Development and evaluation of machine-learning methods in whole-body magnetic resonance imaging with diffusion weighted imaging for staging of patients with cancer: the MALIBO diagnostic test accuracy study

Andrea Rockall,[1,7*] Xingfeng Li,[1] Nicholas Johnson,[2] Ioannis Lavdas,[1] Shalini Santhakumaran,[2,3] A Toby Prevost,[2] Dow-Mu Koh,[4] Shonit Punwani,[5] Vicky Goh,[6] Nishat Bharwani,[1,7] Amandeep Sandhu,[7] Harbir Sidhu,[5,8] Andrew Plumb,[5] James Burn,[7] Aisling Fagan,[7] Alf Oliver,[5] Georg J Wengert,[1,10] Daniel Rueckert,[9] Eric Aboagye,[1] Stuart A Taylor,[5,8] Ben Glocker[9] and The MALIBO Investigators

[1]Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK
[2]Nightingale-Saunders Clinical Trials and Epidemiology Unit, King's College London, London, UK
[3]King's Cancer Prevention Group, School of Cancer and Pharmaceutical Sciences, King's College, London, UK
[4]Royal Marsden Hospital and The Institute of Cancer Research, Sutton, UK
[5]Centre for Medical Imaging, University College London, London, UK
[6]Cancer Imaging, School of Biomedical Engineering and Imaging Sciences, King's College London and Department of Radiology, Guy's and St Thomas' Hospitals NHS Foundation Trust, London, UK
[7]Imaging Department, Imperial College Healthcare NHS Trust, London, UK
[8]Department of Radiology, University College London Hospital, London, UK
[9]Faculty of Engineering, Department of Computing, Imperial College London, London, UK
[10]Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Vienna General Hospital, Vienna, Austria

[*]Corresponding author a.rockall@imperial.ac.uk

Published October 2024
DOI: 10.3310/KPWQ4208

Scientific summary

Development and evaluation of machine-learning methods in whole-body magnetic resonance imaging with diffusion weighted imaging for staging of patients with cancer: the MALIBO diagnostic test accuracy study

Efficacy and Mechanism Evaluation 2024; Vol. 11: No. 15
DOI: 10.3310/KPWQ4208

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

Whole-body magnetic resonance imaging (WB-MRI) has been developed in the last decade and it has been proposed as an alternative to multimodality cancer staging pathways. The STREAMLINE study has demonstrated that staging of lung and colon cancer via WB-MRI was of similar accuracy to standard care staging pathways but resulted in fewer tests being required, and a reduction in staging time and cost. However, this technique has not widely translated into clinical practice, being limited to a few expert centres. A barrier to translation may be the complexity of interpretation by inexperienced readers. A machine-learning (ML) algorithm for automated detection of cancer lesions, to assist radiologists, may allow clinical translation of WB-MRI for the benefit of patient care.

Objective

The objective of Phases 1 and 2 was to develop a ML method for automatic detection of cancer lesions on WB-MRI.

The primary objective of Phase 3 was to compare the diagnostic test accuracy of WB-MRI in patients being staged for cancer with and without ML support, when read by independent, experienced readers. The reference standard was the consensus reference panel from the STREAMLINE study, which had recorded the sites of disease in the STREAMLINE patients using all available clinical information for a 12-month follow-up from time of enrolment.

The planned secondary objectives of this study were:

1. to compare the reading time of WB-MRI scans, with and without ML support;
2. to determine the interobserver variability of WB-MRI diagnosis by different radiologists, with and without ML support;
3. to evaluate the diagnostic accuracy of WB-MRI read by non-experienced readers, with and without ML support;
4. to evaluate different combinations of acquired MRI sequences; all the above with and without ML support (not achieved);
5. to determine any difference in costs related to radiology reading time by means of a simple cost-effectiveness analysis (not achieved).

Design

This was an observational study (study limited to working with data), using different patient cohorts, and methodologies, being evaluated in series during three consecutive phases.

Phase 1: Previously acquired WB-MRI scans in 51 healthy volunteers were stitched into imaging volumes. The normal organs and skeleton were segmented by a trained radiologist. We compared and tested several state-of-the-art medical image segmentation methods to train an algorithm to automatically detect and segment the organs. This included a multi-atlas (MA) approach, classification forests (CFs) and convolutional neural networks (CNNs) methods. For the CFs method, we used 50 trees with a maximum tree depth of 30 in the segmentation. The stopping criterion for growing trees was if either the objective function (information gain) could not be further improved or the number of training samples in a leaf fell below a threshold of four samples. For the CNNs method, we employed a dual pathway (2 resolutions), 11-layer deep CNN, where the last 2 layers correspond to fully connected layers, which combine the features extracted on the 2 resolution pathways. We adopted 50–70 feature maps (different kernels) for each layer. The network architecture was fully convolutional and there were no max-pooling layers, which we found to increase segmentation accuracy. The CNN architecture was a balance between model capacity, training efficiency, and memory demands. The third algorithm was based on a MA label propagation approach. MA segmentation uses a set of atlases (images with corresponding segmentations) that represent the interparticipant variability of the anatomy to be segmented. Each atlas was registered to the new image to be segmented using a deformable image registration. The MA approach accounts for anatomical shape variability and is more robust than single-atlas propagation methods, in that at any errors associated with propagation were averaged out when combining multiple atlases. The approach employed here makes use of efficient 3D–3D intensity-based image registration with free-form deformations as the transformation model and correlation coefficient as the similarity measure. Majority voting is used to derive the final tissue label at each voxel.

Phase 2: Available WB-MRI scans from the STREAMLINE study (*n* = 438) were allocated by the study statistician to Phase 2 (model training) and clinical validation (hold-out set of 193 cases). The visible lesions in cases allocated to Phase 2 were segmented by trained radiologists, using the defined consensus reference standard for the sites of disease. Using 226 evaluable and annotated cases, we trained a model for lesion detection and localisation rather than attempting accurate automated segmentation. We found that lesion detection in WB-MRI was suboptimal with the CNN alone. An optimal ‘two-stage’ ML method was developed and tested. In the first stage of the ‘two-stage’ method, the information from Phase 1 was used to identify the position of organs and bones and in stage two, the lesions were detected. Stage two could be modular with respect to the anatomical location in which the suspected lesion could be found. The architecture and configuration of the used CNN were modified to achieve optimal performance.

Phase 3: The final two-stage ML algorithm from Phase 2 was applied to the 193 cases that were held out for Phase 3, generating probability heatmap volumes. WB-MRI scans were reported by 25 independent radiologists (18 radiologists experienced in reading WB-MRI and 7 radiologists inexperienced in WB-MRI). All radiologists took part in the 1st and 2nd round reads which incorporated inter-rater reads, and 8 participated in the 3rd round read for intrarater reads. All reads were undertaken in an NHS radiology reporting room, although using a separate cloud-based picture archiving and communication system (PACS). Based on experience level, 10–16 cases were randomly allocated by the study statistician to the reading lists for inexperienced or experienced radiologists. Each radiologist read both lung and colon cases with and without ML as a paired cohort. There were at least 4 weeks interval between each reading round for individual radiologists. Cases allocated to round 1 would be a random mixture of cases with and without ML support and then in round 2 the cases would be reversed such that each case was read once with and once without ML support. The number of cases with and without ML output was balanced in each reading round. A scribe recorded reading time.

Results

Phase 1 results

We found that CNNs outperformed CFs and the MA algorithm when T2w volumes were used as input to the algorithms and when using pooled overlap-evaluation metrics [Dice similarity coefficient (DSC), recall (RE), precision (PR)] to assess the accuracy of segmentation. When the performance of the algorithms was assessed, with pooled surface distance metrics [average surface distance (ASD), root-mean-square surface distance (RMSSD), and Hausdorff distance (HD)], it was the MA algorithm that performed best. Single misinterpreted voxels in CFs and CNNs can greatly elevate ASD, RMSSD, and HD; these metrics are particularly sensitive to outliers. We then assessed the pooled metrics performance of CFs and CNNs when using all imaging combinations [T2w + T1w + diffusion-weighted imaging (DWI)] as input, arguing that maximisation of training information to the algorithms might improve the performance of segmentation. We found that the performance of CFs was improved, however not significantly, when using all imaging combinations as input for training. The opposite was observed for CNNs.

The findings for the pooled metrics analysis, described above, were corroborated by a 'per-organ' quantitative analysis of the commonly used DSC, to assess the performance of our segmentation algorithms. This analysis confirmed that for all individual anatomical structures (except for the bladder), the algorithm that returned the greatest DSC was CNNs with T2w images only used as input. As our morphological T2w and T1w scans were acquired using breath-holds and the DWI sequence was acquired with free breathing, we found that there was significant displacement between soft tissues in anatomical areas adjacent to the diaphragm between these types of scans. As the employed affine registration method could not fully compensate for nonlinear motions caused by breathing, we assumed that misregistration could be the reason why the performance of CNNs, despite performing better than the other two algorithms when using T2w volumes as input only, was degraded when using all imaging combinations as input for training. A more robust, nonlinear registration method could improve the accuracy of CNNs and further improve the performance of CFs.

The performance of our methods cannot be directly compared to similar methods in the literature because there is no previous work describing automatic, simultaneous segmentation of healthy organs and bones in multiparametric WB-MRI. We believe, however, that our methods may compare favourably to other ML methods for detection and segmentation in medical imaging because our classifiers are inherently multilabel and effective training was achieved when using a relatively small number of data sets, something that is very important in the clinical setting. However, we still need to address the performance limitations of our algorithms when segmenting organs with big variability in appearance (e.g. the gallbladder or the pancreas).

Phase 2 results

We tested different ML methods for lesion detection. We found that using CNNs was not optimal for lesion detection on WB-MRI. This may have been due to the small fraction of lesion volume occupying the scanned space, when compared to the WB volume. It may also have been due to the complexity of intensities in background tissue and the lesion with weak boundaries causing challenges for the CNNs. We, therefore, adapted our process to become an optimal two-stage process, with an initial stage for detection of organs as per the Phase 1 technique followed by a CNN to detect lesions. Stage two could be modular with respect to the anatomical location where the suspected lesion can be found. The architecture and configuration of the used CNN could be modified to achieve optimal performance for lesion detection.

Phase 3 results

All radiology reads were completed between November 2019 and March 2020. Among the 193 cases allocated to Phase 3, 188 WB-MRI scans were evaluable (117 colon and 71 lung cancer) and 50/188 cases had metastases.

Per-patient specificity for detection of metastases within experienced readers was 86.2% (WB-MRI + ML) and 87.7% [WB-MRI + standard deviation (SD)], [difference −1.5%, 95% confidence interval (CI) −6.4%, 3.5%; *p* = 0.387]. Per-patient sensitivity produced results of 66.0% (WB-MRI + ML) and 70.0% (WB-MRI + SD) (difference −4.0%, 95% CI −13.5% to 5.5%; *p* = 0.344). For inexperienced readers (53 reads, 15 with metastases), per-patient specificity was 76.3% in both groups with sensitivities of 73.3% (WB-MRI + ML) and 60.0% (WB-MRI + SD). Per-site specificity remained high within all sites; above 95% (experienced) or 90% (inexperienced). Per-site sensitivity was highly variable due to the low number of lesions in each site, hampering interpretation.

Reading time was lowered under ML by 6.2% (95% CI −22.8% to 10.0%). Read time was primarily influenced by read round with round 2 read times reduced by 32% (95% CI 20.8% to 42.8%) overall with subsequent regression analysis showing a significant effect (p = 0.0281) by using ML in round 2 estimated as 286 seconds (or 11%) quicker.

Interobserver variance for experienced readers suggests moderate agreement, Cohen's κ = 0.64, 95% CI 0.47 to 0.81 (WB-MRI + ML) and Cohen's κ = 0.66, 95% CI 0.47 to 0.81 (WB-MRI + SD).

Conclusion

Phase 1

In Phase 1, we developed and evaluated three state-of-the-art algorithms that automatically segment healthy organs and bones in WB-MRI with accuracy comparable to the one achieved manually by clinical experts, using relatively sparse training data. An algorithm based on CNNs and trained using T2w-only images as input performs favourably when compared to CFs or a MA algorithm, trained with either T2w-only images or a combination of imaging inputs (T2w + T1w + DWI). Using multimodal MRI data as input for training did not improve the segmentation performance in this work, but it is anticipated to improve the segmentation performance if more effective WB registration between the various imaging modalities can be performed. This investigation was the first step towards developing robust algorithms for the automatic detection and segmentation of benign and malignant lesions in WB-MRI scans for staging of cancer patients.

Phase 2

There were many challenges in training the algorithm for lesion detection, with a very heterogeneous data set, acquired at 16 sites on different machines. The low number of metastatic lesions were scattered through many anatomic sites. Identification of lesions of relatively small volume in relation to the large volume of the WB-MRI required a two-stage approach for lesion detection, with initial organ identification followed by lesion detection.

Phase 3

We undertook a robust diagnostic test accuracy study comparing WB-MRI with ML support (index text) with standard WB-MRI, without ML support (comparator test), with paired reads separated by a wash-out period.

Although there was no clear statistical difference with or without ML support in terms of diagnostic test accuracy, either in experienced or inexperienced hands, we found that ML reads were likely to be a little shorter in reading time.

Implications for health care

Machine learning support for image interpretation is a rapidly expanding area of research. With continually increasing demand for diagnostic imaging and a radiology workforce crisis in the NHS, as well as globally, ML techniques may offer support to allow complex imaging modalities, such as WB-MRI, to be ready for translation into clinical care for the benefit of patients. Phase 1 demonstrated that with relatively sparse data, we could develop very successful organ segmentation tools in highly complex data sets and this has many potential future roles. In this study, we used this as a first step prior to lesion detection and this generic method could be applied to a wider range of disease areas, including other tumour types.

We found that ML support slightly shorted the reading time, although with no improvement in diagnostic accuracy. However, we have shown that lesion detection using ML is achievable on heterogeneous and complex multiparametric WB-MRI scans. With further model training, we believe that ML support is feasible in the future, with the potential to improve translation of WB-MRI more widely. In addition, many of the techniques that have been developed in the course of the study have the potential to be applied to other areas of diagnostic imaging.

Recommendations for future research

Future research should investigate:

1. Further improvement in lesion detection accuracy, with evaluation of failure cases for improved model training.
2. Evaluate lesion detection technique on other cancers, notably breast, prostate and myeloma.
3. Further developments in understanding the marrow composition in healthy aging and in metastatic cancer.
4. Developing further understanding and methods to overcome registration challenges between breath-hold and non-breath-hold MRI sequences in order to improve ML tissue characterisation.

Study registration

This study is registered as ISRCTN23068310.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Efficacy and Mechanism Evaluation (EME) programme (NIHR award ref: 13/122/01) and is published in full in *Efficacy and Mechanism Evaluation*; Vol. 11, No. 15. See the NIHR Funding and Awards website for further award information.

Efficacy and Mechanism Evaluation

ISSN 2050-4373 (Online)

A list of Journals Library editors can be found on the NIHR Journals Library website

Efficacy and Mechanism Evaluation (EME) was launched in 2014 and is indexed by Europe PMC, DOAJ, Ulrichsweb™ (ProQuest LLC, Ann Arbor, MI, USA) and NCBI Bookshelf.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full EME archive is freely available to view online at www.journalslibrary.nihr.ac.uk/eme.

Criteria for inclusion in the *Efficacy and Mechanism Evaluation* journal
Manuscripts are published in *Efficacy and Mechanism Evaluation* (EME) if (1) they have resulted from work for the EME programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

EME programme

The Efficacy and Mechanism Evaluation (EME) programme funds ambitious studies evaluating interventions that have the potential to make a step-change in the promotion of health, treatment of disease and improvement of rehabilitation or long-term care. Within these studies, EME supports research to improve the understanding of the mechanisms of both diseases and treatments.

The programme supports translational research into a wide range of new or repurposed interventions. These may include diagnostic or prognostic tests and decision-making tools, therapeutics or psychological treatments, medical devices, and public health initiatives delivered in the NHS.

The EME programme supports clinical trials and studies with other robust designs, which test the efficacy of interventions, and which may use clinical or well-validated surrogate outcomes. It only supports studies in humans and where there is adequate proof of concept. The programme encourages hypothesis-driven mechanistic studies, integrated within the efficacy study, that explore the mechanisms of action of the intervention or the disease, the cause of differing responses, or improve the understanding of adverse effects. It funds similar mechanistic studies linked to studies funded by any NIHR programme.

The EME programme is funded by the Medical Research Council (MRC) and the National Institute for Health and Care Research (NIHR), with contributions from the Chief Scientist Office (CSO) in Scotland and National Institute for Social Care and Health Research (NISCHR) in Wales and the Health and Social Care Research and Development (HSC R&D), Public Health Agency in Northern Ireland.

This article

The research reported in this issue of the journal was funded by the EME programme as award number 13/122/01. The contractual start date was in February 2015. The draft manuscript began editorial review in October 2020 and was accepted for publication in July 2023. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The EME editors and production house have tried to ensure the accuracy of the authors' manuscript and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this article.

This article presents independent research. The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the MRC, the EME programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the EME programme or the Department of Health and Social Care.

This article was published based on current knowledge at the time and date of publication. NIHR is committed to being inclusive and will continually monitor best practice and guidance in relation to terminology and language to ensure that we remain relevant to our stakeholders.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Newgen Digitalworks Pvt Ltd, Chennai, India (www.newgen.co).