University of Dundee

Accuracy of Automated Computer-Aided Diagnosis for Stroke Imaging
Wardlaw, Joanna M.; Mair, Grant; von Kummer, Rüdiger; Williams, Michelle C.; Li, Wenwen; Storkey, Amos J.
Published in:
Stroke
DOI:
10.1161/STROKEAHA.121.036204
Publication date:
2022

Document Version
Peer reviewed version

Link to publication in Discovery Research Portal

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Title: The accuracy of automated computer-aided diagnosis for stroke imaging. A critical evaluation of current evidence.

Authors: Joanna M Wardlaw,1 MD, FRCR, FMedSci
Grant Mair,1 MD, FRCR
Rüdiger von Kummer,2 MD
Michelle C Williams,3 PhD, FRCR
Wenwen Li,1 PhD
Amos J Storkey,4 PhD
Emanuel Trucco,5 PhD
David S. Liebeskind,6 MD
Andrew Farrall,1 PhD, FRC
Philip M Bath,7 DSc, FMedSci
Philip White,8 MD, FRCR

Affiliations: 1: Centre for Clinical Brain Sciences, UK Dementia Research Institute Centre at the University of Edinburgh, Chancellor’s Building, Little France, Edinburgh, EH16 4SB, UK
2: Institute of Diagnostic and Interventional Neuroradiology, Universitätsklinikum Carl Gustav Carus, Dresden, Germany
3: Centre for Cardiovascular Science, University of Edinburgh, Chancellor’s Building, 49 Little France, Edinburgh, EH16 4SB, UK
4: School of Informatics, University of Edinburgh.
5: VAMPIRE project, Computing, School of Science and Engineering, University of Dundee.
6: Neurovascular Imaging Res Core, UCLA, Los Angeles, CA
7: Stroke Trials Unit, Mental Health & Clinical Neuroscience, University of Nottingham, Queen’s Medical Centre campus, Derby Road, Nottingham NG7 2UH, UK
8: Translational and Clinical Research Institute, Newcastle University, Henry Wellcome Building, Framlington Place, Newcastle upon Tyne and Newcastle upon Tyne Hospitals NHS Trust, UK

Correspondence: JMW as above

Running Title: AI in acute stroke diagnosis

Key words: stroke; machine learning; artificial intelligence;

Disclosures: GM received speaker and consulting fees from Canon Medical Ltd, Europe. PMB provided consultancy for Diamedica, Phagenesis, Moleac, Sanofi, Nestle. MCW received speaker fees from Canon Medical Systems. DSL consultants for imaging core lab services to Cerenovus, Genentech, Medtronic, Stryker and Rapid Medical. PMW reports institutional unrestricted educational grants from Medtronic, Stryker and Penumbra; Institutional research funding from Microvention and IBEX and consulting fees/core lab services from Microvention and E-VASC. AJF received institutional educational grants from Medtronic for the development of training materials in stroke radiology. AS has received funding from Huawei, ARM and Microsoft Research, Cambridge not related to the topic of the current paper. JMW holds institutional academic grants from the
Fondation Leducq, EU H2020 programme, Stroke Association, Health Data Research UK, British Heart Foundation, Alzheimer’s Society, Weston Brain Institute, and Row Fogo Charitable Trust for research unrelated to the

**Funding:** JMW is supported by the UK Dementia Research Institute, which receives its funding from DRI Ltd, funded by the UK Medical Research Council, Alzheimer’s Society and Alzheimer’s Research UK. GM is Stroke Association Edith Murphy Foundation Senior Clinical Lecturer (SA L-SMP 18\1000). We acknowledge funding from the Stroke Association (TSA_CR_2017/01) and Medical Research Council Proximity to Discovery fund (MC_PC_17188) for the RITE Study. RvK receives royalties as Editor-in-Chief of *Neuroradiology*. WL is funded by Health Data Research UK which receives its funding from the Medical Research Council. MCW is supported by the British Heart Foundation (FS/ICRF/20/26002). PMB is supported by the British Heart Foundation and NIHR Health Technology Programme.

**Acknowledgements:** PMB is Stroke Association Professor of Stroke Medicine and an Emeritus NIHR Senior Investigator.
Abstract
There is increasing interest in computer applications, using artificial intelligence (AI) methodologies, to perform healthcare tasks previously performed by humans, particularly in medical imaging for diagnosis. In stroke, there are now commercial AI software for use with CT or MR imaging to identify acute ischemic brain tissue pathology, arterial obstruction on CT angiography or as hyperattenuated arteries on CT, brain hemorrhage, or size of perfusion defects. A rapid, accurate diagnosis may aid treatment decisions for individual patients and could improve outcome if it leads to effective and safe treatment; or conversely, to disaster if a delayed or incorrect diagnosis results in inappropriate treatment. Despite this potential clinical impact, diagnostic tools including AI methods are not subjected to the same clinical evaluation standards as are mandatory for drugs. Here, we provide an evidence-based review of the pros and cons of commercially-available automated methods for medical imaging diagnosis, including those based on artificial intelligence (AI), to diagnose acute brain pathology on CT or MRI in patients with stroke.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>artificial intelligence</td>
</tr>
<tr>
<td>CTA</td>
<td>CT angiography</td>
</tr>
<tr>
<td>DL</td>
<td>deep learning</td>
</tr>
<tr>
<td>ICH</td>
<td>intracerebral hemorrhage</td>
</tr>
<tr>
<td>LVO</td>
<td>large vessel occlusion</td>
</tr>
<tr>
<td>ML</td>
<td>machine learning</td>
</tr>
<tr>
<td>MRA</td>
<td>MR angiography</td>
</tr>
<tr>
<td>SAH</td>
<td>subarachnoid hemorrhage</td>
</tr>
</tbody>
</table>
Introduction

Stroke is common, hospitals are busy, delays=lost brain; diagnosis of the cause should be rapid so that appropriate treatment can start and give the patient the best chance of independent survival. Brain imaging is essential to differentiate ischemic from hemorrhagic stroke and stroke mimics. Furthermore, with advances in treatment options for specific patient subgroups, it is not enough just to identify ischemia or hemorrhage, since the size of the acute lesion, presence of other features (obstructed arteries, mass effect), and pre-stroke changes (leukoaraiosis, old infarcts, brain atrophy) also influence management. Most acute general hospitals assess several patients with suspected stroke each day, meaning that all steps in the process, including diagnosis, should be rapid, timely, efficient and accurate.

It takes many years to train a neuroradiologist, they are scarce in many countries, and serve many disease areas in addition to stroke. Vascular neurologists and stroke physicians learn to recognize early features of ischemic brain on scanning and major contraindications to reperfusion treatment. However, early ischemic changes on non-contrast CT can be subtle and complex, with serious implications hanging on their correct identification, fueling interest in ways to improve their recognition and quantification. Methods such as perfusion CT require post-processing to generate a diagnostic image, highlighting abnormalities such as thresholded tissue blood flow. Such ‘cleaned’ images may appear more user friendly and may facilitate rapid interpretation.

Alongside these longstanding pressures to reduce time and increase diagnostic accuracy, there have been substantial advances in computer vision and artificial intelligence (AI) technologies across all walks of life. At its most general, AI refers to use of computers to solve problems in ways that mimic human behavior; machine learning (ML) is of the key technology behind AI where computer algorithms learn from examples (‘ground truth’) without explicit programming which properties of the data are relevant for a given problem (‘feature selection’); and deep learning (DL) is a subset of ML that uses biologically inspired neural networks to learn abstract high-order features in any type of data without requiring predetermined inputs.

Medical imaging has been an obvious target for these developments. Several commercial CT and MR scan diagnostic software for stroke are now in use in many hospitals. Nonetheless, the major demand for accelerated diagnosis in acute stroke, the fascination with the latest imaging tools, and huge potential financial gains for industry, should not cloud the essential need to demonstrate that AI tools are accurate, are improving not impeding healthcare, and that the benefits outweigh the potential harms for patients.

We assess the current evidence for AI diagnostic imaging tools in stroke, commercially available or in clinical use, the motivations, expectations and work needed to underpin their implementation into clinical practice.
Review of Evidence

1. What could automated AI imaging technologies achieve in acute stroke diagnosis?

AI technologies could improve accuracy, speed and standardization of stroke diagnosis, particularly in low throughput Centres, and improve prognostication using quantitative measures of acute and chronic brain injury. AI tools could also improve workflows and communication along thrombectomy referral pathways, reduce time to treatment, improve clinical outcomes and save clinicians and radiologists time.

AI software could be most useful to assess the cause of stroke and its likely pathology:

1. Brain focal ischemia vs. intracranial hemorrhage vs. “stroke mimics” (migraine, seizure), the latter two requiring different management to ischemic stroke. Intracranial hemorrhages (brain hemorrhage, subarachnoid hemorrhage (SAH), subdural hematoma) can be visualized on CT and MRI with high accuracy. “Stroke mimic” is a clinical diagnosis based on the exclusion of brain ischemia and hemorrhage, noting that in patients with ischemic stroke symptoms, CT and MRI may appear normal, or show various degrees of reduced attenuation/angered signal intensity and/or swelling in the affected tissue.

2. Arterial pathology: site of acute embolic or thrombotic arterial occlusion, secondary features (e.g. collateral blood flow), and underlying pathologies (atherosclerosis, dissection, inflammation, vasospasm). This is, perhaps, the most important determinant of acute therapy decisions and related prognosis assessment. Thin slice non-enhanced CT, angiography (CTA, MRA, DSA) and arterial wall imaging may differentiate between embolic or local arterial disease and have an impact on secondary stroke prophylaxis.

3. Early ischemic brain tissue alterations: isolated tissue swelling due to autoregulatory vasodilation, tissue swelling with reduced attenuation (indicating early focal net water uptake or ionic edema triggered by critically low CBF <15 ml/100g/min, which cannot be tolerated by brain tissue for more than about 30 min), ion pump failure triggered by a CBF < 30 ml/100g/min with consequent neuronal dysfunction and cellular edema indicated by ADC decrease/increase of DWI signal intensity. The still limited understanding of ischemia-effected brain tissue alterations makes it a potential mistake that algorithms are trained to detect “brain infarction” on non-enhanced CT or DWI within 6 hours of symptom onset, since focal brain ischemia of up to 6-hour duration does not induce coagulation necrosis.

4. Pre-stroke changes: leukoaraiosis (or white matter hyperintensities), prior infarcts or hemorrhages, and brain atrophy are all associated with worse short-term (haemorrhagic transformation, dependency, death) and long term (dependency, recurrent stroke, death, cognitive impairment, dementia) outcomes.

Commonly used approaches for developing AI tools in stroke start with identifying attenuation or signal change that indicates ischemic tissue and its extent, adapted from visual rating tools. The Alberta Stroke Programme Early CT Score (ASPECTS) is a widely used visual rating tool used to help diagnose acute ischemic stroke and select patients for thrombolysis and thrombectomy. Table 1 lists vendors that currently provide 32 different commercially available AI software packages for ischemic stroke to assess a range of features including: ischemic...
tissue change (CT or MRI); hyperattenuated arteries (a surrogate for arterial thrombus); hemorrhage on CT; large artery occlusion on CTA; and ‘salvageable tissue’ from CT or MR perfusion imaging. ASPECTS is used by seven of eighteen companies currently offering commercial stroke AI software,\(^2\),\(^21\) despite several drawbacks of ASPECTS (not an independent outcome predictor,\(^18\) variable cut points, validity\(^20\)). Four companies provide comprehensive packages for handling non-enhanced, angiographic and perfusion imaging in one workflow (Brainomix, NICO.Lab, RapidAI and Viz.AI), while three companies combine assessment of non-enhanced and angiographic imaging only (Aidoc, Avicenna, Circle Neurovascular Imaging). Other companies assess different combinations or individual components of ischemic stroke or haemorrhagic stroke (Table 1).

2. What drives the increasing use of automated analysis technologies?

There is undoubtedly a real need to aid busy doctors at the hospital front door, and radiologists are also increasingly pressured. However, two decades after early thrombolysis trials, clinical awareness of acute stroke imaging features is much more established. Therefore, any AI diagnostic tools have to be exceptionally fast, easy to use, and accurate on ‘real world’ data (see Section 6) in order to add value. AI tools also require users to be trained in their proper use and interpretation.

Commercially, data and AI offer massive financial gains for successful products.\(^5\) Worldwide spending on AI was estimated at US$38 billion in 2019 and is predicted to rise to $98 billion by 2023.\(^22\) Investments in AI-based medical imaging companies in the USA reached US$1.17 billion between January 2014 and January 2019, doubling since 2012-2017, while the number of companies in the AI market had trebled,\(^5\) including new industries devoted to image classification (section 7). Major medical imaging manufacturers are incorporating AI tools into consoles to retain a marketing edge. AI requires data storage capacity and computing power: between Jan 2014 and Jan 2019, there were over US$435 billion-worth of cloud-based medical imaging transactions, indicating massive investments in these areas.\(^5\)

3. What clinically relevant testing should automated analysis undergo in acute stroke?

Currently, AI software for radiology can only be marketed in the European Economic Area (EEA) after achieving a CE mark (Conformité Européenne), indicating that the technology conforms with European health, safety, and environmental protection standards, and in the USA with FDA (Food and Drug Administration) approval. However, the clinical standards for achieving these certifications are low as compared with licensing a new drug. Both CE and FDA systems have different classes of approval depending on the perceived risk to the patient and to software users. Since radiology software have so far been designed to support rather than replace physicians, they require usually only Class I or II approval denoting low to medium risk. The ‘decision support’ labelling raises important questions about what happens when the clinician disagrees with the AI diagnoses, and who is to blame when one or other gets it wrong. In both territories, Class I approval is awarded without external scrutiny and companies self-declare these products as compliant. While Class II approval usually includes the submission of
evidence assessed by an independent body, companies can provide this evidence from their own internal testing without peer-review or publication. Indeed, a recent independent review of all CE-marked AI software for radiology in Europe found that 64 of 100 products had no published peer-reviewed evidence of efficacy. It is beyond the scope of this article to assess the depth of peer-reviewed evidence for every available software, so we focus on commercial products with the most citations according to recent reviews (Brainomix, RapidAI, and Viz.ai). Reporting guidelines and methodological standards for developing AI in medical imaging are available including SPIRIT-AI, CONSORT-AI, and checklists. Several societies have released their own guidelines and position statements. The FDA https://www.fda.gov/media/145022/download, and British Standards Institute (BSI) and Medicines and Healthcare Regulatory Agency (MHRA) https://standardsdevelopment.bsigroup.com/committees/50299208 provide standards. Few if any evaluations of AI software are meeting these standards, generally or in stroke. A recent systematic review of reporting quality of studies of ML in medical diagnosis found 28 includable studies but none mentioned a reporting guideline, few mentioned the distribution of disease severity or alternative diagnoses, most studies had a long delay between the reference standard and ML diagnoses, and in half of studies the population was of uncertain relevance to the clinical setting. Five of the 28 included studies concerned brain imaging none of which addressed stroke.

4. Accuracy of AI in stroke:

i) How accurate is AI software for differentiating acute ischemic from haemorrhagic stroke and stroke mimics, and on which imaging modalities?

Recent systematic reviews, narrative reviews, a review protocol and a pending combined analysis of nine large stroke trials show that most studies of AI have focused on ischemic not haemorrhagic stroke, and CT not MRI. The systematic reviews identified 20 (tissue and arterial changes) and 68 (non-contrast CT only) includable studies, but most studies were small, provided little documentation of the patient characteristics, recruitment or CT characteristics, and focused on comparing AI against feature detection by humans, not on clinical outcomes. Rates of failed scan processing were often omitted. Many methodological differences between studies precluded formal meta-analyses. All published studies and reviews focused on AI detection of ischemic stroke features including arterial obstruction, with only one pending study assessing if an AI software can differentiate ischemic from haemorrhagic stroke or mimics.

The extent to which AI software may be affected by patient-related (background brain changes or alternative brain pathology, movement and position during scanning, heart failure, metallic artefacts) and imaging-related (scanner manufacturer, acquisition parameters such as slice thickness, or CT gantry position) factors is beginning to emerge. These affect the likelihood of successful image processing, agreement with a reference
standard, and strength of associations with clinical outcome. However, most published evidence excluded difficult cases before analysis, while results of studies that retained difficult cases were often less positive.32,37,38

ii) How good are these technologies for identifying key features of acute ischemic stroke that are of prognostic importance?
Few studies assessed AI software-based diagnoses and clinical outcomes. Amongst commercially available AI tools, we reviewed the published literature for the three most established providers offering comprehensive imaging packages: Brainomix, RapidAI, and Viz.ai, (Table 2). We used published reviews7-9,21 updated by searching Pubmed for company and software names, and the vendor’s websites. We focus on studies with the largest test datasets (ideally >100 patients), and report diagnostic accuracy statistics for stroke feature detection.

Three studies assessed detection of tissue ischemia (all Brainomix, all retrospective, two independent of the company,32,39 total patients n=367),32,39,40 with sensitivities of 44-83% and specificities of 57-93% (Table 2). Six studies assessed detection of large vessel occlusion (LVO) (three vendors, all retrospective, one independent of the vendor,41 total n=2635)38,41-45 with sensitivities of 80-96% and specificities of 90-98%. Only one study each assessed hemorrhage detection46 and MRI diffusion or perfusion imaging.47 Compared with optimal circumstances, the agreement between each of these software and experts was poorer in patients with leukoaraiosis, old infarcts, or other parenchymal defects.7 This underscores the importance of evaluating AI tools in realistic and common clinical settings where patients are often older, have multiple conditions or delayed presentations,48 and not relying on results when tested in the simplified training scenarios common in public datasets.5,49

iii). On a population level, how many false positives or negatives might arise per 100 typical suspected strokes and what are the implications for patient outcomes?
Given the range of published sensitivity and specificity results for stroke feature detection by AI software above, we translate the results as follows, Figure 1. For every 100 patients assessed using these software:
• With ischemic stroke, ischemia will be correctly detected in 44-83 but missed in 17-56.
• Without ischemic stroke, ischemia will be incorrectly detected in 7-43.
• With LVO, occlusion will be correctly detected in 80-96 but missed in 4-20.
• Without LVO, occlusion will be incorrectly detected in 2-10.

iv). Have (any) automated, including AI, technologies that are proposed for use in stroke, undergone proper prospective randomised blinded outcome clinical trial assessment to determine impact on clinical outcomes or health economics?
Randomised controlled trials of AI technologies are scarce and mostly ongoing: a recent survey of trials’ registries and the literature identified only one RCT comparing DL with clinicians in medical imaging (on breast
We identified one ongoing multicenter RCT testing the impact of Viz LVO on stroke workflow and 90 day clinical outcome in 500 participants admitted with stroke and suspected LVO in the USA (Automated Detection and Triage of Large Vessel Occlusions Using Artificial Intelligence for Early and Rapid Treatment, ALERT, NCT04142879). Diagnostic accuracy is not a primary or specified secondary outcome in ALERT. Another ongoing multicenter RCT in China (GOLDEN BRIDGE II, NCT04524624) is testing AI identification of stroke on diffusion imaging plus decision support versus usual care in 21689 patients requiring secondary stroke prevention.

Three studies compared times to thrombectomy before and after introduction of RAPID\textsuperscript{41,50} or Viz LVO\textsuperscript{51}, reporting average reductions of 30 minutes to groin puncture; however, all were retrospective, two only report the small numbers of patients who all received thrombectomy, and before-after comparisons are unable to address many sources of bias.

5. How do stroke AI technologies compare with other medical AI technologies, particularly medical imaging AI, in terms of stage of development and quality and thoroughness of assessment?

Stroke AI is similar to other medical imaging AI – great hopes but important challenges for delivery into clinical practice. These challenges reflect data curation, model development, relevance to clinical practice, potential to introduce and amplify biases, the AI tool’s transparency, and evidence of accuracy, impact on outcomes and cost effectiveness meeting RCT evidence standards.

The quality, quantity, diversity, and provenance of the data used to train a ML model are critical to its utility in clinical practice. Many current papers describe ML models trained on one small dataset from one hospital\textsuperscript{52,53}, insufficient to be useful on the variety seen in clinical practice.\textsuperscript{53} Commercial ML models have similar issues, particularly “black box” models where the training is not described.

Datasets created as part of international challenges, e.g. the RSNA 2019 Brain CT Hemorrhage Challenge\textsuperscript{54} (874035 images, multiple institutions) are limited in diversity, accuracy and reproducibility.\textsuperscript{55} Over-tuning of software to particular datasets (overfitting) impedes generalization. Sometimes AI models identify confounders rather than target disease, e.g. hip fracture detection, where scanner model (emergency department) and “priority request” marker predicted fracture better than the imaging findings.\textsuperscript{56} ‘Explainable AI models’ may help to show the underlying features identified by a DL model to avoid ‘black box’ problems.

Bias emerging during the development of any automated system is common in AI studies,\textsuperscript{57} and can reflect issues with the underlying datasets or model development techniques. Numerous examples of biases have emerged, including those related to sex,\textsuperscript{58} race,\textsuperscript{59} and geography,\textsuperscript{60} which could inadvertently exacerbate underlying healthcare inequalities.\textsuperscript{51}
There are increasing suggestions that AI can ‘perform better than humans’ in medical tasks. A recent systematic review found 10 RCTs testing DL versus clinicians (2 completed, 8 ongoing), only one of which was in medical imaging (breast ultrasound, ongoing). In contrast, they found 81 non-randomised comparisons (nine prospective, six relevant clinical environment). The AI was said to be better than or comparable to the clinician in 47% of the 81 studies. However, development and testing often used the same dataset, had small numbers of human comparators (e.g. five), most studies had high risk of bias, and few adhered to reporting standards.

Another comparison of the diagnostic accuracy of DL versus clinicians identified 82 studies in which it was possible to calculate accuracy in 69. Mean sensitivity was 79.1% (range 9.7–100%), specificity was 88.3% (range 38.9–100%), but many studies did not compare the DL and clinicians on the same data, did not externally validate their results, or report their methods adequately. Amongst the 14 studies with external validation that tested DL and clinicians on the same sample, the pooled sensitivity was 87.0% (95% CI 83.0–90.2%) versus 86.4% (79.9–91.0%), and pooled specificity was 92.5% (85.1–96.4%) versus 90.5% (80.6–95.7%) for DL versus clinicians respectively. Of note, there were no studies of AI in stroke amongst the 82 studies, and only two studies concerned brain imaging (one MRI in dementia, one CT in head trauma).

6. What evidence is there that AI technologies will meet the needs of users, including community practitioner, neurologist, radiologist, and the patient?

Do AI tools reduce ‘door to needle’ time? Or might AI tools worsen treatment delays, or deny some patients effective treatments? There are no completed prospective randomised trials of the impact of stroke AI tools on workflows or clinical outcomes, only before-after evaluations (see 4iv above). It is common in hospitals for new digital systems to slow, not accelerate, workflows. Attention of AI should focus on improving routine medical workflows including image management and electronic case records to reduce time wasted, improve information content and diagnostic utility. Some AI tools perform tasks that are not that helpful, provide clinically irrelevant measures, or operate slower than a seasoned user of existing medical computing systems, delaying uptake of AI into general radiology.

Different algorithms may give different results. Comparison of three AI tools for ASPECTS scoring showed the highest correlation between the expert read and Brainomix (ICC=0.871 (0.818, 0.909), p <0.001) but comparable area-under-curve between the AI applications and expert consensus (Brainomix: AUC 0.759 (0.670–0.848), p <0.001; Frontier V2: AUC 0.752 (0.660–0.843), p <0.001; RAPID: AUC 0.734 (0.634–0.831), p <0.001). AI software may help less experienced doctors interpret acute stroke CT scans, e.g. use of Syngo.via Frontier ASPECT Score Prototype improved the correlation between junior radiologists and the reference standard from good (r=0.680) to excellent (r =0.852) in one small study. Apparently good performance at group level may mask important variation at the individual level. Amongst 12 ML and seven statistical models to predict cardiovascular disease risk using data from 3.6 million patients, the models had similar population level performance (C statistics
of about 0.87), but varied widely in their prediction of individual risks particularly at higher risks, and compared with the risk predicted by a reference model, about 60% of patients would have been classed as lower risk by another model.

AI tool evaluation is usually restricted to single technical measures in controlled settings that only indirectly relate to the intended clinical tool use. However, the clinical need is rarely ‘raw classification’ but rather diagnostic or therapeutic decision support. DL tools are notoriously sensitive to changes in input characteristics, and not customarily stress-tested across different scanners, parameters, clinics, patient groups, pre-processing tools, etc., further hindered by the black-box nature of DL methods and commercial confidentiality. The benefits and challenges of introducing computational innovations into existing clinical ecosystems remains sparsely assessed; AI software may require specific tailoring to suit different settings and institutions.

There is little participant involvement in development of AI software despite concerns. Few people want to receive a terminal diagnosis from a robot, or want major treatment decisions for a life-threatening disease (like stroke) to be based primarily on AI-determined scan characteristics, particularly when these differ between AI software. Increasing availability of AI diagnostic tools including front line use by less experienced doctors risks subverting medical judgement through inflexible application of easy-to-derive threshold values - e.g. treat if ≤70ml core, not if >70 ml. Treatment decisions with the very powerful reperfusion therapies now available for stroke must consider the whole patient and not place inappropriate weight on perfusion maps that only represent brief snapshots in time of a highly dynamic disorder, particularly when use of a different software is very likely to give different results.

7. What could improve translation of AI technologies into benefits for patients and health care systems?

There is a gulf between AI software developers and the intended clinical uses of the software, and a need for consensus-based, interdisciplinary and comprehensive scoping of user needs and constraints to guide effective development of AI tools. The accuracy of an AI tool depends on its data input. AI developers highlight the need for ‘ground truth’ for training the software, meaning images where relevant features have been demarcated, often by hand, and in large datasets. While large collections of medical imaging data are increasingly available, manual annotation is a massive, time consuming task and must be done by humans, therefore not many large annotated datasets exist. Some companies are outsourcing the work to low paid workers in low income countries. The market for data-labelling services may triple to US$5billion by 2023, with companies like ‘Mechanical Turk’ (owned by Amazon) providing freelancers ready to perform ‘micro-tasks’ like tagging images, or ‘Hive’ which runs online data labelling ‘games’ where operators earn money for labelling features, which questions the reliability of the ground truth thereby derived.
Can we accelerate reliable, representative and diverse dataset availability, and, is ‘ground truth’ really essential, or could correlated variables like clinical outcomes be used instead? AI tools could be even more valuable if they could discover novel features or markers of severity, or predict clinically-relevant outcomes and treatment response, and thus improve clinical management. Table 3 lists important factors, including more accessible large-scale data, standardization of pre-processing pipelines, sharing open source codes, adoption of guidelines for reporting of AI development, closer working between ML/DL developers and clinicians, and better standards for evaluating AI tools against relevant clinical outcomes, control of confounders, correlates and colliders that impede AI performance.
Discussion and Conclusions

While AI tools hold great promise in stroke, much more work is required to demonstrate their clinical value and cost-effectiveness to patients, doctors, and health-care providers. AI development requires more focus on multidisciplinary teams including AI experts, IT experts, radiologists and strokologists listening to each other carefully. Currently, without a more cohesive multidisciplinary effort, it seems that the stroke-AI is at risk of Verschlimmbesserung, meaning an 'attempted advance without improvement, or even with worsening'. Imaging and AI image analysis should demonstrate clearly through objective, relevant and reliable evidence that improve patient management and clinical outcomes.

‘Digital’ methods complement but cannot replace the human touch in medicine. Patients should be actively involved in the design and evaluation of AI tools usage, since patient groups are clearly very wary of too much unregulated use of computers in clinical decision-making.

Not all measurement has value. Rather, consider what features would be treatment relevant to detect and quantify. Imaging findings in stroke patients help us chose the most effective treatment. Spontaneous brain hemorrhage requires angiography to find and treat a vascular malformation or aneurysm. Exclusion of hemorrhage but thrombus within the MCA requires immediate thrombectomy. Stroke imaging is complex, not a ‘single feature’ process, and perhaps a more difficult place to initiate AI tool development than it might seem superficially. Low ASPECTS is an independent predictor of poor outcome, but patients may still benefit from treatment, therefore reducing information to a single binary variable such as ASPECTS score would seem to be a retrograde step.

Can costs of AI tools be realistic? Currently one typical commercial AI software costs around US$47,868 for one hospital for one year in the UK, equivalent to about a third of a hospital consultant’s salary, and seems an unreasonable amount of money for something which only identifies a few features in one disease, should only be used by an experienced medic, and thus does not replace anything or anyone, and has a limited evidence base.

The essential next steps are first, to be aware of the limitations where commercial AI tools are in use, and second, to obtain reliable evidence of benefits versus harms of imaging AI tools’ performance. This would best be tested in large scale randomised trials, to minimize bias, and in the clinical settings in which they will be used to ensure applicability to real world clinical practice. It is no longer appropriate to show only diagnostic accuracy on selected retrospective datasets, without reporting the failures. We need to see how AI tools work in clinical practice, how they integrate into patient care, and if and how much they are beneficial, through providing evidence that they change management for the better, improving outcomes, and are cost effective.
References


20. Suss RA. ASPECTS, the mismeasure of stroke: A metrological investigation. *OSF Preprints*. 2019;31st Dec


71. Mittelman M, Markham S, Taylor M. Patient commentary: Stop hyping artificial intelligence—patients will always need human doctors. *BMJ.* 2018;363:k4669


<table>
<thead>
<tr>
<th>Company (Country)</th>
<th>Product</th>
<th>Automated functions (imaging modality)</th>
<th>CE Approval (Class) for EU Marketing</th>
<th>FDA Approval (Class) for US Marketing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aidoc (Israel)</td>
<td>Briefcase</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Aidoc LVO</td>
<td>Detect LVO (CTA)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Avicenna.AI (France)</td>
<td>Cina-ASPECTS</td>
<td>Provide ASPECTS (CT)</td>
<td>Yes (I)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Cina-ICH</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Cina-LVO</td>
<td>Detect LVO (CTA)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Brainomix (UK)</td>
<td>e-ASPECTS</td>
<td>Provide ASPECTS, detect dense MCA, detect hemorrhage (CT)</td>
<td>Yes (Ia)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>e-CTA</td>
<td>Detect LVO, provide collateral scoring (CTA)</td>
<td>Yes (Ia)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>e-CTP</td>
<td>Process perfusion data (CTP)</td>
<td>Yes (Ia)</td>
<td>No</td>
</tr>
<tr>
<td>Cercare Medical (Denmark)</td>
<td>Cercare Stroke</td>
<td>Process perfusion data (CTP, MRP), detect ischemic lesions (MRI)</td>
<td>Yes (Ia)</td>
<td>No</td>
</tr>
<tr>
<td>Circle Neurovascular Imaging (Canada)</td>
<td>StrokeSENS</td>
<td>Provide ASPECTS (CT), provide collateral assessment, detect LVO (CTA)</td>
<td>Yes (II), does not include collaterals</td>
<td>Yes (I)*</td>
</tr>
<tr>
<td>Deep01 (Taiwan)</td>
<td>DeepCT</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>General Electric (USA)</td>
<td>Stroke VCAR</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>CT Perfusion 4D</td>
<td>Process perfusion data (CTP)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Icometrix (Belgium)</td>
<td>Icobrain CVA</td>
<td>Process perfusion data (CTP)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Infervision Med Tech (China)</td>
<td>InferRead CT Stroke.AI</td>
<td>Provide ASPECTS, detect hemorrhage (CT)</td>
<td>Yes (Ia)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>JLK Inc. (South Korea)</td>
<td>JBS-01K</td>
<td>Detect ischemic lesions (MRI), provide stroke classification</td>
<td>Yes (I)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>JBS-04K</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (I)</td>
<td>No</td>
</tr>
<tr>
<td>Keya Medical (China)</td>
<td>CuraRad-ICH</td>
<td>Detect hemorrhage (CT)</td>
<td>No</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Company</td>
<td>Software/Tool</td>
<td>Functionality</td>
<td>CE Risk</td>
<td>FDA Risk</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>MaxQ.ai (Israel)</td>
<td>Accipio IX</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>NICO.Lab (Netherlands)</td>
<td>StrokeViewer</td>
<td>Provide ASPECTS, detect hemorrhage (CT), LVO detection, provide collateral assessment (CTA), process perfusion data (CTP)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Olea Medical (France)</td>
<td>Olea Sphere</td>
<td>Process perfusion data (CTP)</td>
<td>Yes (Ia)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Qure.ai (India)</td>
<td>qER</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (Ia)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>RapidAI (USA)</td>
<td>Rapid</td>
<td>Process perfusion data and provide flow dynamics (CTA, CTP), detect DWI lesions (MRI)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Rapid ASPECTS</td>
<td>Provide ASPECTS (CT)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Rapid ICH</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Rapid LVO</td>
<td>Detect LVO (CTA)</td>
<td>Yes (I)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Siemens (Germany)</td>
<td>syngo.CT ASPECTS</td>
<td>Provide ASPECTS (CT)</td>
<td>Yes (not declared)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>syngo.CT Neuro Perfusion</td>
<td>Process perfusion data (CTP)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Viz.ai (USA)</td>
<td>Viz ICH</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Viz LVO (ContaCT)</td>
<td>Detect LVO (CTA)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td></td>
<td>Viz CTP</td>
<td>Process perfusion data (CTP)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
<tr>
<td>Zebra Medical Vision (Israel)</td>
<td>HealthICH</td>
<td>Detect hemorrhage (CT)</td>
<td>Yes (not declared)</td>
<td>Yes (II)</td>
</tr>
</tbody>
</table>

**Notes:** Both CE and FDA classification use 3 classes depending on risk to the patient and/or user according to the intended use of the device: I=low risk, II=medium risk, III=high risk. Both classification systems have a more stringent process for classifying higher risk devices. For CE, I can be self-certified by manufacturer, II&III require audit of validation results by a notified body. For FDA, I&II require 501k (prove equivalence to device already approved for marketing, or de novo assessment if novel) while III requires premarket approval (PMA) including software validation results.

* Approved for data transfer only, not automated processing.

Details extracted from publicly available [EU](https://ec.europa.eu) and [FDA](https://www.fda.gov) data and vendor websites, correct to 31st Aug 2021, modified 21 Feb 2022.
Table 2. Accuracy of three commercially available AI-based software in stroke.

<table>
<thead>
<tr>
<th>Company</th>
<th>Software</th>
<th>Detection of ischemic brain injury</th>
<th>Detection of LVO</th>
<th>Detection of hemorrhage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CT</td>
<td>MRI/CTP</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brainomix</td>
<td>e-ASPECTS</td>
<td>Retrospective, 132, 44-45%, 91-93%, follow-up CT&lt;sup&gt;10&lt;/sup&gt; Retrospective*, 119, 83%, 57%, follow-up CT&lt;sup&gt;32&lt;/sup&gt; Retrospective*, 116, 75%, 73%, experts with all data including follow-up CT or MRI.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>e-CTA</td>
<td></td>
<td>Retrospective, 301, 84%, 96%, experts with all data including follow-up imaging.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>e-CTP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rapid</td>
<td>Rapid ASPECTS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rapid ICH</td>
<td>Retrospective, 63, 100%, 91%, experts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rapid LVO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viz.ai</td>
<td>Viz ICH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Viz LVO (ContaCT)</td>
<td>Retrospective, 1167, 81-82%, 90-96%, experts&lt;sup&gt;38&lt;/sup&gt; Retrospective, 163, 96%, 94%, NS&lt;sup&gt;45&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Viz CTP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUMMARY</td>
<td>367 patients, 1 software</td>
<td>Sensitivity = 44-83% Specificity = 57-93%</td>
<td></td>
<td>2635 patients, 3 different software</td>
</tr>
</tbody>
</table>

Note: Results are [study design, n, sensitivity, specificity, reference standard], unless otherwise stated. * Indicates study conducted independent of company. Shaded boxes indicate no suitable papers were identified or expected. NS = Not specified.
<table>
<thead>
<tr>
<th>Need</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>More accessible large-scale data repositories:</td>
<td>RSNA (Radiology Society of North America) Head CT Challenge: ICH detection (&gt;800,000 scans); ASFN (American Society of Functional Neuroradiology) Head CT Challenge: CT pathology at different ages (<a href="https://aichallenge.asfnr.org/">https://aichallenge.asfnr.org/</a>); ISLES (Ischemic Stroke Lesion Segmentation) Challenge (n=75); large trials e.g. Third International Stroke Trial (IST-3) (3035 pts, &gt;10,000 scans) UK Biobank (ukbiobank.ac.uk); HDR UK (Health Data Research)</td>
</tr>
<tr>
<td>Guidelines to standardize pre-processing pipelines</td>
<td>Preprocessing steps: reading DICOM data, converting DICOM to other formats, brain extraction from skull, defacing, registration, etc RSNA 2019 ICH detection challenge shows many common problems that can occur e.g., under-labelling data, human errors, imbalanced classes, inappropriate de-identification and anonymization across data sources.</td>
</tr>
<tr>
<td>Guidelines and standards for reporting of machine learning models</td>
<td>Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis - statement specific to Machine Learning (TRIPOD-ML), Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)-AI and Consolidated Standards of Reporting Trials (CONSORT)-AI.</td>
</tr>
<tr>
<td>Collaboration between ML practitioners and all relevant disciplines</td>
<td>Physicians and radiologists benefit from becoming familiar with basic concepts in machine learning. Machine Learning methods must be developed by those familiar with details and context of the real clinical settings including decision triage, practicalities of the specific patient population, typical medical imaging, image interpretation, definition of treatment relevant imaging finding, and data preparation.</td>
</tr>
<tr>
<td>Testing and validation protocols</td>
<td>Should move beyond technical performance evaluation (e.g., ROC, accuracy, sensitivity /specificity) to determine the clinical value of a system, given limited annotator availability, experience, accuracy and interest. The limits of testing criteria used in international challenges are evident.</td>
</tr>
<tr>
<td>Attuned algorithmic development</td>
<td>Move beyond classifiers: methods must control for confounders, correlates and colliders that introduce bias and produce non-robust methods that collapse with potentially dangerous consequences when deployed in different real settings.</td>
</tr>
<tr>
<td>Open-source AI code and data</td>
<td>Proprietary code impedes replication, reproducibility, clinical validation: it is difficult and costly. Commercial development based on ill-established methods that the community cannot verify in clinical settings risks reputational damage.</td>
</tr>
</tbody>
</table>
Figure 1. Potential clinical implications of AI software use for stroke feature detection per 100 patients assessed, derived from data in Table 2.

Note: Orange circles indicate the overlapping range of values provided by different studies.