European Radiology
https://doi.org/10.1007/s00330-018-5445-8

MAGNETIC RESONANCE

Whole-body MRI for staging and interim response monitoring in paediatric and adolescent Hodgkin's lymphoma: a comparison with multi-modality reference standard including ^{18}F-FDG-PET-CT

Arash Latifoltojar[1] · Shonit Punwani[1,2] · Andre Lopes[3] · Paul D. Humphries[1,2] · Maria Klusmann[2] · Leon Jonathan Menezes[4] · Stephen Daw[5] · Ananth Shankar[5] · Deena Neriman[4] · Heather Fitzke[1] · Laura Clifton-Hadley[3] · Paul Smith[3] · Stuart A. Taylor[1,2]

Received: 7 February 2018 / Revised: 16 March 2018 / Accepted: 22 March 2018

Abstract

Objectives To prospectively investigate concordance between whole-body MRI (WB-MRI) and a composite reference standard for initial staging and interim response evaluation in paediatric and adolescent Hodgkin's lymphoma.

Methods Fifty patients (32 male, age range 6–19 years) underwent WB-MRI and standard investigations, including ^{18}F-FDG-PET-CT at diagnosis and following 2–3 chemotherapy cycles. Two radiologists in consensus interpreted WB-MRI using prespecified definitions of disease positivity. A third radiologist reviewed a subset of staging WB-MRIs (n = 38) separately to test for interobserver agreement. A multidisciplinary team derived a primary reference standard using all available imaging/clinical investigations. Subsequently, a second multidisciplinary panel rereviewed all imaging with long-term follow-up data to derive an enhanced reference standard. Interobserver agreement for WB-MRI reads was tested using kappa statistics. Concordance for correct classification of all disease sites, true positive rate (TPR), false positive rate (FPR) and kappa for staging/response agreement were calculated for WB-MRI.

Results There was discordance for full stage in 74% (95% CI 61.9–83.9%) and 44% (32.0–56.6%) of patients against the primary and enhanced reference standards, respectively. Against the enhanced reference standard, the WB-MRI TPR, FPR and kappa were 91%, 1% and 0.93 (0.90–0.96) for nodal disease and 79%, < 1% and 0.86 (0.77–0.95) for extra-nodal disease. WB-MRI response classification was correct in 25/38 evaluable patients (66%), underestimating response in 26% (kappa 0.30, 95% CI 0.04–0.57). There was a good agreement for nodal (kappa 0.78, 95% CI 0.73–0.84) and extra-nodal staging (kappa 0.60, 95% CI 0.41–0.78) between WB-MRI reads

Conclusions WB-MRI has reasonable accuracy for nodal and extra-nodal staging but is discordant with standard imaging in a substantial minority of patients, and tends to underestimate disease response.

Key Points

- *This prospective single-centre study showed discordance for full patient staging of 44% between WB-MRI and a multi-modality reference standard in paediatric and adolescent Hodgkin's lymphoma.*
- *WB-MRI underestimates interim disease response in paediatric and adolescent Hodgkin's lymphoma.*
- *WB-MRI shows promise in paediatric and adolescent Hodgkin's lymphoma but currently cannot replace conventional staging pathways including ^{18}F-FDG-PET-CT.*

Arash Latifoltojar and Stuart A. Taylor contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-018-5445-8) contains supplementary material, which is available to authorized users.

✉ Stuart A. Taylor
stuart.taylor1@nhs.net

[1] Centre for Medical Imaging, University College London, Charles Bell House, 2nd floor, 43-45 Foley Street, London W1W 7TS, UK

[2] Department of Radiology, University College London Hospitals, 235 Euston Road, London NW1 2BU, UK

[3] Cancer Research UK and UCL Cancer Trial Centre, University College London, 90 Tottenham Court Road, London W1T 4TJ, UK

[4] Institute of Nuclear Medicine, University College London and NIHR University College London Hospitals Biomedical Research Centre, 235 Euston Road, London NW1 2BU, UK

[5] Department of Paediatric Haemato-Oncology, University College London Hospitals, 235 Euston Road, London NW1 2BU, UK

Published online: 15 June 2018

Keywords Whole-body scan · Diffusion-weighted MRI · Tumour staging · Treatment · Hodgkin lymphoma

Abbreviations

ADC	Apparent diffusion coefficient
DWI	Diffusion-weighted imaging
FPR	False positive rate
HL	Hodgkin's lymphoma
IQR	Interquartile range
MDT	Multidisciplinary team
TPR	True positive rate
WB-MRI	Whole-body MRI

Introduction

Hodgkin's lymphoma (HL) is the most common adolescent lymphoma [1]. Positron emission tomography computed tomography (^{18}F-FDG-PET-CT) remains the first-line imaging technique [2], providing both structural and functional metabolic information to localise and characterise tumour burden. Furthermore, as a biomarker of glucose metabolism, uptake of the radiotracer ^{18}F-2-fluro-2-deoxy-D-glucose (^{18}F-FDG) provides a more accurate assessment of treatment response than simple structural evaluation [2–4]. ^{18}F-FDG-PET-CT, however, imparts a substantial dose of ionising radiation, which may be associated with increased risk of secondary malignancies [2, 5]. This is a concern in the paediatric age group given the increased sensitivity of tissues to radiation exposure, coupled with the significant improvement in long-term survival [6, 7].

Whole-body magnetic resonance imaging (WB-MRI) is an attractive alternative to ^{18}F-FDG-PET-CT as it does not impart ionising radiation and can provide high quality anatomical images through the body in less than 1 h [6, 8, 9]. Moreover, there is evidence suggesting that diffusion-weighted imaging (DWI) may act as a surrogate for the functional information provided by ^{18}F-FDG [10, 11]. DWI captures water movement within tissue and its functional parameter, the apparent diffusion coefficient (ADC), is a marker of tissue cellularity and related to glucose metabolism [12, 13].

There is increasing supportive literature for implementation of WB-MRI in lymphoma staging pathways [14–18], although such data remains relatively sparse in the paediatric population [19]. Extrapolation from adult studies may be flawed given the complexities of imaging patients with smaller body habitus, the challenges of prolonged WB-MRI protocols for younger patients, and potential differences in disease patterns and behaviours between paediatric and adult patients, and within lymphoma subtypes.

The purpose of this study was to investigate prospectively the concordance between WB-MRI and a composite reference standard based on clinical evaluation, histology and standard staging imaging including ^{18}F-FDG-PET-CT for initial staging and interim treatment response monitoring in paediatric and adolescent Hodgkin's lymphoma.

Material and methods

We conducted a prospective single-arm cohort study in a single tertiary referral centre, following ethical permission (Clinicaltrials.gov number: NCT01459224).

Consent for study investigations, including collection of anonymised patient data, was obtained from patients and or their parents/guardians according to the institutional and ethical committee guidelines.

Patient population

Consecutive patients were prospectively identified between December 2011 and August 2014 inclusive from the paediatric lymphoma service of University College London Hospital.

Inclusion criteria were age 5–20 years (inclusive), histological confirmation of HL or clinically suspected HL (classical HL and nodular lymphocyte predominant HL) undergoing staging investigations pending final biopsy confirmation, and patients/guardian consent. All patients were either recruited to the Euronet PHL-C1 or PHL-LP1 trials [20] or were due to undergo treatment using the chemotherapy regimens of these trials.

Exclusion criteria included previous diagnosis of HL without being disease free for 5 years, previous chemotherapy and or radiotherapy within the previous 2 years, pregnancy or breastfeeding and any known contraindication to MRI.

Summary of study conduct

All recruited patients underwent the standard staging investigations employed at the recruiting institution: (i) whole-body ^{18}F-FDG-PET-CT, (ii) anatomical WB-MRI sequences with single-phase post-contrast acquisition through the upper abdomen, (iii) abdominal ultrasound in cases of equivocal solid organ involvement and (iv) contrast-enhanced chest CT (CE chest CT) scan in case of equivocal lung involvement.

To provide a comprehensive "stand-alone" WB-MRI protocol, for the purposes of the current study, the WB-MRI protocol was extended to include whole-body DWI and dynamic contrast-enhanced (DCE) sequences though the liver/spleen and chest, as well as the standard basic anatomical sequences.

Thereafter, patients underwent interim ^{18}F-FDG-PET-CT (iPET-CT) within 14 days of completing the first two (Euronet PHL-C1) or three (Euronet LP1) cycles of chemotherapy for initial treatment response evaluation. Patients were invited to undergo a second WB-MRI (iWB-MRI) and were followed for a minimum 24 months post chemotherapy.

Imaging protocols

Full descriptions of the WB-MRI and standard imaging protocols are given in the Electronic Supplementary Material (ESM). WB-MRI sequence parameters are shown in Supplemental Table 1.

Staging imaging interpretation

A full description of WB-MRI and standard imaging interpretation is provided in the ESM. In brief, ^{18}F-FDG-PET-CT was interpreted by a nuclear medicine physician (LM with more than 10 years of experience) and basic anatomical WB-MRI sequences including single-phase post-contrast sequences through the upper abdomen (but excluding DWI and DCE sequences), abdominal ultrasound and chest CT images (when available) were evaluated by consultant paediatric radiologist (PH with 11 years of experience in WB-MR imaging). WB-MRI was interpreted by two radiologists (SAT and SP with 5 and 7 years' experience of WB-MRI) in consensus utilising all the available sequences (including the DWI and DCE images). The radiologists were blinded to the clinical history (other than the diagnosis of lymphoma) and all other investigations. A third blinded radiologist (MK with 3 years' experience of WB-MRI) interpreted a subset of 38 WB-MRI data sets to test interobserver agreement with the primary consensus read.

The disease status for 18 nodal and 14 extra-nodal sites was evaluated, as well as the final Ann Arbor stage derived using predefined definitions based on size, ^{18}F-FDG uptake and ADC (based on previous pilot data [21]; see ESM).

Interim treatment response evaluation

Interim ^{18}F-FDG-PET-CT (iPET-CT) and WB-MRI (iWB-MRI) were interpreted by the same individuals who read the initial staging investigations.

For WB-MRI, the ADC criteria for nodal response were based on those derived from previous work investigating ADC changes in responsive and non-responsive nodal disease [21] (Table 1). Extra-nodal response was evaluated by qualitative assessment of iWB-MRI classifying response into four categories: (a) locally undetectable (complete response), (b) locally detectable but reduction in size or number of deposits (partial response), (c) locally unchanged (no change in the number or size of deposits) and (d) locally progressive (increase in size or number of deposits).

A full description of interim response evaluation is provided in ESM.

Primary and enhanced reference standards

A full description is provided in ESM. In brief, the primary reference standard was assigned by a multidisciplinary team (MDT) on the basis of their assessment of all standard imaging tests together with all clinical information, including available histology.

Given the potential limitations of standard imaging in staging HL, which may weaken the primary reference standard, a retrospective enhanced reference standard was also produced by a central expert panel comprising two radiologists, two nuclear medicine physicians and two paediatric haemato-oncologists. The central panel reviewed all staging, interim and end of treatment scans as well as follow-up imaging and clinical outcomes up to 24 months post chemotherapy. The panel corrected simple labelling (boundary) discrepancies that were due to differences in disease site description between tests, and thereafter any perceptual or technical failures in the primary reference standard (Fig. 1). WB-MRI perceptual errors were also noted.

Data analysis and study power

The primary endpoint was based on achieving full (100%) concordance between WB-MRI and the primary reference standard in terms of correct disease classification for each and every anatomical site (i.e. the 18 nodal and 14 extra-nodal sites). A binary classification of each disease status as either negative or positive/equivocal was made as part of the reference standard.

See ESM for the power calculation of the study sample size.

The primary endpoint was summarised in terms of frequency and percentage of patients who had a concordance below 100% for all disease sites combined, and separately for nodal and extra-nodal sites. The median and interquartile range (IQR) discordance rate for each patient was also calculated.

The true positive rate (TPR) (sensitivity) and false positive rate (FPR) of WB-MRI were calculated for nodal and extra-nodal disease sites, along with the kappa statistic. Agreement for Ann Arbor staging and classification of interim treatment response evaluation (for positive/equivocal disease sites that were concordant at initial staging) were summarised in terms of frequency, percentages and kappa.

Table 1 Nodal disease response assessment

Disease response	Definition for standard imaging tests	Definition for WB-MRI scan
Complete response (CR)	Residual tumour volume is < 25% of initial staging or ≤ 2 ml and PET negative	Residual tumour volume is < 25% of initial staging or ≤ 2 ml and ADC > 30% change compared to pretreatment value
Partial response (inadequate) (PRi)	Residual tumour volume < 75% but ≥ 50% of initial staging, or disease is PET avid (focal or diffuse uptake exceeding that of mediastinal blood pool in a location incompatible with normal anatomy or physiology)	Residual tumour volume < 75% but ≥ 50% of initial staging, or fractional change in ADC < 70% compared to pretreatment value
Partial response (adequate) (PRa)	Residual tumour volume ≤ 50% but ≥ 25% of initial staging, and all disease is PET negative (avidity not exceeding that of mediastinal blood pool)	Residual tumour volume ≤ 50% but ≥ 25% of initial staging, and fractional change in ADC ≥ 70% compared to pretreatment value
No change (NC)	Residual tumour volume ≥ 75% but < 125% of initial staging	Residual tumour volume ≥ 75% but < 125% of initial staging
Progression (PRO)	Residual tumour volume ≥ 125%	Residual tumour volume ≥ 125%

WB-MRI whole-body MRI, *PET* positron emission tomography, *ADC* apparent diffusion coefficient

Sensitivity analysis were performed using the outcomes from the central review process, and the enhanced reference standard.

Specifically, the agreement analyses for staging WB-MRI were repeated:

- After correcting for anatomical boundary labelling description discrepancies only
- Against the enhanced reference standard (including correction of boundary labelling description discrepancies)
- Against the enhanced reference standard after removal of WB-MRI perceptual errors

Ann Arbor staging agreement was also assessed after integrating the results of enhanced reference standard and after WB-MRI correction for perceptual errors.

Interobserver agreement between consensus WB-MRI read and the third radiologist was tested using kappa statistics.

Statistical analysis was performed using the Stata software package (Version 14. Stata Corporation LP, College Station, Texas).

Results

Patient characteristics

Fifty-eight patients were recruited (M/F 39:19, median age 16, range 5–19 years). The study flowchart is presented in Fig. 2. Eight patients were excluded. The demographics, disease subtype and treatment regimen of the final 50 patient study cohort are shown in Table 2. Staging WB-MRI was performed within a median 2 days (range 0–20 days) of ^{18}F-FDG-PET-CT without any complication, and before treatment in all patients.

Central review and enhanced reference standard

Across the cohort there were 1527 disease sites [875 nodal (850 predefined sites and 25 "other" sites and 652 extra-nodal sites (650 predefined sites and 2 "other" sites)] evaluated by both WB-MRI and standard imaging.

The central review identified and resolved 44 anatomical boundary labelling description discrepancies. There were 10 nodal and 4 extra-nodal perceptual errors in the primary reference standard, together with 1 technical error.

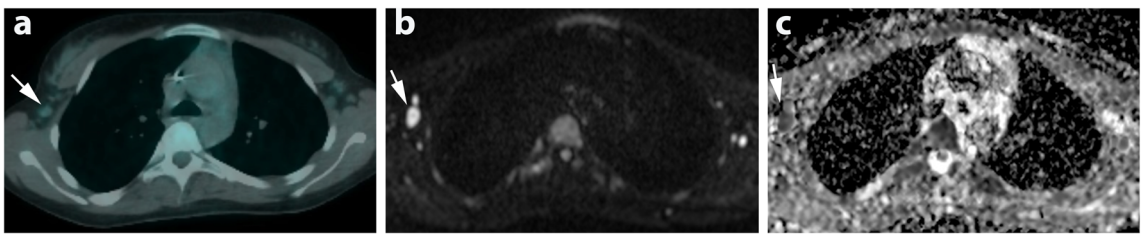

Fig. 1 Example of ^{18}F-FDG-PET-CT perceptual error. Right axillary nodal station was called negative on (**a**) ^{18}F-FDG-PET-CT and positive on WB-MRI. b_{500} diffusion-weighted MRI (**b**) and apparent diffusion coefficient map (**c**) showing restricted diffusion (ADC 1.0 × 10^{-3} mm^2/s). On retrospective evaluation of nodal station, with full follow-up data available, the expert panel judged the right axillary node (arrows) to be a positive nodal site based on ^{18}F-FDG uptake, and thus a perceptual error on initial ^{18}F-FDG-PET-CT interpretation

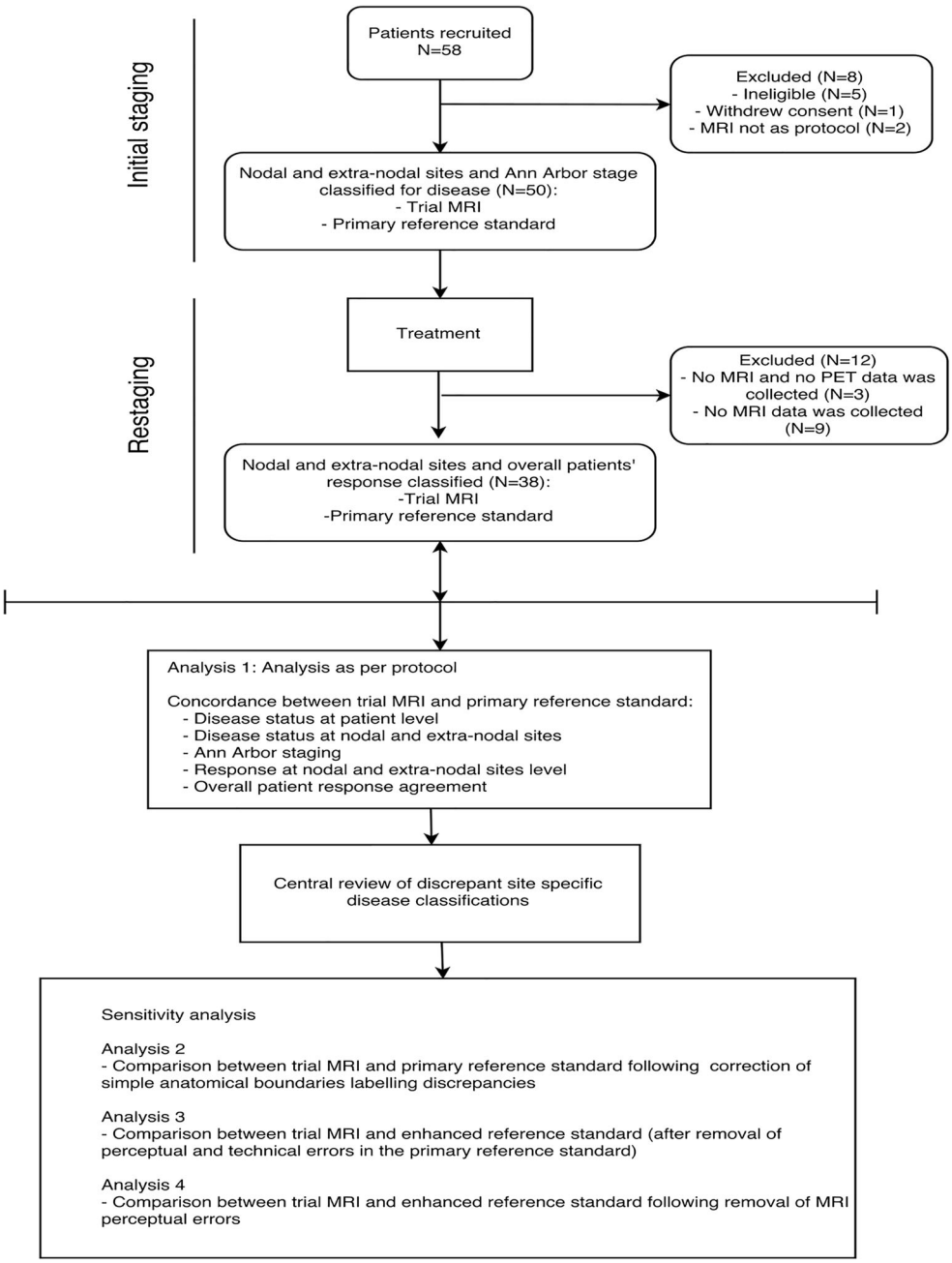

Fig. 2 Study flowchart

There were 20 WB-MRI perceptual errors.

Initial staging agreement: per patient

Per patient concordance rate for each analysis is shown in Table 3 and Fig. 3.

After correcting for labelling discrepancies, the discordance rate was 44% (90% CI exact 32.0–56.6%) for nodal sites and 28% (90% CI exact 17.8–40.3%) for extra-nodal sites.

Against the enhanced reference standard, the equivalent discordance rates fell to 44% (90% CI exact 32.0–56.6%) for all sites, 34% (90% CI exact 23.0–46.5%) for nodal sites and 18% (90% CI exact 9.7–29.3%) for extra-nodal sites. After removal of WB-MRI perceptual errors, the discordance rates for all, nodal and extra-nodal sites were 18% (90% CI exact 9.7–29.3%), 16% (90% CI exact 8.2–27.0%) and 4% (90% CI exact 0.7–12.1%), respectively.

Initial staging agreement: disease site

Absolute agreement rate, TPR, FPR and Cohen's kappa statistic for nodal and extra-nodal disease sites for each analysis are shown in Table 4.

Table 2 Patients' cohort demographics

Baseline characteristics	*N* (%) *n* = 50
Age (years)	
Median (range)	16 (6–19)
Sex	
Female	18 (36%)
Male	32 (64%)
Hodgkin's lymphoma subtype	
Classical	42 (84%)
Nodular lymphocyte predominant	8 (16%)
Chemotherapy	
OEPA	9 (18%)
OEPA/COPDAC	32 (64%)
CVP	7 (14%)
DHAP/OEPA/COPDAC	1 (2%)
Others[a]	1 (2%)

OEPA vincristine, etoposide, prednisolone, doxorubicin; *COPDAC* cyclophosphamide, vincristine, prednisolone, dacarbazine; *CVP* cyclophosphamide, vincristine, prednisolone; *DHAP* dexamethasone, cytarabine, cisplatin

[a] One patient with stage I and single lymph node involvement that was excised for histopathology did not received any treatment

Against the enhanced reference standard, the WB-MRI TPR, FPR and kappa agreement were 91%, 1% and 0.93 (95% CI 0.90–0.96) for nodal disease and 79%, < 1% and 0.86 (95% CI 0.77–0.95) for extra-nodal disease.

Following removal of WB-MRI perceptual errors, the TPR, FPR and kappa agreement were 97%, < 1% and 0.97 (95% CI 0.95–0.99) for nodal and 95%, 0% and 0.97 (95% CI 0.93–1.00) for extra-nodal assessment compared to enhanced reference standard. There were seven WB-MRI false negative nodal sites due to technical failures (i.e. not visible in retrospect), two false positive nodal sites and two false negative extra-nodal sites (Supplemental Table 4).

Ann Arbor staging agreement

Based on enhanced reference standard, there were 2, 26, 5, 14 and 3 patients with Ann Arbor stage 1, 2, 3, 4 and 4E, respectively.

Agreement between WB-MRI and the primary reference standard was substantial (kappa 0.66, 95% CI 0.50–0.83) with staging concordant in 39/50 (78%) patients (Supplemental Table 5).

Prior to removal of WB-MRI perceptual errors, agreement between WB-MRI and the enhanced reference was substantial (kappa 0.72, 95% CI 0.56–0.88) with concordance in 41/50 (82%) patients. After removal of the WB-MRI perceptual errors concordance was achieved in 48/50 patients (96%), (kappa 0.94, 95% CI 0.85–1.00). Two patients were under-staged as a result of technical failure of WB-MRI compared to enhanced reference (Fig. 4 and Supplemental Table 5).

Interim treatment response agreement

Thirty-eight of the 50 patients were evaluable for interim treatment response analysis (Fig. 2). iWB-MRI scans were acquired within a median 1 day (range 0–7 days) of iPET scans.

On a per patient basis, iWB-MRI agreed with the primary reference standard response classification in 25/38 patients (66%, 6 PR and 19 CR), underestimating response in 10 (26%) patients and overestimating response in 3 (8%) patients (kappa 0.30, 95% CI 0.04–0.57) (Table 5).

There were 143 nodal and 26 extra-nodal positive concordant sites evaluable for interim treatment assessment.

iWB-MRI agreed with the primary reference standard response classification in 126/143 (88%) nodal sites, underestimating response in 3 (2%) sites and overestimating response in 14 (10%) (Supplemental Table 6).

iWB-MRI agreed with primary reference standard response classification in 17/26 (66%) of extra-nodal sites. In the remaining 9 (34%) sites, WB-MRI underestimated response (Fig. 5). Specifically, WB-MRI underestimated bone marrow response in four patients (three with reduced but persistent detectable disease and one with unchanged disease), and for spleen and lung in two and three patients respectively (all five with reduced but persistent disease on WB-MRI). All nine sites showed complete response on primary reference standard.

WB-MRI interobserver agreement

There was a good agreement between the consensus WB-MRI and the 3rd radiologist reads for nodal (kappa 0.78, 95% CI 0.73–0.84), extra-nodal staging (kappa 0.60, 95% CI 0.41–0.78) and Ann Arbor staging (kappa 0.62, 95% CI 0.32–0.73).

Discussion

In the current study we compared WB-MRI with a combined multi-modality reference standard based mainly on standard imaging (notably ^{18}F-FDG-PET-CT) but including clinical and histological data for staging and interim treatment response monitoring in paediatric HL.

Overall, we found that WB-MRI has reasonable accuracy for nodal and extra-nodal staging but did not achieve full concordance for all disease sites in a substantial minority of patients, and tends to underestimate disease response.

Our findings of intrinsically high sensitivity and specificity for nodal and extra-nodal staging confirm the data of Littooij et al. who performed a similar staging study in a cohort of 33 paediatric patients with a range of lymphoma phenotypes [19],

Table 3 Per patient concordance rate for each analysis

Concordance rate	Overall (Nodal and extra-nodal sites) $N = 50$ n (%)	Nodal sites $N = 50$ n (%)	Extra-nodal sites $N = 50$ n (%)
Analysis 1[a]			
≤ 60%	–	–	–
> 60% to ≤ 80%	1 (2%)	5 (10%)	–
> 80% to ≤ 90%	8 (16%)	15 (30%)	–
> 90% to < 100%	28 (56%)	16 (32%)	14 (28%)
100%	13 (26%)	14 (28%)	36 (72%)
Analysis 2[b]			
≤ 60%	–	–	–
> 60% to ≤ 80%	–	1 (2%)	–
> 80% to ≤ 90%	4 (8%)	5 (10%)	–
> 90% to < 100%	26 (52%)	16 (32%)	14 (28%)
100%	20 (40%)	28 (56%)	36 (72%)
Sensitivity analysis 1[c]			
≤ 60%	–	–	–
> 60% to ≤ 80%	–	–	–
> 80% to ≤ 90%	1 (2%)	4 (8%)	–
> 90% to < 100%	21 (42%)	13 (26%)	9 (18%)
100%	28 (56%)	33 (66%)	41 (82%)
Sensitivity analysis 2[d]			
≤ 60%	–	–	–
> 60% to ≤ 80%	–	–	–
> 80% to ≤ 90%	–	1 (2%)	–
> 90% to < 100%	9 (18%)	7 (14%)	2 (4%)
100%	41 (82%)	42 (84%)	48 (96%)

[a] Comparison between WB-MRI and primary reference standard before correction of simple anatomical boundaries labelling discrepancies

[b] Comparison between WB-MRI and primary reference standard following correction of simple anatomical boundaries labelling discrepancies

[c] Comparison between WB-MRI and enhanced reference standard (after removal of perceptual and technical errors in the primary reference standard)

[d] Comparison between WB-MRI and enhanced reference standard following removal of WB-MRI perceptual errors

and mirror those of Mayerhoefer et al. [17] who studied a cohort of 140 adult patients. In line with previous work [14], we utilised a rigorous consensus review process taking into consideration all long-term imaging and clinical follow-up to create an enhanced reference standard, thereby correcting deficiencies in standard staging pathways, and providing a more realistic evaluation of the accuracy of WB-MRI. Against this enhanced reference, WB-MRI, sensitivity for extra-nodal disease was still modest at 79%. We also retrospectively corrected WB-MRI perceptual errors to indicate the theoretical "best" technical performance of WB-MRI, which increased nodal sensitivity to 97% and extra-nodal disease sensitivity to 95%. Clearly perceptual errors are unavoidable so such corrected data will overestimate the performance of WB-MRI, but particular emphasis should be made on detecting extra-nodal disease during radiologist training.

Our primary analysis, and one rarely performed in the literature, is how often WB-MRI achieved full concordance with standard imaging for each and every disease site in an individual patient. Such data is clinically highly relevant, as patients with early unfavourable response will often undergo targeted radiotherapy to individual involved nodal stations following chemotherapy [22]. Against the enhanced reference standard, full concordance for nodal disease was achieved in 66% of patients, which increased to 84% after removal of WB-MRI perceptual errors. Although such data is encouraging, there is a substantial minority of patients with discordant findings to standard staging, which may have treatment implications. Our data suggests that using ADC as a surrogate for ^{18}F-FDG uptake, although promising [12, 21], is currently insufficient. It is clear there is overlap in ADC between malignant lymph nodes and normal/reactive lymph nodes and the

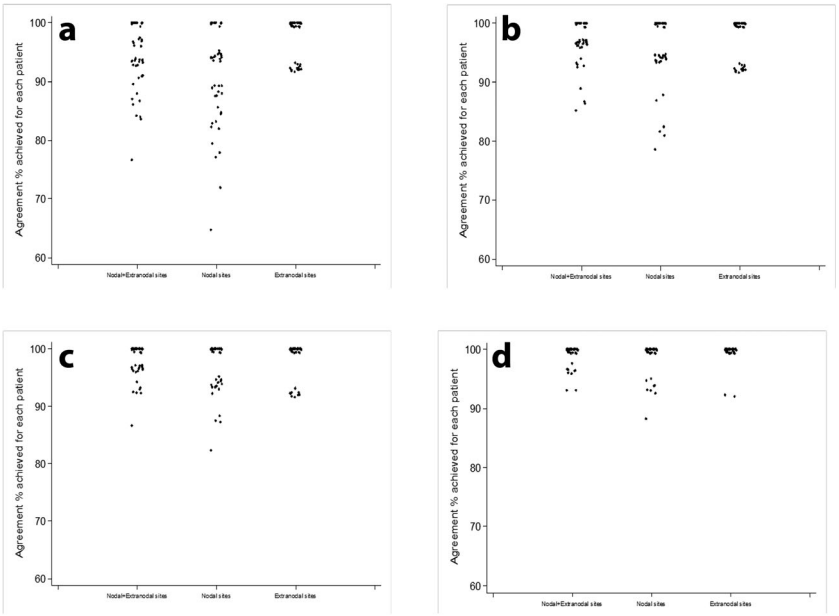

Fig. 3 Per patient concordance rate. Concordance rate for nodal, extra-nodal and combined nodal/extra-nodal sites between **a** WB-MRI and primary reference standard prior to the removal of simple boundary classification labelling discrepancies and **b** following the removal of simple boundary classification labelling discrepancies. **c** WB-MRI and the enhanced reference standard (following removal of ^{18}F-FDG-PET-CT perceptual and technical errors) and **d** WB-MRI and the enhanced reference standard following removal of WB-MRI perceptual errors. Median and interquartile range (IQR) are presented for each analysis tier

optimal ADC cut-off remains unclear, and requires further investigation [23].

Although access to new ^{18}F-FDG-PET-MR technology is currently very limited, this platform my ultimately prove to be the investigation of choice and prospective studies are currently underway [24].

The accuracy of iWB-MRI for interim treatment response assessment is under investigation, but far from proven [11–13, 18, 25].

Using simple visual inspection of DWI images, Mayerhoefer et al. [18] reported that region-based agreement between WB-DWI with ^{18}F-FDG-PET-CT was 99.2% after 1–3 therapy cycles in their cohort of 51 adult patients with various lymphoma types, and Tsuji et al. [11] found that WB-DWI was concordant with ^{18}F-FDG-PET-CT in 100% of cases (n = 19) with lesion negative interim scans.

One potential advantage of applying quantitative ADC cut-offs for response assessment is to improve the specificity of

Table 4 Overall true positive rate, false positive rate, agreement rate and kappa for nodal and extra-nodal staging

Analyses	Agreement rate	TPR	FPR	Kappa (95% CI)
Analysis 1[a]				
Nodal sites	91% (799/875)	81% (184/226)	5% (34/649)	0.77 (0.72–0.82)
Extra-nodal sites	98% (638/652)	72% (28/39)	< 1% (3/613)	0.79 (0.68–0.90)
Analysis 2[b]				
Nodal sites	96% (799/831)	90% (184/204)	2% (12/627)	0.89 (0.86–0.93)
Extra-nodal sites	98% (638/652)	72% (28/39)	< 1% (3/613)	0.79 (0.68–0.90)
Sensitivity analysis 1[c]				
Nodal sites	97% (809/831)	91% (192/210)	1% (4/621)	0.93 (0.90–0.96)
Extra-nodal sites	99% (643/652)	79% (30/38)	< 1% (1/614)	0.86 (0.77–0.95)
Sensitivity analysis 2[d]				
Nodal sites	99% (822/831)	97% (203/210)	< 1% (2/621)	0.97 (0.95–0.99)
Extra-nodal sites	> 99% (650/652)	95% (36/38)	0% (0/616)	0.97 (0.93–1.00)

TPR true positive rate, *FPR* false positive rate, *CI* confidence interval

[a] Comparison between WB-MRI and primary reference standard before correction of simple anatomical boundaries labelling discrepancies

[b] Comparison between WB-MRI and primary reference standard following correction of simple anatomical boundaries labelling discrepancies

[c] Comparison between WB-MRI and enhanced reference standard (after removal of perceptual and technical errors in the primary reference standard)

[d] Comparison between WB-MRI and enhanced reference standard following removal of WB-MRI perceptual errors

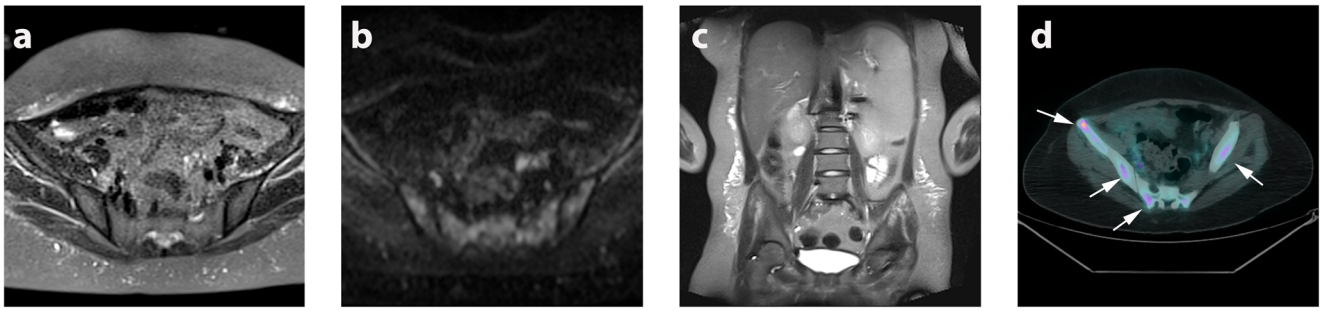

Fig. 4 Example of WB-MRI technical error. False negative WB-MRI technical error resulting in under-staging of a 15-year-old female patient with multifocal bone marrow involvement; **a** axial STIR-HASTE, **b** DWI b_{500} and **c** coronal STIR-HASTE MRI show no discernible bone marrow abnormality. **d** ^{18}F-FDG-PET-CT, however, demonstrates multifocal bone marrow metastasis (arrows)

simple visual assessment. Littooij et al. [13], for example, reported that applying an ADC cut-off value of 1.21×10^{-3} mm^2/s increased specificity for residual nodal disease detection by nearly 30% compared to visual inspection only.

By applying a similar ADC cut-off, we found that iWB-MRI agreed with the reference standard in a moderate 66% of patients.

One particular observation was the persistence of abnormal DWI bone marrow signal after successful treatment, resulting in underestimation of response by MRI and highlighting a limitation of visual response of extra-nodal disease on DWI. Quantitative ADC measurements my aid the differentiation between persistent tumour and treatment necrosis [26] and requires further investigation. For example, post-chemotherapy ADC monitoring in multiple myeloma has already shown promise for response assessment [27]. Such evidence is currently lacking in paediatric lymphoma, although intuitively ADC assessment could also be beneficial, and requires further evaluation.

Our study has some limitations. Our standard staging protocol, although primarily based on ^{18}F-FDG-PET-CT, also includes anatomical MRI sequences. There is a theoretical risk of incorporation bias as these sequences were available to the MDT when they created the primary reference standard [28]. However, DWI and DCE sequences were not available to the MDT, and the complete WB-MRI examination was viewed as a standalone examination by radiologists blinded to all other clinical information. As noted, ^{18}F-FDG-PET-CT is the mainstay of staging at our institution. Any incorporation bias would favour WB-MRI and the fact we report modest WB-MRI performance data suggests that any bias did not influence the overall study outcome.

We used an unblinded expert panel opinion and long-term follow-up data to derive the enhanced reference standard, an approach commonly used in studies of imaging diagnostic accuracy in absence of a single reference standard [14, 15].

We have used the highest *b* value of 500 s/mm^2 for DWI disease assessment. We acknowledge that a higher *b* value between 800 and 1000 s/mm^2 would have been in line with current recommendations on WB-DWI [29]. However, our ADC cut-off parameters were derived from previous pilot work [21] using similar DWI protocol as the current study. It is, however, possible that using a higher *b* value of 800–1000 s/mm^2 instead of 500 s/mm^2 could improve disease detection because of a superior lesion-to-contrast ratio. This could, for example, potentially decrease perceptual errors for extra-nodal disease assessment.

We used both qualitative and quantitative MRI assessment for staging and response monitoring. The generalizability of ADC quantitation across institutions and platforms, however, remains challenging [30, 31]. We also used a consensus reading paradigm for WB-MRI as at the time of the study set-up this mirrored our usual clinical practice and the use of ADC cut-offs was deemed exploratory [32]. We did reassuringly demonstrate good interobserver agreement with a third radiologist (as have others [19]). However, given that consensus reading is not widely used, it cannot be assumed that our data is representative of standard clinical practice where single reading is more common.

It has been shown that quantitative ADC changes following chemotherapy may differ between HL and non-HL subtypes of lymphoma [25] and our data is applicable to paediatric and adolescent HL.

Table 5 Per patient interim treatment response for whole-body MRI compared to combined reference standard

Overall patient response		Combined reference			
		CR (*n*)	PR (*n*)	NC (*n*)	PRO (*n*)
Trial WB-MRI	CR (*n*)	19	2	0	0
	PR (*n*)	9	6	1	0
	NC (*n*)	1	0	0	0
	PRO (*n*)	0	0	0	0

CR complete response, *n* number, *NC* no change, *PR* partial response, *PRO* progression, *WB-MRI* whole-body MRI

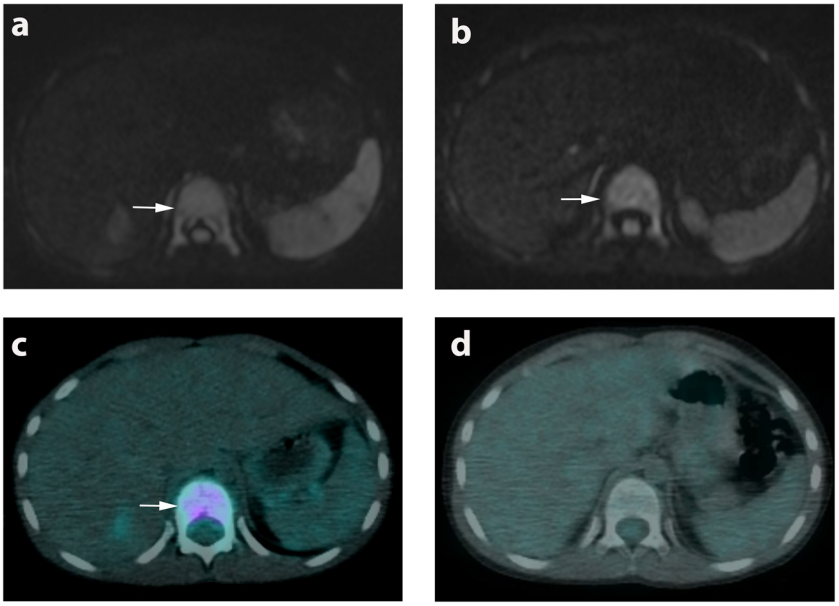

Fig. 5 Example of discrepant interim treatment response classification. WB-MRI and ^{18}F-FDG-PET-CT of an 8-year-old male subject with Ann Arbor stage 4 disease. Baseline WB-MRI (**a**) and ^{18}F-FDG-PET-CT (**c**) showing involvement of entire T11 vertebrae (arrows). Interim WB-MRI (**b**) showing no signal intensity changes (arrow) whilst interim ^{18}F-FDG-PET-CT (**d**) demonstrated complete response. Patient remained in remission following chemotherapy

Finally, although ADC changes as early as 1 week post chemotherapy have been documented for very early response assessment in adult lymphoma [33], the delayed second time point for iWB-MRI in our study was based on institutional guidelines for iPET-CT, Euronet trial [20] and recommendations in the literature [34, 35]. It would now be useful to investigate whether WB-MRI performs better for response assessment if performed at an earlier time point (e.g. 2 weeks) after chemotherapy.

In conclusion, WB-MRI with DWI has reasonable intrinsic diagnostic performance for nodal and extra-nodal staging of paediatric HL. However, in a substantial minority of patients it fails to achieve full concordance with standard imaging for all disease sites. WB-MRI has reasonable accuracy for interim treatment response classification but tends to underestimate disease response, particularly in extra-nodal disease sites. Overall, although promising, WB-MRI with DWI cannot currently replace standard imaging investigations in paediatric and adolescent Hodgkin's lymphoma and further research is required, particularly to derive optimum ADC cut-offs for disease status, and the significance of persistent extra-nodal abnormality following treatment.

Acknowledgements AL was supported by a Cancer Research UK/Engineering and Physical Sciences Research Council (CRUK/EPSRC) award (C1519/A10331 and C1519/A16463) from the University College London/King's College London (UCL/KCL) Comprehensive Cancer Imaging Centre (CCIC).

This work was undertaken at the Comprehensive Biomedical Centre (BRC), University College Hospital London (UCLH), which received a proportion of the funding from the National Institute for Health Research (NIHR). The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health.

ST is an NIHR senior investigator.

The authors would like to acknowledge and thank University College London Cancer Trials Centre (UCL CTC), Dr Darren Edwards and Mrs K M Mak for their contribution towards this manuscript.

The trial was managed by the Cancer Research UK and University College London Cancer Trials Centre. The authors would also like to thank the patients and their families who took part in the study and the investigators and research staff at the participating centre.

Funding This study has received funding from Cancer Research UK, project number CRUK ASC 12707.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Professor Stuart A Taylor.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Andre Lopes) is a statistician with significant statistical expertise.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional review board approval was obtained.

Methodology
- prospective
- diagnostic or prognostic study/observational
- performed at one institution

References

1. Ward E, DeSantis C, Robbins A, Kohler B, Jemal A (2014) Childhood and adolescent cancer statistics. CA Cancer J Clin 64: 83–103
2. Uslu L, Doing J, Link M, Rosenberg J, Quon A, Daldrup-Link HE (2015) Value of 18F-FDG PET and PET/CT for evaluation of pediatric malignancies. J Nucl Med 56:274–286
3. Wahl RL, Jacene H, Kasamon Y, Lodge MA (2009) From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med 50(Suppl 1):122S–150S
4. Cheson BD, Fisher RI, Barrington SF et al (2014) Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J Clin Oncol 32:3059–3068
5. Hall EJ, Brenner DJ (2008) Cancer risks from diagnostic radiology. Br J Radiol 81:362–378
6. Nievelstein RA, Littooij AS (2016) Whole-body MRI in paediatric oncology. Radiol Med 121:442–453
7. Brenner D, Elliston C, Hall E, Berdon W (2001) Estimated risks of radiation-induced fatal cancer from pediatric CT. AJR Am J Roentgenol 176:289–296
8. Davis JT, Kwatra N, Schooler GR (2016) Pediatric whole-body MRI: a review of current imaging techniques and clinical applications. J Magn Reson Imaging 44:783–793
9. Greer MC, Voss SD, States LJ (2017) Pediatric cancer predisposition imaging: focus on whole-body MRI. Clin Cancer Res 23:e6–e13
10. Maggialetti N, Ferrari C, Minoia C et al (2016) Role of WB-MR/DWIBS compared to (18)F-FDG PET/CT in the therapy response assessment of lymphoma. Radiol Med 121:132–143
11. Tsuji K, Kishi S, Tsuchida T et al (2015) Evaluation of staging and early response to chemotherapy with whole-body diffusion-weighted MRI in malignant lymphoma patients: A comparison with FDG-PET/CT. J Magn Reson Imaging 41:1601–1607
12. Punwani S, Taylor SA, Saad ZZ et al (2013) Diffusion-weighted MRI of lymphoma: prognostic utility and implication for PET/MRI? Eur J Nucl Med Mol Imaging 40:373–385
13. Littooij AS, Kwee TC, de Keizer B et al (2015) Whole-body MRI-DWI for assessment of residual disease after completion of therapy in lymphoma: a prospective multicenter study. J Magn Reson Imaging 42:1646–1655
14. Punwani S, Taylor SA, Bainbridge A et al (2010) Pediatric and adolescent lymphoma: comparison of whole-body STIR half-Fourier RARE MR imaging with an enhanced PET/CT reference for initial staging. Radiology 255:182–190
15. Kwee TC, Vermoolen MA, Akkerman EA et al (2014) Whole-body MRI, including diffusion-weighted imaging, for staging lymphoma: comparison with CT in a prospective multicenter study. J Magn Reson Imaging 40:26–36
16. Regacini R, Puchnick A, Shigueoka DC, Iared W, Lederman HM (2015) Whole-body diffusion-weighted magnetic resonance imaging versus FDG-PET/CT for initial lymphoma staging: systematic review on diagnostic test accuracy studies. Sao Paulo Med J 133:141–150
17. Mayerhoefer ME, Karanikas G, Kletter K et al (2014) Evaluation of diffusion-weighted MRI for pretherapeutic assessment and staging of lymphoma: results of a prospective study in 140 patients. Clin Cancer Res 20:2984–2993
18. Mayerhoefer ME, Karanikas G, Kletter K et al (2015) Evaluation of diffusion-weighted magnetic resonance imaging for follow-up and treatment response assessment of lymphoma: results of an 18F-FDG-PET/CT-controlled prospective study in 64 patients. Clin Cancer Res 21:2506–2513
19. Littooij AS, Kwee TC, Barber I et al (2014) Whole-body MRI for initial staging of paediatric lymphoma: prospective comparison to an FDG-PET/CT-based reference standard. Eur Radiol 24:1153–1165
20. Körholz D, Wallace H, Landman-Parker J (2006) EuroNet-Paediatric Hodgkin's Lymphoma Group, first international intergroup study for classical Hodgkin's lymphoma in children and adolescents, radiotherapy manual. https://www.skion.nl/workspace/uploads/euronet-phl-c1_workingcopy_inkl_amendm06_mw_2012-11-14_0.pdf. Accessed 02 Feb 2018
21. Punwani S, Prakash V, Bainbridge A et al (2010) Quantitative diffusion weighted MRI: a functional biomarker of nodal disease in Hodgkin's lymphoma. Cancer Biomarker 7:249–259
22. Eich HT, Diehl V, Görgen H et al (2010) Intensified chemotherapy and dose-reduced involved-field radiotherapy in patients with early unfavorable Hodgkin's lymphoma: final analysis of the German Hodgkin Study Group HD11 trial. J Clin Oncol 28:4199–4206
23. Vandecaveye V, De Keyzer F, Vander Poorten V et al (2009) Head and neck squamous cell carcinoma: value of diffusion-weighted MR imaging for nodal staging. Radiology 251:134–146
24. Afaq A, Fraioli F, Sidhu H et al (2017) Comparison of PET/MRI with PET/CT in the evaluation of disease status in lymphoma. Clin Nucl Med 42:e1–e7
25. Hagtvedt T, Seierstad T, Lund KV et al (2015) Diffusion-weighted MRI compared to FDG PET/CT for assessment of early treatment response in lymphoma. Acta Radiol 56:152–158
26. Padhani AR, Koh DM, Collins DJ (2011) Whole-body diffusion-weighted MR imaging in cancer: current status and research directions. Radiology 261:700–718
27. Latifoltojar A, Hall-Craggs M, Bainbridge A et al (2017) Whole-body MRI quantitative biomarkers are associated significantly with treatment response in patients with newly diagnosed symptomatic multiple myeloma following bortezomib induction. Eur Radiol 27: 5325–5336
28. Kohn MA, Carpenter CR, Newman TB (2013) Understanding the direction of bias in studies of diagnostic test accuracy. Acad Emerg Med 20:1194–1206
29. Barnes A, Alonzi R, Blackledge M et al (2018) UK quantitative WB-DWI technical workgroup: consensus meeting recommendations on optimisation, quality control, processing and analysis of quantitative whole-body diffusion-weighted imaging for cancer. Br J Radiol. https://doi.org/10.1259/bjr.20170577
30. Celik A (2016) Effect of imaging parameters on the accuracy of apparent diffusion coefficient and optimization strategy. Diagn Interv Radiol 22:101–107
31. Koh DM, Collins DJ, Orton MR (2011) Intravoxel incoherent motion in body diffusion-weighted MRI: reality and challenges. AJR Am J Roentgenol 196:1351–1361
32. Bankier AA, Levine D, Halpern EF, Kressel HY (2010) Consensus interpretation in imaging research: is there a better way? Radiology 257:14–17
33. Horger M, Claussen C, Kramer U, Fenchel M, Linchy M, Kaufmann S (2014) Very early indicators of response to systemic therapy in lymphoma patients based on alterations in water diffusivity—a preliminary experience in 20 patients undergoing whole-body diffusion-weighted imaging. Eur J Radiol 83:1655–1664
34. Furth C, Steffen IG, Amthauer H et al (2009) Early and late therapy response assessment with [18F]fluorodeoxyglucose positron emission tomography in pediatric Hodgkin's and adapted treatment guided by interim PET-CT scan in advanced Hodgkin's lymphoma: analysis of a prospective multicenter trial. J Clin Oncol 27:4385–4391
35. Johnson P, Federico M, Kirkwood A et al (2016) Adapted treatment guided by interim PET-CT scan in advanced Hodgkin's lymphoma. N Engl J Med 374:2419–2429

Whole body MRI protocol

Patients were imaged on a 1.5T MRI system (Avanto, Siemens, Erlangen, Germany) from the skull to mid-thigh in the supine position, using the manufacturer's body and spine array coils. Immediately prior to imaging, 0.3 mg/kg of body weight of hyoscine butylbromide (Buscopan; Boehringer Ingelheim, Ingelheim, Germany) was administered.

At the recruitment site, basic anatomical sequences are routinely acquired to supplement the low dose CT component of the ^{18}F-FDG PET-CT protocol. In brief, respiratory and electrocardiographically gated axial and coronal whole body fat suppressed T2-weighted MRI together with axial periodically rotated overlapping lines with enhanced reconstruction (PROPELLER) were acquired through the chest (in maximum inspiration) and complemented by a single-phase post-contrast T1-weighted sequence (3D Fast low angle shot technique, FLASH) through upper abdomen at 25 seconds following gadolinium injection.

As part of the trial intervention, and to provide a comprehensive "stand alone" protocol, the WB-MRI scan was extended and the above sequences were complemented by the addition of axial free-breathing DWI (with four b-values b_0, b_{100}, b_{300} and b_{500}) acquired through the whole-body, together with multiphase breath-hold dynamic contrast enhanced T1-weighted series through the liver and spleen (following a single intravenous dose of 0.1 mmol/kg body weight of gadoterate meglumine (Dotarem; Laboratoire Guerbet, Aulnay-sous-Bois, France) as described previously [1]. Axial and coronal breath-hold post-contrast T1-weighted MRI of the lungs were then acquired following the liver and splenic acquisition. ADC maps were generated using vendor's software.

The full WB-MRI scanning protocol was completed within 60 minutes. WB-MRI scanning parameters are summarised in supplemental Table 1.

^{18}F-FDG PET-CT protocol

^{18}F-FDG PET-CT scans were performed with integrated PET/CT scanners (Discovery ST or Discovery VCT, GE Healthcare, Waukesha Wisconsin, USA). Patients fasted for 6 hours and blood glucose levels were tested to exclude hyperglycemia (levels >180mg/dL). For pediatric patients, the doses were adjusted according to the European Association of Nuclear Medicine (EANM) pediatric dosage card [2]. ^{18}F-FDG (14MBq - 370MBq) was intravenously injected 60 minutes before imaging. Prior to acquiring the whole-body PET 3D emission scan, a non-contrast CT was obtained for attenuation correction (80-120 kVp, modulated mA [10-200mA], pitch 1.375, 3.75mm slice thickness). Images were acquired at 3 min per bed position as per departmental pediatric protocol.

Contrast-enhanced chest CT

CE chest CT was performed in cases of equivocal lung findings on ^{18}F-FDG PET-CT [3], after intravenous contrast administration (2.0ml/kg Omnipaque 300, General Electric Healthcare, Milwaukee, Wisconsin, USA), using a 64-multi-detector row CT scanner (Siemens Somatom 64, Siemens, Erlangen, Germany) (120kVp, 45 ref. mAs, 0.5s rotation time, 64x0.6mm detectors, pitch 1.4, 24x1.2 collimation).

Abdominal ultrasound

Abdominal ultrasound was performed by a consultant radiologist in case of equivocal solid organ involvement on other cross-sectional imaging.

Standard staging imaging interpretation

^{18}F-FDG PET-CT was interpreted by a nuclear medicine physician (Blinded for review) on a dedicated workstation (Xeleris 2; GE Healthcare, Milwaukee, Wisconsin, USA). The basic anatomical WB-MRI sequences including single-phase post-contrast sequences through the upper abdomen (but excluding DWI and dynamic contrast enhanced sequences), abdominal ultrasound and high-resolution contrast-enhanced chest CT images (when available) were evaluated by consultant pediatric radiologist (Blinded for review) using a standard picture archiving and communication system (PACS) (IMPAX version 6.5.1; Agfa-Gevaert, Morstel, Belgium). The readers derived the disease status for the 18 nodal and 14 extra nodal, as well as the final Ann Arbor stage. The 18 nodal disease sites were cervical [right (R) and left (L)], supraclavicular (R and L), subpectoral (R and L), axillary (R and L), mediastinal, splenic hilar, liver hilar, mesenteric, retroperitoneal, iliac (R and L), inguinal (R and L) and “other” sites, and 14 extra-nodal disease sites were lung (R and L), pleura, pericardium, chest wall, liver, spleen, kidney (R and L), stomach, bowel, pancreas, bone marrow and “other” sites.

Definitions of nodal disease positivity were those utilized by the Euronet PHL-C1 or LP1 trials [4], and based on long-axis size and ^{18}F-FDG uptake in comparison to background activity (supplemental Table 2).

Disease volume was derived using all three-axis measurements [(X x Y x Z)/2] for largest nodal mass in each nodal site, or for the conglomerate nodal mass if no discrete nodal tissue was visible, and was used for subsequent treatment response evaluation.

The criteria for extra-nodal disease status using standard imaging is summarised in supplemental Table 3.

Staging Whole body MRI interpretation

Two radiologists (Blinded for review) in consensus reviewed anonymized WB-MRI datasets as a “stand alone staging investigation” using Osirix (Version 4.0, Apple, California, USA) viewing software, utilizing all the available sequences (including the DWI and dynamic contrast enhanced images). The radiologists were blinded to the clinical history (other than the diagnosis of lymphoma) and all other investigations.

The disease status for the same 18 nodal disease sites and 14 extra-nodal disease sites, as well as final Ann Arbor stage were derived according to predefined trial criteria. Specifically for lymph nodes, disease positivity was defined using a combination of size and ADC criteria (supplemental Table 2). Size criteria for disease positivity were based on those used by the Euronet PHL-C1 or LP1 trials [4]. The largest diameter of a single lymph node or a lymph node conglomerate was measured on T2-weighted MRI. Nodal disease volume was derived as described above for standard imaging.

ADC quantitation was performed by placing a region of interest in the largest cross section of the node on the ADC map, guided by anatomically matched axial fat-suppressed T2-weighted MRI. The derived ADC cut offs for nodal positivity were based on previous work [5].

The criteria for extra-nodal disease on WB-MRI was based on structural observations and MRI signal changes (supplemental Table 3).

The disease status for the same 18 nodal and 14 extra-nodal disease sites and Ann Arbor stage were derived according to predefined trial criteria.

Standard Imaging interim response interpretation

The nuclear medicine physician and consultant radiologist who evaluated the standard initial staging imaging interpreted the iPET-CT and anatomical iWB-MRI sequences, and derived the treatment response for all nodal and extra-nodal disease sites using predefined trial criteria based on tumour volume and ^{18}F-FDG uptake (used by the Euronet trials, Table 1) [4]. Consistent with the initial staging measurements, nodal response was assessed for largest diameter of a single lymph node (when visible) or a conglomerate nodal mass (if no discreet node was visible).

Extra-nodal response was classified into four categories: [a] locally undetectable (complete response), [b] locally detectable but reduction in size or number of deposits (partial response), [c] locally unchanged (no change in the number or size of deposits) and [d] locally progressive (increase in size or number of deposits).

The overall interim per patient response for standard imaging was defined using the least responsive nodal and/or extra-nodal disease sites.

Whole body MRI interim response interpretation

The same radiologists who interpreted the initial staging WB-MRI, evaluated the complete iWB-MRI and derived the treatment response for all nodal and extra-nodal disease sites, again blinded to all other investigations, and using predefined criteria based on changes in nodal volume (used by the Euronet trials) [4] and ADC measurement (Table 1).

The ADC criteria for response were based on those derived from previous work [5]. Extra-nodal response was evaluated by qualitative assessment of iWB-MRI and classified into four categories as for the standard imaging described above.

The overall per patient interim response for iWB-MRI was defined using the least responsive nodal and/or extra-nodal disease sites.

Primary reference standard

The primary reference standard for all 32 disease sites, Ann Arbor stage at initial staging, and the interim treatment response evaluation was assigned by a multi-disciplinary team (MDT) meeting attended by a consultant radiologist, nuclear medicine physician, two pediatric haemato-oncologists and a haematopathologist. The panel based their assessment on all standard imaging test results (interpreted as described above), together with all clinical information including clinical examination findings, blood test results and available histology.

Central review of imaging discrepancies and creation of an enhanced reference standard

Given the potential limitations of standard imaging in staging HL, and the risk of radiologist/ nuclear medicine physician perceptual errors adversely influencing the primary reference standard, a retrospective enhanced reference standard was produced to better evaluate the potential accuracy of WB-MRI. Specifically, all discrepancies between WB-MRI and standard imaging tests (including ^{18}F-FDG PET-CT) at initial staging were reviewed by an expert panel comprising two radiologists (one of whom was not involved in the main trial radiological interpretation), two nuclear medicine physicians (one of whom was not involved in the main trial PET scan interpretation) and two pediatric hemato-oncologists. The panel reviewed all staging, interim and end of treatment scans and had access to follow up imaging and clinical outcomes up to 24 months post chemotherapy.

Anatomical boundary description discrepancies

Initially the panel corrected simple labeling discrepancies that were due to differences in disease site description between those interpreting the WB-MRI and those interpreting standard imaging, usually between adjacent anatomical sites. For example, if an involved node was described as “cervical” on WB-MRI but “supraclavicular” on standard imaging, the panel opined if this was a true discrepancy or if both modalities had described the same disease, and just classified its anatomical location differently, in which case the discrepancy was re-classified as concordant.

Correction of perceptual and technical errors in the primary standard reference and creation of an enhanced reference standard

Once simple anatomical boundary labeling discrepancies were corrected, the panel reviewed the remaining discrepancies and decided if the primary reference standard needed correcting based on all the available imaging and follow up data. Initially simple perceptual errors in standard imaging interpretation were corrected, for example unequivocal areas of disease positivity that responded to treatment, but were missed on the original ^{18}F-FDG PET-CT interpretation and visible on the ^{18}F-FDG PET-CT in retrospect on review (Fig. 1).

Thereafter positive findings on WB-MRI not visible on the standard imaging even in retrospect were reviewed to see if any were technical failures of standard imaging. Only unequivocal disease sites with clear response to treatment on WB-MRI were considered technical failures of standard imaging by the panel, otherwise such findings were classified as WB-MRI false positives. In a similar fashion, the panel identified any false positive findings on standard imaging. The creation of this

enhanced reference standard aimed to define the true disease status of the patients as far as possible.

Correction of MRI for perceptual errors

Finally, all the WB-MRI errors against the enhanced reference standard were classified into perceptual errors when the abnormality was visible in retrospect on the WB-MRI, or technical error when it was not. By identifying and correcting MRI perceptual errors it was then possible to assess the theoretical best performance of the WB-MRI protocol.

Sample size calculation

Assuming a discordant rate of 20% (extrapolated from pilot data [6]), a total of 55 patients would be sufficient to exclude a discordance rate of greater than 35% with 80% power and one-sided 5% significance level. A sample size of 55 patients would also be sufficient to form a 95%CI with 20% precision around an assumed kappa of 0.86 and an assumed MRI sensitivity of at least 85% for site-specific disease [6].

Supplemental Table 1: Whole-body MRI sequence parameters

	Axial STIR HASTE	Coronal STIR HASTE	Axial STIR DWI (b0,100,300,500)	Axial T2 PROPELLER (chest)	Axial post-contrast (lung)	Coronal post-contrast (lung)	3D FLASH for DCE (liver & spleen)
TR/TE (ms)	*800/60*	*800/60*	*4900/66*	*3000/133*	*2.85/0.99*	*2.94/1.04*	*2.87/0.93*
Inversion time (ms)	*130*	*180*	*180*	*N/A*	*N/A*	*N/A*	*N/A*
Matrix	*256×192*	*128×96*	*128×96*	*256×256*	*256×88*	*256×128*	*256×176*
Slice Thickness (mm)	*7*	*4*	*4*	*3*	*2.5*	*3.5*	*2.5*
No. of slices	*19*	*27*	*27*	*23*	*104*	*56*	*80*
Averages	*2*	*8*	*8*	*1*	*1*	*1*	*1*
Echo train	*256*	*1*	*1*	*50*	*1*	*1*	*1*
PAT	*2*	*2*	*2*	*1*	*1*	*1*	*2*
Flip angle	*180*	*90*	*90*	*150*	*15*	*15*	*9*
Pixel spacing	*1.56×1.56*	*0.8×0.8*	*0.8×0.8*	*1.25×1.25*	*1.4×1.4*	*1.2×1.2*	*1.56×1.56*

TR: Repetition time
TE: Echo time
PAT: Parallel acquisition technique
STIR: Short tau inversion recovery
HASTE: Half-Fourier single shot turbo spin echo
DWI: Diffusion weighted imaging
FLASH: Fast low angle shot technique
DCE: Dynamic contrast enhanced
PROPELLER: Periodically rotated overlapping lines with enhanced reconstruction

Supplemental Table 2: Pre-defined criteria for nodal assessment

	Standard Imaging		
	Cross sectional imaging (Anatomical MRI sequences, CT component of PET-CT)	**PET-CT Imaging ***	**WB-MR Imaging**
Positive	Nodes > 2cm **	N/A	Nodes > 2cm**
	Nodes 1-2 cm	FDG-PET positive	Nodes 1-2 cm with ADC ≤ 1.2
	Nodes < 1 cm	FDG-PET positive	Nodes < 1 cm with ADC ≤ 0.8
Equivocal	Nodes 1-2 cm	FDG-PET equivocal	Nodes 1-2 cm with ADC >1.2 and <1.8
Negative	Nodes 1-2 cm	FDG-PET negative	Nodes 1-2 cm with ADC ≥ 1.8
	Nodes < 1 cm	FDG-PET negative	Nodes < 1 cm with ADC ≥ 0.8

PET: Positron emission tomography
FDG: ^{18}F-2-fluro-2-deoxy-D-glucose
ADC: Apparent diffusion coefficient
WB-MRI: Whole-body MRI
DWI: Diffusion weighted imaging

***** Involvement defined as uptake above surrounding background in a location incompatible with normal physiological activity

****** Long axis diameter

The largest diameter of all nodes was measured in 3 planes (axial long and short axis and coronal cranio-caudal axis) using the fat-suppressed T2-weighted images and CT scan. Disease volume was derived using all three axis measurements [(X x Y x Z)/2] for subsequent treatment response evaluation.

Supplemental Table 3: Pre-defined criteria for extra-nodal assessment

Sites	Standard Imaging (Anatomical MRI sequences, PET-CT, contrast enhanced CT Chest and abdominal ultrasound)	WB-MRI
Pleura	Involvement of the pleura is assumed if • the lymphoma is contiguous with the pleura without fat lamella or • the lymphoma invades the chest wall or • a pleural effusion occurs which cannot be explained by a venous congestion. Extension: Abnormal nodular tissue within the pleura contiguous with the main nodal mass.	Extension: Abnormal nodular moderate-high signal of equal intensity to nodal tissue within the pleura contiguous with the main nodal mass. Separate: Abnormal high signal of fluid intensity anatomically in keeping with a pleural effusion which cannot be explained by associated pulmonary oedema; or, pleural nodules discrete to the main lymph node mass.
Pericardium	Pericardial involvement is assumed if • the lymphoma has a broad area of close contact towards the heart surface beyond the valve level (ventriculus area) or • a pericardial effusion occurs/ nodules without associated mediastinal lymph node mass. Extension: Extensive contact between mediastinal lymph node mass and pericardium to the level of the ventricles in the presence of a pericardial effusion and / or pericardial nodules.	Extension: Extensive contact between mediastinal lymph node mass and pericardium to the level of the ventricles in the presence of a pericardial effusion and / or pericardial nodules. Separate: Pericardial effusion / nodules without associated mediastinal lymph node mass.
Chest wall	Chest wall infiltration is defined as extension of a mediastinal mass on CT and/or PET positive focal chest wall lesion.	Extension: Moderate-high signal infiltration of the chest wall in continuum with a lymphatic mass/or positive focal chest wall lesion on T2-STIR, DWI and/or post-contrast.

Sites	Standard Imaging (Anatomical MRI sequences, PET-CT, contrast enhanced CT Chest and abdominal ultrasound)	WB-MRI
Lung	A disseminated lung involvement (implying stage IV) is assumed if • there are more than three foci or • an intrapulmonary focus has a diameter of more than 10 mm. Extension: Abnormal infiltration of the lung in continuum with a lymphatic mass.	Extension: Abnormal moderate-high signal infiltration of the lung in continuum with a lymphatic mass. Separate: Abnormal moderate-high signal focus (>1cm diameter) within the lung discrete to lymphatic tissue or more than three foci.
Bone marrow	Bone involvement is assumed if a bone biopsy is positive **or** CT bony window is positive with or without further confirmation by other imaging methods in the same region **or** a positive bone scan is confirmed by either FDG-PET or MRI.	Homogenous moderate-high signal foci within bone on T2-STIR, DWI and/or post-contrast at a site discrete to the bone marrow biopsy.
Liver	Focal changes in the liver structure on ultrasonography that are suspicious of tumour are considered positive – independent of the FDG-PET result. In case of doubtful involvement of liver (e.g. structures atypical of tumour in sonography or MRI) the liver is considered involved if FDG-PET is positive. Extension: Moderate-high signal infiltration of the liver in continuum with an adjacent lymphatic mass	Extension: Moderate-high signal infiltration of the liver in continuum with an adjacent lymphatic mass on T2-STIR, DWI and/or post-contrast. Separate: Low signal (relative to surrounding liver) discrete foci within the liver not in continuation with an adjacent lymphatic mass.
Spleen	Focal changes in the splenic structure on ultrasonography that are suspicious of tumour are considered positive – independent of the FDG-PET result. In case of doubtful involvement of liver (e.g. structures atypical of tumour in sonography or MRI) the liver is considered involved if FDG-PET is positive. Moderate-high signal infiltration of the	Extension: Moderate-high signal infiltration of the spleen in continuum with an adjacent lymphatic mass. Separate: Low signal (relative to surrounding spleen) discrete foci within the spleen on T2-STIR, DWI and/or DCE not in continuation with an adjacent lymphatic mass.

Sites	Standard Imaging (Anatomical MRI sequences, PET-CT, contrast enhanced CT Chest and abdominal ultrasound)	WB-MRI
	spleen in continuum with an adjacent lymphatic mass.	
Kidney	Diffuse enlargement with distortion of the renal parenchyma **or** focal lesion on CT/MRI/US **or** PET/CT positive disease.	Global or focal renal enlargement and / or discrete renal mass.
Stomach	Focal thickening on CT/MRI that also demonstrates PET/CT positivity.	Marked wall thickening in a distended stomach with moderate-high signal.
Pancreas	Diffuse enlargement with distortion of the pancreatic parenchyma **or** focal lesion on CT/MRI/US **or** PET/CT positive disease.	Focal signal change within the pancreas or global pancreatic enlargement.
Bowel	Focal thickening on CT/MRI that also demonstrates PET/CT positivity.	Focal bowel wall thickening and elevated STIR-HASTE signal intensity.

Supplemental Table 4: list of nodal and extra-nodal discrepancies between WB-MRI and the final enhanced reference standard (ERS) following removal of WB-MRI perceptual errors

Site	***WB-MRI***	***ERS***	***Reason***
Supraclavicular (n)	*Negative*	*Positive*	ADC measurement of equivocal node by size criteria
Axillary (n)	*Positive*	*Negative*	ADC measurement of equivocal node by size criteria
Cervical (n)	*Negative*	*Positive*	Subcentimeter LN positive on FDG PET
Axillary (n)	*Negative*	*Positive*	Subcentimeter LN positive on FDG PET
Cervical (n)	*Negative*	*Positive*	Subcentimeter LN positive on FDG PET
Bone marrow (e)	*Negative*	*Positive*	Multi-focal bone marrow involvement missed on WB-MRI
Cervical (n)	*Positive*	*Negative*	ADC measurement of equivocal node by size criteria
Axillary (n)	*Negative*	*Positive*	ADC measurement of equivocal node by size criteria
Lung (e)	*Negative*	*Positive*	Multiple small lung foci detected on CE chest CT scan and missed on WB-MRI
Liver Hilar (n)	*Negative*	*Positive*	ADC measurement of equivocal node by size criteria
Supraclavicular (n)	*Negative*	*Positive*	Subcentimeter LN positive on FDG PET

n: Nodal site**, e:** Extra-nodal site, **LN:** Lymph node, **ERS:** Enhanced reference standard, **PET**: Positron emission tomography
FDG: ^{18}F-2-fluro-2-deoxy-D-glucose, **ADC**: Apparent diffusion coefficient, **CE:** Contrast enhanced, **WB-MRI:** Whole body MRI

Supplemental Table 5: Ann Arbor staging agreement

Ann Arbor Staging		Primary Reference Standard					
		I	II	II E	III	IV	IV E
		N=2	N=28	N=0	N=4	N=13	N=3
WB MRI (Analysis 1)	I	2 (100%)	1 (4%)	-	-	-	-
	II	-	22 (79%)	-	3 (75%)	-	-
	II E	-	-	0 (0%)	-	-	1 (33%)
	III	-	4 (14%)	-	1 (25%)	1 (8%)	-
	IV	-	1 (4%)	-	-	12 (92%)	-
	IV E	-	-	-	-	-	2 (67%)
		Enhanced Reference Standard					
		I	II	II E	III	IV	IV E
		N=2	N=26	N=0	N=5	N=14	N=3
WB MRI (Analysis 2)	I	2 (100%)	1 (4%)	-	-	-	-
	II	-	22 (85%)	-	3 (60%)	-	-
	II E	-	-	0 (0%)	-	-	1 (33%)
	III	-	2 (7%)	-	2 (40%)	1 (7%)	-
	IV	-	1 (4%)	-	-	13 (93%)	-
	IV E	-	-	-	-	-	2 (67%)
		Enhanced Reference Standard					
		I	II	II E	III	IV	IV E
		N=2	N=26	N=0	N=5	N=14	N=3
WB MRI (Analysis 3)	I	2 (100%)	-	-	-	-	-
	II	-	26 (100%)	-	-	-	-
	II E	-	-	0 (0%)	-	-	1 (33%)
	III	-	-	-	5 (100%)	1 (7%)	-
	IV	-	-	-	-	13 (92%)	-
	IV E	-	-	-	-	-	2 (67%)

Analysis 1: Comparison between WB-MRI and primary reference standard

Analysis 2: Comparison between WB-MRI and enhanced reference standard before removal of WB-MRI perceptual errors

Analysis 3: Comparison between WB-MRI and enhanced reference standard following removal of WB-MRI perceptual errors

Supplemental Table 6: List of nodal disease sites with discrepant interim treatment response between WB-MRI and enhanced reference standard.

Disease site	Reference standard response	WB-MRI response	Reason for discrepancy
Cervical	CR	PRa	Percentage residual tumour 34% for WB-MRI and 19% for PET-CT
Inguinal	PRi	CR	Residual PET Positivity
Cervical	CR	PRa	Percentage residual tumour 26% for WB-MRI and 2.7% for PET-CT
Supraclavicular	CR	PRa	Percentage residual tumour 37% for WB-MRI and 3.8% for PET-CT
Mediastinal	PRa	CR	Percentage residual tumour 14% for WB-MRI and 40% for PET-CT
Supraclavicular	PRi	CR	Percentage residual tumour 6% for WB-MRI and 70% for PET-CT
Supraclavicular	PRa	CR	Percentage residual tumour 4.6% for WB-MRI and 38% for PET-CT
Mediastinal	PRi	CR	Residual PET positivity
Cervical	NC	PRi	Percentage residual tumour 53% for WB-MRI and 80% for PET-CT
Supraclavicular	NC	PRa	Percentage residual tumour 32.5% for WB-MRI and 84% for PET-CT
Cervical	PRa	CR	Percentage residual tumour 11% for WB-MRI and 42% for PET-CT
Liver hilar	PRi	CR	Percentage residual tumour 0 for WB-MRI and 57% for PET-CT
Iliac	PRa	CR	Percentage residual tumour 3.3% for WB-MRI and 25% for PET-CT
Mediastinal	PRa	CR	Percentage residual tumour 13% for WB-MRI and 28% for PET-CT
Supraclavicular	PRa	CR	Percentage residual tumour 6.4% for WB-MRI and 42% for PET-CT
Mediastinal	PRi	CR	Residual PET Positivity
Supraclavicular	PRi	CR	Residual PET Positivity

PRa: Partial response adequate
PRi: Partial response inadequate
CR: Complete response
NC: No change
WB-MRI: Whole-body MRI
PET-CT: Positron emission tomography-computed tomography scan

References:

[1] Punwani S, Cheung KK, Skipper N, et al (2013) Dynamic contrast-enhanced MRI improves accuracy for detecting focal splenic involvement in children and adolescents with Hodgkin disease. Pediatr Radiol 43:941-949.

[2] Lassmann M, Biassoni L, Monsieurs M, Franzius C; EANM Dosimetry and Paediatrics Committees (2008) The new EANM paediatric dosage card: additional notes with respect to F-18. Eur J Nucl Med Mol Imaging 35:1666-1668.

[3] Kleis M, Daldrup-Link H, Matthay K, et al (2009) Diagnostic value of PET/CT for the staging and restaging of pediatrics tumors. Eur J Nucl Med Mol Imaging 36:23-36.

[4] Körholz D, Wallace H, Landman-Parker J (2006) EuroNet-Paediatric Hodgkin's Lymphoma Group, First international Inter-Group Study for classical Hodgkin's Lymphoma in Children and Adolescents, Radiotherapy Manual. https://www.skion.nl/workspace/uploads/euronet-phl-c1_workingcopy_inkl_amendm06_mw_2012-11-14_0.pdf

[5] Punwani S, Prakash V, Bainbridge A, et al (2010) Quantitative diffusion weighted MRI: A functional biomarker of nodal disease in Hodgkin's lymphoma. Cancer Biomarker 7:249-259.

[6] Punwani S, Taylor SA, Bainbridge A, et al (2010) Pediatric and adolescent lymphoma: comparison of whole-body STIR half-Fourier RARE MR imaging with an enhanced PET/CT reference for initial staging. Radiology 255:182-190.