Evaluation of medical algorithms

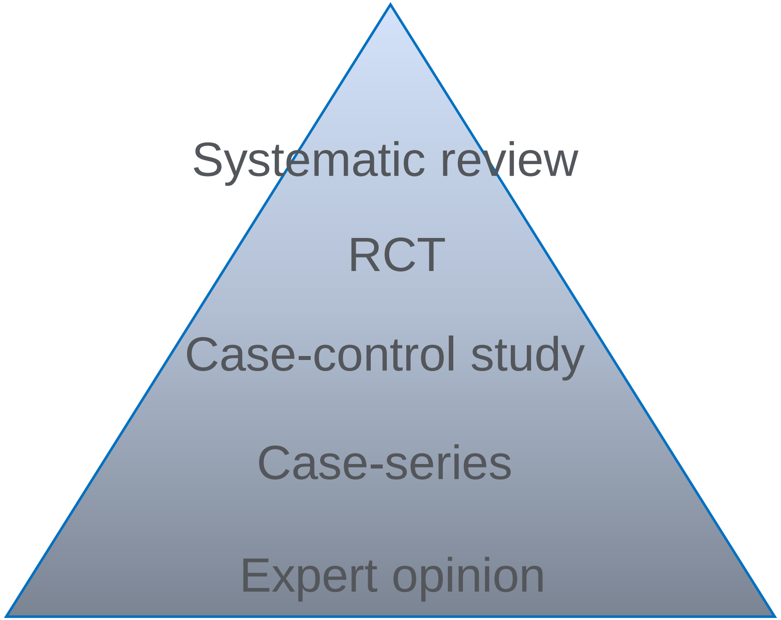
AI for Health, NY

November 2018

Naomi Lee. The Lancet.

Evidence based medicine

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



Systematic review

RCT

Case-control study

Case-series

Expert opinion

Medical journals

Top journals are trusted sources of information:

- Quality assurance
- Peer review
- Standards

- Select practice changing research

What are the standards applied?

ICMJE

ICMJE INTERNATIONAL COMMITTEE *of* MEDICAL JOURNAL EDITORS

Helsinki declaration

EQUATOR NETWORK

equator network

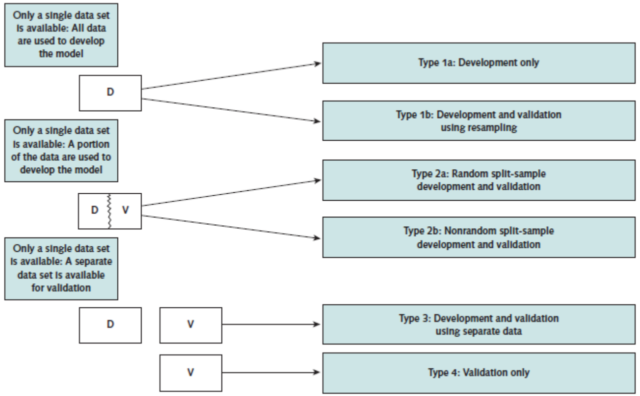
Author guidelines from journals

Information for Authors

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

Figure 1. Types of prediction model studies covered by the TRIPOD statement.


Only a single data set is available: All data are used to develop the model
D
Only a single data set is available: A portion of the data are used to develop the model
D V
Only a single data set is available: A separate data set is available for validation
D
V
V
Type 1a: Development only
Type 1b: Development and validation using resampling
Type 2a: Random split-sample development and validation
Type 2b: Nonrandom split-sample development and validation
Type 3: Development and validation using separate data
Type 4: Validation only

Research

Smartphone-based pathogen diagnosis in urinary sepsis

Research

CHESS1701 trial: radiomics signature for portal hypertension in cirrhosis

Research

Glucocorticoid deficiency reprogrammes glutamine metabolism

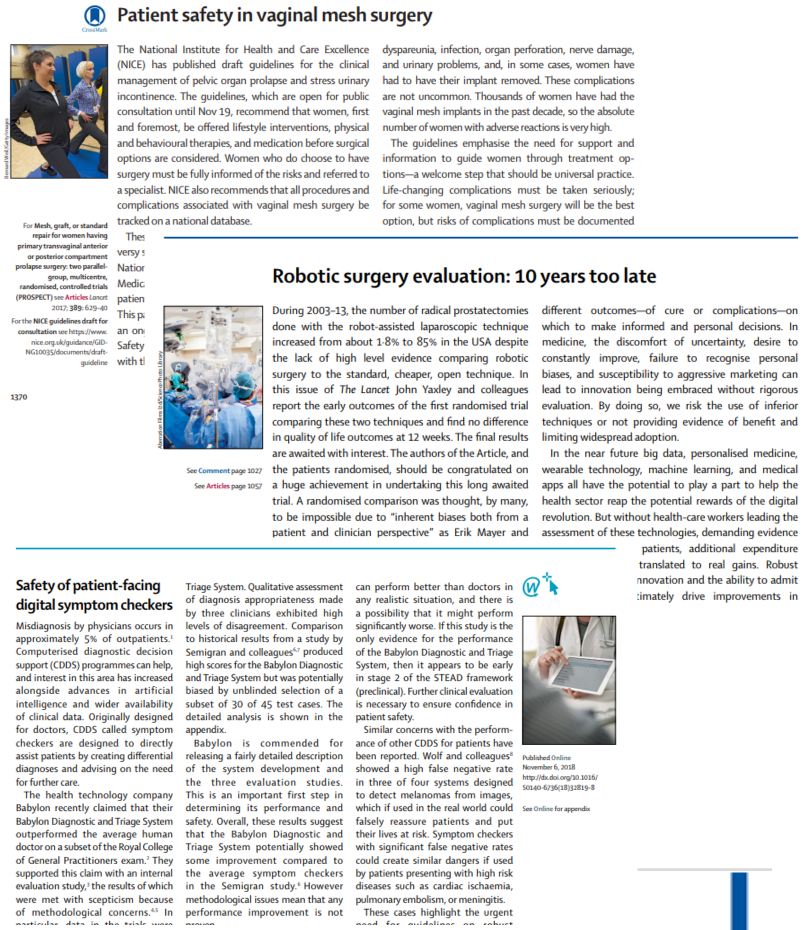
What is ‘practice changing’?

- Accuracy of diagnosis/prediction
- Evidence of efficacy
  - Clinically meaningful endpoint
  - Compared again current standard
- Cost effectiveness

THE LANCET

Why is that a problem?

- Adoption of unassessed technology causes patient harm



Patient safety in vaginal mesh surgery

Bernard Weil/Getty Images

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.

Thes
versy s
Natior
Medic
patien
This p
an on
Safety
with t

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

Aberration Films Ltd/Science Photo Library

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
nnovation and the ability to admit
timately 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 particular, data in the trials were

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 proven

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 need for guidelines on robust

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

See Online for appendix

THE LANCET

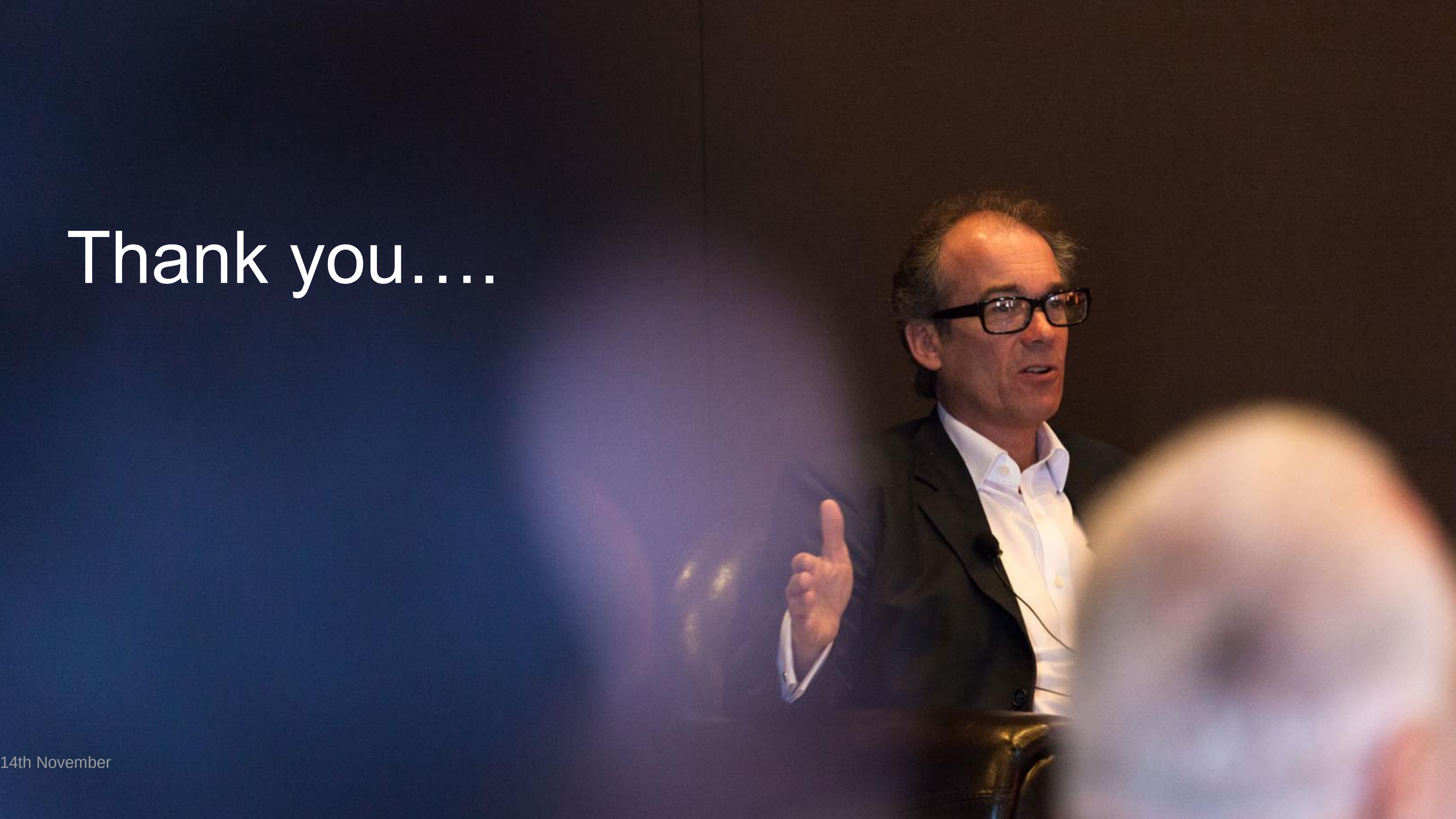
What are the pitfalls for AI?

- A lot of health AI research isn't externally validated
- It doesn't demonstrate clinical efficacy or cost effectiveness

- How do we transition to the mainstream?

- Standards
  - Quality standards
  - Reporting Guidelines

- Framework for assessing efficacy and cost effectiveness



Thank you….