

Comments on Contemporary Uses of Machine Learning for Electronic Health Records

Jordan Bryan, Didong Li

Various decisions concerning the management, display, and diagnostic use of electronic health records (EHR) data can be automated using machine learning (ML). We describe how ML is currently applied to EHR data and how it may be applied in the near future. Both benefits and shortcomings of ML are considered.

Introduction

The use of machine learning (ML) for the analysis of electronic health record (EHR) data has become more frequent over the course of the last two decades. Today, two complementary forces all but ensure that this trend will continue into the near future: EHRs are growing in size and complexity as more patients interact with health care systems and as a broader array of screening technologies are deployed in clinical practice, while at the same time public and private investment is spurring the development of new ML methods that are capable of handling ever more diverse data types. As ML becomes more integrated into decision-making using EHR data, it is important that practitioners and policymakers understand how ML is being applied to EHRs, and how the use of ML may both improve and complicate health care.

ML—sometimes used synonymously with artificial intelligence (AI)—is an evocative term, which may mean different things to different audiences. For the purposes of this commentary, we define ML as a class of methods for deriving decision rules by using a combination of data, mathematical or statistical principles, and computer software. Health care decisions based on ML have the potential to be faster, more precise, less expensive, and less biased than those attainable purely through clinical judgement or case review. Hence, in the last two decades there has been intense interest in applying ML to the analysis of EHR data, where sample sizes tend to be large and where decision rules may inform public health assessments, clinical trial design, optimal treatment regimes, or clinical decision support (CDS).

In many respects, current uses of ML for EHR data represent innovations on themes that emerged in the 1990s as EHRs became more widely adopted by hospital systems. Even at that time, automating aspects of record-keeping

and CDS were seen as key potential benefits of moving from paper records to EHRs [1]. Now, various ML classification and regression models—which take in the demographic information, test results, and biometric or genomic measurements in structured EHR data—are supplementing or replacing clinical decision rules based on common-knowledge health care guidelines. These models may be integrated directly into EHR systems to identify patient subpopulations of interest and to assign diagnostic labels to patients with any number of rare or common disorders in real time [2–7]. The paradigm of reinforcement learning (RL), which constitutes a subset of ML, is also being used to inform health care decisions that must be made sequentially in response to a course of patient outcomes. For instance, RL has been applied to design treatment regimens for patients in intensive care [8] and for those enrolled in clinical trials [9]. In addition to patient-centric uses, physician-centric uses of ML have recently been studied and piloted in clinics. Most commonly, in order to reduce the stress associated with interacting with EHRs, experimental software platforms have been developed to use ML to prioritize only the most relevant information for display on clinicians' EHR interface [10, 11]. ML is also increasingly being used to shape decisions for hospital resource management, such as scheduling hospital admissions from the emergency department (ED) or anticipating 30-day readmission to the hospital [12, 13].

Emerging Trends

While much of the activity in ML research for EHR data is a continuation of what came before, there are at least two factors that distinguish the current state of affairs from that of the previous decade. First, ML is transitioning from being a novelty to being a commonplace tool, particularly for large health care systems with mature EHRs. This means that ML has been around long enough to develop a track record in clinical practice, and that track record has been mixed.

Electronically published July 10, 2024.

Address correspondence to Didong Li, 3107C McGavran-Greenberg Hall, Chapel Hill, NC 27599 (didongli@unc.edu).

N C Med J. 2024;85(4):263-265. ©2024 by the North Carolina Institute of Medicine and The Duke Endowment. All rights reserved. 0029-2559/2024/85413

Certain studies have revealed sources of bias, imprecision, and even increased time-load on physicians interacting with ML recommendation systems that have been deployed in EHR environments [14, 15]. As applications of ML to EHR data become less hypothetical, current and future research efforts in this area will likewise need to become more practical, focusing on the challenges that inevitably arise as ML is integrated into real-time CDS [10, 16–18].

A second defining characteristic of the current era of ML for EHR data is a heavy emphasis on the analysis of so-called unstructured EHR data, which are comprised of physician notes or other text entered by medical scribes. While ML methods for unstructured EHR data are by no means new, the effectiveness of these methods has recently seen an apparent increase due to the construction of large language models (LLMs), which derive decision rules for language generation using internet-scale sources of text data [19]. Refinement of general-purpose LLMs using biomedical text databases has given rise to several open-source LLMs designed for biomedical language synthesis [20–23]. Even more recent work has yielded LLMs specifically designed for question-answer and instruction-response style interaction with EHR data [24]. The remarkable flexibility of LLMs in terms of their ability to process free-text instructions makes them a potentially valuable tool for clinicians who spend hours navigating EHR systems to generate documentation or to search for relevant medical notes, hours which might otherwise be devoted to patient care [25]. However, while other forms of ML have reached the implementation stage in clinical settings, evaluating LLMs in realistic EHR environments remains an open challenge.

State-of-the-art ML models have grown so large in recent years that the amount of data and computing power required to produce effective decision rules from them exceeds the resources of many health care systems. Hence, the next several years are likely to see an increased demand for the use of EHR data by third parties who administer EHR systems and an increased demand for access to EHR data from third parties who have extensive computing resources. The supply of ML tools will also increase, as companies that previously only administered EHR systems begin to develop proprietary ML models, and companies that previously focused on ML for other applications turn their attention to EHR data. To combat the influx of ad-hoc ML tools deployed at the department level, large hospitals with the requisite resources will increasingly turn to “command center” models of hospital administration, which leverage a centralized set of predictive ML tools to monitor and coordinate patient care using live updates of the EHR.

If recent history is any guide, the frenetic activity around ML will create a sense that progress in health care delivery is both rapid and inevitable, but this will not match what is observed in practice. For instance, in a retrospective population-based study of patients who visited the Bradford Royal

Infirmity Hospital in the UK, the use of a hospital command center equipped with ML-powered coordination software was not found to have a positive impact on patient flow or data quality [26]. A widely used ML model developed by Epic Systems Corporation (ESC) for predicting the onset of sepsis was found to have significantly poorer discrimination and calibration than had been previously reported by ESC when validated by researchers at Michigan Medicine, leading to substantial alert fatigue among clinicians using the model to inform their treatment decisions [27]. Even if properly calibrated to naturally occurring EHR data, ML tools can output faulty decisions if their input data have been subtly altered, leaving them vulnerable to cyber attack [28]. Even without explicit malicious intent, corporate agents operating in the health care domain can cause security issues when handling EHR data. For example, in 2019, an aggressive acquisition of EHR data from 50 million Ascension customers by Google gave its employees access to non-anonymous health data, raising concerns that patient confidentiality had been sacrificed for the purpose of creating a proprietary ML model [29].

Cautionary tales like these cast doubt on the prospect that ML can be a panacea for the ills of the health care system, yet they should not necessarily discourage practitioners from considering ML solutions to their problems. Many applications of ML to EHR data do promise a better quality of life, both for the patients who visit hospitals and for the health care professionals who work in them. However, making good on the promise of ML for EHR data will require clearheaded thinking from and coordination between the clinicians, scientists, and policymakers who use, design, and regulate ML tools. As several studies have demonstrated, the performance of ML methods as measured by retrospective EHR analyses tends to exceed that observed in clinical practice [15, 27, 30]. Therefore, when conducting a cost-benefit analysis for the adoption of any ML tool, it is important to keep in mind that the actual performance of the tool may not match its reported performance. Designing prospective evaluation strategies that mimic realistic EHR deployment environments can be a crucial first step toward obtaining realistic estimates of the near-term and long-term performance of ML methods [13, 17]. Defining standards for the coding of medical terminology and the storage of medical data has been a key challenge for the development of EHRs over the last 30 years [1]. Looking forward, defining standards and best practices for ML as applied to EHR data and encouraging their widespread adoption poses new challenges for stakeholders in health care systems, which will ultimately determine whether ML improves or merely complicates how EHRs are used to provide care. NCMJ

Jordan Bryan, PhD Postdoctoral Researcher, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.
Didong Li, PhD Assistant Professor, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Acknowledgments

The authors declare that they have no known conflicts of interest related to the writing of this article or any products or institutions mentioned.

References

1. Evans RS. Electronic health records: then, now, and in the future. *Yearb Med Inform.* 2016;Suppl 1:S48–S61. doi: 10.15265/YIS-2016-s006
2. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc.* 2014;21(2):221–230. URL <https://academic.oup.com/jamia/article-lookup/doi/10.1136/amia-jnl-2013-001935>
3. Jacoba CMP, Celi LA, Silva PS. Biomarkers for progression in diabetic retinopathy: expanding personalized medicine through integration of AI with electronic health records. *Semin Ophthalmol.* 2021;36:250–257. <https://www.tandfonline.com/doi/full/10.1080/08820538.2021.1893351>
4. Robles Ar'evalo A, Maley JH, Baker L, et al. Data-driven curation process for describing the blood glucose management in the intensive care unit. *Sci Data.* 2021;8:80. <https://www.nature.com/articles/s41597-021-00864-4>
5. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. *Ann Rheum Dis.* 2023;82:306–311. <https://ard.bmj.com/lookup/doi/10.1136/ard-2022-222626>
6. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: a methodological review. *J Am Med Inform Assoc.* 2023;30:367–381. <https://academic.oup.com/jamia/article/30/2/367/6839857>
7. Moynihan D, Monaco S, Wah Ting T, et al. Cluster analysis and visualisation of electronic health records data to identify undiagnosed patients with rare genetic diseases. *Sci Rep.* 2024;14:5056. doi: 10.1038/s41598-024-55424-8
8. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med.* 2018;24:1716–1720. <https://www.nature.com/articles/s41591-018-0213-5>
9. Shah KS, Fu H, Kosorok MR. Stabilized direct learning for efficient estimation of individualized treatment rules. *Biometrics.* 2023;79(4):2843–2856. doi: 10.1111/biom.13818
10. Murray L, Gopinath D, Agrawal M, Hong S, Sontag D, Karger DR. MedKnowts: Unified Documentation and Information Retrieval for Electronic Health Records. Presented at: The 34th Annual ACM Symposium on User Interface Software and Technology. ACM, Virtual Event USA, 2021. <https://dl.acm.org/doi/10.1145/3472749.3474814>
11. Jiang S, Lam BD, Agrawal M, et al. Machine learning to predict notes for chart review in the oncology setting: a proof of concept strategy for improving clinician note-writing. *J Am Med Inform Assoc.* 2024;ocae092. <https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocae092/7663875>
12. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges. *Int J Environ Res Public Health.* 2021;18:271. <https://www.mdpi.com/1660-4601/18/1/271>
13. Ryu AJ, Ayanian S, Qian R, et al. A clinician's guide to running custom machine-learning models in an electronic health record environment. *Mayo Clinic Proceedings.* 2023;98:445–45. <https://linkinghub.elsevier.com/retrieve/pii/S0025619622006693>
14. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA.* 2019;322:2377. <https://jamanetwork.com/journals/jama/fullarticle/2756196>
15. Sauer CM, Chen L-C, Hyland SL, Girbes PA, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *The Lancet Digital Health.* 2022;4:e893–e898. <https://linkinghub.elsevier.com/retrieve/pii/S2589750022001546>
16. Ryu AJ, Romero-Brufau S, Qian R, et al. Assessing the generalizability of a clinical machine learning model across multiple emergency departments. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes.* 2022;6:193–199. <https://linkinghub.elsevier.com/retrieve/pii/S2542454822000145>
17. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc.* 2022;29:1631–1636. <https://academic.oup.com/jamia/article/29/9/1631/6596175>
18. Kawamoto K, Finkelstein J, Del Fiol G. Implementing machine learning in the electronic health record: checklist of essential considerations. *Mayo Clinic Proceedings.* 2023;98:366–369. <https://linkinghub.elsevier.com/retrieve/pii/S0025619623000204>
19. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? *Case Study in Medicine.* Cornell University, 2023. <http://arxiv.org/abs/2311.16452>
20. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36:1234–1240. <http://arxiv.org/abs/1901.08746>
21. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2020. <http://arxiv.org/abs/1904.05342>
22. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *npj Digit Med.* 2022;5:194. <https://www.nature.com/articles/s41746-022-00742-2>
23. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pre-training for biomedical natural language processing. *ACM Trans Comput Healthcare.* 2022;3:1–23. <http://arxiv.org/abs/2007.15779>
24. Fleming SL, Lozano A, Haberkorn WJ, et al. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. 2023. <https://arxiv.org/abs/2308.14089>
25. Gaffney A, Woolhandler S, Cai C, et al. Medical documentation burden among US office-based physicians in 2019: a national study. *JAMA Intern Med.* 2022;182:564. <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2790396>
26. Mebrahtu TF, McInerney CD, Benn J, et al. The impact of hospital command centre on patient flow and data quality: findings from the UK National Health Service. *Int J Qual Health Care.* 2023;35:mzad072. <https://academic.oup.com/intqhc/article/doi/10.1093/intqhc/mzad072/7282369>
27. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* 2021;181:1065. <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2781307>
28. Joe B, Park Y, Hamm J, Shin I, Lee J. Exploiting missing value patterns for a backdoor attack on machine learning models of electronic health records: development and validation study. *JMIR Med Inform.* 2022;10:e38440. <https://medinform.jmir.org/2022/8/e38440>
29. Schneble CO, Elger BS, Shaw DM. Google's Project Nightingale highlights the necessity of data science ethics review. *EMBO Mol Med.* 2020;12:e12053. <https://www.embopress.org/doi/10.15252/emmm.202012053>
30. Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal impact of implemented early warning score and best practice alert for patient deterioration*. *Critical Care Med.* 2019;47:49–55. <https://journals.lww.com/00003246-201901000-00007>