VIEWPOINT

Challenges to the Reproducibility of Machine Learning Models in Health Care

Andrew L. Beam, PhD
Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; and Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts.

Arjun K. Manrai, PhD
Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts; and Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts.

Marzyeh Ghassemi, PhD
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada; and Vector Institute, Toronto, Ontario, Canada.

Corresponding Author: Andrew Beam, PhD, Department of Epidemiology, Harvard T.H. Chan School of Public Health, 10 Shattuck St, Boston, MA 02115 (andrew_beam@hms.harvard.edu).

Reproducibility has been an important and intensely debated topic in science and medicine for the past few decades.[1] As the scientific enterprise has grown in scope and complexity, concerns regarding how well new findings can be reproduced and validated across different scientific teams and study populations have emerged. In some instances,[2] the failure to replicate numerous previous studies has added to the growing concern that science and biomedicine may be in the midst of a "reproducibility crisis." Against this backdrop, high-capacity machine learning models are beginning to demonstrate early successes in clinical applications,[3] and some have received approval from the US Food and Drug Administration. This new class of clinical prediction tools presents unique challenges and obstacles to reproducibility, which must be carefully considered to ensure that these techniques are valid and deployed safely and effectively.

Reproducibility is a minimal prerequisite for the creation of new knowledge and scientific progress, but defining precisely what it means for a scientific study to be "reproducible" is complex and has been the subject of considerable effort by both individual researchers and organizations like the National Academies of Science, Engineering, and Medicine. First, it is important to distinguish between the notions of reproducibility and replication. A study is *reproducible* if, given access to the underlying data and analysis code, an independent group can obtain the same result observed in the original study. However, being reproducible does not imply that a study is correct, only that the results were able to be verified by a different group not involved in the original study. A study is *replicable* if an independent group studying the same phenomenon reaches the same conclusion after performing the same set of experiments or analyses after collecting new data.

The discussion around reproducibility and replication has primarily focused on traditional statistical models and the results from randomized clinical trials, but these considerations can and should apply equally to machine learning studies. Challenges to reproducibility and replication include confounding, multiple hypothesis testing, randomness inherent to the analysis procedure, incomplete documentation, and restricted access to the underlying data and code. The last concern, data access, is especially germane for medicine, as privacy barriers are important considerations for data sharing. However, by definition, replication does not require access to the original data or code because a replication exercise examines the extent to which the original phenomenon generalizes to new contexts and new populations.

This Viewpoint focuses on reproducibility, even though it is important to acknowledge that replication is often the ultimate goal. Replication is especially important for studies that use observational data (which is almost always the case for machine learning studies) because these data are often biased, and models could operationalize this bias if not replicated. The challenges of reproducing a machine learning model trained by another research team can be difficult, perhaps even prohibitively so, even with unfettered access to raw data and code.

Unique Challenges to Reproducibility Posed by Machine Learning

Machine learning models have an enormous number of parameters that must be either learned using data or set manually by the analyst. In some instances, simple documentation of the exact configuration (which may involve millions of parameters) is difficult, as many decisions are made "silently" through default parameters that a given software library has preselected. These defaults may differ between libraries and may even differ from version to version of the same library. Thus, 2 researchers using the same code but different versions of a software library could reach substantially different conclusions if important parameters are given different values.

The training of many machine learning models makes use of randomness, and this is especially true for deep learning models,[3,4] which are trained by a process known as stochastic gradient descent. As the name implies, the model is updated using a randomized procedure that will result in different final values for the model parameters every time the code is executed. Thus, if the same model was retrained using the same data, different parameter values will be found each time. The only way to ensure that the results of these models are reproducible is to set a quantity known as the random seed, which controls how random numbers are generated. This value can be any number (eg, 123) but if it is not provided by the user, a combination date, time, or both (eg, 01062020) is often silently used by the system as the random seed. One study[5] found that changing this single, apparently innocuous number could inflate the estimated model performance by as much as 2-fold relative to what a different set of random seeds would yield. Many of the other innumerable, and often silent, parameters that control modern deep learning methods plausibly impart similar influence on the final performance, further complicating reproducibility.

Reproducibility at the Frontier of Machine Learning

Even if these concerns are addressed, the cost to reproduce a state-of-the-art deep learning model from the beginning can be immense. For example, in natural language processing a deep learning model known as the "transformer"[6] has led to a revolution in capabilities across a wide range of tasks, including automatic question answering, machine translation, and algorithms that can write complex and nuanced pieces of descriptive text. Perhaps unsurprisingly,

transformers require a staggering amount of data and computational power and can have in excess of 1 billion trainable parameters.

To automate the process of finding the best transformer for a given application, the machine learning community has developed a technique called neural architecture search, which shifts the task of finding the best transformer to a separate algorithm. This second algorithm, often called a "meta-learner," will evaluate hundreds to thousands of possible configurations to find the transformer with the best predictive performance. A recent study[7] estimated that the cost to reproduce 1 of these models ranged from approximately $1 million to $3.2 million using publicly available cloud computing resources. Thus, simply reproducing this model would require the equivalent of approximately 3 R01 grants from the National Institutes of Health and would rival the cost of some large randomized clinical trials. Perhaps more troubling, the carbon footprint of this approach is substantial. The same study[7] estimated that this process would generate 626 155 lb of CO_2 emissions, which is approximately 5 times the amount of CO_2 that an average car will generate over its entire lifetime on the road. Thus, even if financial resources allow for the reproduction of these models, the environmental implications of doing so raises the question if it even should be considered.

However, high-capacity models at the frontier of machine learning research like the transformer with neural architecture search, do not in general reflect the type of research currently being performed for medical applications of machine learning. Most medical deep learning models are smaller and focused on image recognition, and can be easily reproduced on fairly standard computer hardware. However, models that are adept at dealing with textual data could have immense clinical utility, so researchers may be confronted with reproducing these models at some point in the near future. More optimistically, these large models may provide the basis for productive industry-academic partnerships whereby large, well-resourced tech companies produce a base model that is then provided as a commodity to the research community. Reproducing the base model is expensive, but reproducing small-scale applications of this model is of only moderate expense. Indeed, this is already occurring in medical imaging, where models that were originally trained with enormous databases and computing resources are repurposed and fine-tuned for medical applications, which can then be reproduced with modest resources.

Prospects for Increasing the Reproducibility of Machine Learning Studies

In general, the mainstream machine learning community has embraced fairly radical notions of open science, transparency, and reproducibility. Many reports are first available as preprints, code is usually available as open source, and most articles rely on data sets available in the public domain. Data sources that have been used to train machine learning models are canonically released so that results can be fully reproduced by other researchers. In the medical space, resources such as MIMIC-III, Phillips eICU, CDC NHANES, and the UK Biobank similarly foster reproducibility. This has the added benefit of substantially increasing the pace of research, as results published by one investigative team can be redone and improved on by another within days or weeks.

Medical researchers using machine learning would be well served by adopting some of these practices, including open sharing of data, code, and results whenever possible. While it is not always possible to share data because of privacy concerns, a "walled-garden" approach whereby reviewers are given access to a private network subject to a data use agreement could allow for a reproducibility analysis during the review period for projects with public health implications. Conversely, machine learning researchers moving into medical applications could adhere to standard reporting guidelines such as TRIPOD, CONSORT, and SPIRIT, which are now being adapted for machine learning and artificial intelligence applications.[8] These guidelines set reasonable standards for reporting and transparency, and help communicate how an analysis was done to the larger scientific community.

Determining if machine learning improves patient outcomes remains the most important test, and currently there is scant evidence of downstream benefit. For this, there is likely no substitute for randomized clinical trials. In the meantime, as machine learning begins to influence more health care decisions, ensuring that the foundation on which these tools are built is sound becomes increasingly pressing. In a lesson that is continuously learned, machine learning does not absolve researchers from traditional statistical and reproducibility considerations but simply casts modern light on these historical challenges. At a minimum, a machine learning model should be reproduced, and ideally replicated, before it is deployed in a clinical setting.

ARTICLE INFORMATION

Published Online: January 6, 2020. doi:10.1001/jama.2019.20866

Correction: This article was corrected on January 10, 2020, for an incorrect guideline name in the text and for incomplete information in the funding statement.

Conflict of Interest Disclosures: None reported.

Funding/Support: Dr Beam was supported by National Institutes of Health (NIH) award 7K01HL141771-02. Dr Manrai was supported by NIH award 7K01HL138259-02. Dr Ghassemi was supported by Microsoft Research, a CIFAR AI Chair at the Vector Institute, and a Canada Research Council Chair.

Role of the Funder/Sponsor: The NIH had no role in the preparation, review, or approval of the manuscript and decision to submit the manuscript for publication.

REFERENCES

1. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124

2. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):aac4716. doi:10.1126/science.aac4716

3. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

4. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101-1102. doi:10.1001/jama.2018.11100

5. Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D. Deep reinforcement learning that matters. Presented at: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18); February 2-7, 2018; New Orleans, Louisiana.

6. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, et al, eds. *Advances in Neural Information Processing Systems 30*. Red Hook, NY: Curran Associates Inc; 2017.

7. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv website. https://arxiv.org/abs/1906.02243. Published June 2019. Accessed December 16, 2019.

8. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6