frontiers in Big Data

EDITORIAL
published: 12 July 2021
doi: 10.3389/fdata.2021.727856

Editorial: ML and AI Safety, Effectiveness and Explainability in Healthcare

David Benrimoh[1,2]*, Sonia Israel[2], Robert Fratila[2], Caitrin Armstrong[2], Kelly Perlman[2], Ariel Rosenfeld[3] and Adam Kapelner[4]

[1]Department of Psychiatry, McGill University, Montreal, QC, Canada, [2]Aifred Health, Inc., Montreal, QC, Canada, [3]Department of Information Science, Bar-Ilan University, Ramat Gan, Israel, [4]Department of Mathematics, Queens College (CUNY), New York City, NY, United States

Keywords: healthcare, artificial inteligence, machine learning, safety, effectiveness, explainability

Editorial on the Research Topic

ML and AI Safety, Effectiveness and Explainability in Healthcare

The increasing performance of machine learning and artificial intelligence (ML/AI) models has led to them being encountered more frequently in daily life, including in clinical medicine (Bruckert et al.; Rosenfeld et al., 2021). While concerns about the opaque "black box" nature of ML/AI tools are not new, the need for practical solutions to the interpretability problem has become more pressing as ML/AI devices move from the laboratory, through regulatory processes that have yet to fully catch up to the state-of-the-art (Benrimoh et al., 2018a), and to the bedside. This special edition targets three key domains in which innovation and clearer best practices are required for the implementation of ML/AI approaches in healthcare: ensuring safety, demonstrating effectiveness, and providing explainability. Notably, the first two have long been staples in the evaluation of drugs and medical devices (i.e., in order to be approved for human use, products must prove that they are safe and effective—often compared to a reasonable comparator) (Spławiński and Kuźniar, 2004). The third requirement—that of explainability—appears to be unique to ML/AI, due to the challenge of explaining how models arrive at their increasingly accurate conclusions. Yet, upon closer examination, one might argue that the explainability criterion has been implied in the past: mechanisms of action of drugs and devices are generally described in their product documentation (Health Canada, 2014). However, this can be misleading. For instance, many drugs have known receptor binding profiles and putative mechanisms of actions, although the precise mechanisms by which they produce their effect remain unclear despite their widespread use in clinical practice. Prime examples of this are lithium (Shaldubina et al., 2001) and electroconvulsive therapy (Scott, 2011), both longstanding and highly effective treatments whose mechanisms of action remain controversial. Indeed, even the precise mechanism of general anesthesia is a subject of debate (Pleuvry, 2008). As such, we must consider a compromise-that of *sufficient* explainability (Clarke and Kapelner). This involves answering the question: how much must we know about a model in order to determine that it is safe to use in clinical practice? The articles in this special edition begin to explore possible answers to this as well as other key questions in the application of ML/AI to healthcare contexts.

Bruckert et al. propose a Comprehensible Artificial Intelligence (cAI) framework, which they describe as a "cookbook" approach for integrating explainability into ML/AI systems intended to support medical decision-making. Notably, the authors do not limit explainability to an understanding of general rules a model might use to make predictions, but rather extend it to an example-level approach where human-interpretable semantic information is passed from the

Edited and reviewed by:
*Thomas Hartung,
Johns Hopkins University,
United States*

****Correspondence:***
*David Benrimoh
david.benrimoh@mail.mcgill.ca*

Specialty section:
*This article was submitted to
Medicine and Public Health,
a section of the journal
Frontiers in Big Data*

Received: *19 June 2021*
Accepted: *30 June 2021*
Published: *12 July 2021*

Citation:
Benrimoh D, Israel S, Fratila R, Armstrong C, Perlman K, Rosenfeld A and Kapelner A (2021) Editorial: ML and AI Safety, Effectiveness and Explainability in Healthcare. Front. Big Data 4:727856. doi: 10.3389/fdata.2021.727856

machine to the human user. They also discuss systems which not only provide an explanation to the user, but which receive feedback on this explanation in order to learn the implicit rules which experts may use as part of their routine decision-making. Future research could examine the potential biases which a machine might learn from experts, and how this could be mitigated.

Clarke and Kapelner present a Bayesian ML/AI model that predicts outcomes after lens implant (a surgical treatment for cataracts). Their approach to explainability entails generating a list of the most important features used by the model. This method coheres with pre-ML/AI approaches in ophthalmology, which relied on traditional linear equations with clear variables (e.g. Dang and Raj, 1989), an example of an explainability already standard in the field. Because their approach is Bayesian, their predictions come with uncertainty intervals. Thus, the more uncertain the prediction, the wider the interval. Similar to Bruckert et al., the authors insist upon a "human in the loop," noting that surgeons ought to use the algorithm's predictions and uncertainty intervals as a guide within the context of their clinical judgement. Further, they note that the consequences of an incorrect prediction are relatively minor and simply require corrective medical procedures already employed in standard practice. It is interesting to consider that these kinds of predictions - in which a correct prediction provides a benefit but where a failed prediction carries low risk - are ideal first applications of ML/AI while they remain novel technologies (Benrimoh et al., 2018b).

Desai et al. introduce a ML/AI model for the identification of suicidal ideation in the general population. Using a sensitivity analysis, they demonstrate that a deep learning model, which is traditionally difficult to interpret, can be queried using standard statistical approaches to determine whether relationships between variables and outcomes identified by the model are coherent with the literature. Doing so not only makes deep learning a more viable model architecture for use in healthcare, but in situations where a large body of literature exists allows for a "sanity check" of each model produced to ensure they have not learned "quirks" of a biased or non-representative dataset. This could help address Bruckert et al.'s concern about poor explainability arising from biased data and models.

Finally, Wong et al. offer a comprehensive discussion of challenges and opportunities for machine learning approaches in acute respiratory failure. Their discussion is a prime example of the level of granularity necessary for content experts to provide when building AI/ML models in healthcare. In addition to a general review of ML/AI concepts, the authors discuss domain-specific sources of bias as well as difficulties in operationalizing the outcomes that models should be trained to predict. This discourse serves as a useful entry point to the special issue for the clinically-oriented reader who is less familiar with ML/AI approaches. This article reminds us that each healthcare domain has unique challenges, meaning that a one-size-fits-all approach to explainability and the evaluation of safety and effectiveness is unlikely to succeed.

This special edition provides the reader with both a survey of current approaches to integrating ML/AI into healthcare as well as in-depth discussions of how to determine model safety, measure model effectiveness, and provide model explainability that will be of use to clinicians and regulators.

AUTHOR CONTRIBUTIONS

DB led writing of the manuscript. SI, RF, CA, KP, AR, and AK contributed to the writing and review of the manuscript.

REFERENCES

Benrimoh, D., Fratila, R., Israel, S., Perlman, K., Mirchi, N., Desai, S., et al. (2018b). "Aifred Health, a Deep Learning Powered Clinical Decision Support System for Mental Health," in *The NIPS '17 Competition: Building Intelligent Systems*. Editors S. Escalera and M. Weimer (Heidelberg, Germany: Springer International Publishing), 251–287. doi:10.1007/978-3-319-94042-7_13

Benrimoh, D., Israel, S., Perlman, K., Fratila, R., and Krause, M. (2018a). "Meticulous Transparency—An Evaluation Process for an Agile AI Regulatory Scheme," in *Recent Trends and Future Technology in Applied Intelligence*. Editors M. Mouhoub, S. Sadaoui, O. Ait Mohamed, and M. Ali (Heidelberg, Germany: Springer International Publishing), 869–880. doi:10.1007/978-3-319-92058-0_83

Canada, H. (2014). Guidance Document—Product Monograph (Guidance). Available at: https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/applications-submissions/guidance-documents/product-monograph/product-monograph.html (Accessed June 12, 2021)

Dang, M. S., and Raj, P. P. (1989). SRK II Formula in the Calculation of Intraocular Lens Power. *Br. J. Ophthalmol.* 73 (10), 823–826. doi:10.1136/bjo.73.10.823

Pleuvry, B. J. (2008). Mechanism of Action of General Anaesthetic Drugs. *Anaesth. Intensive Care Med.* 9 (Issue 4), 152–153. Available at: https://www.sciencedirect.com/science/article/pii/S1472029907001774.. ISSN 1472-0299. doi:10.1016/j.mpaic.2007.08.004

Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., et al. (2021). *Big Data Analytics and AI in Mental Healthcare. Applications of Big Data in Healthcare*. Amsterdam, Netherlands: Academic Press

Scott, A. I. F. (2011). Mode of Action of Electroconvulsive Therapy: An Update. *Adv. Psychiatr. Treat.* 17 (1), 15–22. doi:10.1192/apt.bp.109.007039

Shaldubina, A., Agam, G., and Belmaker, R. H. (2001). The Mechanism of Lithium Action: State of the Art, Ten Years Later. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* 25 (4), 855–866. doi:10.1016/s0278-5846(01)00154-3

Spławiński, J., and Kuźniar, J. (2004). Clinical Trials: Active Control vs Placebo-What Is Ethical?. *Sci. Eng. Ethics* 10 (1), 73–79. doi:10.1007/s11948-004-0065-x

Conflict of Interest: DB, SI, RF, CA, and KP are employees, founders, or shareholders of Aifred Health. AR and AK have collaborated with Aifred Health and received honoraria from Aifred Health.