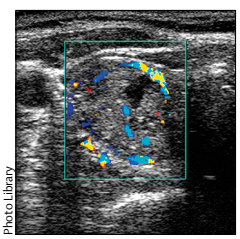
On algorithms, machines, and medicine



Centre Jean Perrin, ISM/Science Photo Library

Published Online
December 21, 2018
http://dx.doi.org/10.1016/S1470-2045(18)30835-0

See Articles page 193

Just as the map is not the territory, so too an algorithm is never the care that is given. Algorithms, neural networks, guidelines, and protocols—these all can only ever model aspects of a reality that is always more complex and fickle.[1] As we move into a world dominated by algorithms and machine-learned clinical approaches, we must deeply understand the difference between what a machine says and what we must do.

A growing number of research papers are reporting impressive diagnostic performance for computer systems built using machine learning. Deep learning techniques in particular are transforming our ability to interpret imaging data.[2] In *The Lancet Oncology*, Xiangchun Li and colleagues[3] report a retrospective preclinical study applying deep learning and statistical methods to diagnose thyroid cancer using sonographic images. Their results are impressive. When compared with six radiologists on unseen data, in an internal validation dataset, the system correctly detected about the same number of cancers (sensitivity 93·4% [95% CI 89·6–96·1] with the algorithm *vs* 96·9% [93·9–98·6] with the radiologists; p=0·003) but had far fewer false-positives (specificity 86·1% [95% CI 81·1–90·2] with the algorithm *vs* 59·4% [53·0–65·6] with the radiologists; p<0·0001).

How generalisable are these results? Training only on patients from one health service or region (in this study, the Tianjin Cancer Hospital, Tianjin, China) runs the risk of overfitting to the training data, resulting in brittle degraded performance in other settings.[4] In this study, although similar machine specificity was achieved on populations from different hospitals, sensitivity dropped to 84·3% (95% CI 73·6–91·9) on an external validation dataset from Jilin and to 84·7% (77·0–90·7) on external validation data from Weihai—substantially below human performance. One might anticipate the system to have weaker performance in non-Chinese populations. One remedy is to retrain the system on patients from new target populations. The problem of biases in training data is, however, foundational,[5] and clinicians must always consider if a machine recommendation is based on data from a population different to their patient.

Automated analysis of cervical smear tests has taught us that computerised image screening is possible, but that it requires solution of many technical and non-technical issues.[6] For example, machine learning might not just learn the intended task, it might inadvertently model clinical workflows or data quality. If these context-specific factors are not replicated when used elsewhere, performance could be poorer. For example, in the study by Li and colleagues, cancer-free images from patients with thyroid cancer were excluded from training. In real-world settings, such images are included, and their presence might distort algorithm performance.

Although this study is preclinical, the authors make commendable efforts to ensure results are as clinically meaningful as possible. Image augmentation was used to artificially distort training data—randomly cropping, scaling, and otherwise distorting images to mimic variations in real-world image quality. Deep learning systems are often criticised because their recommendations come without an explanation, the logic underpinning a diagnosis hidden.[7] In this study, the pixels in an image that most contributed to a diagnosis were highlighted. A clinician could highlight salient parts of an image to help check the computer interpretation.

Yet, focusing just on an algorithm's diagnostic performance or comparison with human beings tells us little about patients' outcomes and might overoptimise what is easy to automate rather than what is important.[8] Decision support must be embedded in a clinical workflow and is but one part of a web of actions and decisions that lead to patients' care. In the case of thyroid cancer, ultrasound is one step in a sequence that can lead to biopsy and treatment. In view of concerns that thyroid cancer is both overdiagnosed and overtreated,[9] improved ultrasound detection might deliver little benefit in terms of patients' outcomes. For example, South Korea has seen a 15-fold increase in thyroid cancer, attributable largely to overdiagnosis,[10] and any diagnostic method that detects more indolent than consequential disease would most likely exacerbate this situation. Certainly, precise automated identification of true negative sonograms might improve a clinician's confidence to do nothing. For this reason, rather than only comparing human to machine, it is more clinically meaningful to measure

the performance of human beings assisted by machine. Such measurements must ultimately take place in clinical trials, recording false-negative identifications and undertreatment as well as overtreatment. Indeed, there is a case that the most pressing decision-support need in thyroid cancer is not in diagnosis but in making the decision to treat.

Thus, excellence in algorithmic performance is essential in our quest for automation, but ultimately we are interested in what a human being decides when using automation in the messy reality of health care. Until our machines are fully embedded in that reality, and see it better than us, our role as clinicians is to be the bridge between machine and decision. At least for now, algorithms do not treat patients, health systems do.

Enrico Coiera
Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia
enrico.coiera@mq.edu.au

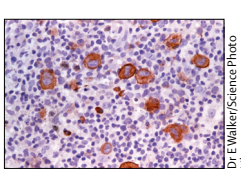
I declare no competing interests. My research is supported by the National Health and Medical Research Council (grant APP1134919, to the Centre for Research Excellence in Digital Health).

1 Coiera E. Basic concepts in informatics: models. In: Coiera E, ed. Guide to health informatics. Boca Raton: CRC Press, 2015: 3–12.
2 Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *Radiographics* 2017; **37:** 2113–31.
3 Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol* 2018; published online Dec 21. http://dx.doi.org/10.1016/S1470-2045(18)30762-9.
4 Chen JH, Asch SM. Machine learning and prediction in medicine: beyond the peak of inflated expectations. *N Engl J Med* 2017; **376:** 2507–09.
5 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; **318:** 517–18.
6 Bengtsson E, Malm P. Screening for cervical cancer using automated analysis of PAP-smears. *Comput Math Methods Med* 2014; **2014:** 842037.
7 Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? Dec 28, 2017. https://arxiv.org/pdf/1712.09923.pdf (accessed Nov 27, 2018).
8 Coiera E. The fate of medicine in the time of AI. *Lancet* 2018; **392:** 2331–32.
9 Ahn HS, Kim HJ, Welch HG. Korea's thyroid-cancer "epidemic": screening and overdiagnosis. *N Engl J Med* 2014; **371:** 1765–67.
10 Furuya-Kanamori L, Bell KJL, Clark J, Glasziou P, Doi SAR. Prevalence of differentiated thyroid cancer in autopsy studies over six decades: a meta-analysis. *J Clin Oncol* 2016; **34:** 3672–79.

Optimisation of adaptive therapy for advanced Hodgkin lymphoma

Management of advanced-stage Hodgkin lymphoma balances tumour control with therapy toxicity. Multiple studies have established the importance of increased-dose bleomycin, etoposide, doxorubicin, cyclophosphamide, vincristine, procarbazine, and prednisone ($BEACOPP_{escalated}$) in providing long-term tumour control for patients with this disease, but this treatment is associated with extended haematological toxicity in normal tissue, including myelodysplasia and secondary malignancies. Some physicians are cautious about giving $BEACOPP_{escalated}$ to patients partly because of concerns that toxicity subtracts from the potential advantage of upfront tumour control. Conventional treatment with doxorubicin, bleomycin, vinblastine, and dacarbazine (ABVD) with the option of dose escalation in patients who respond minimally to upfront therapy is often preferred. Other physicians, however, argue that more aggressive upfront therapy with the option for de-escalation could lead to optimum outcomes for patients.[1–4]

Metabolic imaging with PET has been incorporated into multiple clinical trials of Hodgkin lymphoma with patients of all stages and risk factors. PET with fluorodeoxyglucose has been used to provide uniform thresholds for staging of disease and assessment of response to treatment. The Children's Oncology Group (COG) AHOD0031 trial[4] showed that imaging, including metabolic imaging, could be assessed by central review in real time to make treatment decisions based on response to induction therapy with ABVD. If PET after two cycles of ABVD did not show complete response, chemotherapy was intensified with dexamethasone, etoposide, cytarabine, and cisplatin. Although the COG AHOD0031 study showed that therapeutic titration was advantageous for patients with good responses, patients who needed dose intensification after two cycles of chemotherapy did not achieve the optimum outcome of disease-free survival, which suggests that a more intense therapy at the start of treatment could have been beneficial.


Dr E Walker/Science Photo Library

Published **Online**
January 15, 2019
http://dx.doi.org/10.1016/S1470-2045(19)30005-1

See **Articles** page 202