REVIEW ARTICLE | FOCUS
https://doi.org/10.1038/s41591-018-0300-7
nature medicine

High-performance medicine: the convergence of human and artificial intelligence

Eric J. Topol

The use of artificial intelligence, and the deep-learning subtype in particular, has been enabled by the use of labeled big data, along with markedly enhanced computing power and cloud storage, across all sectors. In medicine, this is beginning to have an impact at three levels: for clinicians, predominantly via rapid, accurate image interpretation; for health systems, by improving workflow and the potential for reducing medical errors; and for patients, by enabling them to process their own data to promote health. The current limitations, including bias, privacy and security, and lack of transparency, along with the future directions of these applications will be discussed in this article. Over time, marked improvements in accuracy, productivity, and workflow will likely be actualized, but whether that will be used to improve the patient–doctor relationship or facilitate its erosion remains to be seen.

Medicine is at the crossroad of two major trends. The first is a failed business model, with increasing expenditures and jobs allocated to healthcare, but with deteriorating key outcomes, including reduced life expectancy and high infant, childhood, and maternal mortality in the United States[1,2]. This exemplifies a paradox that is not at all confined to American medicine: investment of more human capital with worse human health outcomes. The second is the generation of data in massive quantities, from sources such as high-resolution medical imaging, biosensors with continuous output of physiologic metrics, genome sequencing, and electronic medical records. The limits on analysis of such data by humans alone have clearly been exceeded, necessitating an increased reliance on machines. Accordingly, at the same time that there is more dependence than ever on humans to provide healthcare, algorithms are desperately needed to help. Yet the integration of human and artificial intelligence (AI) for medicine has barely begun.

Looking deeper, there are notable, longstanding deficiencies in healthcare that are responsible for its path of diminishing returns. These include a large number of serious diagnostic errors, mistakes in treatment, an enormous waste of resources, inefficiencies in workflow, inequities, and inadequate time between patients and clinicians[3,4]. Eager for improvement, leaders in healthcare and computer scientists have asserted that AI might have a role in addressing all of these problems. That might eventually be the case, but researchers are at the starting gate in the use of neural networks to ameliorate the ills of the practice of medicine. In this Review, I have gathered much of the existing base of evidence for the use of AI in medicine, laying out the opportunities and pitfalls.

Artificial intelligence for clinicians

Almost every type of clinician, ranging from specialty doctor to paramedic, will be using AI technology, and in particular deep learning, in the future. This largely involved pattern recognition using deep neural networks (DNNs) (Box 1) that can help interpret medical scans, pathology slides, skin lesions, retinal images, electrocardiograms, endoscopy, faces, and vital signs. The neural net interpretation is typically compared with physicians' assessments using a plot of true-positive versus false-positive rates, known as a receiver operating characteristic (ROC), for which the area under the curve (AUC) is used to express the level of accuracy (Box 1).

Radiology. One field that has attracted particular attention for application of AI is radiology[5]. Chest X-rays are the most common type of medical scan, with more than 2 billion performed worldwide per year. In one study, the accuracy of one algorithm, based on a 121-layer convolutional neural network, in detecting pneumonia in over 112,000 labeled frontal chest X-ray images was compared with that of four radiologists, and the conclusion was that the algorithm outperformed the radiologists. However, the algorithm's AUC of 0.76, although somewhat better than that for two previously tested DNN algorithms for chest X-ray interpretation[5], is far from optimal. In addition, the test used in this study is not necessarily comparable with the daily tasks of a radiologist, who will diagnose much more than pneumonia in any given scan. To further validate the conclusions of this study, a comparison with results from more than four radiologists should be made. A team at Google used an algorithm that analyzed the same image set as in the previously discussed study to make 14 different diagnoses, resulting in AUC scores that ranged from 0.63 for pneumonia to 0.87 for heart enlargement or a collapsed lung[6]. More recently, in another related study, it was shown that a DNN that is currently in use in hospitals in India for interpretation of four different chest X-ray key findings was at least as accurate as four radiologists[7]. For the narrower task of detecting cancerous pulmonary nodules on a chest X-ray, a DNN that retrospectively assessed scans from over 34,000 patients achieved a level of accuracy exceeding 17 of 18 radiologists[8]. It can be difficult for emergency room doctors to accurately diagnose wrist fractures, but a DNN led to marked improvement, increasing sensitivity from 81% to 92% and reducing misinterpretation by 47% (ref. [9]).

Similarly, DNNs have been applied across a wide variety of medical scans, including bone films for fractures and estimation of aging[10–12], classification of tuberculosis[13], and vertebral compression fractures[14]; computed tomography (CT) scans for lung nodules[15], liver masses[16], pancreatic cancer[17], and coronary calcium score[18]; brain scans for evidence of hemorrhage[19], head trauma[20], and acute referrals[21]; magnetic resonance imaging[22]; echocardiograms[23,24]; and mammographies[25,26]. A unique imaging-recognition study focusing on the breadth of acute neurologic events, such as stroke or head trauma, was carried out on over 37,000 head CT 3-D scans, which the algorithm analyzed for 13 different anatomical findings versus gold-standard labels (annotated by expert radiologists) and achieved an AUC of 0.73 (ref. [27]). A simulated prospective, double-blind, randomized control trial was conducted with real cases from the dataset and showed that the deep-learning algorithm could interpret scans 150 times faster than radiologists (1.2 versus 177 seconds). But the conclusion that the algorithm's diagnostic accuracy in screening acute neurologic scans was poorer than human

Department of Molecular Medicine, Scripps Research, La Jolla, CA, USA. e-mail: etopol@scripps.edu

Box 1 | Deep learning

While the roots of AI date back over 80 years from concepts laid out by Alan Turing[204,205] and Warren McCulloch and Walter Pitts[206], it was not until 2012 that the subtype of deep learning was widely accepted as a viable form of AI[207]. A deep learning neural network consists of digitized inputs, such as an image or speech, which proceed through multiple layers of connected 'neurons' that progressively detect features, and ultimately provides an output. By analyzing 1.2 million carefully annotated images from over 15 million in the ImageNet database, a DNN achieved, for that point in time, an unprecedented low error rate for automated image classification. That report, along with Google Brain's 10 million images from YouTube videos to accurately detect cats, laid the groundwork for future progress. Within 5 years, in specific large data-labeled test sets, deep-learning algorithms for image recognition surpassed the human accuracy rate[208,209], and, in parallel, suprahuman performance was demonstrated for speech recognition.

The basic DNN architecture is like a club sandwich turned on its side, with an input layer, a number of hidden layers ranging from 5 to 1,000, each responding to different features of the image (like shape or edges), and an output layer. The layers are 'neurons,' comprising a neural network, even though there is little support of the notion that these artificial neurons function similarly to human neurons. A key differentiating feature of deep learning compared with other subtypes of AI is its autodidactic quality; the neural network is not designed by humans, but rather the number of layers (Fig. 1) is determined by the data itself. Image and speech recognition have primarily used supervised learning, with training from known patterns and labeled input data, commonly referred to as ground truths. Learning from unknown patterns without labeled input data—unsupervised learning—has very rarely been applied to date. There are many types of DNNs and learning, including convolutional, recurrent, generative adversarial, transfer, reinforcement, representation, and transfer (for review see refs. [210,211]). Deep-learning algorithms have been the backbone of computer performance that exceeds human ability in multiple games, including the Atari video game Breakout, the classic game of Go, and Texas Hold'em poker. DNNs are largely responsible for the exceptional progress in autonomous cars, which is viewed by most as the pinnacle technological achievement of AI to date. Notably, except in the cases of games and self-driving cars, a major limitation to interpretation of claims reporting suprahuman performance of these algorithms is that analytics are performed on previously generated data in silico, not prospectively in real-world clinical conditions. Furthermore, the lack of large datasets of carefully annotated images has been limiting across various disciplines in medicine. Ironically, to compensate for this deficiency, generative adversarial networks have been used to synthetically produce large image datasets at high resolution, including mammograms, skin lesions, echocardiograms, and brain and retina scans, that could be used to help train DNNs[212–216].

performance was sobering and indicates that there is much more work to do.

For each of these studies, a relatively large number of labeled scans were used for training and subsequent evaluation, with AUCs ranging from 0.99 for hip fracture to 0.84 intracranial bleeding and liver masses to 0.56 for acute neurologic case screening. It is not possible to compare DNN accuracy from one study to the next because of marked differences in methodology. Furthermore, ROC and AUC metrics are not necessarily indicative of clinical utility or even the best way to express accuracy of the model's performance[28,29]. Furthermore, many of these reports still only exist in preprint form and have not appeared in peer-reviewed publications. Validation of the performance of an algorithm in terms of its accuracy is not equivalent to demonstrating clinical efficacy. This is what Pearse Keane and I have referred to as the 'AI chasm'—that is, an algorithm with an AUC of 0.99 is not worth very much if it is not proven to improve clinical outcomes[30]. Among the studies that have gone through peer review (many of which are summarized in Table 1), the only prospective validation studies in a real-world setting have been for diabetic retinopathy[31,32], detection of wrist fractures in the emergy room setting[33], histologic breast cancer metastases[34,35], very small colonic polyps[36,37], and congenital cataracts in a small group of children[38]. The field clearly is far from demonstrating very high and reproducible machine accuracy, let alone clinical utility, for most medical scans and images in the real-world clinical environment (Table 1).

Pathology

Pathologists have been much slower at adopting digitization of scans than radiologists[39]—they are still not routinely converting glass slides to digital images and use whole-slide imaging (WSI) to enable viewing of an entire tissue sample on a slide. Marked heterogeneity and inconsistency among pathologists' interpretations of slides has been amply documented, exemplified by a lack of agreement

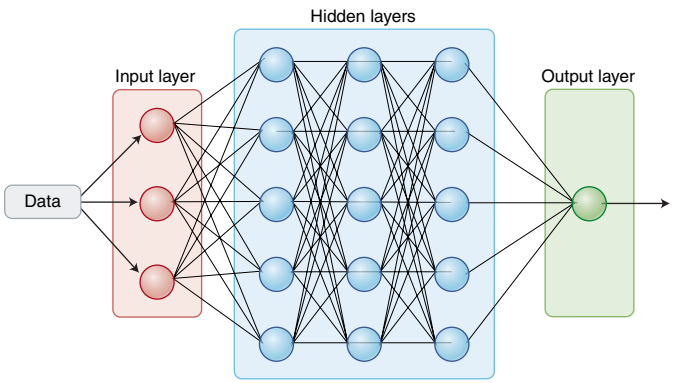

Fig. 1 | A deep neural network, simplified. Credit: Debbie Maizels/Springer Nature

in diagnosis of common types of lung cancer (K = 0.41–0.46)[40]. Deep learning of digitized pathology slides offers the potential to improve accuracy and speed of interpretation, as assessed in a few retrospective studies. In a study of WSI of breast cancer, with or without lymph node metastases, that compared the performance of 11 pathologists with that of multiple algorithmic interpretations, the results varied and were affected in part by the length of time that the pathologists had to review the slides[41]. Some of the five algorithms performed better than the group of pathologists, who had varying expertise. The pathologists were given 129 test slides and had less than 1 minute for review per slide, which likely does not reflect normal workflow. On the other hand, when one expert pathologist had no time limits and took 30 hours to review the same slide set, the results were comparable with the algorithm for detecting noninvasive ductal carcinoma[42].

Table 1 | Peer-reviewed publications of AI algorithms compared with doctors

Specialty	Images	Publication
Radiology/ neurology	CT head, acute neurological events	Titano et al. [27]
	CT head for brain hemorrhage	Arbabshirani et al.[19]
	CT head for trauma	Chilamkurthy et al.[20]
	CXR for metastatic lung nodules	Nam et al.[8]
	CXR for multiple findings	Singh et al.[7]
	Mammography for breast density	Lehman et al.[26]
	Wrist X-ray*	Lindsey et al.[9]
Pathology	Breast cancer	Ehteshami Bejnordi et al.[41]
	Lung cancer (+driver mutation)	Coudray et al.[33]
	Brain tumors (+methylation)	Capper et al.[45]
	Breast cancer metastases*	Steiner et al.[35]
	Breast cancer metastases	Liu et al.[34]
Dermatology	Skin cancers	Esteva et al.[47]
	Melanoma	Haenssle et al.[48]
	Skin lesions	Han et al.[49]
Ophthalmology	Diabetic retinopathy	Gulshan et al.[51]
	Diabetic retinopathy*	Abramoff et al.[31]
	Diabetic retinopathy*	Kanagasingam et al.[32]
	Congenital cataracts	Long et al.[38]
	Retinal diseases (OCT)	De Fauw et al.[56]
	Macular degeneration	Burlina et al.[52]
	Retinopathy of prematurity	Brown et al.[60]
	AMD and diabetic retinopathy	Kermany et al.[53]
Gastroenterology	Polyps at colonoscopy*	Mori et al.[36]
	Polyps at colonoscopy	Wang et al.[37]
Cardiology	Echocardiography	Madani et al.[23]
	Echocardiography	Zhang et al.[24]

Prospective studies are denoted with an asterisk.

Table 2 | FDA AI approvals are accelerating

Company	FDA Approval	Indication
Apple	September 2018	Atrial fibrillation detection
Aidoc	August 2018	CT brain bleed diagnosis
iCAD	August 2018	Breast density via mammography
Zebra Medical	July 2018	Coronary calcium scoring
Bay Labs	June 2018	Echocardiogram EF determination
Neural Analytics	May 2018	Device for paramedic stroke diagnosis
IDx	April 2018	Diabetic retinopathy diagnosis
Icometrix	April 2018	MRI brain interpretation
Imagen	March 2018	X-ray wrist fracture diagnosis
Viz.ai	February 2018	CT stroke diagnosis
Arterys	February 2018	Liver and lung cancer (MRI, CT) diagnosis
MaxQ-AI	January 2018	CT brain bleed diagnosis
Alivecor	November 2017	Atrial fibrillation detection via Apple Watch
Arterys	January 2017	MRI heart interpretation

Other studies have assessed deep-learning algorithms for classifying breast cancer[43] and lung cancer[40] without direct comparison with pathologists. Brain tumors can be challenging to subtype, and machine learning using tumor DNA methylation patterns via sequencing led to markedly improved classification compared with pathologists using traditional histological data[44,45]. DNA methylation generates extensive data and at present is rarely performed in the clinic for classification of tumors, but this study suggests another potential for AI to provide improved diagnostic accuracy in the future. A deep-learning algorithm for lung cancer digital pathology slides not only was able to accurately classify tumors, but also was trained to detect the pattern of several specific genomic driver mutations that would not otherwise be discernible by pathologists[33].

The first prospective study to test the accuracy of an algorithm classifying digital pathology slides in a real clinical setting was an assessment of the identification of presence of breast cancer micrometastases in slides by six pathologists compared with a DNN (that had been retrospectively validated[34]). The combination of pathologists and the algorithm led to the best accuracy, and the algorithm markedly sped up the review of slides[35]. This study is particularly notable, as the synergy of the combined pathologist and algorithm interpretation was emphasized instead of the pervasive clinician-versus-algorithm comparison. Apart from classifying tumors more accurately by data processing, the use of a deep-learning algorithm to sharpen out-of-focus images may also prove useful[46]. A number of proprietary algorithms for image interpretation have been approved by the Food and Drug Administration (FDA), and the list is expanding rapidly (Table 2), yet there have been few peer-reviewed publications from most of these companies. In 2018, the FDA published a fast-track approval plan for AI medical algorithms.

Dermatology. For algorithms classifying skin cancer by image analysis, the accuracy of diagnosis of deep-learning networks has been compared with that of dermatologists. In a study using a large training dataset of nearly 130,000 photographic and dermascopic digitized images, 21 US board-certified dermatologists were at least matched in performance by an algorithm, which had an AUC of 0.96 for carcinoma[47] and of 0.94 for melanoma specifically. Subsequently, the accuracy of melanoma skin cancer diagnosis by a group of 58 international dermatologists was compared with a convolutional neural network; the mean ROCs were 0.79 versus 0.86, respectively, reflecting an improved performance of the algorithm compared with most of the physicians[48]. A third study carried out algorithmic assessment of 12 skin diseases, including basal cell carcinoma, squamous cell carcinoma, and melanoma, and compared this with 16 dermatologists, with the algorithm achieving an AUC of 0.96 for melanoma[49]. None of these studies were conducted in the clinical setting, in which a doctor would perform physical inspection and shoulder responsibility for making an accurate diagnosis. Notwithstanding these concerns, most skin lesions are diagnosed by primary care doctors, and problems with inaccuracy have been underscored; if AI can be reliably shown to simulate experienced dermatologists, that would represent a significant advance.

Ophthalmology. There have been a number of studies comparing performance between algorithms and ophthalmologists in diagnosing

different eye conditions. After training with over 128,000 retinal fundus photographs labeled by 54 ophthalmologists, a neural network was used to assess over 10,000 retinal fundus photographs from more than 5,000 patients for diabetic retinopathy, and the neural network's grading was compared with seven or eight ophthalmologists for all-cause referable diagnoses (moderate or worse retinopathy or macular edema; scale: none, mild, moderate, severe, or proliferative). In two separate validation sets, the AUC was 0.99 (refs. [50,51]). In a study in which retinal fundus photographs were used for the diagnosis of age-related macular degeneration (AMD), the accuracy for DNN algorithms ranged between 88% and 92%, nearly as high as for expert ophthalmologists[52]. Performance of a deep-learning algorithm for interpreting retinal optical coherence tomography (OCT) was compared with ophthalmologists for diagnosis of either of the two most common causes of vision loss: diabetic retinopathy or AMD. After the algorithm was trained on a dataset of over 100,000 OCT images, validation was performed in 1,000 of these images, and performance was compared with six ophthalmologists. The algorithm's AUC for OCT-based urgent referral was 0.999 (refs. [53–55]).

Another deep-learning OCT retinal study went beyond the diagnosis of diabetic retinopathy or macular degeneration. A group of 997 patients with a wide range of 50 retinal pathologies was assessed for urgent referral by an algorithm (using two different types of OCT devices that produce 3-D images) and results were compared with those from experts: four retinal specialists and four optometrists, with an AUC for accuracy of urgent referral triage to replace false alarm of 0.992. The algorithm did not miss a single urgent referral case. Notably, the eight clinicians agreed on only 65% of the referral decisions. Errors on the correct referral decision were reduced for both types of clinicians by integrating the fundus photograph and notes on the patient, but the algorithm's error rate (without notes or fundus photographs) of 3.5% was as good or better than all eight experts[56]. One unique aspect of this study was the transparency of the two neural networks used, one for mapping the eye OCT scans into a tissue schematic and the other for the classifier of eye disease. The user (patient) can watch a video that shows what portions of his or her scan were used to reach the algorithm's conclusions along with the level of confidence it has for the diagnosis. This sets a new bar for future efforts to unravel the 'black box' of neural networks.

In a prospective trial conducted in primary care clinics, 900 patients with diabetes but no known retinopathy were assessed by a proprietary system (an imaging device combined with an algorithm) made by IDx (Iowa City, IA) that obtained retinal fundus photographs and OCT and by established reading centers with expertise in interpreting these images[30,31]. The algorithm was used at primary care clinics up until the clinical trial was autodidactic and thus locked for testing, but it achieved a sensitivity of 87% and specificity of 91% for the 819 patients (91% of the enrolled cohort) with analyzable images. This trial led to FDA approval of the IDx device and algorithm for autonomous detection, that is, without the need for a clinician, of 'more than mild' diabetic retinopathy. The regulatory oversight in dealing with deep-learning algorithms is tricky because it does not currently allow continued autodidactic functionality but instead necessitates fixing the software to behave like a non-AI diagnostic system[30]. Notwithstanding this point along with the unknown extent of uptake of the device, the study represents a milestone as the first prospective assessment of AI in the clinic. The accuracy results are not as good as the aforementioned in silico studies, which should be anticipated. A small prospective real-world assessment of a DNN for diabetic retinopathy in primary care clinics, with eye exams performed by nurses, led to a high false-positive diagnosis rate[32].

While the studies of retinal OCT and fundus images have thus far focused on eye conditions, recent work suggests that these images can provide a window to the brain for early diagnosis of dementia, including Alzheimer's disease[57].

The potential use of retinal photographs also appears to transcend eye diseases per se. Images from over 280,000 patients were assessed by DNN for cardiovascular risk factors, including age, gender, systolic blood pressure, smoking status, hemoglobin A1c, and likelihood of having a major adverse cardiac event, with validation in two independent datasets. The AUC for gender at 0.97 was notable, indicating that the algorithm could identify gender accurately from the retinal photo, but the others were in the range of 0.70, suggesting that there may be a signal that, through further pursuit, could be useful for monitoring patients for control of their risk factors[58,59].

Other less common eye conditions that have been assessed by neural networks include congenital cataracts[38] and retinopathy of prematurity in newborns[60], both with accuracy comparable with that of eye specialists.

Cardiology. The major images that cardiologists use in practice are electrocardiograms (ECG) and echocardiograms, both of which have been assessed with DNNs. There is a nearly 40-year history of machine-read ECGs using rules-based algorithms with notable inaccuracy[61]. When deep learning was used to diagnose heart attack in a small retrospective dataset of 549 ECGs, a sensitivity of 93% and specificity of 90% were reported, which was comparable with cardiologists[62]. Over 64,000 one-lead ECGs (from over 29,000 patients) were assessed for arrhythmia by a DNN and six cardiologists, with comparable accuracy across 14 different electrical conduction disturbances[63]. For echocardiography, a small set of 267 patient studies (consisting of over 830,000 still images) were classified into 15 standard views (such as apical 4-chamber or subcostal) by a DNN and by cardiologists. The overall accuracy for single still images was 92% for the algorithm and 79% for four board-certified echocardiographers, but this does not reflect the real-world reading of studies, which are in-motion video loops[23]. An even larger retrospective study of over 8,000 echocardiograms showed high accuracy for classification of hypertrophic cardiomyopathy (AUC, 0.93), cardiac amyloid (AUC, 0.87), and pulmonary artery hypertension (AUC, 0.85)[24].

Gastroenterology. Finding diminutive (<5 mm) adenomatous or sessile polyps at colonoscopy can be exceedingly difficult for gastroenterologists. The first prospective clinical validation of AI was performed in 325 patients who collectively had 466 tiny polyps, with an accuracy of 94% and negative predictive value of 96% during real-time, routine colonoscopy[36,64]. The speed of AI optical diagnosis was 35 seconds, and the algorithm worked equally well for both novice and expert gastroenterologists, without the need for injecting dyes. The findings of enhanced speed and accuracy were replicated in another independent study[37]. Such results are thematic: machine vision, at high magnification, can accurately and quickly interpret specific medical images as well as or better than humans.

Mental health. The enormous burden of mental health, such as the 350 million people around the world battling depression[74], is especially noteworthy, as there is potential here for AI to lend support to the affected patients and the vastly insufficient number of clinicians. Various tools that are in development include digital tracking of depression and mood via keyboard interaction, speech, voice, facial recognition, sensors, and use of interactive chatbots[75–80]. Facebook posts have been shown to predict the diagnosis of depression later documented in electronic medical records[81].

Machine learning has been explored for predicting successful antidepressant medication[82], characterizing depression[83–85], predicting suicide[83,86–88], and predicting bouts of psychosis in schizophrenics[89].

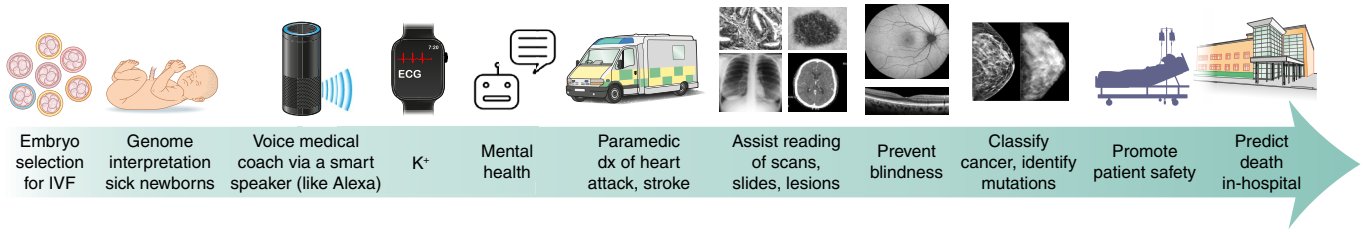

Fig. 2 | Examples of AI applications across the human lifespan. dx, diagnosis; IVF, in vitro fertilization K^+, potassium blood level. Credit: Debbie Maizels/Springer Nature

The use of AI algorithms has been described in many other clinical settings, such as facilitating stroke, autism or electroencephalographic diagnoses for neurologists[65,66], helping anesthesiologists avoid low oxygenation during surgery[67], diagnosis of stroke or heart attack for paramedics[68], finding suitable clinical trials for oncologists[69], selecting viable embryos for in vitro fertilization[70], help making the diagnosis of a congenital condition via facial recognition[71] and pre-empting surgery for patients with breast cancer[72]. Examples of the breadth of AI applications across human lifespan is shown in Fig. 2.There is considerable effort across many startups and established tech companies to develop natural language processing to replace the need for keyboards and human scribes for clinic visits[73]. The list of companies active in this space includes Microsoft, Google, Suki, Robin Healthcare, DeepScribe, Tenor.ai, Saykara, Sopris Health, Carevoice, Orbita, Notable, Sensely and Augmedix.

Artificial intelligence and health systems

Being able to predict key outcomes could, theoretically, make the use of hospital palliative care resources more efficient and precise. For example, if an algorithm could be used to estimate the risk of a patient's hospital readmission that would otherwise be undetectable given the usual clinical criteria for discharge, steps could be taken to avert discharge and attune resources to the underlying issues. For a critically ill patient, a very high likelihood of short-term survival might help this patient and their family and doctor make decisions regarding resuscitation, insertion of an endotracheal tube for mechanical ventilation, and other invasive measures. Similarly, it is possible that deciding which patients might benefit from palliative care and determining who is at risk of developing sepsis or septic shock could be ameliorated by AI predictive tools. Using electronic health record data, machine- and deep-learning algorithms have been able to predict many important clinical parameters, ranging from Alzheimer's disease to death (Table 3)[86,90–107]. For example, in a recent study, reinforcement learning was retrospectively carried out on two large datasets to recommend the use of vasopressors, intravenous fluids, and/or medications and the dose of the selected treatment for patients with sepsis; the treatment selected by the 'AI Clinician' was on average reliably more effective than that chosen by humans[108]. Both the size of the cohorts studied and the range of AUC accuracy reported have been quite heterogeneous, and all of these reports are retrospective and yet to be validated in the real-world clinical setting. Nevertheless, there are many companies that are already marketing such algorithms, such as Careskore, which is providing health systems with estimated of risk of readmission and mortality based on EHR data[109]. Beyond this issue, there are the differences between the prediction metric for a cohort and an individual prediction metric. If a model's AUC is 0.95, which most would qualify as very accurate, this reflects how good the model is for predicting an outcome, such as death, for the overall cohort. But most models are essentially classifiers and are not capable of precise prediction at the individual level, so there is still an important dimension of uncertainty.

In addition to data from electronic health records, imaging has been integrated to enhance predictive accuracy[98]. Multiple studies have attempted to predict biological age[110,111], and this has been shown to best be accomplished using DNA methylation–based biomarkers[112]. With respect to the accuracy of algorithms for prediction of biological age, the incompleteness of data input is noteworthy, since a large proportion of unstructured data—the free text in clinician notes that cannot be ingested from the medical record—has not been incorporated, and neither have many other modalities such as socioeconomic, behavioral, biologic '-omics', or physiologic sensor data. Further, concerns have been raised about the potential

Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

Prediction	*n*	AUC	Publication (Reference number)
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	0.93*0.75+0.85#	Rajkomar et al.[96]
All-cause 3–12 month mortality	221,284	0.93^	Avati et al.[91]
Readmission	1,068	0.78	Shameer et al.[106]
Sepsis	230,936	0.67	Horng et al.[102]
Septic shock	16,234	0.83	Henry et al.[103]
Severe sepsis	203,000	0.85@	Culliton et al.[104]
Clostridium difficile infection	256,732	0.82++	Oh et al.[93]
Developing diseases	704,587	range	Miotto et al.[97]
Diagnosis	18,590	0.96	Yang et al.[90]
Dementia	76,367	0.91	Cleret de Langavant et al.[92]
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al.[98]
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al.[95]
Disease onset for 133 conditions	298,000	range	Razavian et al.[105]
Suicide	5,543	0.84	Walsh et al.[86]
Delirium	18,223	0.68	Wong et al.[100]

LOS, length of stay; *n*, number of patients (training+ validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; ^, all patients; @, structured + unstructured data; ++, for University of Michigan site.

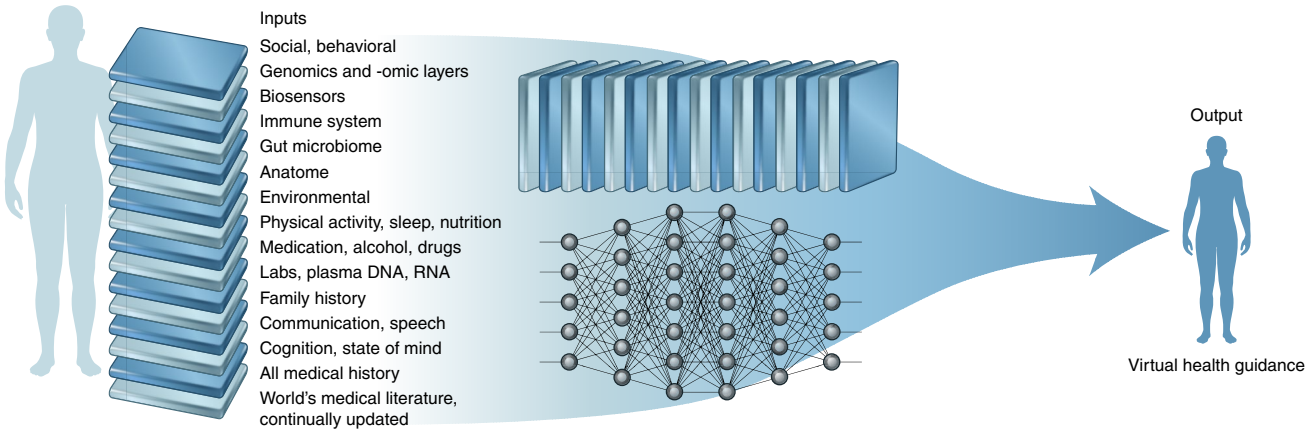

Fig. 3 | The virtual medical coach model with multi-modal data inputs and algorithms to provide individualized guidance. A virtual medical coach that uses comprehensive input from an individual that is deep learned to provide recommendations for preserving the person's health. Credit: Debbie Maizels/Springer Nature

to overfit data owing to small sample sizes in some instances. It has also been pointed out how essential it is to have κ-fold cross-validation of a model through successive, mutually exclusive validation datasets, which is missing from most of these publications. There is also considerable debate about using AUC as the key performance metric, since it ignores actual probability values and may be particularly misleading in regard to the sensitivity and specificity values that are of clinical interest[113].

In summary, it is not yet known how well AI can predict key outcomes in the healthcare setting, and this will not be determined until there is robust validation in prospective, real-world clinical environments, with rigorous statistical methodology and analysis.

Machine vision. Machine vision (also known as computer vision), which uses data from ambient sensors, is attracting considerable attention in health systems for promoting safety by monitoring such activities as proper clinician handwashing[114], critically ill patients in the intensive care unit[115], and risk of falling for patients[116]. Weaning patients in the intensive care unit from mechanical ventilation is often haphazard and inefficient; a reinforcement-learning algorithm using machine vision has shown considerable promise in this regard[117]. There are also ongoing efforts to digitize surgery that include machine vision observation of the team and equipment in the operating room and performance of the surgeon; real-time, high-resolution, AI-processed imaging of the relevant anatomy of a patient; and integration of all of a patient's preoperative data, including full medical history, labs, and scans[118,119]. Extremely delicate microsurgery, such as that inside the eye, has now been performed with AI assistance[120]. There is considerable promise in markedly reducing the radiation and time requirements for image acquisition and segmentation in preparation for radiotherapy via the use of deep-learning algorithms for image reconstruction[121] and of generative adversarial networks to improve the quality of medical scans. These improvements will, when widely implemented, promote safety, convenience, and lower cost[122–124].

Wearables. Of the more than $3.5 trillion per year (and rising) expenditures for healthcare in the United States, almost a third is related to hospitals. With FDA-approved wearable sensors that can continuously monitor all vital signs—including blood pressure, heart rate and rhythm, blood oxygen saturation, respiratory rate, and temperature—there is the potential to preempt a large number of patients being hospitalized in the future. There has not yet been algorithmic development and prospective testing for remote monitoring, but this deserves aggressive pursuit as it could reduce the costs of care without sacrificing convenience and comfort for a patient and family. The reduction of nosocomial infections alone would be an alluring path for promoting safety.

Increased efficiencies. It has been estimated that, per day, AI would process over 250 million images for the cost of about $1,000 (ref. [125]), representing a staggering hypothetical savings of billions of dollars. Besides the productivity and workflow gains that can be derived from AI-assisted image interpretation and clinician support, there is potential to reduce the workforce for many types of back-office, administrative jobs such as coding and billing, scheduling of operating rooms and clinic appointments, and staffing. At Geisinger Health in Pennsylvania, over 100,000 patients have undergone exome sequencing; the results are provided via an AI chatbot (Clear Genetics), which is well-received by most patients and reduces the need for genetic counselors. This demonstrates how a health system can leverage AI tools to provide complex information without having to rely on expansion of highly trained personnel.

Perhaps the greatest long-term potential of AI in health systems is the development of a massive data infrastructure to support nearest-neighbor analysis, another application of AI used to identify 'digital twins.' If each person's comprehensive biologic, anatomic, physiologic, environmental, socioeconomic, and behavioral data, including treatment and outcomes, were entered, an extraordinary learning system would be created. There have been great benefits derived from jet engine[126] digital twins that use an ultrahigh-fidelity model engine to simulate the flight conditions of a particular jet, but such a model has yet to be completed at any scale for patients, who theoretically could benefit from being informed of the best prevention methods, treatments, and outcomes for various conditions by their relevant twin's data[127].

Artifical intelligence and patients

The work for developing deep-learning algorithms to enable the public to take their healthcare into their own hands has lagged behind that for clinicians and health systems, but there are a few such algorithms that have been FDA-cleared or are in late-stage clinical development. In late 2017, a smartwatch algorithm was FDA-cleared to detect atrial fibrillation[128], and subsequently in 2018 Apple received FDA approval for their algorithm used with the Apple Watch Series 4 (refs. [129,130]). The photoplethysmography and accelerometer sensors on the watch learn the user's heart rate at rest and with physical activity, and when there is a significant deviation from expected, the user is given a haptic warning to record an ECG via the watch, which is then interpreted by

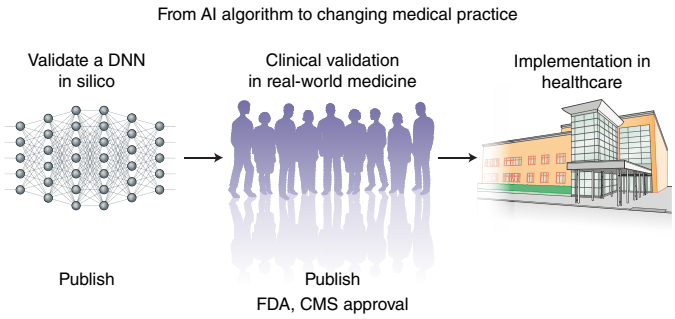

Fig. 4 | Call for due process of AI studies in medicine. The need to publish results in peer-reviewed journals with validation in real-world medicine must be addressed before implementation in patient care can take place. Credit: Debbie Maizels/Springer Nature

an algorithm. There are legitimate concerns that the widescale use of such an algorithm, particularly in the low-risk, young population who wear Apple watches, will lead to a substantial number of false-positive atrial fibrillation diagnoses and prompt unnecessary medical evalautions[131]. In contrast, the deep learning of the ECG pattern on the smartwatch, which can accurately detect whether there is high potassium in the blood, may provide particular usefulness for patients with kidney disease. This concept of a 'bloodless' blood potassium level (Fig. 2) reading via a smartwatch algorithm embodies the prospect of an algorithm able to provide information that was not previously obtainable or discernible without the technology.

Smartphone exams with AI are being pursued for a variety of medical diagnostic purposes, including skin lesions and rashes, ear infections, migraine headaches, and retinal diseases such as diabetic retinopathy and age-related macular degeneration. Some smartphone apps are using AI to monitor medical adherence, such as AiCure (NCT02243670), which has the patient take a selfie video as they swallow their prescribed pill. Other apps use image recognition of food for calorie and nutritional content[132]. In what may be seen as an outgrowth of dating apps that use AI nearest-neighbor analysis to find matches, there are now efforts to use the same methodology for matchmaking patients with primary care doctors to engender higher levels of trust[133].

One study has recently achieved the continuous sensing of blood-glucose (for 2 weeks) along with assessment of the gut microbiome, physical activity, sleep, medications, all food and beverage intake, and a variety of lab tests[134–136]. This multimodal data collection and analysis has led to the ability to predict the glycemic response to specific foods for an individual, a physiologic pattern that is remarkably heterogeneous among people and significantly driven by the gut microbiome. The use of continuous glucose sensors, which now are factory-calibrated, preempting the need for finger-stick glucose calibrations, has shown that post-prandial glucose spikes commonly occur, even in healthy people without diabetes[137,138]. It remains uncertain whether the glucose spikes indicate a higher risk of developing diabetes, but there are data suggesting this possibility[139] along with mechanistic links to gastrointestinal barrier dysfunction[140,141] in experimental models. Nevertheless, the use of AI with multimodal data to guide an individualized diet is a precedent for virtual medical coaching in the future. In the present, simple rules-based algorithms, based upon whether glucose values are rising or falling, are used for glucose management in people with diabetes. While these have helped avert hypoglycemic episodes[142], smart algorithms that incorporate an individual's comprehensive data are likely to be far more informative and helpful. In this manner, most common chronic conditions, such as hypertension, depression, and asthma, could theoretically be better managed with virtual coaching. With the remarkable progress in the accuracy of AI speech recognition and the accompanying soaring popularity of smart speakers, it is easy to envision that this would be performed via a voice platform, with or without an avatar. Eventually, when all of an individual's data and the corpus of medical literature can be incorporated, a holistic, prevention approach would be possible (Fig. 3).

Artificial intelligence and data analysis

While upstream from clinical practice, AI progress in life science has been notably faster, with extensive peer-reviewed publication, an easier path to validation without regulatory oversight, and far more willingness among the scientific community for implementation. As the stethoscope is the icon of doctors, the microscope is the icon of scientists. Using AI, Christiansen et al. [143] developed in silico labeling. Instead of the routine fluorescent staining of microscopic images, which can harm and kill cells and involves a complex preparation, this machine-learning algorithm predicts the fluorescent labels, ushering in 'image-free' microscopy[143–145]. Soon thereafter, Ota et al.[146] reported another image-free flow AI analytic method that they called 'ghost cytometry' to accurately identify rare cells, a capability that was replicated and extended by Nitta et al.[147] with image-activated AI cell sorting. This use of machine learning addresses the formidable problem of identifying and isolating rare cells by rapid, high-throughput, and accurate sorting on the basis of cell morphology that does not require the use of biomarkers. Besides promoting image-free microcopy and cytometry, deep-learning AI has been used to restore or fix out-of-focus images[148]. And computer vision has made possible high-throughput assessment of 40-plex proteins and organelles within a single cell[149,150].

Another challenge confronted by machine and deep learning has been in the analytics of genomic and other -omics biology datasets. Open-source algorithms have been developed for classifying or analyzing whole-genome sequence pathogenic variants[151–158], somatic cancer mutations[159], gene–gene interactions[160], RNA sequencing data[161], methylation[162], prediction of protein structure and protein–protein interactions[163], the microbiome[164], and single cells[165]. While these reports have generally represented a single -omics approach, there are now multi-omic algorithms being developed[166,167] that integrate the datasets. The use of genome editing has also been facilitated by algorithmic prediction of CRISPR guide RNA activity[168] and off-target activities[169].

Noteworthy is the use of AI tools to enhance understanding of how cancer evolves via application of a transfer-learning algorithm to multiregional tumor-sequencing data[170] and of machine vision for analysis of live cancer cells at single-cell resolution via microfluidic isolation[171]. Both of these novel approaches may ultimately be helpful in both risk stratification of patients and guiding therapy.

With the AI descriptor of neural networks, it is not surprising that there is bidirectional inspiration: biological neuroscience impacting AI and vice versa[172]. A couple of examples in *Drosophila* are noteworthy. Robie et al.[173] took videos of 400,00 flies and used machine learning and machine vision to map phenotype with gene expression and neuroanatomy. Whole-brain maps were generated for movement, female aggression, and many other traits. In another study, nearest-neighbor analysis was used to understand how odors are sensed by the flies, that is, their smell algorithm[174].

AI has been used to reconstruct neural circuits, allowing an understanding of connectomics, from electron microscopy[175]. One of the most impressive advances facilitated by AI has been in understanding the human brain's grid cells—which enable perception of the speed and direction of movement of the body, i.e., its place in space[176,177]. Reciprocally, neuromorphic computing, or reverse-engineering of the brain to make computer chips, is not only leading to more efficient computing, but also helping

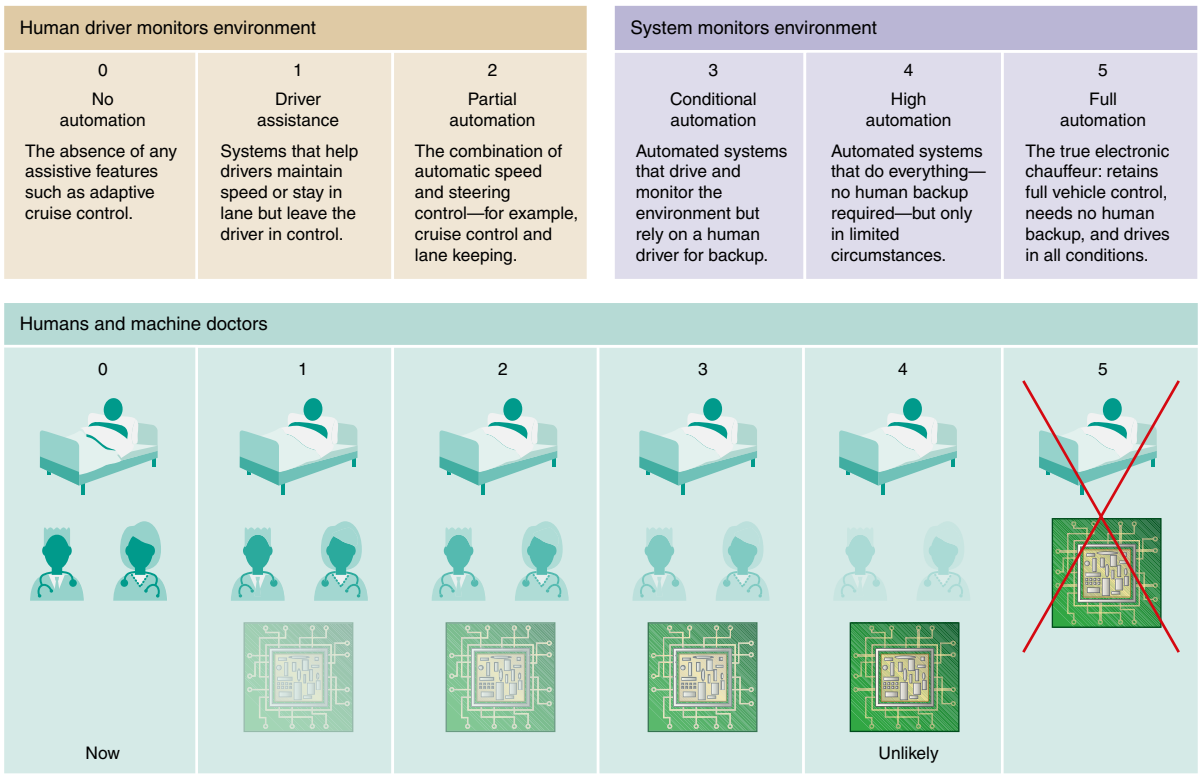

Fig. 5 | The analogy between self-driving cars and medicine. Level 5, full automation with no potential for human backup of clinicians, is not the objective. Nor is Level 4, with human backup in very limited conditions. The goal is for synergy, offsetting functions that machines do best combined with those that are best suited for clinicians. Credit: Debbie Maizels/Springer Nature

researchers understand brain circuitry and build brain–machine interfaces[172,178,179]. Machine vision tracking of human and animal behavior with a transfer-learning algorithm is yet another example of the progress being made[180].

Drug discovery is being revamped with the use of AI at many levels, including sophisticated natural language processing searches of the biomedical literature, data mining of millions of molecular structures, designing and making new molecules, predicting off-target effects and toxicity, predicting the right dose for experimental drugs, and developing cellular assays at a massive scale[181–184]. There is new hope that preclinical animal testing can be reduced via machine-learning prediction of toxicity[185]. AI cryptography has been used to combine large proprietary pharmaceutical company datasets and discover previously unidentified drug interactions[186]. The story of the University of Cambridge and Manchester's robot 'Eve' and how it autonomously discovered an antimalarial drug that is a constituent of toothpaste has galvanized interest in using AI to accelerate the process, with a long list of start-ups and partnerships with major pharmaceutical firms[181,187,188].

Limitations and challenges

Despite all the promises of AI technology, there are formidable obstacles and pitfalls. The state of AI hype has far exceeded the state of AI science, especially when it pertains to validation and readiness for implementation in patient care. A recent example is IBM Watson Health's cancer AI algorithm (known as Watson for Oncology). Used by hundreds of hospitals around the world for recommending treatments for patients with cancer, the algorithm was based on a small number of synthetic, nonreal cases with very limited input (real data) of oncologists[189]. Many of the actual output recommendations for treatment were shown to be erroneous, such as suggesting the use of bevacizumab in a patient with severe bleeding, which represents an explicit contraindication and 'black box' warning for the drug[189]. This example also highlights the potential for major harm to patients, and thus for medical malpractice, by a flawed algorithm. Instead of a single doctor's mistake hurting a patient, the potential for a machine algorithm inducing iatrogenic risk is vast. This is all the more reason that systematic debugging, audit, extensive simulation, and validation, along with prospective scrutiny, are required when an AI algorithm is unleashed in clinical practice. It also underscores the need to require more evidence and robust validation to exceed the recent downgrading of FDA regulatory requirements for medical algorithm approval[190].

There has been much written about the black box of algorithms, and much controversy surrounding this topic[191–193]; especially in the case of DNNs, it may not be possible to understand the determination of output. This opaqueness has led to both demands for explainability, such as the European Union's General Data Protection Regulation requirement for transparency—deconvolution of an algorithm's black box—before an algorithm can be used for patient care[194]. While this debate of whether it is acceptable to use nontransparent algorithms for patient care is unsettled, it is notable that many aspects of the practice of medicine are unexplained, such as prescription of a drug without a known mechanism of action.

Inequities are one of the most important problems in healthcare today, especially in the United States, which does not provide care for all of its citizens. With the knowledge that low socioeconomic status is a major risk factor for premature mortality[195], the disproportionate use of AI in the 'haves,' as opposed to the 'have-nots,' could widen the present gap in health outcomes. Intertwined with this concern of exacerbating pre-existing inequities is embedded bias present in many algorithms due to lack of inclusion of minorities in datasets. Examples are the algorithms in dermatology that diagnose melanoma but lack inclusion of skin color[47] and the use

of the corpus of genomic data, which so far has seriously underrepresented minorities[196]. While there are arguments that algorithm bias is exceeded by human bias[197], much work is needed to eradicate embedded prejudice and strive for medical research that provides a true representative cross-section of the population.

An overriding issue for the future of AI in medicine rests with how well privacy and security of data can be assured. Given the pervasive problems of hacking and data breaches, there will be little interest in use of algorithms that risk revealing the details of patient medical history[198]. Moreover, there is the risk of deliberate hacking of an algorithm to harm people at a large scale, such as overdosing insulin in diabetics or stimulating defibrillators to fire inside the chests of patients with heart disease. It is increasingly possible for an individual's identity to be determined by facial recognition or genomic sequence from massive databases, which further impedes protection of privacy. At the same time, the blurring of truth made possible by generative adversarial networks, with seemingly unlimited capacity to manipulate content, could be highly detrimental for health[198,199]. New models of health data ownership with rights to the individual, use of highly secure data platforms, and governmental legislation, as has been achieved in Estonia, are needed to counter the looming security issues that will otherwise hold up or ruin the chances for progress in AI for medicine[200–202].

Future considerations

A key point that I have emphasized throughout this Review it that the narrative of bringing AI to medicine is just beginning. There has been remarkably little prospective validation for tasks that machines could perform to help clinicians or predict clinical outcomes that would be useful for health systems, and even less for patient-centered algorithms. The field is certainly high on promise and relatively low on data and proof. The risk of faulty algorithms is exponentially higher than that of a single doctor–patient interaction, yet the reward for reducing errors, inefficiencies, and cost is substantial. Accordingly, there cannot be exceptionalism for AI in medicine—it requires rigorous studies, publication of the results in peer-reviewed journals, and clinical validation in a real-world environment, before roll-out and implementation in patient care (Fig. 4). With these caveats, it is also important to have reasonable expectations for how AI will ultimately be incorporated. Piercing through today's widespread hype that doctors will be replaced by machines is the analogy of the self-driving car model for reality testing. Most would agree that autonomous cars represent the pinnacle technical achievement of AI to date, but the term autonomous is misleading. The Society of Automotive Engineers (SAE) has defined five levels of autonomy, with Level 5 indicating full control by the car under all conditions, without any possibility for human backup or taking control of the vehicle (Fig. 5). It is now accepted that this definition of full autonomy is likely to never be attained, as certain ambient or road conditions will prohibit the safe use of such vehicles[203]. By the same token, medicine will unlikely ever surpass Level 3, a conditional automation, for which humans will indeed be required for oversight of algorithmic interpretation of images and data. It is hard to imagine very limited human backup across the board of caring for patients (Level 4). Human health is too precious—relegating it to machines, except for routine matters with minimal risk, seems especially far-fetched.

The excitement that lies ahead, albeit much further along than many have forecasted, is for software that will ingest and meaningfully process massive sets of data quickly, accurately, and inexpensively and for machines that will see and do things that are not humanly possible. This capability will ultimately lay the foundation for high-performance medicine, which is truly data-driven, decompressing our reliance on human resources, and will eventually take us well beyond the sum of the parts of human and machine intelligence. This symbiosis will be preceded by the upstream advances that are already being made in biomedical science and discovery, which have a far less tortuous path to be accepted and widely implemented.

Received: 16 August 2018; Accepted: 12 November 2018;
Published online: 7 January 2019

References

1. Thakrar, A. P. et al. Child mortality in the US and 19 OECD comparator nations: a 50-year time-trend analysis. *Health Aff. (Millwood)* **37**, 140–149 (2018).
2. Roser, M. Link between health spending and life expectancy: US is an outlier. In *Our World in Data* https://ourworldindata.org/the-link-between-life-expectancy-and-health-spending-us-focus (2017).
3. Singh, H. et al. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ Qual. Saf.* **23**, 727–731 (2014).
4. Berwick, D. M. & Hackbarth, A. D. Eliminating waste in US health care. *JAMA* **307**, 1513–1516 (2012).
5. 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).
6. Li, Z. et al. Thoracic disease identification and localization with limited supervision. Preprint at https://arxiv.org/abs/1711.06373 (2017).
7. Singh, R. et al. Deep learning in chest radiography: detection of findings and presence of change. *PLoS ONE* **13**, e0204155 (2018).
8. Nam, J. G. et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* https://doi.org/10.1148/radiol.2018180237 (2018).
9. Lindsey, R., et al. Deep neural network improves fracture detection by clinicians. *Proc. Natl. Acad. Sci. USA* **115**, 11591–11596 (2018).
10. Gale, W. et al. Detecting hip fractures with radiologist-level performance using deep neural networks. Preprint at https://arxiv.org/abs/1711.06504 (2017).
11. Rajpurkar, P. MURA dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. Preprint at https://arxiv.org/abs/1712.06957 (2017).
12. Ridley, E. L. Deep learning shows promise for bone age assessment. In *Aunt Minnie* https://www.auntminnie.com/index.aspx?sec=log&itemID=119011 (2017).
13. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).
14. Bar, A. et al. Compression fractures detection on CT. Preprint at https://arxiv.org/abs/1706.01671 (2017).
15. Ridley, E. L. Deep-learning algorithm can stratify lung nodule risk. In *Aunt Minnie* https://www.auntminnie.com/index.aspx?sec=rca&sub=rsna_2017&pag=dis&ItemID=119166 (2017).
16. Yasaka, K. et al. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* **286**, 887–896 (2018).
17. Liu, F. et al. Joint shape representation and classification for detecting PDAC in abdominal CT scans. Preprint at https://arxiv.org/abs/1804.10684 (2018).
18. Shadmi, R. et al. Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, 2018).
19. Arbabshirani, M. R. et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ Digit. Med.* **1**, 9 (2018).
20. Chilamkurthy, S. et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* **392**, 2388–2396 (2018).
21. Chilamkurthy, S. et al. Development and validation of deep learning algorithms for detection of critical findings in head CT ccans. Preprint at https://arxiv.org/abs/1803.05854 (2018).
22. Lieman-Sifry, J. et al. FastVentricle: cardiac segmentation with ENet. Preprint at https://arxiv.org/abs/1704.04296 (2017).
23. Madani, A.. et al. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**, 6 (2018).
24. Zhang, J. et al. Fully automated echocardiogram interpretation in clinical practice feasibility and diagnostic accuracy. *Circulation* **138**, 1623–1635 (2018).
25. Yee, K. M. AI algorithm matches radiologists in breast screening exams. In *Aunt Minnie* https://www.auntminnie.com/index.aspx?sec=log&itemID=119385 (2017).

26. Lehman, C. D. et al. Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* http://doi.org/10.1148/radiol.2018180694 (2018).
27. Titano, J. J. et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat. Med.* **24**, 1337–1341 (2018).
28. Saito, T. & Rehmsmeier, M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
29. Lobo, J. et al. AUC: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**, 145–151 (2007).
30. Keane, P. & Topol, E. With an eye to AI and autonomous diagnosis. *NPJ Digit. Med.* **1**, 40 (2018).
31. Abramoff, M. et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit. Med.* **1**, 39 (2018).
32. Kanagasingam, Y. et al. Evaluation of artificial intelligence–based grading of diabetic retinopathy in primary care. *JAMA Netw. Open* **1**, e182665 (2018).
33. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
34. Liu, Y. et al. Artificial intelligence–based breast cancer nodal metastasis detection. *Arch. Pathol. Lab. Med.* https://doi.org/10.5858/arpa.2018-0147-OA (2018).
35. Steiner, D. F., et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* **42**, 1636–1646 (2018).
36. Mori, Y. et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy. *Ann. Intern. Med.* **169**, 357–366 (2018).
37. Wang, P. et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat. Biomed. Eng.* **2**, 741–748 (2018).
38. Long, E. et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat. Biomed. Eng.* **1**, 1–8 (2017).
39. Acs, B. & Rimm, D. L. Not just digital pathology, intelligent digital pathology. *JAMA Oncol.* **4**, 403–404 (2018).
40. Yu, K. H. et al. Predicting non–small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**, 12474 (2016).
41. Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
42. Golden, J. A. Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *JAMA* **318**, 2184–2186 (2017).
43. Cruz-Roa, A. et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. *Sci. Rep.* **7**, 46450 (2017).
44. Wong, D. & Yip, S. Machine learning classifies cancer. *Nature* **555**, 446–447 (2018).
45. Capper, D. et al. DNA methylation–based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
46. Yang, S. J. et al. Assessing microscope image focus quality with deep learning. *BMC Bioinformatics* **19**, 77 (2018).
47. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
48. 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).
49. Han, S. S. et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Invest. Dermatol.* **138**, 1529–1538 (2018).
50. Wong, T. Y. & Bressler, N. M. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. *JAMA* **316**, 2366–2367 (2016).
51. 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).
52. Burlina, P. M. et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol.* **135**, 1170–1176 (2017).
53. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e1129 (2018).
54. Ting, D. S. W. et al. AI for medical imaging goes deep. *Nat. Med.* **24**, 539–540 (2018).
55. Rampasek, L. & Goldenberg, A. Learning from everyday images enables expert-like diagnosis of retinal diseases. *Cell* **172**, 893–895 (2018).
56. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
57. Mutlu, U. et al. Association of retinal neurodegeneration on optical coherence tomography with dementia: a population-based study. *JAMA Neurol.* **75**, 1256–1263 (2018).
58. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
59. All eyes are on AI. *Nat. Biomed. Eng.* **2**, 139 (2018).
60. Brown, J. M. et al. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol.* **136**, 803–810 (2018).
61. Willems, J. et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N. Engl. J. Med.* **325**, 1767–1773 (1991).
62. Strodthoff, N. & Strodthoff, C. Detecting and interpreting myocardial infarctions using fully convolutional neural networks. Preprint at https://arxiv.org/abs/1806.07385 (2018).
63. Rajpurkar, P. et al. Cardiologist-level arrhythmia detection with convolutional neural networks. Preprint at https://arxiv.org/abs/1707.01836 (2017).
64. Holme, Ø. & Aabakken, L. Making colonoscopy smarter with standardized computer-aided diagnosis *Ann. Intern. Med.* **169**, 409–410 (2018).
65. Petrone, J. FDA approves stroke-detecting AI software. *Nat. Biotechnol.* **36**, 290 (2018).
66. Hsu, J. & Spectrum. AI could make detecting autism easier. In *The Atlantic* https://www.theatlantic.com/technology/archive/2018/07/ai-autism-diagnosis-screening-bottleneck/564890/ (2018).
67. Lundberg, S. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
68. Peters, A. Having a heart attack? This AI helps emergency dispatchers find out. In *Fast Company* https://www.fastcompany.com/40515740/having-a-heart-attack-this-ai-helps-emergency-dispatchers-find-out (2018).
69. Patel, N. M. et al. Enhancing next-generation sequencing-guided cancer care through cognitive computing. *Oncologist* **23**, 179–185 (2018).
70. De Graaf, M. Will Al replace fertility doctors? Why computers are the only ones that can end the agony of failed IVF cycles, miscarriages, and risky multiple birth. In *Daily Mail* https://www.dailymail.co.uk/health/article-6257891/Study-finds-artificial-intelligence-better-doctor-crucial-stage-IVF.html (2018).
71. Gurovich, Y. et al. DeepGestalt—identifying rare genetic syndromes using deep learning. Preprint at https://arxiv.org/abs/1801.07637 (2017).
72. Bahl, M. et al. High-risk breast lesions: a machine learning model to predict pathologic upgrade and reduce unnecessary surgical excision. *Radiology* **286**, 810–818 (2018).
73. Coiera, E. et al. The digital scribe. *NPJ Digit. Med.* **1**, 58 (2018).
74. The burden of depression. *Nature* **515**, 163 (2014).
75. Cao, B. et al. DeepMood: modeling mobile phone typing dynamics for mood detection. Preprint at https://arxiv.org/abs/1803.08986 (2018).
76. Mohr, D. C. et al. A solution-focused research approach to achieve an implementable revolution in digital mental health. *JAMA Psychiatry* **75**, 113–114 (2018).
77. Frankel, J. How artificial intelligence could help diagnose mental disorders. In *The Atlantic* https://www.theatlantic.com/health/archive/2016/08/could-artificial-intelligence-improve-psychiatry/496964/ (2016).
78. Barrett, P. M. et al. Digitising the mind. *Lancet* **389**, 1877 (2017).
79. Firth, J. et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry* **16**, 287–298 (2017).
80. Fitzpatrick, K. K. et al. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment. Health* **4**, e19 (2017).
81. Eichstaedt, J. C. et al. Facebook language predicts depression in medical records. *Proc. Natl. Acad. Sci. USA* **115**, 11203–11208 (2018).
82. Chekroud, A. M. et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* **3**, 243–250 (2016).
83. Schnyer, D. M. et al. Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder. *Psychiatry Res.* **264**, 1–9 (2017).
84. Reece, A. G. & Danforth, C. M. Instagram photos reveal predictive markers of depression. *EPJ Data Science* **6**, 15 (2017).
85. Wager, T. D. & Woo, C. W. Imaging biomarkers and biotypes for depression. *Nat. Med.* **23**, 16–17 (2017).
86. Walsh, C. G. et al. Predicting risk of suicide attempts over time through machine learning. *Clin. Psychol. Sci.* **5**, 457–469 (2017).
87. Franklin, J. C. et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol. Bull.* **143**, 187–232 (2017).

88. Just, M. A. et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat. Hum. Behav.* **1**, 911–919 (2017).
89. Chung, Y. et al. Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. *JAMA Psychiatry* **75**, 960–968 (2018).
90. Yang, Z. et al. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Sci. Rep.* **8**, 6329 (2018).
91. Avati, A. et al. Improving palliative care with deep learning. Preprint at https://arxiv.org/abs/1711.06402 (2017).
92. Cleret de Langavant, L. et al. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study. *J. Med. Internet. Res.* **20**, e10493 (2018).
93. Oh, J. et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect. Control. Hosp. Epidemiol.* **39**, 425–433 (2018).
94. Bennington-Castro, J. AI can predict when we'll die—here's why that's a good thing. In *NBC News* https://www.nbcnews.com/mach/science/ai-can-predict-when-we-ll-die-here-s-why-ncna844276 (2018).
95. Elfiky, A. et al. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Netw. Open* **1**, e180926 (2018).
96. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
97. Miotto, R. et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).
98. Mathotaarachchi, S. et al. Identifying incipient dementia individuals using machine learning and amyloid imaging. *Neurobiol. Aging.* **59**, 80–90 (2017).
99. Yoon, J. et al. Personalized survival predictions via Trees of Predictors: an application to cardiac transplantation. *PLoS ONE* **13**, e0194985 (2018).
100. Wong, A. et al. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Netw. Open* **1**, e181018 (2018).
101. Alaa, A. M. & van der Schaar, M. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Sci. Rep.* **8**, 11242 (2018).
102. Horng, S. et al. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS ONE* **12**, e0174708 (2017).
103. Henry, K. E. et al. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, 299ra122 (2015).
104. Culliton, P. et al. Predicting severe sepsis using text from the electronic health record. Preprint at https://arxiv.org/abs/1711.11536 (2017).
105. Razavian, N. et al. Multi-task prediction of disease onsets from longitudinal lab tests. *PMLR* **56**, 73–100 (2016).
106. Shameer, K. et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai Heart Failure Cohort. *Pac. Symp. Biocomput.* **22**, 276–287 (2017).
107. Bhagwat, N. et al. Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput. Biol.* **14**, e1006376 (2018).
108. Komorowski, M. et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* **24**, 1716–1720 (2018).
109. Zaidi, D. AI is transforming medical diagnosis, prosthetics, and vision aids. In *Venture Beat* https://venturebeat.com/2017/10/30/ai-is-transforming-medical-diagnosis-prosthetics-and-vision-aids/ (2017).
110. Putin, E. et al. Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging* **8**, 1021–1033 (2016).
111. Wang, Z. et al. Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *J. Biomed. Inform.* **76**, 59–68 (2017).
112. Horvath, S. & Raj, K. DNA methylation–based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
113. Rose, S. Machine Learning for Prediction in Electronic Health Data. *JAMA Netw. Open* **1**, e181404(2018).
114. Haque, A. et al. Towards vision-based smart hospitals: a system for tracking and monitoring hand hygiene compliance. Preprint at https://arxiv.org/abs/1708.00163 (2017).
115. Suresh, H. et al. Clinical intervention prediction and understanding with deep neural networks. Preprint at https://arxiv.org/abs/1705.08498 (2017).
116. Kwolek, B. & Kepski, M. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Comput. Methods Programs Biomed.* **117**, 489–501 (2014).
117. Prasad, N. et al. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. Preprint at https://arxiv.org/abs/1704.06300 (2018).
118. Maier-Hein, L. et al. Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* **1**, 691–696 (2017).
119. Hung, A. J. et al. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg.* **153**, 770–771 (2018).
120. Gehlbach, P. L. Robotic surgery for the eye. *Nat. Biomed. Eng.* **2**, 627–628 (2018).
121. Nikolov, S. et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. Preprint at https://arxiv.org/abs/1809.04430 (2018).
122. Zhu, B. et al. Image reconstruction by domain-transform manifold learning. *Nature* **555**, 487–492 (2018).
123. Harvey, H. Can AI enable a 10 minute MRI? In *Towards Data Science* https://towardsdatascience.com/can-ai-enable-a-10-minute-mri-77218f0121fe (2018).
124. Ridley, E. L. Artificial intelligence guides lower PET tracer dose. In *Aunt Minnie* https://www.auntminnie.com/index.aspx?sec=log&itemID=119572 (2018).
125. Beam, A. L. & Kohane, I. S. Translating artificial intelligence into clinical care. *JAMA* **316**, 2368–2369 (2016).
126. Tuegel, E. J. et al. Reengineering aircraft structural life prediction using a digital twin. *Int. J. Aerosp.* **2011**, 154798 (2011).
127. Tarassenko, L. & Topol, E. Monitoring the health of jet engines and people. *JAMA* https://doi.org/10.1001/jama.2018.16558 (2018).
128. Buhr, S. FDA clears AliveCor's Kardiaband as the first medical device accessory for the Apple Watch. In *TechCrunch* https://techcrunch.com/2017/11/30/fda-clears-alivecors-kardiaband-as-the-first-medical-device-accessory-for-the-apple-watch/ (2017).
129. Victory, J. What did journalists overlook about the Apple Watch 'heart monitor' feature? In *HealthNewsReview* https://www.healthnewsreview.org/2018/09/what-did-journalists-overlook-about-the-apple-watch-heart-monitor-feature/ (2018).
130. Fingas, R. Apple Watch Series 4 EKG tech got FDA clearance less than 24 hours before reveal. In *AppleInsider* https://appleinsider.com/articles/18/09/18/apple-watch-series-4-ekg-tech-got-fda-clearance-less-than-24-hours-before-reveal (2018).
131. Carroll, A. E. That new apple watch EKG feature? There are more downs than ups. In *The New York Times* https://www.nytimes.com/2018/10/08/upshot/apple-watch-heart-monitor-ekg.html (2018).
132. Levine, B. & Brown, A. Onduo delivers diabetes clinic and coaching to your smartphone. In *Diatribe* https://diatribe.org/onduo-delivers-diabetes-clinic-and-coaching-your-smartphone (2018).
133. Han, Q. et al. A hybrid recommender system for patient–doctor matchmaking in primary care. Preprint at https://arxiv.org/abs/1808.03265 (2018).
134. Zmora, N. et al. Taking it personally: personalized utilization of the human microbiome in health and disease. *Cell. Host. Microbe.* **19**, 12–20 (2016).
135. Korem, T. et al. Bread affects clinical parameters and induces gut microbiome–associated personal glycemic responses. *Cell. Metab.* **25**, 1243–1253 e1245 (2017).
136. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
137. Hall, H. et al. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biol.* **16**, e2005143 (2018).
138. Albers, D. J. et al. Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS. Comput. Biol.* **13**, e1005232 (2017).
139. Hulman, A. et al. Glucose patterns during an oral glucose tolerance test and associations with future diabetes, cardiovascular disease and all-cause mortality rate. *Diabetologia* **61**, 101–107 (2018).
140. Thaiss, C. A. et al. Hyperglycemia drives intestinal barrier dysfunction and risk for enteric infection. *Science* **359**, 1376–1383 (2018).
141. Wu, D. et al. Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer. *Nature* **559**, 637–641 (2018).
142. Bally, L. et al. Closed-loop insulin delivery for glycemic control in noncritical care. *N. Engl. J. Med.* **379**, 547–556 (2018).
143. Christiansen, E. M. et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell* **173**, 792–803 e719 (2018).
144. Sullivan, D. P. & Lundberg, E. Seeing more: a future of augmented microscopy. *Cell* **173**, 546–548 (2018).
145. Ounkomol, C. et al. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).
146. Ota, S. et al. Ghost cytometry. *Science* **360**, 1246–1251 (2018).
147. Nitta, N. et al. Intelligent image-activated cell sorting. *Cell* **175**, 266–276 e213 (2018).

148. Weigert, M. et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. Preprint at https://doi.org/10.1101/236463 (2017).
149. Gut, G. et al. Multiplexed protein maps link subcellular organization to cellular states. *Science* **361**, eaar7042 (2018).
150. Sullivan, D. P. et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018).
151. Poplin, R. et al. Creating a universal SNP and small indel variant caller with deep neural networks. Preprint at https://doi.org/10.1101/092890 (2016).
152. Sundaram, L. et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
153. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
154. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
155. Luo, R., et al. Clairvoyante: a multi-task convolutional deep neural network for variant calling in single molecule sequencing. Preprint at https://doi.org/10.1101/310458 (2018).
156. Leung, M. et al. Machine learning in genomic medicine: a review of computational problems and data sets. In *Proceedings of the IEEE* Vol. 104, 176–197 (IEEE, 2016).
157. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
158. Riesselman, A. et al. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
159. Wood, D. E. et al. A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.* **10**, eaar7939 (2018).
160. Behravan, H. et al. Machine learning identifies interacting genetic variants contributing to breast cancer risk: a case study in Finnish cases and controls. *Sci. Rep.* **8**, 13149 (2018).
161. Lin, C. et al. Using neural networks for reducing the dimensions of single-cell RNA-seq data. *Nucleic Acids Res.* **45**, e156 (2017).
162. Angermueller, C. et al. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome. Biol.* **18**, 67 (2017).
163. AlQuraishi, M. End-to-end differentiable learning of protein structure. Preprint at https://doi.org/10.1101/265231 (2018).
164. Espinoza, J. L. Machine learning for tackling microbiota data and infection complications in immunocompromised patients with cancer. *J. Intern. Med.* https://doi.org/10.1111/joim.12746 (2018).
165. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e727 (2018).
166. Zitnik, M. et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Preprint at https://doi.org/10.1111/joim.12746 (2018).
167. Camacho, D. M. et al. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
168. Kim, H. K. et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239–241 (2018).
169. Listgarten, J. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.* **2**, 38–47 (2018).
170. Caravagna, G. et al. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat. Methods* **15**, 707–714 (2018).
171. Manak, M. et al. Live-cell phenotypic-biomarker microfluidic assay for the risk stratification of cancer patients via machine learning. *Nature Biomed. Eng.* **2**, 761–772 (2018).
172. Hassabis, D. et al. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
173. Robie, A. A. et al. Mapping the neural substrates of behavior. *Cell* **170**, 393–406 e328 (2017).
174. Dasgupta, S. et al. A neural algorithm for a fundamental computing problem. *Science* **358**, 793–796 (2017).
175. Januszewski, M. et al. High-precision automated reconstruction of neurons with flood-filling networks. *Nat. Methods* **15**, 605–610 (2018).
176. Savelli, F. & Knierim, J. J. AI mimics brain codes for navigation. *Nature* **557**, 313–314 (2018).
177. Banino, A. et al. Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
178. Adam, G. C. Two artificial synapses are better than one. *Nature* **558**, 39–40 (2018).
179. Wright, C. D. Phase-change devices: crystal-clear neuronal computing. *Nat. Nanotechnol.* **11**, 655–656 (2016).
180. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
181. Smalley, E. AI-powered drug discovery captures pharma interest. *Nat. Biotechnol.* **35**, 604–605 (2017).
182. Schneider, G. Automating drug discovery. *Nat. Rev. Drug. Discov.* **17**, 97–113 (2018).
183. Chakradhar, S. Predictable response: finding optimal drugs and doses using artificial intelligence. *Nat. Med.* **23**, 1244–1247 (2017).
184. Lowe, D. AI designs organic syntheses. *Nature* **555**, 592–593 (2018).
185. Luechtefeld, T. et al. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol. Sci.* **165**, 198–212 (2018).
186. Hie, B. et al. Realizing private and practical pharmacological collaboration. *Science* **362**, 347–350 (2018).
187. Bilsland, E. et al. Plasmodium dihydrofolate reductase is a second enzyme target for the antimalarial action of triclosan. *Sci. Rep.* **8**, 1038 (2018).
188. Artificially-intelligent robot scientist 'Eve' could boost search for new drugs. In *University of Cambridge Research* https://www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs (2015).
189. Ross, C. & Swetlitz, I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. In *Stat News* https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (2018).
190. Miliard, M. As FDA signals wider AI approval, hospitals have a role to play. In *Healthcare IT News* https://www.healthcareitnews.com/news/fda-signals-wider-ai-approval-hospitals-have-role-play (2018).
191. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
192. Knight, W. The dark secret at the heart of AI. In *MIT Technology Review* https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (2017).
193. Weinberger, D. Our machines now have knowledge we'll never understand. In *Backchannel* https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/ (2017).
194. Kuang, C. Can A.I. be taught to explain itself? In *The New York Times* https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html (2017).
195. Stringhini, S. et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women. *Lancet* **389**, 1229–1237 (2017).
196. Wapner, J. Cancer scientists have ignored African DNA in the search for cures. In *Newsweek* https://www.newsweek.com/2018/07/27/cancer-cure-genome-cancer-treatment-africa-genetic-charles-rotimi-dna-human-1024630.html (2018).
197. Miller, A. P. Want less-biased decisions? Use algorithms. In *Harvard Business Review* https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms (2018).
198. Brundage, M. et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. Preprint at https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf (2018).
199. Finlayson, S. et al. Adversarial attacks against medical deep learning systems. Preprint at https://arxiv.org/abs/1804.05296 (2018).
200. Haun, K. & Topol, E. The health data conundrum. In *The New York Times* https://www.nytimes.com/2017/01/02/opinion/the-health-data-conundrum.html (2017).
201. Kish, L. J. & Topol, E. J. Unpatients-why patients should own their medical data. *Nat. Biotechnol.* **33**, 921–924 (2015).
202. Heller, N. Estonia, the digital republic. In *The New Yorker* https://www.newyorker.com/magazine/2017/12/18/estonia-the-digital-republic (2017).
203. Shladover, S. The truth about "self-driving" cars. In *Scientific American* **314**, 53–57 (2016).
204. Turing, A. M. On computable numbers with an application to the Entscheidungsproblem. *P. Lond. Match. Soc.* **s2-42**, 230–265 (1936).
205. Turing, A. M. Computing machinery and intelligence. *Mind* **59**, 433–460 (1950).
206. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
207. Krizhevsky, A. et al. ImageNet classification with deep convolutional neural networks. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems* 1097–1105 (NIPS, 2012).
208. Hu, J. et al. Squeeze-and-excitation networks. Preprint at https://arxiv.org/abs/1709.01507 (2017).
209. Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. Preprint at https://arxiv.org/abs/1409.0575 (2014).
210. Goodfellow, I et al. *Deep Learning* (MIT Press, Cambridge, MA, USA, 2016).
211. Yu, K.-H. et al. Artificial intelligence in healthcare. *Nature Biomed. Eng.* **2**, 719–731 (2018).
212. Korkinof, D. et al. High-resolution mammogram synthesis using progressive generative adversarial networks. Preprint at https://arxiv.org/abs/1807.03401 (2018).
213. Baur, C. et al. Generating highly realistic images of skin lesions with GANs. Preprint at https://arxiv.org/abs/1809.01410 (2018).

214. Kazeminia, S. et al. GANs for medical image analysis. Preprint at https://arxiv.org/abs/1809.06222 (2018).
215. Harvey, H. FAKE VIEWS! Synthetic medical images for machine learning. In *Towards Data Science* https://towardsdatascience.com/harnessing-infinitely-creative-machine-imagination-6801a9fb4ca9 (2018).
216. Madani, A. et al. Deep echocardiography: data-efficient supervised and semisupervised deep learning towards automated diagnosis of cardiac disease. *NPJ Digit. Med.* **1**, 59 (2018).

Acknowledgements

Funding was provided by the Clinical and Translational Science Award (CTSA) from the National Institute of Health (NIH) grant number UL1TR002550.

Competing interests

E.T. is on the scientific advisory board of Verily, Tempus Labs, Myokardia and Voxel Cloud and the board of directors of Dexcoman and is an advisor to Guardant Health, Blue Cross Blue Shield Association, and Walgreens.

Additional information

Correspondence should be addressed to E.J.T.