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

- 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

---

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*


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

**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%).

**sensitivity, 54.9%**

**specificity, 94.1%**
Limited generalizability (loosely referred to as ‘overfitting’)

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

ML/AI terminology
Clinical epi terminology
Threats to generalizability in medical data

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

Importance of external testing

Typical structure of AIM research

“Big” data set

- Training set e.g. 70%
- Tuning set e.g. 15%
- Internal validation test set e.g. 15%

External data set

- 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

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

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

“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

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

<table>
<thead>
<tr>
<th>Key Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are all three image sets (training, validation, and test sets) defined?</td>
</tr>
<tr>
<td>Is an <em>external</em> test set used for final statistical reporting?</td>
</tr>
<tr>
<td>Have multivendor images been used to evaluate the AI algorithm?</td>
</tr>
</tbody>
</table>

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

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

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

“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

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,2 Matthew Sperrin,22 René Spijker,8,9 Ewout W Steyerberg,2 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 Smeden8

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

Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis
Ravi Aggarwal1, Viknesh Sounderajah1, Guy Martin1,2, Daniel S. W. Ting3, Alan Karthikesalingam1, Dominic King1, Hutan Ashrafian4,5 and Ara Darzi1

“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

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

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*

*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.