CellPress

Neuron

Review

Challenges for machine learning in clinical translation of big data imaging studies

Nicola K. Dinsdale,[1,2,*] Emma Bluemke,[3] Vaanathi Sundaresan,[1,4] Mark Jenkinson,[1,5,6] Stephen M. Smith,[1] and Ana I.L. Namburete[1,2]

[1]Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[2]Oxford Machine Learning in NeuroImaging Lab, OMNI, Department of Computer Science, University of Oxford, Oxford, UK
[3]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
[4]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA, USA
[5]Australian Institute for Machine Learning (AIML), School of Computer Science, University of Adelaide, Adelaide, SA, Australia
[6]South Australian Health and Medical Research Institute (SAHMRI), North Terrace, Adelaide, SA, Australia
*Correspondence: nicola.dinsdale@cs.ox.ac.uk
https://doi.org/10.1016/j.neuron.2022.09.012

SUMMARY

Combining deep learning image analysis methods and large-scale imaging datasets offers many opportunities to neuroscience imaging and epidemiology. However, despite these opportunities and the success of deep learning when applied to a range of neuroimaging tasks and domains, significant barriers continue to limit the impact of large-scale datasets and analysis tools. Here, we examine the main challenges and the approaches that have been explored to overcome them. We focus on issues relating to data availability, interpretability, evaluation, and logistical challenges and discuss the problems that still need to be tackled to enable the success of "big data" deep learning approaches beyond research.

INTRODUCTION

The majority of neuroimaging datasets have been limited to small-scale, low-N collections, typically focusing on a specific research question or clinical population of interest. However, large-scale "big data" collections of a wide range of subjects have begun to be collated, many of which are openly available to researchers. This means that if the acquisition protocol, demographic, and non-imaging data meet the requirements of a given study, novel research can be completed without acquiring new scans. Sharing these large-scale datasets has had many benefits: they enable exploration of new research questions and reproducible, rapid methodological prototyping.

Existing large-scale datasets have been curated to explore different research questions, with varying numbers of subjects and imaging sites across studies. For instance, if the research question were about lifespan and aging, datasets to consider would include UK Biobank (Sudlow et al., 2015) and CamCAN (Taylor et al., 2017). Similarly, if considering early development, available datasets include the developing HCP (dHCP) (Hughes et al., 2017) and the adolescent brain cognitive development (ABCD) (Marek et al., 2019); for research on young adults, one could consider HCP Young Adult (Van Essen et al., 2013). Datasets also exist that explore specific clinical groups, such as Alzheimer's disease (ADNI; Jack et al., 2008), schizophrenia, and bipolar disorder (CANDI; Frazier et al., 2008). These datasets allow the exploration of questions that would not be possible with traditional small-scale studies (e.g., with $N < 100$), which will not sufficiently represent variation within the population of interest. Large-scale studies have also enabled the characterization of potential subtypes within patient samples—for example, Young et al. (2018) demonstrated heterogeneity and subtypes in Alzheimer's-related atrophy patterns using data from ADNI.

UK Biobank (Sudlow et al., 2015), the largest of these studies, aims to collect brain imaging data from 100,000 volunteers, including 6 magnetic resonance imaging (MRI) modalities, to study structure, function, and connectivity. It contains a diverse range of lifestyle, genetic, and health measures, allowing researchers to create models of population aging and to explore how genetic and environmental factors interact with aging and disease. For instance, hippocampal atrophy is a well-validated biomarker for Alzheimer's disease; thus, using the UK Biobank, a nomogram of hippocampal volume with normal aging has been created (Nobis et al., 2019), illustrating the progression with age and percentiles of expected volume across for the healthy population, as a reference.

Due to the growth in size of these datasets, sophisticated deep learning models are finally a practical option for neuroimaging analysis, enabling the exploration of new questions in a data-driven manner. Powered by their ability to learn complex, non-linear relationships and patterns from data, deep learning methods have been applied to a wide range of applications, finding success in previously unsolved problems. However, this success has been limited to specific tasks and data domains. Challenges remain for applying deep learning models to the clinical domain, which currently limit the impact that big datasets such as the UK Biobank have on patient care. Work must be undertaken to allow models to extend beyond the research domain. Recent developments in deep learning have begun to tackle the problems faced, but further developments are needed. Here, we discuss the challenges faced, the current approaches being developed to mitigate them, and the barriers that

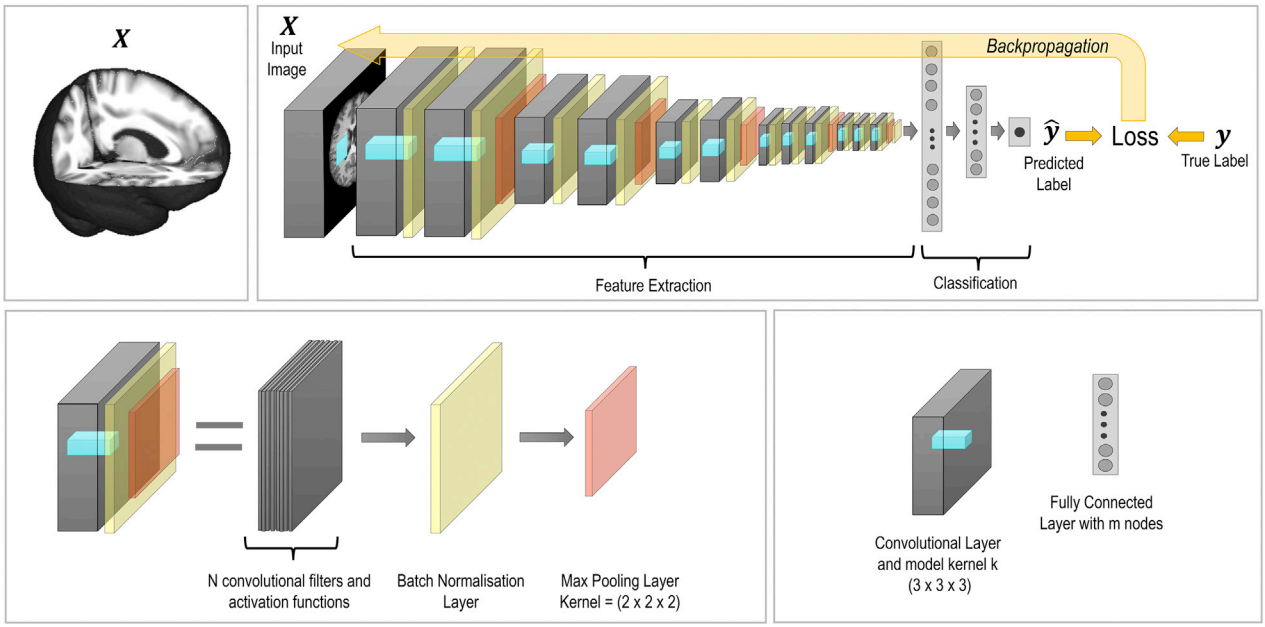

Figure 1. An example network architecture for a convolutional neural network (CNN) for a classification or regression task
The lower panel describes the building blocks used to form the network shown in the upper panel.

remain, including the challenges of data availability, interpretability, and model evaluation, and logistical challenges such as data privacy.

DEEP LEARNING BACKGROUND

To understand the challenges for clinical translatability of deep learning methods, we first require a brief overview of how these methods approach problems (for a more detailed introduction, see LeCun et al., 2015). We will only consider convolutional neural networks (CNNs), which form the vast majority of deep learning methods currently applied in medical imaging, an example architecture of which is shown in Figure 1. The majority are "supervised" approaches (LeCun et al., 2015), meaning that to explore the research question, we need a dataset of images, X, and a set of known true labels, y, for the task in question. This requires an understanding of both the information that we expect to be encoded within the images, and which questions are of interest (defined as "domain knowledge"). An example for the variables could be a structural brain scan (X) with the label being disease prognosis. The task is then to design a neural network architecture capable of mapping from X to y through learning a highly non-linear mapping function $f(X, y; W)$, where W are the trainable weights of the neural network.

The choice of architecture is highly influenced by factors such as the task being explored, the quantity of data, and the computational power available. Nevertheless, most networks are constructed from the same basic building blocks. First, "convolutional filters," which learn features of interest from the data ("feature extraction"). They contain the weights and biases to be learned during the optimization process. Stacks of these layers are placed at different spatial resolutions for a range of different features to be extracted at each level of abstraction. This hierarchical feature extraction allows a rich understanding of the input data. During the "forward pass" of the training procedure, each filter is convolved across the width and height of the input volume. The exact nature of the features is learned through a network optimization procedure that updates the filter weights to find features that are useful contributors to the overall goal of predicting y.

Next are the "activation functions," which play a fundamental role in model training by applying non-linear transformations to the learned features. This non-linearity provides a distinct edge to CNNs, allowing them to learn the complex non-linear relationships (or mapping) between the input and the output. Commonly used activation functions include rectified linear units (ReLUs, e.g., zeroing negative values and keeping positive ones unchanged) and sigmoid (e.g., squashing large values down to a predefined upper bound, typically between 0 and 1). Due to the CNN's sequential data flow, the features at a given depth are a non-linear combination of the previous features and the network parameters, whose values are learned during network training. Without activation functions, CNNs would only be able to train linear models.

Networks then learn features at different spatial resolutions through the inclusion of pooling blocks. Pooling provides a basic invariance to rotations and translations and has been demonstrated to improve the object detection capability of convolutional networks. The final key components of neural networks are "fully connected layers"—essential to many classification or regression architectures—which are normally placed at the end of a network and learn how to classify the extracted features.

By feeding the data through the network, we obtain an output prediction. To render these accurate, the weights of the network

must be optimized through "back propagation." To this end, we evaluate a "loss" or "cost function," which determines the error in the network prediction by comparing the prediction $\hat{y}$ and the true label y. The choice of loss function is task-dependent and plays a crucial role in the network performance.

Thus, we have an optimization problem, the performance of which is highly dependent on two factors: first, the design decisions made about the network architecture and the loss function; second, the data available to train the network. Nearly all relevant techniques have been developed in computer vision, where very large datasets are available and easily curated, for instance, by scraping the internet. In neuroimaging, data have to be labeled by a domain expert. This is one of many differences between neuroimaging and the computer vision field; many challenges are specific to working with neuroimaging data, especially when the aim is clinical translation.

DATA AVAILABILITY

For clinical translatability or for deep learning techniques to be applied to clinical research, data availability is a major limitation. Despite growth in the size of available datasets, the largest are still only of the order of tens of thousands, with a thousand images being commonly regarded as a large dataset. For many specific tasks, datasets exist only in the order of hundreds of subjects due to various factors, including the monetary and time costs of acquiring data, difficulties in sharing and/or pooling data across sites, and the fact that, for some conditions, insufficient patient numbers exist to create a dataset of any great size (Morid et al., 2021). For example, the frequently explored brain tumor segmentation (BraTS) dataset (Menze et al., 2014) only has data from 369 subjects available for training (2020 challenge data), in stark contrast to popular datasets from computer vision, such as ImageNet (Deng et al., 2009) (1,281,167 training examples) and MNIST (LeCun et al., 1998) (60,000 training examples). Simply by considering dataset size, it is clear that we are likely to be underpowered for training neural networks: highly parameterized, deep neural networks are very dependent on the amount of available training data (He et al., 2020). With performance generally improving as the number of data points is increased, they are more affected by the amount of available training data than classical machine learning techniques. This is due to the need to learn the useful features as well as the (highly non-linear) decision boundary (He et al., 2020), and thus techniques to overcome the lack of data are required, especially for clinical applications.

Maximizing the impact of available data

There has, therefore, been an increasing focus on developing techniques to facilitate more effective use of available data. A commonly used technique from computer vision is the use of large natural image datasets (Raghu et al., 2019), with ImageNet (Deng et al., 2009) being the most popular, to "pre-train" the network. This involves training the weights on a related task with more available data, so that the optimization starts from an informed place rather than a random initialization. Clearly, it might be useful to consider the information learned by the network at the different stages (Olah et al., 2018): the early layers learn features such as edges and simple textures, largely resembling Gabor filters, and are thus very general and applicable across different images, regardless of the target tasks (Yosinski et al., 2014). The final layers learn features which are far more task- and dataset-specific. Therefore, we can take a network pre-trained on the large, canonical dataset and use this to extract features that we then pass to a classifier, requiring only the final classifier layers to be trained or, more commonly, fine-tuning (re-training) the deeper layers to the specific task. This requires less data, as not only are we starting the optimization process from an informed point in the "parameter space" but also the very earliest layers can often be frozen (kept at their value and not updated during training), greatly reducing the number of weights in the model that need to be optimized. This process is referred to as "transfer learning"; it is a step frequently used to allow networks to be trained with smaller amounts of training data. Transfer learning can be performed across data domain (dataset), task, or both, depending on the datasets available for pre-training, and so may enable us to train models on the clinical data of interest and thus explore clinical research questions directly; e.g., Peng et al. (2021) showed that by training on UK Biobank data and then fine-tuning the model on the target dataset, they significantly improved age prediction performance.

Although standard practice is to use the huge datasets of natural images for pre-training, natural images have very different characteristics from many medical images, and therefore the features learned are not necessarily the most appropriate for the tasks being considered in neuroimaging (Raghu et al., 2019). For instance, natural images are often stored as RGB 2D images (3-channel images), whereas MR images are encoded as grayscale (single channel) 3D images. Also, in medical images, the location of structures could be informative, which is rarely true in natural images. Creating pre-trained networks for medical images has therefore been a focus, with Model Genesis (Zhou et al., 2019) creating a flexible architecture trained to complete multiple tasks, extracting features which aim to generalize across medical imaging tasks. Similarly, some works pre-train on large datasets, such as UK Biobank, for tasks such as age or sex prediction, where obtaining labels is relatively trivial (Lu et al., 2021), or on datasets for the same task with a dataset where more labels are available (Kushibar et al., 2019). Again, the aim is to learn features from another task that are also useful for the task of interest—features that generalize across tasks and information from a large dataset, which helps us to understand a smaller clinical dataset.

Other studies utilize self-supervised approaches, such as "contrastive representation learning," where general features of a dataset are learned, without labels, by teaching the model which data points are similar or different. These then act as the starting point for further model training on a smaller target dataset rather than pre-training the model on a different dataset. An example approach is presented in Chen et al. (2020), where the data have been "augmented" (small transformations applied to increase the size of the dataset, discussed below); the network is then trained to encode both the original and the augmented images into the same location in the feature space using a contrastive loss function (Hadsell et al., 2006) that learns features, describing the similarity between images. Different self-supervised methods and contrastive loss approaches have

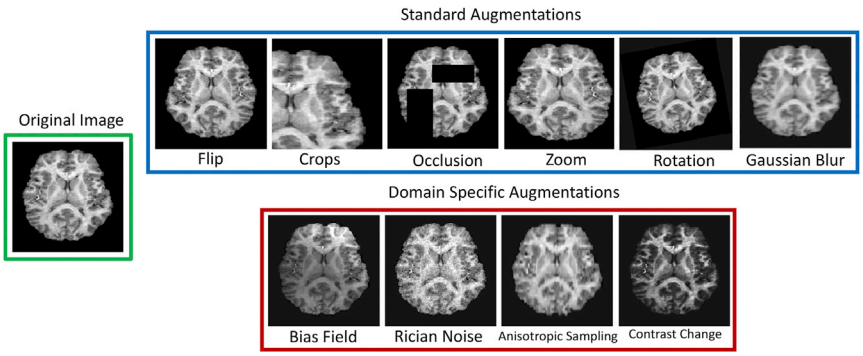

Figure 2. Example augmentations that might be applied to an MRI image
Standard augmentations come directly from computer vision approaches, and domain-specific augmentations for neuroimaging focus on variation that would be likely to be seen in MR images.

been developed and have begun to be applied in medical imaging (Zhang et al., 2020; Chaitanya et al., 2020), including the segmentation of MRI scans of the brain (Chen et al., 2019a).

Data augmentation

CNNs, however, still ultimately require a reasonable amount of data (hundreds or thousands of samples) in the target data domain, as at least some of the network parameters must be fine-tuned to optimize the prediction performance for the specific dataset and task. Even though the amount of data required is likely to be reduced, the degree of reduction will be determined by the similarity between the proxy and target tasks (He et al., 2019); the amount required may remain greater than is available. In this circumstance, data augmentation is often applied (Simard et al., 1998), artificially increasing the size and diversity of the training dataset by applying transformations, creating slightly perturbed versions of the data. Fundamentally, data augmentation enables us to artificially create a larger dataset which can be used to train the model, potentially enabling exploration directly with clinical data.

Augmentations (Figure 2) can take the form of basic transformations, such as flips and rotations (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015), as standardly applied in computer vision tasks, to more extreme examples such as Mixup (Zhang et al., 2017), which merges images from different classes to form hybrid classes, or generative networks such as conditional generative adversarial networks (GANs)—networks trained to generate simulated data (Mirza and Osindero, 2014). Although most deep learning studies apply data augmentation during training, some studies explore this for neuroimaging specifically; e.g., augmentation can be achieved through GANs being used to generate additional meaningful datapoints (Wu et al., 2020), or registration to templates (Nguyen et al., 2020), which generate biologically plausible transformations of the data. Similarly, they can be produced by identifying augmentations that are plausible across sites and scanners (Billot et al., 2020a), such as applying bias field.

Existing literature suggests that performing augmentations, even transformations that create images beyond realistic variation (Billot et al., 2020b), helps the network to generalize better to unseen data at test time. However, data augmentation must be used cautiously, so that the transformations applied do not change the validity of the label associated with the image. Consider, for instance, classifying Alzheimer's disease from structural MRI: the key indicator could be the atrophy of the hippocampus, so if any transformations are applied during the augmentation process that affect this region (e.g., local elastic deformations), it must be ensured that the level of atrophy is not affected and, thus, the true label changed. Ensuring this requires high levels of specific domain knowledge and can limit the augmentations that can be applied.

Patch or slice-based sampling

Other approaches to solving the shortage of available training data focus on breaking the input data down into patches, e.g., Wachinger et al. (2018), or slices (where the data are 3D), with many studies treating MRI data as 2D inputs, where each slice is treated as a separate training sample, e.g., Livne et al. (2019). This approach can vastly increase the amount of available data and can be especially effective for segmentation tasks where we have voxel-level labels. However, fragmenting the image can lead to the loss of global information; when they can be implemented, fully 3D networks have in most cases provided better results (Kamnitsas et al., 2017). Patch-wise or slice-wise approaches cannot necessarily be applied to classification tasks; where a single label is provided for the whole image, it may not hold for a given patch or slice of the image (Khagi et al., 2018).

Differences between datasets or data domain shift

Having sufficient data to train the model, however, is only the first difficulty for clinical application. The flexibility that allows deep learning methods to learn complex and highly non-linear mappings between the input images and the labels comes at a cost: deep learning methods are prone to overfitting to the training data (Srivastava et al., 2014); this is exacerbated if the amount of training data is insufficient. Further, although a well-trained model should interpolate well to data that falls within the same distribution as that seen during training, the performance degrades quickly once it must extrapolate to out-of-distribution data. Even perturbations unnoticeable to the human eye can cause network performance to collapse (Papernot et al., 2017). For clinical translatability, we need generalizability from the training set to all other reasonable datasets, including future datasets as yet uncollected; otherwise, a result may be a function of domain drift rather than of the subject's pathology.

Multisite datasets, such as the ABIDE study (Di Martino et al., 2014), still show an increase in non-biological variance when we pool data across sites and scanners (Yu et al., 2018). A demonstration of this variance leading to performance degradation for a segmentation task is shown in Figure 3. Multiple studies have confirmed this variation, identifying causes (batch effects) from

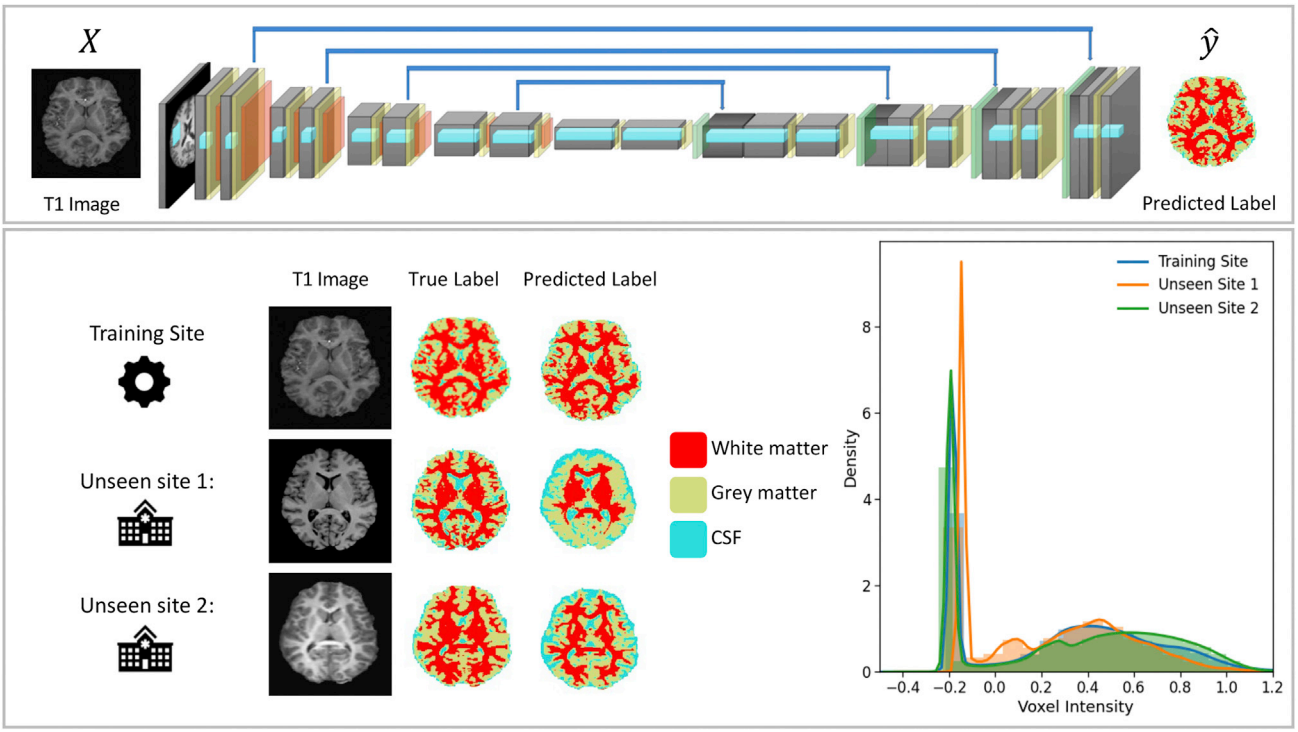

Figure 3. Demonstration of domain shift

To demonstrate the effect of the difference between domain datasets or domain shift, tissue segmentation was carried out on data from three sites collected as part of the ABIDE (Di Martino et al., 2014) multisite dataset. Although the data were all collected as part of one study, differences exist between the data collected at different sites due to scanner differences. The architecture used was a 3D UNet (Cicek et al., 2016) with T1 as the input image; only images from one site were used during training. The predictions can be seen, for example, images from three sites—one seen during training and two unseen. The segmentation for the site seen during training is good but suffers significant degradation when applied to the unseen sites, despite their being collected for the same study and having similar (normalized) voxel intensities, demonstrating the potential difficulties caused by domain shift.

scanner and acquisition differences, including scanner manufacturer (Han et al., 2006), scanner upgrade (Han et al., 2006), scanner drift (Takao et al., 2011), scanner magnet strength (Han et al., 2006), and gradient non-linearities (Jovicich et al., 2006). The removal of scanner-induced variance is therefore vital for neuroimaging studies, especially if models are to be applied to clinical datasets with a small number of subjects for any given site. Most deep learning approaches either use generative methods to output harmonized versions of the input data (Cetin Karayumak et al., 2019; Dewey et al., 2019) or aim to remove the scanner-related information from the features used to produce the predictions, for instance, using adversarial learning (Dinsdale et al., 2021b). These methods succeed in removing the scanner effects from the predictions but hold no guarantees for scanners not seen during training. Further, any harmonized output images are hard to validate without "traveling heads datasets" (images from the same subjects acquired on different scanners) (Moyer et al., 2020).

The domain shift experienced with multisite data is less than might be expected when we move between research and clinical data, or even just two datasets collected independently. The domain shift here can come from two sources: the scanner and acquisition or the demographics of the studies. First, MRI scans collected for research are often at a higher resolution and field-strength than clinical scans, which are designed to be more time efficient—both in terms of the time required for acquisition and for visual inspection—and are often collected at lower resolutions and field strengths. Also, research scans frequently have isotropic voxel sizes, whereas anisotropic voxels are still the norm in the clinic and present in the majority of legacy data (Iglesias et al., 2020). Unfortunately, due to the aforementioned paucity of training data, we are unlikely to be able to train sophisticated models directly and solely on clinical data in the near future.

Thus, methods being developed that consider this domain shift (e.g., between clinical and research data), focus either on domain adaptation approaches to create shared feature representations for the different datasets, or on synthesizing data to enable us to use the clinical domain. Domain adaptation techniques normally consider the situation where there is a large source dataset (e.g., a research dataset such as UK Biobank; Sudlow et al., 2015) and a much smaller target dataset (e.g., the clinical dataset of interest), and generally aim to force the learned features to have the same distribution from across sites such that information can be shared across datasets. Domain adaptation approaches have been applied for segmentation (Kamnitsas et al., 2017; Sundaresan et al., 2021) and classification problems (Guan et al., 2020). These methods can perform well on the target clinical data, through harnessing information from a large dataset to improve our understanding of a clinical dataset of interest, but further work is required to enable them to adapt reliably to higher numbers of datasets simultaneously.

Domain adaptation methods, at the extreme, essentially have the end goal that the network would work regardless of the acquisition, which is an active area of research (Billot et al., 2020c; Thakur et al., 2020). The other approach that has been explored is to use generative methods to convert the data from one domain to the other (Iglesias et al., 2020), such that the transformed data can be used in the existing model. Any generated images must be carefully validated to ensure that they convey the same information as the originals and that the outcomes are the same.

Data composition and algorithmic biases

Finally, we must consider that the demographics of study data frequently do not fully represent the population as a whole, so a domain shift is experienced when we attempt to move from, for example, the research domain to the clinical domain. Because research data are usually acquired with the targeted exploration of a certain study question in mind, the datasets rarely contain subjects with co-morbidities or incidental findings. For example, patients with advanced Alzheimer's disease are unlikely to be recruited for a general imaging study due to ethical implications such as the inability to consent (Clement et al., 2019). Also, a strong selection bias exists in both recruitment and completion, with studies having demonstrated biases in age, education, ancestry, geographic location, and health status (Clement et al., 2019). Furthermore, people with family connections to a given condition are more likely to volunteer for a study as a healthy control, leading to certain genetic markers being more prevalent in a study dataset than in the population as a whole (Hostage et al., 2013). Therefore, associations learned when considering research data may not generalize, and care must be taken in extrapolating any model trained on these datasets to clinical populations.

Models therefore suffer from algorithmic bias: that is, the outcomes of the model may potentially be systematically less favorable to, or have a lower performance on, individuals within a particular group, where no relevant difference (e.g., pathology) between groups exists to justify such effects (Paulus and Kent, 2020). Erroneous or unsuitable outcomes may be produced for groups less likely to be represented in the training data. As networks simply learn the patterns in the data, any bias in the data may be learned and encoded into the models.

Inevitably, when considering complicated questions with extremely heterogeneous populations, the datasets used to train the deep learning methods will be incomplete and insufficient in terms of spanning all possible modes of variability (Ning et al., 2020). For instance, pathologies will occur against a background of normal aging, with differences being present between individuals due to both processes. Sufficiently encompassing all of this variation is infeasible, due both to the number of subjects that would be required and to the difficulty in recruiting subjects from some specific groups. Thus, when models are developed for clinical translation, the limitations of the models must be understood; wherever groups are underrepresented, the appropriateness of the application of the model must be considered and any limitations identified.

INTERPRETABILITY AND TRUST

Performance degradation experienced with domain shift would be potentially less problematic were it not for another issue of deep learning methods: models will output a prediction for any data, but that prediction may not necessarily be meaningful. Lacking a "do not know" option, a neural network will output a prediction even if it is meaningless or the input nonsensical. For instance, if a random noise image is fed into a network trained to predict brain age, the network will predict an apparently valid age for the random noise (see Figure 4). While here, visually identifying the pure-noise image is trivial, where the network is trained for a more complicated classification tasks, identifying erroneous results is more difficult and requires both clinical and domain knowledge, leading to a critical question: can the results be trusted?

The majority would agree that, for deep learning methods to be used to determine patient care, they must be interpretable and interrogable. Interpretability is often defined as "the ability to provide explanations in understandable terms to a human." The explanations should, therefore, be logical decision rules that lead to a given diagnosis or patient care being chosen. This is especially important because neural networks have no semantic understanding of the problem they are being asked to solve. Thus, if spurious information (or "confounders") in X exists, which can aid in this mapping, then this information will probably be used, misleading the predictive potential of the network. Consider, for instance, the case where all subjects with a given pathology were collected on the same scanner. A network could then achieve 100% recall accuracy for this pathology by fitting to the scanner signal rather than by learning any information about the pathology (Winkler et al., 2019). It would then, in all probability, identify a healthy control from the same scanner as having the same pathology.

The effect of confounders would not be observed without further probing the behavior of the trained model—and probing networks is non-trivial. This has led to neural networks being commonly described as "blackbox" methods. There is a need for interpretable networks, allowing both understanding and scrutiny of decisions made, which with existing techniques is currently not possible. Although this may be acceptable for many computer vision tasks, interrogability is indispensable for clinical neuroimaging tasks. Approaches have been developed to try to enable some insight, which have broadly focused on two main areas: visualization and uncertainty.

Visualization

Visualization methods generally attempt to show which aspects of the input image led to the given classification—the salient regions—often by creating a "heatmap" of importance within the input image. Many of these methods are post hoc, taking a pretrained model and testing which regions of the image drove the model prediction. Most commonly, they analyze the gradients or activations of the network for a given input image, such as saliency maps (Simonyan et al., 2014) or layer-wise relevance propagation (Binder et al., 2016), and have been applied in a range of MRI analysis tasks to explain decision making (Böhle et al., 2019) and brain age prediction (Dinsdale et al., 2021a). However, concerns remain that these methods do not pass basic sanity checks and are not providing a valid insight into the model (Adebayo et al., 2018). Other methods are occlusion- or perturbation-based, where parts of the image are removed or altered in the input, whereafter heatmaps are generated to evaluate the effect of this perturbation on the network's performance (Zeiler and Fergus, 2014). Most of

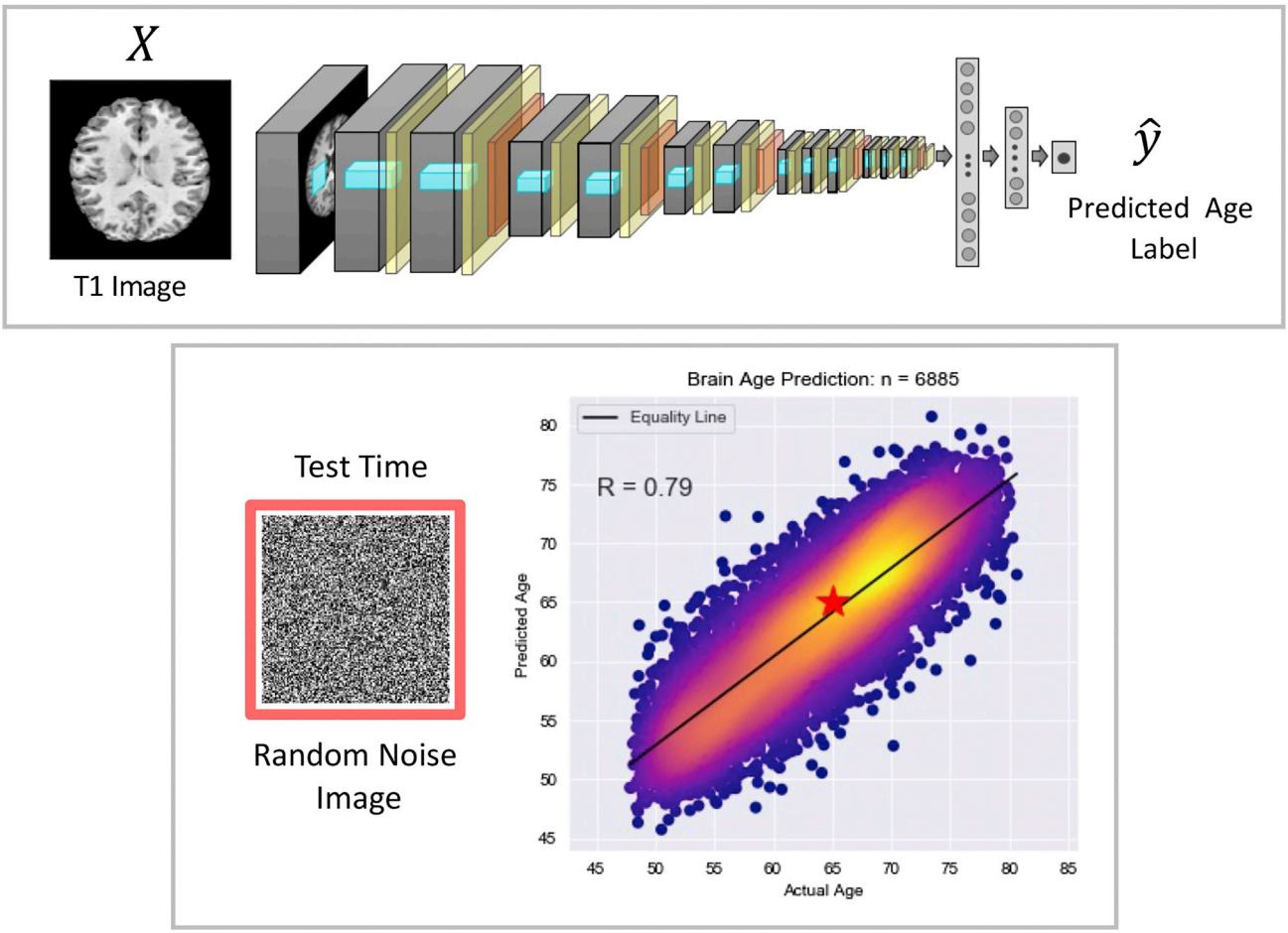

Figure 4. Model prediction for random noise

When a model trained to predict brain age (Dinsdale et al., 2021a) from T1 structural images was presented with an image of random noise, as it was unable to output an unknown class, the network predicted the random noise to have an age of 65—around the average age of the dataset's subjects. Although we can easily identify the random noise image by eye, there are many situations where the model, if presented with an image outside of the distribution it was trained on, would still output a (meaningless) result, which would be much harder for a user to identify.

these methods, however, provide coarse and low-resolution attribution maps and are computationally very expensive (Bass et al., 2020), especially when working with 3D medical images.

These post hoc methods do not require any model training in addition to the original network; however, it appears that they often fail to identify all the salient regions of a class, especially in medical imaging applications (Bass et al., 2020). Classifiers base their results on certain salient regions rather than the object as a whole, and a classifier may therefore ignore a region if the information there is redundant, i.e., if it can be provided by a different region of the image that is sufficient to minimize the loss function. Therefore, the regions of interest highlighted by these methods may not fully match a clinician's expectations (see Figure 5). Further, the prediction results might be virtually unchanged if the network were retrained with supposedly salient areas removed. Generally, although many methods have been developed to produce saliency or "heatmaps" from CNNs, limited effort has been focused on their evaluation with end-users (Alqaraawi et al., 2020). Fundamentally, these methods at best only highlight the important content of the image rather than uncovering the internal mechanisms of the model, and thus only indicate what is important, not *why*. Further, they are limited by the fact that CNNs are highly non-linear systems, so it is unlikely that, in general, there will be a mapping between regions of the input image and the task output that are understandable to humans.

Attention gates are components of the network that aim to focus a CNN on the target region of the image (the salient regions) by suppressing irrelevant feature responses in feature maps *during* training rather than post hoc (Park et al., 2018). This provides the user with attention maps, which again highlight the regions of the input image driving the network predictions. However, these methods, similarly to saliency or gradient-based methods, may not highlight all the expected regions in the image and can only indicate regions, not elucidate why. Attention gates have been applied to a range of imaging tasks, both for classification (Dinsdale et al., 2021a) and segmentation (Schlemper et al., 2019). Other methods have been developed to allow the visualization of

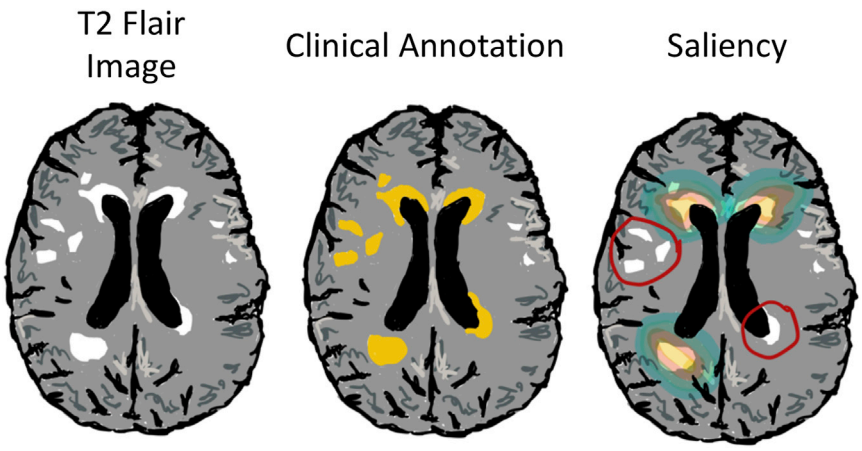

Figure 5. Schematic of the limitation of using saliency

When identifying the presence of white matter hyper-intensities, the neural network might only need to focus on a few of them to make the prediction. Thus, not all of the white matter hyper-intensities would be indicated in the saliency map, so the prediction would not match the clinician's expectation.

the differences between classes directly, rather than analyzing the model post hoc (Bass et al., 2020; Lee et al., 2020a, 2020b).

The methods discussed thus far enable visualization of the regions of the input image that drive the predictions, but do not provide insight into *how* the underlying filters of the network create decision boundaries, or why the regions were important, and are vulnerable to confirmation bias. In addition, in neuroimaging, patients with a given pathology are typically heterogeneous, and any changes they cause probably occur simultaneously. There are also significant amounts of healthy and normal variation in shape and appearance, so the interpretation of feature attribution maps to understand network predictions is difficult. Given the millions of parameters in many deep learning networks, despite our ability to visualize individual filters, and weights helping us to understand the hierarchical image composition, it is difficult to interrogate why decisions were made. Without some understanding of the model decision-making process, application across neuroimaging tasks in a clinical setting is likely to be limited due to the lack of trust that could be placed on the decisions. This is less of a concern in some settings, such as lesion segmentation, where the outputs can potentially be validated manually, but for tasks such as disease prediction, there may be greater concerns about model interpretability, which is yet to be solved by existing approaches.

Uncertainty

The use of uncertainties is an approach that aims to address the problem that, regardless of the input image, neural networks will always output a prediction, however inaccurate. Thus, by providing an estimate of the uncertainty associated with the prediction we can help the user to make an informed decision about whether or not to trust the model prediction. The softmax values output by a neural network are not true probabilities (Gal and Ghahramani, 2016), and networks often output high, incorrect softmax values, especially when presented with noisy or ambiguous data, or when the data presented to them differs from the distribution of the training data. As such, uncertainties are needed to allow proper quantification of the confidence of the prediction.

Uncertainties in deep learning can be split into two distinct groups (Kendall, 2017): aleatoric uncertainty, due to the ambiguity and noise in the data, and epistemic uncertainty, due to the uncertainty in the model parameters. The majority of methods in the literature focus on epistemic uncertainty, using Bayesian approaches to quantify the degree of uncertainty. The goal here is to estimate the posterior distribution of the model parameters. However, due to the very high dimensional parameter space, analytically computing the posterior directly is infeasible. Therefore, most methods use Monte Carlo dropout (Gal and Ghahramani, 2016), where dropout is applied to each of the convolutional layers and kept at test time; thus, we are able to sample from the distribution of possible model architectures. The uncertainty is then quantified through the variance of the predictive distribution, resulting from multiple iterations of the prediction stage with dropout present at test time, as demonstrated in Figure 6. This approach can readily be applied to existing CNNs; in medical imaging, it has primarily been used for segmentation tasks (Roy et al., 2018), where the segmentation is predicted alongside an uncertainty map. Other works have studied disease prediction, where the uncertainty is associated with the predicted class (Tousignant et al., 2019), and image registration (Bian et al., 2020). However, care must be taken with choice of the hyperparameters to ensure that the model assumptions are reasonable.

Some methods focus on the aleatoric uncertainty instead, estimated by having augmentation at test time (Ayhan and Berens, 2018; Wang et al., 2019a). Understanding of the uncertainty introduced by data varying from the training distribution is vital for the clinical translation of deep learning techniques. Given the degree of variation present in clinical data between sites and scanners, it is vital to understand how this contributes to predictions, both to mitigate against it and to develop user confidence in the predictions. Correlation between erroneous predictions and high uncertainties exists, so this could be used to improve the eventual predictions (Jungo et al., 2018).

However, further work in this area is still needed to ensure that the uncertainties produced would be meaningful at deployment, for instance, across dataset shifts. Calibration of uncertainties is also necessary so that they are comparable across methods (Thagaard et al., 2020). Furthermore, uncertainty values are only as good as the model and only meaningful alongside a well-validated model that is sufficiently powerful to discriminate the class of interest.

Interrogating the decision boundary

For many applications in neuroimaging, the output of a deep learning algorithm, if applied clinically, could potentially directly

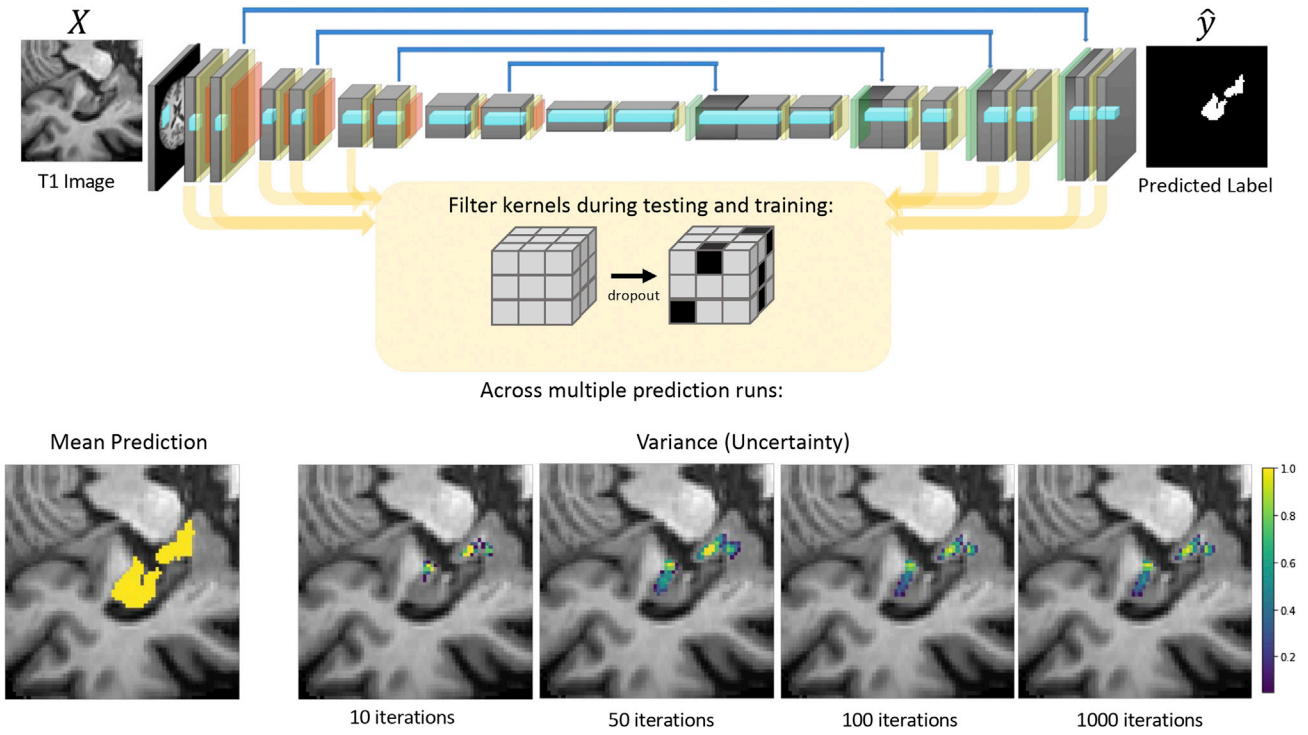

Figure 6. Model uncertainty

Most uncertainty methods have dropout applied at training and test time. Weights in the convolutional kernels are removed, which is approximated to represent the distribution of possible model architectures at test time. To demonstrate this, we trained a standard 3D UNet to complete hippocampal segmentation, with a dropout value of 0.5 applied on all convolutional layers. The HarP dataset (Frisoni and Jack, 2015) was used in this experiment, pre-processed as in Dinsdale et al. (2019). For each subject, we obtained a mean prediction and an uncertainty map, indicating the regions where the predictions between models were most varied and so approximated to be least certain.

influence patient care and outcomes. Thus, there is a clear need to be able to interrogate how decisions were made (Shah et al., 2019). Although visualization methods allow the inspection of which regions of the image influenced the prediction, and uncertainties grant us insight as to the confidence we should place in a prediction, for many applications we need to know precisely which characteristics led to a given prediction, and what would need to change for the outcome to be different and to help identify any bias driving predictions.

Our ability to interrogate the decision boundary is currently limited. Counterfactual analysis is one of the few existing approaches, which, given a supervised model where the desired prediction has not been achieved, shows what would have happened if the input were altered slightly (Verma et al., 2020). Simply, it identifies what altered characteristics would have led to a different model prediction. However, applications to neuroimaging (Pawlowski et al., 2020) are currently few, and exploration of its utility across neuroimaging tasks is required to ascertain its viability in a clinical setting.

EVALUATION

Availability of training labels

The evaluation of metrics requires labels: the "ground truth." We generally regard the ground truth as labels created by "domain experts"; these labels are key for training models but do not necessarily form part of standard clinical practice. Labels are required both for evaluation of the model performance and to train supervised methods. This exacerbates the problem of the shortage of data, as we need both large amounts of data and equal amounts of labels. These labels are expensive to obtain, requiring a large allocation of expert time to curate as well as expert domain knowledge, and are unlikely to be available for every clinical imaging site. Thus, we need methods that work when low numbers of labeled data points are available.

Few- and zero-shot learning methods work in very low-data regimes and are beginning to be applied to medical imaging problems (Feyjie et al., 2020). They are unlikely to generalize well to images from other sites and scanners, as the variation seen will not span the expected variation of the data, but they can help to begin to learn clusters of similar subjects where few labels are available. Unsupervised domain adaptation has been applied more widely, including for neuroimaging problems, to help cope with a lack of labels, with information from one dataset being leveraged to help us perform the same or a related task on another dataset (Sundaresan et al., 2021).

Other methods to overcome the lack of available labels focus on working with approximations for labels, which are cheaper to acquire (Tajbakhsh et al., 2020). Many methods propose pre-training the network using auxiliary labels generated using automatic tools, and then fine-tuning the model on the small number

of manual labels (Roy et al., 2018), or registration of an atlas to propagate labels from the atlas to the subject space (Hesse et al., 2022). Other approaches are weakly supervised, utilizing quick annotations, such as image-level labels (Feng et al., 2017), or bounding box annotations.

Other approaches to allow us to utilize deep learning when we have limited numbers of training labels include active learning and omnisupervised learning, both of which are trying to make the most effective use of the limited number of labels available. Active learning aims to minimize the quantity of labeled data required to train the network by prompting a human labeler to produce additional manual labels only where they might provide the greatest performance improvements, thereby minimizing the total number of annotations that need to be provided but giving a better performance than random annotation of the same number of samples (Yang et al., 2017). In omnisupervised learning (Radosavovic et al., 2018), automatically generated labels are created to improve predictions, starting from a small, labeled training set. By combining data diversity through applying data augmentation, and model diversity through the use of multiple different models, a consensus of labels is produced, which can be used to train the final model (Huang et al., 2018).

The difficulty in acquiring good quality manual labels is exacerbated by the variance caused when we pool data. The labels themselves provide an additional source of variance: when working in neuroimaging, the labels are frequently complicated and ambiguous (Shwartzman et al., 2020), often open to interpretation or with subjects having multiple labels that could be attributed due to co-morbidities (Graber, 2013). Despite this, we usually assume them to be 100% accurate (Cabitza et al., 2020)—the "gold" standard. If there is no objective answer, we cannot expect networks to provide one. Furthermore, this also leads to "inter-rater" variability, which generates a degree of uncertainty in the produced ground truth. The effect that this variability has on the predictions of the network needs to be understood and mitigated against. The uncertainty in the labels is also amplified by the lack of available data for rare conditions, which are therefore less represented in datasets, resulting in raters having less experience assessing them—particularly problematic if trying to quantify longitudinal changes with different raters (Visser et al., 2019).

Approaches need to consider three factors (Cabitza et al., 2020): agreement—the degree to which raters agree on a given label; confidence—how certain a rater is in their label; and competence—how accurate a given rater is. Research directions into the effects of rater variance have largely focused either on quantifying the reliability of the labels (Cabitza et al., 2020) or quantifying its effect on network performance (Shwartzman et al., 2020). Before any algorithm is deployed in practice, the limitations due to the labels must be understood and its consideration become a standard part of any deployment pipeline, remembering the "garbage in, garbage out" principle.

Choice of loss function

When training and evaluating model performance, we must choose a loss or cost function that we aim to minimize. Although some works design bespoke, task-specific cost functions, the majority are based on standard functions, such as categorical cross entropy for classification and segmentation tasks, Dice (an overlap metric) for segmentation and mean square error (MSE) for regression-based tasks.

These metrics are normally chosen because of their well-understood and characterized behavior (Maier-Hein et al., 2018). For the clinical translation of deep learning methods, we need to consider which measures are most important for the clinical application (Shah et al., 2019; Keane and Topol, 2018). Metrics only tell us part of the story: it is crucial to ensure that all vital information for clinical assessment is provided by the reported metrics. For instance, in many cases, false negatives are more problematic than false positives, resulting in a patient failing to receive the necessary care. Developing networks and loss functions with each specific application in mind is vital.

Furthermore, when training neural networks, we generally maximize the average performance. In practice, however, we are more likely to care about the performance on the hardest examples being acceptable than the performance on the easiest set of examples being improved slightly (Shu et al., 2020). Trading off a small amount of performance on easier examples in return for better performance on harder examples, which may give the same average performance overall, is probably preferable. Thus, the standard practice of minimizing the average performance may not be appropriate.

LOGISTICAL CHALLENGES

Computational resource

The final category of challenges is more logistical. Many of the most successful methods applied in imaging challenges involve large ensemble models such as the nnU-Net (Isensee et al., 2021), leading to many parameters and, therefore, calculations that must be stored and computed. Although successful in challenges, these methods are often not implementable on the hardware available in practice. Therefore, for clinical translatability, methods need to be developed which consider that computational limitations will be present on deployment and seek to create solutions that work within these constraints. Student-teacher networks (Hinton et al., 2015) and model distillation (Murugesan et al., 2020) aim to create smaller networks capable of mimicking the performance of the original large model (teacher), thus reducing the number of parameters in the final network which is deployed (student). Other approaches use separable convolutions, which drastically reduce the number of parameters in the network. Model pruning (LeCun et al., 1990; Dinsdale et al., 2022a) acknowledges that the parameters in neural networks are sparse and, therefore, by removing those that contribute least to the final prediction, we can reduce the size of the model architecture while maintaining performance.

Data sharing and data privacy

If we want CNNs that work for patients in real clinical applications, we need to be able to train our models on medical data that are relevant, realistic, and representative. Many current approaches focus on pooling anonymized data from across sites and patient groups through removing identifiable features such as name, birth date, and faces from the images. However, neural networks are still capable of extracting identifiable features from

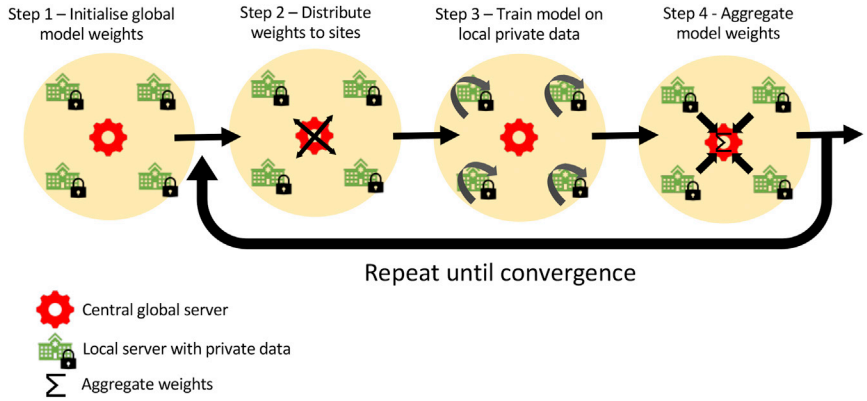

Figure 7. Illustration of a centralized federated learning framework

In the framework, the data for training the model is stored in local servers and not shared with the central server to ensure data security. Although the global model is available in the central server, the model parameters are shared with the local nodes 1,2... N, where training and parameter updates happen. The updated weights are then received at the central server, where the incoming updates are aggregated and applied to the global model. This learning and update happens in an iterative manner; both up and down transfer of model parameters are encrypted for data security.

these anonymized images such as age and sex (Pawlowski et al., 2020), which, in combination with other features such as hospital location and illness, could be identifying (Sweeney, 2002). The ability of neural networks to extract this information is only likely to increase. Furthermore, a proportion of identification risk comes from the presence of other auxiliary information—for instance, in neuroimaging, the scanner used to acquire the image. This is known as "linkage attack" and it is increasingly difficult to protect against fields using classic anonymization techniques (Sweeney, 2002).

Although de-identifying these data may just seem like an extra task for medical researchers, there are parties whose core business model is to de-anonymize medical data that have been sold for research purposes and sell that information to insurance companies (Tanner, 2017). De-anonymization research is a rapidly advancing field—for instance, reconstructing the faces of defaced medical images (Abramian and Eklund, 2019). Thus, to avoid future data privacy problems, approaches avoiding the aggregation of private medical information are valuable.

Fortunately, medical research is not the only field to face difficulties regarding the handling of sensitive, personal information. For instance, banking and mobile phone companies have faced this problem before. Therefore, we can take advantage of the privacy-preserving data analysis techniques that have rapidly developed in recent years. These techniques allow models to be trained without having direct access to the data and prevent these models from inadvertently storing sensitive information about the data. The most popular of these techniques are "federated learning," "differential privacy," and various forms of encrypted computation (Al-Rubaie and Chang, 2019; Kaissis et al., 2021). Here, we will focus on federated learning and differential privacy, as they currently show the most practical relevance in a neuroscience research setting (Rieke et al., 2020).

Federated learning (Figure 7) means training or testing your model on data that is stored on different devices or servers across the world, without having to centrally collect the data samples into one local aggregate dataset (Li et al., 2020). Instead of moving the data to the model, copies of the global model are sent to where the data are located; the data remain on the hospital server. The model is then trained on the local data, after which the newly improved model with its updated parameters is sent back to the main server to be aggregated with the main model. This preserves privacy in the sense that the data have not been moved from the device, and this method is therefore gaining popularity in various healthcare applications (Sheller et al., 2019). However, federated learning is limited by the fact that the content of the local data can sometimes be inferred from the weight updates or improvements in the models (Wang et al., 2019b) or due to the large numbers of parameters memorizing information about individuals.

Differential privacy helps overcome these drawbacks by injecting statistical noise to obscure the data contributions from individuals in the dataset (Dwork and Roth, 2014; Ziller et al., 2021). This is performed while ensuring that the model still gains insight into the overall population and thus provides predictions that are accurate enough to be useful. Ultimately, the use of differential privacy is a careful trade-off between privacy preservation and model utility (Dwork and Roth, 2014). A critical aspect of differential privacy is its inherent robustness to linkage attacks (Sweeney, 2002). As methods are developed, consideration of these approaches and future developments will be vital for ensuring privacy is maintained.

Conclusions

The combination of deep-learning-based methods and large-scale imaging datasets, such as UK Biobank, offers many opportunities to neuroimaging. Clearly, however, for the full impact of these methods to be experienced in the clinical domain there are challenges that must still be overcome.

Our key recommendations for future directions are discussed in Box 1. Ultimately, for models to be able to be deployed successfully, the clinical needs and limitations must be considered central to model design, so that the models produced are robust, reliable, and able to improve patient outcomes. In this article, we have discussed issues relating to data availability, interpretability, model evaluation and data privacy. Deep-learning-based methods are beginning to receive FDA approval for applications in medical imaging, but it is yet to be seen what impact or uptake these methods will have. The challenges for neuroimaging are, however, likely to differ in focus to those of the computer vision field. In particular, interpretability—the ability to interrogate decision making and trust

Box 1. Recommendations for future directions

Throughout this review, we have discussed the key barriers for the success of deep learning in neuroimaging, and the current approaches and directions being explored to overcome them. Although various methods are being explored, the barriers remain significant and thus, we here briefly discuss our recommendations for future research directions.

- **Data availability**: current challenges are interlinked to the data available: inevitably, the data we have collected can only ever be a snapshot of the populations we wish to study. We need to better understand the limitations of the data we have available, for instance, through risk analysis (Zendle et al., 2015) to identify the underrepresented demographics in the data. Where underrepresented groups or other forms of training-sample-bias are identified, this should enable exploration of mitigation approaches, such as oversampling underrepresented groups, creating simulated subjects using generative methods or targeted data collection.
- **Data pooling and harmonization**: the relative cost and difficulty of acquiring imaging data (compared with simpler, smaller forms of subject-level data such as simple demographics) makes it hard to build up large-N imaging datasets for training deep learning models; furthermore, the cost, size and complexity of imaging data further exacerbates this. It will often be necessary therefore to pool datasets, creating privacy and harmonization issues. More work on robust multi-modal image processing, applied before deep learning training, will be needed to reduce problems of data harmonization (for example, reducing variations due to imaging hardware and acquisition protocol). However, as this is unlikely to be perfect, deep learning models will still likely need to include a harmonization component, an important area of future research.
- **Interpretability and trust**: we need to develop better methods to explain *why* a prediction has been made—in addition to *what* drives the predictions. Until it is possible to train truly interpretable deep learning networks, in safety-critical applications, such as methods to predict diagnosis or suggest treatment options, methods should be used which are inherently interpretable, such that patients and clinicians can interrogate outcomes, or we risk long-term damage to the trust in deep learning models (Rudin, 2019).
- **Evaluation**: after training, the model evaluation must be thorough, to ensure that the model works as expected and is robust to the expected variation. Better robust evaluation procedures should be developed which maximize the impact of the available labels, for instance through test time augmentation to simulate variation (Hendryck and Dietterich, 2019), multiple evaluation metrics to capture different aspects of the prediction, or stratifying results to understand performance across different demographic groups such as age or sex.
- **Logistical challenges**: the use of representative clinical data will be vital for the success of deep learning models; thus, privacy-preserving approaches are important for new methodological development. Therefore, future developments should be built around federated frameworks (i.e., non-centralized data stores), despite the increased constraint on architectures and training procedures (Dinsdale et al., 2022b). New methods will need to minimize the amount of information which needs to be shared (to the centralized training process), and properly understand the dangers of de-anonymization (e.g., understanding differential privacy).

the decision-making process—is likely to be a significant barrier for translatability and will likely require specific efforts beyond those in the general computer vision field.

The code for the examples in this paper can be found at: github.com/nkdinsdale/challenges_review.

ACKNOWLEDGMENTS

This work was supported in part by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) (grant number EP/L016052/1) (N.K.D. and E.B.), by the Clarendon Scholarship fund (E.B.), and by a Wellcome Trust Collaborative Award (215573/Z/19/Z) (S.M.S.). The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z) (V.S.). A.I.L.N. is grateful for support from the UK Royal Academy of Engineering under the Engineering for Development Research Fellowships scheme. N.K.D. and A.I.L.N. are also supported by an Academy of Medical Sciences Springboard Grant (SBF005/1136). M.J. is supported by the National Institute for Health Research (NIHR) and the Oxford Biomedical Research Centre (BRC). This research has been conducted in part using the UK Biobank Resource under Application Number 8107. We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible. Analysis was carried out on the clusters at the Oxford Biomedical Research Computing (BMRC) facility. BMRC is a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute, supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

DECLARATION OF INTERESTS

M.J. receives royalties from licensing of FSL to non-academic, commercial parties.

REFERENCES

Abramian, D., and Eklund, A. (2019). Refacing: reconstructing anonymized facial features using GANS. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 1104–1108.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Bianchi-Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. IUI '20: 25th International Conference on Intelligent User Interfaces, 275–285.

Al-Rubaie, M., and Chang, J.M. (2019). Privacy-preserving machine learning: threats and solutions. IEEE Secur. Privacy *17*, 49–58.

Ayhan, M., and Berens, P. (2018). Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. Proceedings of the Medical Imaging with Deep Learning.

Bass, C., da Silva, M., Sudre, C.H., Tudosiu, P.D., Smith, S.M., and Robinson, E.C. (2020). ICAM: interpretable classification via disentangled representations and feature attribution mapping. 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S., Ma, K., Yuan, J., and Zheng, Y. (2020). Uncertainty-aware domain alignment for anatomical structure segmentation. Med. Image Anal. *64*, 101732.

Billot, B., Bocchetta, M., Todd, E., Dalca, A.V., Rohrer, J.D., and Iglesias, J.E. (2020a). Automated segmentation of the hypothalamus and associated subunits in brain MRI. NeuroImage *223*, 117287.

Billot, B., Robinson, E., Dalca, A.V., and Iglesias, J.E. (2020b). Partial volume segmentation of brain {MRI} scans of any resolution and contrast. Medical Image Computing and Computer Assisted Intervention—MICCAI 2020.

Billot, B., Greve, D., Van Leemput, K., Fischl, B., Iglesias, J., and Dalca, A. (2020c). A learning strategy for contrast-agnostic MRI segmentation. Proceedings of the Machine Imaging with Deep Learning (MIDL).

Binder, A., Bach, S., Montavon, G., Müller, K.-R., and Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. Lecture Notes in Electrical Engineering, 913–922.

Böhle, M., Eitel, F., Weygandt, M., and Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front. Aging Neurosci. *11*, 194.

Cabitza, F., Campagner, A., Albano, D., Aliprandi, A., Bruno, A., Chianca, V., Corazza, A., Di Pietto, F., Gambino, A., Gitto, S., et al. (2020). The elephant in the machine: proposing a new metric of data reliability and its application to a medical case to assess classification reliability. Appl. Sci. *10*, 4014.

Cetin Karayumak, S., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., and Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. NeuroImage *184*, 180–200.

Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv:2006.10511v2.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. Med. Image Anal. *58*, 101539.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv:2002.05709v3.

Cicek, Ozgun, Abdulkadir, A., Lienkamp, S.S., Brox, T., and Ronneberger, O. (2016). 3D U-net: learning dense volumetric segmentation from sparse annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 424–432.

Clement, C., Selman, L.E., Kehoe, P.G., Howden, B., Lane, J.A., and Horwood, J. (2019). Challenges to and facilitators of recruitment to an Alzheimer's disease clinical trial: a qualitative interview study. J. Alzheimers Dis. *69*, 1067–1075.

Deng, J., Dong, W., Socher, R., Li-Jia, L., Li, K., and Li, F.-F. (2009). ImageNet: a large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, 248–255.

Dewey, B.E., Zhao, C., Reinhold, J.C., Carass, A., Fitzgerald, K.C., Sotirchos, E.S., Saidha, S., Oh, J., Pham, D.L., Calabresi, P.A., et al. (2019). DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. Magn. Reson. Imaging *64*, 160–170.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al. (2014). The autism brain imaging data exchange: Towards large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry *19*, 659–667.

Dinsdale, N.K., Jenkinson, M., and Namburete, A.I.L. (2019). Spatial warping network for 3D segmentation of the hippocampus in MR images. 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2019, 284–291.

Dinsdale, N.K., Bluemke, E., Smith, S.M., Arya, Z., Vidaurre, D., Jenkinson, M., and Namburete, A.I.L. (2021a). Learning patterns of the ageing brain in MRI using deep convolutional networks. Neuroimage *224*, 117401.

Dinsdale, N.K., Jenkinson, D., and Namburete, A.I.L. (2021b). Deep learning-based unlearning of dataset bias for {MRI} harmonisation and confound removal. NeuroImage, 117689.

Dinsdale, N.K., Jenkinson, M., and Namburete, A.I.L. (2022). STAMP: simultaneous training and model pruning for low data regimes in medical image segmentation. Med. Image Anal. *81*, 102583.

Dinsdale, N.K., Jenkinson, M., and Namburete, A.I.L. (2022). FedHarmony: unlearning scanner bias with distributed data. arXiv:2205.15970v1.

Dwork, C., and Roth, A. (2014). The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. *9*, 211–407.

Feng, X., Yang, J., Laine, A.F., and Angelini, E.D. (2017). Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2017.

Feyjie, A.R., Azad, R., Pedersoli, M., Kauffman, C., Ben Ayed, I., and Dolz, J. (2020). Semi-supervised few-shot learning for medical image segmentation. https://doi.org/10.48550/arXiv.2003.08462.

Frazier, J.A., Hodge, S.M., Breeze, J.L., Giuliano, A.J., Terry, J.E., Moore, C.M., Kennedy, D.N., Lopez-Larson, M.P., Caviness, V.S., Seidman, L.J., et al. (2008). Diagnostic and sex effects on limbic volumes in early-onset bipolar disorder and schizophrenia. Schizophr. Bull. *34*, 37–46.

Frisoni, G.B., and Jack, C.R. (2015). HarP: the EADC-ADNI Harmonized Protocol for manual hippocampal segmentation. A standard of reference from a global working group. Alzheimers Dement. *11*, 107–110.

Gal, Y., and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 1050–1059.

Graber, M.L. (2013). The incidence of diagnostic error in medicine. BMJ Qual. Saf. *22*, ii21–ii27.

Guan, H., Yang, E., Yap, P.-T., Shen, D., and Liu, M. (2020). Attention-guided deep domain adaptation for brain dementia identification with multi-site neuroimaging data. Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, 31–40.

Hadsell, R., Chopra, S., and Lecun, Y. (2006). Dimensionality reduction by learning an invariant mapping. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 1735–1742.

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., et al. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. NeuroImage *32*, 180–194.

He, K., Girshick, R., and Dollar, P. (2019). Rethinking ImageNet pre-training. 2019/CVF International Conference on Computer Vision (ICCV), 4917–4926.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., and Yeo, B.T.T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage *206*, 116276.

Hendryck, D., and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and pertubations. ICLR.

Hesse, L.S., Aliasi, M., Moser, F., INTERGROWTH-21(st) Consortium, Haak, M.C., Xie, W., Jenkinson, M., and Namburete, A.I.L. (2022). Subcortical segmentation of the fetal brain in 3D ultrasound using deep learning. NeuroImage *254*, 119117.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. NIPS.

Hostage, C.A., Roy Choudhury, K., Doraiswamy, P.M., and Petrella, J.R. (2013). Dissecting the gene dose-effects of the APOE eta4 and eta2 alleles

on hippocampal volumes in aging and Alzheimer's disease. PLoS One *8*, e54483.

Huang, R., Noble, J.A., and Namburete, A.I.L. (2018). Omni-supervised learning: scaling up to large unlabelled medical datasets. International Conference on Medical Image Computing and Computer-Assisted Intervention, 572–580.

Hughes, E.J., Winchman, T., Padormo, F., Teixeira, R.P., Wurie, J., Sharma, M., Fox, M., Hutter, J., Cordero-Grande, L., Price, A.N., et al. (2017). A dedicated neonatal brain imaging system. Magn. Reson. Med. *78*, 794–804.

Iglesias, J., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., Alexander, D., Golland, P., Edlow, B., and Fischl, B. (2020). Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast. Neuroimage *237*, 118206.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods *18*, 203–211.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al. (2008). The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging *27*, 685–691.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., Macfall, J., et al. (2006). Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage *30*, 436–443.

Jungo, A., McKinley, R., Meier, R., Knecht, U., Vera, L., Pérez-Beteta, J., Molina-García, D., Pérez-García, V.M., Wiest, R., and Reyes, M. (2018). Towards uncertainty-assisted brain tumor segmentation and survival prediction. BrainLes 2017: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 474–485.

Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, I., Mancuso, J., Jungmann, F., Steinborn, M., et al. (2021). End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat. Mach. Intell. *3*, 473–484.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. International Conference on Information Processing in Medical Imaging, 597–609.

Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. *36*, 61–78.

Keane, P.A., and Topol, E.J. (2018). With an eye to AI and autonomous diagnosis. npj Digit. Med. *1*, 40.

Kendall, A., and Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Proceedings of the 31st international conference on neural information processing systems, 5580–5590.

Khagi, B., Lee, C.G., and Kwon, G.-R. (2018). Alzheimer's disease classification from brain MRI based on transfer learning from CNN. International Conference on Smart Electronics and Communication.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks systems. Advances in Neural Information Processing Systems 25 (NIPS 2012).

Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., and Lladó, X. (2019). Supervised domain adaptation for automatic subcortical brain structure segmentation with minimal user interaction. Sci. Rep. *9*, 6742.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324.

LeCun, Y., Denker, J.S., and Solla, S.A. (1990). Optimal brain damage. Adv. Neural Inf. Process. Syst. *2*, 598–605.

Lee, B., Yamanakkanavar, N., and Choi, J.Y. (2020a). Automatic segmentation of brain {MRI} using a novel patch-wise U-Net deep architecture. PLoS One *15*, 1–20.

Lee, H., Tseng, H.-Y.T., Mao, Q., Huang, J.B., Lu, Y.-D., Singh, M., and Yang, M.-H. (2020b). DRIT++: diverse image-to-image translation via disentangled representations. Int. J. Comput. Vision *128*, 2402–2417.

Li, T., Sahu, A.K., Talwalkar, A., and Smith, V. (2020). Federated learning: challenges, methods, and future directions. IEEE Signal Process. Mag. *37*, 50–60.

Livne, M., Rieger, J., Aydin, O.U., Taha, A.A., Akay, E.M., Kossen, T., Sobesky, J., Kelleher, J.D., Hildebrand, K., Frey, D., et al. (2019). A U-net deep learning framework for high performance vessel segmentation in patients With cerebrovascular disease. Front. Neurosci. *13*, 97.

Lu, B., Li, H.-X., Chang, Z.K., Li, L., Chen, N.-X., Zhu, Z.-C., Zhou, H.-X., Zhou, H.-X., Li, X.-Y., Wang, Y.-W., Cui, S.-X., et al. (2021). A practical Alzheimer disease classifier via brain imaging-based deep learning on 85,721 samples. https://doi.org/10.1101/2020.08.18.256594.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. *9*, 5217.

Marek, S., Tervo-Clemmens, B., Nielsen, A.N., Wheelock, M.D., Miller, R.L., Laumann, T.O., Earl, E., Foran, W.W., Cordova, M., Doyle, O., et al. (2019). Identifying reproducible individual differences in childhood functional brain networks: an {ABCD} study. Dev. Cogn. Neurosci. *40*, 100706.

Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahaniy, K., Kirby, J., Burren, Y., et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging *99*, 1993–2024.

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. https://doi.org/10.48550/arXiv.1411.1784.

Morid, M.A., Borjali, A., and Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. *128*, 104115.

Moyer, D., Ver Steeg, G., Tax, C.M.W., and Thompson, P.M. (2020). Scanner invariant representations for diffusion MRI harmonization. Magn. Reson. Med. *84*, 2174–2189.

Murugesan, B., Vijayarangan, S., Sarveswaran, K., Ram, K., and Sivaprakasam, M. (2020). KD-MRI: a knowledge distillation framework for image reconstruction and image restoration in MRI workflow. Proceedings of the Third Conference on Medical Imaging with Deep Learning, 515–526.

Nguyen, K.P., Fatt, C.C., Treacher, A., Mellema, C., Trivedi, M.H., and Montillo, A. (2020). Anatomically informed data augmentation for functional {MRI} with applications to deep learning. Proc. SPIE Int. Soc. Opt. Eng. *11313*, 113130T.

Ning, L., Bonet-Carne, E., Grussu, F., Sepehrband, F., Enrico, K., Veraart, J., Blumberg, S.B., et al. (2020). Cross-scanner and cross-protocol multi-shell diffusion {MRI} data harmonization: algorithms and results. NeuroImage *221*, 117128.

Nobis, L., Manohar, S.G., Smith, S.M., Alfaro-Almagro, F., Jenkinson, M., Mackay, C.E., and Husain, M. (2019). Hippocampal volume across age: nomograms derived from over 19,700 people in UK Biobank. NeuroImage: Clin. *23*, 101904.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The building blocks of interpretability. Distill *3*, e10.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Berkay Celik, Z., and Swami, A. (2017). Practical black-box attacks against machine learning. ASIA CCS 2017—Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (Association for Computing Machinery), pp. 506–519.

Park, J.C., Woo, S., Lee, J.-Y., and Kweon, I. (2018). BAM: bottleneck attention module. Proceedings of the British Machine Vision Conference (BMVC).

Paulus, J.K., and Kent, D.M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. npj Digit. Med. *3*, 99.

Pawlowski, N., Coelho de Castro, D., and Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference. Advances in Neural Information Processing Systems, 857–869.

Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., and Smith, S.M. (2021). Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. *68*, 101871.

Radosavovic, I., Dollar, P., Girshick, R., Gkioxari, G., and He, K. (2018). Data distillation: towards omni-supervised learning. 2018/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4119–4128.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: understanding transfer learning with applications to medical imaging. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).

Rieke, N., Hancox, J., Milletari, F., Roth, H., Albarqouni, S., Bakas, S., Galtier, M., et al. (2020). The future of digital health with federated learning. npj Digit. Med. *3*, 119.

Roy, A.G., Conjeti, S., Navab, N.N., and Wachinger, C. (2018). Inherent brain segmentation quality control from fully ConvNet Monte Carlo sampling. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, 664–672.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. *1*, 206–215.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. (2019). Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. *53*, 197–207.

Shah, N.H., Milstein, A., and Bagley, S.C. (2019). Making machine learning models clinically useful. JAMA *322*, 1351–1352.

Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., and Bakas, S. (2019). Multi-institutional deep learning modeling Without sharing patient data: A feasibility study on brain tumor segmentation. Brainlesion *11383*, 92–104.

Shu, M., Liu, C., Qiu, W., and Yuille, A. (2020). Identifying model weakness with adversarial examiner. Proceedings of the AAAI conference on artificial intelligence *34*, 11998–12006.

Shwartzman, O., Gazit, H., Shelef, I., and Riklin-Raviv, T. (2020). The worrisome impact of an inter-rater bias on neural network training. https://doi.org/10.48550/arXiv.1906.11872.

Simard, P., Lecun, Y., and Denker, J. (1998). Transformation invariance in pattern recognition tangent distance and tangent propagation. In Neural Networks: Tricks of the Trade (Springer).

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep Inside convolutional networks: visualising image classification models and saliency maps. Workshop at International Conference on Learning Representations.

Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: an open access resource for identifying the causes of a Wide Range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

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

Sweeney, L. (2002). K-anonymity: a model for protecting privacy. Int. J. Unc. Fuzz. Knowl. Based Syst. *10*, 557–570.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., and Ding, X. (2020). Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. *63*, 101693.

Takao, H., Hayashi, N., and Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. J. Magn. Reson. Imaging. *34*, 438–444.

Tanner, A. (2017). Our bodies, Our Data: How Companies Make Billions Selling Our Medical Records (Beacon Press).

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-Can, C.A.N., and Henson, R.N. (2017). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage *144*, 262–269.

Thagaard, J., Hauberg, S., van der Vegt, B., Ebstrup, T., Hansen, J.D., and Dahl, A.B. (2020). Can you trust predictive uncertainty Under real dataset shifts in digital pathology? International Conference on Medical Image Computing and Computer-Assisted Intervention, 824–833.

Thakur, S., Doshi, J., Pati, S., Rathore, S., Sako, C., Bilello, M., Ha, S.M., Shukla, G., Flanders, A., Kotrotsou, A., et al. (2020). Brain extraction on MRI scans in presence of diffuse glioma: multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. NeuroImage *220*, 117081.

Tousignant, A., Lemaitre, P., Precup, D., Arnold, D.L., and Arbel, T. (2019). Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, 483–492.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., and Ugurbil, K.; WU-Minn HCP Consortium (2013). The WU_Minn Human connectome Project: an overview. Neuroimage *80*, 62–79.

Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: a review. https://doi.org/10.48550/arXiv.2010.10596.

Visser, M., Müller, D.M.J., van Duijn, R.J.M., Smits, M., Verburg, N., Hendriks, E.J., Nabuurs, R.J.A., Bot, J.C.J., Eijgelaar, R.S., Witte, M., et al. (2019). Inter-rater agreement in glioma segmentations on longitudinal MRI. Neuroimage Clin. *22*, 101727.

Wachinger, C., Reuter, M.R., and Klein, T. (2018). DeepNAT: deep convolutional neural network for segmenting neuroanatomy. NeuroImage *170*, 434–445.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019a). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing *335*, 34–45.

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., and Qi, H. (2019b). Beyond inferring class representatives: user-level privacy leakage From federated learning. IEEE Conference on Computer Communications, 2512–2520.

Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al. (2019). Association Between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. *155*, 1135–1141.

Wu, W., Lu, Y., Mane, R., and Guan, C. (2020). Deep learning for neuroimaging segmentation with a novel data augmentation strategy. 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 1516–1519.

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D.Z. (2017). Suggestive annotation: A deep active learning framework for biomedical image segmentation. Lecture Notes in Computer Science, 399–407.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems.

Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., et al. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nat. Commun. *9*, 4273.

Yu, M., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., and Sheline, Y.I. (2018). Statistical

harmonization corrects site effects in functional connectivity measurements from multisite fMRI data. Hum. Brain Mapp. *39*, 4213–4227.

Zeiler, M., and Fergus, R. (2014). Visualizing and understanding convolutional neural networks. IEEE European Conference on Computer Vision (ECCV).

Zendle, O., Murschitz, M., Humenberger, M., and Herzner, W. (2015). CV-HAZOP: introducing test data validation for computer vision. ICCV.

Zhang, H., Cisse, M., Dauphin, Y., and Lopez-Paz, D. (2017). mixup: beyond empirical risk minimization. https://doi.org/10.48550/arXiv.1710.09412.

Zhang, Y., Jiang, H., Miura, Y., Manning, C., and Langlotz, C. (2020). Contrastive learning of medical visual representations from paired images and text. https://doi.org/10.48550/arXiv.2010.00747.

Zhou, Z., Sodha, V., Mahfuzur Rahman Siddiquee, M.d., Feng, R., Tajbakhsh, N., Gotway, M.B., and Liang, J. (2019). Models Genesis: generic autodidactic models for {3D} medical image analysis. Medical Image Computing and Computer Assisted Intervention—MICCAI, 384–393.

Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D., and Kaissis, G. (2021). Medical imaging deep learning with differential privacy. Sci. Rep. *11*, 13524.