Artificial intelligence (AI) for neurologists: do digital neurones dream of electric sheep?

Joshua Au Yeung,[1,2] Yang Yang Wang,[3] Zeljko Kraljevic,[4] James T H Teo [1,2,4]

[1]CogStack team, Guy's and St Thomas' NHS Foundation Trust, London, UK
[2]CogStack team, King's College Hospital NHS Foundation Trust, London, UK
[3]Medicine, Guy's and St Thomas' Hospitals NHS Trust, London, UK
[4]Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

Correspondence to
Professor James T H Teo, Neurology, King's College Hospital NHS Foundation Trust, London, London, UK; jamesteo@nhs.net

Accepted 29 August 2023
Published Online First 17 November 2023

To cite: Au Yeung J, Wang YY, Kraljevic Z, *et al*. *Pract Neurol* 2023;**23**:476–488.

ABSTRACT

Artificial intelligence (AI) is routinely mentioned in journals and newspapers, and non-technical outsiders may have difficulty in distinguishing hyperbole from reality. We present a practical guide to help non-technical neurologists to understand healthcare AI. AI is being used to support clinical decisions in treating neurological disorders. We introduce basic concepts of AI, such as machine learning and natural language processing, and explain how AI is being used in healthcare, giving examples its benefits and challenges. We also cover how AI performance is measured, and its regulatory aspects in healthcare. An important theme is that AI is a general-purpose technology like medical statistics, with broad utility applicable in various scenarios, such that niche approaches are outpaced by approaches that are broadly applicable in many disease areas and specialties. By understanding AI basics and its potential applications, neurologists can make informed decisions when evaluating AI used in their clinical practice. This article was written by four humans, with generative AI helping with formatting and image generation.

INTRODUCTION

Artificial intelligence (AI) has become ubiquitous in the media and in our daily lives, from self-driving cars and voice assistants to sophisticated game-playing agents, internet search engines, facial recognition and AI chatbots. Its significant impact on industries such as transport, finance, science and entertainment is undeniable; healthcare is poised as the next industry for transformation.

While AI-related health research has grown exponentially in recent decades (figure 1), its adoption into healthcare is still nascent due to the industry's high-stakes and highly regulated nature. Important healthcare considerations include patient safety, data governance, data bias and ethico-legal implications of AI models. Additionally, the increasing interest in deep learning and generative AI models has led to growing concerns over the interpretability of AI models that are being developed.

Neurology has been at the forefront of AI innovation. Practising neurologists are key stakeholders in adopting AI into their specialty, AI could vastly improve patient care, but this needs to be implemented safely, ethically and conscientiously. It is crucial that neurologists gain an understanding of clinical AI and critically appraise the effectiveness, safety and value of the various AI software that they will eventually encounter.

This article serves as a practical guide for non-technical neurologists to understand healthcare AI. By considering the factors discussed, clinical neurologists can help to shape the direction of AI development and implementation in healthcare for years and decades to come.

AI AND MACHINE LEARNING—NAVIGATING THE MINEFIELD OF MISUSED DEFINITIONS

In 1955, John McCarthy coined the term 'Artificial Intelligence' and defined it as 'the science and engineering of making intelligent machines'.[1] More specifically, AI involves machines that demonstrate more-human-like intelligence; however, there is no universally accepted definition of human intelligence. An expert panel consensus in 1997 tried to define 'intelligence' as involving 'among other things, the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas and learn from experience'.[2] In the last century, these were usually rules-based approaches, such as decision flow charts. Some are even implemented as an

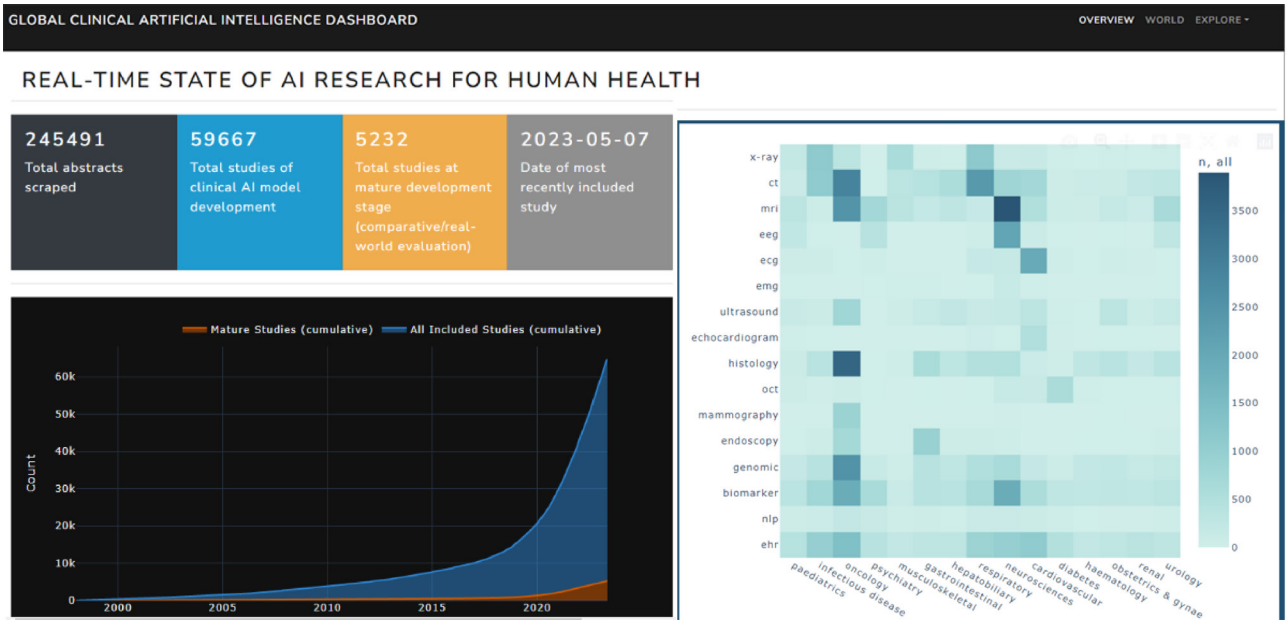

Figure 1 Exponential increase in artificial intelligence (AI)-related publications in healthcare.[34] A real-time dashboard using natural language processing to determine if a research publication claims to use AI at: https://aiforhealth.app/.

automated process in a smartphone app or triaging software. Most experts would not consider these systems to be AI as they are deterministic, overly rigid and prone to errors in missing context. They are essentially scripted flow charts with which clinicians are familiar as 'protocols' or 'algorithms' for choosing investigations or treatments. Most hospitals have plenty of these in the appendices of clinical guidelines, but often they are patchily adopted because they are too voluminous or complex.

Subsequent progress has led to other methods, including machine learning and deep learning (figure 2), which are probabilistic to handle information processing and self-adjusting responses and behaviour to inputs, and significantly less scripted.

Machine learning is a subfield of AI that involves developing computer systems or models that enable computers to 'learn' from data and to improve their performance over time without being explicitly programmed. In other words, machine learning algorithms can automatically identify patterns and relationships in data and use this information to make predictions. Machine learning algorithms can range from those derived from classical statistics, such as logistic regression, to complex deep learning algorithms. Machine learning has been responsible for most of the recent progress in AI. By the 2010s, the term 'AI' had begun to take on a totemic significance, accelerated by advertising of 'the magic sauce'; it is easy to become cynical about the role of marketing when we see toothbrushes or sneakers being 'AI-powered'.

WILL AI REPLACE NEUROLOGISTS?

Concern about job displacement dominates societal discourse about AI systems. Previous industrial technologies have primarily affected blue-collar jobs involving repetitive manual labour, but AI technologies have the potential to impact white-collar jobs.[3 4] However, for neurologists and clinicians, fears of job automation and resultant unemployment are unfounded, as a their task workflows and responsibilities comprise many non-homogeneous activities. A neurologist's work includes many categories of activity including:

- diagnostic and multimodal interpretation for patient care (classical thinking of work of a neurologist);
- clinical decision-making (eg, patient management that blends pragmatism and clinical evidence);
- procedural tasks (eg, delivering medications such as neurotoxins, performing lumbar punctures); and
- human communication (eg, relaying information such as breaking bad news, explaining the diagnosis and treatment options).

As well as these core activities, modern industrialised healthcare includes many related activities including administration (eg, diagnostic coding, writing prescriptions, clinical documentation), para-clinical activities (consensus-building with multidisciplinary teams), and research and education (eg, conducting clinical research, clinical audits, teaching peers and students).

Each of these activities comprises subtasks with different combinations of cognitive, communication and motor contributions, which are not easily segregable. This means that even if an AI system could fully automate two or three steps, it would be very difficult to replicate all the different steps to a uniform degree, and therefore extremely improbable that AI would fully substitute for a neurologist. Instead, as AI automates or improves subtasks, a neurologist's workflow will change as they use AI for different steps of

Au Yeung J, *et al*. *Pract Neurol* 2023;**23**:476–488. doi:10.1136/pn-2023-003757

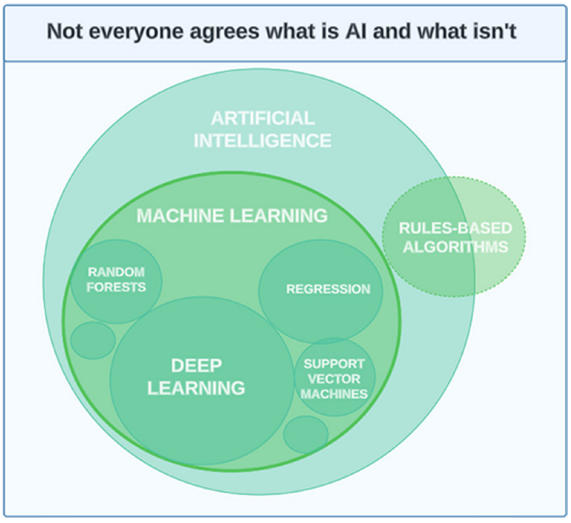

Figure 2 Venn diagram interpretation of the relationship between artificial intelligence (AI), machine learning and deep learning.

clinical activity. If implemented well, this process will help neurologists, who become the decision-making orchestrators and patient communicators.

The easiest parts to automate are those in the non-core activities. For example, AI could alleviate the administrative burden, help to interpret investigations like imaging, summarise patient data and prepopulate forms. Voice dictation for letters is the best example of an activity that has evolved, from individual dictation into tapes for a personal medical secretary, to dictation into a typing pool, then to an online remote typing pool, and then to AI-powered dictation. This has replaced a medical secretary's scribing activity but has not replaced their other roles. Recognising this, most clinical AI systems are designed to be 'assistive' rather than 'autonomous'.[5]

While improving efficiency is a priority for healthcare institutions and economies, neurologists often wish to improve care for their specific individual patient. While this is easier to benchmark in highly standardised diagnostic disciplines such as radiology, it is much harder in clinical neurology, since the neurological history and examination is a semiscripted fully integrated cognitive–verbal–motor activity that relies on examiner skill and experience.[6 7]

A systematic review evaluating the diagnostic accuracy of the neurological examination in diagnosing lumbosacral radiculopathy showed poor sensitivity and specificity in sensory examination (ranging between 0.61–0.63). Motor examination sensitivity was also poor (ranging between 0.13 and 0.61), while reflex testing was moderate (specificity ranging between 0.60 and 0.93).[7] A study reviewing the performance of neurological examination in 46 patients with focal cerebral hemisphere lesions showed poor sensitivity, with only 61% being correctly identified.[8] This high variability means that machine learning and feature-ranking algorithms will drop the individual findings from a traditional neurological examination as being noisy and unreliable. Consequently, most current AI technologies that include features from the clinical assessment use standardised scales, such as the National Institutes of Health Stroke Scale (NIHSS), Rankin, ADAS-Cog, MoCA, MMSE; these scales are designed to 'grade' impairments or function more consistently, rather than to 'diagnose' pathology.

The consequence is that the less consistent components of the clinical examination will remain human delivered while the more consistent aspects of clinical evaluations will feed AI algorithms. AI algorithms will perform better at less clinician-dependent tasks (combining different multimodal findings, high-dimensional image interpretation, biomarker tracking).

Case 1 Using stroke imaging AI

A 65-year-old man presented to an emergency department at 03:00 with sudden onset right-sided weakness and difficulty speaking. He had last been seen well at 17:00 the previous day. There was limited history and his NIHSS score was 12. He underwent urgent CT scan of head, with CT angiogram and CT perfusion scanning. A commercial AI product in the hospital CT scanner processed the images immediately in <10 s, outputting to the picture archiving and communication system and to the stroke neurologist's smartphone (figure 3).

The stroke neurologist reviewed the AI report and images, which showed no early ischaemic changes (upper left image), a large vessel occlusion on the left M1 segment of the middle cerebral artery (upper right image) and a perfusion mismatch of 40 mL, indicating a salvageable ischaemic penumbra. The stroke neurologist decided to thrombolyse immediately (the patient therefore received this treatment within 30 min of arriving in hospital). The AI data then prompted further discussion with an interventional neuroradiologist at a distant tertiary neuroscience centre whether or not to transfer the patient for thrombectomy.

Case 2 Using stroke imaging AI

A 50-year-old man presented with a left-sided hemiparesis and gaze deviation. His NIHSS score was 13 at 6 hours from stroke onset. The stroke AI imaging algorithm again detected a large vessel occlusion and perfusion analysis showed no mismatch, and no reversible penumbra. The neurologist reviewed the source images and affirmed this. Thrombolysis was therefore unlikely to help and might have increased the risk of harm. The size of the infarct core meant that the patient was at risk for malignant middle cerebral artery syndrome, so he was closely monitored for early neurosurgical intervention if necessary (figure 4).

Cases 1 and 2 demonstrate the assistive nature of AI that automates only subtasks of clinical activity, assisting with neuroimaging interpretation in acute

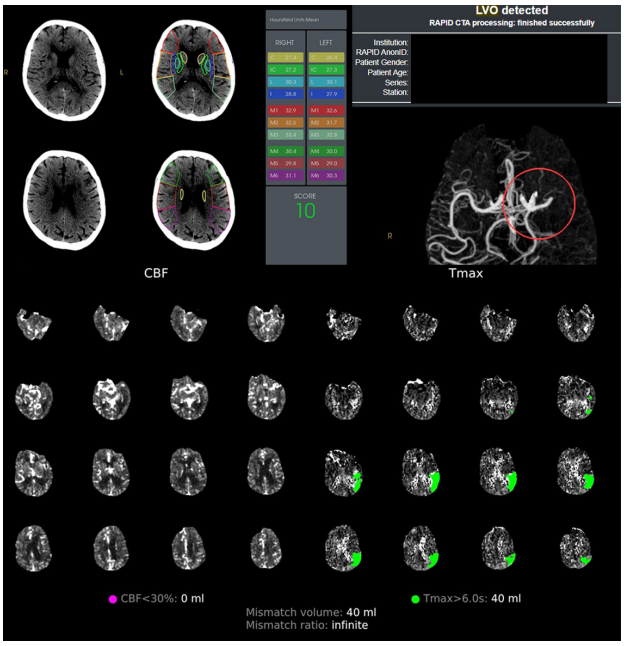

Figure 3 : Screenshots of AI-generated reports of hyperacute stroke imaging to support clinical decision-making: haemorrhage detection (upper left); ASPECTS score for early ischaemic change (bottom left); large vessel occlusion on CT-angiogram, and perfusion mismatch on CT-perfusion (right)

stroke and summarising complex data, with the neurologist taking responsibility for integrating clinical findings and contextual aspects for final decision-making.

Both cases show the value of integrating AI as subtasks into a specific care pathway for a particular disease scenario. Like the above, many of the current health AI technologies are niche solutions that provide a very specific function as a subtask for a very specific clinical need. This works when clinicians have clearly defined that the problem being solved is safe for responsible adoption. This is less straightforward for other uses of AI and AI products, especially for systems claiming to be general purpose across many different clinical scenarios.

EXAMPLES OF AI USES BEING TESTED IN NEUROLOGY

The healthcare areas in which AI are the most developed and ready for use are computer vision (eg, neuroradiology) and natural language processing (eg, extracting meaning and understanding from text). We highlight a few areas in which machine learning is being applied in neurology.

Computer vision/medical imaging analysis

AI-supported analysis of stroke imaging was one of the earliest uses, with multiple algorithms often built within a product suite: (1) to detect early ischaemic change on CT, such as the ASPECTS scores (Alberta stroke program early CT score); (2) to detect haemorrhage; (3) to detect large-vessel occlusion on CT angiograms; and (4) to evaluate ischaemic penumbra on CT-perfusion imaging. Several of these commercial products are already widely used in UK hospitals, for example, Rapid.AI, Brainomix and Viz.ai (figure 5). These can accelerate scan interpretation, especially out of hours and in emergency departments, and when located far away from neuroradiological centres. The models can flag up imaging abnormalities in real time, raising alerts for intervention in stroke, and thereby reducing the door-to-needle time. Real-world evaluations show reasonable performance heavily weighted towards high sensitivity but low specificity, and high false positive rate.[9] The next generation of AI algorithms will not just detect but will aim to prognosticate.[10] Other approaches include use in glioma imaging markers[11] and volumetric and lesion analysis for neurodegeneration, epilepsy and multiple sclerosis. Most approaches rely on 'segmentation', or highlighting, a region of an image with a label to allow further interpretation by a clinician.

A particularly practical use is general anomaly detection algorithms.[12] These are trained to detect that an image is not similar to an average or 'normal' image, and can bring it to the neuroradiologist's attention,

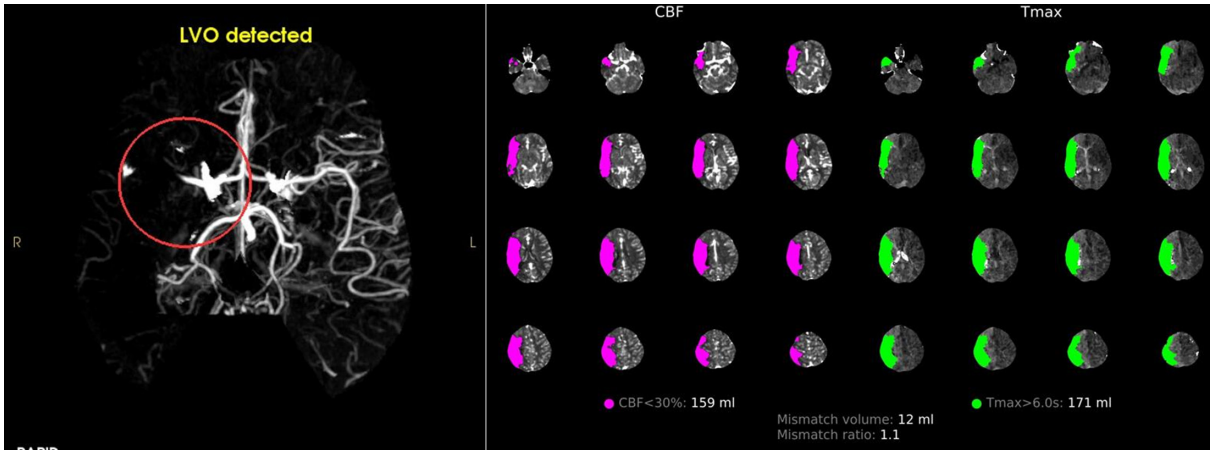

Figure 4 AI-generated segmentation of right middle cerebral artery large vessel occlusion with associated large core and minimal penumbra, indicative of little salvageable tissue

Au Yeung J, *et al*. *Pract Neurol* 2023;**23**:476–488. doi:10.1136/pn-2023-003757

potentially forming an important tool for triaging and managing volume work.

Machine learning in electronic health records

Electronic health records contain self-generating real-world patient data as well as curated data formats in research study case report forms. Machine learning works on these tabulated data to perform outcome prediction, diagnostic classification, anomaly detection, risk estimation, triaging and pathway optimisation. The simplest form uses a set of inputs, and then outputs a risk score, analogous to traditional clinical risk calculators such as NEWS2 or Sepsis Scores. While simple in concept, most published machine learning risk models are underpowered, poorly generalisable and simply badly trained.[13] Machine learning algorithms such as random forest and decision trees can predict prognosis in patients with traumatic brain injury.[14] While there is some marginally improved accuracy compared with traditional warning scores,[15] the systems are often overly complex; it is easy to underestimate the scale and quality of the data required for generalisable algorithms.

The main (but often overlooked) limitation of electronic health records data is their quality; many UK hospital electronic health records do not use internationally standardised data formats for tabulated data. Healthcare professionals often create electronic health records forms using poorly standardised data entry in pursuit of short-term clinical utility for audits or checklists. Furthermore, the burden of completing forms might lead clinicians to leave blank fields or to mis-enter data, and risks causing clinician burnout. The result may be useless and unusable data, unless clinicians concurrently use language AI (or natural language processing).

Natural language processing

Natural language processing is a branch of AI that aims to use machines to interpret, understand or generate

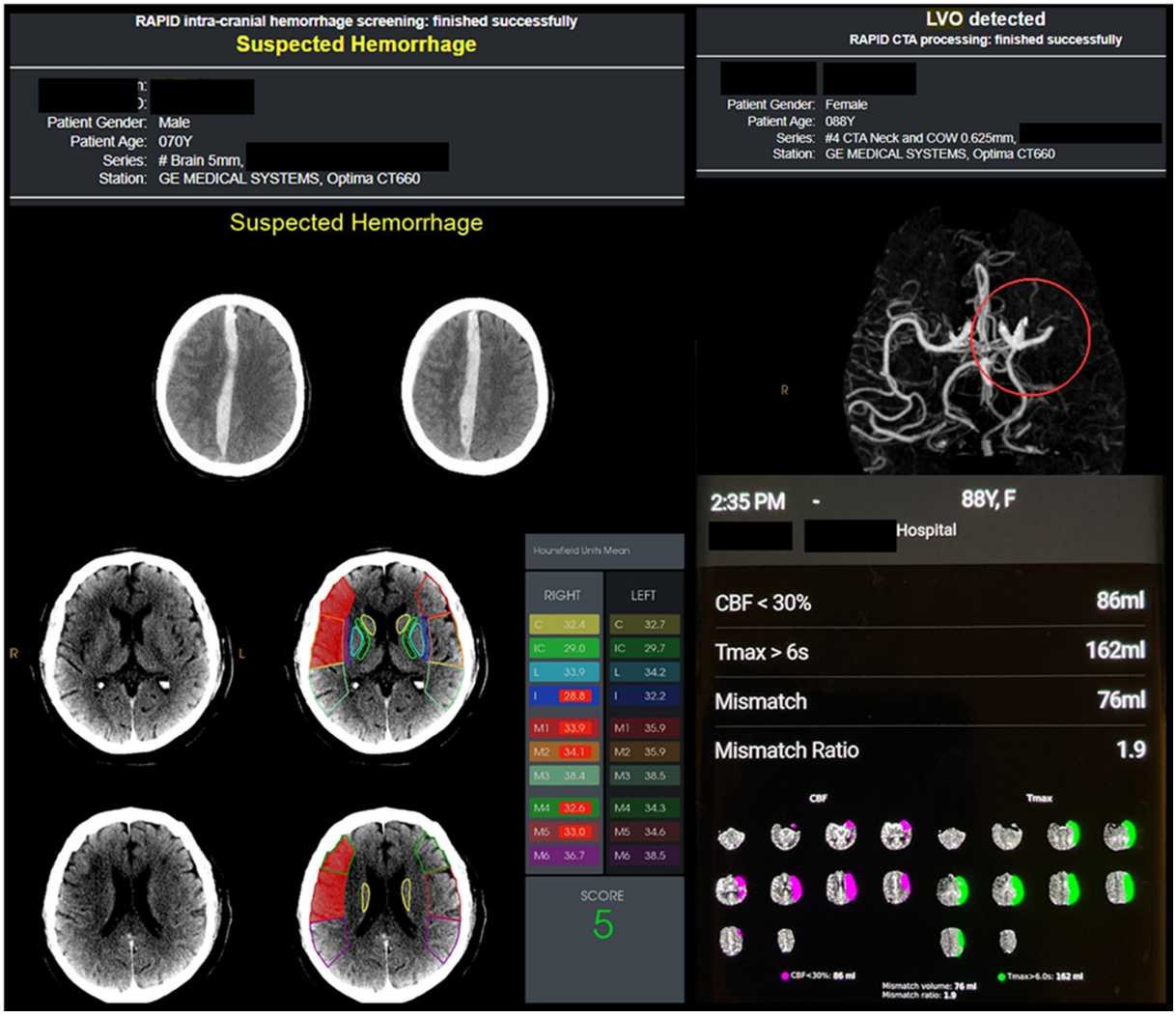

Figure 5 Screenshots of artificial intelligence (AI)-generated reports of hyperacute stroke imaging to support clinical decision-making: haemorrhage detection (upper left); ASPECTS score for early ischaemic change (bottom left); large vessel occlusion on CT angiogram and perfusion mismatch on CT perfusion scanning (right).

human language. The world took notice of this in December 2022, when OpenAI announced ChatGPT, though natural language processing technology has been used and tested for many years. The initial hype was about its performance in answering examination questions, and now many are speculating about its uses in healthcare.[16] This requires careful consideration of the risks if directly involving patients.[17] The greater short-term value to a clinician is in its help in writing letters, and efficiently summarising information.[18]

Separately from large language models for conversations, natural language processing is used for various activities. Our National Health Service (NHS)-grown group (CogStack) uses natural language processing in various clinical workflows including machine-learning enabled clinical coding of stroke comorbidities,[19] whole hospital analysis of disease comorbidities,[20] tracking disease trends,[21] temporal modelling of patient trajectories using natural language processing document text[22] and providing National Institute of Health and Care Excellence (NICE) guideline-based advice (figure 6). Over time, natural language processing will probably be integrated into natural language processing-based structured data systems as well as being combined with machine vision tasks, such as evaluating radiology reports.

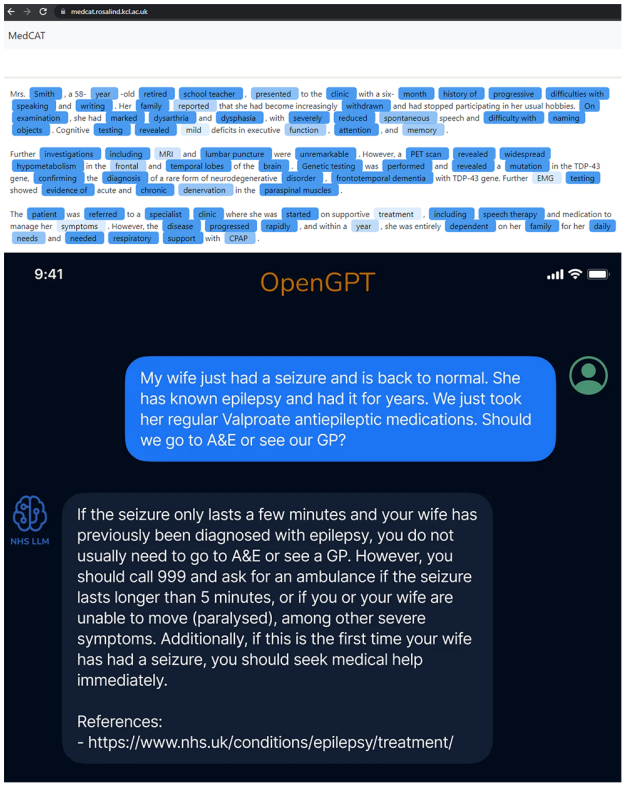

Figure 6 The use of natural language processing to detect words and phrases and to develop ability to allow computable semantics, that is, computers can read meaning (top). The development of large language models trained on NICE guidelines and healthcare data by the authors' team (bottom).

Case 3: Finding missing cases using language AI

Genomic sequencing is made available by Genomics England for specific rare diseases, syndromes and cancers; neurologists may use this service for cases they see. A few have kept patient lists on a spreadsheet or logbook. One senior consultant used to have dozens of such cases, but many have been discharged after making no genetic diagnosis, with no log of suitable cases.

Natural language processing was applied to the entire hospital health record to find patients where the neurologist mentioned a possible cause in a clinic letter but never found the diagnosis. Search terms were applied as well as phrases to identify phenotypes and syndromes without any diagnosis. Patients were ascertained including many 'lost to follow-up', whom the neurologists now contact for genetic sequencing for diagnosis.

A similar approach was also applied to finding women of childbearing age who were taking sodium valproate. Initially, epilepsy clinics were manually audited, but with several non-epilepsy indications for sodium valproate, so this natural language processing was applied hospitalwide in an afternoon, saving months of manual auditing.

Most recently, we used an AI-boosted technique during an upgrade of an electronic healthcare records system at two large London NHS hospital trusts (Kings College Hospital and Guys & St Thomas Hospitals), to detect diagnostic codes in outpatient letters. We then inserted these into a new electronic healthcare records system for clinician reconciliation into internationally standardised problem lists. This task, involving >400k diagnosis codes, would have taken 2 years if done manually.

Voice analysis algorithms

Speech-to-text software can be used to autodictate or summarise consultations and so reduce documentation burden; this is widely used for consumer dictation systems like Siri in Apple smartphones, Alexa in Amazon, and in healthcare, Dragon Dictate by Nuance such that it no longer seems astounding. Neurally based AI models such as Whisper have been freely released so that anyone can tune a voice recognition model. In disease detection, preneural network approaches with signal-processing algorithms were piloted 10 years ago.[23] Neural network 'AI' has substantially increased accuracy in detecting Parkinson's disease[24] and Alzheimer's disease,[25] but large-scale testing has not yet been completed.

Video, movement and remote evaluations

There has been substantial work on this area in healthcare but most are small-scale pilots. Small-scale projects have tested video analysis of tremor, motion detection in seizures and movement disorders, and tracking using sensor remote devices. This will change

Au Yeung J, *et al*. *Pract Neurol* 2023;**23**:476–488. doi:10.1136/pn-2023-003757

very soon, as technology industry and the open-source community have built huge datasets of human motion data (called 'poses') outside of healthcare. Healthcare teams supplemented by the acceleration of consumer hardware for motion capture will soon be able to classify subtle aspects of motion using AI.[26] In neurology, early efforts will be to detect and classify types of abnormal gait from video footage as well as other concurrent information from video (eg, pulse rate, respiratory rate, audio) and other on-patient sensors. It remains to be seen if this is the right approach for technological translation into neurology.

Neurophysiology and AI

Machine learning has shown promise in automating the detection of seizure activity in electroencephalogaphic (EEG) recordings.[27 28] These rely on highly curated public datasets and its generalisability remains to be determined systematically. Alternative approaches have also been explored using machine-learning algorithms to interpret EEG signals for brain–computer interfaces, which is hoped might provide adaptive technologies to improve independence of people with neurodisability.[29] AI-assisted analysis of electromyography signals has also been attempted in motor neurone disease.[30] Nonetheless, there is lack of standardisation in the use of AI in the field of neurophysiology, so most of the described are still in the proof-of-concept phase and some way from wider real-world application.

Histopathology and AI

The application of AI in digital histopathology is another form of machine vision in the automated analysis of histopathology slides. The most obvious use would be in glioma or brain tumour specimens. AI models can be trained to detect and quantify specific pathological features accurately, such as amyloid plaques or Lewy bodies, which indicate these diseases.

The process of doing this training needs semantic sophistication and relies on labelling images carefully for the 'features' rather than classifying into pathology versus non-pathology (figure 7). It is helpful to label with clinically recognised grading systems for mitotic activity, nuclear atypia and microvascular proliferation in neurooncology, as an AI model would learn the relevant features, rather than mislearn other aspects of the images (often related to hospital-specific preparations). AI adds the most value to high-volume specialties.

Biomarker discovery and proteogenomics

This is an exciting area of new discovery and future treatments. Machine learning approaches speed up the interpretation of genomic, proteomic and biomarker information. The most significant impact is Google Deepmind developing an AI model for predicting protein folding, called Alphafold.[31] This has had significant impact in structural biology, including drug discovery, protein function modelling, protein design and target prediction. Another area of development is synthetic chemistry, where molecular designs are imagined using AI, and then best fit for molecular targets are found. In neurology, a knowledge-based approach was used to identify novel defective RNA-binding proteins in amyotrophic lateral sclerosis.[32] Similarly, machine learning has identified prognostic biomarkers in people with high-grade glioma.[33] The pace of preclincial development is accelerating as AI-designed drugs by biotech companies such as Exscienta have begun or completed phase I clinical trials in oncology and immunology.

Most of this biomarker work does not directly affect clinicians until new medicines and therapies (discovered with AI support) become available after relevant regulatory approvals.

MACHINE LEARNING—HOW DOES AN ALGORITHM OR MODEL LEARN?

For a machine learning model to 'learn', it must be provided with data; it can then generate patterns based on the data inputted. The main methods of learning are: supervised, unsupervised, reinforcement and deep learning figure 6. The choice of learning method depends on the nature of the problem and the availability of labelled data. It is essential that training data are in the correct format, type and quality for the model to learn. Clinicians typically underestimate the amount of data required, which is usually a data volume of tens of thousands to billions. Most machine learning advances in medicine have come through supervised learning, although recently unsupervised deep learning methods have attracted great interest as they need minimal or no labels.

- *Supervised learning*: this involves training a machine learning model with a labelled dataset, where there are known input and output variables. The model establishes a relationship between input and output variables to predict unseen data accurately. For example, in predicting Parkinson's disease, supervised learning can be used to train a machine learning model on a dataset of MR brain scans that have been labelled as being either healthy or as having Parkinson's disease. The model learns to recognise data patterns that differentiate healthy and diseased brains, enabling predictions about new, unseen scans.
- *Self-supervised (or unsupervised) learning*: this involves training the model on an unlabeled dataset where output variables are unknown. The model identifies patterns or structures in the data by implementing clustering or dimensionality reduction techniques. For example, a machine learning model can be trained on a dataset of clinical records for patients with Parkinson's disease, without any explicit labels. The model can then identify common patterns of symptoms among patients and use them to cluster and identify distinct clinical subtypes of Parkinson's disease.
- *Reinforcement learning*: this involves training the model through trial and error by receiving feedback in the form

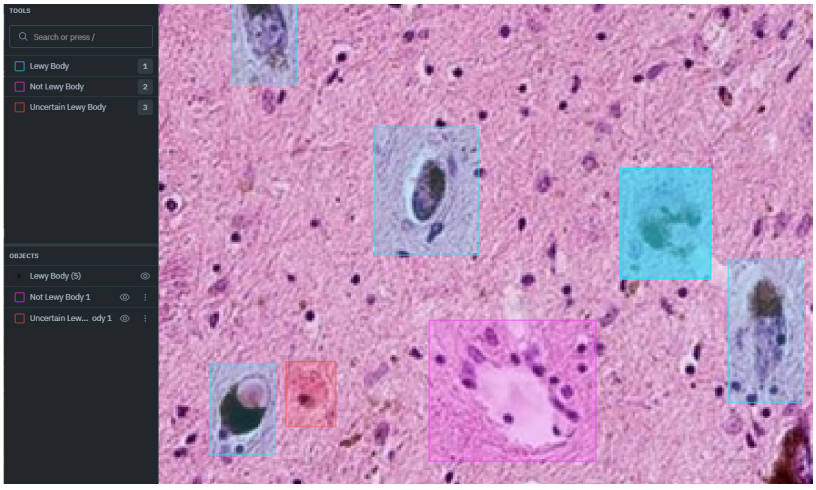

Figure 7 Example of a software tool for labelling a Lewy body, to train AI-automated labelling of Lewy bodies on central nervous system histopathology (image produced by author).

of rewards or penalties while interacting with an environment. This approach has famously achieved superhuman level performance in games such as chess, Go and other board games. A clinical analogy would be a deep-brain stimulation system that is 'rewarded' by reducing a patient's parkinsonian tremor. By learning and adapting to patient's feedback, the model can learn to alter the stimulation parameters that are tailored to the patient's individual needs.

The above paradigms describe how to train algorithms, and the specific algorithms are wide-ranging mathematical constructs, example: Support Vector Machines, Random Forests, Logistic Regression, XGBoost. Since 2020s, Artificial Neural Networks and Deep Learning are the dominant types. These are inspired by neuroscience—where digital neurones are 'activated' based on a summation of its inputs from other digital neurones, analogous to a biological neurone reaching a threshold to produce an action potential. These digital neurones are connected in layers modelled after the visual neocortex, and the degree to which one neurone influences improves through machine learning, for example, machine vision to understand what a picture sees (figure 7).

The engineering of many layers of neurones allows very complex mathematical processes to be accomplished and to represent data in very abstract forms. A detailed survey of the different types of artificial networks is beyond the scope of this article, but interested readers can review online encyclopaedias (https://www.asimovinstitute.org/neural-network-zoo/). Newer architectures such as transformers, autoencoders and diffusion models combine many of these into very complex architectures. As neural networks are computationally demanding, those with access to nanoscale AI-chip technologies physically or through data centres or the cloud dominate the space as they can produce, fine-tune and use new AI models. Farsighted NHS organisations and universities have invested in these specialised AI-computers and internal infrastructure.

Machine learning versus traditional statistics

The line between traditional statistical methods and machine learning is blurred. Many machine learning techniques build on classical statistical methods to create intelligent and adaptable models. In general, machine learning refers to models and algorithms that learn and adapt to new data by adjusting their weights and biases; this is distinct from traditional statistical methods that tend to rely on prespecified models modelled to fit the static data.

A key difference is the intended goal; classical medical statistics are used primarily to analyse data to uncover patterns, relationships and trends. In contrast, machine learning is focused on developing algorithms and models that learn to incorporate the above patterns, relationships and trends to make predictions or decisions based on that learning. Machine learning is also more suited to having many variables (dimensions) of input data (thousands to millions of pixels in images, millions of characters or words in text) whereas classical statistics are often limited by the degrees of freedom and repeated hypothesis statistical testing.

CRITICAL APPRAISAL OF MACHINE LEARNING AND AI MODELS

New 'state-of-the-art' machine learning models are being released on a weekly basis in code repositories, preprints, peer-reviewed research articles and as commercial products. At this pace, it may seem hard to keep up with each new iteration of technology. However, all the new models are generally built on similar architectures or recipes with small changes that lead to a difference in benchmarking performance.

Numerous international frameworks have been developed for reporting clinical studies on AI

Au Yeung J, *et al*. *Pract Neurol* 2023;**23**:476–488. doi:10.1136/pn-2023-003757

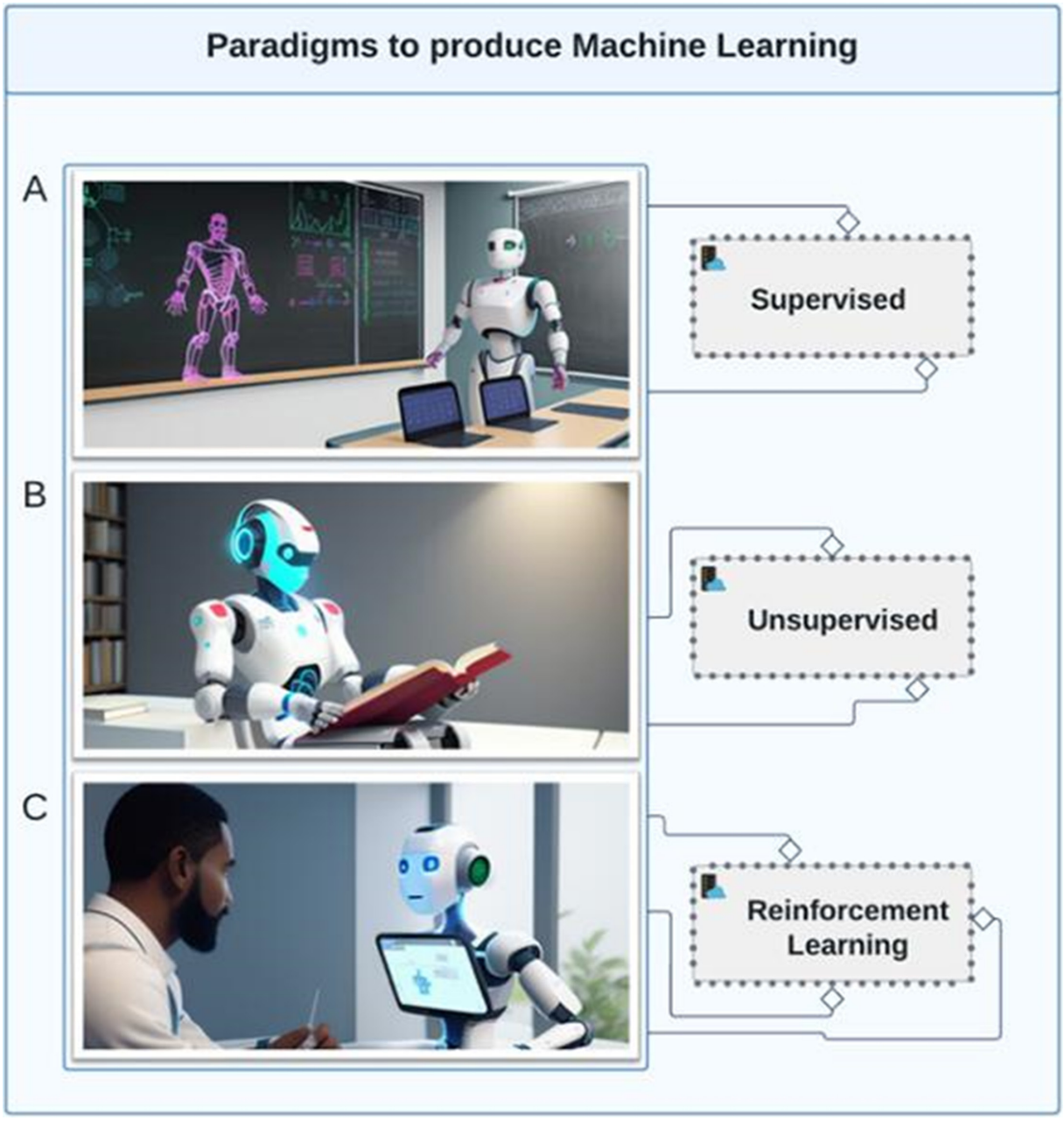

Figure 8 Paradigms to produce machine learning.

development (eg, CONSORT-AI, STARD-AI, SPIRIT-AI, TRIPOD-AI, PROBAST-AI, DECIDE-AI),[13 34–38] but there is variable compliance on this especially with the number of different frameworks. For a clinician, the machine learning paper should then review the following:

1. *Understanding the problem*: the first step is to understand the problem the AI model is attempting to solve. Is this a real clinical problem that clinicians encounter? Does it specifically require AI to solve? Is the problem niche or widespread? A local departmental problem may be due to structural issues or human factors, and so creating an AI software to solve this will be costly, inappropriate and will not scale across the organisation.

2. *Reviewing the methods and source training data:* the characteristics and quality of the training data used to produce the AI model is key (see Jargon Box). An AI model attempts to learn the patterns and trends within its training data, and so the quality of data is often more important than the type of AI architecture used. Are there intrinsic biases within the data (ie, do the data adequately mimic prevalence of clinical scenarios? are only the appropriate biases captured in the training data?).

3. *Evaluating the performance*: the results of the AI model should be evaluated using appropriate performance measures. Datasets are usually split into a test set and a training set. The model performance is first evaluated on the training set, and then on the test set

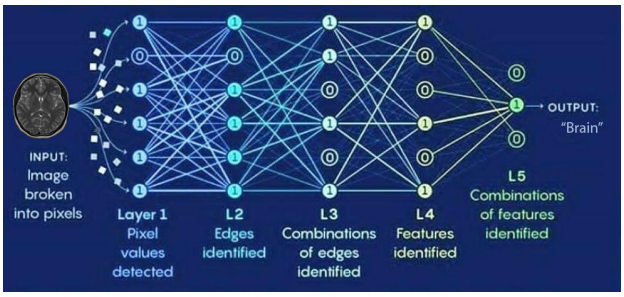

Figure 9 The base design of artificial neural networks. An image is broken up into pixels and the brightness of each pixel is fed into a digital neurone in layer 1 of the artificial neural network. The digital neurones are connected to layer 2 digital neurones, and if a certain threshold of activation occurs, the layer 2 neurone activates as well, which then leads onwards to layer 3 and layer 4. The amount of activation needed to activate the next layer is called the 'model weights' and these weights are generated through a process similar to long-term potentiation in biological neurones (following the Hebbian Rule). Each layer then produces an abstraction of the preceding layer, such that lines and edges are captured, followed by whole shapes and polygons, finally ending in an output layer.

(to which the model was not exposed during training). An AI-driven diagnostic algorithm will need sensitivity and specificity data and a standard area under the receiver operator characteristic curve analysis (see jargon box). Beyond these simple performance measures, the model needs additional performance benchmarking against established benchmarks (eg, model performance compared with standardised retrospective datasets, benchmarking against expert clinicians). As gold standard, a prospective real-world multisite trial would appraise how the model performs in the real-world clinical setting and test generalisability across healthcare settings.

4. *Understanding implementation dependencies in the real world*: an additional consideration is the cost of implementing and maintaining AI in the real world. Often, there are dependencies and critical bits of infrastructure or digital ecosystems that render the system not reproducible in practice. In the case of AI models, poor performance could be due to issues with poor generalisability of the AI model, lack of compatibility with hospital IT infrastructure, barriers to data access and confidentiality or poor workforce adoption due to model complexity. This is often not mentioned in research studies, and an experienced clinician would want more prospective machine learning/deep learning clinical trials that evaluate the model's real-world effectiveness prospectively.[39]

5. *Looking beyond the model*: even an AI model with a perfect performance will not necessarily be adopted. Workforce adoption relies on many factors. First, an understanding and trust of the technology, as it is much easier to trust a transparent model than a 'black-box' model. Second, the AI application needs to be implemented seamlessly into a neurologist's workflow so that it does not detract and burden their existing work. This is no easy task, but requires an in-depth understanding of clinician workflows, healthcare ecosystem and attention to user interface design. Furthermore, just like any other technology in health services, it needs to demonstrate improved health outcomes, reduced costs and improved patient and clinician satisfaction. Given the finite resources of healthcare systems, it is only logical that it adopts and maintains interventions that are cost saving.

Software development and evidence evaluations have phases that are analogous to clinical drug trials (table 1). Various NHS AI Lab and NIHR clinical trials funding schemes already mirror this format (https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ai-health-and-care-award-winners/).

HOW DOES AI SOFTWARE GO FROM 'BENCH' TO 'BEDSIDE'?

Most clinicians would not confuse the use of AI-based systems in consumer or general population with patient care; there is often a grey area between the use of AI *in healthcare* and the use of AI *for healthcare*. Using spell-checking AI for emailing a colleague or dictating an outpatient letter is very different from using AI in delivering care to patients, with its substantially more regulatory safeguards.

AI technologies used for healthcare are considered to be 'software as medical device' (SaMD), and so must comply with a series of regulatory requirements. In most cases, the AI is a non-adaptive algorithm that is in a frozen state of learning (the algorithm's internal parameters and performance do not change over time); adaptive machine learning algorithms remain an evolving regulatory landscape.

The crucial first step in the regulatory approval process is to define the intended use of an AI algorithm or SaMD. This involves describing the device's specific purpose, function and target population. The intended use statement serves as the basis for classification, evaluation and approval of the SaMD.

There are also different approvals based on the regulatory domains; the Conformitè Europëenne marking in Europe, the Food and Drug Administration for the USA and the UK Conformity Assessment marking for the UK. In the UK, this is regulated by the Medicines and Healthcare Products Regulatory Agency (MHRA) approval. The MHRA regulates SaMD in the UK in accordance with principles of safety, performance, quality, traceability and vigilance. Importantly, the MHRA requires clinical evaluation of the safety and performance of a product (premarket and post market) https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device. Other applicable quality standards include ISO13485 (and DCB0129), which provides clinical safety assurance from the device or software manufacturer for distribution.

JARGON BOX: Data considerations for artificial intelligence (AI)

Data quality is crucial for the training and validation as well as for implementing machine learning models. Data of good quality have been sorted, standardised, cleaned and contain all the attributes necessary for the model's learning process.

Accuracy and relevancy: training data should provide meaningful and accurate information to the model that reflect the real world (so the model can capture the data trend correctly rather than introducing noise or bias).

Consistency: for supervised training datasets, human labellers must label data consistently for the model to learn. Interlabeller agreement can be a useful metric to evaluate if labellers are consistent using the same approach.

Ground truth/gold standard: machine learning algorithms need to be trained on outcomes with minimal inconsistency and interobserver variation, to generate outputs. This is often not available in many aspects of neurology where features and diagnoses are opinion based.

Standardised and interoperable: healthcare data standards allow data recorded in one system to be interrogated by another system. Good quality data used in AI should follow data standards so that any AI model is interoperable. Examples of healthcare data standards include SNOMED-CT, HL7v2, FHIR, DICOM and LOINC.

Representativeness and data diversity: the data should represent the target population or the specific context in which the AI model will be deployed. It should cover the diversity of the data distribution to avoid biased or skewed models. This can be challenging in rare diseases.

Completeness: data should be complete, containing all the necessary information required for the AI model training task. Missing or incomplete data can hinder the model's performance. This can occur if hospital's electronic health records are siloed or fragmented.

Data privacy and information governance: health data are privileged special category data, and is governed by multiple legislations, in the UK specifically by the Data Protection Act (2018), the Health Service (Control of Patient Information) Regulations 2002, and Common Law of Confidentiality. Data protection and information security measures also apply.

JARGON BOX: Measures of AI model performance

AUROC (area under the receiver operator characteristic curve) is a metric of performance used to evaluate the quality of binary classification models, which measures the ability of the model to distinguish between positive and negative classes. A score of 1.0 is perfect.

Precision (also known as positive predictive value) is a measure of a model's accuracy in identifying true positives out of all the positive predictions made by the model. A score of 1.0 is perfect.

Recall (also known as sensitivity) is a performance metric that measures the ability of a model to identify all positives in a dataset. A score of 1.0 is perfect.

F1 score is a combined measure of a model's precision and recall, which provides a balanced evaluation of a model's performance. A score of 1.0 is perfect.

Overfitting is where an algorithm is overtrained on too few data; initial accuracy measures are absurdly high but the algorithm fails to perform as well on other datasets from other hospitals.

Benchmarks in standardised datasets are established and often large datasets which groups create for machine learning researchers to train and evaluate performance of new machine learning models. This includes imaging datasets such as Chest X-ray14,[41] multimodal electronic health record datasets such as MIMIC,[42] question-answering datasets such as MedMCQA.[43]

Benchmarks in real-world evidence where the performance is evaluated against real world

Once an SaMD meets regulatory requirements for distribution, its use in clinical care and clinical systems requires distinct regulatory compliance from the healthcare providers, specifically ISO14971 (and DCB0160) compliance for clinical safety and risk management.

Finally, to achieve clinical impact, it is critical to integrate such AI-powered software into clinical IT systems. Integration means connecting all the relevant aspects of the software into existing systems so that data flow seamlessly with accurate data transformations. Often, and especially pilot projects, this integration is not done as it is too expensive or too laborious to perform for a time-limited project. It may also sometimes be too niche for a hospital IT dept to expend resources on doing especially if the software is not designed around international digital standards. This failure of integration to international standards is the most common obstacle to many digital projects in healthcare (and neurology).

For clinical impact, health AI should be an organisation-led exercise (not as a single individual researcher or clinician) and requires collaboration and convergence of people outside of traditional healthcare professions, including engineers. Once implemented, it is important that the deployed AI algorithm is then reviewed regularly for data and model drift, as well as performance and bias. AI software needs constant tending and ongoing overhead costs (possibly more so than traditional software).

FUTURE CHALLENGES

Healthcare has undergone a digital transformation in the last two decades, and patients generate an enormous amount of data with each clinical visit in the form of free text, images and health data. AI models thrive with large amounts of data, which reveal complex patterns

Table 1 Comparison of simplified phases of evaluation of a drug versus artificial intelligence (AI) in healthcare

Study phases	Drug	AI in healthcare[44–46]
Phase 0 preclinical/discovery	Compound/drug target development Preclinical/lab studies	Proof-of-concept studies (usually on a static/ retrospective dataset) Algorithm development and performance metrics evaluation
Phase I safety	Safety assessment Evaluating metabolism and optimal therapeutic dosage Adverse effects	Feasibility to implement into an existing workflow 'Real world' evaluation of algorithm performance Safety evaluation
Phase II efficacy and safety	Prospective efficacy and safety evaluation/ clinical trial (in a larger study group, >100 patients with controls)	Prospective efficacy and safety evaluation/ clinical trial (in a larger study group, potentially multi-departmental or hospital-wide)
Phase III therapeutic efficacy	Efficacy and safety clinical trial (>1000 patients with controls) Medium to long term adverse event monitoring	Efficacy and safety clinical trial (potentially hospitalwide or multitrust, with controls) Medium-term to long-term performance evaluation compared with control/existing non-AI workflows
Phase IV safety and effectiveness	Postmarket surveillance	Postdeployment surveillance

and trends with the potential to transform patient care in neurology. Understandably, there is significant hype around AI in both the news and academia declaring that AI performance has surpassed clinicians and that AI will soon replace doctors. This invariably often wilfully neglects to elaborate on the many subtasks involved in the work of a neurologist and other doctors.

Many significant key challenges lie ahead that are beyond the scope of this article. These challenges include tackling inequalities and data biases in AI. There are also large discrepancies between continents and their AI expertise, capital and data access to train large AI models. This means that a few big players tend to dominate the AI conversation and direction of development. Ethical concerns around AI-related medical errors require careful thought; current AI systems are merely assistive, so decision-making and responsibility remain with the clinician. However, if we were to advance to autonomous AI systems, how would we build a safe system and who would take responsibility when mistakes occur? Furthermore, AI models should be trained and deployed in ways that protect patient data and privacy. Patient engagement groups may help guide development of medical AI software. Finally, the advent and rapid marketing of proprietary large language models such as ChatGPT (OpenAI) and Bard (Google) have transformational potential in healthcare, but can lead to unintended risk of misuse or abuse. Some experts have called for a temporary pause in the development of powerful large language models and a closer focus on AI safety.[40]

Many neurologists are probably already using AI systems in their day-to-day life outside of medicine, and many will also already be using some forms of healthcare AI in their clinical practice. Healthcare AI is an applied science, and so relies heavily on real-world applications by front-line staff. Practising neurologists will be key to adopting AI in healthcare, with a crucial role to ensure that it is implemented effectively, safely and responsibly. Doctors that can use AI systems safely will likely eventually displace those who do not. Neurologists must embrace these technologies to shape the way AI is used in neurology, and not let these systems be designed only by technologists.

Key points

- Artificial intelligence (AI) is a general-purpose technology.
- Neurologists will not be replaced, but will orchestrate these different assistive AI tools for different tasks.
- AI software use is most mature in hyperacute stroke, and prototypes proliferate in most other neurology specialties.
- AI software is regulated as a 'Software as a Medical Device' (SaMD) and there are international frameworks of evidence of benefit.
- It is essential to have high-quality data standardisation.

Further reading

Due to the rapid nature of AI development, most material is outdated by the time it is published in textbooks or journals. Most up-to-date AI developments are on digital publishing platforms such as blogs, online gazettes (Substack.com, Medium.com or Towards Data Science) articles), podcasts and online video tutorials. To get started, we recommend the following resources:

- Kings Innovation Scholars, Big Data and AI skills for the NHS health workforce https://innovationscholars.er.kcl.ac.uk/.
- Andrew Ng's Machine Learning course on Coursera. This will cover fundamentals of ML and mathematics that underlie ML models https://www.coursera.org/specializations/deep-learning.

Au Yeung J, *et al*. *Pract Neurol* 2023;**23**:476–488. doi:10.1136/pn-2023-003757

- Both MIT and Stanford have online courses freely available on YouTube, you may want to view MIT Introduction to Deep Learning 6.S191 or Stanford CS230: Deep Learning.
- AI for healthcare substack, short blogs with natural language processing focus from our AI group: https://aiforhealthcare.substack.com/.

For those seeking technical literacy, we recommend learning about Python, Git repositories, shell scripting, Docker containers, REST API's and cloud environments. As with learning a foreign language, this will take 9–12 months of immersion to get basic literacy. Proficiency or mastery takes many years and is as much a life-long learning process as medicine.

Acknowledgements Thanks to the many informatics, data, NHS and academic departments of Kings Health Partner organisations that has enabled some of the work described.

Collaborators Not applicable

Contributors JAY and YYW drafted the first version. ZK provided critical review. JTHT provided critical review, edits. Generative AI (personalised Stable Diffusion) helped with image generation.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests JTHT is a member of the Confidentiality Advisory Group of the UK Health Research Authority, his views here are his own and do not reflect the view of the Health Research Authority.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; externally peer reviewed by Helen Oram, London, UK, and Gavin Giovannoni, London, UK.

Data availability statement No data are available. Not applicable.

ORCID iD
James T H Teo http://orcid.org/0000-0002-6899-8319

REFERENCES

1 McCarthy J, Minsky ML, Rochester N, *et al*. A proposal for the dartmouth summer research project on artificial intelligence. *AI Mag* 1955;27:12.
2 Gottfredson LS. Mainstream science on intelligence: an editorial with 52 signatories, history, and bibliography. *Intelligence* 1997;24:13–23.
3 Eloundou T, Manning S, Mishkin P, *et al*. Gpts are Gpts: an early look at the labor market impact potential of large language models. 2023
4 Eloundou T*et al*. Gpts are Gpts: an early look at the labor market impact potential of large language models; 2023.
5 Bitterman DS, Aerts H, Mak RH. Approaching autonomy in medical artificial intelligence. *Lancet Digit Health* 2020;2:e447–9.
6 Hansen M, Sindrup SH, Christensen PB, *et al*. Interobserver variation in the evaluation of neurological signs: observer dependent factors. *Acta Neurol Scand* 1994;90:145–9.
7 Teo JTH, Dinu V, Bernal W, *et al*. Real-time clinician text feeds from electronic health records. *NPJ Digit Med* 2021;4:35.
8 Anderson NE, Mason DF, Fink JN, *et al*. Detection of focal cerebral hemisphere lesions using the neurological examination. *J Neurol Neurosurg Psychiatry* 2005;76:545–9.
9 Chan N, Sibtain N, Booth T, *et al*. Machine-learning algorithm in acute stroke: real-world experience. *Clin Radiol* 2023;78:e45–51.
10 Bonkhoff AK, Xu T, Nelson A, *et al*. Reclassifying stroke lesion anatomy. *Cortex* 2021;145:1–12.
11 Booth TC, Williams M, Luis A, *et al*. Machine learning and glioma imaging biomarkers. *Clin Radiol* 2020;75:20–32.
12 Pinaya WH, Tudosiu P, Gray RJ, *et al*. Unsupervised brain anomaly detection and Segmentation with transformers. International Conference on Medical Imaging with Deep Learning; 2021
13 Zhang J, Whebell S, Gallifant J, *et al*. An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. *Lancet Digit Health* 2022;4:e212–3.
14 Khalili H, Rismani M, Nematollahi MA, *et al*. Prognosis prediction in traumatic brain injury patients using machine learning algorithms. *Sci Rep* 2023;13:960.
15 Goh KH, Wang L, Yeow AYK, *et al*. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun* 2021;12:711.
16 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. reply. *N Engl J Med* 2023;388:2400.
17 Au Yeung J, Kraljevic Z, Luintel A, *et al*. AI Chatbots not yet ready for clinical use. *Front Digit Health* 2023;5:1161098.
18 Searle T, Ibrahim Z, Teo J, *et al*. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *J Biomed Inform* 2023;141:104358.
19 Shek A, Jiang Z, Teo J, *et al*. Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing. *Eur J Neurol* 2021;28:4090–7.
20 Bean DM, Kraljevic Z, Shek A, *et al*. Hospital-wide natural language processing summarising the health data of 1 million patients. *PLOS Digit Health* 2023;2:e0000218.
21 Tsanas A, Little MA, McSharry PE, *et al*. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Trans Biomed Eng* 2010;57:884–93.
22 Kraljevic Z, Bean DM, Shek A, *et al*. Foresight -- generative pretrained transformer (GPT) for modelling of patient timelines using Ehrs. arXiv Preprint arXiv:2212.08072; 2022.
23 Wang X, Peng Y, Lu L, *et al*. Chestx-ray: hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common Thorax diseases. In: Lu L, Wang X, Carneiro G, et al., eds. *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics. Advances in Computer Vision and Pattern Recognition*. Cham: Springer, 2019.
24 Tannemaat MR, Kefalas M, Geraedts VJ, *et al*. Distinguishing normal, neuropathic and Myopathic EMG with an automated machine learning approach. *Clin Neurophysiol* 2023;146:49–54.
25 Haulcy R, Glass J. Classifying Alzheimer's disease using audio and text-based representations of speech. *Front Psychol* 2020;11:624137.

26 Bourached A, Griffiths RR, Gray R, *et al*. Generative model-enhanced human motion prediction. *Appl AI Lett* 2022;3:e63.
27 Ahmad I, Wang X, Zhu M, *et al*. EEG-based epileptic seizure detection via machine/deep learning approaches: a systematic review. *Comput Intell Neurosci* 2022;2022:6486570.
28 Maimaiti B, Meng H, Lv Y, *et al*. An overview of EEG-based machine learning methods in seizure prediction and opportunities for Neurologists in this field. *Neuroscience* 2022;481:197–218.
29 Aggarwal S, Chugh N. Review of machine learning techniques for EEG based brain computer interface. *Arch Computat Methods Eng* 2022;29:3001–20.
30 Tawa N, Rhoda A, Diener I. Accuracy of clinical neurological examination in diagnosing Lumbo-sacral radiculopathy: a systematic literature review. *BMC Musculoskelet Disord* 2017;18:93.
31 Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with alphafold. *Nature* 2021;596:583–9.
32 Bakkar N, Kovalik T, Lorenzini I, *et al*. Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. *Acta Neuropathol* 2018;135:227–47.
33 Shu C, Wang Q, Yan X, *et al*. Whole-genome expression microarray combined with machine learning to identify prognostic biomarkers for high-grade glioma. *J Mol Neurosci* 2018;64:491–500.
34 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.
35 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2:e549–60.
36 Rivera SC, Liu X, Chan A-W, *et al*. The SPIRIT-AI and CONSORT-AI working group. guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370:m3210.
37 Collins GS, Dhiman P, Andaur Navarro CL, *et al*. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.
38 Sounderajah V, Ashrafian H, Aggarwal R, *et al*. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. *Nat Med* 2020;26:807–8.
39 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
40 Future of Life Institute. Pause giant AI experiments: an open letter. Future of Life Institute, . 2023
41 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
42 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
43 Pal A, Umapathi LK, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice Dataset for medical domain question answering. Proceedings of the Conference on Health, Inference, and Learning, in Proceedings of Machine Learning Research 174:248-260; 2022 Available: https://proceedings.mlr.press/v174/pal22a.html
44 Park Y, Jackson GP, Foreman MA, *et al*. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 2020;3:326–31.
45 MHRA. Software and Artificial Intelligence (AI) as a Medical Device, 23 July . 2023Available: https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device
46 FDA.gov. Artificial Intelligence and Machine Learning in Software as a Medical Device, January . 2021Available: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd7