Human versus machine in medicine: can scientific literature answer the question?

In *The Lancet Digital Health*, Xiaoxuan Liu and colleagues present a systematic review and meta-analysis in an attempt to answer the question of whether deep learning is better than human health-care professionals across all imaging domains of medicine. Despite the plethora of headlines proclaiming how the latest artificial intelligence (AI) has outperformed a human physician, the authors found surprisingly few studies that compare the performance of humans and these models. From more than 20,000 unique abstracts, fewer than 100 studies met their eligibility criteria for the systematic review and only 25 met their inclusion criteria for the meta-analysis. These 25 studies compared the performance of deep learning solutions to health-care professionals for 13 different specialty areas, only two of which—breast cancer and dermatological cancers—were represented by more than three studies. The meta-analysis suggests equivalent performance of deep learning algorithms and health-care professionals in the 14 studies that used the same out-of-sample validation dataset to compare their performances, showing a pooled sensitivity of 87·0% (95% CI 83·0–90·2) for deep learning models and 86·4% (79·9–91·0) for health-care professionals, and a pooled specificity of 92·5% (85·1–96·4) for deep learning models and 90·5% (80·6–95·7) for health-care professionals. This work nicely illustrates the challenge of attempting to compare AI with humans for medical applications, and the authors rightly qualify their conclusion with a detailed list of potential confounders and limitations. The eventual sample size representing a broad swath of the domain of medicine underlines the need for a deeper dive into the literature.

Evaluation of diagnostic accuracy—whether for AI or otherwise—requires a ground truth. In the absence of a perfect ground truth, inherent biases are introduced into a study. This is particularly problematic when evaluating AI tools to decide if they perform better than humans. As Liu and colleagues point out, there is a wide spectrum of what constitutes expert consensus or ground truth in the literature, yet these datasets with inconsistent, imperfect, or even incorrect labels become training and testing data for AI models. If researchers cannot all agree on what it means to agree, how can we know if model A is better than human B? More importantly, how can an AI model be trained when experts themselves disagree on the correct answer to a question?

AI cannot yet replicate the essence of the diagnostic process. In medicine, different datapoints become available at different times during a work-up. One test might be ordered because of the result of another. So, when AI algorithms are trained on a complete corpus of retrospective data that eliminates both the temporal variation and the dependency within the data, can it actually be compared with the human physician who made a series of related decisions to create that comprehensive dataset? Additionally, formulation of a differential diagnosis often gets tossed aside when training an AI algorithm, because the focus shifts to making a single diagnosis rather than highlighting the relevant data that lead a physician to a particular set of diagnoses with associated likelihoods—ie, the differential diagnosis.

Can AI developed and trained in silico effectively be compared with the human physician functioning in the real world, where data are messy, elusive, and imperfect? As Liu and colleagues discovered, only a handful of studies evaluate the performance of AI models in the presence of a-priori knowledge about the patient. This is contrary to medical practice, which begins with the history of the present illness, the review of systems, and the physical exam, and uses these data to determine what diagnostic testing is needed. The presence or absence of these data introduces additional bias, either towards a broader differential or a specific diagnosis. Again, there is also the question of diagnosis versus detection—if an AI is trained on a limited, curated dataset, while the human diagnostician technically has the entire medical record at their fingertips, what task is the AI actually performing? And when the human physician is asked to make a diagnosis with a fraction of the data they would normally use (eg, “I cannot tell you if the patient has a productive cough, shortness of breath, fever, or is immunocompromised, but is that opacity on the chest radiograph pneumonia?”), are we actually evaluating the human physician’s true performance?
The scientific literature is known to be incomplete because negative studies—ie, those that do not disprove the null hypothesis—are less frequently published. This adds to the complexity of evaluating the performance of AI compared with human physicians, because the results might be skewed in favour of those AI models that do perform well. More negative studies and studies that reproduce results need to be added to the existing body of knowledge on AI in medicine to balance and mature the literature. Furthermore, the techniques used to evaluate model performance (confusion matrices and F scores) do not always take into account the reality of medical practice, where the relative costs of false negatives and false positives differ according to disease and scenario.

From the beginning, AI has been criticised because of the black-box nature of the tool: data go in and an answer comes out, with little understanding of what occurs in between. Fortunately, the burgeoning subfield of explainable AI has begun to offer some tools to better interpret AI models. The challenge remains that although explainable AI can tell us what aspects of the input data it used to determine its output, it still cannot tell us why the end result is produced.

Liu and colleagues conclude that the accuracy of deep learning is similar to that of health-care professionals. With the increasing hype of the potential of AI in medicine, this result could be misconstrued as machine diagnosis being better than human diagnosis: why have a human doctor when a digital one would be just as good, maybe better? Given the extensive discussion of the limitations of their study, claiming equivalence or superiority of AI over humans could be premature. Perhaps the better conclusion is that, in the narrow public body of work comparing AI with human physicians, AI is no worse than humans, but the data are sparse and it might be too soon to tell. AI in many forms, including explainable AI and augmented intelligence, continues to climb towards the peak of inflated expectations on the 2019 Gartner Hype Cycle. As scientists and physicians, we should take on a responsible assessment of this new and rapidly developing technology, and stick to the facts, rather than risking a drop into the trough of disillusionment and a third major AI winter.

Tessa S Cook
Department of Radiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
Tessa.Cook@pennmedicine.upenn.edu

I report current or recent grants from the American College of Radiology (ACR), ACR Imaging Network, AUR Strategic Alignment Award, Radiological Society of North America (RSNA), RSNA Education Scholar Award, and Society for Imaging Informatics in Medicine (SIIM); personal fees from SIIM, the Pennsylvania Radiological Society, University of Wisconsin, Cornell University, Thomas Jefferson University, Emory University, and ACR, and other support from ACR, Osher Institute, University of Wisconsin, Cornell University, Thomas Jefferson University, and Emory University, outside of the submitted work.

Copyright © 2019 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

1. Liu X, Faes L, Kale A, et al. A comparison of deep learning performance against health care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digital Health 2019; published online Sept 24. https://doi.org/10.1016/S2589-7500(19)30123-2
2. Challen R, Denney J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019; 28: 231–37.
3. Crawford K, Calo R. There is a blind spot in AI research. Nature 2016; 538: 311–13.
4. Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. Dermatol Pract Concept 2018; 8: 249–51.
5. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnosis of pediatric diseases using artificial intelligence. Nat Med 2019; 25: 333–38.
6. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018; 15: e1–17.
7. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005; 4: e168.
8. Loh E. Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. BMJ Lead 2018; 2: 59–63.
9. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018; 286: 800–09.
10. Kind A, Azopardi G. An explainable AI-based computer aided detection system for diabetic retinopathy using retinal fundus images. In: Vento M, Peronnella G, eds. Computer analysis of images and patterns. Cham: Springer International Publishing, 2019: 457–68.