for model performance to shift when the continuity of care of such a patient with a system differs (Huang et al., 2023). Therefore, until external validation across the United States health systems becomes standard practice, single-site AUCs cannot be read or interpreted as a guarantee of real-world performance, but a proof of concept (Suleman et al., 2025; Ara & Abdullah, 2026).

For clinicians, there is a narrower and more actionable near-term opportunity than the logic that AI will predict neurological disease. Some of these tools, such as a stroke-flagging model in the electronic record or a dementia-risk calculator running quietly in the background of a primary care visit, are effective as a decision support today. Meanwhile, they need bias checks built in, including a clear plan for what happens when a patient gets flagged as high-risk (Labkoff et al., 2024). For funders and health systems, the SHAP-based finding that headache and sleep apnea are underappreciated dementia risk signals is a testable and concrete hypothesis that could be incorporated by primary-care screening protocols before the formal deployment of an algorithm (Awais, 2026).

V. LIMITATIONS

Rather than present a new, unified analysis, this review synthesizes published findings to ensure that differences in cohort definitions, outcome labelling, and prediction periods across the underlying studies limit exact cross-study comparison. Reported AUCs are derived from studies of varying geography, size, and validation depth; as the highest figures cited here are from a broad systematic review that spans the U.S. and international studies, and thus should be treated as the upper ceiling of a real range, rather than a typical result. To date, most fairness studies, including the disparities literature, focus primarily on Black-White comparisons because those groups have had the largest sample sizes in the U.S. health system data for many decades, while the bias affecting Asian American, Hispanic, Native American, and other populations is relatively understudied. Lastly, by construction, all models reviewed in the studies were essentially developed using data from patients with consistent contact with a health system, which excludes from the training data and eventual benefit the uninsured populations and others that avoid care.

VI. CONCLUSION

Machine learning models trained on ordinary electronic health record data can predict neurological diseases including stroke, Alzheimer's disease, and other related diseases risks before they become clinically obvious. This often outperforms the risk tools currently used in practice in the U.S. Contrary to popular beliefs, the capability is no longer speculative, but has been demonstrated across dozens of U.S. studies using tens of thousands of real patient records. However, an unresolved issue remains, not pertaining to the effectiveness of the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep309

prediction, but whether they work for everyone (inclusive and equitable). The same historical inequities documented in dementia and stroke outcomes are detectable in the training data that the models learn from, and where deliberate bias auditing is lacking, a predictive tool can encode the disparities it might instead help close. In the future, the goal is to continue to make fairness testing a routine as much as accuracy testing before the models are translated from research paper to applications in the clinic.

REFERENCES

[1]. Abedi, V., Misra, D., Chaudhary, D., Avula, V., Schirmer, C. M., Li, J., & Zand, R. (2024). Machine learning-based prediction of stroke in emergency departments. Therapeutic Advances in Neurological Disorders.

[2]. Akter, S., Liu, Z., Simoes, E. J., & Rao, P. (2025). Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's Disease and Related Dementias. The journal of prevention of Alzheimer's disease, 12(7), 100169.

[3]. Ara, J., & Abdullah, S. B. (2026). AI-Driven Diagnostic Modelling Frameworks for Enhancing Accuracy and Privacy Protection in US Healthcare Analytics Systems. American Journal of Scholarly Research and Innovation, 5(01), 109-155.

[4]. Asadi, et al. (2024). The most efficient machine learning algorithms in stroke prediction: A systematic review. Health Science Reports, Wiley.

[5]. Ashburner, J. M., Chang, Y., Porneala, B., Singh, S. D., Yechoor, N., Rosand, J. M., ... & Atlas, S. J. (2024). Predicting post-stroke cognitive impairment using electronic health record data. International Journal of Stroke, 19(8), 898-906.

[6]. Awais, M. (2026). A Multimodal Feature Fusion and Explainable AI-Driven Framework for Early Detection and Staging of Dementia (Doctoral dissertation, CAPITAL UNIVERSITY).

[7]. Batarseh, F. A., & Latif, E. A. (2016). Assessing the quality of service using big data analytics: with application to healthcare. Big Data Research, 4, 13-24.

[8]. Chase, H. S., Mitrani, L. R., Lu, G. G., & Fulgieri, D. J. (2017). Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC medical informatics and decision making, 17(1), 24.

[9]. Darkis, A. T., & Cerna, P. D. (2025, December). Performance Analysis of Machine Learning Model of Speech and Tic Recognition for Early Diagnosis of Tourette Syndrome. In 2025 6th International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS) (pp. 1736-1740). IEEE.

[10]. Ehrenstein, V., Kharrazi, H., Lehmann, H., & Taylor, C. O. (2019). Obtaining data from electronic health records. In Tools and technologies for registry interoperability, registries for evaluating patient outcomes: A user's guide, 3rd edition, Addendum 2 [Internet]. Agency for Healthcare Research and Quality (US).

[11]. Felix, C., Johnston, J. D., Owen, K., Shirima, E., Hinds, S. R., Mandl, K. D., Milinovich, A., & Alberts, J. L. (2024). Explainable machine learning for predicting conversion to neurological disease: Results from 52,939 medical records. DIGITAL HEALTH.

[12]. Haick, H., & Tang, N. (2021). Artificial intelligence in medical sensors for clinical decisions. ACS nano, 15(3), 3557-3567.

[13]. Heseltine-Carp, W., Courtman, M., Browning, D., Kasabe, A., Allen, M., Streeter, A., ... & Mullin, S. (2025). Machine learning to predict stroke risk from routine hospital data: A systematic review. International Journal of Medical Informatics, 196, 105811.

[14]. Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. JMIR Medical Informatics, 10(5), e36388.

[15]. Huang, Y., Guo, J., Chen, Z., Xu, J., Donahoo, W. T., Carasquillo, O., ... & Shenkman, E. A. (2023). The impact of electronic health records (EHR) data continuity on prediction model fairness and racial-ethnic disparities. arXiv preprint arXiv:2309.01935.

[16]. Ilapaka, A., & Ghosh, R. (2026). An Agentic AI Framework for Explainable Health Insurance Decision Support. In 2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0224-0233). IEEE.

[17]. Kaswan, K. S., Gaur, L., Dhatterwal, J. S., & Kumar, R. (2021). AI-based natural language processing for the generation of meaningful information electronic health record (EHR) data. In Advanced AI techniques and applications in bioinformatics (pp. 41-86). CRC Press.

[18]. Kitamura, T., Miyazaki, S., Sulaiman, H. B., Akaike, R., Ito, Y., & Suzuki, H. (2020). Insomnia and obstructive sleep apnea as potential triggers of dementia: is personalized prediction and prevention of the pathological cascade applicable? EPMA Journal, 11(3), 355-365.

[19]. Labkoff, S., Oladimeji, B., Kannry, J., Solomonides, A., Leftwich, R., Koski, E., ... & Quintana, Y. (2024). Toward a responsible future: recommendations for AI-enabled clinical decision support. Journal of the American Medical Informatics Association, 31(11), 2730-2739.

[20]. Li, Q., Yang, X., Xu, J., Guo, Y., He, X., Hu, H., & Bian, J. (2023). Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimer's & Dementia, 19(8), 3506-3518.

[21]. Newman, T. B., & Kohn, M. A. (2020). Evidence-based diagnosis: an introduction to clinical epidemiology. Cambridge University Press.

[22]. 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.
[23]. Siddique, S. M., Tipton, K., Leas, B., Jepson, C., Aysola, J., Cohen, J. B., ... & Mull, N. K. (2024). The impact of health care algorithms on racial and ethnic disparities: a systematic review. Annals of Internal Medicine, 177(4), 484-496.
[24]. Suleman, M. U., Mursaleen, M., Khalil, U., Saboor, A., Bilal, M., Khan, S. A., ... & Tahir, M. (2025). Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. Annals of Medicine and Surgery, 87(12), 8803.
[25]. Tang, A. S., Woldemariam, S. R., Miramontes, S., Norgeot, B., Oskotsky, T. T., & Sirota, M. (2024). Harnessing EHR data for health research. Nature Medicine, 30(7), 1847-1855.
[26]. Tanner, A. (2017). Our bodies, our data: how companies make billions selling our medical records. Beacon Press.
[27]. Turaev, S., Al-Dabet, S., Babu, A., Rustamov, Z., Rustamov, J., Zaki, N., ... & Loo, C. K. (2023). Review and analysis of patients' body language from an artificial intelligence perspective. IEEE Access, 11, 62140-62173.
[28]. Wang, Y., Wang, L., Zhou, Z., Laurentiev, J., Lakin, J. R., Zhou, L., & Hong, P. (2024). Assessing fairness in machine learning models: A study of racial bias using matched counterparts in mortality prediction for patients with chronic diseases. Journal of Biomedical Informatics, 156, 104677.