Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study

Livia Faes, Siegfried K Wagner*, Dun Jack Fu, Xiaoxuan Liu, Edward Korot, Joseph R Ledsam, Trevor Back, Reena Chopra, Nikolas Pontikos, Christoph Kern, Gabriella Moraes, Martin K Schmid, Dawn Sim, Konstantinos Balaskas, Lucas M Bachmann, Alastair K Denniston, Pearse A Keane*

Summary

Background Deep learning has the potential to transform health care; however, substantial expertise is required to train such models. We sought to evaluate the utility of automated deep learning software to develop medical image diagnostic classifiers by health-care professionals with no coding—and no deep learning—expertise.

Methods We used five publicly available open-source datasets: retinal fundus images (MESSIDOR); optical coherence tomography (OCT) images (Guangzhou Medical University and Shiley Eye Institute, version 3); images of skin lesions (Human Against Machine [HAM] 10000), and both paediatric and adult chest x-ray (CXR) images (Guangzhou Medical University and Shiley Eye Institute, version 3 and the National Institute of Health [NIH] dataset, respectively) to separately feed into a neural architecture search framework, hosted through Google Cloud AutoML, that automatically developed a deep learning architecture to classify common diseases. Sensitivity (recall), specificity, and positive predictive value (precision) were used to evaluate the diagnostic properties of the models. The discriminative performance was assessed using the area under the precision recall curve (AUPRC). In the case of the deep learning model developed on a subset of the HAM10000 dataset, we did external validation using the Edinburgh Dermofit Library dataset.

Findings Diagnostic properties and discriminative performance from internal validations were high in the binary classification tasks (sensitivity 73·3–97·0%; specificity 67–100%; AUPRC 0·87–1·00). In the multiple classification tasks, the diagnostic properties ranged from 38% to 100% for sensitivity and from 67% to 100% for specificity. The discriminative performance in terms of AUPRC ranged from 0·57 to 1·00 in the five automated deep learning models. In an external validation using the Edinburgh Dermofit Library dataset, the automated deep learning model showed an AUPRC of 0·47, with a sensitivity of 49% and a positive predictive value of 52%.

Interpretation All models, except the automated deep learning model trained on the multilabel classification task of the NIH CXR14 dataset, showed comparable discriminative performance and diagnostic properties to state-of-the-art performing deep learning algorithms. The performance in the external validation study was low. The quality of the open-access datasets (including insufficient information about patient flow and demographics) and the absence of measurement for precision, such as confidence intervals, constituted the major limitations of this study. The availability of automated deep learning platforms provide an opportunity for the medical community to enhance their understanding in model development and evaluation. Although the derivation of classification models without requiring a deep understanding of the mathematical, statistical, and programming principles is attractive, comparable performance to expertly designed models is limited to more elementary classification tasks. Furthermore, care should be placed in adhering to ethical principles when using these automated models to avoid discrimination and causing harm. Future studies should compare several application programming interfaces on thoroughly curated datasets.

Funding National Institute for Health Research and Moorfields Eye Charity.

Introduction

Diagnosis, an understanding of the probability for presence of illness, depends on data: its collection, integration, and interpretation enables accurate classification of clinical presentations into an accepted disease category. Human diagnosticians achieve acceptable accuracy in such classification tasks through the learning of diagnostic rules (patterns recorded by other human diagnosticians) followed by training on real cases for which the diagnostic labels are provided (supervised clinical experience). In artificial intelligence (AI), the technique of deep learning uses artificial neural networks—so-called because of their superficial resemblance to biological neural networks—as a computational model to discover intricate structure and patterns in large, high-dimensional datasets such as medical

Lancet Digital Health 2019; 1: e232–42

See **Comment** page e198

*Contributed equally

Department of Ophthalmology, Cantonal Hospital Lucerne, Lucerne, Switzerland (L Faes MD, Prof M K Schmid MD); **National Institute of Health Research Biomedical Research Center, Moorfields Eye Hospital National Health Service Foundation Trust, and University College London Institute of Ophthalmology, London, UK** (S K Wagner BMBCh, X Liu MBChB, R Chopra BSc, N Pontikos PhD, D Sim PhD, K Balaskas MD, Prof A K Denniston PhD, P A Keane MD); **Medical Retina Department, Moorfields Eye Hospital National Health Service Foundation Trust, London, UK** (L Faes, S K Wagner, D J Fu PhD, E Korot MD, R Chopra, C Kern MD, G Moraes MD, D Sim, K Balaskas MD, P A Keane); **Department of Ophthalmology, University Hospitals Birmingham National Health Service Foundation Trust, Birmingham, UK** (X Liu, Prof A K Denniston); **Academic Unit of Ophthalmology, Institute of Inflammation & Ageing, University of Birmingham, Birmingham, UK** (X Liu, Prof A K Denniston); **Beaumont Eye Institute, Royal Oak, Michigan** (E Korot MD); **DeepMind, London, UK** (J R Ledsam MBChB, T Back PhD, R Chopra); **Department of Ophthalmology, University Hospital of Ludwig Maximilian University, Munich, Germany** (C Kern); **Medigntion, Zurich, Switzerland** (Prof L M Bachmann PhD); **and Centre for Patient Reported Outcome Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK** (Prof A K Denniston)

Correspondence to:
Dr Pearse A Keane, National Institute of Health Research Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital National Health Service Foundation Trust and University College London Institute of Ophthalmology, London EC1V 2PD, UK
pearse.keane1@nhs.net

For **Google Cloud AutoML** see https://cloud.google.com/automl/

Research in context

Evidence before this study

We did a systematic search of the literature to identify classical deep learning models (with bespoke architectures developed by human experts) to provide a comparison to models constructed using the automated deep learning platform Google Cloud AutoML. We searched MEDLINE, Embase, Science Citation Index, Conference Proceedings Citation Index, Google Scholar, and arXiv from Jan 1, 2012, until Oct 5, 2018. We prespecified a cut-off of 2012, on the basis of a generally recognised step-change in deep learning performance since that time. No language restrictions were applied. Studies were included if the authors developed a deep learning algorithm on the datasets used in their study. The search strategy is available in the appendix. The performance of these existing models served as a direct comparison to the model we developed.

Our search resulted in five open-source datasets: retinal fundus images (MESSIDOR), optical coherence tomography images (Guangzhou Medical University and Shiley Eye Institute, version 3), images of skin lesions (Human Against Machine HAM10000), paediatric chest x-ray images (Guangzhou Medical University and Shiley Eye Institute, version 3), and adult chest x-ray images (NIH CXR14 dataset).

Added value of this study

Currently, the highly specialised technical expertise necessary to develop artificial intelligence (AI) models is scarce. Even transfer learning, which builds on existing algorithms, requires substantial machine learning experience to achieve adequate results on new image classification tasks. This currently limits the use of deep learning to a growing, but small, community of computer scientists and engineers.

We show, to our knowledge, a first of its kind automated design and implementation of deep learning models for health-care application by non-AI experts, namely physicians. Although comparable performance to expert-tuned medical image classification algorithms was obtained in internal validations of binary and multiple classification tasks, more complex challenges, such as multilabel classification, and external validation of these models was insufficient.

Implications of all the available evidence

We believe that AI might advance medical care by improving efficiency of triage to subspecialists and the personalisation of medicine through tailored prediction models. The automated approach to prediction model design improves access to this technology, thus facilitating engagement by the medical community and providing a medium through which clinicians can enhance their understanding of the advantages and potential pitfalls of AI integration. We believe that this Article is novel in its discussion of the scientific and clinical features. Moreover, it addresses health-care delivery aspects and challenges of this technology.

images.[1] A key feature of these networks is their ability to fine-tune on the basis of experience, allowing them to adapt to their inputs, thus becoming capable of evolving. This characteristic makes them powerful tools for pattern recognition, classification, and prediction. In addition, the features discovered are not predetermined by human engineers, but rather by the patterns they have learned from input data.[2] Although first espoused in the 1980s, deep learning has come to prominence in the past 10 years, driven in large part by the power of graphics processing units originally developed for video gaming, and the increasing availability of large datasets.[3] Since 2012, deep learning has brought profound changes to the technology industry, with discoveries in areas as diverse as computer vision, image caption, speech recognition, natural language translation, robotics, and even self-driving cars.[4–9] In 2015, *Scientific American* listed deep learning as one of their "world changing ideas" of the year.[10]

Until now, the development and implementation of deep learning methodology into health care has faced three main blockers. First, access to large, well curated, and well labelled datasets is a major challenge. Although numerous institutions around the world have access to large clinical datasets, far fewer have them in a computationally tractable form and with reliable clinical labels for learning tasks. Second, highly specialised computing resources are needed, because the performance of deep learning models depend on recent advances in parallel computing architectures, termed graphic processing units. The architecture of silicon customised to these tasks is rapidly evolving with software companies increasingly designing their own hardware chips, such as tensor processing units, and field-programmable gate arrays.[11,12] Thus, it is already clear that it will be difficult for small research groups, working alone in hospital and university settings, to accommodate these financial costs and the rapidly evolving landscape. Third, specific technical expertise and mathematical knowledge is required to develop deep learning models. This proficiency is still uncommon. A 2019 report[13] by Element AI concluded that although the number of self-reported AI experts worldwide had increased to 36 000, the "supply of top-tier AI talent does not meet the demand".

One approach to combat these obstacles is the increasingly popular technique called transfer learning, where a model developed for a specific task is repurposed and leveraged as a starting point for training on a novel task. Although transfer learning mitigates some of the substantial computing resources required in designing a bespoke model from inception, it nevertheless demands deep learning expertise to deliver effective results. With this in mind, several companies released application

programming interfaces (API) in 2018, claiming to have automated deep learning to such a degree that any individual with basic computer competence could train a high-quality model.[14,15]

Because programming is not a common skill among health-care professionals, automated deep learning is a potentially promising platform to support the dissemination of deep learning application development in health care and medical sciences. In the case of classification tasks, these products automatically match generic neural network architectures to a given imaging dataset, fine tune the network aiming at optimising discriminative performance, and create a prediction algorithm as output. In other words, the input is a (labelled) image dataset, and the output is a custom classifying algorithm. Yet, the extent to which people without coding experience can replicate trained deep learning engineers' performance with the help of automated deep learning remains unclear.

In this study, two physicians without any deep learning expertise explored the feasibility of automated deep learning model development and investigated the performance of these models in diagnosing a diverse range of disease from medical imaging. More precisely, we identified medical benchmark imaging datasets for diagnostic image classification tasks and their corresponding publications on deep learning models; used these datasets as input; replaced the classic deep learning models with automated deep learning models; and compared the discriminative performance of the classic and the automated deep learning models. Moreover, we sought to evaluate the interface that was used for automated deep learning model development (Google Cloud AutoML Vision API, beta release) for its use in prediction model research.[16–18]

Methods

Study design and data source

We used five distinct open-source datasets comprising medical imaging material to automatically develop five corresponding deep learning models for the diagnosis of common diseases or disease features. Namely, we trained deep learning models on retinal fundus images (the MESSIDOR dataset; hereafter referred to as retinal fundus image set); retinal optical coherence tomography (OCT) images (Guangzhou Medical University and Shiley Eye Institute Version 3; hereafter referred to as the retinal OCT set); paediatric chest x-ray (CXR) images (Guangzhou Medical University and Shiley Eye Institute; hereafter referred to as the paediatric CXR set); adult CXR images (National Institute of Health [NIH] CXR14 dataset; hereafter referred to as the adult CXR set); and dermatology images (the Human Against Machine [HAM] 10000 dataset; hereafter referred to as the dermatology image set).[19–22] Moreover, in a proof-of-principle evaluation, we aimed to test one of the models out of sample, as recommended by current guidelines.[23] The current version of the API only allows single image upload for model prediction, limiting the feasibility of large-scale external validation. However, in an effort to emulate external validation in one exemplary use-case, the authors used the dermatology image set for training and tuning of a deep learning model and tested its performance by using a separate skin lesion image dataset, the Edinburgh Dermofit Library (hereafter referred to as the dermatology validation set).[24]

Training of health-care professionals using the Graphical User Interface

Two physicians (LF and SKW) with no previous coding or machine learning experience did the model development and analysis after a period of self-study. This self-study consisted of basic competence training in shell script programming to allow expedient transfer of the large medical imaging datasets into a cloud-based Google bucket; familiarisation with the Google Cloud AutoML online documentation and graphical user interface; and preparation of benchmark datasets with randomisation of images to corresponding training, validation, or testing datasets as applicable. In total, each researcher invested approximately 10 h of training time. Because of the release cycle evolution of the Google Cloud AutoML Vision API during the study (alpha release May, 2018, beta release July, 2018), they adopted an iterative approach when executing the analyses. All analytic steps and interpretations of results were done jointly. Interaction with the AutoML Cloud Vision platform was through a graphical user interface (video).

See Online for video

Patient recruitment and enrolment

We accessed five de-identified open-source imaging datasets that were collected from retrospective, non-consecutive cohorts, showing diseases or disease features of common medical diagnoses. Eligibility criteria, patient demographics and patient workflow for each of these datasets are published elsewhere.[19–22]

Index test: AutoML Cloud Vision API

The term automated machine learning commonly refers to automated methods for model selection or hyperparameter optimisation. This is the concept that led to the idea of allowing a neural network to design another neural network, through the application of a neural architecture search.[16–18] In deep learning, designing and choosing the most suitable model architecture requires a substantial amount of time and experimentation even for those with extensive deep learning expertise. This time and experimentation is because the search space of all possible model architectures can be exponentially large (eg, a typical ten-layer network could have approximately 1×10^{10} candidate networks). To make this model design process easier and more accessible, an approach known as neural architecture search has been described.[25] Neural

For more on the **retinal optical coherence tomography images** see http://dx.doi.org/10.17632/rscbjbr9sj.3

architecture search is typically achieved using one of two methods: reinforcement learning algorithms, and evolutionary algorithms. Reinforcement learning algorithms forms the basis of the commercially available API evaluated in this study.[26]

Data handling and analytic approach

We uploaded images of five open-source datasets to a Google Cloud bucket in conjunction with a comma-separated value file indicating the image label, file path, and dataset distribution (ie, training, validation, or test dataset) using shell script programming. Upload, and model development and evaluation are shown in the video. Images were allocated to the training, validation, and test datasets (80%, 10%, and 10%, respectively) using a random number function in Microsoft Excel. In the case of the retinal OCT images, where a specific test set had been stipulated in a previous report, we mirrored the same test set to both provide a direct comparison between model performance and to uphold the patient-wise independence between the training and test sets.[20] Duplicate images were automatically detected and excluded by the API. We did not relabel any of the used datasets. All models were trained for a maximum of 24 compute hours. Except for the retinal OCT set, the discriminative performance of each deep learning model was evaluated using the randomly specified test dataset, and in the case of the deep learning model developed on a subset of the dermatology image set, additionally in an external validation using an independent open-source dermatology dataset (Edinburgh Dermofit Library).

Comparison with benchmark classic deep learning models

To be able to make a direct comparison with the performance of classic deep learning models developed using traditional non-autoML techniques (deep learning models with bespoke architectures for a data and problem set developed by human experts), we did a systematic search of the literature to identify classical deep learning models, composed by deep learning experts, which have been trained or validated, or both, on the five open-source datasets. The performance of these existing models served as a direct comparator with the API. We searched MEDLINE, Embase, Science Citation Index, Conference Proceedings Citation Index, Google Scholar, and arXiv from Jan 1, 2012, until Oct 5, 2018. Studies were included if they developed a deep learning algorithm on the datasets used in this study. No language restrictions were applied. The search strategy is available in the appendix (p 1). We prespecified the cut-off of 2012, on the basis of a step-change in deep learning performance; a deep learning model called AlexNet won a visual recognition challenge, the ImageNet Large-Scale Visual Recognition Challenge, for the first time.[27] If a study provided contingency tables for the same or for separate algorithms tested in a specific classification task, we assumed these to be independent from each other. We accepted this, because we were interested in providing an overview of the results of various studies rather than providing accurate point estimates.

See Online for appendix

Statistical analysis

The AutoML Cloud Vision API provides metrics that are commonly used by the AI community. These are recall (sensitivity) or precision (positive predictive value) for given thresholds and the area under the precision recall curve (AUPRC). Additionally, confusion matrices are depicted for each model, cross-tabulating ground truth labels versus the labels predicted by the deep learning model. Where possible, we extracted binary diagnostic accuracy data and constructed contingency tables and calculated specificity at the threshold of 0·5. Contingency tables consisted of true-positive, false-positive, true-negative, and false-negative results. If a dataset tackled a multiclass problem, we constructed two-by-two tables for each of the disease labels versus normal. For consistency, we adhered to the typical test accuracy terminology: sensitivity (recall), specificity, and positive predictive value (precision). The classification tasks were chosen according to their popularity in the current AI literature for the purpose of comparability to state-of-the-art deep learning models. Where possible, we plotted contingency tables against the ones reported by other studies using the same benchmark datasets to develop deep learning models.

A priori, we attempted to compare the classification performance between state-of-the-art deep learning studies and our results. However, although the published reports provided areas under the receiver operating characteristic curve (AUC ROC), the AutoML Cloud Vision API reports the AUPRC. Although the points of the two types of curves can be mapped one-to-one and hence curves can be translated from the ROC space to the prediction space (if the confusion matrices are identical) differences in the confusion matrices and the level of reporting impeded us from doing a comparison on the level of AUC. Instead, we compared the performance on the level of sensitivity and specificity at the same threshold as had been used in the previous reports.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

We assessed performance of automated deep learning in archetypal binary classification tasks, and found diagnostic properties and discriminative performance were comparable in the case of the investigated binary classification tasks.

Task 1 involved classification of diabetic retinopathy versus normal retina on fundus images. The retinal

	Prevalence*	True positives	False positives	True negatives	False negatives	Area under the precision-recall curve	Positive predictive value	Sensitivity	Specificity
MESSIDOR: fundus images (2014)									
Presence *vs* absence of DR	55%	48	18	36	18	0·87	73%	73%	67%
Guangzhou Medical University and Shiley Eye Institute: retinal OCT images (2018)									
Overall	100%	NR	NR	NR	NR	0·99	98%	97%	100%
CNV *vs* others	25%	246	2	973	1	NR	99%	100%	100%
Drusen *vs* others	24%	208	0	975	23	NR	100%	90%	100%
DMO *vs* others	25%	247	1	974	0	NR	100%	100%	100%
Normal *vs* others	26%	250	0	975	0	NR	100%	100%	100%
Guangzhou Medical University and Shiley Eye Institute: paediatric CXR images (2018)									
Pneumonia *vs* normal	74%	412	6	153	10	1	97%	97%	100%
NIH CXR14: adult CXR images (2017)									
Overall	NR	NR	NR	NR	NR	0·57†	71%	38%	NR
HAM10000: dermatology image set (2018)									
Overall	100%	NR	NR	NR	NR	0·93	91%	91%	NR
Actinic keratosis *vs* others	3%	25	9	961	8	NR	74%	76%	99%
Basal cell carcinoma *vs* others	5%	46	8	943	6	NR	85%	88%	99%
Nevus *vs* others	67%	655	47	286	15	NR	93%	98%	86%
Melanoma *vs* others	11%	75	12	879	37	NR	86%	67%	99%
Dermatofibroma *vs* others	1%	7	3	988	5	NR	70%	58%	100%
Vascular lesion *vs* others	1%	13	0	989	1	NR	100%	93%	100%
Benign keratosis *vs* others	1%	91	10	883	19	NR	90%	83%	99%

NR=not reported. DR=diabetic retinopathy. CNV=choroidal neovascularisation. DMO=diabetic macular oedema. OCT=optical coherence tomography. *Number of given cases as percentage of test dataset. †Averaged across different classes of this multiclass model.

Table 1: **Summary of the diagnostic properties and the discriminative performance of all five automated deep learning models**

fundus image dataset involved 1187 images, with 533 normal fundus images (R0 cases), 153 images showing mild diabetic retinopathy (R1 cases), 247 moderate diabetic retinopathy (R2 cases), and 254 severe diabetic retinopathy (R3 cases). 13 duplicate images were automatically excluded by the API. The automated deep learning model trained to distinguish healthy fundus images from fundus images showing diabetic retinopathy (R0 from R1, R2, and R3 cases) reached an AUPRC of 0·87, and best accuracy at a cut-off value of 0·5 with a sensitivity of 73·3% and a specificity of 67% (table 1).

Task 2 involved classification of pneumonia versus normal on paediatric CXR. The paediatric CXR set provided by Guangzhou Medical University and Shiley Eye Institute involved 5827 of 5232 patients' CXR images (1582 showing normal paediatric chest x-rays, and 4245 showing pneumonia). The API detected and excluded eight duplicate images. The AUPRC of this automated deep learning model was 1, best accuracy was reached at a cut-off value of 0·5 with a sensitivity of 97% and a specificity of 100%.

We next assessed performance of automated deep learning in multiple classification tasks. The two models trained to distinguish multiple classification tasks showed high diagnostic properties and discriminative performance.

Task 1 involved classification of three common macular diseases and normal retinal OCT images. The retinal OCT set provided by Guangzhou Medical University and Shiley Eye Institute involved 101418 images of 5761 patients. 31882 images depicted OCT changes related to neovascular age-related macular degeneration (791 patients), 11165 to diabetic macular oedema (709), 8061 depicted drusen (713 patients), and 50310 were normal (3548 patients). 175 images were identified as duplicates and excluded by the API. The AUPRC of the automated deep learning model trained to distinguish these four categories was 0·99, while best accuracy was reached at a cut-off value of 0·5, with a sensitivity of 97·3%, a specificity of 100%, and a positive predictive value (PPV) of 97·7%.

Task 2 involved classification of seven distinct categories of skin lesions using dermatoscopic images. The dermatology image set involved 10013 images of

	Model architecture	Threshold	Sensitivity	Specificity
MESSIDOR: fundus images (2014)				
Google Cloud AutoML	Automated deep learning	0·5	73%	67%
Li et al[30]	VGG-s and Conv1-Fc8	0·5	86%	97%
Guangzhou Medical University and Shiley Eye Institute: retinal OCT images				
Google Cloud AutoML	Automated deep learning	0·5	97%	100%
Kermany et al[20]	Inception V3	NR	98%	97%
Guangzhou Medical University and Shiley Eye Institute: paediatric CXR images				
Google Cloud AutoML	Automated deep learning	0·5	97%	100%
Kermany et al[20]	Inception V3	NR	93%	90%
NIH CXR14: adult CXR images (2017)				
Google Cloud AutoML	Automated deep learning	0·5/0·7	38%/23%	NR
Guan et al[29]	ResNet-50 and DenseNet-121	0·7	NR	NR

NR=not reported. OCT=optical coherence tomography.

Table 2: **Image classification performance of algorithms trained using automated deep learning compared to best performing algorithms found in the literature**

	Drusen	CNV (predicted label)	DMO	Normal
Drusen	0·9	0·1	..	..
CNV	..	0·996	0·004	..
DMO (true label)	..	..	1	..
Normal	..	..	..	1

CNV=Choroidal neovascularisation. DMO=Diabetic macular oedema.OCT=optical coherence tomography.

Table 3: **The confusion matrix for the model developed on the OCT image dataset provided by Guangzhou Medical University and Shiley Eye Institute**

skin lesions of 10 013 patients (327 images depicted actinic keratosis, 514 basal cell carcinoma, 6703 naevus, 1113 melanoma, 115 dermatofibroma, 142 vascular lesion, and 1099 benign keratosis consisting of seborrheic keratosis, solar lentigo, and lichen-planus like keratoses). There were no duplicate images detected. The AUPRC of the automated deep learning model trained to distinguish these seven categories was 0·93, while best accuracy was reached at a cut-off value of 0·5, with a sensitivity of 91% and a positive predictive value of 91%.

We then assessed performance of automated deep learning in multilabel classification tasks, and found the automated deep learning model trained to perform this task on the adult CXR dataset showed poor diagnostic properties and a discriminative performance near chance (AUPRC 0·57, best accuracy at a cut-off value of 0·5, with a sensitivity of 38% and a positive predictive value of 71%). The NIH CXR14 comprised 11 542 cases of atelectasis, 2399 of cardiomegaly, 3323 of consolidation, 1862 of oedema, 8036 of effusion, 1734 of emphysema, 1215 of fibrosis, 156 of hernia, 11 785 of infiltration, 2923 of mass, 3009 of nodule, 1216 of pleural thickening, 325 of pneumonia, and 2199 of pneumothorax, and 60 304 with no findings of 112 120 patients. 12 duplicates were detected and excluded by the API.

We compared the diagnostic properties and the diagnostic performance of algorithms trained using automated deep learning on the retinal fundus image, retinal OCT, paediatric CXR, adult CXR, and dermatology image datasets compared with best performing deep learning algorithms (table 2). Interestingly, all best performing algorithms used transfer learning. Some automated deep learning models showed comparable diagnostic properties at a threshold of 0·5 to state-of-the-art deep learning algorithms in published literature. For example, using the OCT dataset, automated deep learning achieved a sensitivity of 97% and a specificity of 100% (*vs* a sensitivity of 98% and a specificity of 97% published by Kermany and colleagues[20]); using the paediatric CXR dataset, automated deep learning reached a sensitivity of 97% and a specificity of 100% (*vs* a sensitivity of 93% and a specificity of 90% published by Kermany and colleagues[20]). Other models showed lower diagnostic properties; ie, using the multilabel classification task of the NIH CXR14 dataset in which automated deep learning reached sensitivity of 38%, a positive predictive value of 71%, and an AUPRC of 0·57 (*vs* an AUC of 0·87 published by Guan and colleagues[28]); or using the retinal fundus image dataset, in which automated deep learning reached a sensitivity of 73% and a specificity of 67% (*vs* a sensitivity of 86% and a specificity of 97% published by Li and colleagues[29]). The thresholds were reported in two cases.[28,30] Although it is difficult to determine why multilabel classification model performance was significantly worse, a number of factors might have contributed to this poor performance, including dataset image or labelling ground truth quality, or less likely an inherent weakness of the AutoML platform for multilabel classification.

The AutoML Cloud Vision API provides confusion matrices in the case of single-label classification tasks to uncover label categories in which the model performs insufficiently. The model trained to distinguish the four ophthalmic diagnoses from OCT images (Guangzhou Medical University and Shiley Eye Institute), classified drusen as choroidal neovascularisation (CNV) in 10% of cases, implicating a more urgent referral than needed. The model trained on the dermatology image set on the other hand, misclassified 28·6% of melanomas as naevus, which in a real-world setting would result in less urgent referral for further work-up and delayed, or worse, missed diagnosis. Moreover, this model also had a high misclassification rate (41·7%) for images showing dermatofibromas. Tables 3 and 4 show the corresponding confusion matrices for the OCT and dermatology image set and the figure shows cases from each model where the incorrect label was predicted are shown in tables 3 and 4.

In the case of the deep learning model developed on a subset of the dermatology image set, we additionally did an external validation using the dermatology validation set. As the latter set did not include benign keratosis as a

	Melanoma	Naevus	Benign keratosis	Actinic keratosis (predicted label)	Basal cell carcinoma	Dermatofibroma	Vascular skin lesion
Melanoma	0·67	0·286	0·027	..	0·018	..	..
Naevus	1	0·978	0·003	0·001	0·004	0·003	..
Benign keratosis	0·027	0·091	0·827	0·045	..	0·009	..
Actinic keratosis (true label)	..	0·061	0·121	0·758	0·061	..	..
Basal cell carcinoma	0·019	0·019	0·038	0·038	0·885	..	..
Dermatofibroma	0·167	0·167	..	0·083	..	0·583	..
Vascular skin lesion	..	..	..	..	0·071	..	0·929

Table 4: **The confusion matrix for the model developed on the dermatology image set.**

label, these images were removed from the dermatology image set used for training.

The automated deep learning model showed poor diagnostic properties and a discriminative performance near coin tossing (AUPRC 0·47, best accuracy at a cut-off value of 0·5, with a sensitivity of 49% and a positive predictive value of 52%). Of note, the sensitivity for melanoma classification is 11% with a misclassification rate of 63·7%.

Interestingly, naevus was the most likely classification in all cases, followed by the ground truth. The only exception was the case of actinic keratosis, where its ground truth diagnosis was the third most probable diagnosis to be predicted (10·6% likelihood). The prevalence of images showing naevus was 76% in the developmental dataset, compared with 36% in the dataset used for external validation. To investigate the effect of class imbalance between the dermatology image set and dermatology validation set, we undersampled the naevus class by 3000 images and assessed the resulting change in accuracy. Only minimal improvements were noted in its discriminative performance and diagnostic properties (data available on request). The external validation of the model trained on a subset of the dermatology image set are shown in tables 5 (discriminative performance and diagnostic properties) and table 6 (confusion matrix).

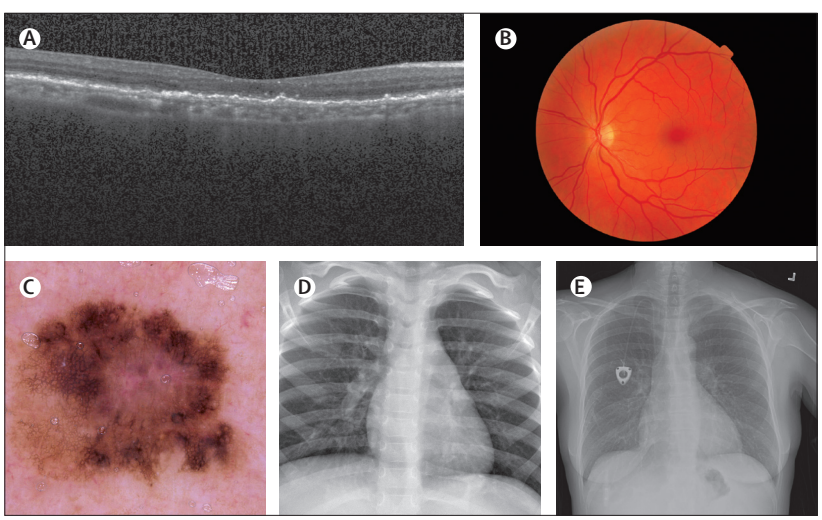

Figure: **Cases from each model where the incorrect label was predicted**
(A) Case of drusen, which was predicted as neovascular age-related macular degeneration. (B) Presence of diabetic retinopathy predicted as normal. (C) A melanoma predicted as a naevus. (D) Pneumonia predicted as normal. (E) A pleural effusion predicted as normal. Case B does not have detectable features of its label while E is equivocal.

Discussion

This Article shows that physicians with no coding experience can use automated deep learning to develop algorithms that can do clinical classification tasks to a level comparable with traditional deep learning models that have been applied in existing literature. Most of the automatically developed deep learning models, except for that trained on the multilabel classification task of the adult CXR set, showed comparable discriminative performance and diagnostic properties to state-of-the-art performing deep learning algorithms. The web interface was intuitive to use (video), although a substantial limitation was the inability to batch-test data after the model was created.

From a methodological viewpoint, our results—as is also the case with the results reported in state-of-the-art deep learning studies—might be overly optimistic, because we were not able to test all the models out of sample, as recommended by current guidelines.[23] Moreover, for external validation, the present version of the API only allows single image upload for prediction, limiting large scale external validation. This reduces its usability for systematic evaluation in prediction model research considerably, given the high numbers of images that these datasets comprise.

To circumvent this issue, we emulated an external validation of the model constructed using the dermatology image set and found a substantial reduction in the performance of the deep learning model. The limited performance of automated deep learning models in that case (also in the multilabel classification task) might be related to idiosyncrasies of the datasets used to train the models. To obviate concerns over class imbalance in our external validation, we undersampled accordingly; however, this did not substantially alter the model's discriminative performance. The reasons for the weak performance in the external validation remain unclear: one possibility is the variation in image resolution (although the dermatology validation set images were acquired in a standardised digital fashion, the dermatology

	Prevalence*	True positives	False positives	True negatives	False negatives	Area under the precision-recall curve	Positive predictive value	Sensitivity	Specificity
Overall	100%	NR	NR	NR	NR	0·47	52%	49%	NR
Actinic keratosis	13%	13	30	772	110	NR	30%	11%	96%
Basal cell carcinoma	26%	67	39	647	172	NR	63%	28%	94%
Nevus	36%	322	360	234	9	NR	47%	97%	39%
Melanoma	8%	8	8	846	68	NR	50%	11%	99%
Dermatofibroma	7%	21	5	855	44	NR	81%	32%	99%
Vascular lesion	10%	33	46	788	58	NR	42%	36%	94%

NR=not reported. *Number of given cases as percentage of test dataset.

Table 5: **The diagnostic properties and discriminative performance of the external validation of the algorithms trained using automated deep learning on the dermatology image set**

	Basal cell carcinoma	Naevus	Actinic keratosis	Melanoma (predicted label)	Vascular skin lesion	Dermatofibroma
Basal cell carcinoma	0·28	0·552	0·113	0·008	0·042	0·004
Naevus	0·003	0·973	0·003	..	0·009	0·012
Actinic keratosis	0·203	0·683	0·106	0·008	..	..
Melanoma (true label)	0·039	0·789	0·013	0·118	0·039	..
Vascular skin lesion	0·033	0·582	..	0·022	0·363	..
Dermatofibroma	0·108	0·477	0·015	0·031	0·046	0·323

Table 6: **The confusion matrix of the external validation of HAM10000-trained algorithm on the Edinburgh Dermofit Library dataset**

image set was more heterogeneous with a large proportion of digitised images), an issue which can be addressed, to some extent, through image preprocessing. Although the Google Cloud AutoML platform will select an appropriate network architecture, less attention is paid to adjustment of the input data format, such as levels per pixel and image aspect ratio.

To our knowledge, this is the first assessment of the feasibility and usefulness of automated deep learning technology in medical imaging classification tasks done by physicians with little programming experience. In this study, we showed best effort to comply with the reporting guidelines for prediction model research, and for developing and reporting machine learning predictive models.[23,31] A strength of the study is that we tested one exemplary model for robustness in an out-of-sample cross validation, because internal validation and random-split sample validation has been claimed to overestimate test performance. Another strength is that our results can easily be explored by others, given the use of public datasets and the available free trial use of the AutoML Cloud Vision API.

The sampling used in the in-sample cross validation has been claimed to introduce bias and exaggerate estimates of diagnostic performance.[32] Furthermore, the API was not able to depict saliency maps, and consequently we were not able to interrogate the model for the image areas it considered most important for its prediction.[33] This black box classification does not provide any information useful for clinical purposes.[34] Moreover, the specifics of the models used by the API are not transparent, which constitutes a barrier to their evaluation and the reproducibility of this study.[35,36] We were not able to extract or calculate all metrics and measures of uncertainty conventionally used in prediction model research in all cases (ie, specificity or confidence intervals), which impedes comparison to the current best technology other than deep learning. Moreover, explicit information on how the API determined certain metrics was scarce; for example, it was not clear how the AUPRC for a multiclassification model was calculated. There was also no ability to randomise training, test, and validation groups, while maintaining patient grouping in cases where multiple images might have come from the same patient. We noted that best accuracy was consistently achieved at a threshold of 0·5 by the API. This is likely the default threshold to which the API is optimised: a setting that is inaccessible to the user during model development. In clinical practice, thresholds should be set according to the role of the diagnostic test and the consequences of a misdiagnosis. Therefore, the ability to adjust the preferred threshold is an important function for creating a fit-for-purpose API. Model interpretability is an active area of research within the field of AI and machine learning. Although possible solutions have been proposed, further work is needed to reach a consensus solution. Such research is outside the scope of this work.[37,38]

Besides Google's AutoML Cloud Vision API, a number of vendors have released similar automated deep learning platforms, including both established technology corporations (eg, Amazon SageMaker, Baidu EZDL, IBM Watson Studio) and start-ups (Oneclick.ai, Platform.ai). Our study pertained to only one API, Google's AutoML, because this was among the first publicly available neural architecture search-based engines released, and was freely available on a trial basis.

Although this report is a proof-of-concept evaluation of health-care professional-led deep learning, it is unclear whether other APIs might provide greater discriminative performance. Assessment of other platforms is an objective of our future research.

Currently, studies on AutoML, including ours, have to rely on publicly available datasets. Although using them allows for comparison of performance between different algorithms, these are not without concern. For many classification tasks, and particularly for validation purposes, the existing datasets tend to be too small and not representative of the general population. Moreover, data quality in general could be improved. A full evaluation of dataset limitations is beyond the scope of this Article; however, inconsistent labelling and duplicate images bore direct pertinence to our study. Issues such as equivocal labels (figure), image confounders (presence of chest drain in images of pneumothorax), and label semantics (nodule *vs* mass, consolidation *vs* pneumonia) have been noted previously in datasets used for deep learning.[39] Apart from the dermatology image set, all datasets contained duplicate images. Conveniently, Google's AutoML Cloud Vision API will automatically detect, indicate, and exclude the relevant images; however, clinicians need to be cognisant to this possibility, because other APIs might not have this feature and generate spurious evaluation metrics. Because the quality of the results obtained using deep learning models substantially depends on the quality of the dataset used in the model development, it is imperative that patient demographics and information about the way the data was collected (ie, patient flow and whether multiple images from the same patient were present) is presented, because the validity and generalisation of the models would otherwise be difficult to assess. The splitting of test and training when done on a per-patient or per-test set basis might substantially affect model accuracy. However, in our study we were only able to provide limited dataset descriptions, using what has been published by their creators.

With the availability of new and carefully administered datasets, many validity problems could be resolved. Great hopes lay in data-sharing initiatives, as promoted by many peer-reviewed journals, or those from the Dataverse Network Project and the UK Data Archive.[40,41] On the other hand, these initiatives struggle with issues of confidentiality and anonymity when publishing or sharing data relating to individuals. Moreover, regulatory restrictions still remain. Fortunately, developments both in the UK with the NHS Digital Guidance and the call for Health Insurance Portability and Accountability Act compliance[24] in the USA have clarified the framework for many public Cloud systems. The EU General Data Protection Regulation is another possible barrier to an efficient use of published data; however, because many studies will be dealing with ephemeral processing of de-identified data, we do not believe that the General Data Protection Regulation is likely to pose a substantial hindrance.

We confirmed feasibility, but encountered various methodological problems that are well known in research projects performing classification tasks and predictions. We believe that concerted efforts in terms of data quality and accessibility are needed to make automated deep learning modelling successful. Moreover, as the technology evolves, transparency in the reporting of the models and a careful reporting of their performance in methodologically sound validation studies will be pivotal for a successful implementation into practice. Finally, the extent to which automated deep learning algorithms must adhere to regulatory requirements for medical devices is unclear.[35]

Although the development of deep learning prediction models was feasible for health-care professionals without any deep learning expertise, we would recommend the following developments for automated deep learning: transparency of the model architectures and technical specifications in use; reporting of established performance parameters in prediction model research, such as sensitivity, specificity reporting of the label distribution in random-split sample validations done automatically (in cases in which the subsets have not been stipulated by the user explicitly); depiction of all incorrectly and correctly classified images (ie, true-positive, false-negative, false-positive, and true-negative cases); a robust solution to allow systematic external validation; and the addition of granular tools to randomise data splitting for training and validation, while maintaining grouping of images on a patient or visit basis.

The availability of automated deep learning might be a cornerstone for the democratisation of sophisticated algorithmic modelling in health care. It allows the derivation of classification models without requiring a deep understanding of the mathematical, statistical, and programming principles. However, the translation of this technological success to meaningful clinical effect requires concerted efforts and a careful stepwise approach to avoid biasing the results. Deep learning experts and clinicians will need to collaborate in ensuring the safe implementation of artificial intelligence. The sharp contrast of the model's discriminative performance in internal versus external validation might foretell the ultimate use case for automated deep learning software once the technology matures. As researchers and clinicians have excellent access to images and patient data within their own institutions, they might be able to design automated machine learning models for internal research, triage, and customised care delivery. Automating these processes might avert the need for costly external prospective validation across imaging devices, patient populations, and clinician practice patterns. By contrast, large-scale, standardised, deep learning algorithms will necessitate worldwide, multivariable validations of expertly tuned models. Thus, there is considerable value to these small data approaches customised to a specific geographical patient population that a given clinic might encounter. This might be where automated deep learning

finds its niche in the medical field. Importantly, this could make models susceptible to selection bias, overfitting, and a number of other issues from imprecise model training or patient selection. Novices should adhere to ethical principles when designing these models to avoid discrimination and causing harm.[42,43] Therefore, regulatory guidelines are needed for both medical deep learning and clinical implementation of these models before they might be used in clinical practice. In summary, although our approach seems rational in this early evaluation, the results of this study cannot yet be extrapolated into clinical practice.

Contributors

All authors contributed to the conception and design of the study. LF and SW trained the automated deep learning models and collected the data, which were analysed by LF, SKW, DJF, LMB, and PAK. LF and SKW drafted the Article, which was revised with substantial input from all authors. All authors have approved the final version of the Article.

Declaration of interests

JRL and TB are employees of DeepMind Technologies, a subsidiary of Alphabet. RC is an intern at DeepMind. EK is a contractor employee of Google Health, a subsidiary of Alphabet. PAK is an external consultant for DeepMind. Alphabet is the parent company of Google, which developed Google Auto ML Vision. All other authors declare no competing interests.

Acknowledgments

PAK is supported by a National Institute for Health Research (NIHR) Clinician Scientist Award (NIHR-CS--2014–14–023). EK is supported by a Moorfields Eye Charity Springboard Grant. 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. The Messidor images are kindly provided by the Messidor program partners

For more on the **Messidor Program** see http://www.adcis.net/en/third-party/messidor/

References

1 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.
2 Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *Dig Med* 2018; **1:** 5.
3 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature* 2015; **518:** 529–33.
4 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. *arXiv* 2014; published online Nov 17. https://arxiv.org/abs/1411.4555 (preprint).
5 Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process Mag* 2012; **29:** 82–97.
6 Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. *J Mach Learning Res* 2011; **12:** 2493–537.
7 Hadsell R, Erkan A, Sermanet P, Scoffier M, Muller U, LeCun Y. Deep belief net learning in a long-range vision system for autonomous off-road driving. 2008 IEEE/RSJ International Conference on Intelligent Robot Systems, IROS; Nice, France; Sept 22–26, 2018 (4651217).
8 Hadsell R, Sermanet P, Ben J, et al. Learning long-range vision for autonomous off-road driving. *J Field Rob* 2009; **26:** 120–44.
9 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems* 2012; **1:** 1097–105.
10 Scientific American. World changing ideas of 2015. *Scientific American* Dec 2,2015. https://www.scientificamerican.com/article/worldchanging-ideas-2015/ (accessed Aug 18).
11 Jouppi N. Google supercharges machine learning tasks with TPU custom chip. *Google Cloud*, May 18, 2016. https://cloud.google.com/blog/products/gcp/google-supercharges-machine-learning-tasks-with-custom-chip (accessed Feb 20, 2019).
12 Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 2019; **25:** 65–69.
13 ElementAI. 2019 Global AI talent report. *ElementAI*, April 2, 2019. https://www.elementai.com/news/2019/2019-global-ai-talent-report (accessed July 22, 2019).
14 MSV J. Why AutoML is set to become the future of artificial intelligence. *Forbes*, April 15, 2018. https://www.forbes.com/sites/janakirammsv/2018/04/15/why-automl-is-set-to-become-the-future-of-artificial-intelligence/#337d90ae780a (accessed Feb 26, 2019).
15 AI Multiple. Auto machine learning software/tools in 2019: in-depth guide. *AI Multiple*, Jan 1, 2019. https://blog.aimultiple.com/auto-ml-software/ (accessed Feb 26, 2019).
16 Zoph B, Vasudevan V, Shlens J, Le Q. AutoML for large scale image classification and object detection. *Google AI Blog*, Nov 2, 2017. https://ai.googleblog.com/2017/11/automl-for-large-scale-image.html (accessed July 19, 2018).
17 Thomas R. fast.ai, making neural nets uncool again. http://www.fast.ai/2018/07/23/auto-ml-3/2018 (accessed July 24, 2018).
18 Zoph B, Le QV. Neural architecture search with reinforcement learning. *arXiv* 2016; published online Nov 5. https://arxiv.org/pdf/1611.01578.pdf (preprint).
19 Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed image database: the Messidor database. *Image Anal Stereol* 2014; **33:** 231–34.
20 Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018; **172:** 1122–31.
21 Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA; July 21–26, 2017 (abstr 1304). DOI:10.1109/CVPR.2017.369.
22 Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 2018; **5:** 180161.
23 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350:** g7594.
24 HIPAA Journal. What covered entities should know about cloud computing and HIPAA compliance. https://www.hipaajournal.com/cloud-computing-hipaa-compliance/. *HIPAA J* Feb 19, 2018.
25 Elsken T, Metzen JH, Hutter F. Neural architecture search: a survey. *arXiv* 2018; published online Aug 16. https://arxiv.org/abs/1808.05377 (preprint).
26 Le Q, Zoph B. Using Machine Learning to Explore Neural Network Architecture. *Google AI Blog*, May 17, 2017. https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html (accessed July 22, 2019).
27 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 2012 https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks (July 22, 2019).
28 Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification. *arXiv* 2018; published online Jan 30. https://arxiv.org/abs/1801.09927 (preprint).
29 Li X, Pang T, Xiong B, Liu W, Liang P, Wang T. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics; Shanghai, China; Oct 14–16, 2017 (abstract). DOI: 10.1109/CISP-BMEI.2017.8301998.
30 Codella NC, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging; Washington DC, USA; April 4–7, 2018: 168–72.

31 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *JIMR* 2016; **18:** e323.

32 Kohn MA, Carpenter CR, Newman TB. Understanding the direction of bias in studies of diagnostic test accuracy. *Acad Emerge Med* 2013; **20:** 1194–206.

33 Oakden-Rayner L. Explain yourself, machine. Producing simple text descriptions for AI interpretability. https://lukeoakdenrayner.wordpress.com/2018/06/05/explain-yourself-machine-producing-simple-text-descriptions-for-ai-interpretability/2018 (accessed July 22, 2019).

34 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.

35 Nicholson Price N. Regulating Black-Box Medicine. *Mich L Rev* 2017; 116.

36 Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *Digit Med* 2019; **2:** 2.

37 Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *ArXiv* 2019; published online May 13. https://arxiv.org/abs/1905.05134 (preprint).

38 Lipton ZC. The Mythos of model interpretability. *ArXiv* 2016; published online June 10. https://arxiv.org/abs/1606.03490 (preprint).

39 Oakden-Rayner L. Exploring the ChestXray14 dataset: problems. https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/2017 (accessed July 22, 2019).

40 Open source research data repository software. https://dataverse.org/about (accessed Aug 15, 2019).

41 Li C, Wang X, Liu W, Latecki LJ. DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. *Med Image Anal* 2018; **45:** 121–33.

42 Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005; **5:** 142–49.

43 Char DS, Shah NH, Magnus D. Implementing machine learning in health care–addressing ethical challenges. *N Engl J Med* 2018; **378:** 981–83.