Machine Learning Applied to Electronic Health Record Data: Opportunities and Challenges

Riccardo Bellazzi
Department of Electrical, Computer and Biomedical Engineering, University of Pavia, IRCCS ICS Maugeri, Pavia, Italy

Abstract: The increasing success of application of machine and deep learning in many areas of medicine, in particular in imaging diagnostics (Rajpurkar et al., 2020), is pushing towards the implementation of AI-based approaches to extract knowledge from health records data (EHR) (Li et al., 2020). The potential of sophisticated strategies to derive regularities from very large collection of textual data, such as language models, is also generating strong expectations about the capability of extracting information unstructured textual notes as well as in generating biomedical texts (Segura-Bedmar et al., 2022; Luo et al., 2022). The COVID-19 pandemics, being one of the most relevant healthcare challenges synchronously happened worldwide, has represented a strong push towards the timely use of EHR data to characterize the clinical course of the COVID-19 disease. Successful examples are represented by cooperative international efforts, such as the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) initiative (Brat et al., 2020). However, EHR data are particularly complex, due to their multifaceted nature and inherent relationship with the health care organizations generating the data. In a recent paper, Kohane and colleagues summarizing the experience carried on in leading 4CE have identified six main challenges that have proven to be crucial for running EHR-based projects (Kohane et al., 2021): i) data completeness, ii) data collection and handling, iii) data type, iv) robustness of methods against EHR variability (within and across institutions, countries, and time), v) transparency of data and analytic code, and vi) the need of multidisciplinary approach. Those topics, in the context of structured EHR data, have been recently further systematized by a consensus paper by the European Society of Cardiology and the BigData@Heart consortium that has defined the CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research (Kotecha et al., 2022). When applying ML to EHR data the above-mentioned aspects become even more important, since data-driven approaches may easily suffer from biases, incompleteness, and lack of contextual information. These problems may lead to models that, even if evaluated with a rigorous statistical testing, can be hardly applicable in practice. As a matter of fact, the "local" nature of the EHR data collection may lead to models that cannot be easily exported in clinical settings other than the one that have generated the training data. For this reason, it is important to provide ML models with additional strategies for self-assessment during clinical use. Recently, reliability has been proposed as an instrument to verify the quality of point predictions, based on two principles: the density principle and the local fit principle (Nicora et al., 2022). The density principle verifies if the case to be evaluated by the model is similar to examples the training set. The local fit principle verifies that the trained model performs well on training subsets that are similar to the instance under evaluation. Reliability and explainability can be seen as safeguards and instruments towards a more trustworthy use of AI and Machine learning. In this talk all these aspects will be discussed through some examples and a few suggestions will be given for future research in this area.

REFERENCES

Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31-38. doi: 10.1038/s41591-021-01614-0. Epub 2022 Jan 20. PMID: 35058619.

Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020 Apr 28;10(1):7155. doi: 10.1038/s41598-020-62922-y. PMID: 32346050; PMCID: PMC7189231.

Segura-Bedmar I, Camino-Perdones D, Guerrero-Aspizua S. Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. BMC Bioinformatics. 2022 Jul 6;23(1):263.

Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu TY. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022 Nov 19;23(6):bbac409.

Bellazzi, R.
Machine Learning Applied to Electronic Health Record Data: Opportunities and Challenges.
DOI: 10.5220/0011950600003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, pages 23-24
ISBN: 978-989-758-631-6; ISSN: 2184-4305

Brat GA, Weber GM, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020 Aug 19;3:109.

Kohane IS, Aronow BJ, et al); Weber GM, Cai T. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2021 Mar 2;23(3):e22219.

Kotecha D, Asselbergs FW, et al. CODE-EHR best-practice framework for the use of structured electronic health-care records in clinical research. Lancet Digit Health. 2022 Oct;4(10):e757-e764. doi: 10.1016/S2589-7500(22)00151-0. Epub 2022 Aug 29. PMID: 36050271.

Nicora G, Rios M, Abu-Hanna A, Bellazzi R. Evaluating pointwise reliability of machine learning prediction. J Biomed Inform. 2022 Mar;127:103996.