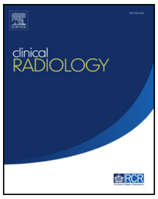
Clinical Radiology 74 (2019) 346–356

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net



Review

Machine learning in whole-body MRI: experiences and challenges from an applied study using multicentre data

I. Lavdas [a,*], B. Glocker [b], D. Rueckert [b], S.A. Taylor [c], E.O. Aboagye [a], A.G. Rockall [a,d]

[a] *Imperial College Comprehensive Cancer Imaging Centre (C.C.I.C.), Hammersmith Campus, Commonwealth Building Main Office, Ground Floor, Du Cane Road, London W12 0NN, UK*
[b] *Biomedical Image Analysis Group, Department of Computing, Huxley Building, 180 Queen's Gate, Imperial College London, London SW7 2AZ, UK*
[c] *Department of Imaging, University College London Hospitals NHS Foundation Trust, Euston Road, London NW1 2BU, UK*
[d] *Department of Radiology Hammersmith Hospital, Imperial College Healthcare NHS Trust, Du Cane Road, London W12 0NN, UK*

ARTICLE INFORMATION

Article history:
Received 5 October 2018
Accepted 22 January 2019

Machine learning is now being increasingly employed in radiology to assist with tasks such as automatic lesion detection, segmentation, and characterisation. We are currently involved in an National Institute of Health Research (NIHR)-funded project, which aims to develop machine learning methods to improve the diagnostic performance and reduce the radiology reading time of whole-body magnetic resonance imaging (MRI) scans, in patients being staged for cancer (MALIBO study). We describe here the main challenges we have encountered during the course of this project. Data quality and uniformity are the two most important data traits to be considered in clinical trials incorporating machine learning. Robust data pre-processing and machine learning pipelines have been employed in MALIBO, a task facilitated by the now freely available machine learning libraries and toolboxes. Another important consideration for achieving the desired clinical outcome in MALIBO, was to effectively host the resulting machine learning output, along with the clinical images, for reading in a clinical environment. Finally, a range of legal, ethical, and clinical acceptance issues should be considered when attempting to incorporate computer-assisting tools into clinical practice. The road from translating computational methods into potentially useful clinical tools involves an analytical, stepwise adaptation approach, as well as engagement of a multidisciplinary team.

Introduction

Machine learning applications are ever-present in our daily activities, whether the beneficiary is aware of it or

[*] Guarantor and correspondent: I. Lavdas, Comprehensive Cancer Imaging Centre (C.C.I.C.), Hammersmith Campus, Commonwealth Building Main Office, Ground Floor, Du Cane Road, London W12 0NN, UK. Tel.: +44 020 83838598; fax: +44 020 8383 1783.
E-mail address: ilavdas@imperial.ac.uk (I. Lavdas).

https://doi.org/10.1016/j.crad.2019.01.012
0009-9260/© 2019 The Royal College of Radiologists. Published by Elsevier Ltd. All rights reserved.

not. Medical imaging and, more specifically, clinical radiology could not have remained unaffected by these advances.[1–3] The development and application of machine learning methods in radiology has the potential to support a series of clinical tasks, such as automatic lesion detection and segmentation, lesion classification, patient risk stratification or patient outcome prediction, and may apply to radiological images of different modalities. Recently, driven by the rapid progress in computational power and speed and the availability of big datasets, the use of deep learning and, more specifically, convolutional neural networks has revolutionised the field of automated analysis of radiological images by accomplishing some of the aforementioned tasks with remarkable accuracy.[4–6]

The developed machine learning methodologies seek to improve the diagnostic and predictive performance of radiological scans and generate an, "up to the hilt", time-efficient and error-proof workflow for the reporting radiologist. The role of computational tools is intended to be complementary and supportive to the radiologist, potentially performing time-consuming tasks such as quantitative measurements; the experienced radiologists' judgement remains the reference standard, taking many other factors and non-imaging information into account; however, to quote Curtis Langlotz of Stanford from the Radiological Society of North America (RSNA) meeting in 2017: "*radiologists who use artificial intelligence, will replace those who don't*".

Recent technological advances in magnetic resonance imaging (MRI), have allowed whole-body MRI (WB-MRI) to be performed clinically with acceptable image quality and within reasonable time. The addition of diffusion-weighted imaging (DWI) in WB protocols, means that WB-DWI is now becoming an increasingly important tool in oncology for cancer diagnosis, staging, and treatment-response monitoring.[7–9] A significant challenge when reading WB-MRI images is the increased volume of resulting imaging data, especially when multiparametric acquisitions are used. The reading process can then become rather time-consuming, with increased risk of misinterpretations. In addition, WB-DWI for staging cancer patients has limitations with respect to its diagnostic performance,[10] as it may be prone to false-positives resulting from tissues with normally occurring restricted diffusivity.[11]

The National Institute of Health Research (NIHR) has funded a project (EME project 13/122/01), which aims to develop state-of-the art machine learning algorithms for the automatic detection of malignant and benign lesions in multicentre, multiparametric WB-MRI.[12] The study hypothesis is that the developed machine learning tools will have the potential to improve the diagnostic performance and reduce the reading time of WB-MRI. We discuss here our experiences from this study and demonstrate the methodology employed and challenges met in the pathway towards translating our methods into a potentially useful clinical tool.

The MALIBO (machine learning in body oncology) study

MALIBO is a prospective, observational study, which aims to develop machine learning methods and validate them by comparing the diagnostic performance and reading time of WB-DWI, when assessed alone and when assessed in conjunction with machine learning output. The study does not collect patient imaging data, but relies on data collected by other NIHR and CRUK-funded trials, referred to as "contributing studies".[13,14] MALIBO is funded by the NIHR, Efficacy and Mechanism Evaluation programme (EME project: 13/122/01) and is a collaboration between the Imperial College Comprehensive Cancer Imaging Centre (C.C.I.C.) and the Department of Computing at Imperial College. Data from the contributing studies are provided by the University College London (UCL) and University College London Hospitals NHS Foundation Trust (UCLH).

The study is divided into three phases, whereby in Phase 1 algorithms are developed and evaluated for their accuracy to identify normal structures in WB-MRI scans from healthy volunteers. In Phase 2, the developed algorithms will be further trained to identify benign lesions and then tested and further refined for detecting cancer lesions. Finally, in Phase 3 the algorithms will be tested in a large cohort of "unseen" WB-MRI data. As far as we are aware, MALIBO is the first study that applies machine learning techniques in WB-(DW)-MRI.

The MALIBO study relies on WB-MRI data from a range of multi-centre trials, and includes a range of cancer types, and thus the setting of the study is truly pragmatic in clinical terms. As a result, the imaging data are relatively heterogeneous, or "messy", which poses significant challenges to applying any statistical image analysis approach. Current machine learning methodology requires the data to be fairly homogeneous, in the sense that the training data from which task-specific features are learned should be similar to the unseen test data, on which one wishes to make predictions for. Fig 1 shows a block diagram identifying the MALIBO phases, during which the most significant challenges have been encountered to date and for which our methodology required adaptation.

Data acquisition

The use of big datasets is a desirable feature for either clinical outcome-driven imaging studies or purely machine learning outcome-driven imaging studies. A large cohort of examined patients can potentially increase the statistical power of primary and secondary outcomes in clinical trials and can also boost the accuracy of the employed algorithms in machine learning-related imaging studies, where larger datasets are more likely to sufficiently capture the natural variability of both anatomy and pathology. Thus, investigators turn to the use of retrospectively acquired imaging data or look into multicentre collaborations to

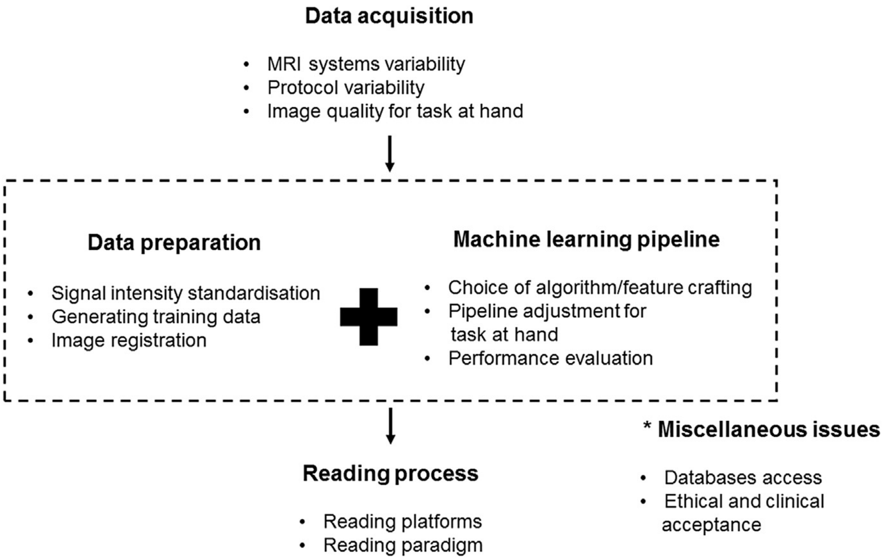

Data acquisition
• MRI systems variability
• Protocol variability
• Image quality for task at hand
Data preparation
• Signal intensity standardisation
• Generating training data
• Image registration
Machine learning pipeline
• Choice of algorithm/feature crafting
• Pipeline adjustment for task at hand
• Performance evaluation
Reading process
• Reading platforms
• Reading paradigm
* Miscellaneous issues
• Databases access
• Ethical and clinical acceptance

Figure 1 Block diagram depicting the methodological components that were considered in MALIBO study.

maximise the amount of available data for their studies; however, this means that there will be data compliance issues. In studies using, for example, CT datasets, the data are likely to be fairly homogeneous, although differences in slice thickness or differences in the use of contrast medium may pose challenges; however, in the MRI setting, as encountered in MALIBO, there may be extra significant variabilities in the data, including differences in imaging sequences, between manufacturers and differences in acquisition parameters posing additional challenges to the training and deployment of machine learning tools, as will be described below.

MRI systems and acquisition protocol variabilities

The MRI systems used in multicentre studies, will very commonly be of different manufacturers and different field strengths, have different coil characteristics, and will be quality checked to different standards, even in the context of well-designed clinical imaging studies. This implies that images of inconsistent appearance and quality will be acquired throughout different centres. These differences are of little consequence to interpretation by the flexible human reader, who is trained to readily adapt to visual differences, but pose significant challenges for current machine learning algorithms. Furthermore, the introduction of functional imaging, which can now be incorporated into WB protocols as in MALIBO, means that the spatial and signal intensity discrepancies between images acquired in different centres can be of particular importance in machine learning-related imaging studies.

This protocol variability in terms of anatomical localisation and signal intensity effects is demonstrated, using MALIBO data, in Fig 2. Methods with which a number of the variability issues mentioned above, were mitigated in MALIBO, are described in the "Data preparation" section.

Image quality

The versatility of MRI is the modality's "blessing and curse". It is very common that image acquisition in the body may be compromised by patient factors such as movement, bowel gas, joint prosthesis, or surgical material, and imaging datasets of compromised quality can be "passed through the sieve" of the clinical workflow, often out of necessity.

Repeating sequences may not always be practicable, because of time constrains or patient exhaustion (especially if incorporating multiple sequences including DW-MRI). It should be stressed, however, that the quality of the acquired datasets might have been suitable for the objectives of the clinical study, involving human readers, and not all of the issues are externally triggered (for example, distortions in echo planar imaging [EPI] DWI acquisitions are unavoidable[15]), but they may cause very significant challenges to the machine learning algorithms and be detrimental to their performance.

This, highlights the importance of having imaging data with readiness level of *"Band A*", appropriate for the task at hand, as described by Lawrence 2017,[16] for machine learning studies. It is acknowledged, however, that when multicentre data are collected, the scenario above is unrealistic, so removal of inappropriate or compromised datasets might be unavoidable for the purposes of algorithm training and also at test time, when predictions are made on new, "unseen'" data. We have estimated that a proportion of the datasets employed in MALIBO, were not suited for machine learning purposes, and had to be discarded. Fig 3 shows some of the image-quality issues we encountered in MALIBO.

It is, therefore, highly recommended that MRI acquisitions for machine learning studies are standardised to the highest possible degree and are performed and monitored by an experienced research radiographer or by the local MRI physicist.

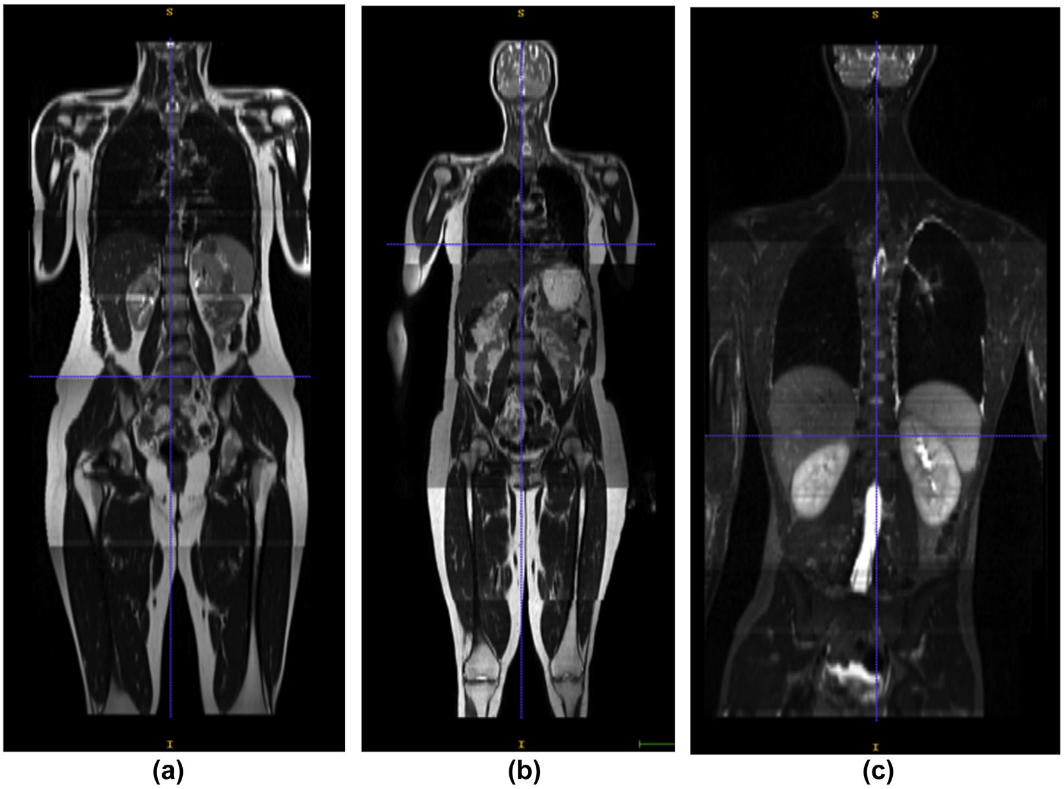


Figure 2 Different variants of a T2W WB-MRI protocol. (a) Non-fat-suppressed T2W images covering the body from the neck to mid-thighs. (b) Non-fat-suppressed T2W images covering the body from the top of the head to mid-calves. (c) Fat-suppressed T2W images covering the body from the middle of the head to the pelvis. Note the anatomical and signal intensity variability, which is of particular importance in machine learning imaging studies.

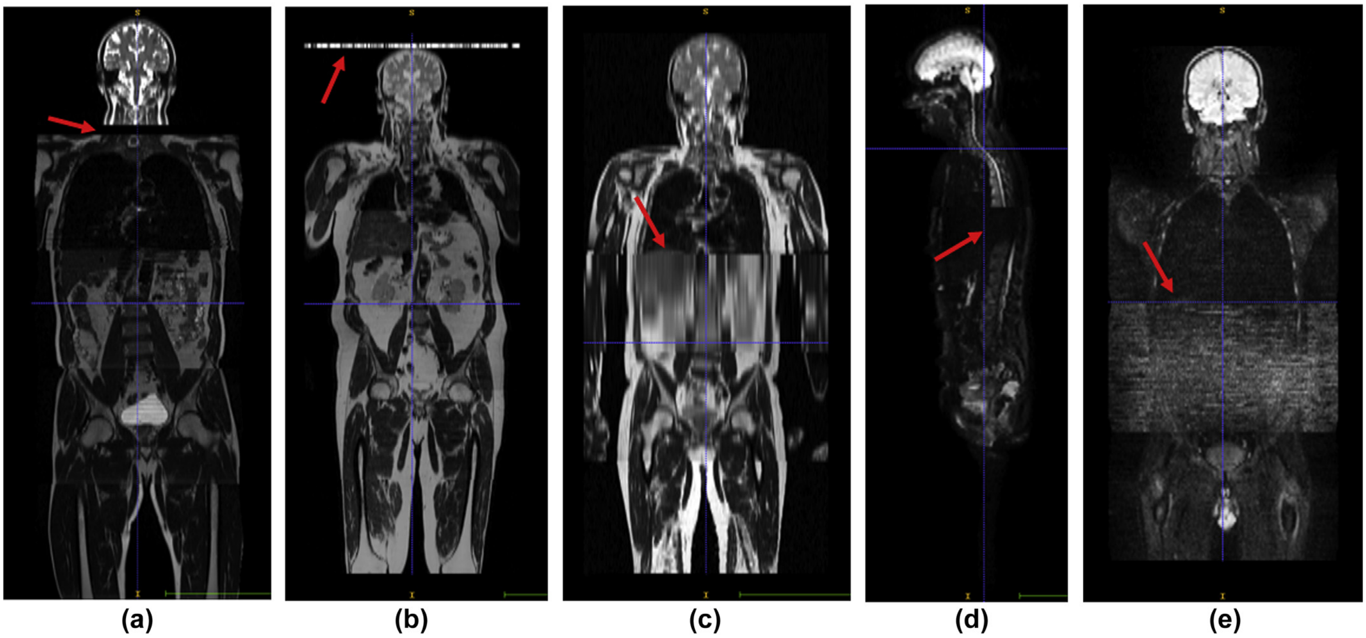


Figure 3 Demonstrating some of the data quality challenges (artefacts) we encountered in the datasets used in MALIBO. Missing slices (a), RF interference (b), and motion artefacts (c) on T2W images. RF field inhomogeneities leading to dielectric shading (d) and RF noise in DW images.

Data preparation

Data preparation or pre-processing is an essential step in any machine learning study, whether related to imaging or not. In MALIBO, where WB-MRI data from multiple imaging stations were acquired, we converted all our datasets in compressed Nifti format (nii.gz), in the interest of space and machine learning pipeline efficiency, after stitching images together according to slice location to form WB volumes. It should be noted that, in case of DICOM data conversion to other "headerless" formats, the original data should be retained so that header information can be "glued" back to the converted images for uploading to the reading platform, as the majority of reading platforms accommodate DICOM data only.

Signal intensity standardisation

As discussed earlier, the richness of acquisition schemes in MRI, comes with a major challenge. Unlike other medical imaging modalities, the image intensities in MRI do not have a fixed interpretation, not even within the same protocol or when acquired in the same body region, using the same scanner for the same patient.[17] In MALIBO, this even applies between imaging stations in WB acquisitions. This lack of a fixed meaning for intensities poses problems, not only when it comes to image quantification, but also in machine learning tasks, such as image segmentation. Therefore, it is essential that an MRI signal intensity standardisation step is incorporated in the preparation pipeline before extracting the features in supervised learning algorithms or feeding the images in deep learning algorithms.

In MALIBO we designed a specific pre-processing pipeline for intensity normalisation across images. We initially experimented with simple intra-subject intensity scaling, based on signal normalisation using the 4th and 94th percentiles of the intensity histogram, a somewhat arbitrary choice, which has been shown to work well for brain imaging[18]; however, in WB imaging there is the challenge of inconsistent anatomical coverage due to protocol variability, as discussed above. A number of WB volumes used in MALIBO, fully included the head and neck regions down to the lower limbs, while others only covered the body from the shoulders down to knees (Fig 2). This violates the assumption that statistics, such as percentiles obtained from the image intensity histograms, correspond to similar anatomical regions. To address this, we make use of a rigid registration technique to approximately align all images to a reference image. In this way, the field of view between the tested and training images is normalised and similarity between the histogram statistics is ensured.

This then allows us to employ Nyul's intensity normalisation technique,[19] which involves two stages. In the learning stage, a standard scale is derived from the intensity histograms of the training images using 10, uniformly distributed, histogram landmarks ranging from the 1st to the 99th percentile. In the testing stage, any new image, following rigid registration to the reference image, can then be mapped to the intensity standard scale, using the learned transformation from the training stage. Fig 4 shows an example of using this pipeline on a whole-body T2-weighted (T2W) volume.

Other histogram-based methods to perform intra- and inter-subject signal intensity standardisation for the same acquisition protocol are currently explored and compared to the existing pipeline.[20]

Generating training data

Generating training data for machine learning algorithms is one of the most important, but also laborious and time-consuming processes. Manual, volumetric segmentations performed by clinical experts, should be used to ensure reliable and accurate algorithmic training. These labelled data, should also be used as the reference standard to compare with, when evaluating algorithmic performance. Semi-automatic or fully automatic methods can also be used to alleviate part of the workload, but it is suggested that these segmentations are always double-checked and finalised by a clinical expert. In MALIBO, we used ITK-SNAP[21] to manually generate annotated WB images. Labelling of heathy structures (23 anatomical structures, including organs and bones) occupied a significant proportion of Phase 1 of the project, but this work was of paramount importance as in Phase 2 we are using a two-stage approach, to identify cancer lesions, as will be discussed below.

Image registration

The use of multimodal MRI data ("multi-channel" data as commonly referred to in computer science terminology) has been shown to improve algorithmic performance in tasks such as brain lesion segmentation[22]; however, using multichannel inputs for algorithm training requires optimally registered imaging datasets between modalities, so that annotated data from a single modality are used—in the interest of time efficiency—when generating training data. Anatomically matched datasets from different modalities, is a task that can be performed efficiently enough in the brain, where minimal gross motion or anatomical deformation is expected between acquisitions, with using a rigid registration algorithm.

In abdominal imaging, where there might be significant organ motion and deformation between acquisitions, a rigid registration might not suffice. The task proved to be even more challenging with WB-MRI data. Furthermore, when we attempted to register DWI volumes to anatomical volumes, we encountered the extra challenge from the geometrically distorted EPI-acquired, high b-value DW volumes.[15] We qualitatively assessed registration between DWI and anatomical volumes, when using a 12 degrees-of-freedom affine registration,[23] but with mixed results. A non-rigid registration using free-form deformations[24] was also tested, but the time required to apply on the tens of WB datasets used in MALIBO was unacceptably long. At this stage of MALIBO, we simply use slice-matched acquisitions,

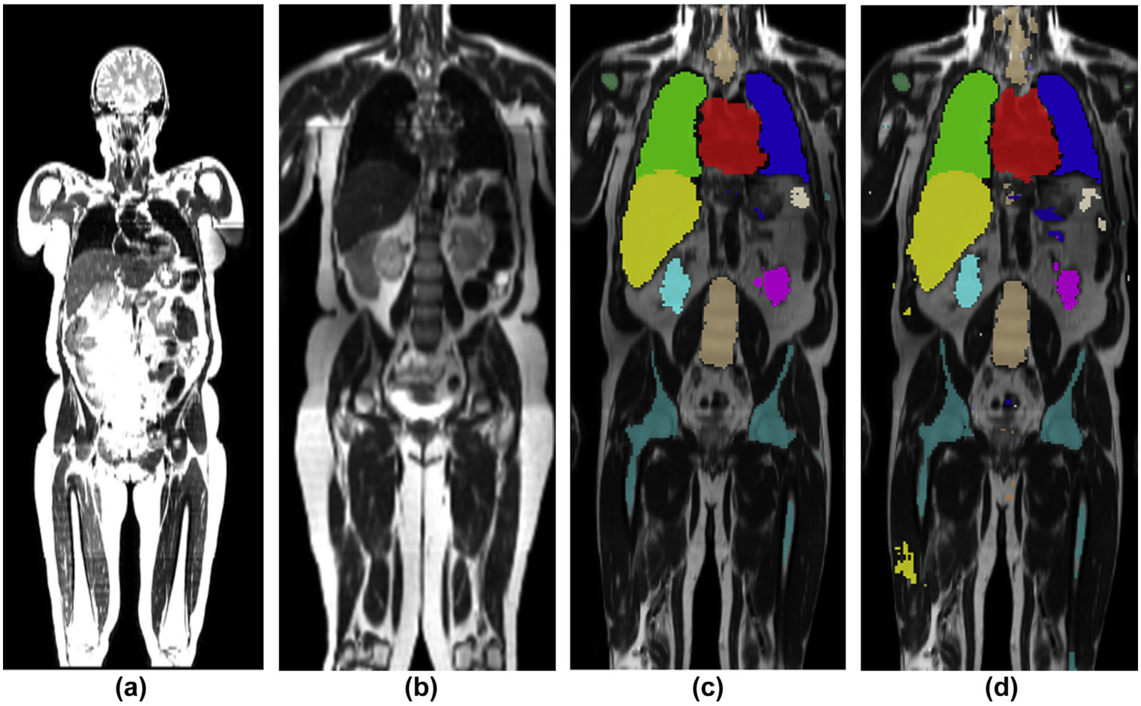


Figure 4 Using intensity normalisation pipeline on a test image. (a) Original T2W volume. (b) Same image, but scale-matched using Nyul's histogram-based method described in the text, following rigid registration. The two volumes are displayed using the same window/level settings. Employing Nyul's histogram-based method improved healthy organ detection on previously unseen T2W images (c), when compared to using the simple signal normalisation based on the 4th and 94th percentiles of the intensity histogram (d).

resampled to match the spatial resolution of the reference (T2-weighted) volumes. This aligns the majority of structures, in particular bones, very well between modalities, but ignores differences due to breathing or other movements of the subjects between scans. A block diagram of the data preparation pipeline for MALIBO, as described above, is shown in Fig 5.

Machine learning pipeline

Choice of algorithm and feature crafting

The choice of machine learning algorithm will depend on the task at hand. Unfortunately, there is no "one-size-fits-all" recipe and so, the choice comes down to a recursive trial-and-error process, until the desirable performance and characteristics are reached. The number of supervised, state-of-the-art, algorithms suited for imaging-related tasks and their variants, but also the choice for the hyper-parameters in each individual method may seem infinite; previous experience, already published results and the quality and quantity of available data for training should provide guidance for a good starting point.

Another important consideration for algorithm selection is whether model interpretability is of interest for the task at hand. Deep learning algorithms have demonstrated great accuracy in imaging-related tasks,[6] but interpreting the extracted features and the complex, non-linear relationships between them, which take place in the hidden layers of the network, remains an almost impossible challenge. Despite the fact that there are now ways to visualise the features that activate specific neurons in a layer,[25] the hidden layers of a deep convolutional neural network still have the traits of a "black box".

In MALIBO, we mainly tested and evaluated two algorithms; one state-of-the-art ensemble algorithm based on classification forests (CFs)[26,27] and one deep learning algorithm based on convolutional neural networks (CNNs).[28] Classification forests are powerful, multi-label classifiers, which facilitate the simultaneous segmentation of multiple organs. They have very good generalisation properties, which means they can be effectively trained using a limited number of datasets. Both of these traits were desirable in MALIBO. Our convolutional neural networks implementation was based on MALIBO,[28,29] an approach that has been shown to perform very well in brain lesion segmentation with multiparametric MRI data.[22] The details of the hyper-parameters used for the CFs and network architecture for the CNNs, can be found elsewhere.[30] CNNs performed consistently better in healthy organ segmentation in Phase 1 of MALIBO, so it was the algorithm of choice for Phase 2 of the project (lesion detection).

Pipeline adjustments for task at hand and performance evaluation

Whether the task at hand is organ or lesion classification, segmentation or detection, the core of the pipeline will most commonly be an accurate and robust classifier. In

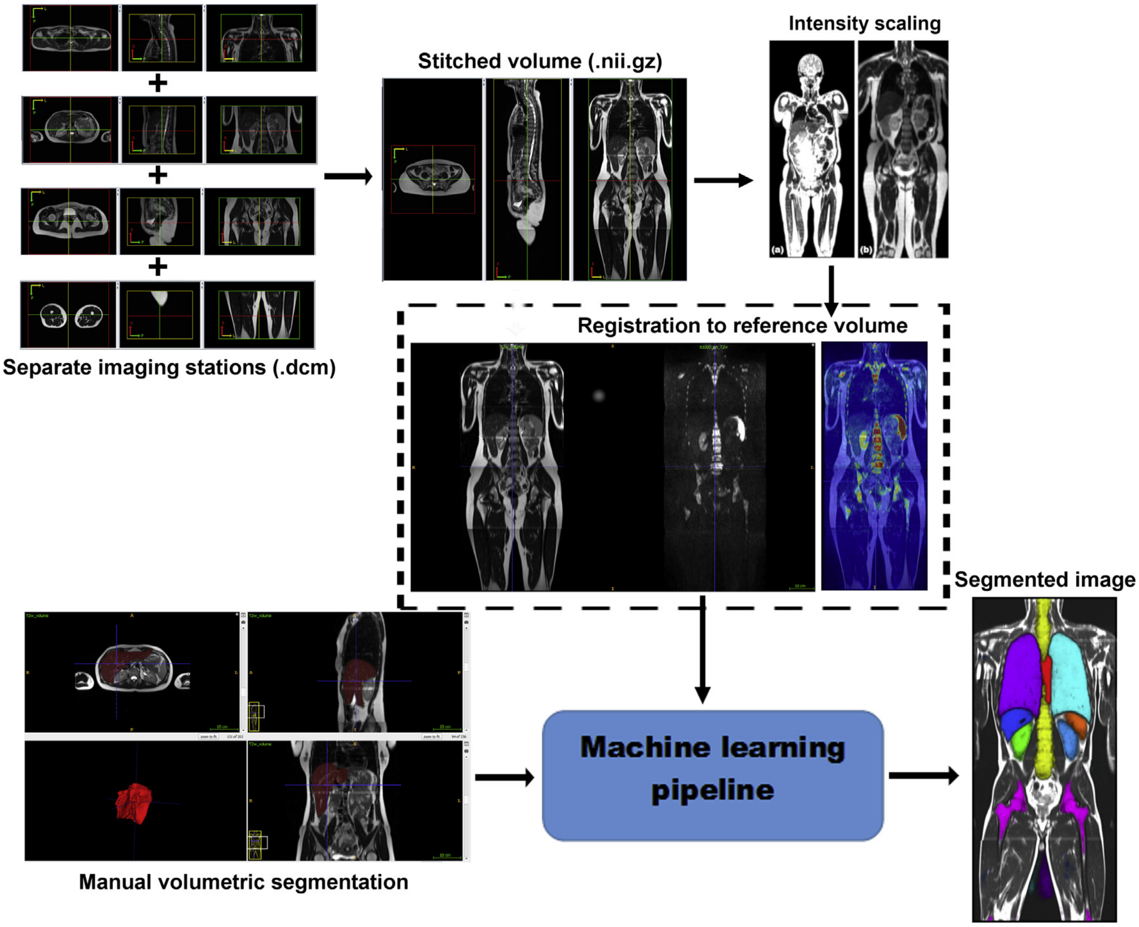

Separate imaging stations (.dcm)
Stitched volume (.nii.gz)
Intensity scaling
(a)
(b)
Registration to reference volume
Segmented image
Manual volumetric segmentation
Machine learning pipeline

Figure 5 Block diagram of the MALIBO data preparation pipeline.

MALIBO Phase 2 we were interested in lesion localisation and characterisation, rather than segmentation. We therefore had to employ a scheme to evaluate the segmentation algorithms used in Phase 1, but now in terms of detection. A specific automatic evaluation procedure was implemented to calculate detection accuracy. This uses as inputs the manual reference segmentation and the detection map from the segmentation algorithm and calculates the true positive rate, positive predictive value and F1 score, based on a user defined distance threshold (in mm). An example plot of the accuracies for a range of detected lesions and manual segmentations distance is shown in Fig 6.

We then used the CNN algorithm, developed in Phase 1 of MALIBO, to evaluate the performance of detected primary colon lesions from colorectal cancer patients, scanned with WB-MRI.[13] We observed that lesion detection in WB scans was suboptimal with the CNNs, presumably due to the small fraction of lesion volume occupying the scanned space, when compared to the WB volume. The complexity of intensities in background tissue and the lesion weak boundaries appeared to be confusing the CNN.[31]

We therefore, had to adapt our approach to become a two-stage process, whereby in the first stage, the information from Phase 1 healthy organs/bones is used to identify normality and in stage two the lesion is detected (Phase 2 of MALIBO). Stage two can be modular with respect to the anatomical location that the suspected lesion can be found. According to this and the availability of training data, the

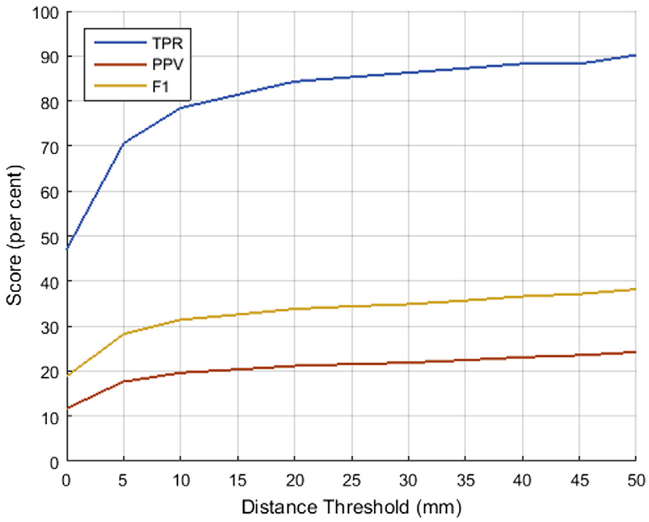


Figure 6 Primary colon lesion detection accuracies (true positive rate [TPR], positive predictive value [PPV], and F1 score) for different ground truth-detection distances, when using the CF algorithm.

architecture and configuration of the used CNN can be modified to achieve optimal performance. This work is now ongoing and the aforementioned process is depicted in Fig 7.

Finally, post-processing steps are required to prepare the machine learning output for reading. In MALIBO, the final probability maps obtained from the CNN were smoothed, normalised and “thresholded” to reduce false positives and improve visual appearance for the reading process.

An integrated machine learning pipeline should also incorporate an objective performance evaluation stage. The choice of performance assessment metrics will, once again, depend on the examined data availability and the task at hand. In MALIBO, we evaluated segmentation tasks using cross-validation and a range of overlap and distance metrics[32] and detection, using the scheme described above.

Reading process

Reading platforms

Traditionally, the picture archiving and communications system (PACS) is used for hosting medical images and associated reader’s reports; however, PACS is not flexible enough to accommodate hanging protocols for machine learning outputs and also, access from readers external to the hosting institution is not possible. In MALIBO, we have used a secure central imaging server (3Dnet), provided by Biotronics3D (London, UK),[33] to ensure that images and related machine learning output, are hosted in an environment where customised hanging protocols can be created and images are accessible by all readers via a standard internet connection.

A hanging protocol was created for MALIBO readers in Biotronics3D, so that stitched volumes from different imaging modalities, alongside the machine learning output, are opened and browsed simultaneously, as shown in Fig 8. This setting also allows for the anatomical localisation using cross-hairs and also fusion between the colour-mapped machine learning output and any of the MRI modalities.

Reading paradigm and reading process

In MALIBO, we have used a similar reading paradigm and case report forms (CRFs) to the contributing studies,[13,14] with slight modifications to account for the machine learning output effects in the source study’s diagnostic performance and reading time. Pilot testing of case report forms (CRFs) used randomised reads of anonymised scans from colorectal cancer patients,[13] which were performed by six independent readers. Before the reading process, it was essential that the involved study readers met and reached a consensus as to how the machine learning output will be interpreted (based on suspicious lesion’s size and location, detection probability value, etc.).

Miscellaneous issues

Data and databases access

In the era of machine learning in radiology, there is a need for well-organised, suitably anonymised, and accurately annotated database of images, annotations, and metadata throughout all stages of such studies. File nomenclature, which should be clearly defined, needs to be available to all those involved with password-controlled access to data. This may include multiple radiologists undertaking human expert segmentation and standardisation of file names, which is essential for proper management of the large number of files. In addition, version control is an important concern, which needs attention during the iterative training process. As described in Kohli 2017[34] ideal datasets for radiology machine learning studies should be FAIR (Findable, Accessible, Interoperable and Reusable). In

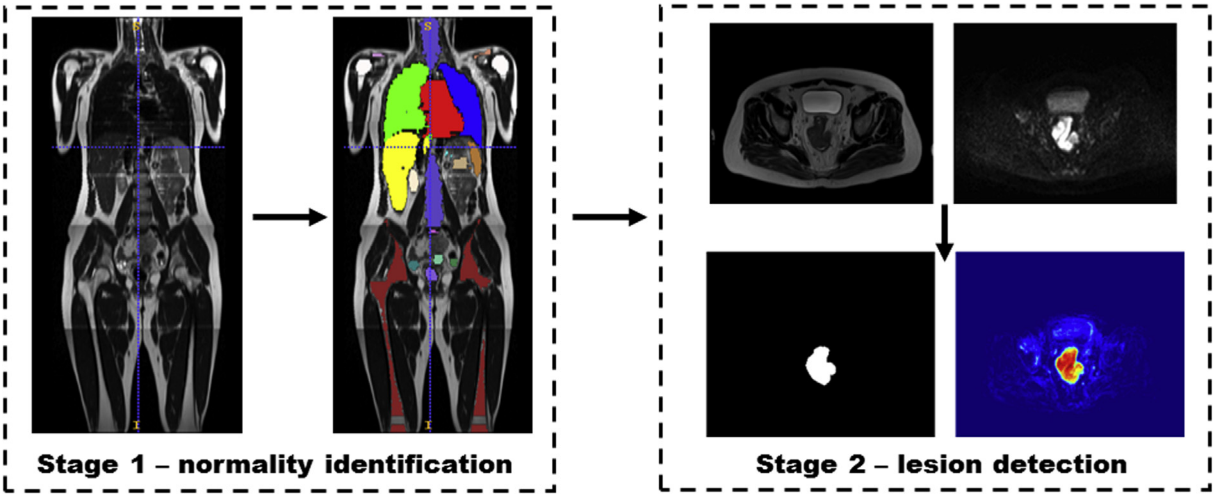

Stage 1 – normality identification
Stage 2 – lesion detection

Figure 7 Two-stage lesion detection process, employed in MALIBO Phase 2. During stage 1, the normal organs/bones are identified, based on Phase 1 training. During stage 2, lesion detection takes place. Stage 2 can be modular, with each module algorithm training, depending on anatomical position.

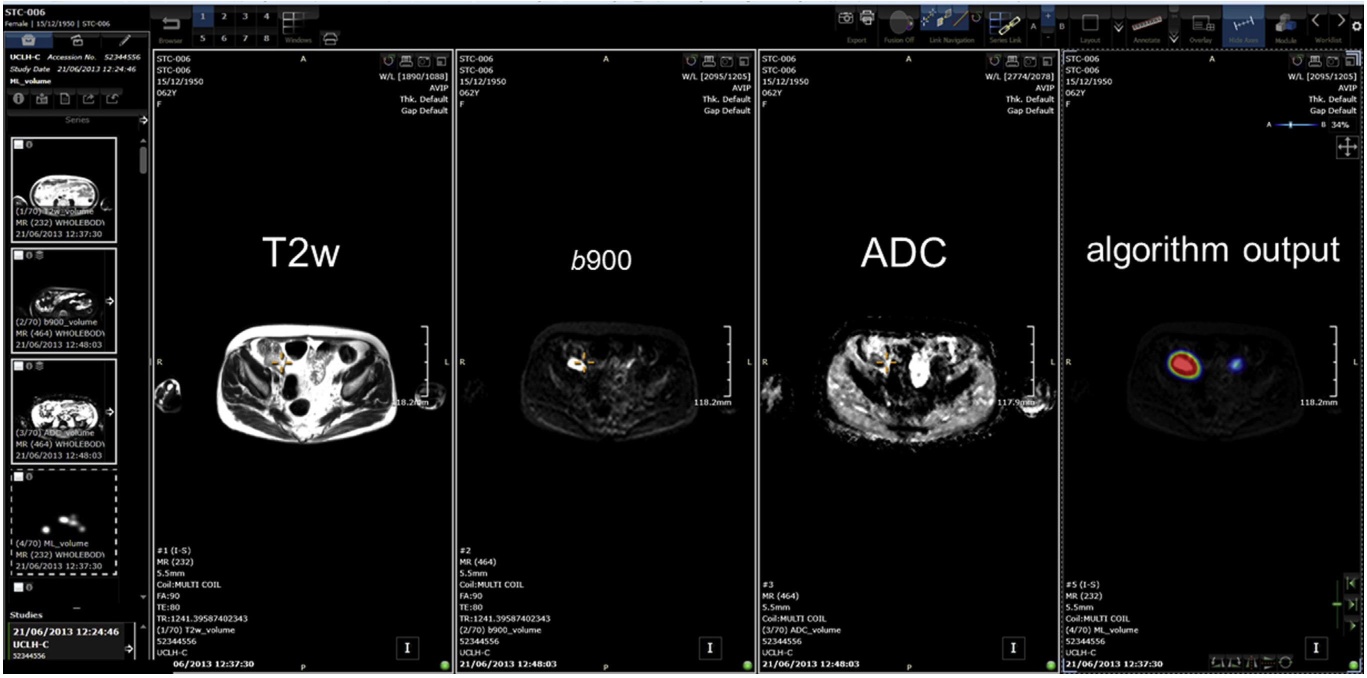


Figure 8 Biotronics 3D view of the WB volumes from different modalities and the algorithm output, fused with the DWI image from a colon lesion.

MALIBO, imaging data, metadata and annotations were stored in a dedicated, secure workstation. Data sharing and reporting was accomplished via Biotronics3D.

In another NIHR-funded study involving WB–MRI data (MAchine Learning In MyelomA Response: MALIMAR study, EME project 16/68/34), the use of XNAT[35] for the aforementioned tasks is currently being optimised. XNAT is an open-source, extensible and flexible database system that allows for image, annotations and metadata storage, sharing, and management.

Legal, ethical, and clinical acceptance

Data sharing agreements are an essential step in studies where data are being shared between collaborators. Each involved party, needs to be clear and transparent concerning the data to be shared and agreements with respect to background and foreground intellectual property should also be in place. Local contract negotiations are required prior to study commencement. Agreement for data sharing from the source study funders, trial management group, trial steering committee, and sponsor should be obtained in writing.

Ethics considerations will vary depending on the arrangements of the primary source studies. For the MALIBO study, ethics approvals were available from each of the contributing studies for use of the data and, in addition, an institutional research and development approval with information governance agreement were all in place for the MALIBO protocol at the start of the study. Public and patient representation in the trial management group is important to ensure that the patient's voice is heard in the planning of the study and in the dissemination of the findings and public acceptance of the use of machine learning support tools.

Clinical acceptance is also an important consideration in machine learning-related imaging studies. The validation of the developed machine learning tools needs to stand up to scrutiny and the methods used for testing the tools need to be clear to clinical radiologists. In MALIBO, we have devised a viewing framework that is widely used by radiologists and incorporates the machine learning tools into a typical clinical environment for testing.

Conclusion

Machine learning algorithms can now perform image analysis tasks with performance equal, or even superior, to the one achieved by human experts. Automatically derived measurements and visual guides, obtained with machine learning techniques will serve as a valuable aid in many clinical tasks and, most certainly, will transform the ways we see and use medical imaging analysis tools.

We have used MALIBO, a study that is looking into developing machine learning methods for improving the diagnostic performance and reducing the reading time of WB-MRI data, as a platform for identifying some of the main challenges encountered in a clinical study involving machine learning. Our experiences are described in this manuscript. Given the pragmatic setting of MALIBO, we believe that the methodological steps and challenges described here, can be of invaluable assistance, and can serve as a guide, to groups who would like to apply similar studies in the future, not only for MRI, but in radiology generally.

One of the most important considerations when designing a clinical study involving machine learning is data readiness. Acquired and used data should be assessed in the context of appropriateness with quality and uniformity being the two most important parameters to be considered. If these data traits cannot be assured upon design, then appropriate steps towards upgrading the data level readiness should be taken or even manually identify the appropriate datasets if necessary. A robust machine learning pipeline should be designed and implemented, a task that should now be straightforward to accomplish, given that robust machine learning libraries, modules and toolboxes are now freely available, to implement a vast amount of algorithms and preparation/evaluation schemes. An important consideration for achieving the desired clinical outcome is to effectively host the resulting machine learning output, along with the clinical images, for reading. Once again, there are now a range of cloud-based services available to facilitate this process. The reading paradigm and reading process should be agreed by the readers in consensus. Finally, a range of legal, ethical, and clinical acceptance issues should be considered when attempting to incorporate computer-assisting tools into clinical trials.

In conclusion, clinical studies involving the development and use of machine learning methodology require careful design, if the study objectives are to be accomplished and the employed methods to reach their full potential. The road from translating computing methods into potentially useful clinical tools involves an analytical, stepwise adaptation approach, as well as engagement of a multidisciplinary team.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank the MALIBO readers and co-investigators for their contribution to the study design and implementation, and Konstantinos Kamnitsas for sharing his expertise in implementing DeepMedic to MALIBO data. Thanks to Dr Fahdi Kanavati for the interesting conversations regarding machine learning. Finally, the authors thank researchers at Biotronics3D, London, UK, for their assistance with accommodating the MALIBO images and output in 3Dnet. The MALIBO project is funded by the Efficacy and Mechanism Evaluation (EME) Programme, an MRC and NIHR partnership. The views expressed in this publication are those of the author(s) and not necessarily those of the MRC, NHS, NIHR, or the Department of Health and Social Care. The authors also acknowledge infrastructure support from Imperial College Biomedical Research Centre and Experimental Cancer Medicines Centre. This study was supported by the National Institute of Health Research (EME Project: 13/122/01).

References

1. Wang S, Summers RM. Machine learning and radiology. *Med Image Anal* 2012;**16**(5):933–51.
2. Erickson BJ, Korfiatis P, Akkus Z, *et al*. Machine learning for medical imaging. *RadioGraphics* 2017;**37**(2):505–15.
3. Kohli M, Prevedello LM, Filice RW, *et al*. Implementing machine learning in radiology practice and research. *AJR Am J Roentgenol* 2017;**208**(4):754–60.
4. Chartrand G, Cheng PM, Vorontsov E, *et al*. Deep learning: a primer for radiologists. *RadioGraphics* 2017;**37**(7):2113–31.
5. Erickson BJ, Korfiatis P, Kline TL, *et al*. Deep learning in radiology: does one size fit all? *J Am Coll Radiol* 2018;**15**(3 Pt B):521–6.
6. Mazurowski M, Buda M, Saha AR, *et al*. *Deep learning in radiology: an overview of the concepts and a survey of the state of the art*. 2018. arXiv.org; arXiv:1802.08717.
7. Takahara T, Imai Y, Yamashita T, *et al*. Diffusion weighted whole body imaging with background body signal suppression (DWIBS): technical improvement using free breathing, STIR and high resolution 3D display. *Radiat Med* 2004;**22**(4):275–82.
8. Koh DM, Collins DJ. Diffusion-weighted MRI in the body: applications and challenges in oncology. *AJR Am J Roentgenol* 2007;**188**(6):1622–35.
9. Schmidt GP, Reiser MF, Baur-Melnyk A. Whole-body MRI for the staging and follow-up of patients with metastasis. *Eur J Radiol* 2009;**70**(3):393–400.
10. Wu L-M, Gu H-Y, Zheng J, *et al*. Diagnostic value of whole-body magnetic resonance imaging for bone metastases: a systematic review and meta-analysis. *J Magn Reson Image* 2011;**34**(1):128–35.
11. Padhani AR, Koh D-M, Collins DJ. Whole-body diffusion-weighted MR imaging in cancer: current status and research directions. *Radiology* 2011;**261**(3):700–18.
12. Rockall AG, Glocker B, Rueckert D, et al. Development and evaluation of machine learning methods in whole body magnetic resonance imaging (MRI), with diffusion-weighted imaging (DWI), for staging of patients with cancer (Machine Learning In Body Oncology: MALIBO). Study protocol. January 2019. Manuscript in preparation.
13. Taylor SA, Mallett S, Miles A, *et al*. Streamlining staging of lung and colorectal cancer with whole body MRI; study protocols for two multicentre, non-randomised, single-arm, prospective diagnostic accuracy studies (Streamline C and Streamline L). *BMC Cancer* 2017;**17**(1):299.
14. Latifoltojar A, Punwani S, Lopes A, *et al*. Whole-body MRI for staging and interim response monitoring in paediatric and adolescent Hodgkin's lymphoma: a comparison with multi-modality reference standard including 18F-FDG-PET-CT. *Eur Radiol* 2019;**29**(1):202–12.
15. Le Bihan D, Poupon C, Amadon A, *et al*. Artifacts and pitfalls in diffusion MRI. *J Magn Reson Image* 2006;**24**(3):478–88.
16. Lawrence ND. *Data readiness levels*. 2017. arXiv.org; arXiv:1705.02245.
17. Nyul LG, Udupa JK, Xuan Z. New variants of a method of MRI scale standardization. *IEEE Trans Med Image* 2000;**19**(2):143–50.
18. Sun X, Shi L, Luo Y, *et al*. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *BioMed Eng OnLine* 2015;**14**(1):73.
19. Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. *Magn Reson Med* 1999;**42**(6):1072–81.
20. Madabhushi A, Udupa JK. New methods of MR image intensity standardization via generalized scale. *Med Phys* 2006;**33**(9):3426–34.
21. Yushkevich PA, Piven J, Hazlett HC, *et al*. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* 2006;**31**(3):1116–28.
22. Geremia E, Zikic D, Clatz O, *et al*. Classification forests for semantic segmentation of brain lesions in multi-channel MRI. In: Criminisi A, Shotton J, editors. *Decision forests for computer vision and medical image analysis*. London: Springer; 2013. p. 245–60.
23. Studholme C, Hill DLG, Hawkes DJ. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recog* 1999;**32**(1):71–86.
24. Rueckert D, Sonoda LI, Hayes C, *et al*. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Image* 1999;**18**(8):712–21.

25. Karpathy A. Convolutional neural networks for visual recognition. Available at: http://cs231n.github.io/understanding-cnn/. Accessed September 2018.
26. Breiman L. Random forests. *Machine Learn* 2001;**45**(1):5–32.
27. Glocker B, Konukoglu E, Haynor DR. Random forests for localization of spinal anatomy. In: Zhou S, editor. *Medical recognition, segmentation and parsing*. London: Academic Press; 2015. p. 94–109.
28. Kamnitsas K, Ledig C, Newcombe VFJ, *et al*. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 2017;**36**:61–78.
29. Kamnitsas K. *DeepMedic source code 2016*. 2016 [cited 2016 Accessed November, 2016]; DeepMedic source code. Available from: https://github.com/Kamnitsask/deepmedic.
30. Lavdas I, Glocker B, Kamnitsas K, *et al*. Fully automatic, multiorgan segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs), and a multi-atlas (MA) approach. *Med Phys* 2017;**44**(10):5210–20.
31. Valindria V, Lavdas I, Cerrolaza J, *et al*. *Small organ segmentation in whole-body MRI using a two-stage FCN and weighting schemes*. 2018. arXiv.org; arXiv:1807.11368.
32. Heimann T, van Ginneken B, Styner MA, *et al*. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Image* 2009;**28**(8):1251–65.
33. Biotronics3D. Biotronics3D, Analyze—collaborate—discover. Available at: https://www.biotronics3d.com/public/. Accessed September 2018.
34. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning. Whitepaper from the 2016 C-MIMI Meeting Dataset Session. *J Digit Image* 2017;**30**(4):392–9.
35. XNAT. XNAT, The most widely-used informatics platform for imaging research. Available at: https://www.xnat.org/. Accessed September 2018.