HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2021 May 25.

Published in final edited form as:

Nature. 2020 October ; 586(7829): E14–E16. doi:10.1038/s41586-020-2766-y.

Transparency and reproducibility in artificial intelligence

Benjamin Haibe-Kains[1,2,3,4,5,$], **George Alexandru Adam**[3,5], **Ahmed Hosny**[6,7], **Farnoosh Khodakarami**[1,2], **Massive Analysis Quality Control (MAQC) Society Board of Directors**[*], **Levi Waldron**[16], **Bo Wang**[2,5,17], **Chris McIntosh**[2,5,17], **Anna Goldenberg**[3,5,18,19], **Anshul Kundaje**[20], **Casey S. Greene**[21,22], **Tamara Broderick**[23], **Michael M. Hoffman**[1,2,3,5], **Jeffrey T. Leek**[24], **Keegan Korthauer**[25], **Wolfgang Huber**[26], **Alvis Brazma**[27], **Joelle Pineau**[28,29], **Robert Tibshirani**[30,31], **Trevor Hastie**[30,31], **John P.A. Ioannidis**[30,31,32,33,34], **John Quackenbush**[35,36,37], **Hugo J.W.L. Aerts**[6,7,38,39], **Thakkar Shraddha**[8], **Rebecca Kusko**[9], **Susanna-Assunta Sansone**[10], **Weida Tong**[8], **Russ D. Wolfinger**[11], **Christopher Mason**[12], **Wendell Jones**[13], **Joaquin Dopazo**[14], **Cesare Furlanello**[15]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada [3]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada [4]Ontario Institute for Cancer Research, Toronto, Ontario, Canada [5]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada [6]Artificial Intelligence in Medicine (AIM) Program, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA [7]Radiation Oncology and Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA [8]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA [9]Immuneering Corporation, Cambridge, Massachusetts, USA [10]Engineering Science Department, Oxford e-Research Centre, University of Oxford, Oxford, UK [11]SAS Institute Inc., Cary, North Carolina, USA [12]Weill Cornell Medicine, New York, New York, USA [13]Q2 Solutions, Morrisville, North Carolina, USA [14]Hospital Virgen del Rocio, Sevilla, Spain [15]Fondazione Bruno Kessler, Trento, Italy [16]Department of Epidemiology and Biostatistics and Institute for Implementation Science in Population Health, CUNY Graduate School of Public Health and Health Policy, New York, NY, USA [17]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada [18]SickKids Research Institute, Toronto, Ontario, Canada [19]Child and Brain Development Program, CIFAR, Toronto, Ontario, Canada [20]Department of

[$]**Corresponding Author** Benjamin Haibe-Kains: bhaibeka@uhnresearch.ca.
[*]A list of authors and their affiliations appears at the end of the paper

Author Contributions

BHK and GAA wrote the first draft of the manuscript. BHK and HJWLA designed and supervised the study. AH, FK, TS, RK, SAS, WT, RDW, CN, WJ, JD, CF, LW, BW, CM, AG, AK, CSG, TB, MMH, JTL, KK, WH, AB, JP, RT, TH, JPAI and JQ contributed to the writing of the manuscript.

Competing Interests

AH is a shareholder of and receives consulting fees from Altis Labs. MMH received a GPU Grant from Nvidia. HJWLA is a shareholder of and receives consulting fees from Onc.AI. BHK is a scientific advisor for Altis Labs. CM holds an equity position in Bridge7Oncology and receives royalties from RaySearch Laboratories. GAA, FK, LW, BW, AK, CSG, JTL, WH, AB, JP, RT, TH, JPAI and JQ declare no other competing interests related to the manuscript.

Data Availability

No data have been generated as part of this manuscript.

Code Availability

No computer code has been generated as part of this manuscript.

Genetics, Stanford University School of Medicine, Stanford, CA, USA [21]Dept. of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA [22]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA [23]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [24]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore MD, USA [25]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada BC Children's Hospital Research Institute, Vancouver, British Columbia, Canada [26]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany [27]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK [28]McGill University, Montreal, QC, Canada [29]Montreal Institute for Learning Algorithms, QC, Canada [30]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA [31]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA [32]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA [33]Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA, USA [34]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA [35]Department of Biostatistics, Harvard T.H Chan School of Public Health, Boston, MA, USA [36]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA [37]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA [38]Radiology and Nuclear Medicine, Maastricht University, Maastricht, Netherlands [39]Cardiovascular Imaging Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Breakthroughs in artificial intelligence (AI) hold enormous potential as it can automate complex tasks and go even beyond human performance. In their study, McKinney et al. showed the high potential of AI for breast cancer screening. However, the lack of methods' details and algorithm code undermines its scientific value. Here, we identify obstacles hindering transparent and reproducible AI research as faced by McKinney et al., and provide solutions to these obstacles with implications for the broader field.

The work by McKinney et al.[1] demonstrates the potential of AI in medical imaging, while highlighting the challenges of making such work reproducible. The authors assert that their system improves the speed and robustness of breast cancer screening, generalizes to populations beyond those used for training, and outperforms radiologists in specific settings. Upon successful prospective clinical validation and approval by regulatory bodies, this new system holds great potential for streamlining clinical workflows, reducing false positives, and improving patient outcomes. However, the absence of sufficiently documented methods and computer code underlying the study effectively undermines its scientific value. This shortcoming limits the evidence required for others to prospectively validate and clinically implement such technologies. By identifying obstacles hindering transparent and reproducible AI research as faced by McKinney et al., we provide potential solutions with implications for the broader field.

Scientific progress depends upon the ability of independent researchers to (1) scrutinize the results of a research study, (2) reproduce the study's main results using its materials, and (3) build upon them in future studies[2]. Publication of insufficiently documented research does not meet the core requirements underlying scientific discovery[3,4]. Merely textual descriptions of deep learning models can hide their high level of complexity. Nuances in the computer code may have dramatic effects on the training and evaluation of results[5], potentially leading to unintended consequences[6]. Therefore, transparency in the form of the actual computer code used to train a model and arrive at its final set of parameters is essential for research reproducibility. The authors state *"The code used for training the models has a large number of dependencies on internal tooling, infrastructure and hardware, and its release is therefore not feasible"*. Computational reproducibility is indispensable for high-quality AI applications[7,8]; more complex methods demand greater transparency[9]. In the absence of code, reproducibility falls back on replicating methods from textual description. Although, the authors claim that *"all experiments and implementation details are described in sufficient detail in the* Supplementary Methods *section to support replication with non-proprietary libraries"*, key details about their analysis are lacking. Even with extensive description, reproducing complex computational pipelines based purely on text is a subjective and challenging task[10].

In addition to the reproducibility challenges inherent to purely textual descriptions of methods, the authors' description of the model development as well as data processing and training pipelines lacks critical details. The definitions of multiple hyperparameters for the model's architecture (composed of three networks referred to as the Breast, Lesion, and Case models) are missing (Table 1). In their original publication, the authors did not disclose the settings for the augmentation pipeline; the transformations used are stochastic and can significantly affect model performance[11]. Details of the training pipeline were also missing. Without this key information, independent reproduction of the training pipeline is not possible.

There exist numerous frameworks and platforms to make artificial intelligence research more transparent and reproducible (Table 2). For the sharing of code, these include Bitbucket, GitHub, and GitLab among others. The multiple software dependencies of large-scale machine learning applications require appropriate control of the software environment, which can be achieved through package managers including Conda, as well as container and virtualization systems, including Code Ocean, Gigantum, Colaboratory, and Docker. If virtualization of the McKinney et al. internal tooling proved to be difficult, they could have released the computer code and documentation. The authors could have also created small artificial examples or used small public datasets[12] to show how new data must be processed to train the model and generate predictions. Sharing the fitted model (architecture along with learned parameters) should be simple aside from privacy concerns that the model may reveal sensitive information about the set of patients used to train it. Nevertheless, techniques for achieving differential privacy exist to alleviate such concerns. Many platforms allow sharing of deep learning models, including TensorFlow Hub, ModelHub.ai, ModelDepot, and Model Zoo with support for multiple frameworks such as PyTorch and Caffe, as well as the TensorFlow library used by the authors. In addition to improving accessibility and

transparency, such resources can significantly accelerate model development, validation, and transition into production and clinical implementation.

Another crucial aspect of ensuring reproducibility lies in access to the data the models were derived from. In their study, McKinney et al. used two large datasets under license, properly disclosing this limitation in their publication. Sharing of patient health information is highly regulated due to privacy concerns. Despite these challenges, sharing of raw data has become more common in biomedical literature, increasing from under 1% in the early 2000s to 20% today[13]. However, if the data cannot be shared, the model predictions and data labels themselves should be released, allowing further statistical analyses. Above all, concerns about data privacy should not be used as a smokescreen to distract from the requirement to release code.

Although sharing of code and data is widely seen as a crucial part of scientific research, the adoption varies across fields. In fields such as genomics, complex computational pipelines and sensitive datasets have been shared for decades[14]. Guidelines related to genomic data are clear, detailed, and most importantly, enforced. It is generally accepted that all code and data are released alongside a publication. In other fields of medicine and science as a whole, this is much less common, and data and code are rarely made available. For scientific efforts where a clinical application is envisioned and human lives would be at stake, we argue that the bar of transparency should be set even higher. If a dataset cannot be shared with the entire scientific community, because of licensing or other insurmountable issues, at a minimum a mechanism should be set so that some highly-trained, independent investigators can access the data and verify the analyses.

The lack of access to code and data in prominent scientific publications may lead to unwarranted and even potentially harmful clinical trials [15]. These unfortunate lessons have not been lost on journal editors and their readers. Journals have an obligation to hold authors to the standards of reproducibility that benefit not only other researchers, but also the authors themselves. Making one’s methods reproducible may surface biases or shortcomings to authors before publication[6]. Preventing external validation of a model will likely reduce its impact, as it also prevents other researchers from using and building upon it in future studies. The failure of McKinney et al. to share key materials and information transforms their work from a scientific publication open to verification and adoption by the scientific community into a promotion of a closed technology.

We have high hopes for the utility of AI methods in medicine. Ensuring that these methods meet their potential, however, requires that these studies be scientifically reproducible. The recent advances in computational virtualization and AI frameworks are greatly facilitating the implementations of complex deep neural networks in a more structured, transparent, and reproducible way. Adoption of these technologies will increase the impact of published deep learning algorithms and accelerate the translation of these methods into clinical settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Scott McKinney and colleagues for their prompt and open communication regarding the materials and methods of their study.

References

1. McKinney SM, Sieniek M, Godbole V & Godwin J. International evaluation of an AI system for breast cancer screening. Nature (2020).
2. Nature Research Editorial Policies. Reporting standards and availability of data, materials, code and protocols. Springer Nature https://www.nature.com/nature-research/editorial-policies/reporting-standards.
3. Bluemke DA et al. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board. Radiology 192515 (2019) doi:10.1148/radiol.2019192515.
4. Gundersen OE, Gil Y & Aha DW On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. AI Magazine 39, 56–68 (2018).
5. Crane M. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. Transactions of the Association for Computational Linguistics 6, 241–252 (2018).
6. Sculley D. et al. Hidden Technical Debt in Machine Learning Systems. in Advances in Neural Information Processing Systems 28 (eds. Cortes C, Lawrence ND, Lee DD, Sugiyama M & Garnett R.) 2503–2511 (Curran Associates, Inc., 2015).
7. Stodden V et al. Enhancing reproducibility for computational methods. Science 354, 1240–1241 (2016). [PubMed: 27940837]
8. Hutson M. Artificial intelligence faces reproducibility crisis. Science 359, 725–726 (2018). [PubMed: 29449469]
9. Bzdok D & Ioannidis JPA Exploration, Inference, and Prediction in Neuroscience and Biomedicine. Trends Neurosci. 42, 251–262 (2019). [PubMed: 30808574]
10. Gundersen OE & Kjensmo S State of the art: Reproducibility in artificial intelligence. in Thirty-second AAAI conference on artificial intelligence (2018).
11. Shorten C & Khoshgoftaar TM A survey on Image Data Augmentation for Deep Learning. Journal of Big Data 6, 60 (2019).
12. Lee RS et al. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data 4, 170177 (2017). [PubMed: 29257132]
13. Wallach JD, Boyack KW & Ioannidis JPA Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. PLoS Biol. 16, e2006930 (2018).
14. Amann RI et al. Toward unrestricted use of public genomic data. Science 363, 350–352 (2019). [PubMed: 30679363]
15. Carlson B. Putting oncology patients at risk. Biotechnol. Healthc. 9, 17–21 (2012). [PubMed: 23091430]

Table 1:

Essential hyperparameters for reproducing the study for each of the three models (Lesion, Breast, and Case), including those missing from the description in Mckinney et al.

	Lesion	Breast	Case
Learning rate	Missing	0.0001	Missing
Learning rate schedule	Missing	Stated	Missing
Optimizer	Stochastic gradient descent with momentum	Adam	Missing
Momentum	Missing	Not applicable	Not applicable
Batch size	4	Unclear	2
Epochs	Missing	120,000	Missing

Table 2:

Frameworks and platforms to share code, software dependencies and deep learning models to make artificial intelligence research more transparent and reproducible.

Resource		URL
Code		
BitBucket		https://bitbucket.org
GitHub		https://github.com
GitLab		https://about.gitlab.com
Software dependencies		
Conda		https://conda.io
Code Ocean		https://codeocean.com
Gigantum		https://gigantum.com
Colaboratory		https://colab.research.google.com
Deep learning models		
TensorFlow Hub		https://www.tensorflow.org/hub
ModelHub		http://modelhub.ai
ModelDepot		https://modeldepot.io
Model Zoo		https://modelzoo.co
Deep learning frameworks		
TensorFlow		https://www.tensorflow.org/
Caffe		https://caffe.berkeleyvision.org/
PyTorch		https://pytorch.org/