Updates in AI in medicine from the perspective of scientific editing and peer review

The 7th Asian Science Editors' Conference and Workshop (July 12, 2022)

Seong Ho Park, MD, PhD

Professor

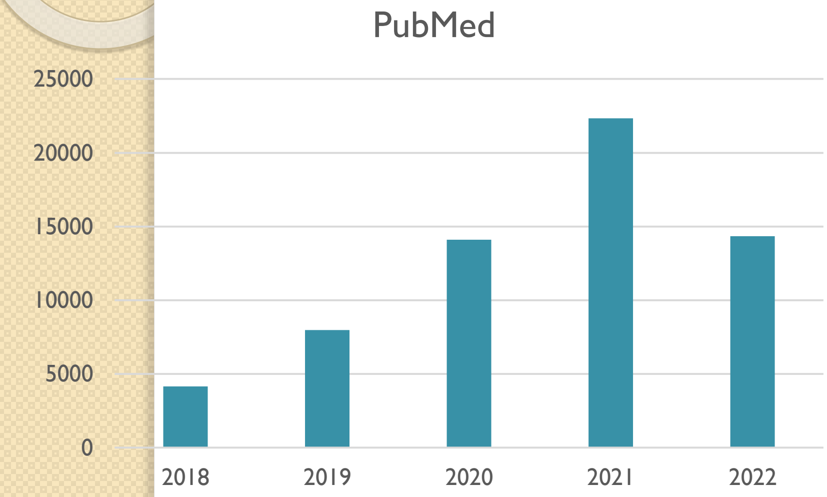
Department of Radiology

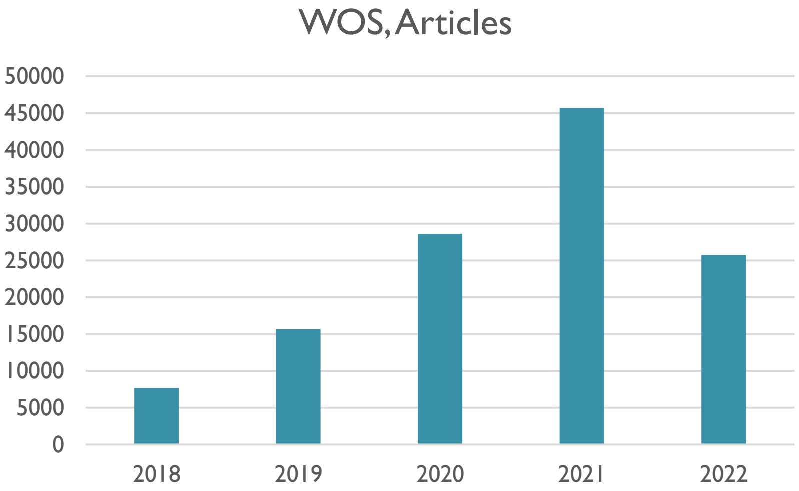
Univ. of Ulsan, Asan Medical Center

Scope

- **AI in Health and Medicine**, not AI in general

AI papers, #/year


PubMed


WOS, Articles

- Search date: June 25, 2022
- Keywords
  - “artificial intelligence” OR “augmented intelligence” OR “deep learning”

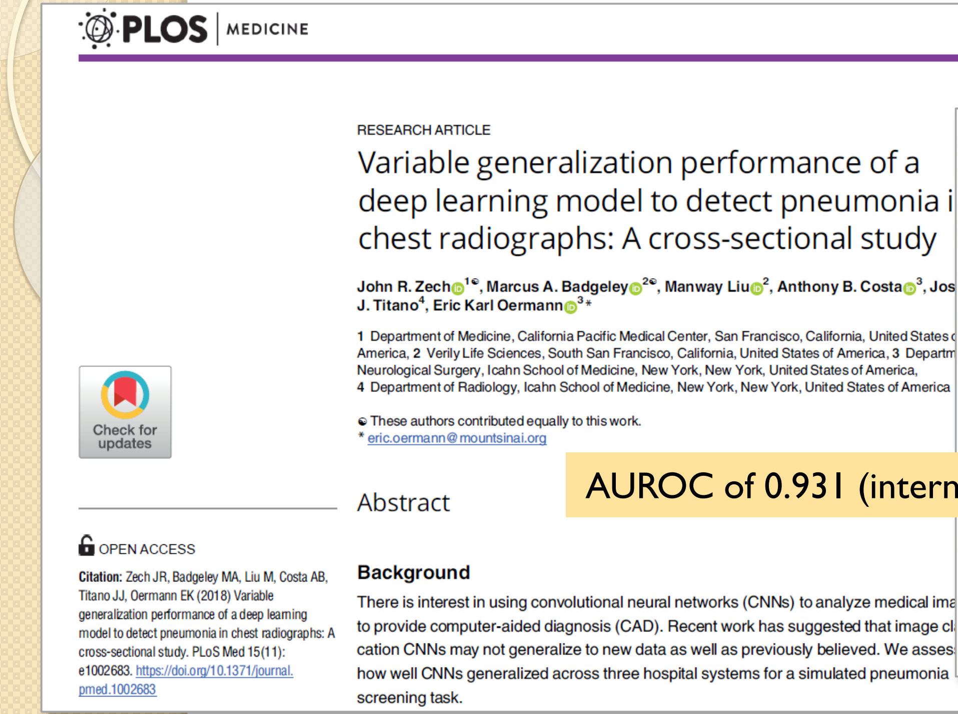
Key issues regarding scientific editing and peer review of AI research

- Priority to external testing and use of unbiased data
- Transparency on the acquisition & nature of data, testing, generalizability, and potential bias
- Algorithm sharing with manuscript submission
- Various reporting guidelines for AI Studies
- Clear use of terminology: validation, overfitting

Two critical characteristics of current data-driven AI (≈ deep learning)

- Data dependency
  - Limited generalizability
  - Bias in, bias out (e.g., biases against historically underserved socioeconomic, ethnic, or gender groups)[1,2]
- Black-box nature

[1] Larrazabal et al. *PNAS* 2020;117(23):12592-12594.
[2] Seyyed-Kalantari et al. *Nat Med* 2021;27(12):2176-2182.



PLOS | MEDICINE

RESEARCH ARTICLE

Variable generalization performance of a deep learning model to detect pneumonia i chest radiographs: A cross-sectional study

John R. Zech[1☯], Marcus A. Badgeley[2☯], Manway Liu[2], Anthony B. Costa[3], Jos J. Titano[4], Eric Karl Oermann[3]*

1 Department of Medicine, California Pacific Medical Center, San Francisco, California, United States o America, 2 Verily Life Sciences, South San Francisco, California, United States of America, 3 Departm Neurological Surgery, Icahn School of Medicine, New York, New York, United States of America, 4 Department of Radiology, Icahn School of Medicine, New York, New York, United States of America

☯ These authors contributed equally to this work.
* eric.oermann@mountsinai.org

Check for updates

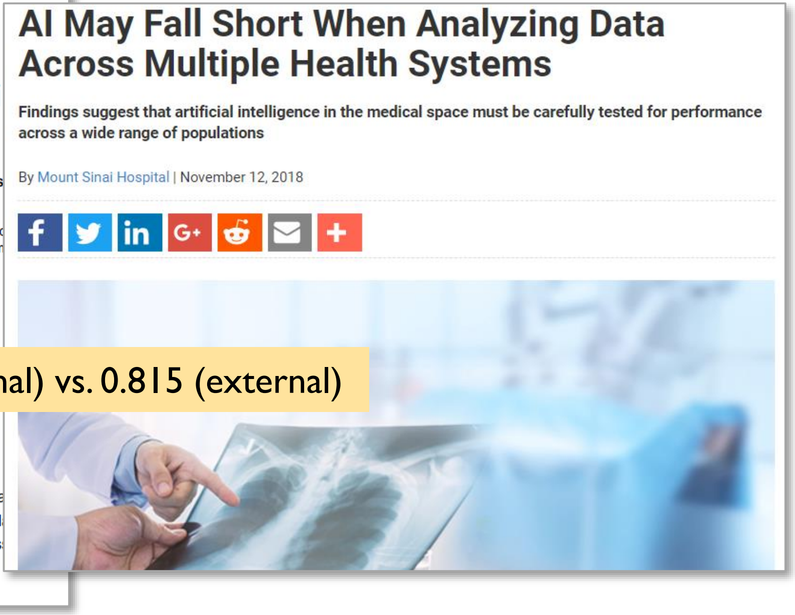
OPEN ACCESS

Citation: Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15(11): e1002683. https://doi.org/10.1371/journal.pmed.1002683

Abstract

Background

There is interest in using convolutional neural networks (CNNs) to analyze medical ima to provide computer-aided diagnosis (CAD). Recent work has suggested that image cl cation CNNs may not generalize to new data as well as previously believed. We asses how well CNNs generalized across three hospital systems for a simulated pneumonia screening task.

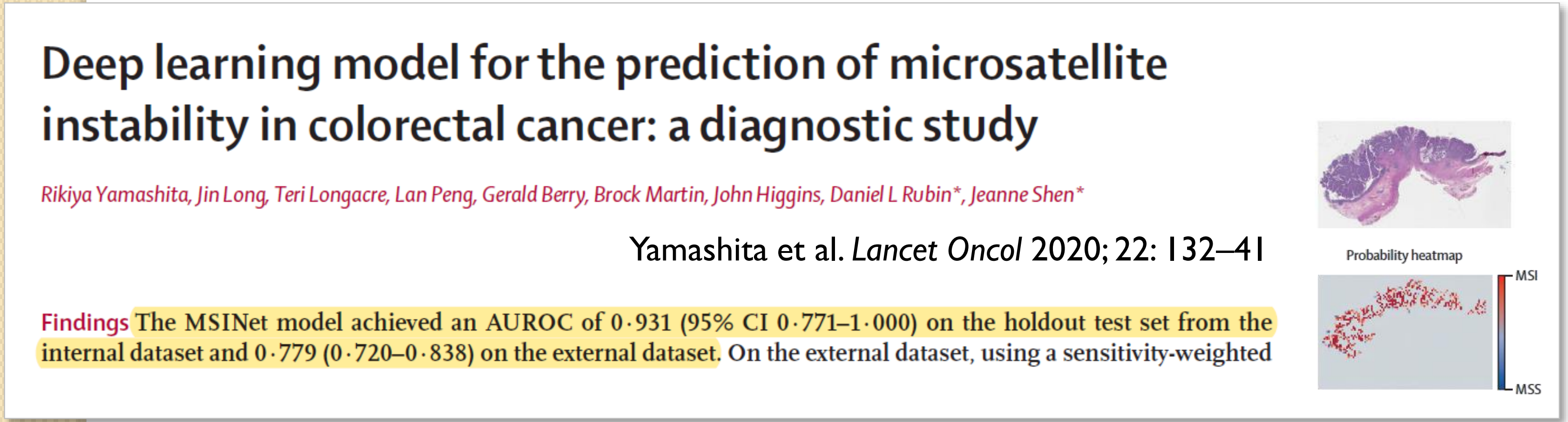


AI May Fall Short When Analyzing Data Across Multiple Health Systems

Findings suggest that artificial intelligence in the medical space must be carefully tested for performance across a wide range of populations

By Mount Sinai Hospital | November 12, 2018

AUROC of 0.931 (internal) vs. 0.815 (external)



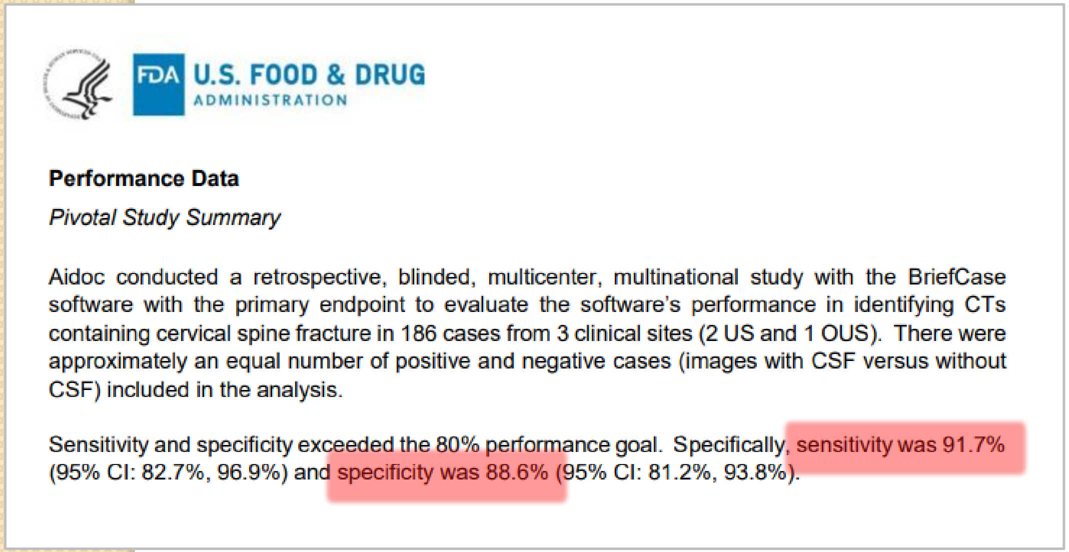
Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study

Rikiya Yamashita, Jin Long, Teri Longacre, Lan Peng, Gerald Berry, Brock Martin, John Higgins, Daniel L Rubin, Jeanne Shen**

Yamashita et al. *Lancet Oncol* 2020; 22: 132–41

Findings The MSINet model achieved an AUROC of 0·931 (95% CI 0·771–1·000) on the holdout test set from the internal dataset and 0·779 (0·720–0·838) on the external dataset. On the external dataset, using a sensitivity-weighted

Voter et al. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures. *AJNR Am J Neuroradiol*. 2021;42(8):1550-1556.



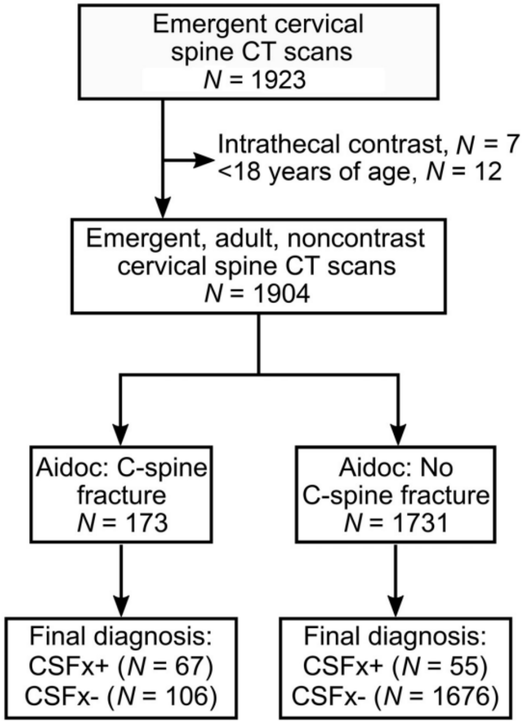
FDA U.S. FOOD & DRUG ADMINISTRATION

Performance Data

Pivotal Study Summary

Aidoc conducted a retrospective, blinded, multicenter, multinational study with the BriefCase software with the primary endpoint to evaluate the software's performance in identifying CTs containing cervical spine fracture in 186 cases from 3 clinical sites (2 US and 1 OUS). There were approximately an equal number of positive and negative cases (images with CSF versus without CSF) included in the analysis.

Sensitivity and specificity exceeded the 80% performance goal. Specifically, sensitivity was 91.7% (95% CI: 82.7%, 96.9%) and specificity was 88.6% (95% CI: 81.2%, 93.8%).



Emergent cervical spine CT scans *N* = 1923
Intrathecal contrast, *N* = 7
<18 years of age, *N* = 12
Emergent, adult, noncontrast cervical spine CT scans *N* = 1904
Aidoc: C-spine fracture *N* = 173
Aidoc: No C-spine fracture *N* = 1731
Final diagnosis: CSFx+ (*N* = 67) CSFx- (*N* = 106)
Final diagnosis: CSFx+ (*N* = 55) CSFx- (*N* = 1676)

sensitivity, 54.9%
specificity, 94.1%

Limited generalizability (loosely referred to as 'overfitting')

imprecise term

- Difference in training data and testing data
  - Out-of-distribution data
  - Covariate shift
  - Domain shift
  - Label shift

ML/AI terminology

  - Spectrum effect
  - Prevalence effect

Clinical epi terminology

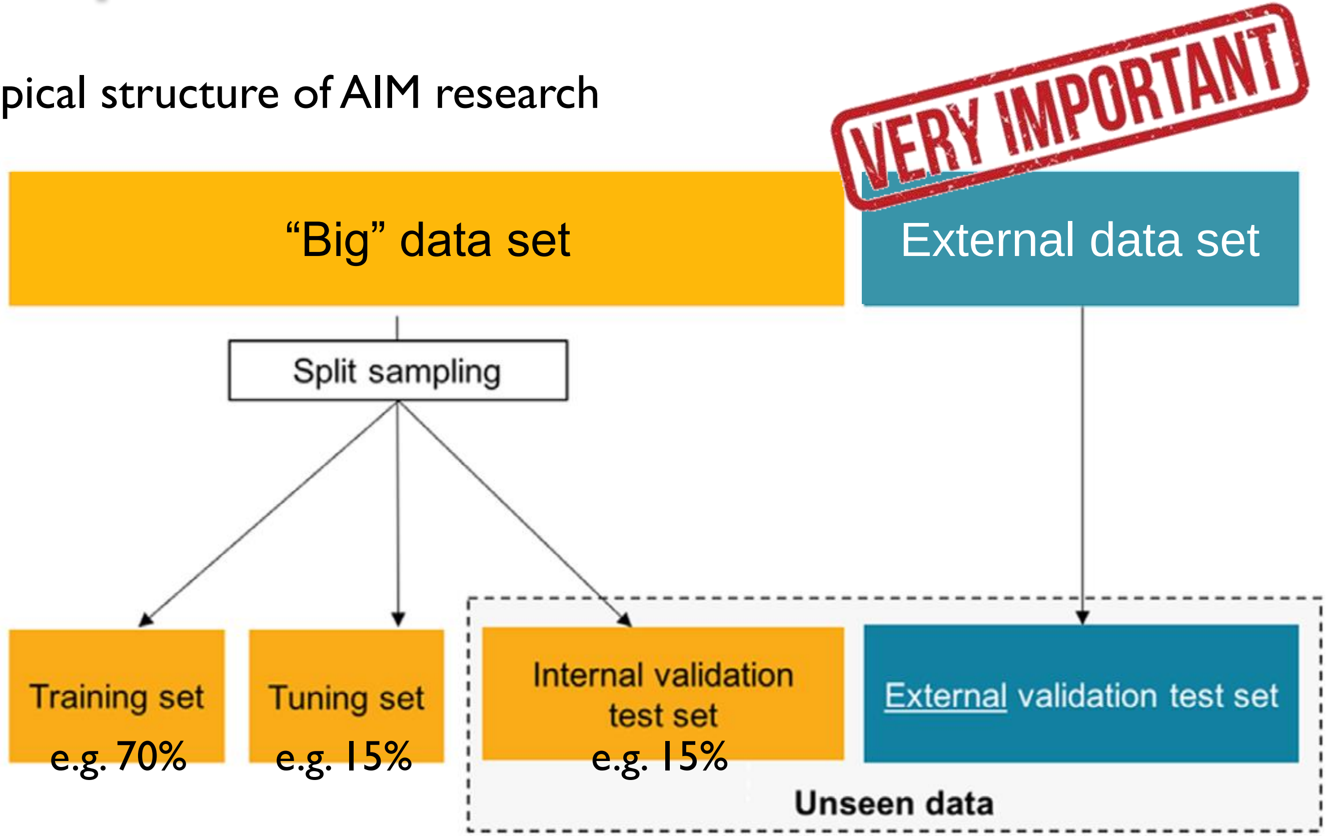
Threats to generalizability in medical data[1]

1. Changes in the practice pattern over time
2. Differences in practice between health systems
3. Patient demographic variations
4. Patient genotypic and phenotypic variations
5. Variations in the hardware and software used for data capture
6. Variations in other determinants of health and disease

[1] Futoma et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:e489-e492.

Importance of external testing

Typical structure of AIM research



VERY IMPORTANT

"Big" data set

External data set

Split sampling

Training set
e.g. 70%

Tuning set
e.g. 15%

Internal validation test set
e.g. 15%

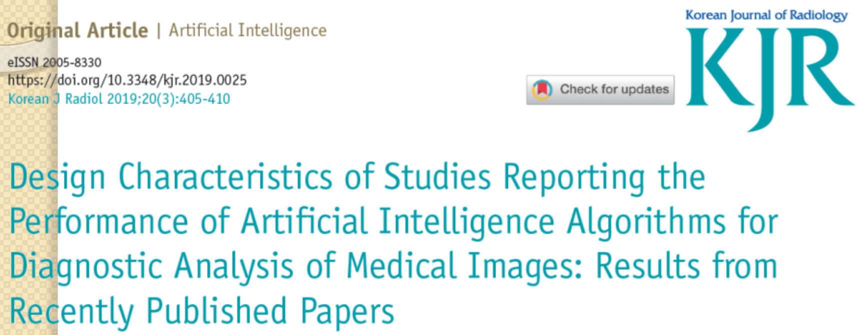
External validation test set

Unseen data

Medical lit.	Training	Tuning	Int. Validation	Ext. Validation
ML/AI lit.	Training	Validation	Int. Testing	Ext. Testing

Faes et al. *Transl Vis Sci Technol* 2020;9(2):7
Kim et al. *PLOS ONE* 2020;15: e0238908

Deficiencies in the published literature



Original Article | Artificial Intelligence

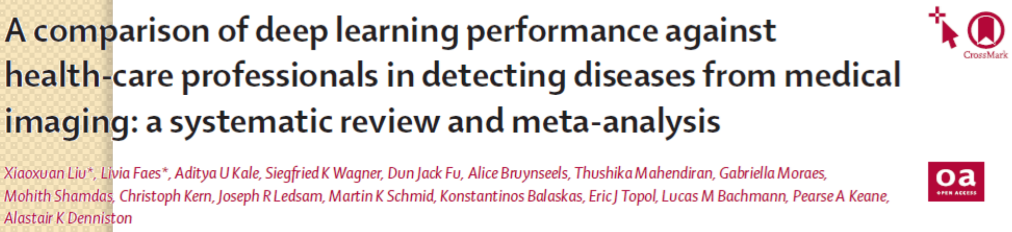
eISSN 2005-8330
https://doi.org/10.3348/kjr.2019.0025
Korean J Radiol 2019;20(3):405-410

Check for updates

Korean Journal of Radiology
KJR

Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers

"Only 6% performed external validation... Nearly all did not have the design features that are recommended..." — *Korean J Radiol* 2019;20(3):405-410



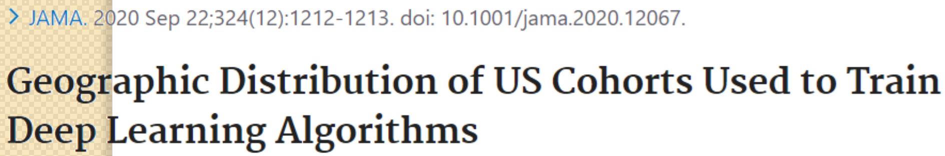
A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Xiaoxuan Liu*, Livia Faes*, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, Alastair K Denniston

CrossMark

oa OPEN ACCESS

"few studies presented externally validated results" — *Lancet Digital Health* 2019;
1: e271–97



› JAMA. 2020 Sep 22;324(12):1212-1213. doi: 10.1001/jama.2020.12067.

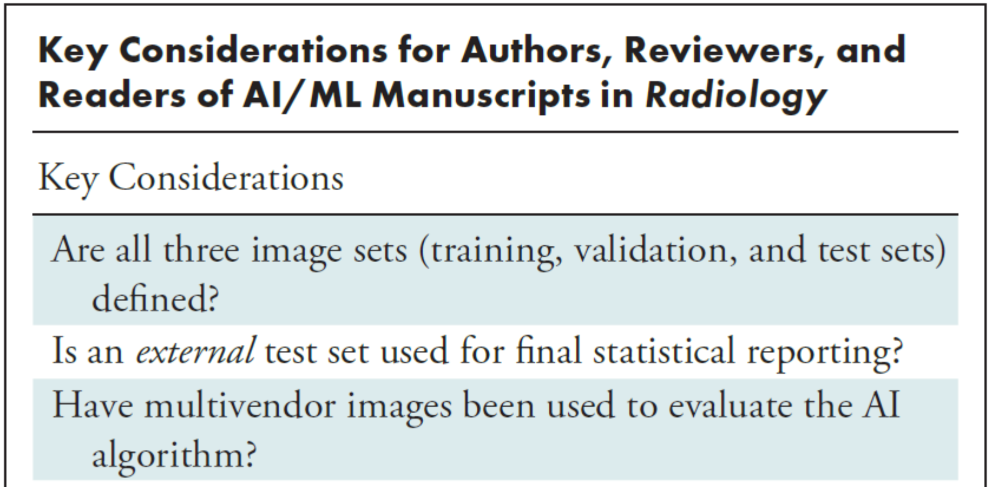
Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms

"algorithms trained on US patient data were disproportionately trained on cohorts from CA, MA, and NY, with little to no representation from the remaining 47 states" — *JAMA* 2020;324(12):1212-1213

Editorial counter measures

- Promotion of external testing and use of unbiased data by giving priority to them
- Request for transparency on the acquisition & nature of data, testing, generalizability, and potential bias

Bluemke et al.
Radiology
2020;294(3):487-489



Key Considerations for Authors, Reviewers, and Readers of AI/ML Manuscripts in *Radiology*

Key Considerations
Are all three image sets (training, validation, and test sets) defined?
Is an *external* test set used for final statistical reporting?
Have multivendor images been used to evaluate the AI algorithm?

Editorial counter measures (cont.): algorithm sharing with manuscript submission

> *"All AI algorithms should be made publicly available via a website such as GitHub. Commercially available algorithms are considered publicly available."*[1]

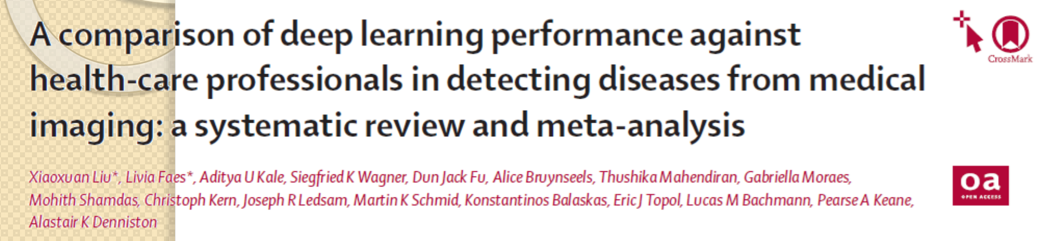
- Promotion of external testing: enabling independent verification of algorithm performance by third parties
- Minimum proof of a real study
  - Dry-bench work using digital big data
  - More opaque "physical" integrity of the study
  - Inability to independently reproduce the study
  - Black-box nature = inability to interrogate

[1] Bluemke et al. *Radiology* 2020;294(3):487-489

Reporting Guidelines for AI Studies

- EQUATOR-related
  - CLAIM (2020)
  - CONSORT-AI (2020)
  - SPIRIT-AI (2020)
  - DECIDE-AI (2022)
  - STARD-AI (pending)
  - TRIPOD-AI (pending)
- Many others

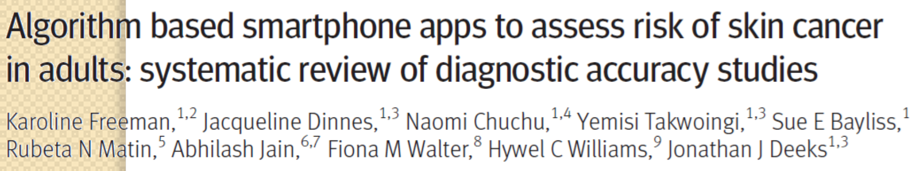
Deficiencies in the published literature



A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Xiaoxuan Liu*, Livia Faes*, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, Alastair K Denniston

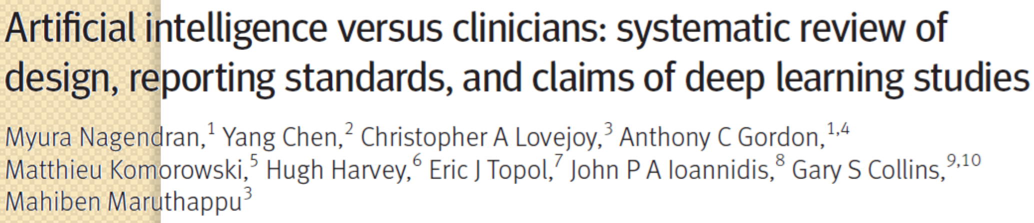
"Poor reporting is prevalent in deep learning studies" — *Lancet Digital Health* 2019;
1: e271–97



Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies

Karoline Freeman,[1,2] Jacqueline Dinnes,[1,3] Naomi Chuchu,[1,4] Yemisi Takwoingi,[1,3] Sue E Bayliss,[1] Rubeta N Matin,[5] Abhilash Jain,[6,7] Fiona M Walter,[8] Hywel C Williams,[9] Jonathan J Deeks[1,3]

"Test performance is likely to be poorer than reported here when used in clinically relevant populations and by the intended users of the apps." — *BMJ* 2020;368:m127

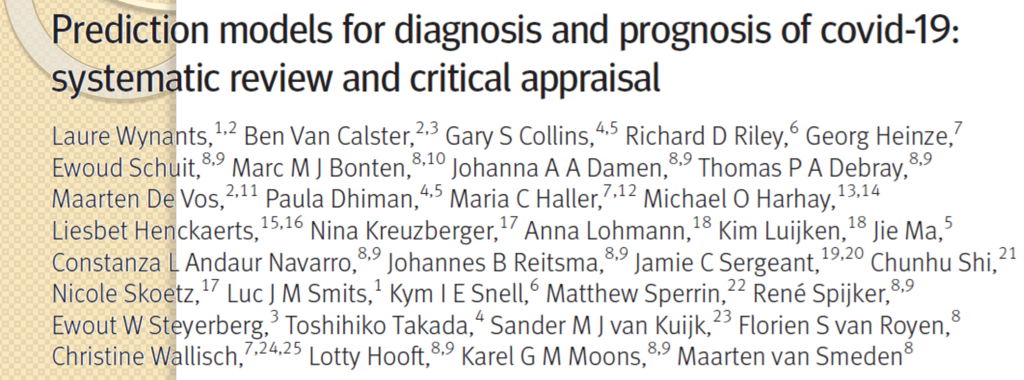


Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies

Myura Nagendran,[1] Yang Chen,[2] Christopher A Lovejoy,[3] Anthony C Gordon,[1,4] Matthieu Komorowski,[5] Hugh Harvey,[6] Eric J Topol,[7] John P A Ioannidis,[8] Gary S Collins,[9,10] Mahiben Maruthappu[3]

Future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions. — *BMJ* 2020;368:m689

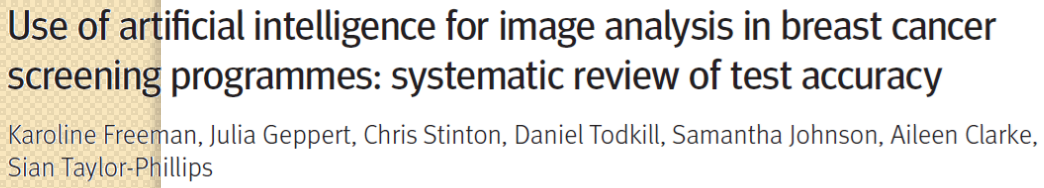
Deficiencies in the published literature



Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Johanna A A Damen,[8,9] Thomas P A Debray,[8,9] Maarten De Vos,[2,11] Paula Dhiman,[4,5] Maria C Haller,[7,12] Michael O Harhay,[13,14] Liesbet Henckaerts,[15,16] Nina Kreuzberger,[17] Anna Lohmann,[18] Kim Luijken,[18] Jie Ma,[5] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[19,20] Chunhu Shi,[21] Nicole Skoetz,[17] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[22] René Spijker,[8,9] Ewout W Steyerberg,[3] Toshihiko Takada,[4] Sander M J van Kuijk,[23] Florien S van Royen,[8] Christine Wallisch,[7,24,25] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

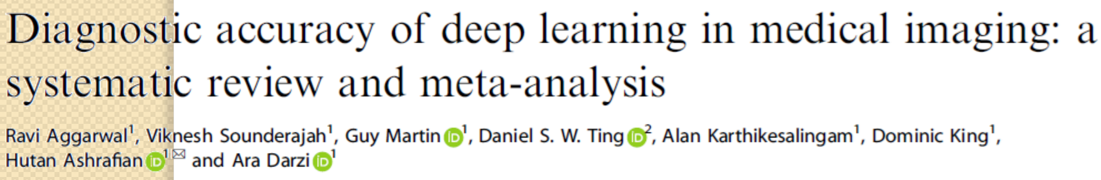
"This review indicates that proposed models are poorly reported, at high risk of bias, and their reported performance is probably optimistic." — *BMJ* 2020;369:m1328



Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy

Karoline Freeman, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, Sian Taylor-Phillips

Current evidence for AI does not yet allow judgement of its accuracy in breast cancer screening programmes — *BMJ* 2021;374:n1872

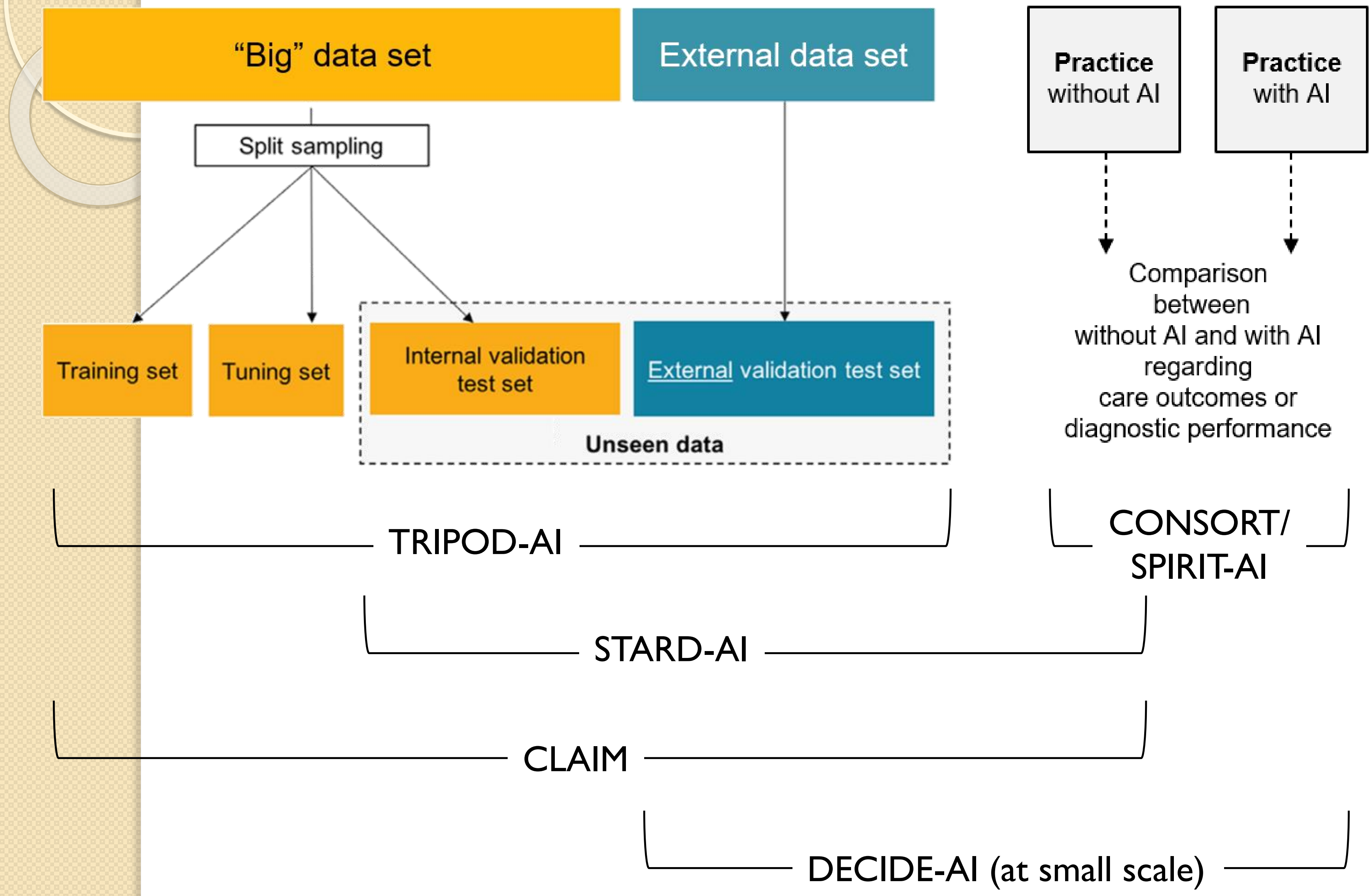


Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis

Ravi Aggarwal[1], Viknesh Sounderajah[1], Guy Martin[1], Daniel S. W. Ting[2], Alan Karthikesalingam[1], Dominic King[1], Hutan Ashrafian[1]✉ and Ara Darzi[1]

Heterogeneity was high between studies and extensive variation in methodology, terminology and outcome measures was noted. — *NPJ Digit Med* 2021;4(1):65.

Scope of the EQUATOR-related Reporting Guidelines*


"Big" data set
External data set
Practice without AI
Practice with AI
Split sampling
Comparison between without AI and with AI regarding care outcomes or diagnostic performance
Training set
Tuning set
Internal validation test set
External validation test set
Unseen data
TRIPOD-AI
CONSORT/ SPIRIT-AI
STARD-AI
CLAIM
DECIDE-AI (at small scale)

*According to the speaker's assessment

Summary: Key issues regarding scientific editing and peer review of AI research

- Priority to external testing and use of unbiased data
- Transparency on the acquisition & nature of data, testing, generalizability, and potential bias
- Algorithm sharing with manuscript submission
- Various reporting guidelines for AI Studies
- Clear use of terminology: validation, overfitting

Thank you for your attention.