Ophthalmol Ther (2022) 11:69–80
https://doi.org/10.1007/s40123-021-00430-6

REVIEW

Translating the Machine: Skills that Human Clinicians Must Develop in the Era of Artificial Intelligence

Tariq M. Aslam · David C. Hoyle

Received: October 20, 2021 / Accepted: November 9, 2021 / Published online: November 22, 2021

ABSTRACT

In coming decades, artificial intelligence (AI) platforms are expected to build on the profound achievements demonstrated in research papers towards implementation in real-world medicine. The implementation of AI systems is likely to be as an adjunct to clinicians rather than a replacement, but it still has the potential for a revolutionary impact on ophthalmology specifically and medicine in general in terms of addressing crucial scientific, socioeconomic and capacity challenges facing populations worldwide. In this paper we discuss the broad range of skills that clinicians should develop or refine to be able to fully embrace the opportunities that this technology will bring. We highlight the need for an awareness to identify AI systems that might already be in place and the need to be able to properly assess the utility of their outputs to correctly incorporate the AI system into clinical workflows. In a second section we discuss the need for clinicians to cultivate those human skills that are beyond the capabilities of the AI platforms and which should be just as important as ever. We describe the need for such an awareness by providing clinical examples of situations that might in the future arise in human interactions with machine algorithms. We also envisage a harmonious future in which an educated human and machine interaction can be optimised for the best possible patient experience and care.

Keywords: Artificial intelligence; Machine learning; Ophthalmology; Human skills

Key Summary Points

Artificial intelligence (AI) platforms are likely to increasingly penetrate to direct clinical care in upcoming years as an adjunct rather than replacement for human clinicians.

It is therefore incumbent on human clinicians to arm themselves with the knowledge and skills necessary to effectively interact with AI systems of the future.

T. M. Aslam · D. C. Hoyle
School of Pharmacy and Optometry, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UKD. C. Hoyle
e-mail: dchoyle01@gmail.com

T. M. Aslam (✉)
Manchester Royal Eye Hospital, Manchester University NHS Foundation Trust, Oxford Road, Manchester, UK
e-mail: Tariq.aslam@cmft.nhs.uk

This undertaking involves developing an understanding of the algorithms themselves, their strengths and weaknesses and the precise meaning of their outputs with relevance to individual clinical scenarios.

Human skills will remain critically important, and these skills should also be nurtured with respect to particular aspects of their importance in the AI-enabled clinics of the future.

INTRODUCTION

There has been a rapid proliferation of research in recent years demonstrating the immense power of artificial intelligence (AI) and, in particular, deep learning, to perform complex clinical tasks that had hitherto been regarded as the exclusive domain of human experts [1, 2]. The pace and breadth of developments in this discipline, including in the fields of mathematics, medicine, computing, statistics, and research methodology, would challenge the capacity of most clinicians to comprehend them all, with pertinent questions left unanswered on clinical implications and impact.

The most fundamental of these questions is whether the abilities of current AI systems will effectively replace trained human experts, leaving them redundant in a brave new world. Our article demonstrates that, to the contrary, there are areas where human intelligence provides crucial services beyond the capability of current AI systems and yet other areas where although AI systems may have proven equality or superiority, expert human skill is still required to interpret and interface with the computer systems. Indeed, studies demonstrating such effectiveness of AI systems may not be as promising as they appear and must be subjected to further assessment to ensure that they are based on research conducted under strict methodological guidelines [3]; for example, a recent study highlights that impressive-appearing outcomes in some scientific reports are significantly flawed when subjected to close scrutiny [4]. The findings also need to be externally validated with independent prospective clinical trials in a real-world setting that considers workflows and addresses any legal or ethical implications. The algorithms will ultimately need regulatory approval, and subsequent clinical deployment is likely to require clinician training and support as well as regular audits, including close attention to adverse events and evidence for clinical benefit.

This complex and necessary development pathway requires considerable time and finances to bridge the results of any widely released, albeit impressive, published algorithm results to real-world utility. The many less common conditions and less stable situations that are more complex to train and less lucrative may therefore remain in the human domain for many years to come, even if they are cases humans might appreciate more help with. Ultimately, most authorities concur that AI is likely to involve platforms as useful adjuncts to help clinicians rather than replace them, both in medicine in general [5–8] and in ophthalmology specifically [9, 10].

The profound results achieved by DeepMind and other systems do however indicate that this adjunctive potential of AI could have a revolutionary impact on ophthalmology specifically and medicine in general [5], potentially addressing many of the crucial scientific, socioeconomic and capacity challenges facing populations worldwide. Thus, the important question for human clinicians becomes not whether AI will replace them but how should they work with AI to ensure that this potential is fully realised. Whilst AI has the potential within the field of ophthalmology to be an all-pervasive powerful technology, it is also one that requires careful implementation as its foreign methodologies may be a degree more opaque and impenetrable to humans than other technologies we have adapted to.

Given the vast field of AI and machine learning research, in this article we will not explain the technical details of developing AI systems or guidelines on conducting detailed research critique, issues which have been amply addressed elsewhere, both in medicine in

general [11, 12] and in ophthalmology specifically [13, 14]. Instead, this article presents a template for human clinicians to prepare for a future landscape where validated AI systems need to be appropriately exploited and combined with the best of human intelligence to provide the safest, most effective and efficient healthcare system for their patients. This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

There are two broad areas of this template where we believe human clinician awareness and training will ensure they are prepared for the new era of AI. The first is in the appropriate identification, assessment and incorporation of AI systems into their clinical workflows; the second is in the cultivation of their human skills that are beyond the capabilities of the AI system. We will address these two aspects in turn.

ASSESSMENT AND INCORPORATION OF AI INTO CLINICAL WORKFLOW

Understanding Where AI has been Employed

We may imagine AI systems as being discrete, highly visible and publicised tools that we need to opt-in to use, but their presence may not necessarily be obvious and the first challenge for human clinicians involves awareness and vigilance to this presence. Physicians should ideally be aware of all AI processes in their workflow so that these can, if necessary, be muted to serve the raw data to the clinician for analysis.

For example, AI may be integrated into patient electronic health records (EHR), using natural language processing (NLP) to filter out noise and extract information determined to be important for the clinician [15]. Whilst these steps may be overall of great benefit, if they are not understood, there is a risk of complex or rare cases being mismanaged due to the inadvertent filtering out of pertinent data. Useful information for an ophthalmic surgeon from a large set of patient data, for example, may be that the patient is myopic and diabetic; however, this crucial information might not necessarily be highlighted by a deficient NLP system, and any presumption by the attending physician that qualified clinicians had procured and shared all relevant data could lead to errors.

Similarly, the images we see from modern ophthalmic scanners, such as optical coherence tomography (OCT) angiography, are highly processed from raw collected data, and some of these steps may involve machine learning. In some cases, these applications are clearly indicated, and associated studies have demonstrated their benefit [16]. However, this may not always be the case for future applications where it may be possible that images apparently enhanced by machine learning could also cause degradation of relevant information in certain circumstances.

Understanding the Accuracy of the AI Algorithm Employed

Even if an algorithm is clinically approved, its level of validity will need to be fully understood to be properly incorporated into any decision-making process in an individual clinical practice; therefore, understanding and interpreting the nuances behind accuracy metrics may be a key critical skill for a clinician working with AI.

As an example, consider a theoretical AI algorithm that is being used in a clinic. From an analysis of an OCT image this algorithm estimates that there is an 81% probability that the macular neovascular membrane in a patient's left eye is active. This 81% probability is essentially a measure of how confident the algorithm is that the patient's membrane is active. The clinician, however, feels the membrane is not necessarily active and is unsure how to proceed, especially as the patient had a severe adverse event from an injection in the other eye and does not want to be treated unless absolutely necessary.

The first thing to note is that although the 81% figure is one that is easy for the clinician to grasp, it only represents the developed algorithm's estimate. We know, however, that more

pertinent information for our clinician would be the accuracy of this prediction based upon its subsequent performance in independent trials with patients in a clinical setting.

In these clinical trials the output of such an algorithm should be tested against a gold standard; in this case the best possible human expert decision, which involves using all available data on whether the OCT image was indeed one of an active membrane or not. This then becomes a comparison of the output probabilities from the algorithm with a gold standard binary classification of active or inactive. To allow for this comparison, the clinical trial investigator could choose a threshold for the output probability above which the prediction is considered to be for an active membrane and below which it is considered inactive. For example, if the threshold applied to the algorithm's output is 50%, then if the algorithm determines a 50% or greater probability of an active membrane, its classification is given as 'active'. Depending on which threshold is chosen, the performance of the algorithm measured in clinical studies against the gold standard classification will naturally change. For example, by choosing a high threshold for membranes to be thought of as active the specificity of detection would increase. The sensitivity, however, will also become lower as some of the patients with active membranes would not be detected by such a high threshold. The sensitivity is also known as the true positive rate (TPR), and specificity is related to the false positive rate (FPR = 1 − specificity). In order to show the full range of the algorithm's ability, a research paper will often visually show how the TPR changes with the FPR as we change the cutoff threshold at which we define a 'positive'—in this case an active membrane. This is typically done in the form of a plot of a receiver-operator-characteristic (ROC) curve [17]. The ROC curve provides a summary assessment of the overall predictive performance of the algorithm, typically quantified through reporting the area under that ROC curve, i.e. the area-under-the-curve (AUC or AUROC) statistic. Less common, but related and perhaps more useful, is a plot of the precision–recall curve [18, 19]. Figure 1 shows the ROC curve for a data set of 1000 algorithm predictions against actual trial results that is consistent with the algorithm from our example. We see two potential threshold locations highlighted, and for each location we have added to the curve the crucial data of results from comparing actual clinical findings in the trial to the 1000 test-set predictions, if that threshold was reached. The data at each location are displayed in a matrix, known as a confusion matrix; in this case it compares the gold standard actual outcomes for an OCT to the clinical trial results for the algorithm's prediction for those same OCTs and when that threshold is applied.

For our example in the clinic, the algorithm gave an output probability of 81%. The actual confusion matrix for the surpassed threshold of 80% is shown in Fig. 2 and represents the clinical trial comparisons and, therefore, most closely the actual values we might expect in the scenario for our clinician and their patient. The numbers in this matrix show that with this threshold the algorithm's classifications in the clinical trial have an overall accuracy of 85%, which is the proportion represented by the diagonal elements in the confusion matrix.

This overall accuracy figure is reassuring to our clinician. However, even this figure of accuracy with a threshold appropriate to their patient can be misleading. The overall prevalence of active membranes in this population is low, and so even a naïve algorithm that always predicted an inactive membrane would achieve a high degree of accuracy. What is more relevant to the AI-aware clinician using the algorithm in their practice is specifically how likely is their patient to have an active membrane given the algorithm has estimated an active membrane probability higher than the 80% threshold—what is called the positive predictive value (PPV; also called precision). From the right-hand column in Fig. 2 we see that with the confusion matrix at the cutoff threshold of 80%, when the algorithm predicts an active membrane its accuracy is actually 108/(108 + 58) = 65%; thus, with the application of the 80% cutoff, 65% of predicted active membranes actually turned out to be active in the sample of 1000 patients in the clinical trial. Its estimated PPV is therefore 65%. This figure of

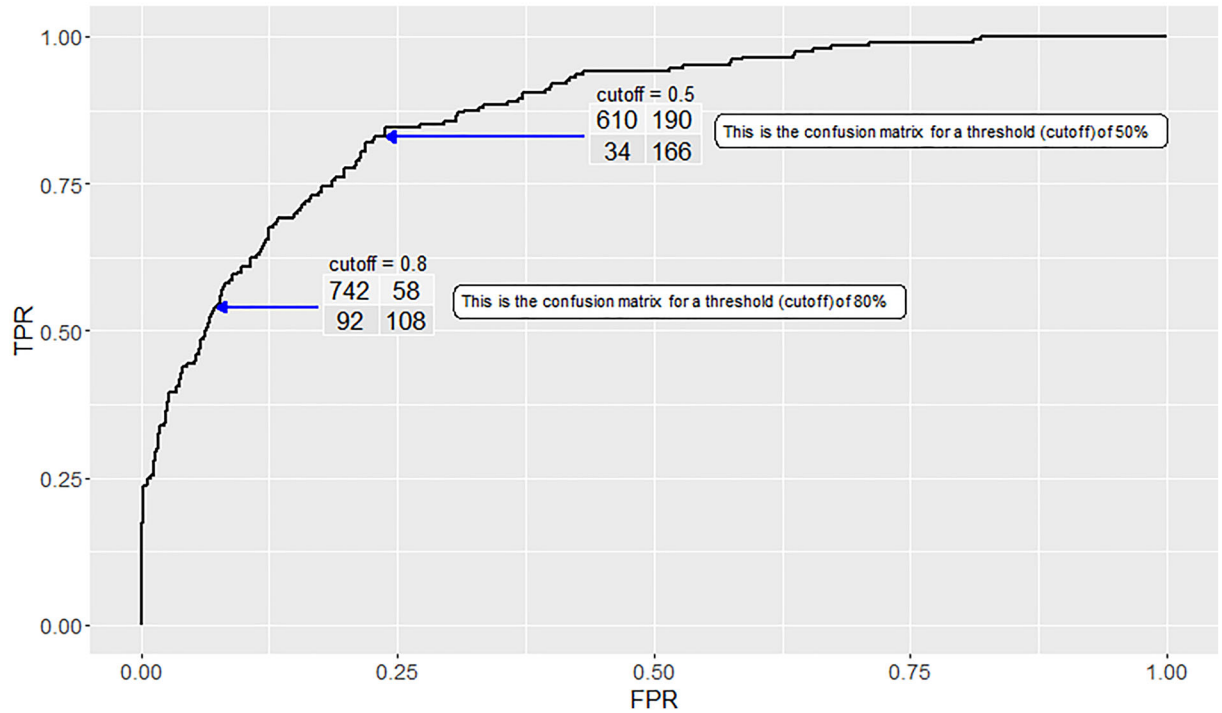

Fig. 1 Example receiver-operator-characteristic curve for test data that are consistent with the algorithm being used by our clinician in the example. Two different thresholds, or cutoffs, are highlighted. At each, we display the resultant data in a confusion matrix. The 80% (0.8) threshold cutoff provides the closest appropriate data relevant for the output of the algorithm used in our clinical scenario. *FPR* False positive rate, *TPR* true positive rate

		Predicted	
		Inactive	Active
Actual	Inactive	742	58
	Active	92	108

Fig. 2 Test-set confusion matrix for our artificial intelligence algorithm taken at a threshold of 80%. The overall accuracy in the test set is total correct predictions/total predictions. In this case, the total accuracy is (742 + 108)/1000 = 85%

65% is the closest yet described to one the clinician should use to judge the validity of the AI algorithm's prediction for their patient—and is somewhat different to that which was first presented.

Although this is a hypothetical scenario, it demonstrates the variety of statistics that might be used to present algorithm validity and how they can easily be misinterpreted. Although a plot of the ROC curve or precision–recall curve is an excellent overall indicator of an algorithm's performance that is very appropriate for publications validating algorithms, it still does not allow the clinican, in a clinical setting, to determine the PPV for their specific patient. Even the PPV has deficiencies, and other refined indicators may become commonplace, but the importance of understanding the principles of these statistics is illustrated here with a final most appropriate value that is very different from the one first presented.

Understanding the Compatibility of the AI Algorithm with Current Clinical Scenario

Even when we have a clinically approved algorithm and can confirm its highly impressive predictive value with reference to our clinical problem, this may still not be enough to warrant a clinician deploying or deferring to its decision-making. As an overarching principle, it is well recognised that AI algorithms are limited by the data they are trained on. As such, if the current clinical scenario is different from the situation for which the algorithm was originally trained, then the algorithm will not achieve the reported level of predictive accuracy. Consequently, the clinician needs to ensure the current context corresponds with that of the algorithm training and testing process.

At a basic machine level, for example, when using an algorithm that interprets colour fundus images it would be important to know the technical quality of the fundus camera and what the mix was of patients that the algorithm was trained on. If a clinician is seeing an Asian patient with a much higher quality resolution image then they should be less likely to completely rely on a computer algorithm that was trained on white patients with lower quality images and may wish to give greater credence to human expert opinion. This example highlights that successful use of an AI algorithm for diagnostic purposes requires the identification of any material contextual differences between the algorithm's training population and the current clinical scenario. Any differences need to be accounted for by the human interpreting the algorithmic output into a clinical decision.

Understanding How AI and Clinician Diagnoses may be Combined

When an algorithm has been robustly developed and shown to have excellent outcomes with similar populations and characteristics to those in your specific workplace, the decision may be made to deploy it in this clinical workplace. At the heart of the public and media fervor on the future roles of AI algorithms are claims that they are markedly more accurate than clinicians in making a diagnosis. However, a recent meta-analysis from Nagendran et al. [20] suggests that the reported evidence for the superiority of deep-learning systems over clinicians is not clear-cut, primarily due to a lack of well-designed prospective trials.

Furthermore, even if headline figures suggest equality in trials comparing human against machine decision-making, it should be borne in mind that mistakes that are made by humans may be very different from those made by AI systems, with the latter having no backing of general intelligence and therefore having the potential to result in more devastating consequences [21, 22]. For example, if we take a hypothetical scenario of an algorithm screening patients for exudative age-related macular degeneration (AMD) to decide if they should have a hospital visit, we could find that only one error out of 100 occurred for the algorithm and one error out of 100 occurred for a human expert, suggesting that they are equal in terms of screening for AMD. However, in this scenario the algorithm followed its specific training, and its single error was in defining a patient who had a very large and obvious melanoma in the periphery of imaging as not needing medical care—a condition which even a junior clinician would have easily detected. This melanoma was not in its learning database and was missed repeatedly as it grew as there is no automatic common-sense check in AI. Conversely, the human error was in missing a small area of paracentral subretinal fluid which at the next visit of the patient had grown in size until it was more visible and easily noted. Clearly, errors made by systems without general intelligence have the potential to be of different consequential importance to those by humans [21]. Equal sensitivity and specificity do not tell the whole story, and workflows may need to have human checks in place for such eventualities.

Finally, it has been reported that even when algorithms appear to show superiority in formal statistical testing, combining human and algorithmic decisions can improve results beyond those achieved by either system alone [23–25].

There is therefore a need, for all of these reasons, to understand how to combine the diagnosis made by the AI algorithm and the

clinician into a consensus. But what is the best practice for doing this?

Mathematically combining predictions from multiple algorithms to improve overall accuracy is standard practice within the machine learning field [26] and is variously referred to as 'classifier combination', 'classifier fusion', 'decision fusion', 'mixture of experts' or 'ensemble learning'. This can be a complex task, for which heuristic approaches exist [27], as well as more principled Bayesian approaches [28, 29]. Most methods require the measurement of the predictive accuracy of the individual algorithms to determine appropriate weights when combining the predictions. The assessment of the correlation between the output of the different algorithms may also be required: consider if we had two algorithms whose predictions are perfectly correlated, i.e. identical, one of the algorithms would be redundant as it provides no new information to us.

To apply such techniques to derive a consensus of the AI algorithm and clinician would then require quantifying the historical diagnostic accuracy of the clinician—and importantly—evaluated on the same patient samples on which the AI algorithm was evaluated. In the longer term, such a combination of AI and clinician expertise may be possible, guided by software that has already accessed the historical audit information on clinician diagnostic accuracy or taken the clinician through an appropriate calibration process as part of the setup of the new diagnostic tool. However, the difficulty of obtaining AI versus clinician comparisons on the same samples makes it unlikely that such a combination of AI and clinician diagnoses will be routine in clinical settings anytime soon, and clinicians will need to learn how incorporate in a non-mathematical way the prediction of an AI system—or even possibly reject it.

As an example, a clinician reviews the OCT scan of a patient with AMD and feels that although the condition is largely stable, there are some discrete choroidal signs that have been shown in literature to indicate inactive disease. A machine learning system known to use the appropriate and relevant learning input data is then applied, and after processing the patient's OCT the AI system states that the wet AMD is actually active, with 90% probability and with a PPV also of 90%—i.e. it disagrees with the clinician's diagnosis. The clinician is not sure and estimates their human clinician level of certainty just above 50:50. Superficially, it appears that the AI algorithm is highly likely to be correct. How should the clinician proceed?

To fully optimise the power of AI, clinicians should demand systems that feature explainable AI [30]. Due to significant work in the AI field on this, most image-based AI prediction algorithms will be capable of attaching to each pixel a numerical measure of how that pixel contributed to the algorithm's prediction through so-called saliency maps [31, 32]. From this the clinician can see whether the algorithm is basing its prediction upon the same areas of the OCT as they are. Despite processing the same input data, the AI algorithm and clinician may still be using different features to make their diagnoses.

In our clinical example above, the clinician may become aware of basing the diagnosis on a relatively small area of the OCT that can be indicative of inactive disease but with no appreciation of other discrete and faint areas that the highlighted pixels have now demonstrated. This information helps the clinician to decide on a management plan.

Rather than the clinician attempting some form of mathematical computation or disregarding the AI algorithm in cases of disagreement, the preferred pathway is therefore likely to be close examination of the explainable AI data and, combined with knowledge of the AI training and development, to make an empirical judgment on the algorithm. It may be as simple as noting that clearly artefactual data were used in the algorithmic decision to overrule it. Alternatively, it may reveal, such as in our example, that areas of the OCT that the human physician had not seen contain useful information, thereby validating the algorithm decision over the initial human one and reassuring the physician of the decision to follow the algorithm. This important step of challenging AI decisions using explainability would likely rely on the human making decisions before seeing the decision of the AI system to

avoid a gradual complacency of aligning with the view of the computer without independent human thought. Clearly this attribute of explainability should be sought out in any clinically available systems of the future.

CULTIVATING HUMAN SKILLS THAT EXCEED AI

In the preceding section we have explored how humans should adapt in order to be able to incorporate AI decisions into their clinical work streams. We now address the importance of humans in their own right. Where are the abilities of humans likely to continue to be greater than those of computers?

Firstly, we address the need for humans in the incorporation of AI predictions into the wider clinical management strategy. In the preceding sections we discussed issues around understanding the utility of AI-based analyses applied to discrete clinical data, such as images, and combining these assessments with clinician assessments of the same data. However, physicians' patient management strategies often need to incorporate more information than that derived from structured planned history, examination and investigations. We clinicians must also consider more vague concepts, such as the patient's willingness and tolerance of treatment, preference or insistence on seeing particular physicians, reliability of future attendances, convenience for timing of treatment, general discomfort and faith in the treatment and service. Only with knowledge of such human factors are clinicians able to balance the holistic impact of their treatment decisions—the decision-making process then involves the integration of all these pieces of information, with the clinician ultimately making trade-offs to identify the optimal course of action for the patient. In principle, calculating the optimal trade-offs could be performed by another algorithm, provided all risks and benefits have been suitably quantified . In reality, however, a patient may only give vague qualitative statements about their views on treatment or even only non-verbal cues, and it would still require a clinician to translate the consultation dialogue and behaviour into a quantitative point on a scale for each of the many factors. Then, other considerations outside of those programmed for could arise, such as local travel problems. Overall, there are multiple points during this process where the clinician's expertise is needed, and the nuances of the various trade-offs mean that decisions around whether and how to treat are likely to be always better made by a human clinician.

For example, a patient with diabetic retinopathy attends clinic in the early stages of a viral pandemic. The hospital cancelled the patient's last appointment, and staffing levels are low. The patient admits to poor glucose control. The patient is also a carer and must leave their partner to attend clinics. An algorithm suggests the likelihood of the retinopathy being proliferative is low at this visit, recommending a review in 4 months. However, the retinopathy is still very severe and worsening. You have access to the laser today and the patient seems willing. Do you proceed to laser, or just to review? If you proceed to the laser treatment, you may be overtreating according to the algorithm and possibly cause unnecessary retinal damage. However, you may decide to perform the laser treatment anyway because if the patient fails to make the next appointment and develops neovascular complications the consequences could be devastating and so overall it may be the safer option.

The discussion here is not about the details of the AI algorithm, its accuracy, nor the nuances in how to interpret its output. Instead, the discussion is about the decision to treat, the wider inputs to that decision and the various prognoses with their upsides and downsides. The AI algorithm that predicts exclusively the probability of proliferative retinopathy is only a small part of this wider decision and should be recognised as such. Decision-making in nuanced and highly variable situations is precisely the sort of thing humans excel at whilst algorithms struggle with [33], reinforcing an ongoing need for humans to ensure they maintain and nurture their broad clinical skills. Humans should therefore be reminded of the importance of taking the AI algorithms as a report only on the very precise clinical question

posed to the algorithm. Physicians must use this report together with all the additional information they have on the patient's overall external clinical factors to make a final clinical decision, which may or may not follow the algorithmic advice.

In our final segment on how we as humans must learn to assimilate and interact with the tools that AI provides us, we turn appropriately to a discussion of the importance of further nurturing of those characteristics that make us distinct as humans. We must further remember that even if extremely accurate and appropriate information is available from a computer terminal, patients will often prefer to hear it from a clinician. The patient may be anxious and need to hear the voice of a confident and respected individual to convey what may be alien information. They must hear this from an empathic and knowledgeable voice they feel they can trust. As we discussed in preceding sections, AI decisions may not necessarily be followed by the practicing human clinician, any such deviations from computer recommendations would need to be understood by the clinicians and duly explained.

Once the information is conveyed, the patient will very likely have questions, and these may or may not be about the algorithm. Explainable AI may theoretically allow less well-trained individuals to provide some information on how the algorithm arrived at its conclusions but is unlikely to be appropriate for many questions, such as how the disease will affect detailed aspects of the patients' specific lives. If AI can be used to effectively enhance and speed up some aspects of consultations, then we should take the opportunity to enhance these other aspects of communication with patients to improve their overall experience and well-being.

Finally, there is a more general argument over patients' overall acceptability to have machines making decisions, a concept that has been wrestled with in the context of automated cars [34]. As well as the general principle of important decisions being made by machines, we have become accustomed to seeking apologies, explanations, retribution or forgiveness when humans are responsible for errors. However, machines also have a level of fallibility. Patients may not accept these without the intervention of a responsible individual, and there is some evidence of people favouring human judgment in inherently uncertain domains such as in medicine [35].

In this section we have detailed the human skills that are superior to those derived from machine learning and highlighted how important these will be. They are also skills that have become somewhat degraded over time as clinician–patient time has become gradually eroded by economic pressures in all fields of medicine. These economic pressures may well impact the use of AI systems merely to speed up and automate patient care so that greater numbers of patients can be seen without compromising safety and at reduced cost. It is our opinion, however, that at least part of the time savings should be reinvested in resurrecting greater amounts of patient–physician time to realise the full potential of AI in the workplace. In an ideal future world, tasks can be delegated such that some of the more automated ones of image screening can be done by machines, leaving more time for human–human interaction in the explaining and nuanced decision-making tailored to the individual patient needs.

CONCLUSIONS

Medical science in general and ophthalmology in particular have been blessed in recent years with transformative technologies such as OCT imaging that have improved diagnosis and the management of patients. These technologies have typically required an ongoing process to educate and update clinicians, for example with training on OCT and interpretation of OCT angiography.

Inevitably AI algorithms in future clinical use will similarly emerge with their own individualised instructions for correct use. However, there are more fundamental principles and mindsets that will apply more generally to the use of all AI systems in modern clinical practice. This paper demonstrates how, just as with other revolutionary technologies, there are several distinct principles of AI algorithm use that

clinicians must be educated on to ensure correct use to improve patient care. We have examined these principles and explained the potential pitfalls if these principles are not understood.

With the increasing complexity and pervasiveness of AI, this knowledge will become a crucial foundational requirement for clinicians of the future. Indeed, a recent study suggests there is already an appetite among clinicians to understand more than just the cursory details of the AI systems they use [36].

It should be clear from this article that AI will not replace clinicians—rather, it will augment aspects of their work. Certainly, there are—at least in the short term—also significant non-technical barriers to the widespread adoption of AI into healthcare systems [37]. In the longer term, clinicians willneed to learn to interact with AI systems. However, it should also be remembered that as AI hopefully frees up clinicians' time, there will be an increasing need for clinicians to focus on tasks requiring empathy and communication, and also for higher-level tasks, such as the complex integration of multiple pieces of nuanced clinical information.

It is our hope that AI will indeed allow use of machines to do what they do well whilst at least to some extent freeing up humans for more time to do what we do well.

In paraphrasing a much-used expression, AI will not replace clinicians, but clinicians who understand AI and the principles outlined in this paper will perhaps replace those who do not.

ACKNOWLEDGEMENTS

Funding. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors. No external funding was provided in support of this work. No funding or sponsorship was received for this study or publication of this article. Neither author has a proprietary interest in the contents of this paper.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Author contributions. TA and DH contributed equally to the concept and design, drafting and final approval of the manuscript.

Disclosures. Tariq Aslam has received funding and educational grants from Bayer, Novartis Laboratories Thea, Oraya, Bausch and Lomb and , Roche. David C. Holyle reports no financial disclosures.

Compliance with ethics guidelines. This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

Data availability. Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

REFERENCES

1. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.

2. 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(9): 1342–50.

3. Begg C, Cho M, Eastwood S et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA. 1996;276(8): 637–9.

4. Roberts M, Driggs D, Thorpe M et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199–217.

5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.

6. Topol EJ. Deep medicine: how artificial intelligence can make healthcare human again. 1st ed. New York: Basic Books; 2019.

7. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;7:e7702.

8. Rampton V. Artificial intelligence versus clinicians. BMJ. 2020;369:m1326.

9. Zarbin MA. Artificial intelligence: quo vadis? Transl Vis Sci Technol. 2020;9(2):1.

10. Korot E, Wagner SK, Faes L et al. Will AI replace ophthalmologists? Transl Vis Sci Technol. 2020;9(2):2.

11. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.

12. Esteva A, Robicquet A, Ramsundar B et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.

13. Choi RY, Coyner AS, Kalpathy-Cramer J et al. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):14.

14. Faes L, Liu X, Wagner SK et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Transl Vis Sci Technol. 2020;9(2):7.

15. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. J Am Med Inform Assoc. 2015;22(5):938–47.

16. McGrath O, Sarfraz MW, Gupta A et al. Clinical utility of artificial intelligence algorithms to enhance wide-field optical coherence tomography angiography images. J Imaging. 2021;7(2):32.

17. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.

18. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on machine learning. Pittsburg: Association for Computing Machinery; 2006. p. 233–40.

19. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE. 2015;10(3):e0118432.

20. Nagendran M, Chen Y, Lovejoy CA et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.

21. Firestone C. Performance vs. competence in human–machine comparisons. Proc Natl Acad Sci USA. 2020;117(43):26562–71.

22. Finlayson SG, Bowers JD, Ito J et al. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287–9.

23. Nam JG, Park S, Hwang EJ et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology. 2019;290(1): 218–28.

24. Patel BN, Rosenberg L, Willcox G et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med. 2019;2:111.

25. Tschandl P, Rinner C, Apalla Z et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229–34.

26. Mohandes M, Deriche M, Aliyu SO. Classifiers combination techniques: a comprehensive review. IEEE Access. 2018;6:19626–39.

27. Tulyakov S, Jaeger S, Govindaraju V and Doermann D. Review of classifier combination methods. In: Marinai S, Fujisawa H, editors., et al., Machine learning in document analysis and recognition. Berlin: Springer; 2008. p. 361–86.

28. Kim, H.-C, Ghahramani, Z. Bayesian classifier combination. Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research 2012;22:617–27.

29. Pirš G, Štrumbelj E. Bayesian combination of probabilistic classifiers using multivariate normal mixtures. J Mach Learn Res. 2019;20(51):1–18.

30. Belle V, Papantonis I. Principles and practice of explainable machine learning. Front Big Data. 2021;4:39.

31. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham.

32. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. In proceedings of 2014 Workshop at International Conference on Learning Representations

33. Korteling JE, van de Boer-Visschedijk GC, Blankendaal RAM et al. Human- versus artificial intelligence. Front Artif Intell. 2021;4:14.

34. Othman K. Public acceptance and perception of autonomous vehicles: a comprehensive review. AI Ethics. 2021;1(3):355–87.

35. Dietvorst BJ, Bharti S. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci. 2020;31(10):1302–14.

36. Cai CJ, Winter S, Steiner D et al. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum Comput Interact. 2019;3(CSCW):104.

37. Singh RP, Hom GL, Abramoff MD et al. Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Transl Vis Sci Technol. 2020;9(2):45. https://doi.org/10.1167/tvst.9.2.45.