HAL open science

# Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse

Sophie Loizillon, Simona Bottani, Aurélien Maire, Sebastian Ströer, Lydia Chougar, Didier Dormont, Olivier Colliot, Ninon Burgos

## To cite this version:

Sophie Loizillon, Simona Bottani, Aurélien Maire, Sebastian Ströer, Lydia Chougar, et al.. Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse. Medical Image Analysis, 2025, 103, pp.103617. 10.1016/j.media.2025.103617 . hal-05059257

**HAL Id: hal-05059257**

**https://inria.hal.science/hal-05059257v1**

Submitted on 7 May 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

# Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse

Sophie Loizillon[a], Simona Bottani[a], Aurélien Maire[b], Sebastian Ströer[c], Lydia Chougar[c], Didier Dormont[c,d], Olivier Colliot[a], Ninon Burgos[a,*], the APPRIMAGE Study Group[1]

[a] *Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris 75013, France*
[b] *AP-HP, Innovation & Données – Département des Services Numériques,Paris 75012, France*
[c] *AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, Paris 75013, France*
[d] *Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, DMU DIAMENT, Paris 75013, France*

**Abstract**

Clinical data warehouses, which have arisen over the last decade, bring together the medical data of millions of patients and offer the potential to train and validate machine learning models in real-world scenarios. The quality of MRIs collected in clinical data warehouses differs significantly from that generally observed in research datasets, reflecting the variability inherent to clinical practice. Consequently, the use of clinical data requires the implementation of robust quality control tools.

By using a substantial number of pre-existing manually labelled T1-weighted MR images (5,500) alongside a smaller set of newly labelled FLAIR images (926), we present a novel semi-supervised adversarial domain adaptation architecture designed to exploit shared representations between MRI sequences thanks to a shared feature extractor, while taking into account the specificities of the FLAIR thanks to a specific classification head for each sequence. This architecture thus consists of a common invariant feature extractor, a domain classifier and two classification heads specific to the source and target, all designed to effectively deal with potential class distribution shifts between the source and target data classes. The primary objectives of this paper were: (1) to identify images which are not proper 3D FLAIR brain MRIs; (2) to rate the overall image quality.

For the first objective, our approach demonstrated excellent results, with a balanced accuracy of 89%, comparable to that of human raters. For the second objective, our approach achieved good performance, although lower than that of human raters. Nevertheless, the automatic approach accurately identified bad quality images (balanced accuracy >79%). In conclusion, our proposed approach overcomes the initial barrier of heterogeneous image quality in clinical data warehouses, thereby facilitating the development of new research using clinical routine 3D FLAIR brain images.

*Keywords:* Clinical Data Warehouse, Deep Learning, Quality Control, MRI, Domain Adaptation, FLAIR

## 1. Introduction

In recent years, clinical data warehouses (CDWs) have emerged, bringing together large amounts of medical data that can be reused for research purposes (Gehring and Eulenfeld, 2018; Doutreligne et al., 2023; Nordlinger et al., 2020). In France, several hospitals have been working for more than ten years to create CDWs. Of the 32 French university hospitals, 14 have a CDW in production (Lamer et al., 2023; Artemova et al., 2019; Wack, 2017). One of these is that of the Greater Paris University Hospitals (Assistance publique-hôpitaux de Paris [AP-

*Corresponding author: ninon.burgos@cnrs.fr
[1] Members of the APPRIMAGE study group can be found at https://www.aramislab.fr/apprimage

Paper published in Medical Image Analysis, doi:10.1016/j.media.2025.103617

HP]), which gathers all the medical data from 39 Parisian hospitals, including more than 25 million medical images. This ecosystem provides an exceptional opportunity to develop and validate machine learning algorithms on heterogeneous clinical images. In the past year, numerous research projects have emerged from the AP-HP CDW on different imaging modalities and tasks. Berenbaum et al. (2023) developed an automated classification model for whole-body $^{18}$F-Fluorodeoxyglucose positron emission tomography images. Vanderbecq et al. (2024) trained a deep learning model to automatically identify bowel obstructions on abdominal computed tomography scans. In addition, Beaufrère et al. (2024) proposed a classification model for primary liver cancer from routine tumour biopsies using weakly supervised deep learning.

If CDWs provide many research opportunities, they also come with challenges. In a previous study (Bottani et al., 2023), we highlighted the challenges of translating computer-aided diagnosis systems from research to clinical routine, with a focus on the diagnosis of dementia based on T1-weighted (T1w) MRI. In particular, we demonstrated that the performance of such systems is strongly biased upwards by confounders such as image quality (due to a "short-cut" learning phenomenon). This study highlights the importance of first developing a robust automatic image quality control (QC) tool before delving into any study involving CDW images. Indeed these images are routine clinical data, which are characterised by a great heterogeneity in terms of machines, manufacturers, magnetic field and overall image quality. These clinical routine images greatly differ from research dataset images. We therefore developed a supervised model for the automatic QC of 3D T1w brain MRIs based on the manual annotation of more than 5,500 images (Bottani et al., 2021). Our model was able to accurately detect poor quality images (balanced accuracy >80%), making it possible to perform an initial quality filter to conduct further studies and analyses using the T1w MR images of the CDW (Bottani et al., 2023, 2024).

*FLAIR quality control*

While 3D T1w brain MRI is a valuable sequence for observing structural changes and assessing the pattern and extent of brain atrophy, its diagnostic power can be enhanced by combining it with other sequences, such as fluid attenuated inversion recovery (FLAIR) MRI. Visualisation of white matter lesions is facilitated in this type of sequence, which suppresses the confounding signal from cerebrospinal fluid. Thus, FLAIR is a very popular sequence to detect white matter lesions that may occur in diverse disorders such as multiple sclerosis or leukoaraiosis (Karadeli et al., 2016; Bink et al., 2006). Recently, 3D FLAIR sequences have become part of most routine clinical protocols. Despite the increased acquisition time, they have the advantage of high spatial resolution with high signal-to-noise ratio (SNR) and the ability to obtain multi-planar reconstruction, allowing simultaneous evaluation of the lesion in three orthogonal planes (Naganawa, 2015).

As with the T1w sequence, FLAIR images can suffer from different types of artefacts including motion, noise, poor contrast, distortion artefacts, flow artefacts, magnetic susceptibility artefacts and ghosting artefacts (Krupa and Bekiesińska-Figatowska, 2015; Lavdas et al., 2014). These different artefacts can affect the overall image quality, highlighting the critical need for quality control. While significant efforts have been dedicated to develop quality control tools for T1w brain MRIs (Esteban et al., 2017; Alfaro-Almagro et al., 2018; Sujit et al., 2019; Bottani et al., 2021; Hendriks et al., 2024; Griffanti et al., 2023), works targeting FLAIR are more limited. We can still distinguish two main types of approaches: quantitative and qualitative. Peltonen et al. (2018) proposed a quantitative quality control method for clinical FLAIR MRI. Four metrics – image resolution, contrast-to-noise ratio (CNR), quality index and bias index – were introduced to evaluate the overall image quality. The evaluation of these metrics was based on the segmentation of different brain tissues by SPM (Penny et al., 2011). Similarly, Duchesne et al. (2019), as part of their work to harmonise the Canadian dementia cohort, have developed a QC approach for the FLAIR sequence based on SNR and CNR measurements between different brain tissues. In their study, Storelli et al. (2019) used both quantitative and qualitative approaches for QC of FLAIR MRIs from multiple sclerosis patients within the Italian Neuroimaging Network initiative. The quantitative approach involved the measurement of various metrics, including relative image contrast and head positioning quality, while the qualitative approach involved the visual assessment of different types of artefacts, such as motion, noise and ghosting.

The pipelines proposed in these studies are quite extensive and require segmentation steps using specific software such as SPM to extract features. Although these approaches are efficient on research datasets, their successful extension to large, heterogeneous routine clinical databases is challenging. Such databases often contain a large number of low-quality images, making the use of QC tools based on extensive pre-processing pipelines impossible, since such pipelines are unreliable on low-quality images. In (Loizillon et al., 2024a), we showed that although `FAST`, the automated tissue segmentation tool of FSL (Jenkinson et al., 2012), accurately segmented most good quality, non-injected, T1w clinical routine images, its segmentation success rate declined as image quality deteriorated. In particular, failures occurred in 24% of T1w MRIs with severe poor contrast problems and in 44% of T1w MRIs with severe noise. This prevents tools such as MRIQC (Esteban et al., 2017) from being used in routine clinical MRI.

Thus, to overcome these limitations, some studies based on visual inspection of the FLAIR have been developed. Corcuera-Solano et al. (2015), in their comparative study of FLAIR image quality from two different types of acquisition (fast spin-echo PROPELLER vs conventional PROPELLER), manually assessed the quality of the images, considering notably the contrast, i.e., the ability to distinguish between grey and white matter, on a three-point scale. The presence of blurb, due to the acquisition parameters of the multi-echo sequences, was noted as a binary flag. Thanks to the use of the PROPELLER motion correction technique (Pipe, 1999), there was no need of grading the motion. Tanenbaum et al. (2017) conducted a comparative study between synthetic and conventional MRIs for routine neuroimaging. In this context, they established the following procedure for quality control of FLAIR sequences. Each image was given a score between 1 and 5 for the overall quality: from unacceptable for diagnostic use (1) to excellent overall quality (5). Tsuchida et al. (2021) performed quality control on a multimodal brain database consisting of more than 1,800 MRI scans of healthy young adults. Three raters manually annotated the presence and severity of different types of artefacts on a three-point scale, including ringing artefacts, contrast of subcortical structures and contrast between grey and white matter.

Although automatic quality control methods trained using labels established on qualitative assessment appear better suited to CDW, as they do not require extensive pre-processing pipelines, they are nonetheless costly in terms of manual annotations. In our previous work (Bottani et al., 2021), two raters manually annotated 5,500 T1w MRIs in order to obtain ground truth for developing and testing an effective deep learning model for automating the QC of T1w MRIs. To efficiently extend this work to the FLAIR MRI sequence, we aim to reuse the annotations made on the T1w MRIs to limit the number of FLAIR images to be annotated. This is commonly referred to as domain adaptation in the literature.

*Domain adaptation*

Unsupervised domain adaptation aims to minimise the gap between the source (e.g., T1w) and target (e.g., FLAIR) domains using only labelled data from the source domain (e.g., T1w) (Ganin et al., 2016; HassanPour Zonoozi and Seydi, 2022). Promising results have been obtained using adversarial learning to extract domain-invariant feature maps. This method consists in trying to fool the domain classifier that distinguishes between source and target features (HassanPour Zonoozi and Seydi, 2022). An effective strategy in adversarial learning is the Domain Adversarial Neural Network (DANN), which consists of a shared feature extractor and a domain discriminator (Ganin et al., 2016). The role of the domain discriminator is to discriminate between samples from the source and target domains. At the same time, the feature extractor is trained to fool the domain discriminator, facilitating the acquisition of domain-invariant feature maps. However, Zhao et al. (2019) caution against the DANN (Ganin et al., 2016), warning that the acquisition of domain invariant features does not guarantee robust generalisation to the target domain, especially if there are differences in the distribution of classes between domains.

Saito et al. (2019) introduced the concept of semi-supervised domain adaptation (SSDA), where in addition to the source label annotations, some labelled target samples are available and can help improve model performance. Based on the mini-max paradigm, their mini-max entropy approach consists of first estimating domain-invariant prototypes by maximising entropy, and then clustering features around the estimated prototypes, this time by minimising the same entropy. Some SSDA

approaches focus on reducing intra-domain divergence in the target domain (Jiang et al., 2020; Kim and Kim, 2020). For instance, Jiang et al. (2020) introduced bidirectional adversarial training to attract unaligned sub-distributions from the target domain to the corresponding source sub-distributions. Although many current SSDA methods use adversarial techniques, there has been a recent surge of interest in the application of contrastive learning for domain adaptation (Singh, 2021; Thota and Leontidis, 2021).

Domain adaptation presents a significant challenge in medical imaging, particularly due to the multitude of modalities and sequences involved. The use of adversarial learning in the medical field for classification or segmentation purposes has been explored in a large number of studies (Roels et al., 2019; Sundaresan et al., 2021; Feng et al., 2023). Roels et al. (2019) presented an extension of the adversarial classification architecture proposed by (Ganin et al., 2016) to an encoder-decoder segmentation architecture in volume electron microscopy imaging. First, the network was trained in an unsupervised manner and then fine-tuned on the target domain. Sundaresan et al. (2021) led a comparative analysis of DANN (Ganin et al., 2016), a semi-supervised DANN, and a strategy consisting in fine-tuning on the target labelled data for segmenting white matter lesions. Best results were obtained with the semi-supervised DANN. However, all source and target samples were passed through the same label classifier, which seems sub-optimal, especially when there is a distribution shift between the source and target data.

*Contributions*

In this paper, we present a large-scale dataset of clinical brain 3D FLAIR MRIs originating from a network of 39 university hospitals around Paris (AP-HP) and propose a novel semi-supervised domain adaptation method for automatic quality control of FLAIR brain MRI for a large clinical data warehouse. The proposed adversarial architecture, composed of a classification head for both source and target labels, effectively addresses potential class distribution shifts between the source and target data classes. Our approach allows an assessment of overall image quality using a limited number of manually annotated FLAIR images, leveraging the annotations from 5,500 manually annotated T1w MRIs. Preliminary work was published in the proceedings of the MICCAI Workshop on Domain Adaptation and Representation Transfer (DART) (Loizillon et al., 2023). Contributions specific to this paper include i) a novel learning strategy that uses artefact-specific models which are subsequently recombined to determine the overall image quality; ii) the implementation and application of the proposed SSDA approach for the detection of images not considered as 3D FLAIR brain MRIs, and for the detection of images of good, medium and bad quality and iii) the in-depth presentation of the large-scale clinical dataset of brain 3D FLAIR images.

## 2. Materials and methods

### *2.1. Dataset description*

This study is based on a large routine clinical dataset that includes 3D FLAIR brain MRI scans of adult patients acquired in AP-HP hospitals. As soon as an image is acquired at one of the 39 Parisian hospitals that are part of the AP-HP, it is stored directly in a single central clinical picture archiving and communication system (PACS). Images stored in the clinical PACS are regularly imported in the CDW datalake for research purposes. In addition to these imaging data, the AP-HP CDW also stores clinical data associated with patients within the CDW datalake. For example, clinical data may include prescriptions and administration of medicines, results of laboratory tests and diagnostic codes. Both images and clinical data stored in the datalake are pseudonymised. Specifically, DICOM fields corresponding to patient and physician information are removed, and the examination and birth dates are shifted. Data were only accessible remotely by connecting to the AP-HP network and all the experiments were done using the CDW computing resources. Exporting data outside the hospital network was not allowed. The workflow is described in Fig. 1.

*FLAIR dataset*

The 3D FLAIR MRIs were selected in the AP-HP CDW according to DICOM attributes. A first filter was applied to the CDW datalake to list all DICOM attributes corresponding to MRIs. Then the DICOM attributes 'series description', 'body parts examined' and 'study description' were listed. With the help of a neuroradiologist, we manually selected all attribute values that could refer

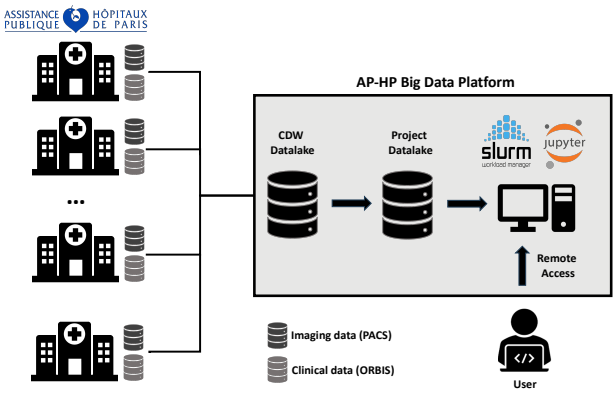


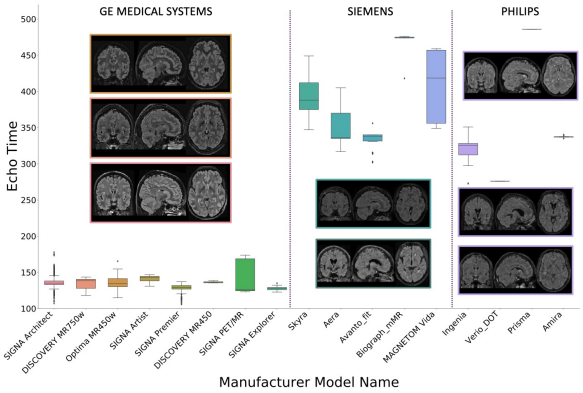

Figure 1: Workflow illustrating the clinical data warehouse ecosystem. Imaging and clinical data from the 39 AP-HP hospitals are pseudonymised, copied and aggregated into the CDW datalake within the big data platform. According to the research project, the AP-HP Innovation and Data Division is giving access to a subset of the CDW datalake (Project Datalake). Research teams can only access data through a JupyterHub and connect to cluster machines within the AP-HP network.

to 3D FLAIR brain MRI (e.g., 'Sag CUBE FLAIR HS', '3D FLAIR HYDROPS', '3D SPACE FLAIR CAIPI', 'FLAIR 3D VISTA', 'SPACE FLAIR'). In the end, 331 relevant attribute values were selected, corresponding to more than 120,000 potential FLAIR MRIs.

Of all the 3D FLAIR brain MRIs from the AP-HP, a batch of 13,703 images corresponding to 10,999 patients was delivered by the AP-HP Innovation and Data Division. These images were acquired on 25 different scanner models from three manufacturers: Siemens Healthineers (n = 5718), GE Healthcare (n = 7894) and Philips (n = 91). Our 3D brain FLAIR MRI dataset was predominantly composed of images acquired with 3 Tesla (T) machines, with 11,341 images, while only 2,362 images were acquired with 1.5 T machines. Among the 13,703 images, the distribution between women and men was balanced, consisting of 7,120 women and 6,583 men, with an average age of 56.53 ± 18.00 years (min: 18, max: 99). Table 1 reports all the scanner models present in the FLAIR dataset with the corresponding manufacturer and magnetic field strength, the mean age and sex. A Sankey plot analysis of the FLAIR MRIs highlighting the scanner model, manufacturer and site distribution is available in the supplementary materials (Fig. S1).

This dataset is characterised by its great heterogeneity with images acquiring over 20 different hospitals of the Greater Paris area using 25 different scanner models with no homogenisation in the acquisition parameters. Fig. 2 shows the echo time (TE) distribution from the 13,703 clinical FLAIR MRIs as a boxplot. Similar plots are available in the supplementary materials (Fig. S2) for other acquisition parameters: the inversion time (TI) and the repetition time (TR). While the original FLAIR sequence used TI values of 2000–2500 ms to cancel the cerebrospinal fluid signal, coupled with very long TR (8000 ms) and TE (140 ms) values to create a strong T2 weighting (De Coene et al., 1992), we have observed a huge heterogeneity in the parameters used in clinical routine to acquire a 3D FLAIR brain MRI.

Figure 2: Boxplot of the echo times used for the acquisition of the 13,703 FLAIR images from the CDW composing our clinical routine dataset.




### *T1w dataset*

For the 3D T1w clinical dataset, we reused the same dataset as in our previous study, where we randomly selected 5,500 T1w MRIs of the AP-HP CDW, corresponding to 4,177 patients, acquired in Parisian hospitals using 30 different scanner models from four different manufacturers: Siemens Healthineers (n = 3752), GE Healthcare (n = 1710), Philips (n = 33) and Toshiba (n = 5) (Bottani et al., 2021).

## 2.2. *Image preprocessing*

The 3D T1w and FLAIR brain MRIs were converted from DICOM to NIfTI files using the dicom2niix soft-

Table 1: Model name of scanners, grouped by manufacturer, with the corresponding magnetic field strength in Tesla, the number of images, the age (mean ± std) and sex (number of females / males)

| | Model Name | Field strength (T) | N images | Age (mean ± std) | Sex (F/M) |
|---|---|---|---|---|---|
| GE | SIGNA Architect | 3 | 4101 | 54.20 ± 18.0 | 2222/1879 |
| | DISCOVERY MR750w | 3 | 1851 | 54.86 ± 18.0 | 961/890 |
| | Optima MR450w | 1.5 | 860 | 61.59 ± 18.0 | 412/448 |
| | SIGNA Artist | 1.5 | 617 | 58.79 ± 17.0 | 308/309 |
| | SIGNA Premier | 3 | 377 | 59.44 ± 18.0 | 165/212 |
| | DISCOVERY MR450 | 1.5 | 55 | 57.24 ± 14.0 | 24/31 |
| | SIGNA PET/MR | 3 | 19 | 63.47 ± 14.0 | 12/7 |
| | SIGNA Explorer | 1.5 | 14 | 65.21 ± 9.0 | 5/9 |
| Siemens | Skyra | 3 | 4669 | 58.12 ± 19.0 | 2396/2271 |
| | Aera | 1.5 | 633 | 53.16 ± 18.0 | 368/265 |
| | Avanto_fit | 1.5 | 162 | 56.44 ± 17.0 | 78/84 |
| | Biograph_mMR | 3 | 121 | 62.83 ± 17.0 | 66/55 |
| | MAGNETOM Vida | 3 | 44 | 52.27 ± 18.0 | 18/26 |
| | Verio_DOT | 3 | 41 | 56.00 ± 17.0 | 17/24 |
| | Prisma | 3 | 30 | 48.60 ± 16.0 | 18/12 |
| | Amira | 1.5 | 12 | 42.50 ± 20.0 | 4/8 |
| | Spectra | 3 | 2 | 46.00 ± 0.0 | 1/1 |
| | MAGNETOM Sempra | 1.5 | 2 | 62.00 ± 1.0 | 2/0 |
| | Sempra | 1.5 | 1 | 46.00 ± 0.0 | 1/0 |
| | MAGNETOM Sola | 1.5 | 1 | 52.00 ± 0.0 | 0/1 |
| Philips | Ingenia | 3 | 85 | 58.16 ± 16.0 | 39/46 |
| | Ingenia Ambition S | 1.5 | 3 | 66.33 ± 8.0 | 2/1 |
| | Multiva | 1.5 | 1 | 34.00 ± 0.0 | 1/0 |
| | Achieva dStream | 1.5 | 1 | 71.00 ± 0.0 | 0/1 |
| | Ingenia Elition X | 3 | 1 | 66.00 ± 0.0 | 0/1 |

ware (Li et al., 2016) and stored following the Brain Imaging Data Structure specification (Gorgolewski et al., 2016). To facilitate annotation, we slightly pre-processed the T1w and FLAIR MRIs. The T1w MRIs were already pre-processed in our previous study (Bottani et al., 2021) using the `t1-linear` pipeline from the Clinica software (Routier et al., 2021). A new pipeline called `flair-linear` was implemented in the Clinica software to pre-process the FLAIR MRIs. This pipeline is similar to the `t1-linear` pipeline, starting with the same bias field correction based on the N4ITK method (Tustison et al., 2010). However, it changes during affine registration to the MNI space, performed using ANTs (Avants et al., 2014), as a specific FLAIR template is used (Winkler). The spatially normalised images were rescaled by clipping the intensity values to the [2,98] percentiles and cropped to remove background, resulting in images of size 169×208×179, with 1 mm isotropic voxels.

### 2.3. Manual labelling of the dataset

To develop an effective automatic quality control tool for 3D FLAIR images, two trained manual annotators assessed image quality by focusing only on the central axial, sagittal and coronal slices of the brain.

#### 2.3.1. Quality criteria

With the help of a neuroradiologist, we adapted the criteria previously defined for T1w MRIs (Bottani et al., 2021). The aim was to establish criteria for the FLAIR sequence that were close to and consistent with those previously defined for the T1w sequence, but also relevant and applicable to FLAIR MRI. Motion, noise and contrast were graded on a three-point scale according to the following criteria:

- **Motion** 0: no motion, 1: some motion but the structures of the brain are still distinguishable, 2: severe motion, the cortical and subcortical structures are difficult to distinguish. The spatial resolution of the image (before pre-processing) was displayed to know if it was anisotropic. In that case, the small blur that could be seen on only one slice due to resampling was ignored and the image scored as motion 0.

- **Noise** 0: no noise, 1: presence of noise that does not prevent identifying structures, 2: severe noise that does prevent identifying structures.

- **Contrast** 0: good contrast, 1: medium contrast (grey matter and white matter are difficult to distinguish in some parts of the image), 2: bad contrast (grey matter and white matter are difficult to distinguish everywhere in the brain). In case of severe diseases (e.g., important stroke or tumour), the notation of contrast was relaxed as the diseases may severely impact the overall contrast of the image.

Using the scores assigned to various types of artefacts, we have employed the concept of tiers, which was introduced in (Bottani et al., 2021), to categorise the overall quality of MR images. Table 2 recapitulates the notion of tier 1 (good quality), tier 2 (medium quality) and tier 3 (bad quality). If the brain was truncated or not distinguishable or if the image did correspond to a 3D FLAIR MRI (e.g., images of segmented tissue), we labelled it as a straight reject.

Table 2: Description of the rules of the quality control tiers

| Tier | Description | Determination rule |
|---|---|---|
| Tier 1 | Good quality | Grade 0 for all characteristics (motion, contrast and noise) |
| Tier 2 | Medium quality | At least one characteristic among motion, contrast and noise with grade 1 and none with grade 2 |
| Tier 3 | Bad quality | At least one characteristic among motion, contrast and noise with grade 2 |

In our previous study (Bottani et al., 2021), due to the large volume of T1w MRIs annotated (5,500), we adopted a simple automatic consensus approach for motion, noise, and contrast scores in case of disagreement between the two annotators. This approach consisted in considering the most severe score given by one of the two annotators as the consensus score. Here, as we only annotated a small subset of the entire FLAIR cohort, we used a different approach to reach a consensus label. In case of disagreement, the two manual annotators re-evaluated the images together to reach a consensus. This approach allows us to obtain even more accurate image annotations, thereby increasing the reliability of the labels assigned to each image.

#### 2.3.2. Annotation software

Our goal was to facilitate the annotation process of the FLAIR MRIs in the restricted environment of the CDW, which only offered a Jupyter Hub access without a proper medical image viewer. We reused our previous graphical interface in a Jupyter notebook (`https://github.com/SimonaBottani/Quality_Control_Interface`) and adapted it to the new annotation process (Fig. S3). In particular, as FLAIR sequences are very manufacturer-specific, we display the scanner model and manufacturer during annotation to inform the annotator and thus limit the introduction of bias during the annotation process. To guide the manual annotators, we also display the ICD-10 codes when the codes correspond to a diagnosis associated with the presence of brain lesions and when they

were assigned within three months of the MRI acquisition date. This was an important information, particularly for the contrast rating, which was relaxed in the case of severe disease, as this can severely affect the overall contrast of the image.

### 2.4. Automatic QC method

We developed an automated quality control method using a semi-supervised domain adaptation technique to perform three classification tasks: 1) identifying images that are straight reject, i.e., that are not 3D FLAIR MRIs of the whole brain (straight reject: yes vs no); 2) distinguishing between images of bad quality from those of medium to good quality (tier 3 vs tiers 2-1); 3) discriminating between images of medium and good quality (tier 2 vs tier 1).

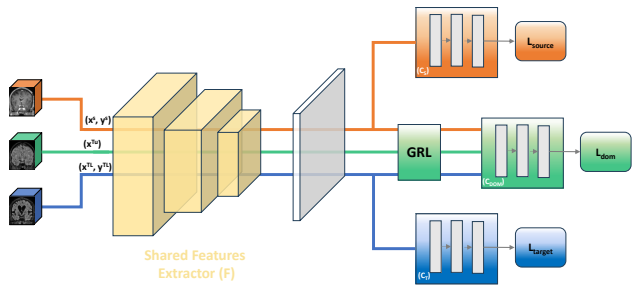
#### 2.4.1. Two-head domain-agnostic classifier

We present a new semi-supervised domain adaptation approach based on adversarial learning for automatic QC of clinical routine FLAIR MRIs. Our proposed architecture, shown in Fig. 3, consists of a common feature extractor ($F$), a source label classifier ($C_S$), a target label classifier ($C_T$) and a domain classifier ($C_{dom}$). We force the shared feature extractor to learn domain invariant features by trying to fool the domain classifier into discriminating between samples from the source and target domains. The introduction of a dedicated classifier for the target data ($C_T$) aims to effectively address the class distribution shift that may exist between the source and target data classes.

We denote the source dataset, comprising labelled T1w MRIs, by $Ds = (x_i^S, y_i^S)_{i=1}^{N_S}$. Regarding the target domain, two datasets are distinguishable: the one comprising the labelled target samples $D_{T_L} = (x_i^{T_L}, y_i^{T_L})_{i=1}^{N_{T_L}}$, and the one with unlabelled target samples denoted by $D_{T_U} = (x_i^{T_U})_{i=1}^{N_{T_U}}$. The labelled data from both domains will be referenced as $D_L = D_S \cup D_{T_L}$. The overall loss is

$$L = \underbrace{L_{pred}(F, C_S, C_T, D_L)}_{\text{Label prediction loss}} - \lambda \cdot \underbrace{L_{dom}(F, C_{dom}, D_L, D_{T_U}, d)}_{\text{Domain confusion loss}} \tag{1}$$

Figure 3: Proposed SSDA architecture composed of a domain invariant feature extractor (F), a source label classifier ($C_S$), a target label classifier ($C_T$) and a domain classifier ($C_{dom}$). A gradient reverse layer (GRL) multiplies the gradient by a negative value when backpropagating to maximise the loss of the domain discriminator.




where

$$L_{\text{pred}} = \underbrace{-\sum_{i=1}^{N_s}\sum_{k=1}^{K}(y_i^S)_k \log(C_S(F(x_i^S))_k)}_{\text{Source label prediction}} \underbrace{-\sum_{i=1}^{N_t}\sum_{k=1}^{K}(y_i^{T_L})_k \log(C_T(F(x_i^{T_L}))_k)}_{\text{Target label prediction}} \tag{1a}$$

$$L_{\text{dom}} = \sum_{i=1}^{N} d_i \log \frac{1}{C_{dom}(F(x_i))} + (1-d_i) \log \frac{1}{1 - C_{dom}(F(x_i))} \tag{1b}$$

with $d_i$ the domain label of the i-th sample, which indicates whether it is a T1w (source domain) or a FLAIR (target domain), and $K$ is the number of classes for the source and target label classifier.

As in (Ganin et al., 2016), the domain adaptation parameter $\lambda$ within the overall loss function is initialised at 0 and progressively increases to 1 using the following equation:

$$\lambda = \frac{2}{1 + \exp(-10p)} - 1 \tag{2}$$

where $p$ is the training progress, which is linearly changing from 0 to 1 over the epochs. This gradual adjustment enables the domain classifier to become less sensitive to noisy signals during the first epochs of the training procedure.

We initialise our model using a pre-trained model that has been trained without the adversarial part (i.e., without the domain classifier). This initialisation is done using all available labelled data from both the source and target domains. During the training of our adversarial network, three different mini-batches were designed within each iteration. The first mini-batch contains labelled samples from the source domain, the second contains labelled samples from the target domain, and the final mini-batch contains unlabelled samples from the target domain. This structure ensures the inclusion of labelled target data in each batch, effectively influencing the training procedure.

#### 2.4.2. *Indirect tier classification*

Rather than directly training a model to assess the overall quality of an image by predicting its tier, we used an indirect tier classification approach that proved successful in our previous study (Loizillon et al., 2024a). Our proposed method is divided into two main steps. First, we train three CNNs using our SSDA architecture tailored to detect specific artefacts (motion, noise and poor contrast) on the FLAIR MRIs from the CDW. Then, we aggregate the results of these three models to predict the overall image quality, .i.e., the tier. We refer to this approach as "indirect quality tier classification" (the tiers are not directly determined by a SSDA CNN but inferred from the results of three models that each detect a specific type of artefact). We compare this approach to the "direct quality tier classification", where the tiers are determined directly by a single CNN. Figure S4 illustrates the two approaches.

#### 2.4.3. *Experiments*

Before starting the experiments, we created a target test set by randomly selecting 480 manually annotated FLAIR MR images. We ensured that they had the same distribution of tiers, manufacturers and field strengths as the images in the training/validation set. The remaining 303 annotated FLAIR MR images were divided into training and validation sets using 5-fold cross-validation. For the source dataset consisting of 5,500 T1w labelled samples, we used the same splits as in (Bottani et al., 2021). The test set is composed of 500 MRIs that have the same distribution of tiers, manufacturers and field strengths as the 5,000 images in the training/validation set, which was further split between training and validation using a 5-fold cross-validation. Our semi-supervised method also allowed us to take advantage of the unlabelled 12,777 FLAIRs during training, which facilitated efficient adversarial training of the feature extractor and domain classifier.

We trained the proposed SSDA architecture using the indirect tier classification methods for two target tasks: tier 3 vs tiers 2-1 and tier 2 vs tier 1. This was done by first training two models for each type of artefact (motion, noise, poor contrast) to detect respectively moderate and severe artefacts. The moderate artefact detection models enable performing the tier 2 vs tier 1 task (detection of medium quality images) and the severe artefact detection models enable performing the tier 3 vs tiers 2-1 (detection of bad quality images). More details about the train/validation/test splits for the moderate and severe artefact detection experiments are given in Table S1 and S2. The inference step is divided into two steps: (1) the prediction of the severity of each artefact with the three artefact-specific networks; (2) the re-combination of the overall quality tier based on the three grades corresponding to the score for motion, noise and contrast. In each experiment, the final model was selected from the five cross-validation models based on the lowest loss over the validation set.

All experiments were performed using the ClinicaDL open-source software implemented in PyTorch (Thibeau-Sutre et al., 2022). Our adversarial semi-supervised domain adaptation architecture consists of a feature extractor comprising five convolutional layers and three classifiers – a source label classifier, a target label classifier and a domain classifier. These classifiers are composed of three fully connected layers each. For the hyperparameters, the Adam optimiser was chosen with a momentum of 0.9. The batch size was set to 10 and the number of epochs to 100. As in (Ganin et al., 2016), an adaptive learning rate, $\mu_p$, was used following Eq. 3 with $\mu_0 = 0.01$, $\alpha = 10$ and $\beta = 0.75$

$$\mu_p = \frac{\mu_0}{(1 + \alpha p)^{\beta}} \quad . \tag{3}$$

## 3. Results

### 3.1. *Manual QC*

Two raters manually annotated 926 3D FLAIRs in accordance with the quality criteria described above. To as-

sess the agreement between the two annotators, we calculated the weighted Cohen's kappa (Watson and Petrie, 2010) for the motion, contrast and noise characteristics. The results for the newly annotated FLAIR are presented in Table 3, together with our previous results for the T1w sequence.

Table 3: Weighted Cohen's kappa between the two annotators for the T1w and FLAIR modalities

| Characteristics | Weighted Cohen's kappa | |
|---|---|---|
| | T1w | FLAIR |
| Contrast (0 vs 1 vs 2) | 0.79 | 0.75 |
| Motion (0 vs 1 vs 2) | 0.68 | 0.65 |
| Noise (0 vs 1 vs 2) | 0.70 | 0.59 |

The inter-rater agreement for motion and contrast is substantial, ranging from 0.65 to 0.75, which is close to the agreement levels previously obtained for T1w MRI. However, we observe a decrease of 10 percent points in agreement for the noise characteristic compared to T1w MRI due to the Kappa dependence on the prevalence (Cicchetti and Feinstein, 1990). For the source modality (T1w), the proportion of images with a noise score of 0 was 69% compared to 84% for the target modality (FLAIR). Nevertheless, the overall level of agreement remains moderate and provides a sufficiently accurate assessment of FLAIR image quality.

We display the distribution of the consensus labels for the 926 manually annotated FLAIR MRIs in Fig. 4. 47% of the images are labelled as good quality (tier 1), 30% as medium quality (tier 2), and 15% as bad quality (tier 3). Only 8% of the MRIs were annotated as straight reject. Figure 5 shows examples of 3D FLAIR brain images along with their corresponding labels. Further examples of straight reject images can be found in the supplementary material (Fig. S5). A poor contrast score was responsible for most of the images being of medium or poor quality. 70% of images in tier 2 had a contrast grade of 1, indicating difficulty distinguishing between grey matter and white matter in some parts of the brain. More than 93% of poor quality FLAIR MRIs (tier 3) were affected by very poor contrast.

We investigated the impact of sex, age, field strength and manufacturer on the different quality tiers. Table 4

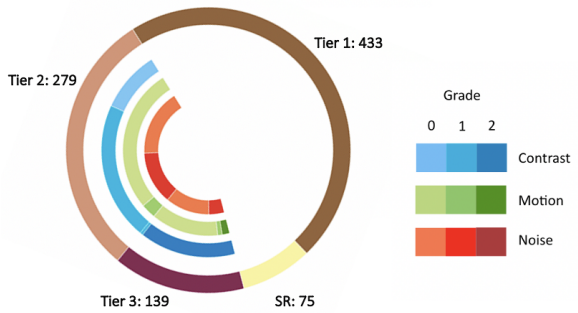



Figure 4: Distribution of the consensus labels for the manually annotated dataset of 926 FLAIR images. Outermost circle: straight reject (SR) images and images in the different tiers. For every tier, we plot the grade distribution of the contrast, motion and noise characteristics.

presents the manufacturer, field strength, age and sex distributions for each quality tier. Tiers 1 and 2 were dominated by 3 T GE Healthcare machines, whereas tier 3 was dominated by 3 T Siemens Healthineers machines. In contrast to age, there was no significant difference in sex between the tiers. We observe that the average age increases with decreasing image quality. For example, the average age difference between good (tier 1) and bad (tier 3) quality images is more than 20 years. It should be noted that a loss of grey-white contrast is observed with age, which is consistent with the high average age of tier 3 images, which account for over 90% of poor quality images.

### 3.2. Automatic QC

The results obtained by our proposed SSDA method for our three tasks of interest (straight reject, tier 3 vs tiers 2-1, tier 2 vs tier 1) are presented in Table 5 and S4. To analyse our results, we compare them with the ones obtained by the annotators and the direct tier classification approach. The balanced accuracy of the annotators is defined as the average of the balanced accuracy between each rater and the consensus. To evaluate the potential statistical difference between the proposed and compared approaches, we performed a McNemar's test. Significance level was set at $p < 0.05$ corrected for multiple comparisons using Bonferroni's approach.

For the straight reject detection task, our model achieved an excellent balanced accuracy of 89%, close to that of the manual annotators (94%). For the tier 3 vs tiers 2-1 classification task using the indirect approach, the

Table 4: Distribution of the manufacturers, field strength, age and sex according to quality grading (performed by the human raters) and on the overall population. We report the percentage of each manufacturer, field strength and sex, and the mean ± standard deviation with the range for age.

| | Manufacturer (%GE, %Siemens, %Philips) | Field strength (%3T, %1.5T) | Age (mean ± std [range]) | Sex (%F, %M) |
|---|---|---|---|---|
| Tier 1 (n=433) | 55.89%, 42.26%, 1.85% | 81.76%, 18.24% | 48.75 ± 16.00 [18, 92] | 52.19%, 47.81% |
| Tier 2 (n=279) | 50.18%, 48.38%, 1.43% | 81.72%, 18.29% | 57.59 ± 18.06 [19, 95] | 47.67%, 52.33% |
| Tier 3 (n=139) | 28.77%, 67.63%, 3.60% | 89.21%, 10.79% | 71.53 ± 14.89 [20, 95] | 45.32%, 54.67% |
| Straight reject (n=75) | 60.00%, 38.67%, 1.33% | 38.67%, 61.33% | 59.60 ± 18.85 [20, 92] | 38.67%, 61.33% |
| Total (n=926) | 50.43%, 47.62%, 1.95% | 79.37%, 20.62% | 55.71 ± 18.51 [18, 95] | 48.70%, 51.30% |

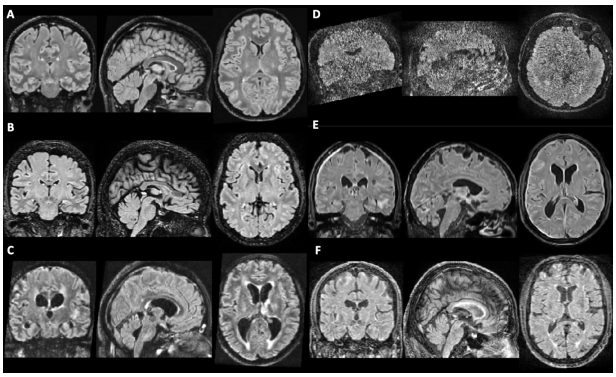



Figure 5: Examples of 3D FLAIR brain images from the clinical data warehouse and the corresponding labels. A: Image of good quality (tier 1); B: Image of medium quality with a noise grade of 1 (tier 2); C: Image of bad quality with a motion grade of 2 (tier 3); D: Straight reject image; E: Image of bad quality with a contrast grade of 2 (tier 2); F: Image of bad quality with a motion and contrast grade of 2 (tier 3).

Table 5: Balanced accuracy of the indirect (proposed) and direct classification approaches for the three tasks: straight reject detection, tier 3 vs tiers 2-1 and tier 2 vs tier 1. We also report the agreement between human raters and the consensus (annotators), which is defined as the average of the balanced accuracy between each rater and the consensus . The table displays the 95% confidence interval computed using bootstrapping on the independent test set (1000 resamples) within brackets. Results with ** indicate a statistically significant difference (corrected p<0.05, McNemar's test) between the proposed (indirect tier classification) and the compared (direct tier classification) approach.

| | Straight reject (yes vs no) | Tier 3 vs tiers 2-1 | Tier 2 vs tier 1 |
|---|---|---|---|
| Annotators | 94.67 [91.32, 97.40] | 86.31 [84.25, 88.25] | 84.43 [82.54, 86.50] |
| Indirect tier classification (proposed) | – | 79.81 [74.65, 84.71] | 73.92 [70.73, 77.22] |
| Direct tier classification | 89.27 [79.69, 97.08] | 76.81** [70.23, 81.26] | 68.08** [63.85, 72.87] |

SSDA performs well with a balanced accuracy of 79%, although lower than that of the annotators (86%). On the other hand, for the more complex tier 2 vs tier 1 task, the balanced accuracy was only 73%, well below the performance of the annotators (84%). We compared our proposed indirect tier classification approach with the direct tier classification method. The indirect approach yielded better performance with a gain of 3 percent points on the tier 3 vs tiers 2-1 task and more than 7 percent points on the tier 2 vs tier 1 task over the direct tier classification.

The intermediate results obtained by our artefact specific models for the moderate and severe tasks are presented in Table 6. We compared them to those obtained by the annotators and to a baseline trained only on the FLAIR images. This baseline network, called Conv5FC3, is composed of five convolutional layers and three fully connected layers to be consistent with the SSDA architecture. This architecture matches that of the network used in our previous work on T1w MRI (Bottani et al., 2021). For the severe artefact detection task, the proposed SSDA approach achieved a balanced accuracy of 82% and 79% for motion and poor contrast artefacts respectively, outperforming the baselines and closely matching the performance of the annotators. However, we have to moderate the significance of the results obtained by the severe motion detection model considerably, as the test set contained only a small number of images with severe motion (6). The detection of moderate artefacts using our SSDA

approach produced mixed results. For the noise, the balanced accuracy was satisfactory and 7 percentage points higher than the baseline, but lower than that of the annotators (77% vs 85%). For motion and contrast detection, the balanced accuracy was low (63% and 69%, respectively), and lower than that of the annotators (90% and 84%, respectively), but still higher than the baselines (61% and 64%).

Table 6: Detection of motion, noise and poor contrast artefacts on the CDW. For both the detection of severe and moderate artefacts, we report the average balanced accuracy (1) of each human rater with the consensus (annotators); (2) of the baseline (Conv5FC3 trained using only FLAIR images); and (3) of the proposed SSDA architecture (SSDA). Sev: severe artefact detection; Mod: moderate artefact detection. † indicates the balanced accuracy obtained with a limited number of images with severe motion artefacts in the test set (6).

| | | Motion | Noise | Contrast |
|---|---|---|---|---|
| Mod | Annotators | 90.00 | 85.73 | 84.65 |
| | Baseline | 61.29 | 71.15 | 64.36 |
| | SSDA (proposed) | 63.54 | 77.13 | 69.33 |
| Sev | Annotators | 76.94† | – | 83.74 |
| | Baseline | 50.00† | | 70.36 |
| | SSDA (proposed) | 82.91† | – | 79.06 |

## 4. Discussion

We have developed a semi-supervised domain adaptation method for automatic quality control of FLAIR brain MRI for a large clinical data warehouse. Our approach allows, with a limited amount of manually annotated FLAIR, an assessment of the overall image quality through the use of 5,500 manually annotated T1w MRIs. To do this, different artefact-specific models were trained to detect severe and moderate artefacts. By recombining the outputs of these models, we were able to assess the overall quality of FLAIR MRIs.

The quality of the images available in the CDW is an initial barrier to their use for research purposes. Therefore, the development of automatic QC tools is a crucial first step in making the most of CDW. Our initial research focused on the development of automated QC tools for T1w MR images, which are widely used to assess structural changes and especially the presence of atrophy when studying neurodegenerative diseases (Bottani et al., 2021). The follow-up to this research was therefore to look at other types of sequences commonly used in clinical practice. We focused on the FLAIR MRI sequence, which is known for its ability to easily visualise white matter lesions, facilitated by the suppression of the cerebrospinal fluid signal. The approach considered was to use domain adaptation to make the most of our previous work on the T1w sequence, in particular the 5,000 manually labelled MRIs for training. Domain adaptation allowed us to significantly reduce the number of new annotations required for the new target sequence (FLAIR), limiting us to just 303 training images. We adapted our manual annotation criteria to the FLAIR, while maintaining consistency with those defined for the annotation of T1w MRIs.

Our study was based on the manual annotation of T1w and FLAIR images, allowing us to compare image quality between these two sequences. The results showed a significant disparity: almost half of the FLAIR images (49%) were classified as good quality (tier 1), while only 16% of the T1w images reached this tier. In addition, 30% of T1w images were of poor quality (tier 3), compared to only 15% in the FLAIR cohort. Overall, the FLAIR cohort had better image quality than the T1w cohort. This difference may be due to the recent introduction of the 3D FLAIR sequence into clinical routine, whereas some of the T1w images in our dataset can be more than 20 years old.

Regarding our automatic QC tools, for the detection of poor quality images (i.e., tier 3 vs tiers 2-1), our semi-supervised domain adaptation approach allows us to identify these bad quality images with a satisfactory balanced accuracy of 79%. This is particularly important for the accurate detection of these types of images, as they are generally the ones on which extensive image processing pipelines can fail. It is interesting to compare the performance of this model (79%) with that obtained in our previous work on T1w with 5,000 annotated images in the training/validation set (83%). Thanks to our domain adaptation technique, we were able to exploit the T1w annotations and reduce the number of annotations for the new modality (FLAIR) by more than 16 times, while maintaining similar performance.

For the task of distinguishing between medium and high quality images (tier 2 vs tier 1), our proposed model shows a low accuracy (73.92%), which is in line with the

results previously obtained on the T1w (71.65%). Although the detection of medium quality images (tier 2) can be useful, it is much less crucial than that of tier 3. Indeed, images of medium quality are likely to contribute to reliable diagnostic predictions. We are therefore confident that these quality control tools can be reliably applied to T1w and FLAIR MRI from large clinical data warehouses and will help to make full use of CDW for research purposes. Nevertheless, we believe that performance could be improved by incorporating additional training data, exploiting synthetic artefacts (Loizillon et al., 2024a,b), refining the quality annotation criteria to ensure greater consistency, and exploring more advanced modelling approaches capable of capturing subtle variations in artefacts. We leave these improvements for future work.

The proposed two-head domain-agnostic classifier, which includes a label classifier specific to the target domain, appears to enable overcoming the covariate shift and the class distributions shift between the source and target domains. To demonstrate the impact of our SSDA architecture, in our preliminary work (Loizillon et al., 2023) we compared our approach with state-of-the-art architectures: semi-supervised DANN (Sundaresan et al., 2021), mini-max entropy (Saito et al., 2019) and the entropy minimisation algorithm (Grandvalet and Bengio, 2004). The results are available in the supplementary materials (Table S3) and show a systematic improvement of our approach by more than 10 percentage points on the Tier 3 vs Tiers 2-1 task.

One of the main limitations of this study is the image annotation process, which is based on the visualisation of only three central slices, potentially limiting the detection of localized artefacts. The annotation procedure was driven by the technical constraints of our restricted working environment, as the annotation was performed using a JupyterHub interface, which severely limited the visualization and interaction with full 3D volumes. As a practical compromise, we selected the central slices, which generally provide a representative view of image quality, but may overlook artefacts located in other regions. This limitation may introduce bias into the models and affect their ability to generalize to certain types of local artefacts. Future work should explore more flexible annotation tools that allow for evaluation of the entire 3D volume, allowing for a more comprehensive assessment of image quality, especially in anatomically challenging areas. In addition, in an attempt to follow the protocol established for T1w, we annotated the noise, motion and grey-white matter contrast of the image in the FLAIR sequence. However, in FLAIR, which is widely used to observe white matter lesions, the assessment of contrast between healthy and lesional tissue may be also relevant. Moreover, this work focused exclusively on 3D FLAIR MRI. However, 2D FLAIR images are still widely used in clinical practice, especially in time-critical contexts such as stroke diagnosis. Extending our method to 2D acquisitions is an important direction for future work.

Thanks to the exceptional heterogeneity of our clinical dataset, including data from multiple sites, different manufacturers and different machines, we were able to develop quality control tools for two MRI sequences widely used in clinical practice to study brain disorders, T1w and FLAIR. It will be interesting to extend our semi-supervised domain adaptation methodology to other sequences commonly used in the clinic (susceptibility-weighted imaging, T2*, etc.), which would then allow developing analysis tools for a wide spectrum of sequences. Another promising future direction is to test and compare our approach with SynthSeg+ (Billot et al., 2023), a segmentation tool specifically designed for processing heterogeneous brain MRI scans from clinical routine. The latest version of SynthSeg+ enables scalable quality control by automatically detecting faulty segmentations. We plan to incorporate this comparison in our future studies.

## 5. Conclusion

We proposed a semi-supervised domain adaptation framework for the automatic quality control of 3D brain FLAIR MRI for a large clinical data warehouse. Using the manual annotations of 5,000 T1w images, we trained CNNs to detect specific artefacts such as motion, noise and poor contrast using only a limited number of FLAIR annotations. Evaluation on an independent test set of 480 images showed good performance in detecting poor quality images, with a balanced accuracy of 79%. We believe that our quality control models will help to overcome the first barrier of heterogeneity of image quality in clinical data warehouses and thus advance research using images acquired in clinical routine.

## Acknowledgments

The research was done using the Clinical Data Warehouse of the Greater Paris University Hospitals. The authors are grateful to the members of the AP-HP WIND and URC teams, and in particular Yannick Jacob, Julien Dubiel, Antoine Rozès, Cyrina Saussol, Rafael Gozlan, Stéphane Bréant, Florence Tubach, Jacques Ropers, Christel Daniel, and Martin Hilka. They would also like to thank the "Collégiale de Radiologie of AP-HP" as well as, more generally, all the radiology departments from AP-HP hospitals.

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the 'France 2030' program (reference ANR-23-IACL-0008, PRAIRIE-PSAI) and as part of the 'Investissements d'avenir' program (reference ANR-19-P3IA-0001, PRAIRIE 3IA Institute and reference ANR-10-IAIHU-06, Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) and was supported by a grant from Inserm and the French Ministry of Health in the context of the MESSIDORE 2023 call operated by IReSP (reference AAP-2023-MSDR-341011).

## Disclosure statement

Competing financial interests related to the present article: none to disclose for all authors.

Competing financial interests unrelated to the present article: OC reports having received consulting fees from AskBio and Therapanacea and having received fees for writing a lay audience short paper from Expression Santé. Members from his laboratory have co-supervised a PhD thesis with myBrainTechnologies and with Qynapse. O.C. holds a patent registered at the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices).

O.C. is an Associate Editor of Medical Image Analysis.

## APPRIMAGE Study Group

Olivier Colliot, Ninon Burgos, Simona Bottani, Sophie Loizillon [1]
Didier Dormont [1,2], Stéphane Lehéricy[2,21,22], Samia Si Smail Belkacem, Sebastian Ströer [2]
Nathalie Boddaert [3]
Farida Benoudiba, Ghaida Nasser, Claire Ancelet, Laurent Spelle [4]
Aurélien Maire, Stéphane Bréant, Christel Daniel, Martin Hilka, Yannick Jacob, Julien Dubiel, Cyrina Saussol, Rafael Gozlan [19]
Florence Tubach, Jacques Ropers, Antoine Rozès, Camille Nevoret [20]
Hubert Ducou-Le-Pointe[5], Catherine Adamsbaum[6], Marianne Alison[7], Emmanuel Houdart[8], Robert Carlier [9,17], Myriam Edjlali[9], Betty Marro[10,11], Lionel Arrive[10], Alain Luciani[12], Antoine Khalil[13] , Elisabeth Dion[14], Laurence Rocher[15], Pierre-Yves Brillet[16] , Paul Legmann, Jean-Luc Drape [18]

[1] Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, Inria, Aramis project-team, F-75013, Paris, France
[2] AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France
[3] AP-HP, Hôpital Necker, Department of Radiology, F-75015, Paris, France
[4] AP-HP, Hôpital Bicêtre, Department of Radiology, F-94270, Le Kremlin-Bicêtre, France
[5] AP-HP, Hôpital Armand-Trousseau, Department of Radiology, F-75012, Paris, France
[6] AP-HP, Hôpital Bicêtre, Department of Pediatric Radiology, F-94270, Le Kremlin-Bicêtre, France
[7] AP-HP, Hôpital Robert-Debré, Department of Radiology, F-75019, Paris, France
[8] AP-HP, Hôpital Lariboisière , Department of Neuroradiology, F-75010, Paris, France
[9] AP-HP, Hôpital Raymond-Poincaré, Department of Radiology, F-92380, Garches, France
[10] AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75012, Paris, France
[11] AP-HP, Hôpital Tenon, Department of Radiology, F-75020, Paris, France
[12] AP-HP, Hôpital Henri-Mondor, Department of Radiology, F-94000, Créteil, France
[13] AP-HP, Hôpital Bichat, Department of Radiology, F-75018, Paris, France
[14] AP-HP, Hôpital Hôtel-Dieu, Department of Radiology, F-75004, Paris, France
[15] AP-HP, Hôpital Antoine-Béclère, Department of Radiology, F-92140, Clamart, France
[16] AP-HP, Hôpital Avicenne, Department of Radiology, F-93000, Bobigny, France
[17] AP-HP, Hôpital Ambroise Paré, Department of Radiology,

F-92100 104, Boulogne-Billancourt, France
[18] AP-HP, Hôpital Cochin, Department of Radiology, F-75014, Paris, France
[19] AP-HP, WIND department, F-75012, Paris, France
[20] AP-HP, Unité de Recherche Clinique, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, F-75013, Paris, France
[21] ICM, Centre de NeuroImagerie de Recherche – CENIR, Paris, France
[22] Sorbonne Université, Institut du Cerveau – Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

## References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage 166, 400–424. doi:https://doi.org/10.1016/j.neuroimage.2017.10.034.

Artemova, S., Madiot, P.E., Caporossi, A., Mossuz, P., Moreau-Gaudry, A., Group, P., et al., 2019. Predimed: clinical data warehouse of grenoble alpes university hospital, in: MEDINFO 2019: Health and Wellbeing e-Networks for All. IOS Press, pp. 1421–1422.

Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., Gee, J.C., 2014. The Insight ToolKit image registration framewor. Frontiers in Neuroinformatics 8.

Beaufrère, A., Ouzir, N., Zafar, P.E., Laurent-Bellue, A., Albuquerque, M., Lubuela, G., Grégory, J., Guettier, C., Mondet, K., Pesquet, J.C., et al., 2024. Primary liver cancer classification from routine tumour biopsy using weakly supervised deep learning. JHEP Reports , 101008.

Berenbaum, A., Delingette, H., Maire, A., Poret, C., Hassen-Khodja, C., Bréant, S., Daniel, C., Martel, P., Grimaldi, L., Frank, M., et al., 2023. Performance of ai-based automated classifications of whole-body fdg pet in clinical practice: The clariti project. Applied Sciences 13, 5281.

Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E., 2023. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proceedings of the National Academy of Sciences 120, e2216399120. doi:10.1073/pnas.2216399120.

Bink, A., Schmitt, M., Gaa, J., Mugler, J.P., Lanfermann, H., Zanella, F.E., 2006. Detection of lesions in multiple sclerosis by 2d flair and single-slab 3d flair sequences at 3.0 t: initial results. European radiology 16, 1104–1110.

Bottani, S., 2022. Machine learning for neuroimaging using a very large scale clinical data warehouse. Ph.D. thesis. Sorbonne Université. URL: https://theses.hal.science/tel-03671129.

Bottani, S., Burgos, N., Maire, A., Saracino, D., Ströer, S., Dormont, D., Colliot, O., 2023. Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse. Medical Image Analysis 89, 102903. doi:10.1016/j.media.2023.102903.

Bottani, S., Burgos, N., Maire, A., Wild, A., Strer, S., Dormont, D., Colliot, O., 2021. Automatic quality control of brain T1-weighted magnetic resonance images for a clinical data warehouse. Medical Image Analysis , 102219doi:10.1016/j.media.2021.102219.

Bottani, S., Thibeau-Sutre, E., Maire, A., Ströer, S., Dormont, D., Colliot, O., Burgos, N., 2024. Contrast-enhanced to non-contrast-enhanced image translation to exploit a clinical data warehouse of T1-weighted brain MRI. BMC Medical Imaging 24, 67. doi:10.1186/s12880-024-01242-3.

Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa: Ii. resolving the paradoxes. Journal of clinical epidemiology 43, 551–558.

Corcuera-Solano, I., Doshi, A., Pawha, P., Gui, D., Gaddipati, A., Tanenbaum, L., 2015. Quiet propeller mri techniques match the quality of conventional propeller

brain imaging techniques. American Journal of Neuroradiology 36, 1124–1127.

Daniel, C., Salamanca, E., 2020. Hospital Databases, in: Healthcare and Artificial Intelligence. Springer, pp. 57–67.

De Coene, B., Hajnal, J.V., Gatehouse, P., Longmore, D.B., White, S.J., Oatridge, A., Pennock, J., Young, I., Bydder, G., 1992. Mr of the brain using fluid-attenuated inversion recovery (flair) pulse sequences. American journal of neuroradiology 13, 1555–1564.

Doutreligne, M., Degremont, A., Jachiet, P.A., Lamer, A., Tannier, X., 2023. Good practices for clinical data warehouse implementation: A case study in france. PLOS Digital Health 2, e0000298.

Duchesne, S., Chouinard, I., Potvin, O., Fonov, V.S., Khademi, A., Bartha, R., Bellec, P., Collins, D.L., Descoteaux, M., Hoge, R., et al., 2019. The canadian dementia imaging protocol: harmonizing national cohorts. Journal of Magnetic Resonance Imaging 49, 456–465.

Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., Gorgolewski, K.J., 2017. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PLoS ONE 12, e0184661. doi:10.1371/journal.pone.0184661.

Feng, Y., Wang, Z., Xu, X., Wang, Y., Fu, H., Li, S., Zhen, L., Lei, X., Cui, Y., Ting, J.S.Z., 2023. Contrastive domain adaptation with consistency match for automated pneumonia diagnosis. Medical Image Analysis 83, 102664.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. Journal of Machine Learning Research 17, 2096–2030.

Gehring, S., Eulenfeld, R., 2018. German medical informatics initiative: unlocking data for research and health care. Methods of information in medicine 57, e46–e49.

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., et al., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific data 3, 1–9.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization, in: Saul, L., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems, MIT Press. pp. 1–8. URL: https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf.

Griffanti, L., Bhalerao, G.V., Gillis, G., Dembele, M., Suri, S., Ebmeier, K.P., Klein, J.C., Hu, M., Mackay, C., 2023. Automated quality control of structural brain mri scans from aging and dementia datasets. Alzheimer's & Dementia 19, e081937.

HassanPour Zonoozi, M., Seydi, V., 2022. A Survey on Adversarial Domain Adaptation. Neural Processing Letters , 1–41.

Hendriks, J., Mutsaerts, H.J., Joules, R., Peña-Nogales, Ó., Rodrigues, P.R., Wolz, R., Burchell, G.L., Barkhof, F., Schrantee, A., 2024. A systematic review of (semi-) automatic quality control of t1-weighted mri scans. Neuroradiology 66, 31–42.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62, 782–790.

Jiang, P., Wu, A., Han, Y., Shao, Y., Qi, M., Li, B., 2020. Bidirectional Adversarial Training for Semi-Supervised Domain Adaptation., in: IJCAI, pp. 934–940.

Karadeli, H.H., Giurgiutiu, D.V., Cloonan, L., Fitzpatrick, K., Kanakis, A., Ozcan, M.E., Schwamm, L.H., Rost, N.S., 2016. Flair vascular hyperintensity is a surrogate of collateral flow and leukoaraiosis in patients with acute stroke due to proximal artery occlusion. Journal of Neuroimaging 26, 219–223.

Kim, T., Kim, C., 2020. Attract, perturb, and explore: Learning a feature alignment network for

semi-supervised domain adaptation, in: Computer Vision–ECCV 2020, Springer. pp. 591–607.

Krupa, K., Bekiesińska-Figatowska, M., 2015. Artifacts in magnetic resonance imaging. Polish journal of radiology 80, 93.

Lamer, A., Moussa, M.D., Marcilly, R., Logier, R., Vallet, B., Tavernier, B., 2023. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. Journal of Clinical Monitoring and Computing 37, 461–472.

Lavdas, E., Tsougos, I., Kogia, S., Gratsias, G., Svolos, P., Roka, V., Fezoulidis, I.V., Kapsalaki, E., 2014. T2 flair artifacts at 3-t brain magnetic resonance imaging. Clinical imaging 38, 85–90.

Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C., 2016. The first step for neuroimaging data analysis: Dicom to nifti conversion. Journal of neuroscience methods 264, 47–56.

Loizillon, S., Bottani, S., Mabille, S., Jacob, Y., Maire, A., Ströer, S., Dormont, D., Colliot, O., Burgos, N., 2024a. Automated MRI Quality Assessment of Brain T1-weighted MRI in Clinical Data Warehouses: A Transfer Learning Approach Relying on Artefact Simulation. Machine Learning for Biomedical Imaging 2, 888–915. doi:10.59275/j.melba.2024-7fgd.

Loizillon, S., Bottani, S., Maire, A., Ströer, S., Dormont, D., Colliot, O., Burgos, N., 2024b. Automatic motion artefact detection in brain T1-weighted magnetic resonance images from a clinical data warehouse using synthetic data. Medical Image Analysis 93, 103073. doi:10.1016/j.media.2023.103073.

Loizillon, S., Colliot, O., Chougar, L., Stroer, S., Jacob, Y., Maire, A., Dormont, D., Burgos, N., 2023. Semi-supervised domain adaptation for automatic quality control of flair mris in a clinical data warehouse, in: MICCAI Workshop on Domain Adaptation and Representation Transfer, Springer. pp. 84–93.

Naganawa, S., 2015. The technical and clinical features of 3d-flair in neuroimaging. Magnetic Resonance in Medical Sciences 14, 93–106.

Nordlinger, B., Villani, C., Rus, D., 2020. Healthcare and Artificial Intelligence. Springer International Publishing.

Peltonen, J.I., Mäkelä, T., Salli, E., 2018. Mri quality assurance based on 3d flair brain images. Magnetic Resonance Materials in Physics, Biology and Medicine 31, 689–699.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical parametric mapping: the analysis of functional brain images. Elsevier.

Pipe, J.G., 1999. Motion correction with propeller mri: application to head motion and free-breathing cardiac imaging. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 42, 963–969.

Roels, J., Hennies, J., Saeys, Y., Philips, W., Kreshuk, A., 2019. Domain adaptive segmentation in volume electron microscopy imaging, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE. pp. 1519–1522.

Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., Wen, J., Wild, A., Habert, M.O., Durrleman, S., Colliot, O., 2021. Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. Frontiers in Neuroinformatics 15, 689675. doi:10.3389/fninf.2021.689675.

Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K., 2019. Semi-supervised domain adaptation via minimax entropy, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8050–8058.

Singh, A., 2021. Clda: Contrastive learning for semi-supervised domain adaptation. Advances in Neural Information Processing Systems 34, 5089–5101.

Storelli, L., Rocca, M.A., Pantano, P., Pagani, E., De Stefano, N., Tedeschi, G., Zaratin, P., Filippi, M., Valsasina, P., Sibilia, M., Preziosa, P., Gallo, A., Bisecco, A., Docimo, R., Petsas, N., Ruggieri, S., Tommasin, S.,

Stromillo, M.L., Brocci, R.T., for the INNI Network, 2019. Mri quality control for the italian neuroimaging network initiative: moving towards big data in multiple sclerosis. Journal of Neurology 266, 2848–2858.

Sujit, S.J., Coronado, I., Kamali, A., Narayana, P.A., Gabr, R.E., 2019. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. J. Magn. Reson. Imaging 50, 1260–1267.

Sundaresan, V., Zamboni, G., Dinsdale, N.K., Rothwell, P.M., Griffanti, L., Jenkinson, M., 2021. Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain MR images. Medical Image Analysis 74, 102215.

Tanenbaum, L.N., Tsiouris, A.J., Johnson, A.N., Naidich, T.P., DeLano, M.C., Melhem, E.R., Quarterman, P., Parameswaran, S., Shankaranarayanan, A., Goyen, M., et al., 2017. Synthetic mri for clinical neuroimaging: results of the magnetic resonance image compilation (magic) prospective, multicenter, multireader trial. American journal of neuroradiology 38, 1103–1110.

Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., Burgos, N., 2022. ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing. Computer Methods and Programs in Biomedicine 220, 106818. doi:10.1016/j.cmpb.2022.106818.

Thota, M., Leontidis, G., 2021. Contrastive domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2209–2218.

Tsuchida, A., Laurent, A., Crivello, F., Petit, L., Joliot, M., Pepe, A., Beguedou, N., Gueye, M.F., Verrecchia, V., Nozais, V., et al., 2021. The mri-share database: brain imaging in a cross-sectional cohort of 1870 university students. Brain Structure and Function 226, 2057–2085.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: Improved N3 Bias Correction. IEEE T. Med. Imaging 29, 1310–1320. doi:10.1109/TMI.2010.2046908.

Vanderbecq, Q., Gelard, M., Pesquet, J.C., Wagner, M., Arrive, L., Zins, M., Chouzenoux, E., 2024. Deep learning for automatic bowel-obstruction identification on abdominal ct. European Radiology , 1–12.

Wack, M., 2017. Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy: déploiement technique, intégration et gouvernance des données. Ph.D. thesis. Université de Lorraine.

Watson, P., Petrie, A., 2010. Method agreement analysis: a review of correct methodology. Theriogenology 73, 1167–1179.

Winkler, Kochunov P, G.D., . Flair templates. URL: http://brainder.org.

Zhao, H., Des Combes, R.T., Zhang, K., Gordon, G., 2019. On learning invariant representations for domain adaptation, in: International Conference on Machine Learning, PMLR. pp. 7523–7532.

# Automatic quality control of brain 3D FLAIR MRIs for a clinical data warehouse

Sophie Loizillon, Simona Bottani, Aurélien Maire, Sebastian Ströer, Lydia Chougar, Didier Dormont, Olivier Colliot, Ninon Burgos and the APPRIMAGE Study Group

-

# Supplementary Material

## S1. Supplementary figures and tables

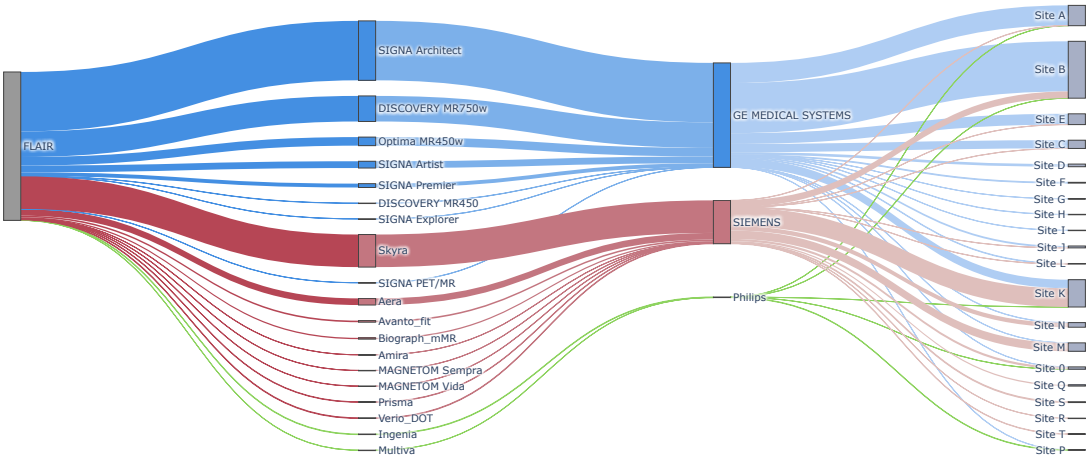



Figure S1: Sankey plot analysis of FLAIR MRIs highlighting the scanner model, manufacturer and site distribution. The 13,703 FLAIR images available for analysis were acquired on 25 different scanner models from three manufacturers, across 20 hospital sites.

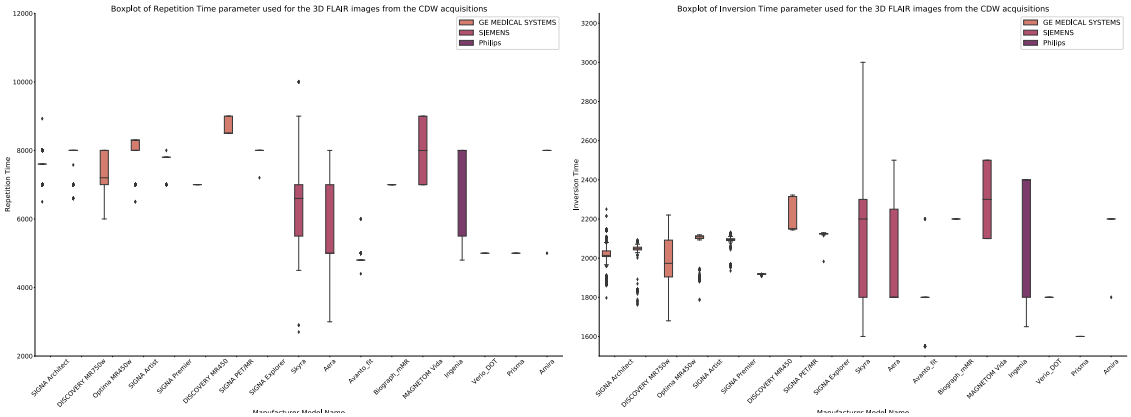



Figure S2: Boxplots of repetition time (left) and inversion time (right) acquisition parameters used for the 13,703 FLAIR images from the CDW composing our clinical routine dataset.

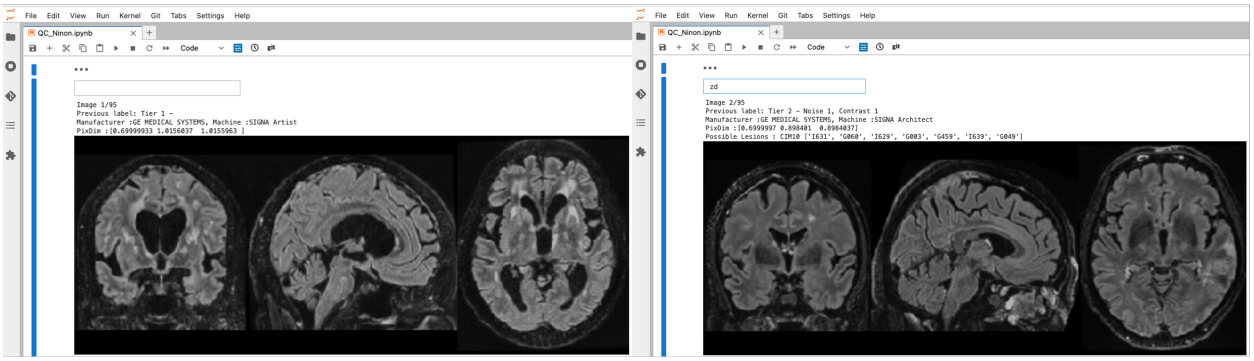



Figure S3: Graphical interface in a Jupyter notebook used to perform the manual image quality annotations. We display the scanner model and the manufacturer during annotation to inform the annotator and thus limit the introduction of bias during the annotation process. To guide the manual annotators, we also provide the ICD-10 codes corresponding to the presence of brain lesions associated with the patients within three months of the MRI acquisition date.

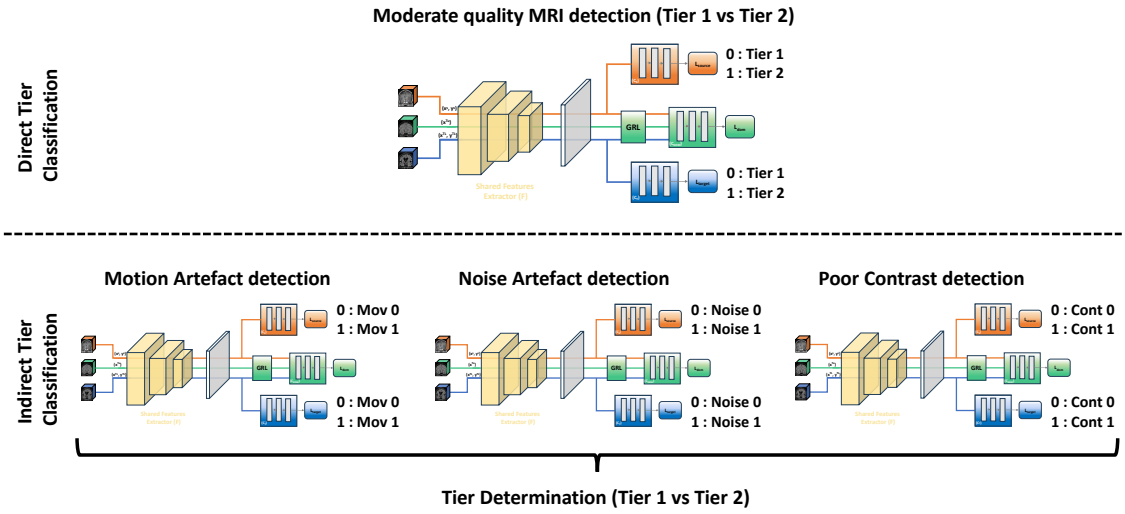



Figure S4: Comparison of the direct and indirect tier classification workflows. In the direct tier classification, our proposed semi-supervised domain adaption architecture is directly trained to detect the tier of the images. In the indirect tier classification, three artefact-specific models are trained to detect each type of artefacts. The quality tiers are then determined using the outputs of the three artefact-specific models following the same rules as for the manual annotation (see Table 2).

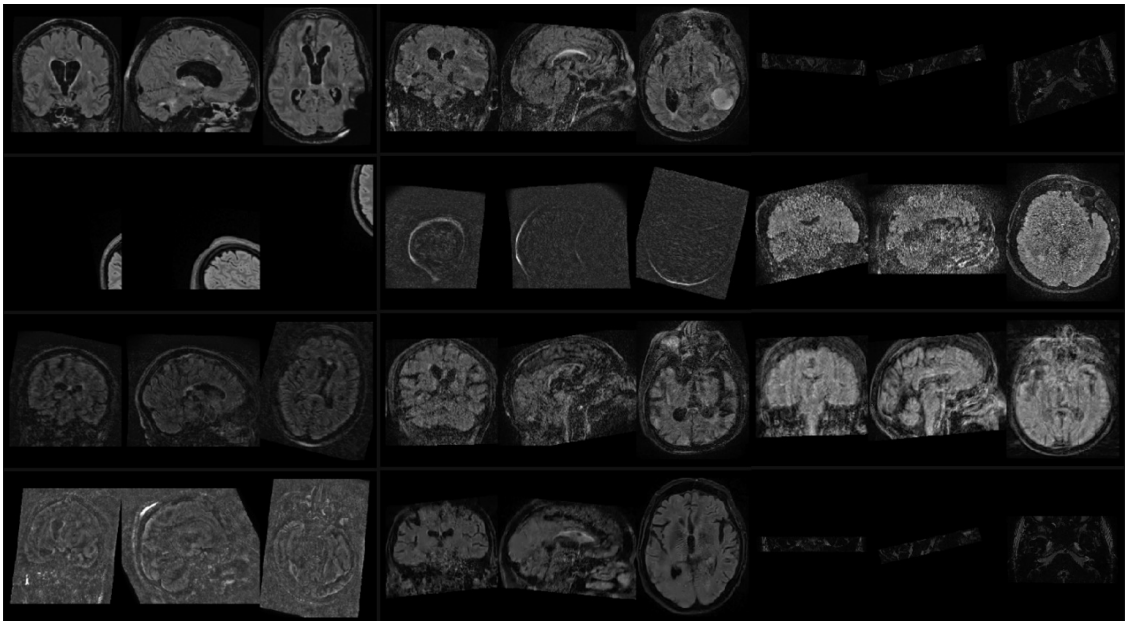


Figure S5: Example of manually annotated straight reject images from the AP-HP clinical data warehouse.

Table S1: Distribution of the training, validation and test sets separately for the moderate artefact detection task using the clinical dataset.

| | | Motion detection | | Contrast detection | | Noise detection | |
|---|---|---|---|---|---|---|---|
| Modality | Split | Mov0 | Mov1 | Cont0 | Cont1 | Noise0 | Noise1 |
| T1 | Train | 1681 | 859 | 1111 | 765 | 1949 | 865 |
| | Validation | 428 | 219 | 273 | 185 | 496 | 232 |
| | Test | 210 | 118 | 135 | 103 | 258 | 125 |
| FLAIR | Train | 318 | 66 | 151 | 53 | 166 | 35 |
| | Validation | 35 | 8 | 34 | 12 | 38 | 8 |
| | Test | 474 | 20 | 305 | 121 | 422 | 78 |

Table S2: Distribution of the training, validation and test sets separately for the severe artefact detection task using the clinical dataset.

| | | Motion detection | | Contrast detection | |
|---|---|---|---|---|---|
| Modality | Split | Mov0/1 | Mov2 | Cont0/1 | Cont2 |
| T1 | Train | 2540 | 379 | 1876 | 1055 |
| | Validation | 647 | 94 | 458 | 271 |
| | Test | 328 | 57 | 238 | 147 |
| FLAIR | Train | 238 | 5 | 204 | 46 |
| | Validation | 54 | 1 | 41 | 9 |
| | Test | 494 | 6 | 426 | 74 |

Table S3: Results for the detection of bad quality images (*tier 3 vs tiers 2-1*) within the T1w and FLAIR test sets from the CDW. We report the mean and empirical standard deviation across the five folds of the balanced accuracy, which is defined as the mean of the specificity and sensitivity. For Manual Annotation, the balanced accuracy corresponds to the average balanced accuracy of the two annotators with the consensus.

| Approaches | T1w | FLAIR |
|---|---|---|
| Manual annotation | 91.56 | 86.54 |
| Semi-supervised DANN (Sundaresan et al., 2021) | 81.97 ± 1.39 | 66.91 ± 2.37 |
| Entropy minimisation (Grandvalet and Bengio, 2004) | 80.07 ± 1.99 | 62.21 ± 4.45 |
| Mini-max entropy (Saito et al., 2019) | 77.39 ± 3.82 | 63.31 ± 4.16 |
| Proposed approach | 80.59 ± 1.62 | 76.81 ± 0.68 |

Table S4: Balanced accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (PPV) of the indirect (proposed) and direct classification approaches for the three tasks: straight reject detection, tier 3 vs tiers 2-1 and tier 2 vs tier 1.

| Approach | Indirect tier classification | | Direct tier classification | | |
|---|---|---|---|---|---|
| Task | Tier 2 vs tier 1 | Tier 3 vs tiers 2-1 | Tier 2 vs tier 1 | Tier 3 vs tiers 2-1 | Straight reject |
| Balanced Accuracy | 73.92 | 79.81 | 68.08 | 76.13 | 89.27 |
| Sensitivity | 81.18 | 82.54 | 84.90 | 82.13 | 98.54 |
| Specificity | 66.67 | 77.09 | 51.27 | 70.13 | 80.00 |
| PPV | 57.50 | 41.27 | 72.98 | 93.50 | 99.16 |
| NPV | 86.44 | 95.77 | 68.64 | 42.86 | 69.57 |

## S2. Ethics approval, procedure and regulations allowing the access and the use of patient data

The AP-HP obtained the authorization of the CNIL (Commission Nationale de l'informatique et des Libertés, the French regulatory body for data collection and management) in 2017 to share data for research purposes in compliance with the MR004 reference methodology (Daniel and Salamanca, 2020). The MR004 reference control data processing for the purpose of studying, evaluating and/or researching that does not involve human persons (in the sense of not involving an intervention or a prospective collection of research data in patients that would not be necessary for clinical evaluation, but which allows retrospective use of data previously acquired in patients). The goals of the clinical data warehouse are the development of decision support algorithms, the support of clinical trials and the promotion of multi-centre studies. According to French regulation, and as authorised by the CNIL, patients' consent to use their data in the projects of the CDW can be waived as these data were acquired as part of the clinical routine care of the patients. At the same time, AP-HP committed to keep patients updated about the different research projects of the clinical data warehouse through a portal on the internet [2] and individual information is systematically provided to all the patients admitted to the AP-HP. In addition, a retrospective information campaign was conducted by the AP-HP in 2017: it involved around 500,000 patients who were contacted by e-mail and by postal mail to be informed of the development of the CDW.

Accessing the data is possible with the following procedure. A detailed project must be submitted to the Scientific and Ethics Board of the AP-HP. If the project participants are external to AP-HP, they have to sign a contract with the Clinical Research and Innovation Board (Direction de la Recherche Clinique et de l'Innovation). The project must

[2] https://eds.aphp.fr/recherches-en-cours

include the goals of the research, the different steps that will be pursued, a detailed description of the data needed, of the software tools necessary for the processing, and a clear statement of the public health benefits. Once the project is approved, the research team is granted access to the Big Data Platform (BDP), which was created by a sub-department of the IT of the AP-HP. The BDP is a platform internal to the AP-HP where data are collected and that external users can access to perform all their analyses, in accordance with the CNIL regulation. It is strictly forbidden to export any kind of data and each user can access only a workspace that is specific to their project. Each person of the research team can access the BDP with an AP-HP account after two-factor authentication. If the research team includes people that are not employed by the AP-HP, a temporary account associated to the project is activated. The project on which the proposed work is based is called APPRIMAGE, it is led by the ARAMIS team (current AP-HP PI: Didier Dormont; initial AP-HP PI: Anne Bertrand, deceased March 2nd 2018) at the Paris Brain Institute and it was approved by the Scientific and Ethics Board of the AP-HP in 2018 (Bottani, 2022).