Keywords: whole body magnetic resonance imaging, machine learning, deep learning, random forests, convolutional neural networks, lesion detection, cancer

Introduction

Machine learning applications are ever-present in our daily activities, whether the beneficiary is aware of it or 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 whole body protocols, means that WB-(DW)-MRI is now becoming an increasingly important tool in oncology for cancer diagnosis, staging and treatment response monitoring [7-9]. A significant challenge when reading whole body MRI scans, is the increased volume of resulting imaging data, especially when multi-parametric acquisitions are used. The reading process can then become rather time-consuming, with increased risk of misinterpretations. Also, whole body 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 XXXXX), which aims to develop state-of-the art machine learning algorithms for the automatic detection of malignant and benign lesions in multi-centre, multi-parametric whole body MRI scans [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 whole body MRI scans. 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 XXXXXX (MAchine Learning In Body Oncology) study

XXXXXX 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-(DW)-MRI, 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]. XXXXXX is funded by the NIHR, Efficacy and Mechanism Evaluation programme (EME project: XXXX) and is a collaboration between the XXXXX and the XXXXX. Contributing studies’ data are provided by the XXXXX and XXXXX.

The study is divided into three phases, whereby in Phase 1 algorithms are developed and evaluated for their accuracy to identify normal structures in whole body 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’ whole body MRI data. As far as we are aware, XXXXXX is the first study that applies machine learning techniques in WB-(DW)-MRI.

The XXXXXX study relies on whole body 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 is 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. Figure 1 shows a block diagram identifying the XXXXXX phases, during which the most significant challenges have been encountered to date and for which our methodology required adaptation.

1. 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 multi-centre collaborations to 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 is likely to be fairly homogeneous, although differences in slice thickness or differences in the use of contrast may pose challenges. However, in the MRI setting, as encountered in XXXXXX, 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.

1.1 MRI systems and acquisition protocol variabilities

The MRI systems used in multi-centre 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 whole body protocols as in XXXXXX, 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 XXXXXX data, in Figure 2. Methods with which a number of the variability issues mentioned above, were mitigated in XXXXXX, are described in the ‘Data preparation’ section.

1.2 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 multi-centre 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 XXXXXX, were not suited for machine learning purposes and had to be discarded. Figure 3 shows some of the image quality issues we encountered in XXXXXX.

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. This issue also raises the much wider question of acquisition uniformity throughout the radiology community, in order to harness the potential benefits of applying machine learning techniques in the future.

2. Data preparation

Data preparation or pre-processing is an essential step in any machine learning study, whether related to imaging or not. In XXXXXX, where whole body 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 whole body 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 these accommodate almost exclusively DICOM data.

2.1 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 XXXXXX, this even applies between imaging stations in whole body 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 XXXXXX 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 whole body imaging there is the challenge of inconsistent anatomical coverage due to protocol variability, as discussed in Section 1.1. A number of whole body volumes used in XXXXXX, fully included the head and neck regions down to the lower limbs, while others only covered the body from the shoulders down to knees (Figure 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 ten, 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. Figure 4 shows an example of using this pipeline on a whole body 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].

2.2 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 XXXXXX, we used ITK-SNAP [21] to manually generate annotated whole body 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 in Section 3.2.

2.3 Image registration

The use of multi-modal MRI data ('multi-channel' data as commonly referred to in computer science terminology) has been shown to improve algorithmic performance in tasks like brain lesion segmentation [22]. However, using multi-channel 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 which 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 whole body 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 whole body datasets used in XXXXXX was unacceptably long. At this stage of XXXXXX, we simply use slice-matched acquisitions, 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 XXXXXX, as described in Section 2, is shown in Figure 5.

3. Machine learning pipeline

3.1 Choice of algorithm and feature crafting

The choice of machine learning algorithm will depend on the task at hand. Unfortunately, there is no ‘one-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 the 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 XXXXXX, 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 XXXXXX. Our convolutional neural networks implementation was based on XXXXX [28, 29], an approach which has been shown to perform very well in brain lesion segmentation with multi-parametric MRI data [22]. The details of the hyperparameters 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 XXXXXX, so it was the algorithm of choice for Phase 2 of the project (lesion detection).

3.2 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 XXXXXX 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 Figure 6.

We then used the CNN algorithm, developed in Phase 1 of XXXXXX, to evaluate the performance of detected primary colon lesions from colorectal cancer patients, scanned with whole body MRI [13]. We observed that lesion detection in whole body scans was suboptimal with the CNNs, presumably due to the small fraction of lesion volume occupying the scanned space, when compared to the whole body 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 XXXXXX). 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 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 Figure 7.

Finally, post-processing steps are required to prepare the machine learning output for reading. In XXXXXX, 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 XXXXXX, we evaluated segmentation tasks using cross-validation and a range of overlap and distance metrics [32] and detection, using the scheme described above.

4. Reading process

4.1 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 XXXXXX, 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 XXXXXX readers in Biotronics3D, so that stitched volumes from different imaging modalities, alongside the machine learning output, are opened and browsed simultaneously, as shown in Figure 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.

4.2 Reading paradigm and reading process

In XXXXXX, 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 6 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.).

5. Miscellaneous issues

5.1 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 XXXXXX, 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 whole body MRI data (MAchine Learning In MyelomA Response - XXXXX study, EME project XXXXX), 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.

5.2 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 XXXXXX 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 XXXXXX 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 XXXXXX, 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.

Discussion- 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 XXXXXX, a study that is looking into developing machine learning methods for improving the diagnostic performance and reducing the reading time of whole body 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 XXXXXX, 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 which 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 multi-disciplinary team.

References

1. Wang S and Summers RM.Machine learning and radiology. Medical Image Analysis 2012; 16(5): 933-951.
2. Erickson BJ, Korfiatis P, Akkus Z, and Kline TL.Machine Learning for Medical Imaging. Radiographics : a review publication of the Radiological Society of North America, Inc 2017; 37(2): 505-515.
3. Kohli M, Prevedello LM, Filice RW, and Geis JR.Implementing Machine Learning in Radiology Practice and Research. American Journal of Roentgenology 2017; 208(4): 754-760.
4. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al.Deep Learning: A Primer for Radiologists. RadioGraphics 2017; 37(7): 2113-2131.
5. Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, and Weston AD.Deep Learning in Radiology: Does One Size Fit All? Journal of the American College of Radiology : JACR 2018; 15(3 Pt B): 521-526.
6. Mazurowski M, Buda M, Saha A, and R. Bashir M, *Deep learning in radiology: an overview of the concepts and a survey of the state of the art*. 2018.
7. Takahara T, Imai Y, Yamashita T, Yasuda S, Nasu S, and Van Cauteren M.Diffusion weighted whole body imaging with background body signal suppression (DWIBS): technical improvement using free breathing, STIR and high resolution 3D display. Radiation Medicine 2004; 22(4): 275-282.

8. Koh DM and Collins DJ.Diffusion-Weighted MRI in the Body: Applications and Challenges in Oncology. American Journal of Roentgenology 2007; 188(6): 1622-1635.

9. Schmidt GP, Reiser MF, and Baur-Melnyk A.Whole-body MRI for the staging and follow-up of patients with metastasis. European Journal of Radiology 2009; 70(3): 393-400.

10. Wu L-M, Gu H-Y, Zheng J, Xu X, Lin L-H, Deng X, et al.Diagnostic value of whole-body magnetic resonance imaging for bone metastases: a systematic review and meta-analysis. Journal of Magnetic Resonance Imaging 2011; 34(1): 128-135.

11. Padhani AR, Koh D-M, and Collins DJ.Whole-Body Diffusion-weighted MR Imaging in Cancer: Current Status and Research Directions. Radiology 2011; 261(3): 700-718.

12. XXXXX

13. XXXXX

14. XXXXX

15. Le Bihan D, Poupon C, Amadon A, and Lethimonnier F.Artifacts and pitfalls in diffusion MRI. Journal of Magnetic Resonance Imaging 2006; 24(3): 478-488.

16. D. Lawrence N, *Data Readiness Levels*. 2017.

17. Nyul LG, Udupa JK, and Xuan Z.New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging 2000; 19(2): 143-150.

18. Sun X, Shi L, Luo Y, Yang W, Li H, Liang P, et al.Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. BioMedical Engineering OnLine 2015; 14(1): 73.

19. Nyúl LG and Udupa JK.On standardizing the MR image intensity scale. Magnetic Resonance in Medicine 1999; 42(6): 1072-1081.

20. Madabhushi A and Udupa JK.New methods of MR image intensity standardization via generalized scale. Medical Physics 2006; 33(9): 3426-3434.

21. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al.User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 2006; 31(3): 1116-1128.

22. Geremia E, Zikic D, Clatz O, Menze BH, Glocker B, Konukoglu E, et al., *Classification Forests for Semantic Segmentation of Brain Lesions in Multi-channel MRI*, in *Decision Forests for Computer Vision and Medical Image Analysis*, A. Criminisi and J. Shotton, Editors. 2013, Springer London: London. p. 245-260.

23. Studholme C, Hill DLG, and Hawkes DJ.An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition 1999; 32(1): 71-86.

24. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, and Hawkes DJ.Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging 1999; 18(8): 712-721.

25. *Convolutional Neural Networks for Visual Recognition. Available via: http://cs231n.github.io/understanding-cnn/. . Accessed in September 2018.*

26. Breiman L.Random Forests. Machine Learning 2001; 45(1): 5-32.

27. Glocker B, Konukoglu E, and Haynor DR, *Random Forests for Localization of Spinal Anatomy*, in *Medical Recognition, Segmentation and Parsing*, S. Zhou, Editor. 2015, Academic Press, Elsevier: London. p. 94-109.

28. XXXXX

29. XXXXX

30. XXXXX

31. Valindria V, Lavdas I, Cerrolaza J, O. Aboagye E, G. Rockall A, Rueckert D, et al., *Small Organ Segmentation in Whole-body MRI using a Two-stage FCN and Weighting Schemes*. 2018.

32. Heimann T, van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, Bauer C, et al.Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. Medical Imaging, IEEE Transactions on 2009; 28(8): 1251-1265.

33. Biotronics3D. *Biotronics3D, Analyze-Collaborate-Discover. Available via https://www.biotronics3d.com/public/. Accessed in September 2018.* 2018.

34. Kohli MD, Summers RM, and Geis JR.Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session. Journal of Digital Imaging 2017; 30(4): 392-399.

35. XNAT. *XNAT, the most widely-used informatics platform for imaging research. Available via https://www.xnat.org/. Accessed in September 2018.* 2018.

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

Figure 2. Different variants of a T2-weighted whole body 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 and (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 XXXXXX. 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.

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).

Figure 5. Block diagram of the XXXXXX data preparation pipeline.

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.

Figure 7. Two-stage lesion detection process, employed in XXXXXX Phase 2. During stage one, the normal organs/bones are identified, based on Phase 1 training. During

stage two, lesion detection takes place. Stage two can be modular, with each module algorithm training depending on anatomical position.

Figure 8. Biotronics3D view of the whole body volumes from different modalities and the algorithm output, fused with the diffusion-weighted image from a colon lesion.