tvst

Special Issue

Will AI Replace Ophthalmologists?

Edward Korot[1], Siegfried K. Wagner[1], Livia Faes[1,2], Xiaoxuan Liu[3,4], Josef Huemer[1], Daniel Ferraz[1,5], Pearse A. Keane[1], and Konstantinos Balaskas[1,6]

[1] NIHR Biomedical Research Center at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK
[2] Department of Ophthalmology, Cantonal Hospital Lucerne, Lucerne, Switzerland
[3] Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[4] Academic Unit of Ophthalmology, Institute of Inflammation & Ageing, University of Birmingham, Birmingham, UK
[5] Federal University of Sao Paulo, Sao Paulo, Brazil
[6] School of Biological Sciences, University of Manchester, Manchester, UK

Correspondence: Konstantinos Balaskas, NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, 162 City Rd, London EC1V 2PD, UK. e-mail: k.balaskas@nhs.net

Introduction

Teamwork, creativity, adaptability, empathy—all traits that physicians employ on a daily basis to effectively deliver patient care. One may argue that these are elements of physician-patient interaction that artificial intelligence (AI) could never replicate. However, others would contend that AI models have already demonstrated some of these features. Recent notable examples include AI mastering cooperative gameplay and generative adversarial networks creating novel artwork and melodic music.[1–4] These advances were all made possible due to the recent proliferation of deep neural networks, which have also ushered a stepwise improvement in machine learning performance in ophthalmology.[5–8] However, it is crucial to clarify that these and similar AI models that show creativity, teamwork, and adaptability are examples of "narrow" AI. These algorithms are typically validated in constrained testing environments and have limited generalizability. Furthermore, when evaluated outside their test environments in a more abstract fashion or presented with intentional adversarial counterfactuals, they often fail with unfortunate consequences.[9]

AI Challenges

Even considering the aforementioned examples, which mimic certain elements of human behavior, there has not yet been a demonstration of empathy by an AI algorithm. In the context of medicine, empathy comprises not only understanding a patient's feelings but, more important, also responding by delivering care in an appropriate manner. A health professional's relationship with the patient helps guide the patient's care in the context of his or her unique physical, emotional, and social environment. Furthermore, the doctor-patient relationship itself has been shown to have a therapeutic effect in a systematic review of 25 randomized controlled trials.[10] The patient-clinician interaction is innately human and, in the words of patients themselves, depends on "two humans who both can fully contextualise and appreciate the patient's values, wishes, and preferences."[11]

Beyond the human interaction component, translating AI from laboratory experiment to a real-world tool entails additional challenges. "Do no harm," the first line of the Hippocratic Oath, signifies that physicians employing tools such as AI in patient care delivery must maintain safety as the first priority. As Luke Oakden-Rayner,[12] a radiologist and critical AI blogger explains, Silicon Valley's ethos of "move fast break things" can be especially dangerous in the context of medical AI. When AI-assisted medical care transitions from triage to diagnostic systems (Fig. 1), so too the inherent risk increases. In ophthalmology, we currently lie at the "dotted line," as triage systems such as the ones developed by IDX and Google DeepMind are precursors of future diagnostic and predictive systems.[5,13] The Moorfields DeepMind algorithm already has a diagnostic component, and predictive systems are just around the corner. Arcadu et al.[14] describe a system capable of predicting two-step worsening of diabetic patients' Early Treatment of Diabetic Retinopathy

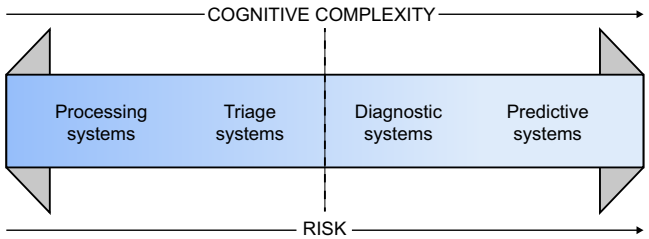

Figure 1. Risks of medical AI. AI model risk increases as cognitive complexity increases. We currently lie at the dotted line, a tipping point between triage and diagnostic systems. Permission received for reproduction from Luke Oakden Rayner.[16]

Study (ETDRS) scores at 12 months with an area under the receiver operating characteristic curve (AUC) of 0.79, and a Google group recently described an algorithm that could predict acute kidney injury 48 hours in advance.[15]

As these AI systems are poised to influence clinical decision making, the risks become more apparent, and prospective validation becomes more important. As Oakden-Rayner[12] suggests, validation studies should focus on clinical outcomes of AI system implementation and not simply on prospective algorithm performance validation. Ophthalmologists need to aim for patient-specific outcomes such as vision loss and disability while maintaining focus on health care–specific outcomes such as reducing money spent per intravitreal injection or surgery.[16,17] Laboratory performance does not equal outcomes, and this was highlighted by the case of mass adoption of computer-aided diagnosis (CAD) for mammography screening. Although early reader studies showed that computers working with radiologists led to better accuracy than radiologists alone, subsequent clinical trials demonstrated that false-positive rates increased after CAD adoption. This led to an almost 20% increase in the rate of biopsies, confirming the potential disconnect between diagnostic accuracy and clinical effectiveness.[16,17]

Furthermore, as predictive algorithms become implemented in clinical decision making, false positives may become a self-fulfilling prophecy.[18] An ophthalmologist will not know whether the patient predicted to develop proliferative diabetic retinopathy, who was subsequently treated with a prophylactic anti-Vascular endothelial growth factor (VEGF) injection, will have ever developed the disease if he or she was not treated. When used in this manner, the algorithm's false-positive rates would be extremely difficult to detect—even in large outcomes-based clinical trials. If that so-called false-positive patient then went on to develop an iatrogenic complication, unnecessary harm would be inflicted and would be difficult to attribute to the AI's recommendation in any post hoc analyses.

Limitations of Deep Learning in Ophthalmology

Even with the success of deep learning–based AI models in the research setting, we must be cautious by not inferring similar performance in real-world use. Although deep learning has led to substantial advances in image classification, it is not without shortcomings. Gary Marcus[19] has well summarized the limitations; specific criticisms that are relevant to ophthalmology include insufficient transparency, poor integration with prior hierarchical knowledge, and inflexibility. Furthermore, a model's evaluation metric may not be indicative of the product's or patient's clinical goals.[20] In an illustrative clinical scenario, a deep learning AI model that was optimized for reducing macular thickness may prove irrelevant, as evidence becomes available that visual acuity may not correlate with this goal. This inflexibility and inability to encode hierarchical prior knowledge could quickly lead to a model's gross underperformance and rapid obsolescence.

AI models may break down when they encounter dissimilar image acquisition and patient-specific variables from those the model was trained on. Although prospective observational clinical validation is crucial to ensuring real-world model performance, this kind of data is often lacking.[21] It is frequently difficult to determine the precise reason that these models fail, and they are therefore often termed "black boxes."[20,22,23] Furthermore, if these models fail in screening settings, which often affect larger numbers of people, unnecessary additional interventions have the potential for even higher impact. This was exemplified by post hoc analyses of prior large-scale computer-aided screening programs.[17,24]

Novel Uses of Ophthalmic AI

Nevertheless, advances in AI have not only shown levels of performance that may supersede human ophthalmologists but have also demonstrated proficiency in tasks that were not previously thought possible for ophthalmologists to perform. Perhaps the most striking is a recent demonstration that a deep learning algorithm could accurately predict cardiovascular risk factors and demographics from fundus photos.[4,25] This AI model predicted age with a mean absolute error (MAE) of 3.26 years, sex with an AUC of 0.97,

smoking status with an AUC of 0.71, and systolic blood pressure with a MAE of 11.23 mm Hg. Although the ophthalmoscope was invented in the mid-1800s, and ophthalmologists have been looking at the fundus for as many years, these insights were not previously conceivable.

Recent evidence continues to support the view that the eye is a window to the vascular and central nervous systems of the body. Associations between retinal findings and neurodegenerative and cardiovascular diseases such as Alzheimer disease and hypertension have been increasingly validated.[26–28] Future advances in AI-based ophthalmic image analysis will undoubtedly demonstrate unforeseen disease associations and their ophthalmic correlates. The ability of AI systems to detect pixel-level patterns among millions of pixels per image, comprised in data sets approaching millions of patients, will never be matched by ophthalmologists. These occult patterns may enable not only earlier systemic disease detection but also novel insights into the pathophysiology of ophthalmic and systemic diseases.

The Potential of an Ophthalmologist-AI Partnership

Effective clinical medicine and ophthalmology with its large data sets of longitudinal imaging will ultimately benefit from collaborating with AI. Verghese et al.[29] describe humans working with machines and emphasize the lead time that predictive models can offer for diagnosis and action. However, these models can only lead to effective clinical decisions if they keep human intelligence "in the loop" to bring context. As with other imaging-heavy specialties such as radiology, ophthalmology is positioned to lead the uptake of medical AI. However, unlike radiologists, ophthalmologists additionally employ specialized examination skills and perform complex microsurgery. Therefore, ophthalmology is both uniquely positioned to take advantage of AI yet also uniquely protected against obsolescence to machines.

Concerns about physician unemployment have historically been raised with any stepwise improvements in automation and are often out of proportion to reality. Verghese et al.[29] reference an editorial from 1981 on using predictive risk factors from a then-novel computer database, stating that "proper interpretation and use of computerized data will depend as much on wise doctors as any other source of data in the past."[30] While a recent US report states 47% of jobs are at risk for automation, the risk for physicians and surgeons is estimated to be 0.4%.[31,32]

As various automations in medicine have freed clinicians from menial tasks, AI will continue that trend by integrating the ever increasing volumes of clinical, genomic, and imaging data. This will allow the ophthalmologist to focus on providing effective and compassionate clinical care. Currently, the majority of time is spent collating and synthesizing data and a minority interacting with the patient. However, with such vast volumes of data, the clinician would in effect be forced to use an aid to perform the data processing and thus have more time to be "deeply human" with the patient.

One can imagine a "clinic of the future"—a term described by Eric Topol,[33] in which a patient presents with multimodal high-resolution images, functional testing, genomic sequencing, metabolomic/proteomic information, and sensor data from home monitoring. Ophthalmologists would be provided a concise summary comprising structural and functional trends. They may also have access to richly annotated imaging segmented and highlighted for trending changes. Additional AI-synthesized predicted images could be presented of disease course depending on various treatment regimens. Consequently, ophthalmologists would then use this information as an additional tool and, together with the patient, formulate a treatment plan. They would subsequently compare this plan with the AI's recommendations, as well as its predictions of disease course and treatment response to select the best course of action based on the patient's unique circumstances.

Humans (and Human Ophthalmologists) Are Underrated

Time is a precious commodity for both patients and ophthalmologists. Patients often complain of insufficient doctor contact; similarly, physicians are increasingly burnt out from more time spent on clerical tasks than patient care.[34] Although new technology often promises efficiency improvements, as with the case of the electronic health records, one can see that such promises frequently fail to deliver. If implemented correctly, AI is unique in its potential to save time by processing large longitudinal data volumes and efficiently representing the patterns identified. Ophthalmologists will have more time for physical patient contact—everting an eyelid to discover a hidden conjunctival melanoma, performing a thorough gonioscopy or cranial nerve examination, or perfecting their surgical technique.

As described by Geoff Colvin,[35] human brains were designed for social interaction. No patient would want to be informed that they have a terminal disease or that they are going blind by their AI assistant. The new high-value skills will become those that "literally define us as humans," sensing the thoughts and feelings of patients losing vision, coordinating assistive devices with family members, and allowing the patients to express themselves about how their eyesight affects their lives. Although many ophthalmologists disagree with the concept of patient satisfaction influencing reimbursement, this relatively new development is an example of the increasing value being placed on such human-metrics. Colvin[35] states, "It used to be that you had to be good at being machinelike. Now, increasingly, you have to be good at being a person. Great performance requires us to be intensely human beings."

Conclusion

We currently lie in a stage between AI demonstration and deployment. Next comes ongoing evaluation, learning, model adjustment, and finally meaningful human-AI interaction. Ophthalmologists should leverage the primary strength of AI, its ability to glean insights from large volumes of multivariate data, with their abilities to interpret the AI's recommendations in a clinical and societal context. In doing so, the field will be well positioned to lead the transformation of health care in a positive and personalized direction. As more time will become available for human-suited tasks, ophthalmologists will have more time to be human—we will just use a digital helping hand from AI.

Acknowledgments

Supported by a Springboard Grant from the Moorfields Eye Charity (EK), UK National Institute for Health Research (NIHR) Clinician Scientist Award (NIHR-CS–2014-12-023; PAK), and CAPES Foundation, Ministry of Education of Brazil, Brasília, DF, Brazil (DAF). "This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001" - Federal University of Sao Paulo. The views expressed are those of the author and not necessarily those of the NHS, the NIHR, or the Department of Health.

Disclosure: **E. Korot,** Google Health (E); **S.K. Wagner,** None; **L. Faes,** None; **X. Liu,** None; **J. Huemer,** None; **D. Ferraz,** None; **P.A. Keane,** Heidelberg Engineering (F), Topcon (F), Carl Zeiss Meditec (F), Haag-Streit (F), Allergan (F), Novartis (F, S), Bayer (F, S), DeepMind (C), Optos (C); **K. Balaskas,** Alimera (F), Allergan (F), Bayer (F), Heidelberg Engineering (F), Novartis (F), TopCon (F)

Citation: Korot E, Wagner SK, Faes L, Liu X, Huemer J, Ferraz D, Keane PA, Balaskas K. Will AI replace ophthalmologists?. Trans Vis Sci Tech. 2020;9(2):2, https://doi.org/10.1167/tvst.9.2.2

References

1. Jaderberg M, Czarnecki WM, Dunning I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*. 2019;364:859–865.
2. Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*. 2018;362:1140–1144.
3. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. *arXiv [csLG]*. September 2018. http://arxiv.org/abs/1809.11096.
4. Engel J, Resnick C, Roberts A, et al. Neural audio synthesis of musical notes with WaveNet autoencoders. *arXiv [csLG]*. April 2017. http://arxiv.org/abs/1704.01279.
5. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24:1342–1350.
6. Schlegl T, Waldstein SM, Bogunovic H, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*. 2018;125:549–558.
7. Bojikian KD, Lee CS, Lee AY. Finding glaucoma in color fundus photographs using deep learning [published online September 12, 2019]. *JAMA Ophthalmol*. doi:10.1001/jamaophthalmol.2019.3512
8. Wen JC, Lee CS, Keane PA, et al. Forecasting future Humphrey visual fields using deep learning. *arXiv [csCV]*. April 2018. http://arxiv.org/abs/1804.04543.
9. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv [statML]*. December 2014. http://arxiv.org/abs/1412.6572.

10. Di Blasi Z, Harkness E, Ernst E, Georgiou A, Kleijnen J. Influence of context effects on health outcomes: a systematic review. *Lancet*. 2001;357:757–762.
11. Mittelman M, Markham S, Taylor M. Patient commentary: stop hyping artificial intelligence-patients will always need human doctors. *BMJ*. 2018;363:k4669.
12. Oaken-Rayner L. *Medical AI safety: we have a problem*. July 11, 2018. Available at: https://lukeoakdenrayner.wordpress.com/2018/07/11/medical-ai-safety-we-have-a-problem/. Accessed September 29, 2019.
13. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96:63–68.
14. Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digital Medicine*. 2019;2:92.
15. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572:116–119.
16. Oaken-Rayner L. *Medical AI safety: doing it wrong*. January 21, 2019. Available at: https://lukeoakdenrayner.wordpress.com/2019/01/21/medical-ai-safety-doing-it-wrong/. Accessed September 29, 2019.
17. Lehman CD, Wellman RD, Buist DSM, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175:1828–1837.
18. Hiremath S. *The rise and rise of quantitative Cassandras— NephJC*. September 24, 2019. Available at: http://www.nephjc.com/news/ai-commentary. Accessed September 29, 2019.
19. Marcus G. Deep learning: a critical appraisal. *arXiv [csAI]*. January 2018. http://arxiv.org/abs/1801.00631.
20. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *arXiv [csLG]*. February 2016. http://arxiv.org/abs/1602.04938.
21. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*. 2019;1:e271–e297.
22. Bansal A, Farhadi A, Parikh D. Towards transparent systems: semantic characterization of failure modes. Available at: http://www.cs.cmu.edu/~aayushb/pubs/characterizing_mistakes_eccv2014.pdf. Accessed January 21, 2019.
23. Baehrens D, Schroeter T. How to explain individual classification decisions. *J Mach Learn Res*. 2010;11:1803–1831.
24. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. *N Engl J Med*. 1998;338:1089–1096.
25. Poplin R, Varadarajan AV, Blumer K, et al. Predicting cardiovascular risk factors from retinal fundus photographs using deep learning. *arXiv [csCV]*. August 2017. http://arxiv.org/abs/1708.09843.
26. Schrijvers EMC, Buitendijk GHS, Ikram MK, et al. Retinopathy and risk of dementia: the Rotterdam Study. *Neurology*. 2012;79:365–370.
27. Chan VTT, Sun Z, Tang S, et al. Spectral-domain OCT measurements in Alzheimer's disease: a systematic review and meta-analysis. *Ophthalmology*. 2019;126:497–510.
28. Wong TY, Klein R, Klein BE, Tielsch JM, Hubbard L, Nieto FJ. Retinal microvascular abnormalities and their relationship with hypertension, cardiovascular disease, and mortality. *Surv Ophthalmol*. 2001;46:59–80.
29. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319:19–20.
30. Califf RM, Rosati RA. The doctor and the computer. *West J Med*. 1981;135:321–323.
31. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? Available at: https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf. Accessed January 21, 2019.
32. Will "physicians and surgeons" be replaced by robots? Available at: https://willrobotstakemyjob.com/29-1060-physicians-and-surgeons. Accessed September 18, 2019.
33. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY: Hachette; 2019.
34. Shanafelt TD, Dyrbye LN, West CP. Addressing physician burnout: the way forward. *JAMA*. 2017;317:901–902.
35. Colvin G. *Humans Are Underrated: What High Achievers Know That Brilliant Machines Never Will*. New York, NY: Penguin; 2015.