tvst

Special Issue

A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies

Livia Faes[1,2], Xiaoxuan Liu[1,3,4,5], Siegfried K. Wagner[6], Dun Jack Fu[1], Konstantinos Balaskas[1,6], Dawn A. Sim[1,6], Lucas M. Bachmann[7], Pearse A. Keane[6], and Alastair K. Denniston[3,4,5,6,8]

[1] Medical Retina Department, Moorfields Eye Hospital NHS Foundation Trust, London, UK
[2] Eye Clinic, Cantonal Hospital of Lucerne, Lucerne, Switzerland
[3] Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK
[4] Academic Unit of Ophthalmology, Institute of Inflammation & Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK
[5] Health Data Research UK, London, UK
[6] NIHR Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK
[7] Medignition Inc, Research Consultants, Zurich, Switzerland
[8] Centre for Patient Reported Outcome Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK

Correspondence: Alastair K. Denniston, University Hospitals Birmingham NHS Foundation Trust, Birmingham B15 2TH, UK. e-mail: a.denniston@bham.ac.uk

Received: October 1, 2019
Accepted: October 4, 2019
Published: February 12, 2020

Keywords: artificial intelligence; machine learning; critical appraisal

Citation: Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, Bachmann LM, Keane PA, Denniston AK. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. Trans Vis Sci Tech. 2020;9(2):7, https://doi.org/10.1167/tvst.9.2.7

In recent years, there has been considerable interest in the prospect of machine learning models demonstrating expert-level diagnosis in multiple disease contexts. However, there is concern that the excitement around this field may be associated with inadequate scrutiny of methodology and insufficient adoption of scientific good practice in the studies involving artificial intelligence in health care.
This article aims to empower clinicians and researchers to critically appraise studies of clinical applications of machine learning, through: (1) introducing basic machine learning concepts and nomenclature; (2) outlining key applicable principles of evidence-based medicine; and (3) highlighting some of the potential pitfalls in the design and reporting of these studies.

Introduction

Machine learning (ML), a form of artificial intelligence (AI), has generated considerable excitement in recent years, particularly through a number of prominent publications demonstrating the ability of these ML models to achieve expert-level diagnosis in multiple disease contexts.[1–6]

The very first AI-based technology approved by the US Food and Drug Administration was an ophthalmic application, IDxDR, an algorithm for screening diabetic retinopathy.[7] This was approved in April 2018 by the US Food and Drug Administration under its breakthrough device program, and has been followed by an increasing number of applications across a range of health-related indications.[8] Interest in this area is intense, but there is a need to ensure that this excitement is tempered by scientific rigor and critical appraisal. In a recent systematic review, we conducted an evaluation of the "state-of-the-art" AI for disease diagnosis using medical imaging, focusing

on deep learning models (an advanced subfield of ML characterized by neural networks).[9] Although this review identified more than 20,000 studies in the field, less than 1% of these studies had sufficiently high quality design and reporting to be included in the meta-analysis. Clear and transparent reporting of methodology and results, fit for AI studies, are needed. Without this, readers cannot judge whether reported findings are justified in the context of potential sources of bias, and the extent to which the findings of such studies are reproducible and generalizable.

With the introduction of reporting guidance, such as the Consolidated Standards of Reporting Trials (CONSORT)[10] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses,[11] substantial improvements have been made in the completeness of reporting across the medical literature. The application of similar reporting standards in the diagnostic field has been more challenging, with no single standard applicable to all diagnostic models. The Standards for Reporting Diagnostic accuracy studies[12] guidelines addresses accuracy studies of single test evaluations only, whereas multivariate diagnostic probability functions are better addressed by the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).[13] With the rise of AI in medicine, researchers from other fields with diverse research backgrounds and publication cultures have entered the medical field. Whereas the medical community has become accustomed to complying with agreed international standards of reporting, this appears to be much less prominent in other fields such as statistics, mathematics, or computational science.

Ophthalmology has been a leader in the AI health revolution, with particular interest being in the generation of algorithms that can perform diagnostic or grading tasks from imaging. Eye health has therefore become a test bed of innovation in the AI sector, and provides a rich source of case examples to illustrate the potential of ML algorithms in medical applications but also the pitfalls around the designs and reporting of such studies. Key information that should be reported on includes: technical specifications (e.g., which optical coherence tomography [OCT] device has been used); contextual study setting and cohort information (e.g., eligibility/selection criteria, demographics, clinical setting, time period, geographic location, the manner of enrolment, patient flow, missing data); and how data were processed (e.g., file image enhancements, cropping, storage file format).[14] Notably, ML models feature additional technical aspects that have yet to be comprehensively addressed yet in current reporting guidance. In response, extensions to several reporting guidelines (including TRIPOD-ML, CONSORT-AI and SPIRIT-AI) are in development.[15–17]

In this article, we hope to provide a reader's guide for those wanting to critically appraise studies of clinical applications of ML, particularly in the high-priority area of classification tasks to support its use in screening, diagnosis, and monitoring based on medical imaging.

Questions to Consider

Was the Study Methodology Prespecified?

Prespecification of study methodology should include: a description of the unmet need, the intended place of the model within a diagnostic pathway, the inclusion/exclusion criteria, the approach to validation, primary and secondary outcomes that will be evaluated, power calculation, and the statistical analysis plan.

A fundamental aspect of the study, key to the interpretation of its results, is to understand the intended use of the ML model in the diagnostic process. Will this test be used for triage or diagnosis? If used in a triage situation, specific test requirements relevant to mass screening could apply. Will this model be used as an isolated test, used in combination with other diagnostic elements (e.g., multimodal imaging), or used as an add-on or replacement test during the workup? If the ML model is a component of the diagnostic decision-tree, researchers should define how the information arising from the model fits within the overall diagnostic probability function.[18,19] We call for the need to clarify a priori the purpose of the ML model and for authors to use this to guide selection of the optimal study design.[20]

A priori reporting of the study methodology helps tackle a number of biases, including publication bias, where "negative" studies (i.e., those failing to reject the null hypothesis) are less likely to be published, and where the evidence base may be skewed in favor of models showing high performance.[21] Additionally, selective reporting of outcomes may occur, whereby the study is reported but only includes those outcomes that show the model in the best light.[22] This may be a particular pressure where a company holds a financial interest in a model and may profit from the exclusive reporting of positive outcomes. Both challenges may be addressed by the prospective registration of studies.[23] Documenting the intended study methodology a priori enhances transparency.[24]

Sample size and a statistical analysis plan should be prespecified. In a recent systematic review of deep learning studies, reporting of prespecified sample size

calculation to ensure sufficient power was scarce.[9] Although there is currently a lack of consensus on how to consider sample size in studies of ML models, it should still be prespecified according to the minimal clinical significant difference and the hypothesis of the study.[25,26]

Is the Model Being Evaluated in its Intended Stage in the Care Pathway?

Diagnosis is a process of integrating information derived from various stages in the patient pathway. Each stage, whether it be the presenting history, clinical examination, or a series of investigations, constitutes an individual data point along a stepwise diagnostic process. At each step, there is a transition from a pretest to posttest disease probability, and it is the combination of information derived at each step that makes up the final diagnosis decision.[27,28]

It is therefore important to understand where the dataset was generated from within a care pathway. For example, fluorescein angiography might only be ordered in those where OCT reveals suspicious findings. Consequently, pretest disease probability in patients undergoing fluorescein angiography is likely higher than those undergoing OCT. In the case of retrospective datasets containing routinely collected imaging data, this information is usually unavailable or is poorly recorded. Similarly, open-source data libraries (in ophthalmology, most notably the MESSIDOR dataset for fundus photographs[29] and the OCT dataset from Guangzhou Medical University and Shiley Eye Institute[30]) may provide vast volumes of images, yet details on indication for the investigation is typically absent.

Any new test, including ML models, developed using a given dataset should not be considered in isolation from its clinical pathway. When considering the validation of models based on a precurated dataset, it is important to ask: for what purpose was this dataset originally curated? And, does the disease probability within this cohort differ to the setting in which the model will be deployed?

Do the Authors Provide Sufficient Clarity on How the Data Were Split?

The terminology around datasets has been a common source of confusion in ML studies as authors have used many of the key terms interchangeably. Common practice in developing ML diagnostic algorithms is to split a dataset for development into training, tuning, and internal validation test sets

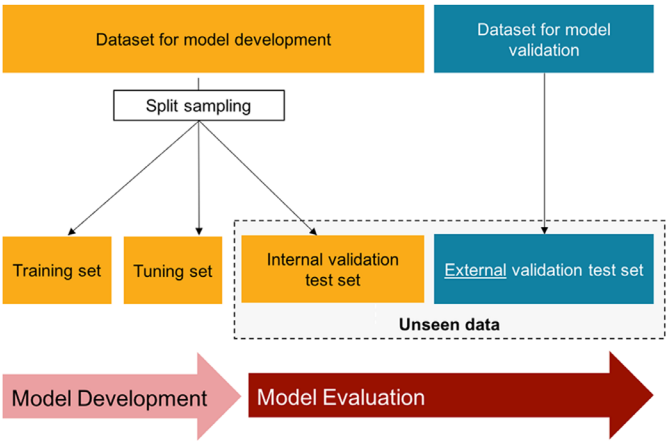

Figure 1. Overview of datasets involved in a machine learning diagnostic algorithm: model development and evaluation.

(split sample validation). Subsequent external validation test sets, for out-of-sample external validations, are also often sought to test for generalizability of the model. We recommend a standard nomenclature comprising the terms "development" and "validation" for the stages of development and evaluation, and the terms "training set," "tuning set," and "validation test set" (either an internal validation test set or external validation test set) for the datasets used (Fig. 1).[9]

Model Development

The training set contains "seen" data used to fit the model in an iterative fashion, and this is where most of the learning occurs. The tuning set is usually a smaller dataset containing examples separate to the training set. It provides an opportunity for ML engineers to observe the performance of a model and fine-tune the model weights (hyperparameters). In that sense, the tuning set also contains seen data. Both training and tuning should be considered part of the model development process because changes in observed performance prompts new adjustments to the model architecture. The internal validation test set on the other hand, whilst also part of the model development process, contains "unseen" imaging examples of the same patient cohort with which to test the performance of the finalized model. There should be no overlap between the seen and unseen datasets; therefore, description of the splitting between the training and validation/tuning datasets needs to be clearly documented.

Model Validation

Robust external validation of a model should be performed in an out-of-sample external validation test dataset.[13] This dataset should be distinctly separate from the development dataset (temporally and/or preferably geographically) and validation should be

performed by independent investigators. Model validation is discussed further in Question 8.

Are the Image Labels Likely to Reflect the True Disease State?

To assess the accuracy of any model, we need to assess against the ground truth (more commonly known to clinicians as the gold standard). Knowing the provenance of the ground truth is critical, and is more often an issue in ML compared with other diagnostic studies because of the sizes of datasets involved and the demands this may therefore place on any manual labeling process. Considerations include: Are the labels added manually or automatically generated from associated records (e.g., electronic health systems)? Are any ground truth labels missing? Were the images labeled prospectively or retrospectively? In some situations, retrospective labeling may be beneficial as it benefits from additional information (such as further follow-up data confirming a diagnosis).

A fundamental question is, how confident we are that these labels are indeed ground truth? It is important to know if these labels are based on a single note in a linked electronic health record or something more robust such as a linked definitive result (such as biopsy), an independent review of the image, or expert consensus? In cancer, the ground truth normally has a high degree of certainty because it is based on histology (and in some cases cancer-free survival). For most ophthalmic applications, the best ground truth available is usually expert opinion. However, the reliability of expert opinion should be critically appraised. It may vary considerably in robustness from, single expert to multiple expert majority vote, multiple expert consensus and multiple independent expert opinion with disagreements escalated to an adjudicator. Even the term "expert" could have different meanings: subspecialist for a certain number of years, board-certified specialist, or certified readers from a reading center.

A useful measure for the reliability of ground truth labels is interobserver agreement between the labelers, and it would be good practice to prespecify a threshold for inclusion of cases where there is nonconsensus. By reporting interobserver agreement, readers can at least make a judgment on the likelihood that the ground truth label is correct. For example, low agreement may signify ambiguous cases, and a decision should be made about whether it is appropriate to train an algorithm based on unreliable ground truth labels or whether these images should be excluded. Other methods to enhance the labeler's likelihood to accurately diagnose disease may also be used, such as presenting extra clinical information alongside the image, using combined information from multimodal imaging or another diagnostic test to confirm findings, and/or providing access to previous images that may give additional contextual information and provide a more appropriate benchmark. The additional benefits are that the comparator against which an algorithm is measured is also more representative of real-world practice, making the study results more clinically applicable. In our review of ML diagnostic algorithms, only 4 of 82 included studies considered the provision of additional clinical information, such as clinical vignettes or historical images, to the health care professional.[9]

How Is Diagnostic Accuracy Reported?

The terminology used to report diagnostic accuracy in ML may initially seem inaccessible to readers used to medical studies, however many of the statistical concepts are familiar. For example, recall is equivalent to sensitivity, precision is equivalent to positive predictive value, and confusion matrix is equivalent to contingency table. There is a clear need for both communities to understand each other's terminology: in medical applications, diagnostic accuracy is usually reported as sensitivity, specificity, and area under the curve; in ML applications, models are also commonly reported in terms of accuracy, F1 score, and dice coefficient. The provision of the actual contingency tables ensures clarity, and to some extent bypasses this issue.

Our recommendation is that, at a minimum, the contingency tables (true positives, true negatives, false positives, and false negatives), should be reported at a justified prespecified threshold. This information allows the calculation of all other relevant measures familiar to both medical and ML communities (Fig. 2). Contingency tables also provide additional value in multiclass classification tasks, where a model is trained to predict three or more disease classes, as opposed to a binary classification task (disease present/not present, class A or B). In multiclass classification, it may be of clinical interest to see if systematic misclassifications are occurring (for example, if subretinal fluid was systematically misclassified as intraretinal fluid).

Is the Dataset Used in Model Development Reflective of the Setting in Which the Model Will Be Applied?

The performance of a model is highly dependent on the training data. When considering a suitable training

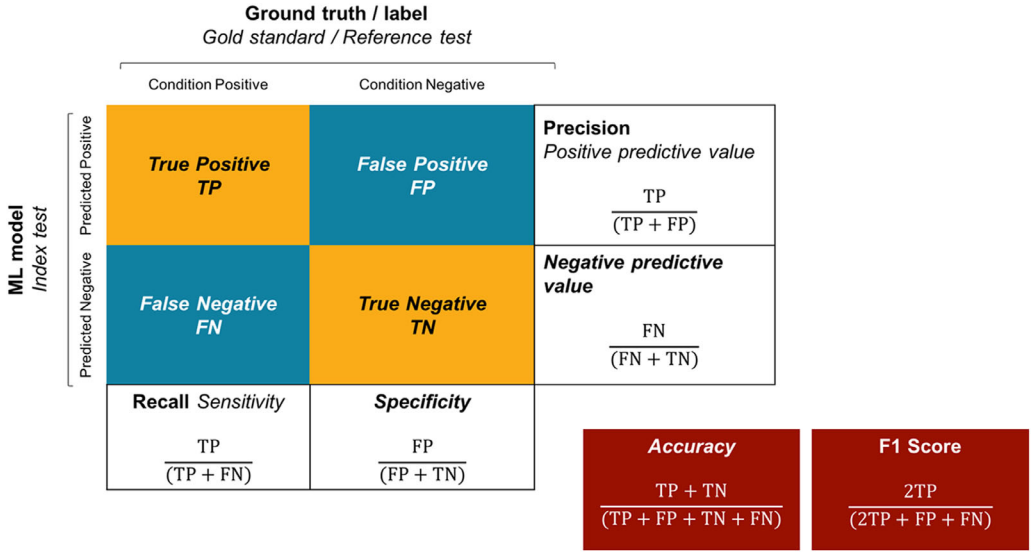

Figure 2. Overview of confusion matrix/contingency table. Differences in nomenclature for machine learning (boldface type) and classical statistics (italic type) and where overlapping (boldface and italic) are highlighted.

set, the need for size in a dataset is well recognized, but the need for data quality and appropriateness is often overlooked. It is important that the dataset is appropriate for training the algorithm for deployment in a specific real-world situation. Underrepresentation of important diagnostic features or disease states during development may profoundly limit its performance once it is released into its intended clinical arena.

One such consideration is whether the dataset represents the complete spectrum of diagnostic cues for the target population. Spectrum bias, where disease manifestations within a dataset (i.e., disease severity, stage, distribution of alternate diagnoses) do not adequately reflect the target patient population, is a common problem, particularly because many investigators may opt for datasets which represent extremes (i.e., normal vs severe disease).[25,31] For example, if a diagnostic algorithm was developed for the presence or absence of diabetic retinopathy (DR) and model was validated using a sample population with normal eyes and only severe cases of DR eyes, the model's ability to discriminate diabetic eyes from normal eyes will be overestimated ("two-gate design" or diagnostic case-control design). If this algorithm was released in a real-world situation, it may perform well on detecting severe nonproliferative DR (NPDR) or proliferative DR cases but classify an unacceptable number of mild or moderate grade NPDR as normal.

A related problem that is well described in the ML literature is class imbalance, where classes (or disease categories) are not equally represented in the training dataset.[32] This may cause the opposite problem to that seen in the previous example. For example, a population-based dataset of fundus images from patients with diabetes reflecting normal prevalence will mostly be normal, mild NPDR, or moderate NPDR; only a small number will be proliferative DR cases. A model trained on these data can learn to favor the more prevalent mild or moderate DR class as a diagnosis based on disease probability, rather than salient pathological features in the image. To tackle this problem, algorithm developers often adopt various methods to balance the classes (either adding copies of the underrepresented class: oversampling or taking away instances of the overrepresented class: undersampling). Although this commonly used technique is helpful in algorithm training, investigators sometimes replicate the class distribution in the validation test set, which is most likely to ensure optimum model performance, even if it is an unrealistic disease prevalence. Reporting results in this way is somewhat unhelpful because it becomes difficult to extrapolate whether the same level of accuracy can be replicated in a real patient cohort.

Therefore, important considerations include: is the disease prevalence in the internal validation test dataset representative of the target population in the real world? Are there under- or overrepresented subgroups within the training dataset? Have the authors applied any inclusion or exclusion criteria which create a selection bias? Have the authors applied a sampling method (i.e., random sampling) to reduce the risk of spectrum bias?

Is the Output of the Model Interpretable and Can it Be Interrogated? Are Differential Diagnoses and Estimates of Confidence Provided?

As a clinician, knowledge about the probability of disease presence within a list of differential diagnoses (and its associated estimates of confidence) is a necessity for good clinical decision-making. Adequate reflection of the diagnostic process by providing supporting evidence alongside a diagnosis suggested by an ML-based tool may complement the capability to interrogate the system thereby facilitating clinical adoption. Imagine a situation where, despite robust diagnostic validation, there remains disagreement between a clinician's and an ML algorithm's diagnosis. When rationalizing treatment decisions and managing risks, having the ability to interpret and interrogate recommendations made by ML models is vital. For successful clinical adoption of ML-based tools, clinician trust is key.[31,33]

In non-ML predictive modeling, input parameters of a model may have been chosen in a hypothesis-driven and rule-based manner.[27,34] For instance, considering today's pathophysiological understanding of diabetic retinopathy (based on a biological model), hemoglobin A1c would be a reasonable indicator to include in a model for predicting progression of retinopathy.[35] On the contrary, common ML techniques for image-based diagnosis in ophthalmology, such as deep learning, may potentially use thousands of inscrutable input parameters fed into a complex model of weighted connections to create data-driven predictions without any supporting evidence.[33,36,37]

Understandably, this way of modelling stays fairly abstract to the human mind ("black box" decision making) and in turn makes it harder to detect bias, overfitting, and confounding. A recent example of a deep learning model for detection of pneumonia demonstrates this point nicely. The resulting model performed very well, but was found to be exploiting confounding variables in the images such as noting if the scan was taken on a portable machine (exclusively used in sicker patients).[38] Similarly, an algorithm for the detection of skin cancer from dermoscopic images of skin lesions was found to be using the presence of skin markings within the photograph as an indicator of likely malignancy.[39]

Concerns have therefore been raised regarding trustworthiness, particularly within the field of medicine, because of the potentially impactful nature of decisions.[33] Several methods evolving around visualization have been suggested to mitigate this issue (i.e., occlusion, saliency, or class activation maps).[40,41] These techniques provide a way of visualizing key predictive features within an image, and can at least allow a degree of recognition on whether irrelevant features were used as predictors.

Is the Performance Reproducible and Generalizable?

For the successful adoption of an ML diagnostic tool in clinical practice, it is important that the predictive accuracy has been shown to be robust beyond the cohort they have been developed in (external validation). It is a known phenomenon that classification performance of predictive models, including ML models, can be overestimated in internal validation alone.[42] External validation should be considered as a continuum rather than a single event. External validation may include: evaluation in a dataset that is independent of the original dataset but similar in terms of its setting and population; evaluation in a dataset that is independent, but differs in either the population (e.g., ethnicity, socioeconomic status) or the setting (e.g., screening, primary care, secondary care); evaluation in the same or new populations over time to test for degradation of the model performance as the population evolves; evaluation in a dataset that differs for technical reasons (e.g., images taken on different scanners). These factors may profoundly affect the performance of a model and highlight the need for reproducibility and generalizability to be evaluated. For instance, an ML model developed on an unselected database of images from a DR screening service is likely to perform optimally in that setting, and less well in either a primary optometry setting or a hospital eye service.

For diagnostic tests outside ML, it has already been established that there is a lack of validation studies.[42] External validation in ML-based diagnostic models is arguably even more important because of the "black box nature" of these systems and the inability to interrogate the models' decisions. In a recent systematic review assessing studies that used deep learning to diagnose diseases from medical imaging, only 6 of 18 studies within ophthalmology reported on external validity. Vollmer and colleagues suggested that insufficient reporting on the particularities of ML models may delay opportunities for external validation by an independent research group.[31]

A specific pitfall for ML applications developing models based on high-dimensional data (such as ophthalmic images) is "overfitting." Imaging is considered high-dimensional data, because theoretically, each

pixel or voxel may be considered as a separate input parameter into the model. To develop an ML model on images or scans may therefore require a substantial amount of training data to avoid overparameterized classification models in which the input parameters are too numerous in relation to the sample size. In turn, overparameterized classification models may be prone to overfitting, which refers to a situation in which there are insufficient outcome events relative to the number of predictors. Overfitting is a common problem in high-dimensional data. The model adjusts to spurious signals and unimportant findings within the training data, and this results in poor generalizability to new data while giving overly optimistic estimates of performance.[25,43,44]

In summary, to test the generalizability of the algorithm's performance, authors should therefore seek to externally validate their results in an out-of-sample external validation to avoid overly optimistic estimates. This should be done in a temporally, or preferably geographically, separate study population, and ideally by an independent research group.[45] Validation should be considered not as a one-off event but as an ongoing process to test performance in any new arenas of intended use, and to ensure that performance is maintained over time.

Conclusion

Machine learning algorithms have shown significant potential for offering expert-level diagnostic capability across a wide range of diseases. Because of the increasing ease with which models can be generated on publicly available datasets, there will be an increasing deluge of reports of AI diagnostics and other interventions claiming impressive sensitivity and specificity. Enthusiasm around this novel technology should not overrule the need for robust critical appraisal, and this will require an increasing community of people with expertise bridging the worlds of medicine, statistics, and computer science. This brief article is intended as an introduction to some of the key points to consider when critically appraising studies reporting ML applications in clinical medicine, particularly around image-based classifiers. New standards specific to reporting studies of ML interventions in health care such as TRIPOD-ML, SPIRIT-AI, and CONSORT-AI are in development, and it is hoped that they will lead to improvements in the design and reporting of such studies.[15–17] Improved understanding of the field will help readers decide whether they can have confidence in the study findings, and whether the results are generalizable and clinically applicable. Although evaluation of diagnostic accuracy is a key step of the validation process, understanding the effect of any given algorithm on patient outcomes requires assessment in the context of the whole patient pathway with a focus on patient outcome, and ideally within the context of a prospective randomized clinical trial. This is particularly important in black box algorithms in which a lack of understanding of the underlying model increases the risk of unexpected negative consequences, which may only be seen after implementation of the device within a clinical pathway. Provided we can achieve the appropriate scientific evaluation and real-world regulation of ML health-related interventions, this exciting tool can fulfil its potential to be a powerful technology for patient benefit and health system improvement.

Acknowledgments

There was no funding source for this study. Keane is supported by a National Institute for Health Research (NIHR) Clinician Scientist Award (NIHR-CS-2014-14-023). The research was supported by the NIHR Biomedical Research Centre based at Moorfields Eye Hospital National Health Service Foundation Trust and University College London Institute of Ophthalmology. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

LF, XL, PAK and AKD contributed equally to this article.

Disclosure: **L. Faes,** Alllergan (F), Bayer (F), Novartis (F); **X. Liu,** None; **S.K. Wagner,** None; **D.J. Fu,** None; **K. Balaskas,** Alimera (F), Allergan (F), Bayer (F), Heidelberg Engineering (F), Novartis (F), TopCon (F); **D.A. Sim,** Haag-Streit (F), Allergan (F), Novartis (F), and Bayer (F). Allergan (S), Bayer (S), Big Picture Eye Health (C); **L.M. Bachmann,** Oculocare (E); **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); **A.K. Denniston,** None

References

1. 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.

2. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
3. Becker AS, Mueller M, Stoffel E, Marcon M, Ghafoor S, Boss A. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *Br J Radiol*. 2018;91:20170576.
4. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med*. 2018;15:e1002699.
5. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. 2017;135:1170–1176.
6. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15:e1002686.
7. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
8. Breakthrough Devices Program. U.S. Food and Drug Administration. http://www.fda.gov/medical-devices/how-study-and-market-your-device/breakthrough-devices-program. Published September 7, 2019. Accessed September 28, 2019.
9. 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. *Lancet Digit Health*. 2019;1:PE271–PE297.
10. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*. 1996;276:637–639.
11. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *PLoS Med*. 2009;6:e1000097.
12. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326(7379):41–44.
13. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement (vol 162, pg 55, 2015). *Ann Intern Med*. 2015;162:600.
14. Cruz-Herranz A, Balk LJ, Oberwahrenbrock T, et al. The APOSTEL recommendations for reporting quantitative optical coherence tomography studies. *Neurology*. 2016;86:2303–2309. doi:10.1212/wnl.0000000000002774
15. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393:1577–1579. doi:10.1016/s0140-6736(19)30037-6
16. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. 2019;25:1467–1468. doi:10.1038/s41591-019-0603-3
17. Liu X, Faes L, Calvert MJ, Denniston AK, CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. *Lancet*. 2019;394:1225. doi:10.1016/S0140-6736(19)31819-7
18. Bachmann LM, ter Riet G, Weber WEJ, Kessels AGH. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. *J Clin Epidemiol*. 2009;62:357–361.
19. Khan KS, Bachmann LM, ter Riet G. Systematic reviews with individual patient data meta-analysis to evaluate diagnostic tests. *Eur J Obstet Gynecol Reprod Biol*. 2003;108:121–125.
20. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006;332:1089–1092.
21. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
22. Korevaar DA, Ochodo EA, Bossuyt PMM, Hooft L. Publication and reporting of test accuracy studies registered in ClinicalTrials.gov. *Clin Chem*. 2014;60:651–659.
23. Korevaar DA, Hooft L, Askie LM, et al. Facilitating Prospective Registration of Diagnostic Accuracy Studies: a STARD Initiative. *Clin Chem*. 2017;63:1331–1341. doi:10.1373/clinchem.2017.272765
24. Goldacre B, Drysdale H, Powell-Smith A, et al. The COMPare trials project. *COMPare-trials org*. 2016.
25. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286:800–809.
26. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312:1342–1343.
27. Miettinen OS, Caro JJ. Foundations of medical diagnosis: what actually are the parameters

involved in Bayes' theorem? *Stat Med*. 1994;13:201–209; discussion 211–215.
28. Miettinen OS, Henschke CI, Yankelevitz DF. Evaluation of diagnostic imaging tests: diagnostic probability estimation. *J Clin Epidemiol*. 1998;51:1293–1298.
29. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the MESSIDOR database. *Image Anal Stereol*. 2014;33:231.
30. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122–1131.
31. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. *arXiv*. 2018. http://arxiv.org/abs/1812.10404.
32. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061–1066.
33. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319:19–20.
34. Miettinen OS, Bachmann LM, Steurer J. Towards scientific medicine: an information-age outlook. *J Eval Clin Pract*. 2008;14:771–774. doi:10.1111/j.1365-2753.2008.01078.x
35. Progression of retinopathy with intensive versus conventional treatment in the Diabetes Control and Complications Trial. Diabetes Control and Complications Trial Research Group. *Ophthalmology*. 1995;102:647–661.
36. Artificial intelligence in health care: within touching distance. *Lancet*. 2018;390:2739.
37. Kahn CE, Jr. From images to actions: opportunities for artificial intelligence in radiology. *Radiology*. 2017;285:719–720.
38. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv*. 2018. http://arxiv.org/abs/1807.00431.
39. Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. 2019.
40. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv [csCV]*. 2013. http://arxiv.org/abs/1312.6034.
41. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainabilty of artificial intelligence in medicine. *WIREs Data Mining Knowl Discov*. 2019;15:e1312.
42. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40.
43. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216–1219.
44. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–W73.
45. Moons KGM, Altman DG, Reitsma JB, Collins GS. New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the TRIPOD Statement. *Adv Anat Pathol*. 2015;22:303–305.