


Artificial intelligence in healthcare

Kun-Hsing Yu¹ , Andrew L. Beam¹ and Isaac S. Kohane^{1,2*}

Artificial intelligence (AI) is gradually changing medical practice. With recent progress in digitized data acquisition, machine learning and computing infrastructure, AI applications are expanding into areas that were previously thought to be only the province of human experts. In this Review Article, we outline recent breakthroughs in AI technologies and their biomedical applications, identify the challenges for further progress in medical AI systems, and summarize the economic, legal and social implications of AI in healthcare.

Artificial intelligence is gradually changing the landscape of healthcare and biomedical research. In the Aravind Eye Care System in India, ophthalmologists and computer scientists are working together to test and deploy an automated image classification system to screen millions of retinal photographs of diabetic patients¹. Diabetic retinopathy (DR) affects more than 90 million people worldwide and is a leading cause of blindness in adults². Fundus photography is an effective method to monitor the extent of DR and to identify patients who will benefit from early treatments³. However, in many parts of the world, there are too few ophthalmologists to read the fundus photographs and to follow up with each diabetic patient⁴. A team of researchers at Google Inc. and collaborating institutions showed that an AI system trained on thousands of images can achieve physician-level sensitivity and specificity in diagnosing referable DR⁵, as well as in identifying previously unrecognized associations between image patterns in the fundus photograph and cardiovascular risk factors⁶. The technology giant is now integrating this AI technology into clinical practice in a chain of eye hospitals in India¹, and a related technology developed by University of Iowa was approved by the US Food and Drug Administration (FDA) for detecting moderate-to-severe DR⁷.

AI has recently re-emerged into the scientific and public consciousness, as new breakthroughs and technologies are announced from technology companies and scientists at a breakneck pace. Stripped of its science-fictional ornamentation and aspirations, AI is, at its core, a branch of computer science that attempts to both understand and build intelligent entities, often instantiated as software programs⁸. AI has a long history that traces its roots back to a conference at Dartmouth in 1956, where the term was used for the first time⁸. The successful development of image classifiers since 2012 has contributed to the recent resurgence of AI⁹. Although much progress has been made in the past decades, AI has suffered from an inconsistent and evolving definition as to what exactly constitutes ‘real AI’. It is a well-recognized property of AI research that success in reaching a specific performance goal soon disqualifies that performance as constituting AI, which makes tracking progress difficult. As an illustration, automated route planners were touted as examples of advanced AI in the 1970s, yet are now so ubiquitous that most people would be surprised to hear them described as AI¹⁰. Consequently, the successes of AI from the 1970s through the 1990s that were once heralded as breakthroughs in medicine, such as the automated interpretation of electrocardiograms (ECGs)¹¹, are now regarded as useful but are hardly considered to be examples of true AI.

Recently, applications of medical-image diagnostic systems have expanded the frontiers of AI into areas that were previously the domain of human experts. This frontier continues to expand into other areas of medicine, such as clinical practice¹², translational medical research^{13–16} and basic biomedical research¹⁷ (Table 1). In this Review Article, we focus on the AI applications that could augment or change clinical practice, offer a historical perspective on AI in medicine to contextualize recent advancements, summarize the successful application areas, identify the potential societal impact arising from the development and deployment of biomedical AI systems, and suggest future research directions. For a glossary of key terms, see Box 1.

A historical overview of AI in medicine

Medicine was identified early as one of the most promising application areas for AI. Since the mid-twentieth century, researchers have proposed and developed many clinical decision support systems^{18,19}. Rule-based approaches saw many successes in the 1970s^{20,21}, and have been shown to interpret ECGs¹¹, diagnose diseases²², choose appropriate treatments²³, provide interpretations of clinical reasoning²⁴ and assist physicians in generating diagnostic hypotheses in complex patient cases²⁵. However, rule-based systems are costly to build and can be brittle, as they require explicit expressions of decision rules and require human-authored updates, just like any textbook. In addition, it is difficult to encode higher-order interactions among different pieces of knowledge authored by different experts, and the performance of the systems is limited by the comprehensiveness of prior medical knowledge²⁶. Moreover, it was difficult to implement a system that integrates deterministic and probabilistic reasoning to narrow down relevant clinical context, prioritize diagnostic hypotheses, and recommend therapy²⁷.

Unlike the first generation of AI systems, which relied on the curation of medical knowledge by experts and on the formulation of robust decision rules, recent AI research has leveraged machine-learning methods, which can account for complex interactions²⁸, to identify patterns from the data. According to the types of task that they intend to solve²⁸, basic machine-learning algorithms fall roughly into two categories: supervised and unsupervised. Supervised machine-learning methods work by collecting a large number of ‘training’ cases, which contain inputs (such as fundus photographs) and the desired output labels (such as the presence or absence of DR). By analysing the patterns in all of the labelled input–output pairs, the algorithm learns to produce the correct output for a given input on new cases²⁹. Supervised machine-learning algorithms are designed to

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ²Boston Children’s Hospital, Boston, MA, USA.

*e-mail: Isaac_Kohane@hms.harvard.edu

Box 1 | A glossary of key terms

Artificial intelligence. A branch of computer science that attempts to both understand and build intelligent entities, often instantiated as software programs⁸.

Deep learning. A subfield of the larger discipline of machine learning. Deep learning employs artificial neural networks with many layers to identify patterns in data³².

Dimensionality reduction. The process of reducing the number of variables in the data. The raw data can contain a great number of redundant and non-informative variables. For example, nearby pixels of an image may be of similar or identical colour. By reducing the number of variables, statistical analyses can be conducted more effectively and sophisticated machine-learning models developed without running out of computer memory.

Feedforward neural network. A type of artificial neural network in which neural layers only connect to the next layer and do not form cycles.

Floating-point operations per second (FLOPS). A measure of computation performance. In computers, numbers with decimal points are represented as 'floating-point numbers'. FLOPS measures the number of floating-point arithmetic operations that a computer system can complete per second; the larger the value, the more powerful the computer system is.

Graphical processing unit (GPU). Computer hardware initially designed to handle computational tasks related to images and to produce outputs to display devices. Since modern GPUs have many computation cores and can provide fast parallel computation, they have become the workhorses for training artificial neural networks.

Machine learning. A field of computer science that uses algorithms to identify patterns in data¹⁵¹.

Perceptron. A binary classifier developed in the 1950s. It classifies samples by using the following function: given an input vector \mathbf{x} , the output is 1 if $\mathbf{w} \cdot \mathbf{x} + \mathbf{b} > 0$, where \mathbf{w} and \mathbf{b} are two vectors of parameters; otherwise, the output is 0.

Supervised machine learning. A type of machine-learning task that aims at predicting the desired output (such as the presence or absence of DR) on the basis of the input data (such as fundus photographs). Supervised machine-learning methods work by identifying the input–output correlation in the 'training' phase and by using the identified correlation to predict the correct output of the new cases.

Unsupervised machine learning. A type of machine-learning task that aims at inferring underlying patterns in unlabelled data. For example, it can find sub-clusters of the original data, identify outliers in the data, or produce low-dimensional representations of the data.

Table 1 | A non-exhaustive list of current and potential AI applications in medicine

Basic biomedical research	Translational research	Clinical practice
Automated experiments	Biomarker discovery	Disease diagnosis
Automated data collection	Drug–target prioritization	Interpretation of patient genomes
Gene function annotation	Drug discovery	Treatment selection
Prediction of transcription factor binding sites	Drug repurposing	Automated surgery
Simulation of molecular dynamics	Prediction of chemical toxicity	Patient monitoring
Literature mining	Genetic variant annotation	Patient risk stratification for primary prevention

identify the optimal parameters in the models to minimize the deviations between their predictions for the training cases and the observed outcomes in these cases, with the hope that the identified associations are generalizable to cases not included in the training dataset. The generalizability of the model can be estimated by the test set (Fig. 1a). Classification, regression and characterization of the similarity among instances of similar outcome labels are among the most widely used tasks of supervised machine-learning models. Unsupervised learning infers the underlying patterns in unlabelled data to find sub-clusters of the original data, to identify outliers in the data, or to produce low-dimensional representations of the data (Fig. 1b). Note that the identification of low-dimensional representations for labelled instances could be more effectively achieved in a supervised fashion. Machine-learning methods enable the development of AI applications that facilitate the discovery of previously unrecognized patterns in the data without the need to specify decision rules for each specific task

or to account for complex interactions among input features. Machine learning has thus become the preferred framework for building AI utilities^{28,30}.

The recent renaissance in AI has to a large extent been driven by the successful application of deep learning — which involves training an artificial neural network with many layers (that is, a 'deep' neural network) on huge datasets — to large sources of labelled data. Since 2012, deep learning has shown substantial improvements in image classification tasks⁹. Figure 2 shows a sampling of deep neural network architectures (there are also automated approaches to network architecture design³¹). The basic architecture of deep neural networks consists of an input layer and an output layer, and a number of hidden layers in between. Perceptron and feedforward neural networks are among the simplest designs (Fig. 2a,b). Autoencoders are used for dimension reduction, whereas sparse autoencoders can generate additional useful features (Fig. 2c,d). Recurrent neural networks are useful for handling time-series data (Fig. 2e). Deep residual neural networks have improved the traditional deep feed-forward neural network by allowing skip connections, avoiding the saturation of model performance^{32,33} (Fig. 2f,g).

Many modern neural networks have more than 100 layers. Neural networks with many layers can model complex relations between the input and output but may require more data, computation time or advanced architecture designs so as to achieve optimal performance. Many types of layers, mathematical operations for the neurons and regularization methods have been designed (Table 2). For example, convolutional layers are useful for extracting spatial or temporal relations, whereas recurrent layers use circular connections to model temporal events. Also, various initialization and activation functions can enhance model performance. The combination of these components enables neural networks to handle various input data with and without spatial or temporal dependence.

Modern neural networks can have tens of millions to hundreds of millions of parameters and take huge amounts of computational resources to train. Fortunately, recent advances in computer-processor design provide the computational power required for deep

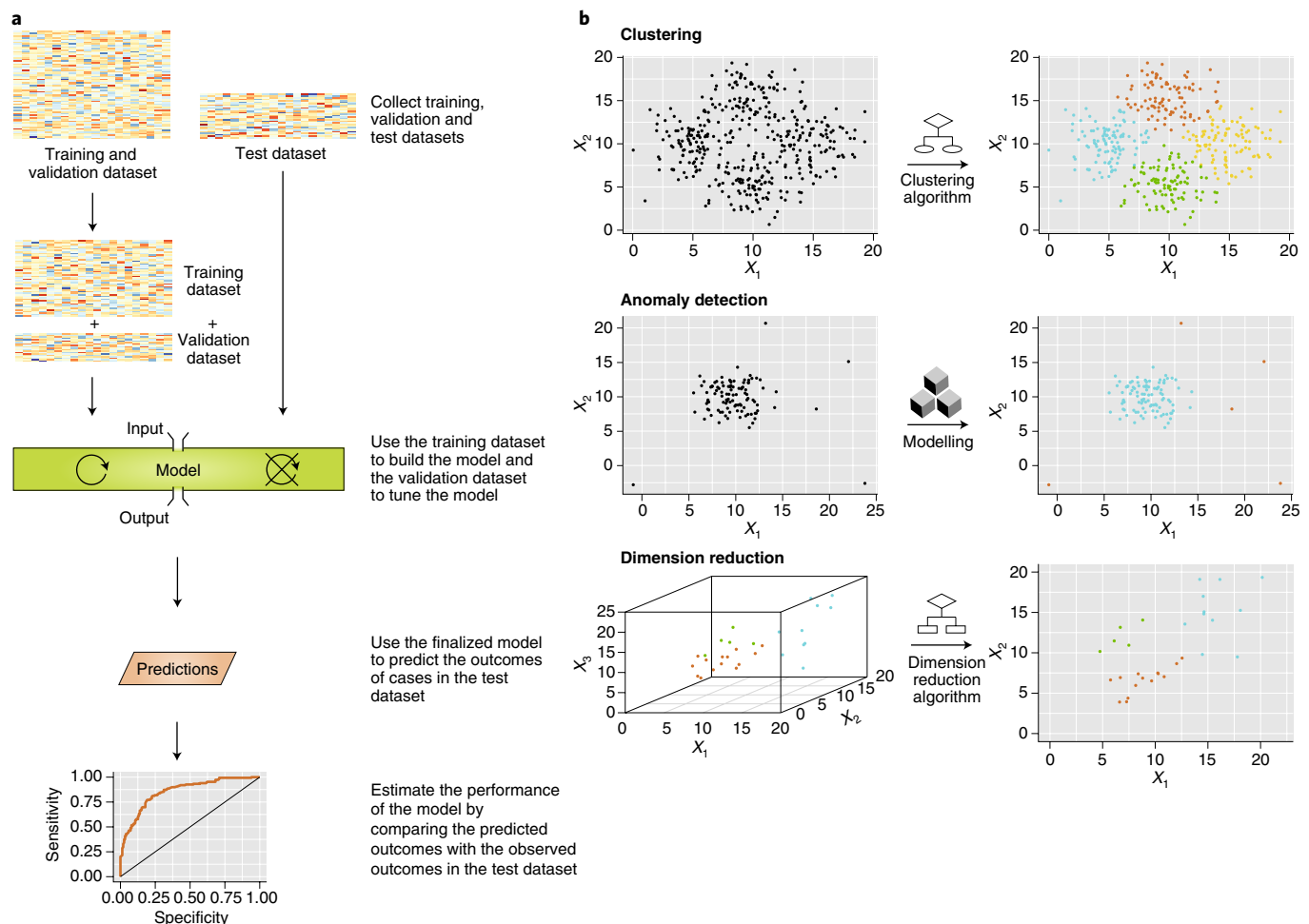


Fig. 1 | Supervised and unsupervised machine learning. **a**, General workflow of supervised machine-learning approaches. First, training and test datasets are collected. Next, part of the training set is used to build the prediction model, and the other part to tune and validate the model (circular arrow). After the machine-learning model is finalized (crossed-out circular arrow), the established model is used to generate predictions on the test dataset and the model's performance is estimated by comparing the predicted outcomes with the observed outcomes for the test dataset. **b**, Unsupervised machine learning includes clustering, anomaly detection and dimensionality reduction. Clustering algorithms group data points with similar measurements into clusters. Anomaly detection identifies outliers in the dataset. Dimensionality reduction reduces the number of random variables used to describe the data; for example, by representing an image with thousands of parameters as a smaller vector of summary features. The resulting summary vector preserves the important information in the raw data; for example, summary vectors from similar images will bear more resemblance than those obtained from irrelevant images.

learning. As an illustration, current graphical processing units (GPUs) can perform more than 7 trillion floating-point operations per second. Any such GPU, which would have ranked among the fastest 100 supercomputers in the world in 2006 (ref. ³⁴), is capable of processing hundreds of millions of medical images per day at relatively low cost³⁵. By harnessing computation power, large datasets and 'convolutional' neural networks (CNNs), deep learning has transformed not just medical-image analysis, but the whole field of computer vision. CNNs use a special type of layer (that is, a convolutional layer) to summarize and transform clusters of pixels in images to extract high-level features. Before the advent of CNNs, features from the images had to be defined and extracted³⁶, and the performance of the machine-learning models depended on the quality of features. In CNNs, a key improvement is that they can operate on the raw image and learn useful features from the training sets, thus simplifying the training process and facilitating the identification of image patterns. CNNs have proven crucial to the success of deep learning in image analysis and are also responsible for the subsequent revolution in medical imaging. There is an on-going community effort to compile neural network applications in biology and

medicine³⁷. Table 3 summarizes the performance, reproducibility, comprehensibility, dependency on prior knowledge, development, running costs, update costs and around-the-clock availability for human practitioners and for different types of AI approaches.

However, deep-learning algorithms are extremely 'data hungry' for labelled cases. Only recently, large sources of medical data that can be fed into these algorithms have become widely available, owing to the establishment of many large-scale studies (in particular, the Cancer Genome Atlas³⁸ and the UK Biobank³⁹), data-collection platforms (such as the Broad Bioimage Benchmark Collection⁴⁰ and the Image Data Resources⁴¹) and the Health Information Technology for Economic and Clinical Health (HITECH) Act, signed in 2009. The HITECH Act provided financial incentives for the adoption of electronic health records (EHRs). In a national survey in 2008, only 13% of physicians reported having a basic EHR system⁴²; by the end of 2012, 72% of physicians had adopted some type of EHR system and 40% of physicians reported having capabilities that met the criteria for a basic system⁴³. The increasing adoption of EHR systems has not only expedited the collection of large-scale clinical data, but also permits smoother integration of AI systems into

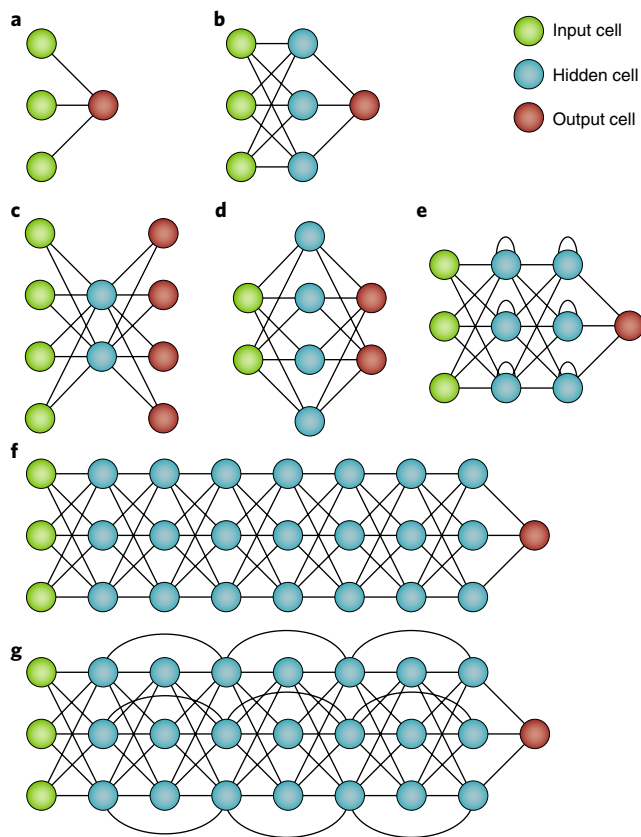


Fig. 2 | The general architecture of artificial neural networks.

a, Perceptron. Given an input vector \mathbf{x} from the input cells, the output is 1 if $\mathbf{w} \cdot \mathbf{x} + \mathbf{b} > 0$, where \mathbf{w} and \mathbf{b} are two vectors of parameters; otherwise, the output is 0. **b**, A two-layer feedforward neural network. The input layer takes in the data and forwards the data to the cells of the hidden layer. Each cell in the hidden layer serves as a function that integrates its inputs and transmits the output of the function into the cells in the next layer. There are no cyclic computations. **c**, Autoencoder. Autoencoding is an unsupervised technique that uses neural networks to learn a representation of the input data. An autoencoder is often used for dimensionality reduction. **d**, Sparse autoencoder. An autoencoder with a large number of cells in the hidden layers and a sparsity constraint, which forces most hidden cells to be inactive given the input. This strategy is useful for deriving features for classification tasks. **e**, Recurrent neural network. A type of neural network that allows connections between the nodes to form directed cycles. It is useful for handling time series. **f**, Deep feedforward neural network. There can be many hidden layers between the input layer and the output layer to model the complex relations between the inputs and the outputs. The last hidden layer connects to the output layer, which generates model outputs. **g**, Deep residual neural network. In this architecture, skip connections are allowed, which in deep neural networks helps avoid performance saturation or degradation.

the clinical workflow (Fig. 3). Conventionally, medical practitioners collect medical information from the patients, make clinical judgments and record their diagnoses and treatment plans in the health records (Fig. 3a). Since the 1970s, decision support systems that collect medically relevant information and provide suggestions to clinicians have been developed (Fig. 3b). There are a number of ways to integrate decision support systems into the clinical workflow; for instance, decision support systems can actively gather information from patients and EHRs, present suggestions to clinicians and store the system outputs in EHRs (Fig. 3c). In many proposed fully automated clinical systems, the autonomous tools

collect information from the patients, make decisions and output the results into EHRs (Fig. 3d) — although such integration has so far been slight. Data from EHR systems provide detailed information about the patients, including clinical notes and laboratory values, enabling the application of natural-language-processing methods to extract codified vocabularies.

The recent confluence of large-scale annotated clinical data acquisition, advancement in machine-learning methods, open-source machine-learning packages, and affordable and rapidly growing computational power and cloud storage has fuelled the recent exponential growth in AI. This promises to change the landscape of medical practice in the near term. AI systems have specialist-level performance in many diagnostic tasks^{5,44}, can better predict patient prognosis than clinicians^{45,46}, and can assist in surgical interventions⁴⁷. As machine-learning models continue to advance, there is a growing sense that AI could revolutionize medical practice and redefine the roles of clinicians in the process¹².

Image-based diagnosis

Currently, automated medical-image diagnosis is arguably the most successful domain of medical AI applications. Many medical specialties, including radiology, ophthalmology, dermatology and pathology, rely on image-based diagnoses. In what follows, we summarize recent advances in the application of AI to each of these medical fields.

Radiology. Diagnostic radiologists use multiple medical-imaging modalities — the most widely used being X-ray radiography, computed tomography, magnetic resonance imaging (MRI) and positron-emission tomography — to detect and diagnose diseases. In each of these approaches, radiologists use a collection of images for disease screening and diagnosis, to identify the cause of illness and to monitor the patient trajectory during the course of a disease⁴⁸.

Radiological practice relies primarily on imaging for diagnosis and thus is very amenable to deep-learning techniques, as images often contain a large proportion of the information needed to arrive at the correct diagnosis. Most radiology departments maintain a database of historical images in a Picture Archiving and Communication System, which typically provides thousands of examples to train the neural networks. Computational approaches for radiology diagnoses have been proposed and implemented since the 1960s⁴⁹. With the help of modern machine-learning methods, many radiology applications of AI, such as the detection of lung nodules using computed tomography images⁵⁰, the diagnosis of pulmonary tuberculosis and common lung diseases with chest radiography^{51–54} and breast-mass identification using mammography scans^{55,56}, have reached expert-level diagnostic accuracies. These studies employed a technique known as transfer learning, where well-established deep neural networks trained on millions of natural, non-medical images are borrowed, and then the neural network connections are fine-tuned by using thousands of biomedical images. Such a strategy can reduce the number of training samples required to train a neural network with tens of millions of parameters, making it highly effective for medical-image classification where the number of images may only be in the range of thousands to tens of thousands. For researchers to visualize the neural network models, the relevance of each pixel to the output classes can be investigated. For example, saliency maps and gradient-weighted class activation maps visualize the importance of each image region in relation to their classification and are useful for identifying localized image features⁶ (Fig. 4a,b), activation maximization generates images that maximally activate a selected neuron (Fig. 4c), and an individual convolution filter can be visualized by producing synthetic input images that maximize the filter output (Fig. 4d–g). These approaches attempt to make neural network models more interpretable⁵¹.

Table 2 | An overview of the common components of deep neural networks

Common components	Types	Function
Layers	Densely connected layer	To operate on the inputs from the previous layer; too many densely connected layers can result in overfitting, which could be mitigated by randomly setting a fraction of inputs to 0 (also called dropout).
	Convolutional layer	To perform convolution over the input; useful for inputs with spatial or temporal relations.
	Pooling layer	To decrease the number of parameters in the neural network, reducing overfitting.
	Recurrent layer	To allow circular connections between the elements in the neural network; useful for modelling temporal events.
	Embedding layer	To map the input into a dense vector space.
	Normalization layer	To normalize the amount of activation from the previous layer.
	Noise layer	To add random noise to the input; useful for reducing overfitting.
Initialization functions	Deterministic	To initialize the values of the cells in the neural network layers to some constants.
	Random	To initialize the values of the cells in the neural network layers to some random numbers that follow a certain distribution.
Activation functions	Sigmoid, hyperbolic tangent (tanh), softmax, scaled exponential linear unit (SELU), rectified linear units (RELU) and others	To enhance network performance by adding nonlinear factors to the neural network.
Loss functions	Mean squared error, mean absolute error, cosine distance, categorical cross-entropy and others	To evaluate the performance of the neural network; the loss function is a part of the objective function.
Optimization algorithms	Stochastic gradient descent, root-mean-square propagation (RMSprop), AdaGrad, Adam and others	To determine the direction to fine-tune the weights in the neural network
Regularization methods	L1, L2, L1 + L2	To prevent large parameters by including the L1 norm (sum of the absolute value), the L2 norm (sum of squares), or the weighted mean of the L1 and L2 norms of parameters into the objective function.

Many clinical applications of AI are seeking regulatory approval. For instance, a deep-learning system for diagnosing cardiovascular diseases using cardiac MRI images was approved by the FDA in 2018 (refs ^{57,58}). With further validation studies and technology transfer efforts, we expect that more image-based computer-aided detection (CAD) and diagnostic systems will be put into clinical use in the near term.

Dermatology. Inspection plays an important role in diagnosing many types of skin lesion. For instance, typical skin melanoma has visual features that distinguish these lesions from benign moles⁵⁹. For the diagnosis of skin melanoma by inspection, dermatologists have developed rules of thumb, such as the widely known ABCDE rule. The rule is applied to diagnosing pigmented tumours, where criterion A refers to the geometrical asymmetry of the tumour, B to irregular borders, C to colour variegation, D to a diameter equal to or more than 6 mm, and E to the enlargement of the surface of the lesion or evolving lesion^{59,60}. With the exception of the E criterion, the other criteria can be assessed from a single photograph of the lesion.

For many years, researchers have attempted to develop automated diagnostic systems that classify photographs of benign and malignant lesions^{61,62}. Recently, convolutional neural networks

trained on 129,450 clinical images achieved dermatologist-level accuracy in diagnosing skin malignancy⁴⁴. The deep-learning algorithm outperformed the average dermatologist in a comparison of the algorithm predictions and the assessments by 21 dermatologists on a set of photographic and dermoscopic images. Although the training phase of the deep-learning model can be computationally expensive, the finalized diagnostic model can be deployed on mobile devices, potentially improving the accessibility of skin-lesion screening at the expert level globally⁴⁴.

Ophthalmology. Fundus photography is a non-invasive procedure that uses retinal cameras to capture images of the retina, optic disc and macula. It can detect and monitor diseases such as DR, glaucoma, neoplasms of the retina and age-related macular degeneration, and plays a vital role in identifying causes of preventable blindness⁶³. In particular, clinical guidelines from the American Diabetes Association recommend DR screening for diabetic patients with minimal or no retinopathy every year, and more frequent examination for patients with progressing DR⁶⁴. Traditionally, fundus photographs are examined and interpreted by ophthalmologists, which is difficult to scale to the millions of diabetic patients at risk of developing sight-threatening DR⁶⁵.

Table 3 | Comparisons between human evaluations and different types of AI approaches

Approaches	Model comprehensibility	Performance	Reproducibility	Dependency on prior knowledge	Development and training costs ^a	Running costs	Around-the-clock availability	Update costs
Human evaluation	High	Moderate or high	Moderate	High	High	High	Low	High
Rule-based algorithms	High	Moderate or high	High	High	Moderate or high	Low	High	High
Feature-based machine-learning methods	Moderate or high	Moderate or high	High	Moderate ^b	Moderate	Low	High	Moderate ^c
Deep artificial neural networks	Low or moderate	High	High	Low	Moderate	Low	High	Low

^aThe estimated cost of training professionals that carry out the clinical tasks (human evaluation) or of developing the automated system (rule-based, feature-based or deep-artificial-neural-network-based) that performs the tasks. ^bFor feature-based machine-learning methods, prior knowledge may facilitate the derivation of useful features from the raw data. ^cWhen the update requires encoding new features, the update cost of feature-based machine-learning methods includes feature engineering and model retraining.

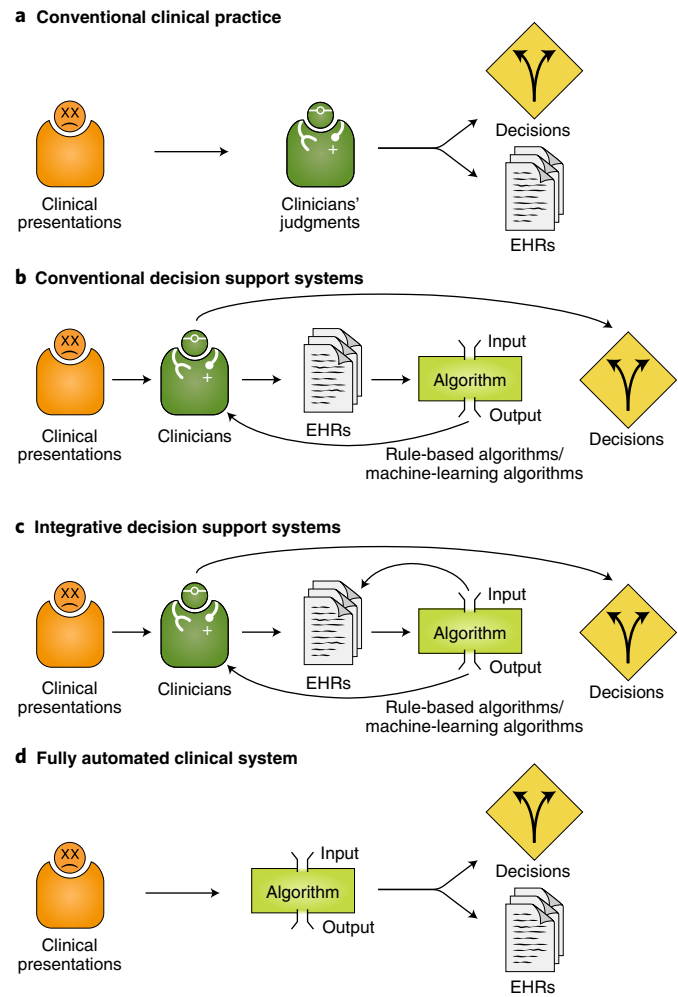


Fig. 3 | Models of information flow in conventional clinical practice, conventional decision support systems, integrative decision support systems and fully automated clinical systems. a. In conventional clinical practice, clinicians collect information from the patients, make clinical decisions using their own judgments, and record their findings in EHRs. **b.** Conventional decision support systems collect information from EHRs and provide suggestions using rule-based algorithms or machine-learning algorithms. Clinicians receive the suggestions and make the final decision. **c.** In integrative decision support systems, the systems can actively request clinically relevant information or gather the data from EHRs, show the results to clinicians, and write them into EHRs automatically. Clinicians still need to make the final decision. **d.** In many proposed fully automated clinical systems, the autonomous systems collect information from the patients, make decisions and output the results to EHRs.

A team of computer scientists and clinicians recently trained convolutional neural network models to identify referable DR and diabetic macular oedema using 128,175 retinal images. In this

retrospective study, the machine-learning model achieved areas under the receiver operating characteristic curve greater than 0.990 in two independent test datasets, which was comparable to the per-

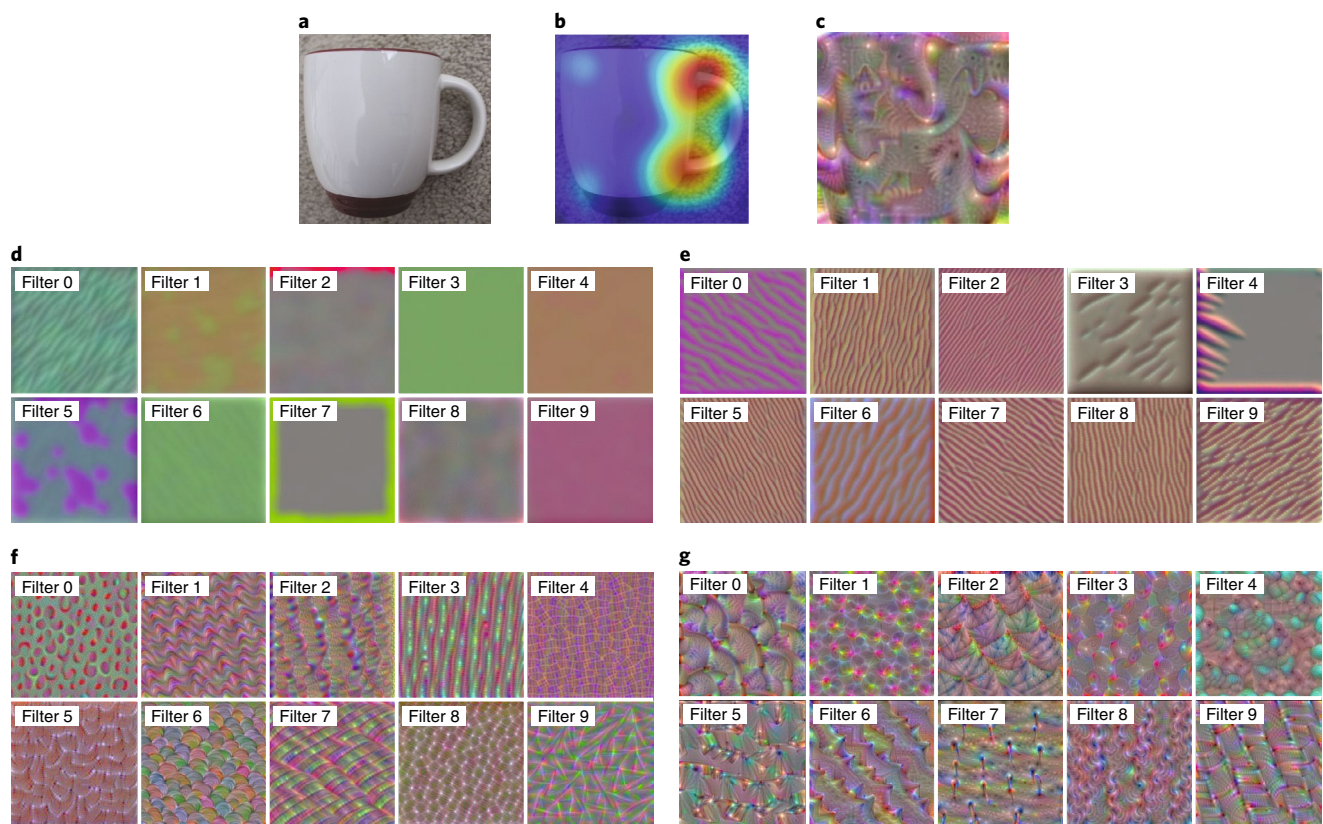


Fig. 4 | Example of the interpretation of convolutional neural networks. **a**, The original image of a coffee mug. **b**, Attention visualization. A gradient-weighted class activation map (Grad-CAM) of VGG16¹⁵⁰ (a neural network with 16 weighted layers that performed well in the Image Net Large Scale Visual Recognition Challenge in 2014) that visualizes the importance of input pixels on the last convolution layer, which summarizes spatial information contained in the image and passes it to the layers that generate the classification. **c**, Activation maximization. An image that maximizes the score for 'coffee mug'. The same technique can produce the images that maximize the activation of any selected neuron in the convolutional neural networks. **d–g**, Visualization of the convolution layers in VGG16. Any convolution filter could be visualized by generating synthetic input images that maximize the filter output. Panel **d** shows the visualization of 10 filters in the second convolution layer of VGG16. It reveals low-level image patterns, such as colour patches and lines; panel **e** shows 10 filters in the fourth convolution layer of the same neural network, revealing various linear patterns; panel **f** shows the patterns that maximize the output of 10 filters in the eighth convolution layer, revealing shape patterns; and panel **g** shows the patterns for 10 filters in the eleventh convolution layer, revealing complex patterns of shapes and objects. All visualizations were generated by the keras-vis package in Python.

formance of ophthalmologists⁵. They also demonstrated that deep learning can extract previously unrecognized associations between retinal image patterns and age, gender, systolic blood pressure and smoking status, as well as major adverse cardiac events⁶, illustrating the utility of machine learning in eliciting new knowledge from the raw data. Another team of researchers showed that the performance of a convolutional neural network exceeded their pre-specified sensitivity (85%) and specificity (82.5%); the system was authorized by the FDA for use by healthcare providers to detect diabetic macular oedema and moderate-to-severe DR (Early Treatment Diabetic Retinopathy Study Severity Scale, level 35 or higher)^{7,66}.

Pathology. Histopathological assessment is the gold standard for the diagnosis of many cancer types^{15,67,68}. The procedure involves processing a biopsy or surgical specimen into tissue slides and staining the slides with pigments, and then expert pathologists interpreting the slides under a microscope on the basis of visual evaluation⁶⁹. However, discrepancies among pathologists have been documented^{69,70}, and the process is not easily scalable. Moreover, some quantitative histopathology image features that are barely noticeable by the human eye can predict the survival outcomes of cancer patients^{45,46,71}, indicating the existence of rich yet previously underutilized information in the pathology slides.

With the advent of deep convolutional neural networks, AI can be useful in the detection of prostate cancer from biopsy specimens⁷², the identification of breast cancer metastasis in lymph nodes^{72,73} and the detection of mitosis in breast cancer⁷⁴. For example, machine learning coupled with a system for imaging live-cell biomarkers can facilitate the risk stratification of prostate and breast cancer patients⁷⁵. As it is estimated that there will be a net deficit of more than 5,700 full-time equivalent pathologists by 2030 (ref. ⁷⁶), an automated system could mitigate this deficit, provide a fast and objective evaluation of the histopathology slides, and improve the quality of care for cancer patients.

Overall, successful applications in radiology, dermatology, ophthalmology and pathology leverage the availability of large labelled data, computation power and deep-learning methods to achieve expert-level diagnostic accuracy⁷⁷. Translating these research results into clinical applications is not straightforward, yet it has the potential to change current medical practice significantly.

Genome interpretation

High-throughput sequencing methods generate terabytes of raw data for genomic studies. Accurate clinical interpretation of these data is key to understanding differences across individuals and paves the way for precision medicine²⁹. However, knowledge about

the human genome is constantly evolving, and it is difficult to systematically compare a patient's genome with known cases and controls solely using human curation. Deep neural networks can annotate pathogenic genetic variants⁷⁸ and identify the functions of non-coding DNA⁷⁹ better than conventional methods, such as logistic regression and support vector machines⁷⁸. Interestingly, one neural-network-based approach that transforms the genomic-variant-calling task into an image classification task achieved better performance than the widely used Genome Analysis Toolkit^{80,81}. Such computational approaches can also be useful for diagnosing complex diseases with genetic components, such as cancer⁸².

Machine learning for biomarker discovery

Biomarker discovery relies on identifying previously unrecognized correlations between thousands of measurements and phenotypes. Omics technologies have enabled the high-throughput measurement of thousands of genes and proteins, and of millions of genomic and epigenomic aberrations²⁹. However, it is almost impossible for researchers to manually analyse and interpret the vast amount of data gathered from omics approaches. Machine-learning methods can identify the molecular patterns associated with disease status and disease subtypes, account for the high-level interactions among the measurements, and derive omics signatures to predict disease phenotypes⁸³. Gene expression, protein abundance levels and DNA methylation profiles can predict the status of a number of diseases, including cancers^{15,84,85}, infectious diseases⁸⁶ and the risk of Down's syndrome⁸⁷. Many of the biomarker panels derived from machine learning have outperformed those selected by experts or by conventional statistical methods. A number of them have been approved by the FDA and can be routinely used to guide treatment selection^{84,85,88}. Non-omics biomarkers, such as neural excitation signals, could facilitate the development of an interface for prosthetic control⁸⁹. The successful deployment of data-driven biomarkers has implications for both clinical management and trial design. Nevertheless, the reproducibility of a few biomarkers has been challenged, and it is methodologically fraught to identify robust biomarkers when the number of measurements or parameters is far larger than the number of samples⁹⁰. With the recent establishment of nationwide biobanks, standardized high-throughput profiling methods and advanced machine-learning methods, more robust and accurate biomarkers are expected to arise.

Clinical outcome prediction and patient monitoring

In addition to identifying biomarkers related to clinical phenotypes, the use of EHRs to predict clinical outcomes shows great promise. Bayesian networks can predict mortality, readmission and length of hospital stay by using EHRs from the emergency department⁹¹. Data from health insurance claims can be used to predict mortality in elderly patients⁹², patient attributes in the medical notes can be employed to classify cancer patients with different responses to chemotherapy⁹³, and clinical predictors for the prognosis of patients receiving thoracic organ transplantation can be identified⁹⁴. These studies characterized a number of robust clinical predictors for patient outcomes, and could be used to help guide patients and their physicians in choosing an individualized treatment strategy.

Patient monitoring is crucial in intensive care units, operating rooms, emergency rooms and cardiac wards where timeliness in clinical decision-making can be measured in seconds. Routine monitoring devices in these high-acuity contexts generate a large amount of data and thus represent a great opportunity for AI-assisted alert systems. By using vital signs and the Modified Early Warning Score, a prediction model for cardiac arrest was established⁹⁵. Demographics, laboratory results and vital signs can also be used to predict cardiac arrest, transfer into the intensive care unit, or death⁹⁶. In addition, an interpretable machine-learning model can assist anaesthesiologists in predicting hypoxaemia events

during surgery⁹⁷. This suggests that, with deep-learning algorithms, raw patient-monitoring data could be better used to avoid information overload and alert overload while enabling more accurate clinical prediction and timely decision-making.

Inferring health status through wearable devices

Modern wearable devices record a plethora of biomedical signals, including heart rate, voice, tremor and limb movement. These biological signals could be useful for detecting diseases and inferring health conditions. By way of illustration, signs of infectious disease and inflammatory responses can be detected early by using heart rate and skin temperature data recorded by wearables⁹⁸. The inclusion of photoplethysmography sensors in wearables enables the monitoring of cardiovascular diseases, pulmonary diseases, anaemia and sleep apnea⁹⁹. Wearable sensors could also detect and quantify symptoms of patients with Parkinson's disease, such as tremor and impaired hand movement, gait, posture and speech patterns¹⁰⁰.

Although personal tracking devices present an opportunity to guide behavioural changes¹⁰¹, the accuracy of the data collected through these devices can be variable^{102,103}. In addition, one-third of all US consumers who have owned wearable devices stopped using them within six months of receiving them, foreshadowing the utility of the devices in fostering long-term behavioural change^{104,105}. More research is needed to identify the ways to maximize the effectiveness of wearables in the promotion and maintenance of health¹⁰⁶.

Autonomous robotic surgery

Robotic systems controlled by AI are routinely used in assembly lines in industry and in many biomedical laboratories¹⁰⁷. However, the development and adoption of autonomous robots in medical interventions has been considerably slower. For many decades, robotic surgery has been synonymous with robotically assisted surgery — systems that facilitate surgical procedures and enable motions smoother than those achievable by human hands, but that still require a surgeon for movement control¹⁰⁸. For instance, in the FDA-approved da Vinci surgical system for minimally invasive operations, surgeons operate the robot from a console¹⁰⁹. Such systems are designed to translate the surgeon's hand movements into the movements of instruments inside the patient¹⁰⁹ and are therefore not autonomous.

As suturing is one of the most common procedures during surgery, autonomous knot-tying robots have been developed¹⁰⁸. Recently, a supervised autonomous robotic system for suturing an intestinal anastomosis showed, in a laboratory setting, better in vivo suturing quality than surgeons⁴⁷. This system employed an autonomous suturing algorithm and a plenoptic three-dimensional near-infrared fluorescent imaging system to perform in vivo open surgery on pigs. The autonomous system had better consistency of suturing, higher quality of anastomosis (measured by the pressure at which the anastomosis leaked), and a fewer number of mistakes that required removing the needle from the tissue than hand-sewn suturing, laparoscopy and robot-assisted surgery with the da Vinci surgical system⁴⁷. Similarly, a number of autonomous robots for cochleostomy have also been proposed¹¹⁰.

With the continuing development of pre-programmed, image-guided and teleoperated surgical robots, more robot-assisted or automated intervention methods are expected to be incorporated in surgical practice¹¹¹.

Taken together, AI is poised to revolutionize many aspects of current clinical practice in the foreseeable future. AI systems can enhance clinical decision-making, facilitate disease diagnosis, identify previously unrecognized imaging or genomic patterns associated with patient phenotypes, and assist in surgical interventions for various human diseases. AI applications also have the potential to bring clinical expertise to remote regions where specialists are

Table 4 | Clinical integration of medical AI at different developmental stages

	Areas where AI performance is more reliable than that of a human expert	Areas where AI performance is at the expert level	Areas where AI performance is reasonable	Areas where AI performance is not yet good enough	Areas where the nature of the clinician–patient interaction is fundamentally different from that of the AI–patient interaction
Examples	Serum analyser ^{144,145} , alert systems (such as drug–drug interaction checkers ^{146,147})	Assessment of certain radiology images (for example, annotation of cardiovascular MRI images ^{57,58} or evaluation of X-ray images for distal radius fracture ¹⁴⁸); dermoscopic melanoma diagnosis ¹⁴⁹ ; fundus photograph evaluation for DR ^{5,7}	ECG reading ⁷¹	Surgery; full interaction with patients	Emotional support and rapport
Potential clinical integrations	Delegate to AI	AI does the majority of the task, clinicians confirm the diagnosis	AI does a portion of the task (such as screening), clinicians confirm the diagnosis	Clinicians lead the clinical evaluations and intervention, AI assists in routine sub-tasks	Clinicians continue to provide the service

scarce or not available^{112,113}. Table 4 summarizes potential clinical integrations of medical AI systems, stratified by their developmental stages.

Technical challenges in AI developments

Although AI promises to revolutionize medical practice, many technical challenges lie ahead. As machine-learning-based methods rely heavily on the availability of large amounts of high-quality training data, care must be taken to compile data that is representative of the target patient population. For example, data from different healthcare environments can contain various types of bias and noise, which may cause a model trained on one hospital's data to fail to generalize to a different one¹¹⁴. When the diagnostic task has an imperfect inter-expert agreement, it has been shown that consensus diagnoses could substantially improve the performance of the machine-learning models trained on the data¹¹⁵. Adequate data curation is necessary for handling heterogeneous data. Also, obtaining the gold standard of patients' clinical status requires clinicians to review their clinical notes individually, which is prohibitively expensive on a population scale. A silver standard¹¹⁶ that employed natural-language-processing techniques and diagnostic codes to impute the true status of the patients has recently been proposed¹¹⁷. Sophisticated algorithms that can address the idiosyncrasies and noises of various datasets will enhance the reliability of the prediction models, and hence safety to use them in life-and-death decisions.

Several high-performing machine-learning models generate results that are difficult to interpret by unassisted humans. Although these models can achieve better-than-human performance, it is not straightforward to convey intuitive notions explaining the conclusions of the models, to identify model weakness, or to extract additional biological insights from these computational 'black boxes'. Recent approaches to explain image classification models include visualizing the convolution filters (Fig. 4d–g) or the relevance of each image region using saliency maps¹¹⁸ (Fig. 4b). However, model interpretation remains much more challenging for deep neural network models trained on data other than images; this is a focus of on-going research effort¹¹⁹.

Much of the recent progress in neural networks has been restricted to well-defined tasks that do not require integration of data across multiple modalities. Methodologies for applying deep neural networks to general diagnostics (such as the interpretation of signs and symptoms, past medical history, laboratory results and

clinical course) and treatment selection are less clear. Although deep learning has been successful in image classification, translation, voice recognition¹²⁰, sound synthesis¹²¹ and even neural network design¹²², clinical diagnostic and treatment tasks often require more context (for example, patient preferences, values, social support and medical history) than the narrow tasks that deep learning has mastered. Additionally, it is unclear how to apply transfer-learning methods to integrate insights learned from large non-medical datasets into algorithms for the analysis of multi-modality clinical data. This implies that larger-scale data-collection and data-annotation efforts are needed to develop end-to-end AI clinical systems.

Implementing a computing environment for collecting, storing and sharing EHRs and other sensitive health data remains a challenge¹²³. Privacy-preserving methods can permit secure data sharing through cloud services (such as third-party-hosted computing environments)¹²⁴. To implement such infrastructure widely, however, development of interoperable applications that meet the standard for the representation of clinical information is required¹²⁵. Deep and smooth data integration across healthcare applications and locations remains spotty and relatively slow. Nonetheless, emerging application-programming interfaces for clinical data are beginning to show significant adoption across multiple EHR vendors, such as the Substitutable Medical Applications and Reusable Technologies on the Fast Health Interoperability Resources platform^{126,127}.

Almost all of the reported medical applications of AI have been conducted on retrospective data collected for research and proof of principle¹²⁸. To validate the real-world utility of medical AI systems, prospective clinical studies that evaluate the systems' performance in clinical settings are needed¹²⁹. Prospective trials will better identify the fragility of AI models in real-world heterogeneous and noisy clinical environments, and point to ways to integrate medical AI into current clinical workflows.

Social, economic and legal challenges

As clinical AI systems mature, there will be an inevitable increase in their clinical use and deployment, which will lead to new social, economic and legal issues^{130,131}. Geoffrey Hinton — one of the pioneers of neural networks — and many AI researchers envision drastic changes in medical practice^{114,132}. AI is likely to improve the quality of care by reducing human error and decreasing physician fatigue arising from routine clinical tasks. However, it may not necessarily reduce physician workload, as clinical guidelines might suggest that examinations be carried out more often for at-risk patients. If AI for

routine clinical tasks is successfully deployed, it could free up time for physicians and allow them to concentrate on more sophisticated tasks and more 'high-touch' time with their patients. As an illustration, AI could facilitate ophthalmologists in triaging and reading the fundus photographs, enabling them to spend more time in the operating room or discussing treatment plans with their patients. Admittedly, AI could potentially replace some healthcare workers in carrying out routine tasks, which might in turn reshape the healthcare workforce and alter the current reimbursement framework in healthcare. Nonetheless, there is currently little empirical evidence of this kind of impact on the clinical workforce.

State-of-the-art AI applications will not reach their full potential unless they are integrated into clinical workflows^{133,134}. However, research has shown that the implementation of AI in healthcare is not trivial. It is widely acknowledged that clinical information systems result in many unintended consequences, including alert fatigue¹³⁵, imposition of additional workloads for clinicians, disruption of interpersonal (including doctor-to-patient) communication styles, and generation of specific hazards that require a higher level of vigilance to detect¹³⁶. For example, when a mammography CAD tool generates a false-negative result, radiologists are more likely to miss the diagnosis than when they are required to interpret the mammography films without CAD¹³⁷. Although many CAD models can be adjusted to balance the sensitivity and specificity needed for each clinical use case, it is challenging to identify the optimal clinical workflow that maximizes the performance of AI-assisted diagnosis¹²⁹. The experience of care providers and their patients shows that careful design and implementation are necessary, yet often lacking, to incorporate information systems into clinical settings^{138,139}.

From a regulatory perspective, clinical AI systems need to be certified before large-scale deployment. The FDA should anticipate an increasing number of premarket-approval submissions (safety and effectiveness evaluations of medical devices that present a potential and unreasonable risk of illness or injury) and 510(k) submissions (premarket submissions made to the FDA to demonstrate that a proposed device is at least as safe and effective as a legally marketed device) describing AI systems with direct clinical impact. Policymakers need to set specific criteria for the process of demonstrating non-inferiority in 510(k) submissions, such as the validation process and quality and representativeness of the validation data. Machine-learning-based models present a unique challenge to regulatory agencies because the models can evolve rapidly as more data and user feedback are collected. It is not clear how the updates should be evaluated. For instance, the new model may be better on average but have worse performance on a subset of patients. The FDA announced in April 2018 that it is moving toward a 'pre-certified approach' for AI software that learns and improves continuously¹⁴⁰. The proposed approach will first look at the technology developer, rather than primarily at the product¹⁴¹. As such, a clear guideline needs to be made regarding the certification of the teams that develop, revise and update the AI systems¹²⁹.

As ubiquitous data collection becomes more commonplace, consensus needs to be reached for a consent framework to guide health-related data sharing. For example, information recorded by mobile sensors¹⁴² may contain sensitive information, such as the location of the patients. It is imperative to involve the most representative and broad range of stakeholders when establishing a privacy policy framework for data collection and sharing.

AI in medicine will inevitably result in legal challenges regarding medical negligence attributed to complex decision support systems. When malpractice cases involving medical AI applications arise, the legal system will need to provide clear guidance on what entity holds the liability¹⁴³. Medical professional malpractice insurance needs to be clear about coverage when healthcare decisions are made in part by an AI system. With the deployment of automated AI for specific

clinical tasks, the credentials needed for diagnostic, therapeutic, supportive and paramedical tasks will need to be updated, and the roles of healthcare professionals will continue to evolve as various AI modules are incorporated into the standard of care.

To address the challenges, AI researchers and medical practitioners need to work together to prioritize and develop the applications that address crucial clinical needs. Hospital administrators would have to evaluate and mitigate clinical workflow disruption when introducing new AI applications. Companies will have to determine the right framework within which they can conduct prospective clinical trials that evaluate the performance of AI systems in the clinical setting. And insurers should assess the value created by medical AI systems and potentially revise their reimbursement policy to reduce the cost of healthcare while improving its quality. Multidisciplinary and multi-sector collaborations will be required to facilitate the development and deployment of medical AI applications.

Outlook

AI has enhanced clinical diagnosis and decision-making performance in several medical task domains. How this performance will translate into impact on the landscape of medical practice, including disease detection and treatments, will depend on how nimbly AI applications co-evolve with a healthcare system that is under tremendous financial strain while accommodating rapid advances in molecular and genomic science. Clinicians will need to adapt to their new roles as information integrators, interpreters and patient supporters, and the medical education system will have to provide them with the tools and methods to do so. Who will end up controlling, certifying or profiting from the application of AI is still to be determined, and therefore the balance of regulatory safeguards and market forces to ensure that patients benefit most must be a high priority.

Received: 12 December 2017; Accepted: 5 September 2018;
Published online: 3 October 2018

References

- Simonite, T. Google's AI eye doctor gets ready to go to work in India. *WIRED* (6 August 2017).
- Lee, R., Wong, T. Y. & Sabanayagam, C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye Vis.* **2**, 17 (2015).
- Lin, D. Y., Blumenkranz, M. S., Brothers, R. J. & Grosvenor, D. M. The sensitivity and specificity of single-field nonmydriatic monochromatic digital fundus photography with remote image interpretation for diabetic retinopathy screening: a comparison with ophthalmoscopy and standardized mydriatic color photography. *Am. J. Ophthalmol.* **134**, 204–213 (2002).
- Zheng, Y., He, M. & Congdon, N. The worldwide epidemic of diabetic retinopathy. *Indian J. Ophthalmol.* **60**, 428–431 (2012).
- Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
- Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
- Abrahamoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit. Med.* **1**, 39 (2018).
- Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach* (Prentice Hall, New Jersey, 2010).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. in *Advances in Neural Information Processing Systems* 1097–1105 (Curran Associates, Nevada, 2012).
- Lewis-Kraus, G. The great A.I. awakening. *The New York Times Magazine* (14 December 2016).
- Kundu, M., Nasipuri, M. & Basu, D. K. Knowledge-based ECG interpretation: a critical review. *Pattern Recognit.* **33**, 351–373 (2000).
- Jha, S. & Topol, E. J. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA* **316**, 2353–2354 (2016).
- Golub, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).

14. Wang, Y. et al. Gene selection from microarray data for cancer classification—a machine learning approach. *Comput. Biol. Chem.* **29**, 37–46 (2005).
15. Yu, K. H. et al. Predicting ovarian cancer patients' clinical response to platinum-based chemotherapy by their tumor proteomic signatures. *J. Proteome Res.* **15**, 2455–2465 (2016).
16. Yu, K. H. et al. Omics Analysis System for Precision Oncology (OASISPRO): a web-based omics analysis tool for clinical phenotype prediction. *Bioinformatics* **34**, 319–320 (2017).
17. Check Hayden, E. The automated lab. *Nature* **516**, 131–132 (2014).
18. Miller, R. A. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. *J. Am. Med. Inform. Assoc.* **1**, 8–27 (1994).
19. Musen, M. A., Middleton, B. & Greenes, R. A. in *Biomedical Informatics* (eds Shortliffe, E. H. & Cimino, J. J.) 643–674 (Springer, London, 2014).
20. Shortliffe, E. *Computer-Based Medical Consultations: MYCIN* Vol. 2 (Elsevier, New York, 2012).
21. Szolovits, P., Patil, R. S. & Schwartz, W. B. Artificial intelligence in medical diagnosis. *Ann. Intern. Med.* **108**, 80–87 (1988).
22. de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P. & Horrocks, J. C. Computer-aided diagnosis of acute abdominal pain. *Br. Med. J.* **2**, 9–13 (1972).
23. Shortliffe, E. H. et al. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Comput. Biomed. Res.* **8**, 303–320 (1975).
24. Barnett, G. O., Cimino, J. J., Hupp, J. A. & Hoffer, E. P. DXplain. An evolving diagnostic decision-support system. *JAMA* **258**, 67–74 (1987).
25. Miller, R. A., McNeil, M. A., Challinor, S. M., Masarie, F. E. Jr & Myers, J. D. The INTERNIST-1/QUICK MEDICAL REFERENCE Project — status report. *Western J. Med.* **145**, 816–822 (1986).
26. Berner, E. S. et al. Performance of four computer-based diagnostic systems. *N. Engl. J. Med.* **330**, 1792–1796 (1994).
27. Szolovits, P. & Pauker, S. G. Categorical and probabilistic reasoning in medical diagnosis. *Artif. Intell.* **11**, 115–144 (1978).
28. Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
29. Yu, K. H. & Snyder, M. Omics profiling in precision oncology. *Mol. Cell. Proteomics* **15**, 2525–2536 (2016).
30. Roberts, K. et al. Biomedical informatics advancing the national health agenda: the AMIA 2015 year-in-review in clinical and consumer informatics. *J. Am. Med. Inform. Assoc.* **24**, 185–190 (2017).
31. *Cloud AutoML^{ALPHA}* (Google Cloud); <https://cloud.google.com/automl/>
32. Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep Learning* 1 (MIT Press, Cambridge, 2016).
33. Gill, N. S. Overview and applications of artificial neural networks. *Xenonstack* <https://www.xenonstack.com/blog/data-science/artificial-neural-networks-applications-algorithms/> (2017).
34. *TOP500 List – November 2006* (TOP500); <https://www.top500.org/list/2006/11/>
35. Beam, A. L. & Kohane, I. S. Translating artificial intelligence into clinical care. *JAMA* **316**, 2368–2369 (2016).
36. Kametsky, L. et al. Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics* **27**, 1179–1180 (2011).
37. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).
38. Tomczak, K., Czerwinski, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, 68–77 (2015).
39. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
40. Ljosa, V., Sokolnicki, K. L. & Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nat. Methods* **9**, 637 (2012).
41. Williams, E. et al. The image data resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).
42. DesRoches, C. M. et al. Electronic health records in ambulatory care—a national survey of physicians. *N. Engl. J. Med.* **359**, 50–60 (2008).
43. Hsiao, C. J. et al. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. *Health Aff.* **32**, 1470–1477 (2013).
44. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
45. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
46. Yu, K. H. et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
47. Shademan, A. et al. Supervised autonomous robotic soft tissue surgery. *Sci. Transl. Med.* **8**, 337ra364 (2016).
48. Reed, J. C. *Chest Radiology: Plain Film Patterns and Differential Diagnoses* (Elsevier Health Sciences, Philadelphia, 2010).
49. Lodwick, G. S., Haun, C. L., Smith, W. E., Keller, R. F. & Robertson, E. D. Computer diagnosis of primary bone tumors: a preliminary report. *Radiology* **80**, 273–275 (1963).
50. van Ginneken, B., Setio, A. A., Jacobs, C. & Ciompi, F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *IEEE 12th International Symposium Biomedical Imaging (ISBI)* 286–289 (IEEE, 2015).
51. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).
52. Wang, X. et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Preprint at <https://arxiv.org/abs/1705.02315> (2017).
53. Yao, L. et al. Learning to diagnose from scratch by exploiting dependencies among labels. Preprint at <https://arxiv.org/abs/1710.10501> (2017).
54. Rajpurkar, P. et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at <https://arxiv.org/abs/1711.05225> (2017).
55. Samala, R. K. et al. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. *Med. Phys.* **43**, 6654–6666 (2016).
56. Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L. & Lopez, M. A. G. Convolutional neural networks for mammography mass lesion classification. In *IEEE 37th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC)* 797–800 (IEEE, 2015).
57. *510(k) Premarket Notification* (FDA, 2017); <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K163253>
58. Marr, B. First FDA approval for clinical cloud-based deep learning in healthcare. *Forbes* (20 January 2017).
59. Rigel, D. S., Friedman, R. J., Kopf, A. W. & Polsky, D. ABCDE—an evolving concept in the early detection of melanoma. *Arch. Dermatol.* **141**, 1032–1034 (2005).
60. Thomas, L. et al. Semiological value of ABCDE criteria in the diagnosis of cutaneous pigmented tumors. *Dermatology* **197**, 11–17 (1998).
61. Ercal, F., Chawla, A., Stoecker, W. V., Lee, H. C. & Moss, R. H. Neural network diagnosis of malignant melanoma from color images. *IEEE Trans. Biomed. Eng.* **41**, 837–845 (1994).
62. Wolf, J. A. et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA Dermatol.* **149**, 422–426 (2013).
63. Panwar, N. et al. Fundus photography in the 21st century — a review of recent technological advances and their implications for worldwide healthcare. *Telemed. J. E. Health* **22**, 198–208 (2016).
64. American Diabetes Association. 10. Microvascular complications and foot care. *Diabetes Care* **40**, 88–98 (2017).
65. Menke, A., Casagrande, S., Geiss, L. & Cowie, C. C. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. *JAMA* **314**, 1021–1029 (2015).
66. Abràmoff, M. D. et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology Visual Sci.* **57**, 5200–5206 (2016).
67. Rorke, L. B. Pathologic diagnosis as the gold standard. *Cancer* **79**, 665–667 (1997).
68. Lakhani, S. R. & Ashworth, A. Microarray and histopathological analysis of tumours: the future and the past? *Nat. Rev. Cancer* **1**, 151–157 (2001).
69. Rubegni, P. et al. Automated diagnosis of pigmented skin lesions. *Int. J. Cancer* **101**, 576–580 (2002).
70. Stang, A. et al. Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. *Lung Cancer* **52**, 29–36 (2006).
71. Yu, K. H. et al. Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Syst.* **5**, 620–627 (2017).
72. Litjens, G. et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci. Rep.* **6**, 26286 (2016).
73. Beijndorf, B. E. et al. Machine learning detection of breast cancer lymph node metastases. *JAMA* **318**, 2199–2210 (2017).
74. Cireşan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2013* (eds Mori, K. et al.) 411–418 (Springer, Berlin, Heidelberg, 2013).
75. Manak, M. S. et al. Live-cell phenotypic-biomarker microfluidic assay for the risk stratification of cancer patients via machine learning. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-018-0285-z> (2018).
76. Robboy, S. J. et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch. Pathol. Lab. Med.* **137**, 1723–1732 (2013).

77. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
78. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).
79. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).
80. DePristo, M. & Poplin, R. DeepVariant: highly accurate genomes with deep neural networks. *Google AI Blog* <https://research.googleblog.com/2017/12/deepvariant-highly-accurate-genomes.html> (2017).
81. Poplin, R. et al. Creating a universal SNP and small indel variant caller with deep neural networks. Preprint at <https://www.biorxiv.org/content/early/2016/12/14/092890> (2018).
82. Kamps, R. et al. Next-generation sequencing in oncology: genetic diagnosis, risk prediction and cancer classification. *Int. J. Mol. Sci.* **18**, 308 (2017).
83. He, Z. & Yu, W. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* **34**, 215–225 (2010).
84. Zhang, Z. et al. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res.* **64**, 5882–5890 (2004).
85. Wallden, B. et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8**, 54 (2015).
86. Sweeney, T. E., Wong, H. R. & Khatri, P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. *Sci. Transl. Med.* **8**, 346ra391 (2016).
87. Huang, T., Hoffman, B., Meschino, W., Kingdom, J. & Okun, N. Prediction of adverse pregnancy outcomes by combinations of first and second trimester biochemistry markers used in the routine prenatal screening of Down syndrome. *Prenat. Diagn.* **30**, 471–477 (2010).
88. Mook, S. et al. Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. *Ann. Surg. Oncol.* **17**, 1406–1413 (2010).
89. Farina, D. et al. Man/machine interface based on the discharge timings of spinal motor neurons after targeted muscle reinnervation. *Nat. Biomed. Eng.* **1**, 0025 (2017).
90. Altman, R. B. Artificial intelligence (AI) systems for interpreting complex medical datasets. *Clin. Pharmacol. Ther.* **101**, 585–586 (2017).
91. Cai, X. et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J. Am. Med. Inform. Assoc.* **23**, 553–561 (2016).
92. Makar, M., Ghassemi, M., Cutler, D. M. & Obermeyer, Z. Short-term mortality prediction for elderly patients using medicare claims data. *Int. J. Mach. Learn. Comput.* **5**, 192–197 (2015).
93. Ng, T., Chew, L. & Yap, C. W. A clinical decision support tool to predict survival in cancer patients beyond 120 days after palliative chemotherapy. *J. Palliat. Med.* **15**, 863–869 (2012).
94. Delen, D., Oztekin, A. & Kong, Z. J. A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artif. Intell. Med.* **49**, 33–42 (2010).
95. Churpek, M. M. et al. Predicting cardiac arrest on the wards: a nested case-control study. *Chest* **141**, 1170–1176 (2012).
96. Churpek, M. M. et al. Multicenter development and validation of a risk stratification tool for ward patients. *Am. J. Respir. Crit. Care Med.* **190**, 649–655 (2014).
97. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* <https://doi.org/10.1038/s41551-018-0304-0> (2018).
98. Li, X. et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biol.* **15**, e2001402 (2017).
99. Majumder, S., Mondal, T. & Deen, M. J. Wearable sensors for remote health monitoring. *Sensors* **17**, 130 (2017).
100. Pastorino, M., Arredondo, M., Cancela, J. & Guillen, S. Wearable sensor network for health monitoring: the case of Parkinson disease. *J. Phys. Conf. Ser.* **450**, 012055 (2013).
101. Mercer, K., Li, M., Giangregorio, L., Burns, C. & Grindrod, K. Behavior change techniques present in wearable activity trackers: a critical analysis. *JMIR Mhealth Uhealth* **4**, e40 (2016).
102. Takacs, J. et al. Validation of the Fitbit One activity monitor device during treadmill walking. *J. Sci. Med. Sport* **17**, 496–500 (2014).
103. Yang, R., Shin, E., Newman, M. W. & Ackerman, M. S. When fitness trackers don't 'fit': end-user difficulties in the assessment of personal tracking device accuracy. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* 623–634 (ACM, 2015).
104. Endeavour Partners. Inside wearables: how the science of human behavior change offers the secret to long-term engagement. *Medium* <https://blog.endeavourpartners.com/inside-wearables-how-the-science-of-human-behavior-change-offers-the-secret-to-long-term-engagement-a15b3c7d4cf3> (2017).
105. Herz, J. C. Wearables are totally failing the people who need them most. *Wired* (11 June 2014).
106. Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D. & Mamykina, L. No longer wearing: investigating the abandonment of personal health-tracking technologies on Craigslist. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* 647–658 (ACM, 2015).
107. Wheeler, M. J. Overview on robotics in the laboratory. *Ann. Clin. Biochem.* **44**, 209–218 (2007).
108. Moustiris, G. P., Hiridis, S. C., Deliparaschos, K. M. & Konstantinidis, K. M. Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature. *Int. J. Med. Robot.* **7**, 375–392 (2011).
109. Gomes, P. Surgical robotics: reviewing the past, analysing the present, imagining the future. *Robot. Comput. Integr. Manuf.* **27**, 261–266 (2011).
110. Majdani, O. et al. A robot-guided minimally invasive approach for cochlear implant surgery: preliminary results of a temporal bone study. *Int. J. Comput. Assist. Radiol. Surg.* **4**, 475–486 (2009).
111. Elek, R. et al. Recent trends in automating robotic surgery. In *2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES)* 27–32 (IEEE, 2016).
112. Liew, C. The future of radiology augmented with artificial intelligence: a strategy for success. *Eur. J. Radiol.* **102**, 152–156 (2018).
113. Jones, L., Golan, D., Hanna, S. & Ramachandran, M. Artificial intelligence, machine learning and the evolution of healthcare: a bright future or cause for concern? *Bone Joint Res.* **7**, 223–225 (2018).
114. Obermeyer, Z. & Emanuel, E. J. Predicting the future — big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).
115. Krause, J. et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* **125**, 1264–1272 (2018).
116. Rebholz-Schuhmann, D. et al. The CALBC silver standard corpus for biomedical named entities—a study in harmonizing the contributions from four independent named entity taggers. In *LREC* 568–573 (2010).
117. Kirby, J. C. et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J. Am. Med. Inform. Assoc.* **23**, 1046–1052 (2016).
118. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at <https://arxiv.org/abs/1312.6034> (2013).
119. Ribeiro, M. T., Singh, S. & Guestrin, C. “Why should I trust you?”: explaining the predictions of any classifier. Preprint at <https://arxiv.org/abs/1602.04938> (2016).
120. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
121. Boulanger-Lewandowski, N., Bengio, Y. & Vincent, P. Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. Preprint at <https://arxiv.org/abs/1206.6392> (2012).
122. Zoph, B. & Le, Q. V. Neural architecture search with reinforcement learning. Preprint at <https://arxiv.org/abs/1611.01578> (2016).
123. Lee, L. M. & Gostin, L. O. Ethical collection, storage, and use of public health data: a proposal for a national privacy protection. *JAMA* **302**, 82–84 (2009).
124. Narayan, S., Gagné, M. & Safavi-Naini, R. Privacy preserving EHR system using attribute-based infrastructure. In *Proceedings of the 2010 ACM Workshop on Cloud Computing Security Workshop* 47–52 (ACM, 2010).
125. Dolin, R. H. et al. HL7 Clinical Document Architecture, Release 2. *J. Am. Med. Inform. Assoc.* **13**, 30–39 (2006).
126. Mandl, K. D. & Kohane, I. S. Escaping the EHR trap—the future of health IT. *N. Engl. J. Med.* **366**, 2240–2242 (2012).
127. Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S. & Ramoni, R. B. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J. Am. Med. Inform. Assoc.* **23**, 899–908 (2016).
128. All eyes are on AI. *Nat. Biomed. Eng.* **2**, 139 (2018).
129. Yu, K. H. & Kohane, I. S. Framing the challenges of artificial intelligence in medicine. *BMJ Qual. Safety* <https://doi.org/10.1136/bmjqs-2018-008551> (2018).
130. Dignum, V. Ethics in artificial intelligence: introduction to the special issue. *Ethics Inf. Technol.* **20**, 1–3 (2018).
131. Price, I. & Nicholson, W. *Artificial Intelligence in Health Care: Applications and Legal Implications* (Univ. Michigan Law School, 2017).
132. Mukherjee, S. A.I. versus M.D. What happens when diagnosis is automated? *The New Yorker* (3 April 2017).
133. Del Beccaro, M. A., Jeffries, H. E., Eisenberg, M. A. & Harry, E. D. Computerized provider order entry implementation: no association with increased mortality rates in an intensive care unit. *Pediatrics* **118**, 290–295 (2006).

134. Longhurst, C. A. et al. Decrease in hospital-wide mortality rate after implementation of a commercially sold computerized physician order entry system. *Pediatrics* **126**, 14–21 (2010).
135. Carspecken, C. W., Sharek, P. J., Longhurst, C. & Pageler, N. M. A clinical case of electronic health record drug alert fatigue: consequences for patient outcome. *Pediatrics* **131**, 1970–1973 (2013).
136. Ash, J. S., Berg, M. & Coiera, E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J. Am. Med. Inform. Assoc.* **11**, 104–112 (2004).
137. Lehman, C. D. et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).
138. Koppel, R. et al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* **293**, 1197–1203 (2005).
139. Middleton, B. et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. *J. Am. Med. Inform. Assoc.* **20**, 2–8 (2013).
140. Gottlieb, S. *Twitter* (12 April 2018); <https://twitter.com/SGottliebFDA/status/984378648781312002>
141. *Digital Health Software Precertification (Pre-Cert) Program* (FDA); <https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.htm>
142. Estrin, D. & Sim, I. Open mHealth architecture: an engine for health care innovation. *Science* **330**, 759–760 (2010).
143. Shortliffe, E. H. Computer programs to support clinical decision making. *JAMA* **258**, 61–66 (1987).
144. Armbruster, D. A., Overcash, D. R. & Reyes, J. Clinical chemistry laboratory automation in the 21st century—amat victoria curam (victory loves careful preparation). *Clin. Biochem. Rev.* **35**, 143–153 (2014).
145. Rosenfeld, L. A golden age of clinical chemistry: 1948–1960. *Clin. Chem.* **46**, 1705–1714 (2000).
146. Kuperman, G. J. et al. Medication-related clinical decision support in computerized provider order entry systems: a review. *J. Am. Med. Inform. Assoc.* **14**, 29–40 (2007).
147. Glassman, P. A., Simon, B., Belperio, P. & Lanto, A. Improving recognition of drug interactions: benefits and barriers to using automated drug alerts. *Med. Care* **40**, 1161–1171 (2002).
148. FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures. <https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm608833.htm> (FDA, 2018).
149. Haenssle, H. A. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842 (2018).
150. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at <https://arxiv.org/abs/1409.1556> (2014).
151. Murphy, K. P. & Bach F. *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, 2012).

Acknowledgements

K.-H.Y. is supported by a Harvard Data Science Postdoctoral Fellowship. I.S.K. was supported in part by the NIH grant OT3OD025466. Figure 4 was generated by using the computational infrastructure supported by the AWS Cloud Credits for Research, the Microsoft Azure Research Award, and the NVIDIA GPU Grant Programme.

Author contributions

K.-H.Y. conceived and designed the article, performed the literature review and wrote and revised the manuscript. A.L.B. and I.S.K. edited the manuscript. I.S.K. supervised the work.

Competing interests

Harvard Medical School (K.-H.Y.) submitted a provisional patent application on digital pathology profiling to the United States Patent and Trademark Office (USPTO).

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to I.S.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2018