Neuroradiology (2022) 64:851–864
https://doi.org/10.1007/s00234-021-02890-w

REVIEW

The R-AI-DIOLOGY checklist: a practical checklist for evaluation of artificial intelligence tools in clinical neuroradiology

Sven Haller[1,2,3,4] · Sofie Van Cauter[5,6,7] · Christian Federau[8,9] · Dennis M. Hedderich[10] · Myriam Edjlali[11,12]

Received: 25 October 2021 / Accepted: 21 December 2021 / Published online: 31 January 2022

Abstract

Artificial intelligence (AI)-based tools are gradually blending into the clinical neuroradiology practice. Due to increasing complexity and diversity of such AI tools, it is not always obvious for the clinical neuroradiologist to capture the technical specifications of these applications, notably as commercial tools very rarely provide full details. The clinical neuroradiologist is thus confronted with the increasing dilemma to base clinical decisions on the output of AI tools without knowing in detail what is happening inside the "black box" of those AI applications. This dilemma is aggravated by the fact that currently, no established and generally accepted rules exist concerning best clinical practice and scientific and clinical validation nor for the medico-legal consequences in cases of wrong diagnoses. The current review article provides a practical checklist of essential points, intended to aid the user to identify and double-check necessary aspects, although we are aware that not all this information may be readily available at this stage, even for certified and commercially available AI tools. Furthermore, we therefore suggest that the developers of AI applications provide this information.

Keywords AI · Artificial intelligence · Neuroradiology · Brain · MRI · CT

Abbreviations

AD	Alzheimer dementia
ADNI	Alzheimer Disease Neuroimaging Initiative
AI	Artificial intelligence
CADASIL	Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy
CNN	Convolutional neural networks
CT	Computed tomography

✉ Sven Haller
sven.haller@me.com

[1] CIMC-Centre d'Imagerie Médicale de Cornavin, Place de Cornavin 18, 1201 Geneva, Switzerland

[2] Department of Surgical Sciences, Radiology, Uppsala University, Uppsala, Sweden

[3] Faculty of Medicine, University of Geneva, Geneva, Switzerland

[4] Department of Radiology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, People's Republic of China

[5] Department of Medical Imaging Ziekenhuis, Oost-Limburg Genk, Schiepse Bos 6, 3600 Genk, Belgium

[6] Department of Radiology, University Hospitals Leuven, Herestraat 49, 3000 Leuven, Belgium

[7] Division of Medicine and Life Sciences, Department Neurosciences, Hasselt University, Campus Diepenbeek, Agoralaan Building D, 3590 Diepenbeek, Belgium

[8] AI Medical AG, Goldhaldenstr 22a, Zollikon CH-8702, Switzerland

[9] Faculty of Medicine, University of Zürich, Pestalozzistrasse 3, CH-8032 Zurich, Switzerland

[10] Department of Neuroradiology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany

[11] Department of Radiology, APHP, Hôpitaux Raymond-Poincaré & Ambroise Paré, DMU Smart Imaging, GH Université Paris-Saclay, U 1179 UVSQ/Paris-Saclay, Paris, France

[12] Laboratoire d'imagerie Biomédicale Multimodale (BioMaps), Université Paris-Saclay, CEA, CNRS, Inserm, Service Hospitalier Frédéric Joliot, Orsay, France

DICOM	Digital Imaging and Communications in Medicine
DL	Deep learning
GDPR	General Data Protection Regulation
MPRAGE	Magnetization prepared rapid gradient-echo
MR	Magnetic resonance
MS	Multiple sclerosis
PHI	Personal health information
RANO	Response assessment in neurooncology
SaMD	Software as a medical device
SL	Supervised learning
SLE	Systemic lupus erythematosus
SVD	Small vessel disease
USL	Unsupervised learning

Introduction

Artificial intelligence (AI) and big data analyses are considered the driving forces of the fourth industrial revolution and many hopes have been raised regarding the application of AI in healthcare, particularly medical imaging [1]. The complex, multidimensional, and often multimodal information in medical images is a challenging, yet attractive starting point for AI-based technology, e.g., aiming at automation of image analysis and interpretation. It is the potential shift from a subjective and qualitative image assessment by the radiologist susceptible to fatigue or distraction to a more objective and quantitative approach by the “machine” that opens the horizon for an improved diagnostic process.

After an exponential increase of scientific publications regarding AI in (neuro)radiology in the last decade, AI-based applications for clinical use have been developed and cleared for clinical purposes [2–4]. However, their clinical implementation is still lagging behind, which may—at least in part—be due to a lack of training and knowledge of neuroradiologists with respect to these emerging tools and implications of their use in clinical practice [5]. A study by Huisman et al. demonstrated that radiologists who had limited knowledge regarding AI were more likely to fear AI. However, those with knowledge of AI were more likely to view the technology positively [6]. This underscores the importance of training curricula and continuing education of medical professionals in order to facilitate the adoption of AI in clinical practice. Providing clear and transparent product-specific information to the radiologist might additionally help adoption of such techniques in clinical routine.

Adopting AI-based tools into clinical practice could potentially help to mitigate an important dilemma in neuroradiology. On the one hand, medicine in general and especially radiology has become more and more complex, leading to the need for detailed reports of subspecialized neuroradiologists. On the other hand, the neuroradiologist is confronted with an ever-increasing number of radiological exams [7] and oftentimes increased financial pressure due to reduced reimbursement in most countries. Furthermore, the complexity of AI-based tools has also increased, which makes it challenging for the clinical neuroradiologists to vet applications for practical clinical use in a time-efficient, yet thorough way.

To fill this gap, we want to give a practical overview of ten essential checkpoints, which will help neuroradiologists to evaluate an AI tool for clinical use in neuroradiology.

Fundamentals of AI tools

In the following, we explain and define some of the main concepts and terminology, which are needed to understand modern and clinically available AI tools for neuroradiology, prior to their use in our daily routine Focusing on the presented clinical checklist, we did not aim to replicate already existing excellent review articles concerning in particular deep learning, included here for examples [8, 9].

Artificial intelligence, machine learning, deep learning

The term artificial intelligence (AI) has been used abundantly in the last years but was initially coined in the 1950s, describing an operation performed by a machine or an algorithm to solve a task that would otherwise have required human intervention [10]. Thus, it can be considered a very broad term, comprising, for example, robotics, natural language processing, and machine learning. Machine learning, in turn, makes use of statistical approaches to learn from data and evidently aims at making predictions with respect to an outcome of interest. Thanks to increasing computational power and algorithmic developments, machine learning has more and more been applied to the analysis of medical imaging. One of these latest developments is called “deep learning” (DL), which is based on different algorithms. The so-called convolutional neural networks (CNN) are currently frequently used algorithms in the domain of imaging, yet alternative algorithms exist and new algorithms will likely be developed. CNNs consist of several network layers with one (or in most cases many more) hidden or “deep” layers. The number of deep layers correlates with the computational complexity that an algorithm can tackle. Increased computational complexity in deep CNNs has enabled them to directly learn from image information as an input without the need for prior definition of task-specific, pertinent image information (feature engineering).

Supervised versus unsupervised AI tools

AI can be classified along several axes. One very important distinction is between supervised and unsupervised tools.

Most current AI tools use a supervised approach. This means that a "ground truth" labeled/annotated dataset exists with respect to the problem that needs to be solved by the algorithm, e.g., detection or segmentation of confirmed brain tumors or MS lesions or flagging intracranial hemorrhage. In general, special medical expertise is needed for annotating the dataset, which requires medical professionals to perform ground truth labeling. This can be rather time-consuming and tedious due to the generally large amount of input data needed. An important advantage of supervised learning is that it incorporates medical knowledge and that in general a relatively smaller dataset is sufficient. An important disadvantage is that supervised learning predominantly produces "narrow" AI that directly addresses a medical question, i.e., an algorithm which can solve one particular task only (e.g., stroke or hemorrhage detection). However, this algorithm will be weak in case of unexpected input data, for example, detecting ongoing seizure as stroke mimic or a brain tumor manifesting brain perfusion changes (Figs. 1, 2).

An unsupervised tool is about the opposite and may be an approach to tackle the wide variety of medical imaging diagnoses and tasks. In an unsupervised tool, there is no prior knowledge or labeling of datasets/cases. The advantage is that no expert labeling of the training data is required, and the tool might detect unexpected or novel diseases/disease patterns. The disadvantage is that in

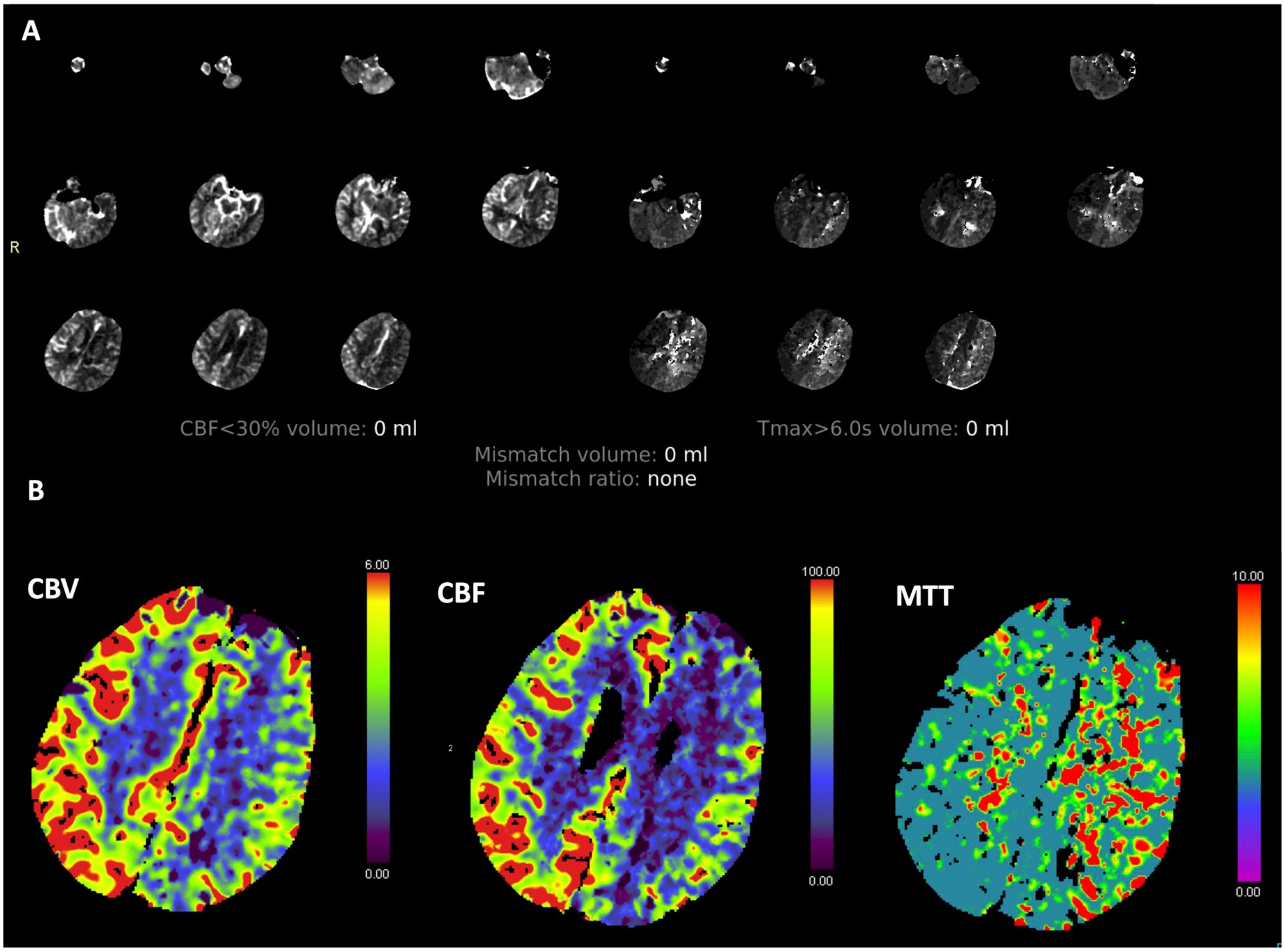

Fig. 1 Male patient, 67 years old, presented in the emergency room with disturbed consciousness and a left hemiparesis, last well seen the night before. Acute stroke workup was initiated with unremarkable non-contrast head CT and CT angiography (not demonstrated). CT perfusion analyzed by AI-based software did not show a core (defined as cerebral blood flow (CBF) values lower than 30%) nor penumbra (defined as a time to the maximum of the residue function (Tmax) over 6 s) (panel **A**). Visual inspection of the perfusion maps demonstrates a general cortical hyper perfusion on cerebral blood volume (CBV) and CBF maps in the right hemisphere and decreased mean transit time (MTT), compatible with ongoing non-convulsive seizure. In this example, the AI tool provides a true negative diagnosis for stroke, but misses the significant diagnosis of non-convulsive seizure

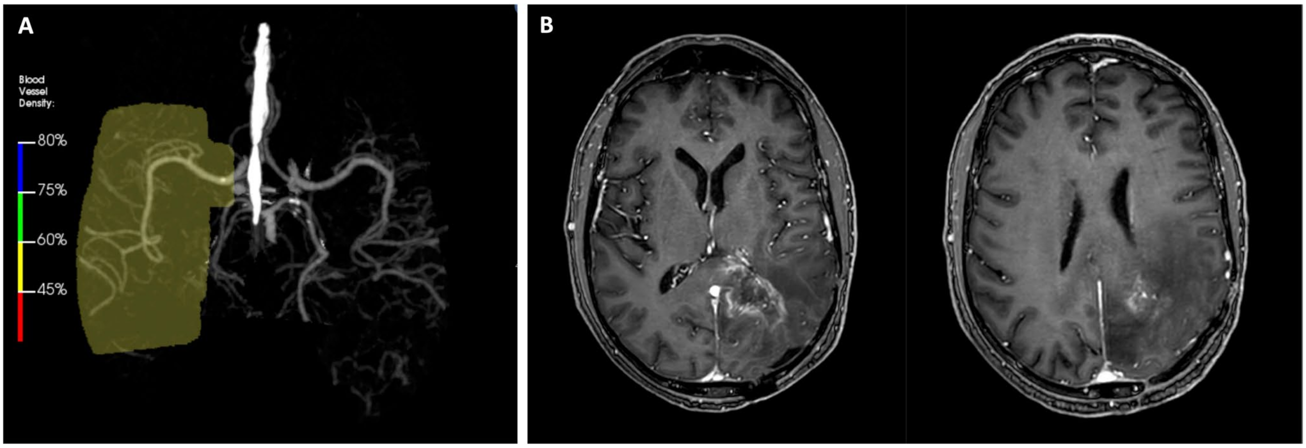

Fig. 2 Male patient, 54 years old, presented in the emergency room with a subacute onset of right hemiparesis. Stroke workup was initiated. Automatic stroke detection software of CT angiography detected an asymmetry in arborization of the arteria cerebri media with reduced amount of detected vessels on the right side (panel **A**). This did not match with his symptomatology of right-sided hemiparesis. Inspection of the patient's medical file demonstrated that he was known with a treated low-grade glioma in the left parietal region. His clinical presentation was explained by tumor progression to a glioblastoma with a distinct neovascularisation (panel **B**), explaining the asymmetry by a relative dominance of vessels on the left side. In this example, the AI tool provides a false positive diagnosis for stroke

general even larger datasets are required for training (big data). Moreover, the tool will detect clusters/patterns in the data, yet this does not necessarily directly correspond to a given medical question, disease or lesion type. Consequently, it must be shown post hoc whether a discovered pattern corresponds, e.g., to a specific brain tumor.

Why it is important

When the radiologist must base clinical decisions or can choose between several AI tools for the same clinical purpose, it is important to be aware how the underlying method of the AI tool might impact the clinical results.

Number of input datasets versus number of features—feature selection

In general, CT and MR input datasets are very large. For example, a simple 3D T1 brain scan of 256 × 256 × 176 voxels includes 11,534,336 input voxels. This is much higher than the typical number of participants available used for training and testing of AI tools (typically in the range of tens or at most hundreds). In the terminology of AI tools, input datapoints are considered features. To avoid this imbalance between input features and cases, and to improve the AI tool, AI tools may include a feature preselection or feature pre-processing including, e.g., tools that detect edges or shapes or approaches such as independent component analyses to reduce the number of input features.

Why it is important

In addition to the type of AI method per se discussed above, also the type of input feature pre-processing may influence the clinical results of AI tools.

Assessment of algorithmic performance: the importance of datasets and how they are created

In order to critically assess algorithmic performance, it is crucial to understand how the data is being used and assigned to certain distinct datasets. The terminology of those datasets may be confusing due to the sometimes inconsistent nomenclature across different areas or publications. In the following, we explain the most commonly used terminology of datasets, which is also used in this review article.

Training dataset The main dataset used to train the AI tool. As mentioned above, this dataset can be labeled/annotated (notably for supervised models) or not (notably for unsupervised models).

Validation dataset/tuning dataset Usually, an additional dataset is used during the training procedure to monitor algorithmic performance and adapt algorithmic parameters including hyperparameters if necessary. The term validation dataset is potentially confusing since it is not meant to "validate" the algorithm, neither in a technical nor in a medical sense, but rather to tune its performance during the training process. Therefore, this dataset is sometimes also referred to as "tuning dataset."

Testing dataset Dataset used to test the performance of the AI tool after training and tuning on unseen, new data. Ideally, the testing dataset should be rigorously separated and distinct from the data used during the algorithm development in order to avoid several degrees of "data leakage." However, this depends on the used validation approach, usually described as cross-validation, internal validation or external validation.

Cross-validation As oftentimes the number of datapoints/cases is limited, many AI tools use a cross-validation approach for testing of the tool, typically tenfold or fivefold cross-validation. In a tenfold approach, the training dataset is split into 10 parts ("folds"); then, 9 parts are used for training and tuning and the remaining part is used for testing. This is repeated 10 times until each part was used once for testing. While this approach is scientifically correct, it has to be noted that the same dataset is used for model selection or hyperparameter tuning and cross-validation; thus, the ability to assess algorithmic performance on truly unseen data is impaired when cross-validation is used.

Internal validation Following this approach, data used during training/validation and testing has been clearly separated, i.e., split before experiments have been started. However, data stem from the same source, e.g., have been acquired in the same center or on the same scanner. Using this approach, it is important to prevent subtle causes of data leakage, e.g., by including a patient's scan in the training procedure and a follow-up scan from the same patient in the testing dataset.

External validation This approach represents the most rigorous way of assessing algorithmic performance on unseen data. As in internal validation, data have been split before starting the experiments and testing data have never been seen by the algorithm during training and validation. In addition, testing data stems from a different source, e.g., a different medical center or hospital.

Reference dataset The reference dataset is not specific to AI tools, but a general term. The reference dataset defines the range of normal values, for example, the classic growth curves for children of brain volumes for MR volumetry. This dataset may or may not be identical or overlapping with the training dataset, i.e., it is in principle possible to train the AI tool using the reference dataset. However, it is also possible to train the AI tool using a completely or partially different training dataset, and then apply the trained AI tool to the reference dataset to create the normative values.

Why it is important

Vetting the chances how useful an algorithm might be in clinical practice, we need to estimate how well it has been shown to work on new, unseen data. In order to do so, we need to assess how data was used and whether there could be any source of data leakage during algorithm development and testing, since this might impact the output of the AI tool and thus potentially also the clinical decision might have lower reported performance values than another tool using a cross-validation approach, yet in clinical use the true performance might actually be higher.

Single versus multiple scanners/sites

As a general rule, one can say that it is already difficult to generalize even basic data parameters (e.g., basic T1w or T2w volumes) between different scanners/sites. The more complex and advanced the dataset (e.g., complex diffusion or perfusion MR parameters), and the smaller the expected effect size, the more difficult yet at the same time the more important it is to make sure that data and results can be generalized between different scanners/sites.

Why it is important

If an AI tool is developed and tested on a single CT or MR scanner from a single site, the results will in general be more optimistic and it will be less evident to generalize this tool in different CT or MR systems at different sites—as compared to an AI tool which was tested and validated on datasets from multiple CT or MR systems and multiple sites.

Single versus multi contrast

Most currently available AI tools use a single MR sequence as input. However, in clinical routine, in almost all MR exams, multiple sequences/contrasts are acquired and analyzed by the radiologist.

Why it is important

The typical scientific evaluation of an AI tool compares the AI tool using a single MR sequence versus the human reader using the same dataset—as this is the direct comparison from a scientific perspective. However, in clinical routine, the radiologist analyzes and integrates multiple MR sequences. Forcing the radiologist to analyze only one MR sequence is not the typical situation and will in most cases reduce the accuracy of the human rater.

The 10-point checklist

Checkpoint 1: Disease application—domain shift

Most current applications of AI for medical imaging rely on supervised learning (SL) approaches. This means that an algorithm needs to learn from "training data" which were labeled according to a ground truth or reference standard. As the purpose of clinical AI tools is to provide accurate results at an individual patient level, generalizability to new, unseen data is of utmost importance. The performance of AI algorithms may be hindered by the so-called domain shift, which is rooted in differences between training data used during algorithm development and test data used either in a clinical study or in clinical practice [11]. This incongruity can be manifold and includes, e.g., disease characteristics of a patient group used for algorithm development [12]. Since perfect generalization of AI-based algorithms has not been achieved yet, it is important to keep differences between patient characteristics used during the development and faced when applying the algorithm as low as possible. It is therefore essential to vet evidence that a given AI tool was approved for use in the desired disease.

Why it is important

Newly developed AI algorithms are capable of directly learning from image information. Thus, the training data of the algorithm influence the domain and type of images on which it can be applied.

Neuroradiology examples

Automated segmentation of WMLs on FLAIR imaging has been a popular application for algorithms and fills a clinical need since counting WML is a notoriously tedious and time-consuming task prone to human error, for example, in patients with multiple sclerosis (MS). However, domain shift with respect to the patients' disease may occur in clinical practice if an algorithm developed on training data from MS patients is being used in WMLs of other type of distribution, such as small vessel disease (SVD) or less common diseases as CADASIL (cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy), SLE (systemic lupus erythematosus), or Susac syndrome. For example, it might be that an AI tool for segmentation of WM lesions trained and optimized for MS also works for SLE, but we cannot assume this as a given, and therefore, such disease or domain shifts should be re-evaluated (Fig. 3).

Checkpoint 2: Preselection of cases and reference dataset

We have previously seen that domain shift occurs due to differences in patients' diseases between training data and real-world application. The distribution of patient data within a specific disease spectrum also matters. This is commonly referred to as spectrum bias or reference dataset bias, which results from an unrealistic distribution of disease severity in a training dataset [13, 14]. It is therefore important to understand which cases were used in the development of the AI tool and how they differ from the patient population the AI tool will potentially be applied to.

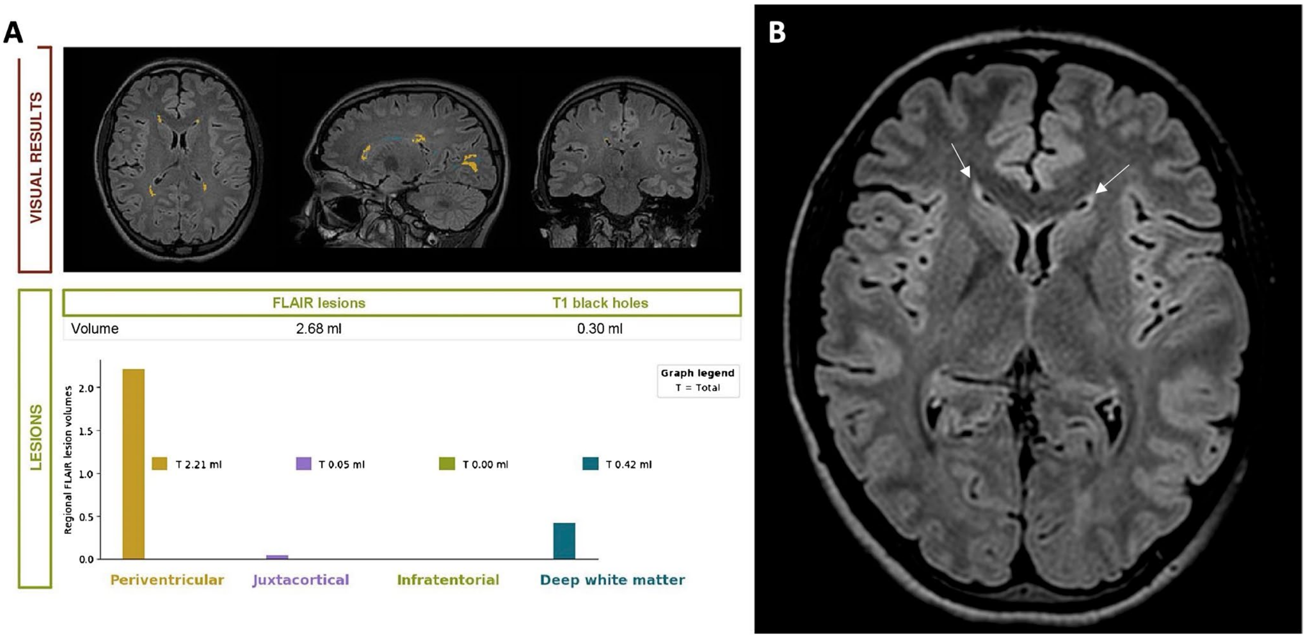

Fig. 3 Female patient, 23 years old, presented with diffuse paresthesia. An MRI was requested to rule out demyelination. The scan was accidentally processed by AI-based software for lesion segmentation in MS patients. The software algorithm annotates the periventricular white matter erroneously as lesions (panel **A**). The MRI exam was normal. The discrete FLAIR hyperintense signal near the ependyma, especially visible around the frontal horns, is a physiological phenomenon (panel **B**, arrows). In this example, the AI tool provides a false positive segmentation of "MS lesions"

Why it is important

In order to design an algorithm for disease classification, it is desirable to have highly accurate labels as the reference standard for algorithm training (e.g., disease yes/no). However, enriching training data with clearly sick patients and clearly healthy controls, i.e., increasing training data at both tails of the distribution spectrum, might increase algorithmic performance on an external test set of a similar distribution spectrum, but may harm the use of an algorithm in clinical practice. There are mainly two reasons for this. First, the contribution to the clinical decision process, e.g., for a classification task is negligible in cases of overt disease or absence of disease because they can be easily identified by visual inspection. Help is especially needed for borderline cases. Second, differences in disease severity between training data and real-world data hamper the diagnostic accuracy in clinical practice. The selection of the ideal training dataset is complex, and currently, no generally accepted rules or recommendations exist. For some algorithms, it is probably better to have a disease spectrum that is "narrow" and similar to the subsequent clinical application. For other algorithms, a "broader" disease spectrum including extreme cases might be beneficial. Ideally, the spectrum of the dataset used to validate the AI tool should match the spectrum of the clinical patients for which the AI tool is used.

Neuroradiology example

One commonly addressed image classification task is the identification of patients with or without Alzheimer's disease (AD). The accuracy of such an algorithm can be increased during the training and validation process by enriching the data with cases of overt dementia and cognitively normal controls. However, this hardly contributes to clinical problem solving, since diagnostic support will mostly be needed for ambiguous cases, which are neither clear on clinical exam nor clear on imaging, e.g., using semiquantitative rating scales. Elaborating on the potential impact of spectrum bias, AI algorithms for disease classification should also be adapted to the incidence of specific diseases in the institution where the algorithm will be deployed. This could mean that algorithmic recalibration is needed depending on whether an algorithm will be used in a highly specialized memory clinic or in a primary care setting.

Checkpoint 3: Data parameters

Even if the focus is only on MR image parameters, varying manufacturer types, magnetic field strengths, different acquisition settings (2D or 3D acquisitions), variability of sequence parameters (spatial resolution, voxel size, TE, TR value, etc.) give a large variability in terms of resulting MR images. This compromises the pooling and the reproducibility of published data, when using independent imaging sets [15]. Currently, most AI tools (notably for scientific publications) are developed using only high-quality research datasets, with imposed and strict imaging acquisition parameters. In contrast, in real-world clinical environment, a large variability of image acquisitions exists leading to inter-patient and intra-patient variability. This potentially affects the performance of AI tools. An interesting approach to overcome this limitation is transfer learning [16], which can apply knowledge from one domain (e.g., medical imaging modality or scanner) and one task (e.g., segmentation, classification) to another related domain and/or another task. A subform of transfer learning with special importance in the medical imaging field is domain adaptation. Here, the source domain (training data) is different from the target domain (test data), a situation which is commonly described as "domain shift." Approximating these two data distributions can be tackled by domain adaptation, which represents a machine learning tasks in itself and can be executed in various forms. Although this approach may be promising to overcome high data variability between the site of development and the site of application, it can currently not be used in clinically practice since it cannot be performed after CE marking.

Why it is important

Several studies have shown the variation in results secondary to the application of an AI tool by just a modification in the image parameters. It is therefore important to check which input data is required for the given AI tool, and to adapt imaging parameters accordingly.

Neuroradiology example

Most MR volumetry tools are developed and trained with high-quality research datasets, e.g., the ADNI (Alzheimer Disease Neuroimaging Initiative). Even modest modification of image contrast in the same MR scanner and same head coil may lead to 5% change in estimated volume using standard segmentation tools (without AI) [17]. Likewise, different scanners impact brain volumetric measurements in multiple sclerosis patients [18]. It can be assumed that even more variable clinical datasets will have even stronger effects also when using AI tools, yet this remains to be elucidated.

Another example is the reliability of segmentation of brain tumors in neuro-oncology, as new methods are being published trying to improve accuracy and reproducibility [19]. Those new methods might confound the accurracy of AI tools, which were developed and tested on older datasets.

Checkpoint 4: Data quality check (motion artifacts, metal artifact)

It is evident and trivial that medical imaging can be affected by unintended artifacts such as motion artifacts or metal artifacts (e.g., dental implants, piercings). A neuroradiologist automatically and intrinsically checks image quality before making a diagnosis. Basically, quality check should be an integral part of any processing of images. Most scientific publications of AI tools exclude images with artifacts. But in reality, imaging artifacts exist. Some, but not all AI tools have an initial data quality check. Those tools without data quality check can still process the data in the "black box," yet this might result potentially erroneous outputs and predictions [20]. Data quality check is fundamental to be able to rely on output of AI tools and to avoid errors.

Why it is important

Rather than assuming that input data quality is adequate, an AI tool should first check data quality to guarantee a correct application of the AI tool avoiding, for example, false classifications or wrong segmentation results.

Neuroradiology example

If a 3D T1-weighted sequence is used to process gray matter segmentation, and if there is motion artifact, the final volume will not have a valid value. But unless the AI tool has a first quality check level prior to the segmentation procedure, one may not notice that there has been an error linked to the quality of the data. This is also true with respect to other types of imaging tasks such as AI applied to perfusion imaging, for which motion artifact will not prevent perfusion maps reconstructions and AI application, leading to inappropriate conclusions [21].

Checkpoint 5: Anonymization, pseudonymization, coding, and de-identification of patient data

The protection of personal health information (PHI) with regard to processing of data and free movement of such information has been established by law. The European Union adopted the General Data Protection Regulation (GDPR) on 14 April 2016 and it became enforceable on 25 May 2018. Other countries have other regulations, and those rules might be changed and adapted over time.

A medical image (typically in DICOM format) file not only contains a viewable image, but it also contains a header with a large variety of data elements that can lead to the identification of the patient [22]. In handling any data, not just medical data, GDPR requires removal of any information that can lead to the identification of an individual person. In this view, it is crucial to use a clear terminology and distinction between anonymization or pseudo-anonymization or coding. Anonymization is the process of irreversibly altering classified data, to make re-identification of the patient impossible. Anonymization in the strict sense is completely impractical for clinical use, as the results cannot be reconnected to the case. Pseudonymization or coding does not remove all identifying information but reduces the obviously evident identity of an individual. The patient's identifiers are replaced by artificial identifiers (codes or pseudonyms) that are kept separately and are subject to technical measures for safety. Coded data still allow for re-identification, which is necessary in the clinical context where imaging data processed by an AI algorithm should be linked back to the original patient. In many cases, the term anonymization is colloquially yet imprecisely used for coding. Coded data are accepted by GDPR but are still considered sensitive data and should be protected accordingly. In clinical practice, this implies that either data are processed on site or rigorous requirements regarding the security of the VPN connection are met.

Of particular concern are 3D datasets of the head, which are routinely used, as they might allow for surface recognition of the individual face. Even if the text of, e.g., name and date of birth is removed, uploading a 3D dataset without removing the face area may be problematic regarding patient's privacy [23, 24]. Before de-facing, 97% of case were correctly identified by face recognition tools. Of note, even after de-facing with popular software, 28–38% of individuals were still successfully identified [24]. Importantly, in some instances, the de-facing may introduce subsequent brain segmentation errors, which are not present without the de-facing [24]. It is not always clearly indicated if and how AI tools anonymize (more precisely code) data, and whether or not face removal ("defacing") is included for 3D datasets.

Informing patients that AI will be used to assist in the interpretation of their images is part of the obligation to obtain inform consent, which may vary between countries, yet currently there is no standard how such information is documented and illustrated.

Why is this important?

In current practice, there is often a delicate balance between the protection of the patient's individual privacy and still ensuring that the data are of sufficient quality to make analytics useful and meaningful. Of note, regulations vary between countries and may change over time.

Neuroradiology example

Data anonymization is an important issue that should be considered in every AI tool. The neuroradiologist should be aware which patient information is sent to the AI provider, how the data are decoded, and how patient information is restored to provide the final results.

Checkpoint 6: Data storage and processing

Data storage and management can be done on a local platform in the hospital or the medical imaging department (on-premise server) or via cloud-based services.

Local platforms have the advantage of data availability and safety. On the other hand, in case of technical issues or upgrade, timely support may be hampered by lack of locally based technicians familiar with the system and maintenance costs are often substantial.

Many AI applications in medical imaging offer cloud-based services, making data easily and ubiquitously accessible and reducing processing times by making data handling and running applications independent of local computer power availability. The cloud-based systems used in AI solutions for neuroradiology should be entirely private. Public or hybrid clouds in this context are not acceptable due to data privacy. Evidently, data protection rules must be respected, as discussed above. Processing speed and turnaround times are other factors to be considered and may potentially differ between local and distant solutions.

Why is this important?

The policy regarding data security varies between centers and countries. Before opting for a cloud-based AI application, one should check with the local IT department if patient data are allowed to leave the imaging center environment to an external cloud environment. Multiple companies provide in-between solutions with in-hospital cloud-like systems that are only accessible by the company providing the AI application on specific demand.

When opting for a cloud-based solution, turnaround times should be taken into account and considered appropriate for the given application.

Lastly, it may be of interest to be aware of how patient data are handled after passing the AI algorithm. According to GDPR, companies are allowed to store the de-identified data after use.

Neuroradiology example

The method of data storage and processing should be known before adopting any AI application and should be compatible with the local regulatory framework. Acceptable turnaround times may differ between applications. For example, AI-based applications concerning stroke require faster turnaround times than volumetry-based methods in the context of neurodegeneration [25–27].

Checkpoint 7: Integration in the radiologist's workflow and patient information

The integration of AI tools in the neuroradiologist's workflow is very important for daily routine, but is challenging for several reasons. The main current medical image (typically DICOM format) visualization software packages were developed before the emergence of AI tools. The integration of AI tools might require significant rewriting of the underlying code and such changes might require regulatory re-certification. Moreover, there is currently no clear standard interface how, e.g., AI tools could be integrated into existing software packages, concerning both input into the AI tool and output of the report. The output of AI tools typically takes the form of a DICOM file, a standalone report in text or pdf format, or an integrated report to the RIS system, but usually lacks interactive features.

Why it is important

Neuroradiologists need and want a smooth workflow and AI tools should be seen as a great opportunity to improve the workflow in many aspects. In light on increasing pressure on cost and time, AI tools should be seamlessly integrated into the clinical radiological workflow.

Neuroradiology example

An increasing number of AI tools are available on the market and can be purchased and adapted to the radiologist's workflow. However, for a smoother workflow, it is important to anticipate one's need. Working with different AI tools in daily clinical practice may potentially complicate the daily clinical workflow, for example, if different tools/graphical user interfaces must be opened for different applications (for example, one tool for volumetry, one for T2 lesion segmentation, one for microbleed detection). Some vendors therefore work on the idea of a unique platform focusing on the simplicity for the user's need, to only open one interface.

Especially in case of disagreement between the result of the AI tool and the final decision of the radiologist, the reasons of disagreement should be kept in the final report.

Checkpoint 8: Update

AI tools are a rapidly emerging field, with new algorithms and optimizations appearing every day. Consequently, it is likely that an AI tool will also be updated in the future, and it is important to know how this update strategy is handled. Updates might be due to new or optimized algorithms, but also to new or updated reference datasets.

Why is this important?

A new software version, which is presumably intended to improve the tool, might change the results. In other words, the results of the analysis of the same dataset might differ depending on the software version. The user should be aware which version of software was used for data analysis, and this should be clearly stated on the report of the AI tool.

Neuroradiology example

A dementia patient undergoes longitudinal follow-up MRI including MR volumetry. If different versions of the AI tool are used for the analysis of the previous and current MR volumetry, there might be a systematic technical bias in the estimated volumes which might be mis-interpreted as disease progression (Fig. 4).

Checkpoint 9: Validation and labels

For a clinical application of AI tools, it is important to realize that several stages of validation exist. While the terminology is sometimes not clear, in general, one can discriminate at least two stages: technical/methodological validation and clinical validation. The technical/methodological validation refers to the fact that a given AI tool is technically valid and consistently reproduces a given output, for example, it reliably estimates the same brain atrophy in a MR volumetry tool. The clinical validation refers to a rigorous clinical evaluation of the AI tool, for example, it evaluates whether the estimated atrophy impacts diagnosis, workflow or patient outcome.

It is important to note that within the respective certification framework leading to either CE marking in the European Union or FDA clearance in the USA, different risk classes exist according to the type and intended use of a medical device. In the European Union, CE risk classes I, IIa, IIb, and III exist and represent an increasing inherent risk of decisions based on the medical device in principal, without considering the probability of these risks. According to the Medical Device Regulation (MDR) which came into force on 26th May 2021, any medical device intended to be used in therapeutic or diagnostic decision-making or for monitoring physiological processes is at least assigned the risk class of IIa [28]. This is an example of how the MDR tends to increase the assigned risk classes for SaMD in contrast to the former medical device directive (MDD) since it is very hard to think of any SaMD which would still be risk class I. Practically, any AI tool will be at least risk class IIa under MDR and could be classified as risk class IIb (if the derived decision impact can cause serious deterioration of health or surgical intervention or in case of SaMD monitoring physiological processes, where the nature of the physiological processes could result in immediate danger to the patient) or even risk class III (if the derived decision impact can cause death or irreversible deterioration of health) [29].

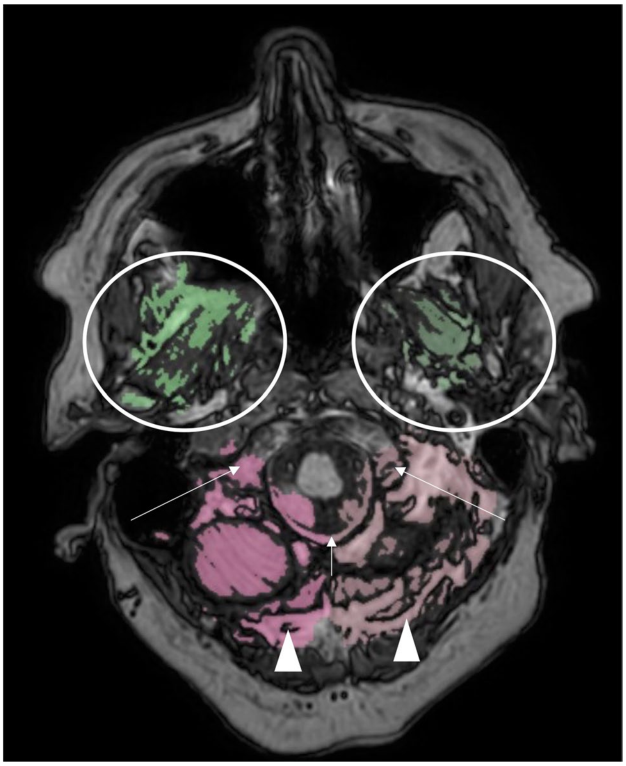

Fig. 4 By means of example, a random 3D T1-weighted gradient-echo sequence (MPRAGE) was analyzed by an AI algorithm which normally has strict requisites for its purpose of whole brain segmentation and brain structure quantification. The results show erroneous segmentation of the infratemporal fossa (circles), the occipital bone (long arrows), the deep muscles of the occipital region (arrowheads), and the dura in the region of the foramen magnum (short arrows). In this example, the AI tool provides a false (positive) segmentation of brain tissue

During the certification process in the EU, a clinical evaluation of the medical device has to be created, and for CE risk classes IIa and higher, an external entity, the so-called notified body, has to be involved. Of note, this clinical evaluation usually does not contain data from clinical studies executed according to rigorous standards; in addition, it is commonly not openly accessible to the public or interest potential users.

The FDA distinguishes three device classes I, II, and III based on the regulatory controls necessary to provide a reasonable assurance of safety and effectiveness and depending on the risk the device poses on the patient and/or the user. Device classes range from I (lowest risk) to III (highest risk) and the assigned risk class largely determines, among other factors, the type of premarketing submission/application needed for FDA clearance to market [30]. These regulatory pathways comprise premarket approval (the most stringent regulatory pathway for class III devices), the 510(k) pathway, where submitters compare their device to an already legally marketed, similar product, or de novo premarket review (for low-risk, class I or II medical devices). In contrary to the EU where CE marking is issued decentralized, private organizations, the FDA is a centralized agency for medical device regulation in the USA. As another difference with respect to the regulations in the EU, the FDA issues a continuously updated list of approved software as a medical device (SaMD) which makes use of AI and the accompanying documentation [31]. Very recently, the FDA has issued an action plan for AI-/ML-enabled SaMD in order to account for the potentially continuously adapting nature of such algorithms by designing a framework, which evaluates both premarket development and post-market performance [32]. Although not particularly focused on AI-/ML-enabled SaMD, this aspect has also been picked up in the EU, since the MDR also emphasizes the need for post-market clinical follow-up studies in order to monitor the performance of medical devices in clinical practice.

Investigating the scientific evidence of 100 CE-marked and commercially available AI tools in radiology, the authors found that for the majority of tools (64/100), peer-reviewed evidence on its efficacy is lacking and only the minority (18/100) have demonstrated (potential) clinical impact [33]. Another recent study reviewed technical and clinical validation of brain morphometry for dementia and found that from 17 products, 11 companies published some form of technical validation, but only 4 used a dementia population [34]. The authors conclude that there is a significant gap between legal certification in the EU and clinical validation, workflow integration, and in-use evaluation of these tools in dementia MRI diagnosis.

Why is this important?

If a neuroradiologist buys a commercially available AI tool with a CE label, the neuroradiologist might have the intrinsic assumption that this AI tool is also validated for clinical use. Consequently, the radiologist might use this AI tool in clinical routine with a good intention. However, in most cases, there is currently not sufficient clinical validation to justify the clinical use of the given AI tool.

Neuroradiology example

Let's take an example of a radiologist who uses an AI tool for automatic detection of brain hemorrhage in CT for triage of cases. The AI tool makes a mistake and does not recognize the presence of an intracranial hematoma, putting the case low on the reading list (Fig. 5). By the time the radiologist reads the case, a complication has occurred. In the absence of a strict clinical validation of the tool and in the absence of commonly accepted medico-legal situation of AI tools, such an error will not only have negative consequences for the patient (and relatives), but may also imply consequences for the radiologist and institution who/which use AI tools in the absence of a clinical validation. Of note, the CE certification of such products typically states that they should be used alongside the usual standard of care. From a practical clinical point of view, it is however not evident to understand how such tools should be used alongside clinical care, e.g., for prioritization of cases, and if the standard of care is not changed by AI tools, what is the point of using them?

Clear criteria for clinical evaluation of AI tools in radiology (but also medicine in general) and clarification of the medico-legal situation are urgently needed, yet currently not available in most countries. The discussion of the complex medico-legal situation is beyond the scope of this article and shall be addressed elsewhere.

Checkpoint 10: Ground truth and reference

Concerning the validation and evaluation of AI tools (see above), a ground truth or reference standard is required. This is however not always obvious and might depend on the task of the AI tool. Eventually, medical imaging is performed with the intention of obtaining a medical diagnosis or therapy response, so, for example, the diagnosis would be the ground truth. However, oftentimes in clinical routine, there is no definite diagnosis, and one can only use the current clinical diagnosis as best approximation knowing that the best current clinical diagnosis is not always the definite correct diagnosis. Concerning volumetry or lesion segmentation tools, the estimated volume should be compared to a reference standard, yet this is not obvious and even manually segmented images have inter-rater disagreement. Concerning image enhancement tools, a gold standard of image quality is generally lacking and oftentimes signal to noise of the resulting enhanced image is used. However, unlike photography which is done with the purpose of creating beautiful images, medical imaging is done with the purpose of providing diagnostic images. It might be that a post-processed image looks nicer with better signal to noise, yet a small but relevant pathology such as micro-metastasis or enhancing MS lesion might be "smoothed" away.

Why is this important?

If there is a choice of several tools performing the same task, it is not obvious for the user how to compare the

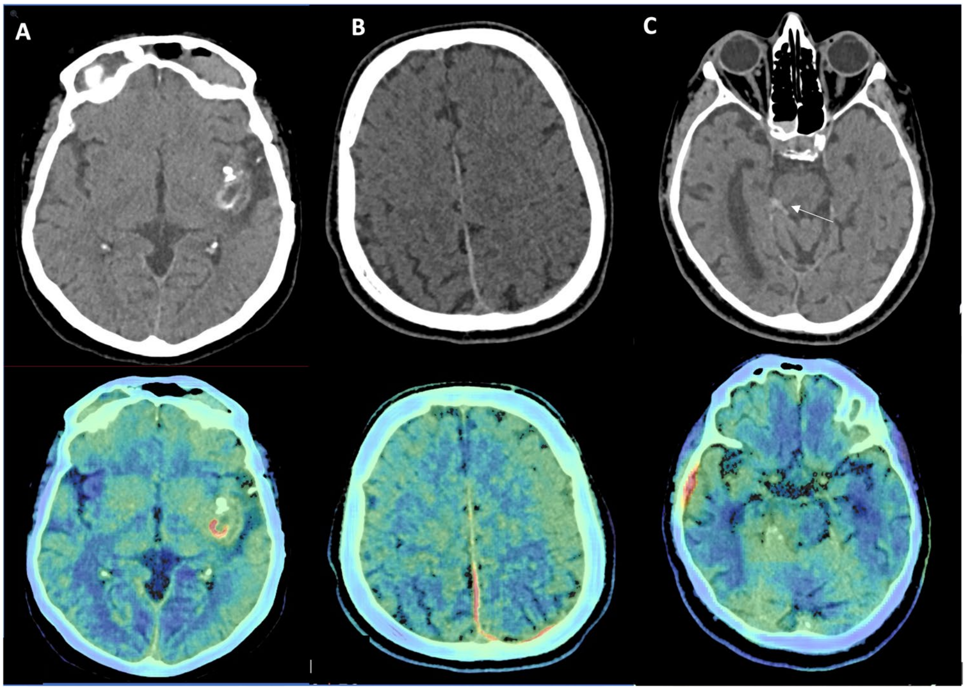

Fig. 5 Three examples of the results of an algorithm for automatic detection of brain hemorrhage, respectively a false positive case (panel **A**), a true positive case (panel **B**), and a false negative case (panel **C**). In the false positive case, the algorithm erroneously detected calcifications in a granuloma due to TBC as hemorrhage. In the true positive case, the radiologist missed the diagnosis of a small subdural hematoma along the falx and in the left parietal region. This was annotated by the algorithm and shown in red color. In the false negative case, the algorithm picked up a small subdural hematoma in the right temporal region but missed the subarachnoid hemorrhage in the ambient cistern (arrow on the native CT images)

tools notably there is no standard for the evaluation of AI tools. One solution could be that the scientific and clinical community provides sample datasets, for example, for volumetry, tumor segmentation of MS lesion segmentation and standards of evaluation, similar to, for example, crash tests for cars. It is obvious that no "artificial" evaluation procedure can reflect the complexity of real life, similar to car crash tests that are evidently not a real accident situation. Nevertheless, a standardized reference dataset and evaluation procedure would represent a simplified but reproducible and comparable evaluation of AI tools.

Neuroradiology example

In the example of AI tools to enhance image quality, the resulting image oftentimes looks smoother than the initial

Table 1 The R-AI-DIOLOGY checklist for clinical use of AI tools in neuroradiology

1	Disease/domain	For which disease/domain was the tool developed and tested? Cleared for use in other diseases?
2	Preselection of cases and reference dataset	Is there a preselection of cases (use for tool development) or preselection of control dataset cases? If yes, is it clearly stated how preselection must be done? Which dataset was used to develop the tool? Definition of controls and patients?
3	Data parameters	Which data acquisition parameters were used? Is there a check that the input data matches the training/reference dataset?
4	Data quality check	Is there a check of the input dataset (e.g. motion, metal or technical artifacts etc..)?
5	Anonymization, coding and de-identification	How are patient data anonymized or more precisely coded? In 3D datasets of the head, is a face removal performed?
6	Data storage and processing	Is the data processing local or on a cloud? Does data processing take place in the same or in a third-party country?
7	Integration in the radiologist's workflow	How is the data transfer into the AI tool? How is the output, and how is this integrated into PACS and/or RIS?
8	Update	Are there automatic or manual software updates? How is the user informed? What if the updated versions provides different results than previous versions (e.g. different reference dataset, different algorithm)
9	Validation and labels	Are there scientific and clinical studies that validate the AI tool for the intended use? Are there regulatory labels?
10	Ground truth and reference	How are ground truth and reference datasets defined?

image. Oftentimes, output measures include signal to noise ratio. However, a smoother image is not necessarily the better image for diagnosis as there is a risk that small but significant lesions are less obvious and “smoothed” away. Different AI tool providers might evaluate their tools in different ways. Standardized criteria for image evaluation of enhanced images are therefore needed.

Another example could be lesion segmentation of MS. Here, oftentimes, some measures of overlap (e.g., DICE score) are provided for the AI tool versus standard clinical evaluation. While such scores are typically used in image analysis evaluation, they are less relevant, e.g., for MS patients. If there is one new lesion, this might indicate a change in treatment even though this lesion might be small and not significantly impact the overall lesion volume. Increase in lesion volume is only beginning to be a marker of disease modification in MS. In contrast, concerning evaluation of brain tumors, increase in lesion volume is typically clinically very relevant and a standard parameter for potential treatment change—although current RANO (response assessment in neurooncology) criteria are based on simple two-dimensional measures. The ground truth and reference criteria of evaluation should therefore be adapted to the clinical needs of different diseases in a standardized way.

Conclusions

The ever-increasing complexity and diversity of available AI tools for clinical neuroradiology make it difficult for the neuroradiologist to capture what is happening inside the “black box”. This practical checklist Table 1 is intended to aid the clinical neuroradiologist to evaluate available AI tools, notably in light of currently unsatisfactorily clarified rules regarding best clinical practice, scientific and clinical validation as well as the medico-legal setting.

Authors' contributions
Sven HALLER
Literature review, manuscript editing, manuscript approval, figures.
Sofie Van Cauter
Literature review, manuscript editing, manuscript approval, figures.
Christian Federau
Literature review, manuscript editing, manuscript approval.
Dennis M. Hedderich
Literature review, manuscript editing, manuscript approval.
Myriam Edjlalis
Literature review, manuscript editing, manuscript approval.

Funding None.

Data availability Not applicable (review article).

Code availability Not applicable (review article).

Declarations

Conflict of interest Christian FEDERAU is working for AI Medical, yet there is no reference to this company in the current review article. The other authors declare no conflict of interest related to the content of this review article.

Ethics approval Not applicable (review article).

Consent to participate Not applicable (review article).

Consent for publication Not applicable (review article).

References

1. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56
2. Muehlematter UJ, Daniore P, Vokinger KN (2021) Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health 3:e195–e203
3. Pesapane F, Codari M, Sardanelli F (2018) Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. Eur Radiol Exp 2(1):35
4. https://grand-challenge.org/aiforradiology/
5. Hedderich DM, Keicher M, Wiestler B, & Gruber… MJ (2021) AI for doctors—a course to educate medical professionals in artificial intelligence for medical imaging. Healthcare
6. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, Coppola F, Morozov S, Zins M, Bohyn C, Koç U, Wu J, Veean S, Fleischmann D, Leiner T, Willemink MJ (2021) An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. Eur Radiol 31(9):7058–7066
7. Lui YW, Chang PD, Zaharchuk G, Barboriak DP, Flanders AE, Wintermark M, Hess CP, Filippi CG (2020) Artificial intelligence in neuroradiology: current status and future directions. AJNR Am J Neuroradiol 41(8):E52–E59
8. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A (2017) Deep learning: a primer for radiologists. Radiographics 37(7):2113–2131
9. Cheng PM, Montagnon E, Yamashita R, Pan I, Cadrin-Chênevert A, Perdigón Romero F, Chartrand G, Kadoury S, Tang A (2021) Deep learning: an update for radiologists. Radiographics 41(5):1427–1445
10. https://phil415.pbworks.com/f/TuringComputing.pdf
11. Shen D, Wu G, Suk HI (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248
12. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9(4):611–629
13. Park SH, Han K (2018) Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 286(3):800–809
14. Sica GT (2006) Bias in research studies. Radiology 238(3):780–789
15. Molina D, Pérez-Beteta J, Martínez-González A, Martino J, Velasquez C, Arana E, Pérez-García VM (2017) Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization. PLoS One 12(6):e0178843
16. Valverde JM, Imani V, Abdollahzadeh A, De Feo R, Prakash M, Ciszek R, Tohka J (2021) Transfer learning in magnetic resonance brain imaging: a systematic review. J Imaging 7(4):66

17. Haller S, Falkovskiy P, Meuli R, Thiran JP, Krueger G, Lovblad KO, Kober T, Roche A, Marechal B (2016) Basic MR sequence parameters systematically bias automated brain volume estimation. Neuroradiology 58(11):1153–1160
18. Biberacher V, Schmidt P, Keshavan A, Boucard CC, Righart R, Sämann P, Preibisch C, Fröbel D, Aly L, Hemmer B, Zimmer C, Henry RG, Mühlau M (2016) Intra- and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. Neuroimage 14:2188–197
19. Chaddad A, Kucharczyk MJ, Daniel P, Sabri S, Jean-Claude BJ, Niazi T, Abdulkarim B (2019) Radiomics in glioblastoma: current status and challenges facing clinical implementation. Front Oncol 9:374
20. Larson DB, Boland GW (2019) Imaging quality control in the era of artificial intelligence. J Am Coll Radiol 16(9 Pt B):1259–1266
21. Copelan AZ, Smith ER, Drocton GT, Narsinh KH, Murph D, Khangura RS, Hartley ZJ, Abla AA, Dillon WP, Dowd CF, Higashida RT, Halbach VV, Hetts SW, Cooke DL, Keenan K, Nelson J, Mccoy D, Ciano M, Amans MR (2020) Recent administration of iodinated contrast renders core infarct estimation inaccurate using RAPID software. AJNR Am J Neuroradiol 41(12):2235–2242
22. Aryanto KY, Oudkerk M, van Ooijen PM (2015) Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol 25(12):3685–3695
23. Schwarz CG, Kremers WK, Therneau TM, Sharp RR, Gunter JL, Vemuri P, Arani A, Spychalla AJ, Kantarci K, Knopman DS, Petersen RC, Jack CR (2019) Identification of anonymous MRI research participants with face-recognition software. N Engl J Med 381(17):1684–1686
24. Schwarz CG, Kremers WK, Wiste HJ, Gunter JL, Vemuri P, Spychalla AJ, Kantarci K, Schultz AP, Sperling RA, Knopman DS, Petersen RC, Jack CR, Alzheimer’s DNI, (2021) Changing the face of neuroimaging research: comparing a new MRI de-facing technique with popular alternatives. Neuroimage 231:117845
25. Neubauer T, Heurix J (2011) A methodology for the pseudonymization of medical data. Int J Med Inform 80(3):190–204
26. Ngiam KY, Khor IW (2019) Big data and machine learning algorithms for health-care delivery. Lancet Oncol 20(5):e262–e273
27. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP (2020) Preparing medical imaging data for machine learning. Radiology 295(1):4–15
28. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745&from=EN; Annex VIII
29. https://ec.europa.eu/health/sites/default/files/md_sector/docs/mdcg_2021-24_en.pdf AOD
30. https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device AOD
31. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices AOD
32. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device AOD
33. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M (2021) Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol 31(6):3797–3804
34. Pemberton HG, Zaki LAM, Goodkin O, Das RK, Steketee RME, Barkhof F, Vernooij MW (2021) Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis-a systematic review. Neuroradiology 63(11):1773–1789