Matters arising

Transparency and reproducibility in artificial intelligence

https://doi.org/10.1038/s41586-020-2766-y

Received: 1 February 2020

Accepted: 10 August 2020

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[8], Bo Wang[2,3,5,9,10], Chris McIntosh[2,5,9], Anna Goldenberg[3,5,11,12], Anshul Kundaje[13,14], Casey S. Greene[15,16], Tamara Broderick[17], Michael M. Hoffman[1,2,3,5], Jeffrey T. Leek[18], Keegan Korthauer[19,20], Wolfgang Huber[21], Alvis Brazma[22], Joelle Pineau[23,24], Robert Tibshirani[25,26], Trevor Hastie[25,26], John P. A. Ioannidis[25,26,27,28,29], John Quackenbush[30,31,32] & Hugo J. W. L. Aerts[6,7,33,34]

ARISING FROM S. M. McKinney et al. *Nature* https://doi.org/10.1038/s41586-019-1799-6 (2020)

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.[1] showed the high potential of AI for breast cancer screening. However, the lack of details of the methods and algorithm code undermines its scientific value. Here, we identify obstacles that hinder transparent and reproducible AI research as faced by McKinney et al.[1], 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.[1], we provide potential solutions with implications for the broader field.

Scientific progress depends on the ability of independent researchers to scrutinize the results of a research study, to reproduce the study's main results using its materials, and to build on them in future studies (https://www.nature.com/nature-research/editorial-policies/reporting-standards). Publication of insufficiently documented research does not meet the core requirements underlying scientific discovery[2,3]. Merely textual descriptions of deep-learning models can hide their high level of complexity. Nuances in the computer code may have marked effects on the training and evaluation of results[4], potentially leading to unintended consequences[5]. 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. McKinney et al.[1] stated that the code used for training the models has "a large number of dependencies on internal tooling, infrastructure and hardware", and claimed that the release of the code was therefore not possible. Computational reproducibility is indispensable for high-quality AI applications[6,7]; more complex methods demand greater transparency[8]. In the absence of code, reproducibility falls back on replicating methods from textual description. Although, McKinney and colleagues[1] claim that all experiments and implementation details were described in sufficient detail in the supplementary methods section of their Article[1] 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[9].

In addition to the reproducibility challenges inherent to purely textual descriptions of methods, the description by McKinney et al.[1] of the model development as well as data processing and training pipelines lacks crucial details. The definitions of several 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 publication, McKinney et al.[1] did not disclose the settings for the augmentation pipeline; the transformations used are stochastic and can considerably affect model performance[10]. Details of the training pipeline were also missing. Without this key information, independent reproduction of the training pipeline is not possible.

[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]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. [9]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada. [10]Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada. [11]SickKids Research Institute, Toronto, Ontario, Canada. [12]Child and Brain Development Program, CIFAR, Toronto, Ontario, Canada. [13]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [14]Department of Computer Science, Stanford University, Stanford, CA, USA. [15]Dept. of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [16]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA. [17]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [18]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [19]Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada. [20]BC Children's Hospital Research Institute, Vancouver, British Columbia, Canada. [21]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [22]European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK. [23]McGill University, Montreal, Quebec, Canada. [24]Montreal Institute for Learning Algorithms, Quebec, Canada. [25]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA. [26]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [27]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [28]Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA, USA. [29]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA. [30]Department of Biostatistics, Harvard T.H Chan School of Public Health, Boston, MA, USA. [31]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [32]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [33]Radiology and Nuclear Medicine, Maastricht University, Maastricht, The Netherlands. [34]Cardiovascular Imaging Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: bhaibeka@uhnresearch.ca

Table 1 | Essential hyperparameters for reproducing the study for each of the three models

	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 to share code, software dependencies and deep-learning models

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/

Numerous frameworks and platforms exist 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 many 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.[1] 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[11] 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 several 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 considerably 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.[1] used two large datasets under license, properly disclosing this limitation in their publication. The sharing of patient health information is highly regulated owing to privacy concerns. Despite these challenges, the sharing of raw data has become more common in biomedical literature, increasing from under 1% in the early 2000s to 20% today[12]. 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 way to distract from the requirement to release code.

Although sharing of code and data are 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[13]. 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 in which 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[14]. 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[5]. 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.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

No data have been generated as part of this manuscript.

1. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
2. Bluemke, D. A. et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the *Radiology* editorial board. *Radiology* **293**, 315–316 (2019).
3. Gundersen, O. E., Gil, Y. & Aha, D. W. On reproducible AI: towards reproducible research, open science, and digital scholarship in AI publications. *AI Mag.* **39**, 56–68 (2018).
4. Crane, M. Questionable answers in question answering research: reproducibility and variability of published results. *Trans. Assoc. Comput. Linguist.* **6**, 241–252 (2018).
5. Sculley, D. et al. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C. et al.) 2503–2511 (Curran Associates, Inc., 2015).
6. Stodden, V. et al. Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241 (2016).

7. Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).
8. Bzdok, D. & Ioannidis, J. P. A. Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).
9. Gundersen, O. E. & Kjensmo, S. State of the art: Reproducibility in artificial intelligence. In *Thirty-second AAAI Conference on Artificial Intelligence (AAAI-18)* 1644–1651 (2018).
10. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **6**, 60 (2019).
11. Lee, R. S. et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**, 170177 (2017).
12. Wallach, J. D., Boyack, K. W. & Ioannidis, J. P. A. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLoS Biol.* **16**, e2006930 (2018).
13. Amann, R. I. et al. Toward unrestricted use of public genomic data. *Science* **363**, 350–352 (2019).
14. Carlson, B. Putting oncology patients at risk. *Biotechnol. Healthc.* **9**, 17–21 (2012).

Acknowledgements We thank S. McKinney and colleagues for their prompt and open communication regarding the materials and methods of their study. This work was supported in part by the National Cancer Institute (R01 CA237170).

Author contributions B.H.-K. and G.A.A. wrote the first draft of the manuscript. B.H.-K. and H.J.W.L.A. designed and supervised the study. A.H., F.K., T.S., R.K., S.-A.S., W.T., R.D.W., C.E.M., W.J., J.D., C.F., L.W., B.W., C. McIntosh, A.G., A.K., C.S.G., T.B., M.M.H., J.T.L., K.K., W.H., A.B., J.P., R.T., T.H., J.P.A.I. and J.Q. contributed to the writing of the manuscript.

Competing interests A.H. is a shareholder of and receives consulting fees from Altis Labs. M.M.H. received a GPU Grant from Nvidia. H.J.W.L.A. is a shareholder of and receives consulting fees from Onc.AI. B.H.K. is a scientific advisor for Altis Labs. C.M. holds an equity position in Bridge7Oncology and receives royalties from RaySearch Laboratories. A.K. is on the SAB of ImmuneAI Inc, a consultant for Biogen Inc., a scientific co-founder of RavelBio Inc. and a shareholder of Freenome Inc. G.A.A., F.K., L.W., B.W., C.S.G., J.T.L., W.H., A.B., J.P., R.T., T.H., J.P.A.I. and J.Q. declare no other competing interests related to the manuscript.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2766-y.
Correspondence and requests for materials should be addressed to B.H.-K.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Massive Analysis Quality Control (MAQC) Society Board of Directors

Thakkar Shraddha[35], Rebecca Kusko[36], Susanna-Assunta Sansone[37], Weida Tong[35], Russ D. Wolfinger[38], Christopher E. Mason[39], Wendell Jones[40], Joaquin Dopazo[41] & Cesare Furlanello[42]

[35]National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA. [36]Immuneering Corporation, Cambridge, MA, USA. [37]Engineering Science Department, Oxford e-Research Centre, University of Oxford, Oxford, UK. [38]SAS Institute Inc, Cary, NC, USA. [39]Weill Cornell Medicine, New York, NY, USA. [40]Q2 Solutions, Morrisville, NC, USA. [41]Hospital Virgen del Rocio, Sevilla, Spain. [42]Fondazione Bruno Kessler, Trento, Italy.

Author Queries

Journal: **Nature**
Paper: **s41586-020-2766-y**
Title: **Transparency and reproducibility in artificial intelligence**

AUTHOR:
The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query Reference	Reference
Q1	Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.
Q2	Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.
Q3	Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly. Note that changes here will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding email address(es).
Q4	You cannot alter accepted Supplementary Information files except for critical changes to scientific content. If you do resupply any files, please also provide a brief (but complete) list of changes. If these are not considered scientific changes, any altered Supplementary files will not be used, only the originally accepted version will be published.
Q5	Owing to the addition of the new affiliations for authors Bo Wang, Anshul Kundake and Keegan Korthauer, subsequent affiliations have been renumbered. Please check all author affiliations are now corrected as shown.

nature research

Corresponding author(s): *Double-blind peer review submissions: write DBPR and your manuscript number here instead of author names.*

Last updated by author(s): *YYYY-MM-DD*

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data have been generated as part of this manuscript.
Data analysis	No computer code has been generated as part of this manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No data have been generated as part of this manuscript.

Field-specific reporting

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/a
Data exclusions	n/a
Replication	n/a
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging