AI for Health

Naomi Lee
Senior Executive Editor, The Lancet
Geneva, May 2019

AI for Health

Global health pressures, explosion of digital health data, AI success in other areas

AI a good fit for medicine

But… health is necessarily conservative

Evaluation framework

Comment

WHO and ITU establish benchmarking process for artificial intelligence in health

Growing populations, demographic changes, and a shortage of health practitioners have placed pressures on the health-care sector. In parallel, increasing amounts of digital health data and information have become available. Artificial intelligence (AI) models that learn from these large datasets are in development and have the potential to assist with pattern recognition and classification problems in medicine—for example, early detection, diagnosis, and medical decision making.[1,2] These advances promise to improve health care for patients and provide much-needed support for medical practitioners.

Over the past decade, considerable resources have been allocated to exploring the use of AI for health. Although there is immense potential, many issues such as regulation, potential for bias, and adequate evaluation of efficacy must first be addressed for safe and ethical implementation of AI in health care.[3]

Modern AI algorithms are complex, and their performance depends on the quality of the training data and learning mechanism. If AI algorithms are poorly designed or the training data are biased or incomplete, errors can occur. There is no agreed framework for assessing or reporting the results of health AI models before deciding whether they are sufficiently robust for application in a population, as there is for new drugs or surgical interventions. The absence of confidence or quality control is a major barrier to the uptake of AI in health care. Creating a rigorous, standardised evaluation framework that leverages the advantages and addresses the limitations of AI models in health is crucial for realising the potential of this technology and limiting risks.

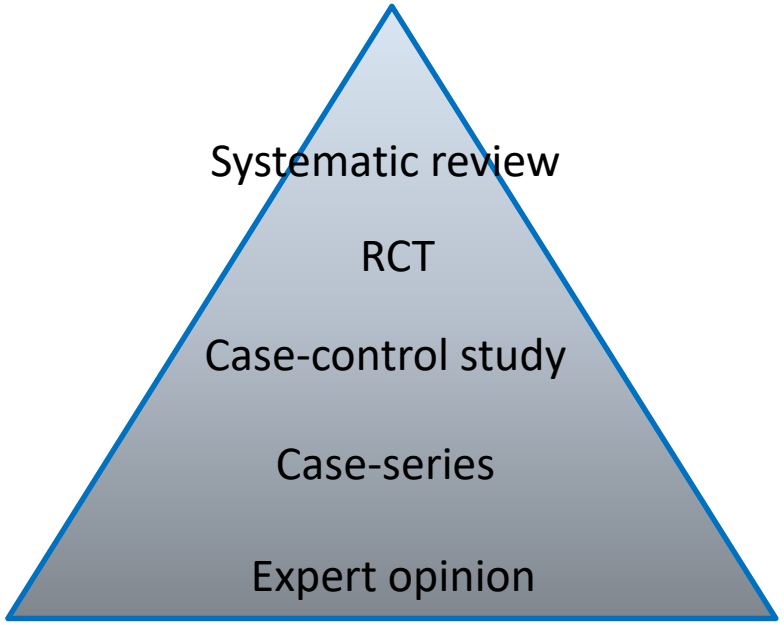
Two UN agencies, WHO and the International Telecommunication Union (ITU), established a Focus Group on Artificial Intelligence for Health (FG-AI4H) in July, 2018. FG-AI4H is developing a benchmarking process for health AI models that can act as an international, independent, standard evaluation framework.

To establish this evaluation and benchmarking process, FG-AI4H is calling for participation from medical, public health, AI, data analytics, and policy experts. Topic groups are being formed by communities of stakeholders allowing FG-AI4H to develop its processes for AI evaluation and benchmarking specific for each health topic. Each topic use case will be reviewed for its relevance and should impact a large and diverse part of the global population or solve a health problem that is difficult or expensive. The AI models are expected to offer improvements over current practices in quality or efficiency that would be expected to lead to better health outcomes or cost-effectiveness. Once formed, topic groups will provide a forum for open collaboration among stakeholders who agree on a pragmatic, best-practice approach for benchmarking each use case, including defining the application scenario and desired output of AI models in that use case, identifying adequate sources of training and testing data, and facilitating the preparation of multisource heterogeneous data. All data for training and testing are expected to be of high quality, ethically generated, and accompanied by detailed information about their format and properties. Thus far, FG-AI4H has developed 11 topic groups in areas such as cardiovascular disease risk prediction, ophthalmology (retinal imaging diagnostics), and AI-based symptom checkers, but this approach is expected to be expanded to other tasks.

The benchmarking process will be done on secure, confidential test data. Ideally, test data will originate from various sources to determine whether the use of an AI model can be generalised across different populations, measurement devices, and health-care settings. The benchmarking process for each use case within a topic needs to be defined. For many use cases it would, at least initially, be meaningful to compare model performance against human performance, or human performance with AI assistance in the same task, whereas for other tasks, comparative performance of algorithms would be more meaningful. Once these requirements are met, AI models can be submitted via an online platform to be evaluated with the test data. Established in this way, the benchmarking process will not only provide a reliable, robust, and independent evaluation system that can demonstrate the quality

Evidence based medicine

- Eminence based medicine
- 1980/1990s
- Medical statistics: the RCT and meta-analysis
- Critical analysis



Stages of phased evaluation

Single intervention should have phased evaluation from pre-clinical to clinical to post marketing surveillance

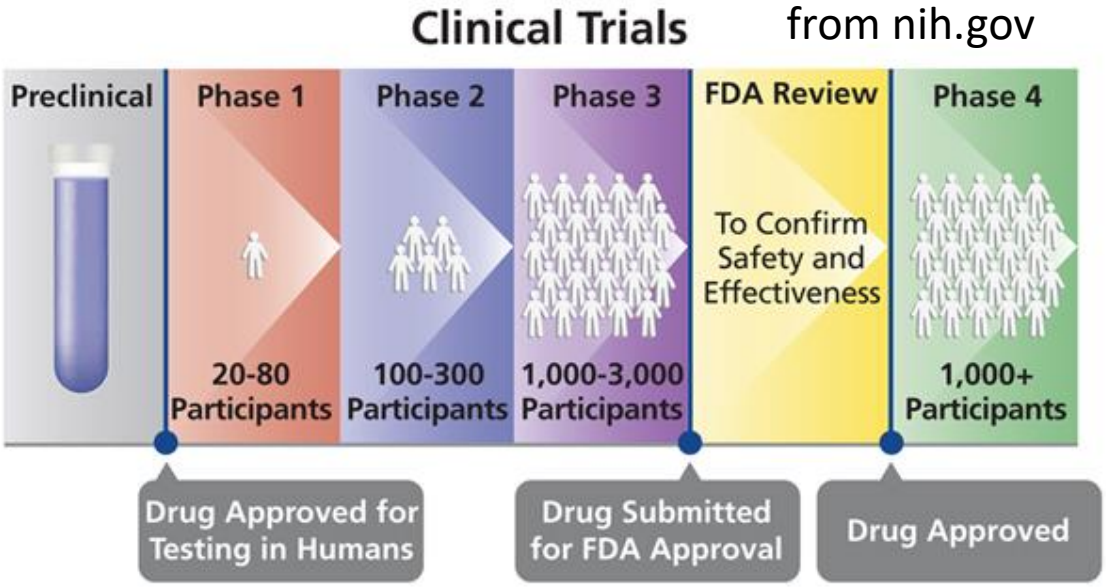


Table Stages of surgical innovation

	1 Idea	2a Development	2b Exploration	3 Assessment	4 Long-term study
Purpose	Proof of concept	Development	Learning	Assessment	Surveillance
Number and types of patients	Single digit; highly selected	Few; selected	Many; may expand to mixed; broadening indication	Many; expanded indications (well defined)	All eligible
Number and types of surgeons	Very few; innovators	Few; innovators and some early adopters	Many; innovators, early adopters, early majority	Many; early majority	All eligible
Output	Description	Description	Measurement; comparison	Comparison; complete information for non-RCT participants	Description; audit, regional variation; quality assurance; risk adjustment
Intervention	Evolving; procedure inception	Evolving; procedure development	Evolving; procedure refinement; community learning	Stable	Stable

The Lancet. VOL 374:9695, P1105-1112, 2009

Quality assurance of evaluation

Helsinki declaration

Good Clinical Practice

Journals:

- EQUATOR NETWORK
- ICMJE/Author guidelines

Regulators/Commissioners:

- Evidence standards framework
- Guidance documents
- Code of Conduct

What should change practice?

- Accuracy of diagnosis/prediction
- Evidence of efficacy
  - Clinically meaningful endpoint
  - Compared again current standard
- Cost effectiveness
- Post market surveillance
- Adoption of poorly evaluated technology causes patient harm and wastes resources

Patient safety in vaginal mesh surgery

The National Institute for Health and Care Excellence (NICE) has published draft guidelines for the clinical management of pelvic organ prolapse and stress urinary incontinence. The guidelines, which are open for public consultation until Nov 19, recommend that women, first and foremost, be offered lifestyle interventions, physical and behavioural therapies, and medication before surgical options are considered. Women who do choose to have surgery must be fully informed of the risks and referred to a specialist. NICE also recommends that all procedures and complications associated with vaginal mesh surgery be tracked on a national database.

dyspareunia, infection, organ perforation, nerve damage, and urinary problems, and, in some cases, women have had to have their implant removed. These complications are not uncommon. Thousands of women have had the vaginal mesh implants in the past decade, so the absolute number of women with adverse reactions is very high.

The guidelines emphasise the need for support and information to guide women through treatment options—a welcome step that should be universal practice. Life-changing complications must be taken seriously; for some women, vaginal mesh surgery will be the best option, but risks of complications must be documented

For **Mesh, graft, or standard repair for women having primary transvaginal anterior or posterior compartment prolapse surgery: two parallel-group, multicentre, randomised, controlled trials (PROSPECT)** see Articles *Lancet* 2017; **389**: 629–40

For the **NICE guidelines draft for consultation** see https://www.nice.org.uk/guidance/GID-NG10035/documents/draft-guideline

1370

Robotic surgery evaluation: 10 years too late

During 2003–13, the number of radical prostatectomies done with the robot-assisted laparoscopic technique increased from about 1·8% to 85% in the USA despite the lack of high level evidence comparing robotic surgery to the standard, cheaper, open technique. In this issue of *The Lancet* John Yaxley and colleagues report the early outcomes of the first randomised trial comparing these two techniques and find no difference in quality of life outcomes at 12 weeks. The final results are awaited with interest. The authors of the Article, and the patients randomised, should be congratulated on a huge achievement in undertaking this long awaited trial. A randomised comparison was thought, by many, to be impossible due to "inherent biases both from a patient and clinician perspective" as Erik Mayer and

different outcomes—of cure or complications—on which to make informed and personal decisions. In medicine, the discomfort of uncertainty, desire to constantly improve, failure to recognise personal biases, and susceptibility to aggressive marketing can lead to innovation being embraced without rigorous evaluation. By doing so, we risk the use of inferior techniques or not providing evidence of benefit and limiting widespread adoption.

In the near future big data, personalised medicine, wearable technology, machine learning, and medical apps all have the potential to play a part to help the health sector reap the potential rewards of the digital revolution. But without health-care workers leading the assessment of these technologies, demanding evidence

patients, additional expenditure

translated to real gains. Robust

novation and the ability to admit

imately drive improvements in

See Comment page 1027

See Articles page 1057

Safety of patient-facing digital symptom checkers

Misdiagnosis by physicians occurs in approximately 5% of outpatients.[1] Computerised diagnostic decision support (CDDS) programmes can help, and interest in this area has increased alongside advances in artificial intelligence and wider availability of clinical data. Originally designed for doctors, CDDS called symptom checkers are designed to directly assist patients by creating differential diagnoses and advising on the need for further care.

The health technology company Babylon recently claimed that their Babylon Diagnostic and Triage System outperformed the average human doctor on a subset of the Royal College of General Practitioners exam.[2] They supported this claim with an internal evaluation study,[3] the results of which were met with scepticism because of methodological concerns.[4,5] In

Triage System. Qualitative assessment of diagnosis appropriateness made by three clinicians exhibited high levels of disagreement. Comparison to historical results from a study by Semigran and colleagues[6,7] produced high scores for the Babylon Diagnostic and Triage System but was potentially biased by unblinded selection of a subset of 30 of 45 test cases. The detailed analysis is shown in the appendix.

Babylon is commended for releasing a fairly detailed description of the system development and the three evaluation studies. This is an important first step in determining its performance and safety. Overall, these results suggest that the Babylon Diagnostic and Triage System potentially showed some improvement compared to the average symptom checkers in the Semigran study.[6] However methodological issues mean that any performance improvement is not

can perform better than doctors in any realistic situation, and there is a possibility that it might perform significantly worse. If this study is the only evidence for the performance of the Babylon Diagnostic and Triage System, then it appears to be early in stage 2 of the STEAD framework (preclinical). Further clinical evaluation is necessary to ensure confidence in patient safety.

Similar concerns with the performance of other CDDS for patients have been reported. Wolf and colleagues[8] showed a high false negative rate in three of four systems designed to detect melanomas from images, which if used in the real world could falsely reassure patients and put their lives at risk. Symptom checkers with significant false negative rates could create similar dangers if used by patients presenting with high risk diseases such as cardiac ischaemia, pulmonary embolism, or meningitis.

These cases highlight the urgent

Published Online November 6, 2018 http://dx.doi.org/10.1016/S0140-6736(18)32819-8

See Online for appendix

Why is AI difficult?

Health and AI communities use different definitions of performance

Medical statistics vs. data science

Association or causation?

How to evaluate a new ability?

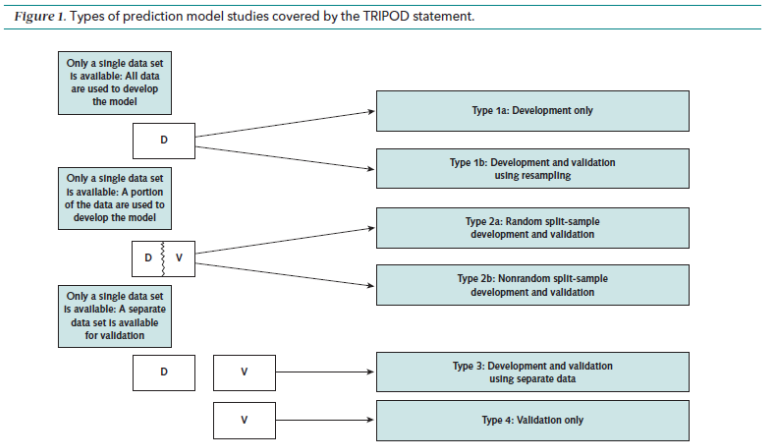
Potential for bias, variable performance

AI models can be adaptive

How necessary is external validation? How transferable are AI models?

Equator Network

- Reporting guidelines for health research
- Transparent reporting of a multivariate prediction model for an Individual Prognosis or Diagnosis (TRIPOD)
- “Gives keys details of how prediction models were developed and validated in order to assess generalizability and risk of bias”
- External validation in a separate dataset



How does this manifest?

Confidence

Published research often doesn't have clinical endpoints, is not externally validated

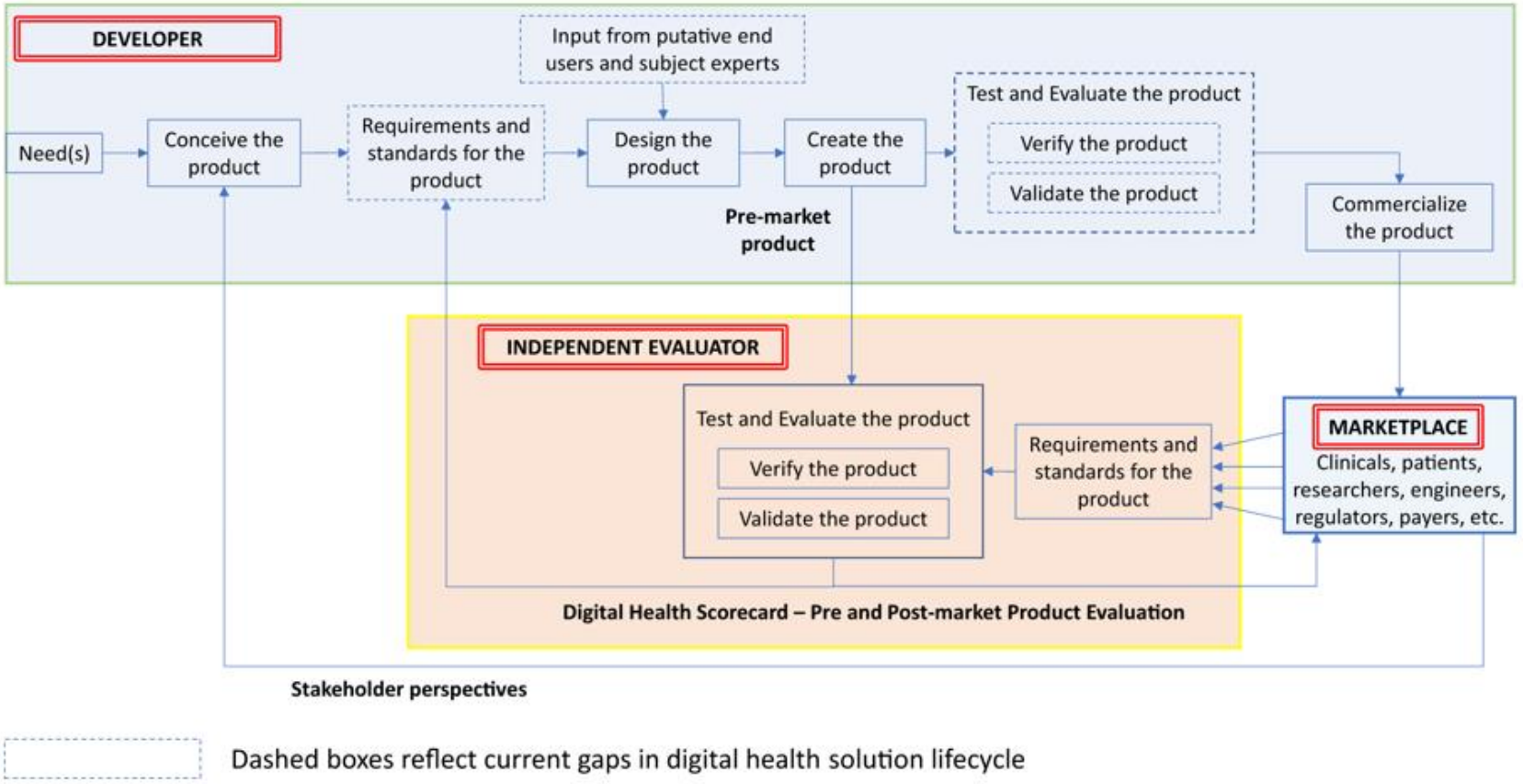
Mismatch between investment and optimism

Slow adoption of AI in health

What should validation look like?

From: Digital health: a path to validation

npj Digital Medicine **volume 2**, Article number: 38 (2019)



What is required for AI?

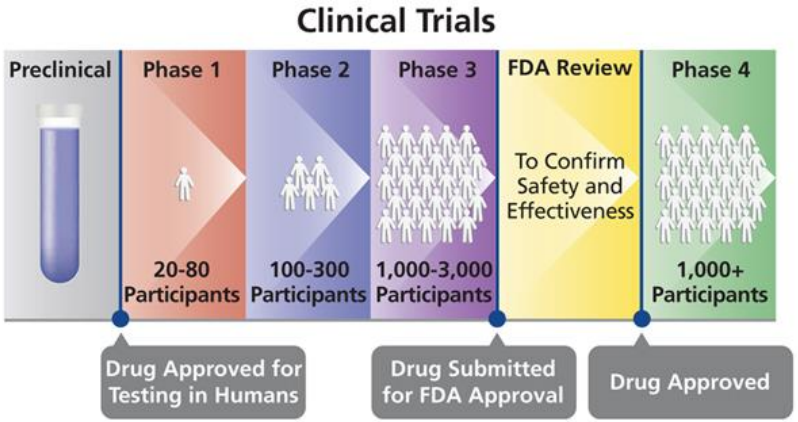
Focus Group will establish benchmarking standard

- Enables validation
- Continuous testing
- International
- Independent
- Comparison with current standard
- Specific to use case

AI for Health
An ITU Focus Group
In partnership with WHO

What is required for AI?

- Community of collaboration
- Focus Group will establish benchmarking standard
- Framework for evaluating
  - Efficacy/cost effectiveness
- Reporting guidelines
- Regulatory framework
- Governance



Thank you.