BIOMEDICAL POLICY

Reproducibility in machine learning for health research: Still a ways to go

Matthew B. A. McDermott[1]*†, Shirly Wang[2,3†], Nikki Marinsek[4], Rajesh Ranganath[5], Luca Foschini[4], Marzyeh Ghassemi[2,6,7]

Machine learning for health must be reproducible to ensure reliable clinical use. We evaluated 511 scientific papers across several machine learning subfields and found that machine learning for health compared poorly to other areas regarding reproducibility metrics, such as dataset and code accessibility. We propose recommendations to address this problem.

INTRODUCTION

Reproducibility is required for scientific research, but many subfields of science have recently experienced a reproducibility crisis, eroding trust in processes and results and potentially influencing the rising rates of scientific paper retractions (*1*, *2*). Reproducibility is also critical for machine learning research (*3*); the goal of which is to develop algorithms to reliably solve complex tasks at scale, with limited or no human supervision. Failure of a machine learning system to consistently replicate an intended behavior in a context different from that in which the behavior was defined may have unfortunate consequences. These risks are particularly high in artificial intelligence (AI) and machine learning applied to health (MLH), where algorithmic findings can directly affect human health care (*4*). As more AI health care tools are deployed in clinical practice, ensuring that manufacturers report reproducible performance metrics is in the public interest. This challenge is further underscored by the lack of randomized control trials for deep learning–based systems in MLH and the high risk of bias that has been found in deep learning nonrandomized clinical trials (*5*).

Unfortunately, several factors related to the availability, quality, and consistency of clinical or biomedical data make reproducibility especially challenging in MLH applications. Here, we analyzed the state of reproducibility in MLH, contrasting this subfield of machine learning to both machine learning in general and to the subspecialties of computer vision and natural language processing. We developed a set of criteria for reproducibility tailored to MLH applications and designed to capture reproducibility goals more broadly. We then used these criteria to define several metrics quantifying the particular challenges in reproducibility faced within MLH. Next, we conducted a review of 511 scientific papers to support our claims and to compare machine learning and AI in health care to machine learning more generally. Last, we build on this analysis by exploring promising areas for further research regarding reproducibility in MLH.

REPRODUCIBILITY CRITERIA

The common understanding of reproducibility in machine learning can be summed up as follows: A machine learning study is reproducible if readers can fully replicate the exact results reported in the paper. We will call this concept technical reproducibility, as it is centrally concerned with whether or not one can exactly reproduce the precise, technical results of a paper under identical conditions. Although intuitive, we argue that technical reproducibility is actually only a small part of the goal of reproducibility more generally. This discrepancy has been noted historically in other domains in various ways (*6*, *7*) and is made apparent by use of the term in the natural and social sciences where attempted reproductions will often occur in different laboratories using different equipment. However, to our knowledge, this discrepancy has not been explored in the context of MLH, and discussions of reproducibility in MLH, a subfield of machine learning where reproducibility is especially critical, have been limited to technical reproducibility. We argue that in order for a study to be fully reproducible, it must meet three reproducibility criteria: (i) technical reproducibility (can results be reproduced under technically identical conditions?), (ii) statistical reproducibility (can results be reproduced under statistically identical conditions?), and (iii) conceptual reproducibility or replicability (can results be reproduced under conceptually identical conditions?).

Technical reproducibility refers to the ability of a result to be fully technically replicated, yielding the precise results reported in the paper. This entails aspects of reproducibility related to code and dataset release. Statistical reproducibility refers to the ability of a result to be upheld under resampled conditions that may yield mildly different numerical results but should not statistically affect the claimed result. For example, if an algorithm is trained on a dataset multiple times with different random initializations or with different random subsamples of the data used to train the algorithm versus evaluate the final model, then the reported results should be statistically equivalent even if they are not technically identical. Note that this is related to the notion of internal validity (*8*), which is commonly used in social science research. Last, conceptual reproducibility, or replicability, describes how well the desired results can be reproduced under conditions that match the conceptual, high-level description of the purported effect. Just as statistical reproducibility is like internal validity, replicability is closely related to external validity (*8*), as it describes the notion of how well the desired results can be reproduced under conditions that match the conceptual description of the purported effect. Replicability is task-definition

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada. [3]Layer 6 AI, TD Bank Group, Toronto, ON M5G 1M1, Canada. [4]Evidation Health Inc., San Mateo, CA 94402, USA. [5]Center for Data Science and Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY 10011, USA ; Department of Population Health, NYU Langone Health, New York, NY, USA 10016. [6]Vector Institute, University of Toronto, Toronto, ON M5G 1M1, Canada. [7]Department of Medicine, University of Toronto, Toronto, ON M5S 1A8, Canada.
*Corresponding author. Email: mmd@mit.edu.
†These authors contributed equally to this work.

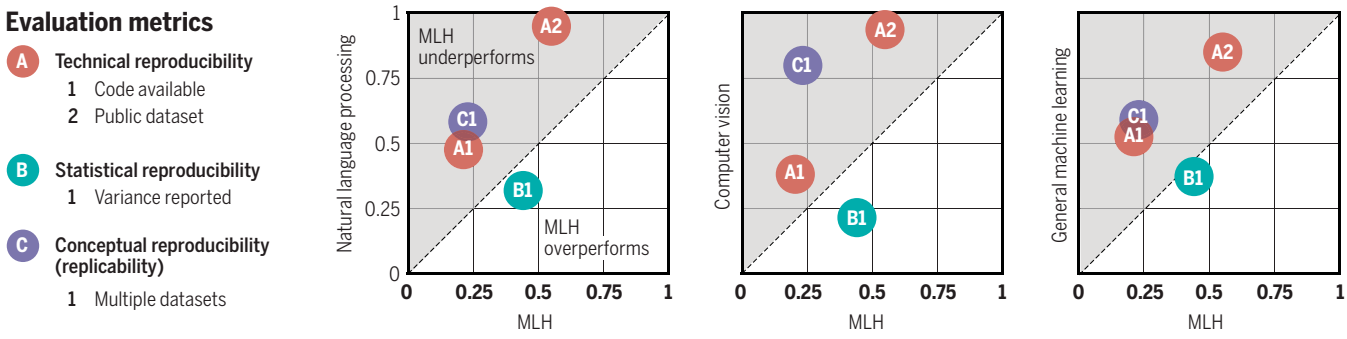

Fig. 1. Reproducibility metrics for machine learning applications. Shown are reproducibility metrics (A, B, and C) for evaluating scientific papers from four machine learning subspecialties: machine learning in health (MLH), natural language processing, computer vision, and general machine learning. Presented is the fraction of papers in a given subspecialty (*y* axis) versus those in MLH (*x* axis) that release their code (A1), release their data (A2), report their variance (B1), and leverage multiple datasets (C1). MLH consistently lags other subfields of machine learning on all measures of reproducibility apart from inclusion of proper statistical variance.

dependent; claiming a task has a greater conceptual horizon of generalizability makes it harder to satisfy this requirement.

All three of these reproducibility criteria are central to full reproducibility. Without technical reproducibility, one's result cannot be demonstrated. Without statistical reproducibility, one's result will not be reproduced under increased sampling and the presence of real-world variance. Without conceptual reproducibility or replicability, one's result does not depend on the desired properties of the data but instead depends on potentially unobserved aspects of the data generation mechanism that, critically, may not be reproduced when deployed in practice. Under each of these lenses, MLH differs from general machine learning domains, such as natural language processing and computer vision, in critical ways and presents unique challenges.

CORE REPRODUCIBILITY CHALLENGES IN MLH

In this section, we illustrate both through qualitative arguments and a quantitative literature review that machine learning in health lags behind other subfields of machine learning on various reproducibility metrics. Our literature review procedure entailed manually extracting and annotating 511 papers presented at various machine learning conferences from 2017 to 2019, spanning the fields of MLH, natural language processing, computer vision, and general machine learning. Each paper was manually annotated regarding reproducibility metrics including whether code and data were publicly available, the kinds of datasets used in the work, and whether variance of the results was reported. Results of this analysis are detailed in the sections below for technical, statistical, and conceptual reproducibility (replicability); Fig. 1 shows the final quantitative results visually. (The full paper annotation and review procedure is detailed in the Supplementary Materials, and the full set of results is available at https://zenodo.org/record/4574378).

Technical reproducibility

MLH papers faced several key challenges regarding technical reproducibility. First, health data are privacy sensitive, making it difficult to release openly without either incurring risks of reidentification or diminishing their usefulness by applying aggressive de-identification techniques. As a result, few public datasets are available, and those that are available are used very frequently, leading to a risk of dataset-specific overfitting. To this point, only ~55% of the MLH papers we examined used public datasets compared to more than 90% of both computer vision and natural language processing papers and ~85% of general machine learning papers (our analysis of the different kinds of datasets used in MLH papers is shown in fig. S1).

MLH papers scored even more poorly when it came to code release, preprocessing specification, and cohort description, with only ~21% of the papers we analyzed releasing their code publicly compared to ~39% of the papers in computer vision and ~48% of the papers in natural language processing. A previous study that reviewed the prevalence of code release in AI reported that only 6% of papers released code (*3*), which is lower than all of our estimates. This discrepancy may be due to sampling papers from different conferences and different time periods (2013–2016 versus 2017–2019 for the current study). A previous study also examined the prevalence of code release and dataset cohort subselection in MLH papers, reporting that even when restricting focus to public datasets, papers often did not release code or included text that was insufficient to enable a full technical reproduction (*9*). Note that code release itself is not necessarily sufficient for full technical reproducibility, because even when code is released, it may not run correctly, it may exclude critical details, or it may fail to generate the results reported in the paper.

Statistical reproducibility

To assess the state of statistical reproducibility in MLH papers, we quantified how often papers described the variance around their results (e.g., by listing both the mean and the SD of a performance metric over several random splits). Interestingly, whereas the rate for this was relatively low (~44%) in MLH papers, it was higher than that for papers in computer vision, natural language processing, or general machine learning (~21, 32, and 37%, respectively).

Although this is an encouraging sign, there is still room for improvement. Even in other fields of machine learning, with arguably less complex data types, repeated studies have shown that published papers fail to be statistically reproduced when appropriate statistical procedures are implemented and fair hyperparameter search/preprocessing methods are used. For example, researchers have reported that published papers evaluated using the public ImageNet test set (a large public dataset and machine learning competition for computer vision studies) show consistent drops in performance when trained

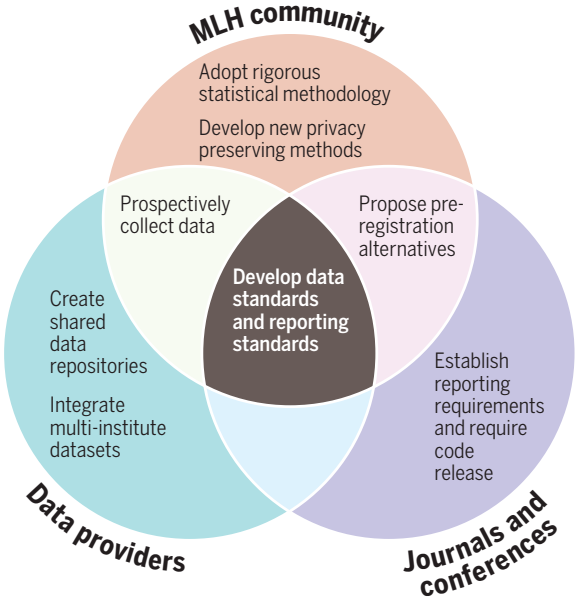

Fig. 2. Improving reproducibility in MLH research. Shown are eight recommendations for improving the reproducibility of machine learning in health research. These recommendations are subdivided according to which primary stakeholders directly drive these changes: the community of MLH researchers, health data providers (e.g., clinical care organizations), and journals and conferences (the primary publishers of machine learning research).

and tested on other random splits within ImageNet (*10*). Issues with statistical reproducibility may also arise due to an effect whereby researchers routinely spend more effort optimizing their method than is spent on the baseline methods against which their method is compared (*11*). It is likely that these problems with statistical reproducibility also plague MLH papers as well, especially given that health datasets tend to be relatively small, have high dimensionality, are noisy, and often suffer from sparse/irregular sampling.

Conceptual reproducibility (replicability)

The critical issue that prevents replicability of results in MLH papers is the lack of multi-institution datasets in health care and the limited usage of those that do exist. Whereas ~80% of computer vision studies and ~58% of natural language processing studies used multiple datasets to establish their results, only ~23% of MLH papers did this. Using only a single health care dataset is not advisable, as it is known that developing machine learning models that attempt to generalize over changing health care practices or health data formats is challenging. Researchers have established that without using manually engineered representations, machine learning models in health exhibit degradation in performance over time as health care patterns evolve (*12*). These results are expected given that health data are rife with hidden confounders, differ markedly between data collection and deployment environments, drift over time, and further differ in structure and concept between different health care institutions (*13*).

OPPORTUNITIES FOR IMPROVEMENT

Here, we present practical suggestions for enhancing reproducibility of results in MLH papers taking into account the MLH research community, health data providers, and associated journals and conferences (Fig. 2).

Create shared research resources

Health data providers such as hospitals, clinical research centers, and government agencies produce vast amounts of valuable health data. Unfortunately, as suggested by our literature review, few health datasets are available for researchers to explore. This is understandable, given the difficulty of ensuring safe, appropriate release of clinical data, the perceived high value of proprietary health data, and the difficulty in designing datasets across noninteroperable health platforms. However, more shared resources would be extremely helpful in enabling reproducible research in the MLH field. We propose more instances of large data trusts where medical institutions can anonymously pool health data for researchers to use and from which algorithms can be created. Several prominent examples of such resources already exist, demonstrating that these challenges are possible to navigate. Such examples include Medical Information Mart for Intensive Care (MIMIC; https://mimic.physionet.org/) (*14*), the national biobanks of the United Kingdom and Japan (*15*, *16*), and the eICU Collaborative Research Database (eICU; https://eicu-crd.mit.edu/) (*17*).

Integrate multi-institute datasets

Multi-institute datasets (i.e., datasets containing health data from multiple care centers or underlying populations) enable studies to assess the ability of algorithms to be translated to new contexts, a critically understudied facet of MLH research. Recent strides have been taken in this regard with the release of the eICU dataset (*17*), one of MLH's first large-scale, multi-institution electronic health record datasets; researchers are already analyzing how to generate generalizable models using this resource (*18*). In addition, Observational Health Data Sciences and Informatics (OHDSI; https://ohdsi.org/) (*19*) provides publicly available code and guidance on best practices to run observational health studies across multiple institutions and countries. Using OHDSI models, researchers have been able to leverage observational health data spanning many health institutions and realize major research goals. Ultimately, however, these efforts only go so far, and we encourage more collaborative efforts from health data providers in this arena to improve reproducibility.

Prospectively collect or directly consent data

Health data collected as a by-product of health care and then later released for research purposes, as is the case for the MIMIC database (*14*), can present serious privacy risks and contain many confounding variables. The landscape of these privacy risks and the nature of the confounding variables change (but do not necessarily lessen) if health data are instead prospectively collected directly from newly consented participants. These types of health data collection regimens are logistically challenging but are possible, as exemplified by the All of Us Research Program (*20*) of the National Institutes of Health, Evidation's DiSCover Project (*21*), and Google's Project Baseline (*22*). In addition, the use of directly consented health data, where patients are able to download their data and make it directly available to research programs, should be considered. This can be enabled by systems such as the Centers for Medicare and Medicaid Services Blue Button (*23*), which allows patients to directly download their health care insurance claims data in readable formats with minimal hassle.

Adopt rigorous statistical methods

MLH researchers should be more rigorous in the development, refinement, and dissemination of statistical best practices, e.g., using the proper procedures for model comparisons. Upholding high standards of statistical rigor, potentially including periodic statistical audits of statistical reproducibility, will

help to ensure mitigation of the problems other fields have found with overfitting of data.

Develop new privacy-preserving analysis techniques

Technological solutions will also be helpful for mitigating privacy concerns by enabling MLH researchers to explore noisy, fully, or partially simulated or encrypted datasets. In cases where health data cannot be released, techniques to train distributed models without sharing data have been proposed (*24*). Synthetic data (i.e., data that are not real but are programmatically generated to resemble real health data) can be an excellent tool to help enable researchers to meaningfully release their code with full end-to-end realization of their pipeline. Technology for producing synthetic patient-level data already exists (*25*).

Preregistration alternatives

In the biomedical sciences, observational studies undergo intense scrutiny to ensure that they are not susceptible to statistical artifacts. Increasingly, these studies are required to be preregistered, meaning they must report their goal and planned analyses before any experiments are performed to avoid intentional or unintentional statistical errors, a move that has both proponents and detractors among scientists and publishers (*26*, *27*). Despite the fact that almost all MLH studies are also observational studies, such prospective checks have so far been absent in MLH research. A verbatim application of the preregistration practices used by the epidemiology community is unlikely to be practical due to the intrinsic exploratory nature of machine learning model development and the fact that datasets used in MLH research are commonly multipurpose. However, other techniques, such as systematic rerelease of new data or rotation of official training/test dataset splits, have been found to help reduce the presence of statistical errors in other fields. MLH researchers and academic publishers should engage in serious conversations regarding the best vehicle for checks and balances for algorithm development.

Establish reporting and code release requirements

Conferences and journals should require data and runnable code release (perhaps through synthetic or sample data, and the use of software containers for easy management) or the provision of statements about additional data and code availability before publication. This would put pressure on the field to address some of the foundational barriers to data reproducibility. Journals have the additional ability to insist on high standards for the reporting of statistical variance around results, hyperparameter search procedures, and evaluation mechanisms, each of which would help to ensure that we maintain high standards of technical and statistical reproducibility. Although code release may be difficult especially in cases of intellectual property restrictions, navigating these issues, which may require a cleaner separation between research and commercial applications, is essential for improving reproducibility in this field. Another important issue regarding code release is the choice of the license used for released software. However, any license that enables reproduction of the stated results (which most standard licenses do) is sufficient for reproducibility, although other important principles may motivate specific licensing choices (e.g., the use of an open source initiative approved license).

Develop data and reporting standards

Collaborative efforts in developing data standards and reporting standards are another avenue for improving reproducibility. Health care analytics organizations have developed data standards such as the Observational Medical Outcomes Partnership standard (*28*) (https://chime.ucsf.edu/observational-medical-outcomes-partnership-omop) and the Fast Healthcare Interoperability Resources standard (*29*) (https://hl7.org/fhir/overview.html), but they are not commonly adopted in MLH research. Increased use of data and reporting standards would make it easier to technically and conceptually replicate MLH studies. Similarly, when datasets are created with the intent to be used in MLH research, better descriptions of their contents, potential confounders and biases, missing data prevalence and distribution, and how they were created should be provided. Increasing the use of "specs" or "datasheets" describing datasets would help to allay these concerns and ensure that datasets are statistically adequate for the kinds of analyses that MLH researchers pursue. Efforts in the broader machine learning community are already achieving this goal (*30*), and these practices should be adopted by the MLH research community.

CONCLUSION

We have framed the question of reproducibility in MLH research around three foundational principles: technical reproducibility, statistical reproducibility, and conceptual reproducibility (replicability). In each of these areas, we argue both qualitatively and quantitatively, through a manual review of the literature, that MLH papers perform worse than do those in other machine learning subfields on several reproducibility metrics. There are intrinsic challenges of data acquisition and use that are inherent to health datasets, but opportunities exist to improve access to health data, expand statistical rigor, and increase the use of multisource data to better enable reproducibility in MLH research.

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/13/586/eabb1655/DC1
Procedures
Table S1. Sources and coverage statistics for our manual literature review.
Fig. S1. A breakdown of datasets used in the MLH papers we analyzed.

REFERENCES AND NOTES

1. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
2. M. Cokol, F. Ozbay, R. Rodriguez-Esteban, Retraction rates are on the rise. *EMBO Rep.* **9**, 2–2 (2008).
3. O. E. Gundersen, S. Kjensmo, State of the art: Reproducibility in artificial intelligence, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (2018), vol. 32.
4. A. L. Beam, A. K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care. *JAMA* **323**, 305–306 (2020).
5. M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, m689 (2020).
6. C. Drummond, Replicability is not Reproducibility: Nor is it Good Science, in *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML* (2009).
7. H. E. Plesser, Reproducibility vs. replicability: A brief history of a confused terminology. *Front. Neuroinform.* **11**, 76 (2018).
8. D. T. Campbell, Relabeling internal and external validity for applied social scientists. *New Dir. Prog. Eval.* **1986**, 67–77 (1986).
9. A. E. Johnson, T. J. Pollard, R. G. Mark, Reproducibility in critical care: A mortality prediction case study, in *Proceedings of the Machine Learning for Healthcare Conference,* (PMLR, 2017), vol. 68, pp. 361–376.
10. B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do ImageNet Classifiers Generalize to ImageNet?, in *Proceedings of the 36th International Conference on Machine Learning,* (PMLR, 2019), vol. 97, pp. 5389–5400.
11. Q. Hu, C. S. Greene, Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics, in *Proceedings of the Pacific Symposium on Biocomputing* (World Scientific, 2019), vol. 24, pp. 362–373.

12. B. Nestor, M. B. A. McDermott, W. Boag, G. Berner, T. Naumann, M. C. Hughes, A. Goldenberg, M. Ghassemi, Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning task, in *Proceedings of the Machine Learning for Healthcare Conference* (PMLR, 2019), vol. 106, pp. 381–405.
13. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), vol. 21, pp. 1721–1730.
14. A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
15. A. Nagai, M. Hirata, Y. Kamatani, K. Muto, K. Matsuda, Y. Kiyohara, T. Ninomiya, A. Tamakoshi, Z. Yamagata, T. Mushiroda, Y. Murakami, K. Yuji, Y. Furukawa, H. Zembutsu, T. Tanaka, Y. Ohnishi, Y. Nakamura; BioBank Japan Cooperative Hospital Group, M. Kubo, Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
16. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
17. T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, O. Badawi, The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
18. A. E. W. Johnson, T. J. Pollard, T. Naumann, Generalizability of predictive models for intensive care unit patients. arXiv:1812.02275 (2018).
19. G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, P. B. Ryan, Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574–578 (2015).
20. All of Us Research Program Investigators, The "All of Us" research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
21. Evidation Health, Digital Signals in Chronic Pain (DiSCover Project), *Clinical Trial NCT03421223* (U.S. National Library of Medicine, 2018).
22. J. Mega, Why Baseline, *Verily Blog* (2017).
23. M. O. Mohsen, H. A. Aziz, The blue button project: Engaging patients in healthcare by a click of a button. *Perspect. Health Inf. Manag.* **12**, 1d (2015).
24. P. Vepakomma, O. Gupta, T. Swedish, R. Raskar, Split learning for health: Distributed deep learning without sharing raw patient data. arXiv:1812.00564 (2018).
25. J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.* **25**, 230–238 (2018).
26. E. Loder, T. Groves, D. MacAuley, Registration of observational studies. *BMJ* **340**, c950 (2010).
27. T. L. Lash, J. P. Vandenbroucke, Commentary: Should preregistration of epidemiologic study protocols become compulsory? Reflections and a counterproposal. *Epidemiology* **23**, 184–188 (2012).
28. J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, P. E. Stang, Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.* **19**, 54–60 (2012).
29. HL7, Introducing HL7 FHIR, *Tech. rep.* (HL7, 2018).
30. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, K. Crawford, Datasheets for datasets. arXiv:1803.09010 (2018).

Acknowledgments: We thank B. Nestor, A. Lu, D. Wu, E. Sergevea, and D. Jin for paper annotation. **Funding:** M.B.A.M. was funded in part by National Institutes of Health/National Institute of Mental Health grant no. P50-MH106933, National Institutes of Health grant no. LM013337, and a Mitacs Globalink Research Award. M.G. was funded in part by a CIFAR AI Chair at the Vector Institute, a Canada Research Council chair, Microsoft Research, and an NSERC Discovery Grant. **Competing interests:** M.G. is a consultant for Radical Ventures and St. Michael's Hospital.

10.1126/scitranslmed.abb1655

Citation: M. B. A. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, M. Ghassemi, Reproducibility in machine learning for health research: Still a ways to go. *Sci. Transl. Med.* **13**, eabb1655 (2021).

Science Translational Medicine

Reproducibility in machine learning for health research: Still a ways to go

Matthew B. A. McDermottShirly WangNikki MarinsekRajesh RanganathLuca FoschiniMarzyeh Ghassemi

Sci. Transl. Med., 13 (586), eabb1655. • DOI: 10.1126/scitranslmed.abb1655

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.