HELMHOLTZ MUNICH

Reproducibility in the context of AI in health care

Sophia J. Wagner, Christian Matek, Sayedali Shetab Boushehri, Melanie Boxberg, Lorenz Lamm, Ario Sadafi, Dominik J. E. Waibel, Carsten Marr, Tingying Peng

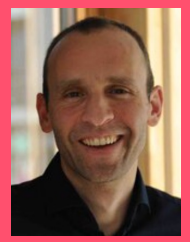
Reproducibility Workshop

February 14, 2023








Reproducibility in AI for health care

This Issue | Views 21,966 | Citations 52 | Altmetric 416

Viewpoint

January 6, 2020

Challenges to the Reproducibility of Machine Learning Models in Health Care

Andrew L. Beam, PhD[1,2]; Arjun K. Manrai, PhD[2,3]; Marzyeh Ghassemi, PhD[4,5]

Matters Arising | Published: 14 October 2020

Transparency and reproducibility in artificial intelligence

Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush & Hugo J. W. L. Aerts

Analysis | Open Access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

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

Machine learning applied to health falls short on several reproducibility metrics compared to other machine learning subfields.

MATTHEW B. A. MCDERMOTT, SHIRLY WANG, NIKKI MARINSEK, RAJESH RANGANATH, LUCA FOSCHINI, AND MARZYEH GHASSEMI Authors Info & Affiliations

SCIENCE TRANSLATIONAL MEDICINE · 24 Mar 2021 · Vol 13, Issue 586 · DOI: 10.1126/scitranslmed.abb1655

Comment | Published: 30 August 2021

Reproducibility standards for machine learning in the life sciences

Benjamin J. Heil, Michael M. Hoffman, Florian Markowetz, Su-In Lee, Casey S. Greene & Stephanie C. Hicks

Comment | Published: 04 October 2021

Avoiding a replication crisis in deep-learning-based bioimage analysis

Romain F. Laine, Ignacio Arganda-Carreras, Ricardo Henriques & Guillaume Jacquemet

Comment | Published: 08 August 2022

Make deep learning algorithms in computational pathology more reproducible and reusable

Sophia J. Wagner, Christian Matek, Sayedali Shetab Boushehri, Melanie Boxberg, Lorenz Lamm, Ario Sadafi, Dominik J. E. Waibel, Carsten Marr & Tingying Peng

Reproducibility of deep learning in digital pathology whole slide image analysis

Christina Fell, Mahnaz Mohammadi, David Morrison, Ognjen Arandjelovic, Peter Caie, David Harris-Birtill

Published: December 2, 2022 • https://doi.org/10.1371/journal.pdig.0000145

TECHNOLOGY FEATURE | 09 January 2023 | Correction 12 January 2023

The reproducibility issues that haunt health-care AI

Health-care systems are rolling out artificial-intelligence tools for diagnosis and monitoring. But how reliable are the models?

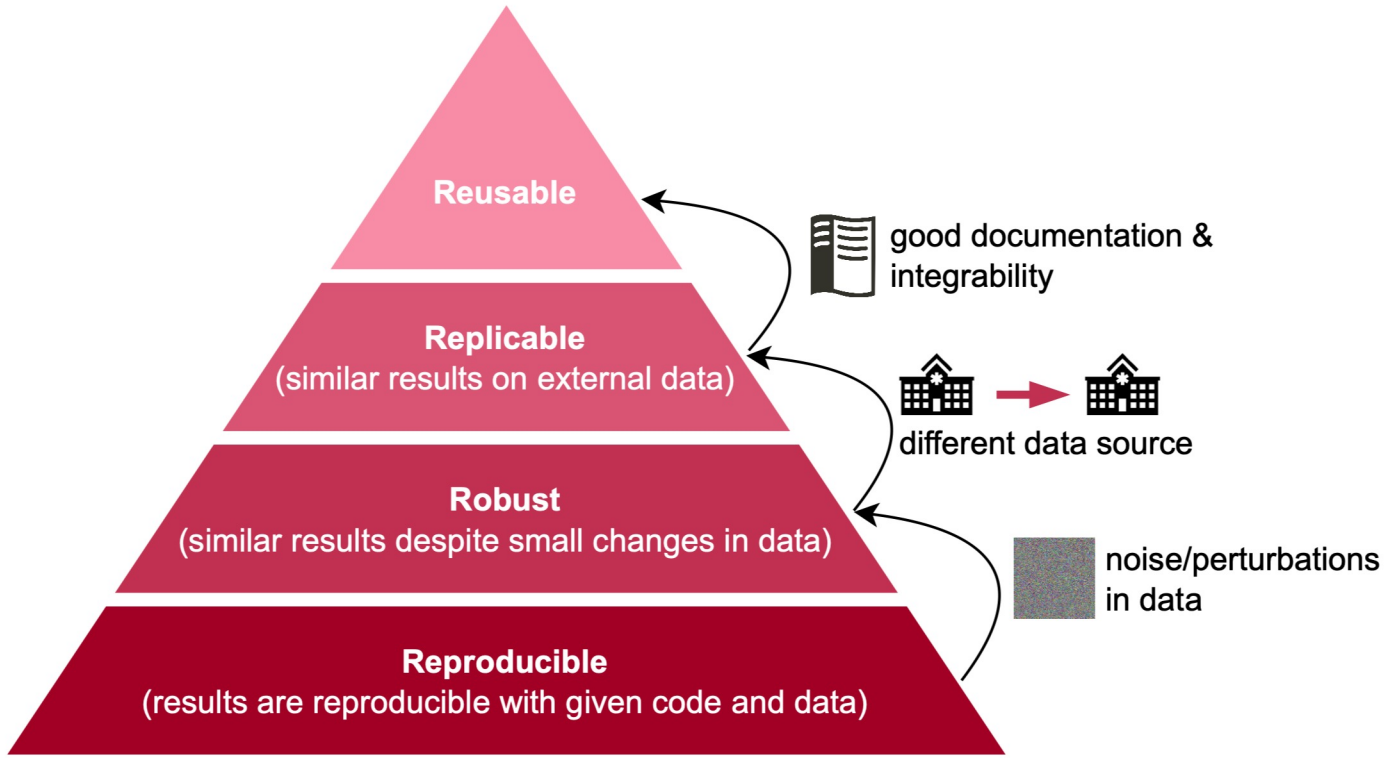
Multiple reproducibility workshops at popular ML conferences such as ICML, ICLR, or NeurIPS

HELMHOLTZ MUNICH
https://doi.org/10.48440/os.helmholtz.062

"Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. […] Reproducibility is a minimum necessary condition for a finding to be believable and informative."

Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.

Reproducibility and reusability



Challenges for AI in health care

Data collection

- Bias in data
- Data leakage
- Annotations
- Varying pre-processing
- Reporting of metadata or clinical data

Evaluation

- Choice of metrics
- Statistical evaluation
- Generalization
- Standardized reporting

Path to the clinic

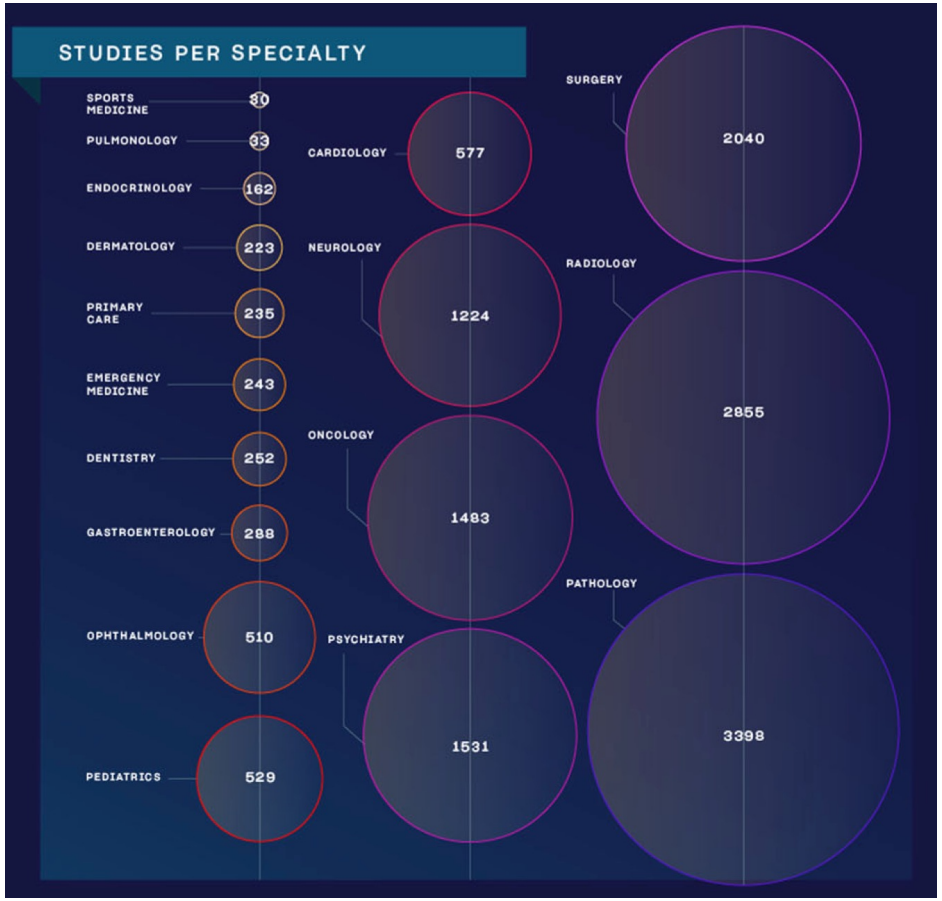
Model development

- Suitable choice of relevant tasks
- Code sharing including all training details
- User interfaces for cliniciancs
- Maintenance

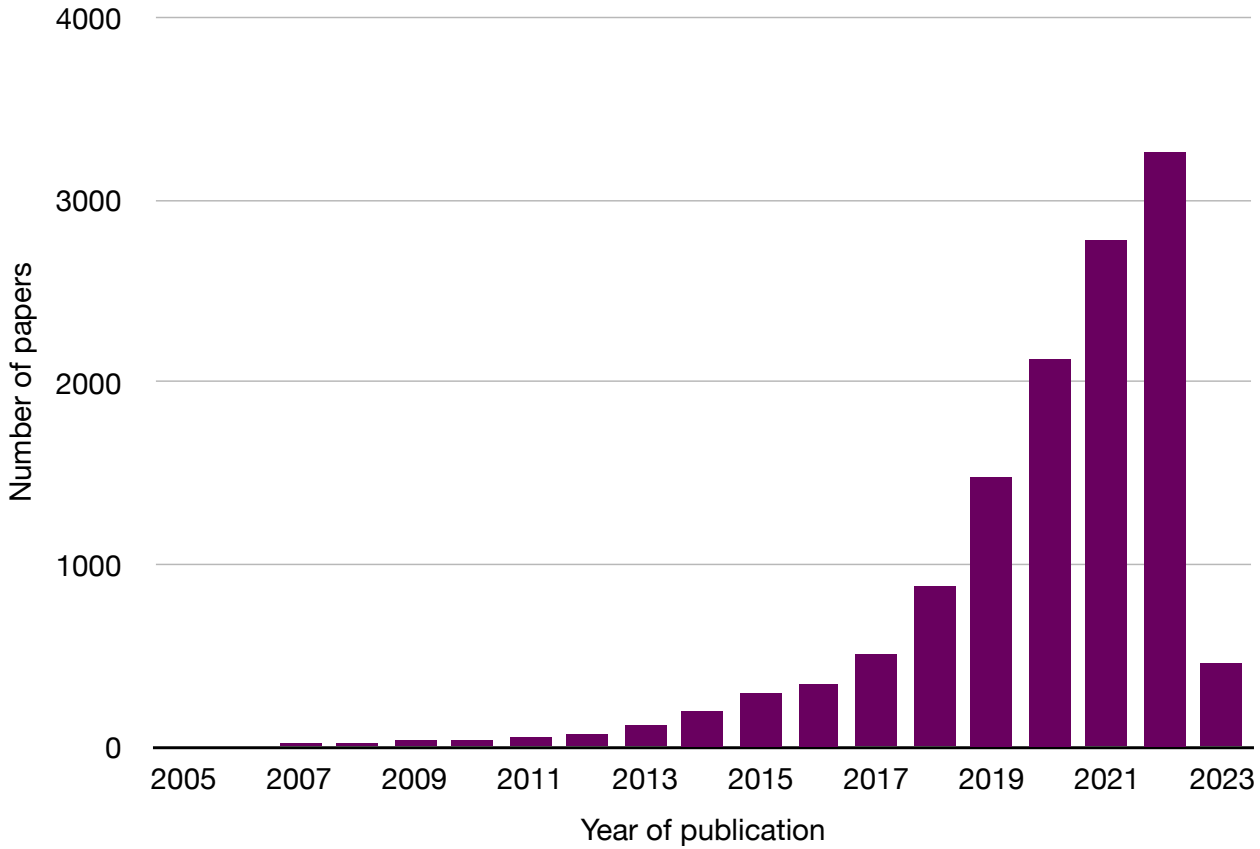
Clinical approval

- Lengthy process
- Untransparent
- Profit-guided

Computational pathology

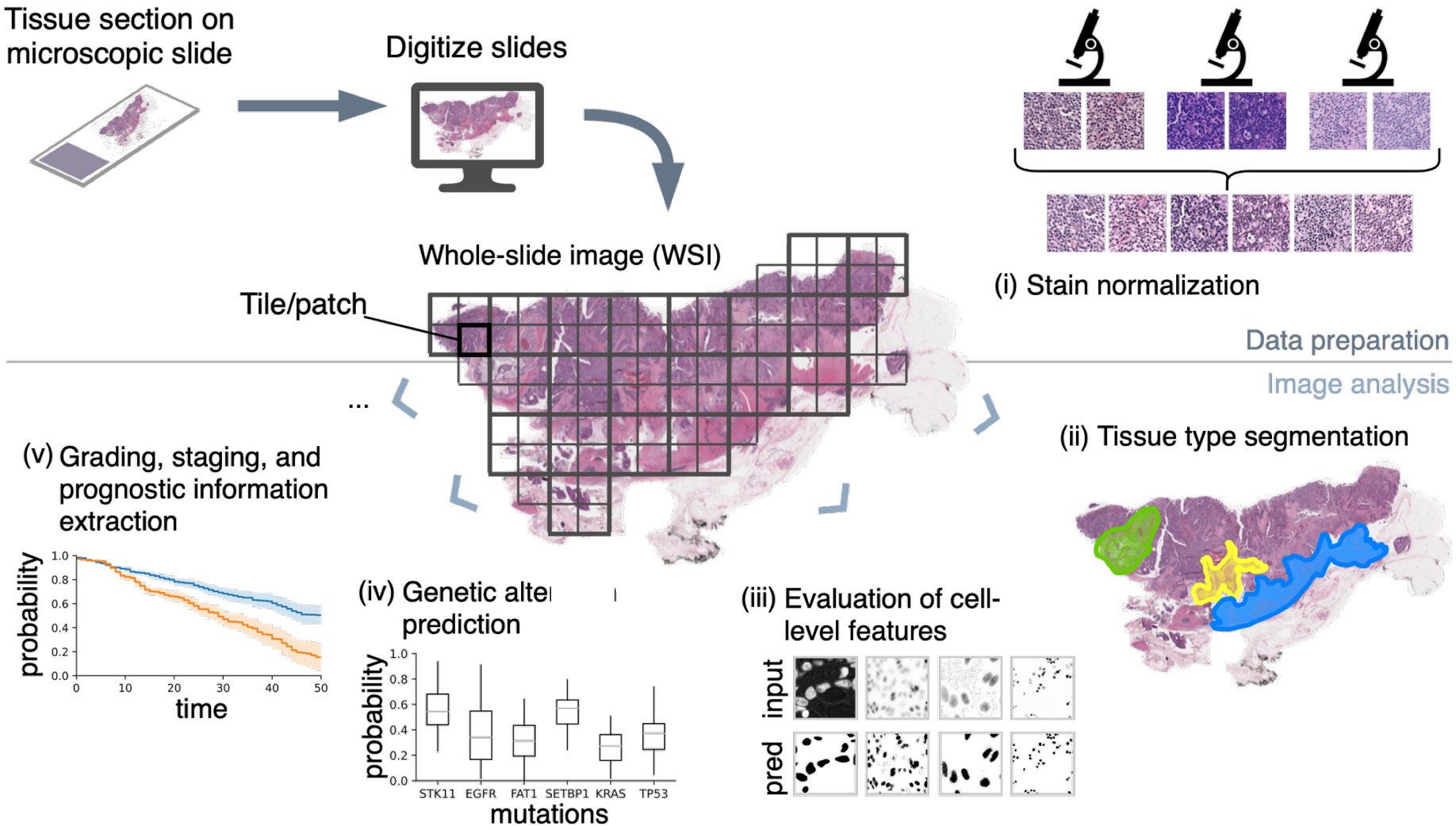


Source: Meskó, B., Görög, M. A short guide for medical professionals in the era of artificial intelligence. npj Digit. Med. 3, 126 (2020). https://doi.org/10.1038/s41746-020-00333-z

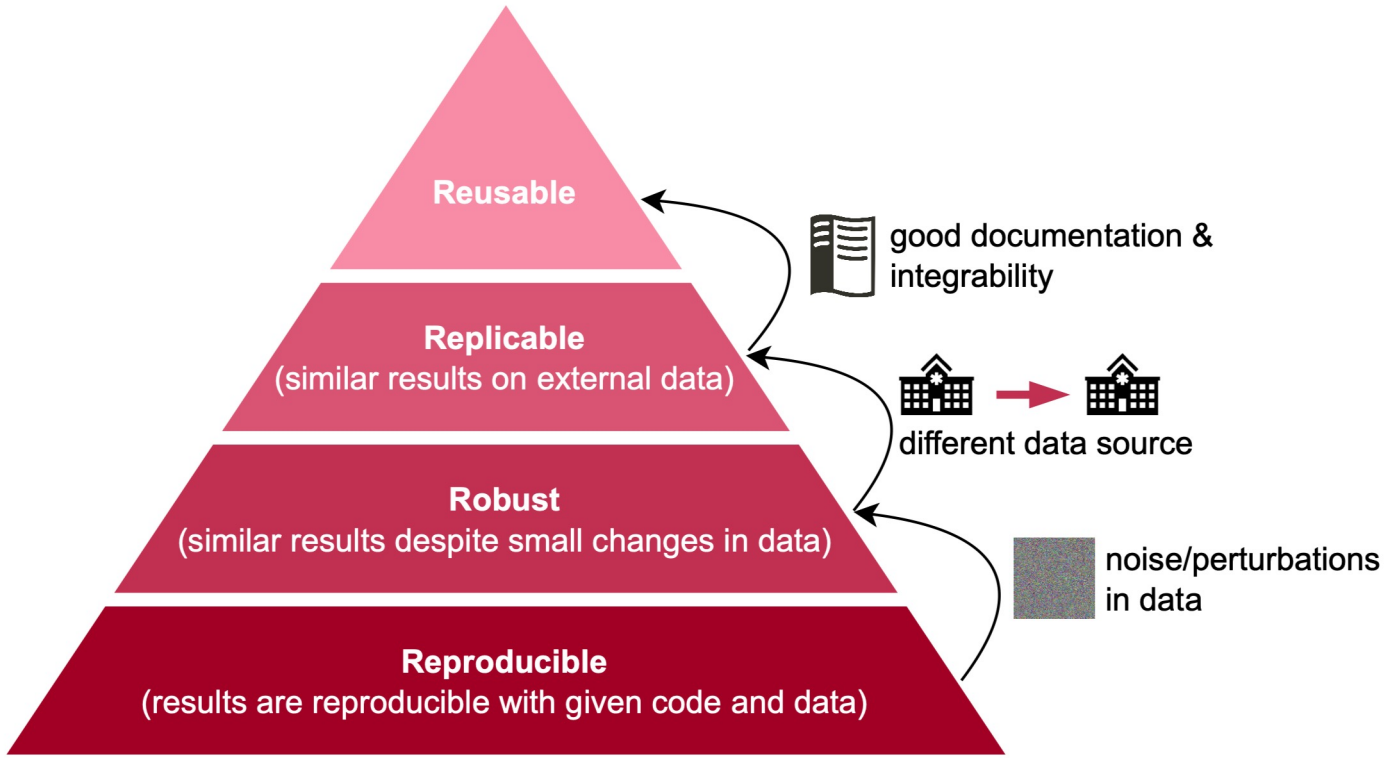


Source: Pubmed

Use cases for AI in computational pathology



Reproducibility and reusability

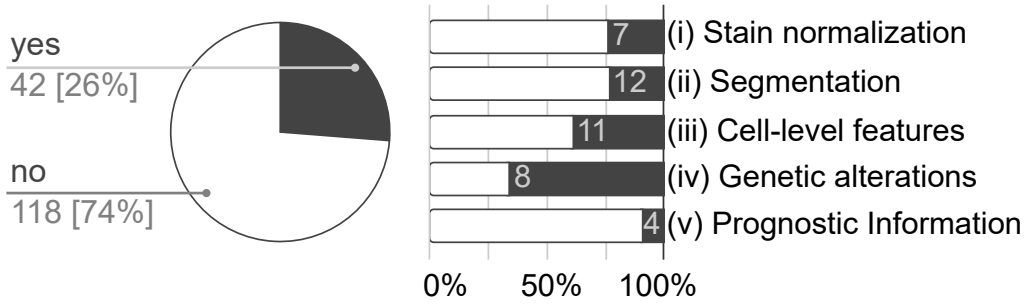


Checklist for
- ✓ Code
- ✓ Data
- ✓ Analysis of statistical variance

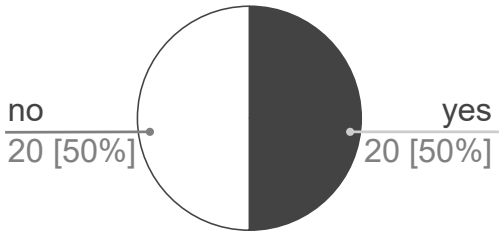
Reproducibility and reusability in CP

160 publications in five use cases

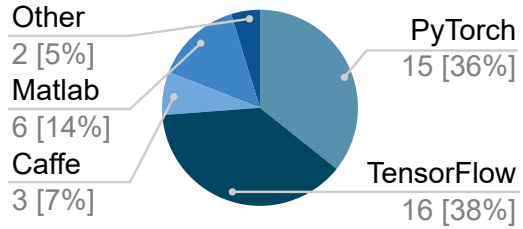
a) Code available?



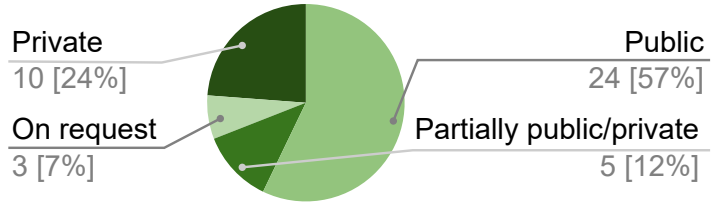
b) Model weights available?



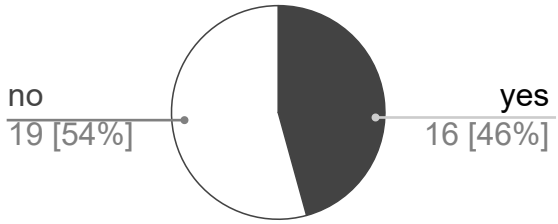
c) Used machine learning frameworks



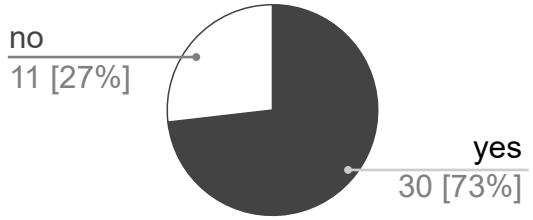
d) Dataset availability?



e) Independent cohort used for evaluation?



f) Variance reported?



Reproducing three publications in CP

- Top-3-performing algorithms of Camelyon17 Challenge
- Reimplemented methods with all given information
- Key technical methods are well described
- No standardized reporting of data pre-processing
- None of the reimplemented algorithms achieved performance close to the performance in the challenge

Compiled checklist for reproducibility:

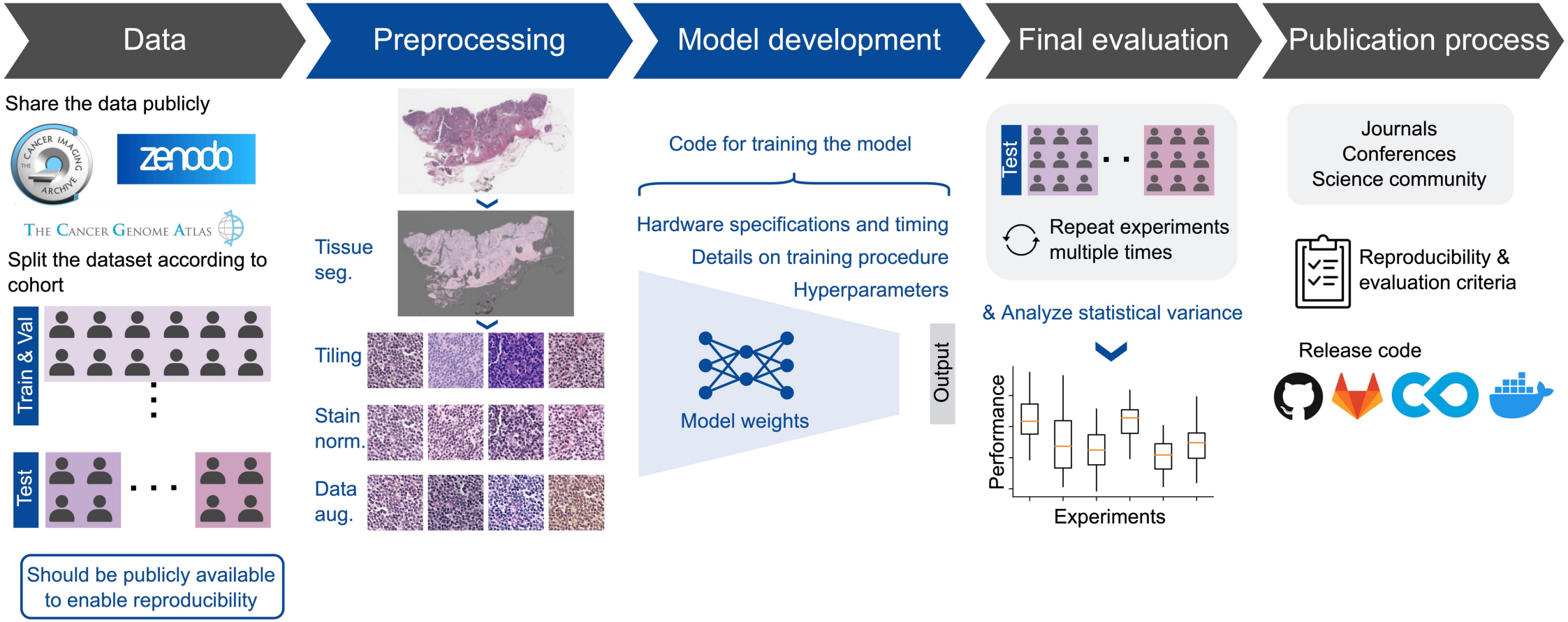
In order to make your work independently reproducible, make sure you have reported all the required details of the following:

1. The hardware and software platform the system was trained and tested on.
2. The source of data and how it can be accessed.
3. How the data was split into train, validation, and testing sets.
4. How or if the slides were normalised.
5. How the background and any artefacts were removed from the slides.
6. How patches were extracted from the image and any data augmentation that was applied.
7. How the patches were labelled.
8. How the patch classifier was trained, including technique, architecture, and hyper-parameters.
9. How the slide classifier was trained, including, pre-processing, technique, architecture, and hyper-parameters.
10. How lesion detection was performed.
11. How the patient classifier was trained, including, pre-processing, technique, architecture, and hyper-parameters.
12. All metrics that are relevant to the all the tasks.

Fell C, Mohammadi M, Morrison D, Arandjelovic O, Caie P, et al. (2022) Reproducibility of deep learning in digital pathology whole slide image analysis. PLOS Digital Health 1(12)

Best Practices and Recommendations

Workflow in Computational Pathology



Wagner, S.J., Matek, C., Shetab Boushehri, S. et al. Nat Med 28, 1744–1746 (2022).

Towards reproducible AI in health care

Paradigm shift towards data sharing

- Multi-institutional datasets
- Diverse datasets
- Standardized metadata

Interdisciplinary collaborations

- Communication of challenges
- AI supporting clinical use
- Software development for end users

Path to the clinic

Code sharing and maintaining

- Publish training details in supplementary
- Use reproducibility checklists
- Collaborate with users