




Comparison of MRI response evaluation methods in rectal cancer: a multicentre and multireader validation study

Najim El Khababi^{1,2} · Regina G. H. Beets-Tan^{1,2} · Renaud Tissier³ · Max J. Lahaye^{1,2} · Monique Maas^{1,2} · Luís Curvo-Semedo⁴ · Raphaëla C. Dresen⁵ · Stephanie Nougaret⁶ · Geerard L. Beets^{2,7} · Doenja M. J. Lambregts^{1,2}  · on behalf of the rectal MRI study group

Received: 7 July 2022 / Revised: 30 September 2022 / Accepted: 29 November 2022 / Published online: 28 December 2022

© The Author(s), under exclusive licence to European Society of Radiology 2022

Abstract

Objectives To compare four previously published methods for rectal tumor response evaluation after chemoradiotherapy on MRI. **Methods** Twenty-two radiologists (5 rectal MRI experts, 17 general/abdominal radiologists) retrospectively reviewed the post-chemoradiotherapy MRIs of 90 patients, scanned at 10 centers (with non-standardized protocols). They applied four response methods; two based on T2W-MRI only (MRI tumor regression grade (mrTRG); split-scar sign), and two based on T2W-MRI+DWI (modified-mrTRG; DWI-patterns). Image quality was graded using a 0–6-point score (including slice thickness and in-plane resolution; sequence angulation; DWI b-values, signal-to-noise, and artefacts); scores < 4 were classified below average. Mixed model linear regression was used to calculate average sensitivity/specificity/accuracy to predict a complete response (versus residual tumor) and assess the impact of reader experience and image quality. Group interobserver agreement (IOA) was calculated using Krippendorff's alpha. Readers were asked to indicate their preferred scoring method(s).

Results Average sensitivity/specificity/accuracy was 57%/64%/62% (mrTRG), 36%/79%/66% (split-scar), 40%/79%/67% (modified-mrTRG), and 37%/82%/68% (DWI-patterns); mrTRG showed higher sensitivity but lower specificity and accuracy ($p < 0.001$) compared to the other methods. IOA was lower for the split scar method (0.18 vs. 0.39–0.43). Higher reader experience had a significant positive effect on diagnostic performance and IOA (except for the split scar sign); below-average imaging quality had a significant negative effect on diagnostic performance. DWI pattern was selected as the preferred method by 73% of readers.

Conclusions Methods incorporating DWI showed the most favorable results when combining diagnostic performance, IOA, and reader preference. Reader experience and image quality clearly impacted diagnostic performance emphasizing the need for state-of-the-art imaging and dedicated radiologist training.

Key Points

- In a multireader study comparing 4 MRI methods for rectal tumor response evaluation, those incorporating DWI showed the best results when combining diagnostic performance, IOA, and reader preference.
- The most preferred method (by 73% of readers) was the “DWI patterns” approach with an accuracy of 68%, high specificity of 82%, and group IOA of 0.43.
- Reader experience level and MRI quality had an evident effect on diagnostic performance and IOA.

Keywords Rectal neoplasms · Magnetic resonance imaging · Neoplasm, residual · Chemoradiotherapy

✉ Doenja M. J. Lambregts
d.lambregts@nki.nl

¹ Department of Radiology, The Netherlands Cancer Institute, P.O. Box 90203, 1106, BE Amsterdam, The Netherlands

² Present address: GROW School for Oncology & Developmental Biology, University of Maastricht, Maastricht, The Netherlands

³ Department of Epidemiology and Biostatistics, The Netherlands Cancer Institute, Amsterdam, The Netherlands

⁴ Department of Radiology, Centro Hospitalar e Universitario de Coimbra EPE, Faculty of Medicine, University of Coimbra, Coimbra, Portugal

⁵ Department of Radiology, University Hospitals Leuven, Leuven, Belgium

⁶ Medical Imaging Department, Montpellier Cancer Institute, Montpellier Cancer Research Institute (U1194), University of Montpellier, Montpellier, France

⁷ Department of Surgery, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Abbreviations

CRT	Chemoradiotherapy
DWI	DIFFUSION-weighted imaging
IOA	Interobserver agreement
mrTRG	MRI tumor regression grade
NPV	Negative predictive value
PPV	Positive predictive value
W&W	Watch & Wait

Introduction

The standard treatment for locally advanced rectal cancer is neoadjuvant chemoradiation (CRT) followed by surgery [1]. Nowadays, there is a paradigm shift to opt for organ-preserving treatment alternatives in patients who respond very well to CRT. Patients with clinical evidence of a complete response after CRT may be entered into a Watch & Wait (W&W) program where patients are deferred from surgery and closely monitored using a combination of imaging and endoscopy. The International Watch & Wait Database (IWWD) recently published the oncologic outcomes of the first 1000 registered W&W patients, showing a good 5-year overall and disease-free survival of 85–94% [2].

The introduction of W&W and other organ-preservation strategies has urged the need for accurate response assessment after CRT to facilitate the patient selection. MRI has an important role in detecting the presence of extraluminal residual disease (e.g. remaining positive lymph nodes) that may render organ preservation unfeasible. MRI is also used as an adjunct to endoscopy to assess the response of the primary tumor in the bowel wall. The diagnostic performance of MRI in this setting is limited owing to difficulties in interpreting fibrotic changes of the tumor bed after CRT [3, 4].

Different methods have been published to address this issue and aid in visually classifying tumor response on MRI after CRT. One of the most well-known is the MRI tumor regression grade (mrTRG), derived from similar TRG scores used in histopathology [5]. The mrTRG can help radiologists classify the degree of fibrotic transformation of the tumor bed on T2-weighted (T2W) MRI to estimate the tumor response [6–9]. Since the introduction and recognition of diffusion-weighted imaging (DWI) sequences as a valuable adjunct to discern viable residual tumor from fibrosis, modified response systems have been reported that combine tumor regression on T2W-MRI with DWI findings [10–12]. Other published methods focus on specific MRI patterns or “signs”. These include the “DWI patterns” approach of Lambregts et al [13], which combines morphological patterns on pre- and post-CRT T2W-MRI with distinct DWI signal patterns post-CRT to differentiate complete responders, and the “split scar” sign published by Santiago et al [14] that describes a typical layered appearance of the tumor bed on T2W-MRI after CRT

(referred to as the “split scar”) as a sign indicating a complete response.

Most of these response methods were published fairly recently. So far they have mainly been tested by expert readers in single-center study settings. Little is known about how well these methods can be reproduced in daily clinical practice, using less curated datasets, and by radiologists with more general expertise.

Therefore, this study aims to validate and compare the above-described methods to assess response after CRT on restaging MRI using a multicentre dataset of clinical MRIs derived from everyday practice, taking into account diagnostic performance, agreement among readers with different expertise levels, and reader preference.

Methods

Patient selection

This study was conducted as a side-study of an institutional review board-approved retrospective multicentre study on multiparametric imaging for tumor response evaluation in locally advanced rectal cancer. Due to the retrospective nature of the study, informed consent was waived. As part of this study, the imaging and clinical outcome data of 1037 patients (2010–2018) were retrospectively collected from 10 centers in the Netherlands, including 1 university hospital, 8 large teaching hospitals, and 1 comprehensive cancer center.

For the current study, we selected from this cohort a semi-random sample of $n = 90$ patients taking into consideration that data of all 10 study centers had to be represented in the cohort and ensuring a clinically representative sample in terms of baseline cTN-stage and response outcomes with a sufficient number of complete responders (being the primary study outcome) to allow meaningful statistical analyses. Study inclusion and exclusion criteria are detailed in Fig. 1.

MRI protocols and quality assessment

MRI examinations were performed according to the local protocols of the participating centers at the time of inclusion. From the full available protocols, we selected a basic set of sequences (representing the main sequences required for rectal cancer restaging according to current guidelines [15]) consisting of 2D-T2W sequences in sagittal, oblique-axial (perpendicular to the tumor axis), and oblique-coronal (parallel to the tumor axis) planes, and an oblique-axial DWI sequence with corresponding apparent diffusion coefficient (ADC) map. T2W slice thickness ranged between 3 and 5 mm and in-plane resolution ranged between 0.35 x 0.35

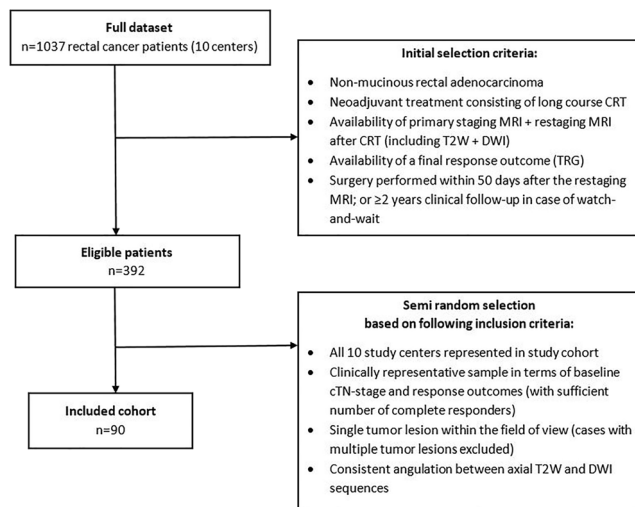


Fig. 1 Patient in- and exclusion flowchart

and 0.94 x 0.94 mm. The DWI sequence included at least one high b -value ranging between 600 and 1200 s/mm².

MR image quality was graded by one of the authors (N.E.K., who did not participate in the further study readings) using a 6-point scoring system developed for the purpose of this study. This scoring was based on current guidelines, other published recommendations on image acquisition, and on previously published scoring systems to grade DWI image quality [15–18], as detailed below:

- Transverse sequence angulation[15]: 0 = not perpendicular to longitudinal tumor axis
1 = perpendicular to longitudinal tumor axis
- T2W slice thickness [15]: 0 = > 3 mm
1 = ≤ 3 mm
- T2W in-plane resolution [16]: 0 = > 0.6 x 0.6 mm
1 = ≤ 0.6 x 0.6 mm
- DWI high b -value[15]: 0 = < 800 s/mm²
1 = ≥ 800 s/mm²
- DWI signal to noise ratio [18]: 0 = Poor – intermediate
1 = Good – excellent
- DWI artefacts [17]: 0 = moderate-severe, hampering interpretation
1 = no-minimal, not hampering interpretation

Scans with a score of $\geq 4/6$ were classified as good quality; scans with a score of $< 4/6$ as below average quality.

Image evaluation

An invitation to participate in the study was sent out to the members of the European Society of Gastrointestinal and Abdominal Radiology (ESGAR) (in particular members with an interest in rectal imaging). This rendered an international group of twenty-two radiologists, including 5 rectal MRI experts (each with ≥ 10 years of dedicated experience in rectal MRI) and 17 abdominal radiologists (or general radiologists with a specific interest in abdominal imaging). Image evaluation was performed using an in-house developed web-based viewing platform (iScore) with embedded electronic case report forms (eCRF) that were designed for the purpose of this study by one of the authors (N.E.K.). The iScore platform incorporates the Open Health Imaging Foundation (OHIF) DICOM viewing platform [19]. An overview of the scoring setup in iScore including the full eCRFS is provided in Supplement 1

The study readers were asked to review the restaging MRIs (T2W, DWI, and ADC map) of the 90 study cases by comparing them to the primary staging MRIs and assessing the response to chemoradiotherapy using four different previously published response methods: mrTRG [6, 8], modified mrTRG [10, 11], DWI patterns score [13], and the split scar sign [14]. Details of these four scoring methods and how they were dichotomized are provided in Table 1. Readers were asked to indicate for each case whether they found the respective scoring methods easy, moderately easy/difficult, or difficult to apply; and to give an overall indication of which scoring method(s) they would prefer to apply in their own daily clinical practice. Readers were blinded to each other's scorings and to the final response outcomes.

Standard of reference

The main study outcome was the differentiation between a complete response and residual tumor, using the pathologic tumor regression grade (pTRG) by Mandard [5] or clinical follow-up during organ preservation as the standard of reference. A complete response was defined as ypT0/pTRG1 after surgery, or a sustained clinical complete response during W&W for at least 2 years. Residual tumor was defined as ypT1-4/pTRG2-5 after surgery.

Statistical analyses

Statistical analyses were performed by one of the authors, a dedicated statistician (R.T.) using R statistics version 4.1.0 (2021) and IBM SPSS version 27 (2020). To assess the impact of reader experience (MRI expert versus abdominal/general

Table 1 Methods of response evaluation

Method	MR sequence(s)	Score	Dichotomized score	
			Complete response	Residual tumor
mrTRG	T2W-MRI (post-CRT)	<ul style="list-style-type: none"> • 1 = no/minimal fibrosis • 2 = dense fibrotic scar without macroscopic tumor signal • 3 = fibrosis predominates but there are obvious areas of tumor signal • 4 = tumor signal predominates with little/minimal fibrosis • 5 = tumor signal only: no fibrosis, includes cases with the progression of tumor 	1–2*	3–5
Modified TRG	T2W-MRI + DWI (post-CRT)	<ul style="list-style-type: none"> • 0 = No tumor signal on T2W-MRI, no diffusion restriction (complete regression) • 1 = Predominant fibrosis on T2W-MRI, focal diffusion restriction (intermediate regression) • 2 = Predominant tumor on T2W-MRI, focal or mass-like diffusion restriction (poor regression) 	0	1–2
DWI pattern	T2W-MRI (pre and post-CRT) + DWI (post-CRT)	<ul style="list-style-type: none"> • A - = normalized rectal wall, no diffusion restriction • A+ = bulky solid residual tumor mass with corresponding diffusion restriction • B = circular/irregular tumors (pre-CRT) with irregular/spiculated fibrosis (post-CRT) <ul style="list-style-type: none"> B- = no diffusion restriction B+ = scattered foci of diffusion restriction • C = semicircular tumors (pre-CRT) with semicircular/focal fibrosis (post-CRT) <ul style="list-style-type: none"> C- = no diffusion restriction C+ = focal diffusion restriction at the inner margin of fibrosis • D = polypoid tumor (pre-CRT) with focal fibrosis at the site of polyp stalk (post-CRT) <ul style="list-style-type: none"> D- = no diffusion restriction D+ = focal diffusion restriction at the site of the stalk 	A-^ C- D-	A+ B+/ C+ D+
Split scar	T2W-MRI (post-CRT)	<ul style="list-style-type: none"> • 0 = Split scar present • 1 = Split scar absent 	0	1

*Cut off chosen based on results of a meta-analysis by Jang et al assessing the mrTRG to diagnose a pathologic complete response, which showed the highest sensitivity for mrTRG1-2 (*Eur Radiol* 2020;30(4):2312-2323)

^ Based on cut-offs recommended in the original publication by Lambregts et al describing the DWI pattern approach (*Dis Colon Rectum* 2018; 61(3):328-337)

radiologist) and MR image quality (good versus below average) on the average sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of each method to predict a complete response (= positive outcome) versus residual tumor a mixed model linear regression was used. Results were additionally compared using receiver operator characteristics (ROC) curves. A patient-level random intercept was used to take into account the repeated measurements of each patient. A significance threshold of 0.05 was used throughout the analyses. Interobserver agreement (IOA) between individual readers was calculated using kappa analysis (κ) [20] with quadratic kappa weighting; group agreement was calculated using Krippendorff's alpha [21, 22].

Results

Baseline characteristics

Baseline patient and study reader data are provided in Table 2. Fifty-two patients (58%) were 52 male, and mean age was 65

\pm 11 years. Twenty-seven patients (30%) were complete responders. The 22 study readers originated from fourteen different countries.

Diagnostic performance and effects of reader experience and image quality

Table 3 shows the average diagnostic performance for the four response methods to discern complete responders from patients with residual tumor, including sub-analyses comparing results for experts versus non-expert readers and for scans with optimal versus below-average image quality. The mrTRG showed the lowest specificity (64% vs. 79–82% for the other methods; $p < 0.001$) but the highest sensitivity (57% vs. 36–40%; $p < 0.001$). NPV was significantly higher ($p = 0.04$) and overall accuracy was significantly lower for mrTRG ($p < 0.001$) compared to the other methods. Overall accuracy ranged between 62 and 68%, with higher accuracy (70–74%) for the expert readers, except for the split scar sign where no significant differences were observed. The area under the ROC-curve (incl. 95%

Table 2 Patient and reader characteristics

Patient characteristics		N=	%
Total		90	100%
Mean age (\pm SD)		65 (\pm 11)	
Sex			
	Male	52	58%
	Female	38	42%
Baseline stage (MRI)			
cT-stage	cT1-2	3	3%
	cT3	68	76%
	cT4	18	20%
cN-stage	cN0	12	13%
	cN1	19	21%
	cN2	59	66%
Final response (pathology)			
yT-stage	yT0 ¹	27	30%
	yT1-2	22	24%
	yT3	37	41%
	yT4	4	4%
yN-stage	yN0 ¹	65	72%
	yN1	17	19%
	yN2	8	9%
TRG	TRG1 ¹	27	30%
	TRG2	17	19%
	TRG3	29	32%
	TRG4	15	17%
	TRG5	2	2%
Complete response vs. residual tumor	Complete response (TRG1 / ypT0)	27	30%
	Residual tumor (TRG2-5 / ypT1-4)	63	70%
Reader characteristics		N=	%
Total		22	100%
Experience			
	Experts	5	23%
	Abdominal/general radiologists	17	77%
Workplace			
	Comprehensive cancer center	8	36%
	University hospital	8	36%
	General hospital	3	14%
	Other	3	14%
Country			
	The Netherlands	4	18%
	United Kingdom	4	18%
	Italy	2	9%
	Switzerland	2	9%
	India	1	5%
	Israël	1	5%
	Denmark	1	5%
	Germany	1	5%
	Portugal	1	5%
	France	1	5%
	Canada	1	5%
	Brazil	1	5%
	Chile	1	5%
	Georgia	1	5%

¹ based on histology after surgery in 21 patients and on a sustained clinical complete response during W&W with > 2 years of clinical follow-up in the remaining 6 patients

Table 3 Diagnostic performance to detect a complete response with specified results demonstrating effects of reader experience level and image quality

		Sensitivity		Specificity		PPV		NPV		Accuracy		
mrTRG	Average (all readers)	57%		64%		44%		77%		62%		
	Expert readers	Non-expert readers	55%	58%	78%	60%	54%*	41%*	80%	77%	71%*	59%*
	Optimal quality	Below average quality	54%	60%	68%#	60%#	45%	43%	79%	76%	64%#	60%#
modTRG	Average (all readers)	40%		79%		50%		75%		67%		
	Expert readers	Non-expert readers	34%	42%	90%*	76%*	63%*	46%*	76%	75%	73%*	66%*
	Optimal quality	Below average quality	34%#	46%#	85%#	74%#	52%	47%	76%	75%	70%#	65%#
DWI patterns	Average (all readers)	37%		82%		52%		75%		68%		
	Expert readers	Non-expert readers	36%	37%	90%*	79%*	67%*	48%*	77%	75%	74%*	67%*
	Optimal quality	Below average quality	31%#	43%#	88%#	76%#	57%#	48%#	76%	74%	71%#	65%#
Split scar	Average (all readers)	36%		79%		46%		75%		66%		
	Expert readers	Non-expert readers	26%	39%	89%	76%	51%	45%	74%	75%	70%	65%
	Optimal quality	Below average quality	36%	37%	84%#	74%#	50%	43%	77%	73%	70%#	63%#

Notes:

- Results were calculated using a complete response as the positive outcome and residual tumor as the negative outcome
- Expert readers ($n = 5$) were MRI experts with ≥ 10 years of dedicated experience in rectal MRI; non-expert readers ($n = 17$) were abdominal radiologists or general radiologists with a specific interest in abdominal imaging. MR image quality was categorized as optimal in $n = 52$ cases and as below average in the remaining $n = 38$ cases.
- Results printed in **boldface** indicate a significant effect size as assessed using mixed model linear regression, with * indicating a significant difference in diagnostic performance between expert and non-expert readers, and # indicating a significant difference in diagnostic performance between scans with optimal and below-average image quality. Effect sizes, confidence intervals, and levels of statistical significance (p -values) are further detailed in Supplement 2

confidence interval) was 0.72 (0.60–0.83) for mrTRG, 0.69 (0.57–0.91) for modified mrTRG, 0.68 (0.55–0.81) for DWI patterns, and 0.74 (0.63–0.85) for the split scar; differences between the four techniques were not statistically significant ($p = 0.17$ – 0.94). Scans with below-average imaging quality had a negative impact on diagnostic performance. Detailed effect sizes and levels of significance are provided in Supplement 2. Selected imaging examples demonstrating the effects of reader experience and image quality are provided in Figs. 2 and 3.

Interobserver agreement and reader preference

Table 4 shows the median IOA (κ), specified results for expert and non-expert readers, and the difficulty and preference scores assigned by the various readers. Detailed IOA results between individual readers are provided in Supplement 3. Group IOA (Krippendorff's alpha) for all readers was 0.39 (mrTRG), 0.40 (modTRG), 0.43 (DWI patterns), and 0.18 (Split scar). Overall, IOA was higher for the expert readers, except for the split scar sign which showed similarly low IOA for all readers. Most readers selected scoring systems incorporating DWI (modified TRG, DWI pattern) as their preferred response method (selected by 68–73% vs. 5–18% for the

mrTRG and split scar). This preference was also reflected in the difficulty scores.

Discussion

This study aimed to validate and compare four previously published methods for rectal tumor response evaluation on MRI after chemoradiotherapy in terms of diagnostic performance to identify complete responders, inter-reader reproducibility, and reader preference. Overall, the most favorable results were found for response methods incorporating DWI, considering their good specificity of $\pm 80\%$, highest overall interobserver agreement, and the fact that the majority of readers preferred the DWI-based methods over the methods based solely on T2W-MRI. Diagnostic performance and interobserver agreement were lower for less expert readers and when MRI image quality was below current clinical standards. These findings emphasize the need for good-quality imaging using state-of-the-art MRI protocols, and the importance of dedicated radiologist training to evaluate restaging MRIs.

The two preferred methods incorporating DWI (the modified mrTRG score and the DWI patterns score) showed a

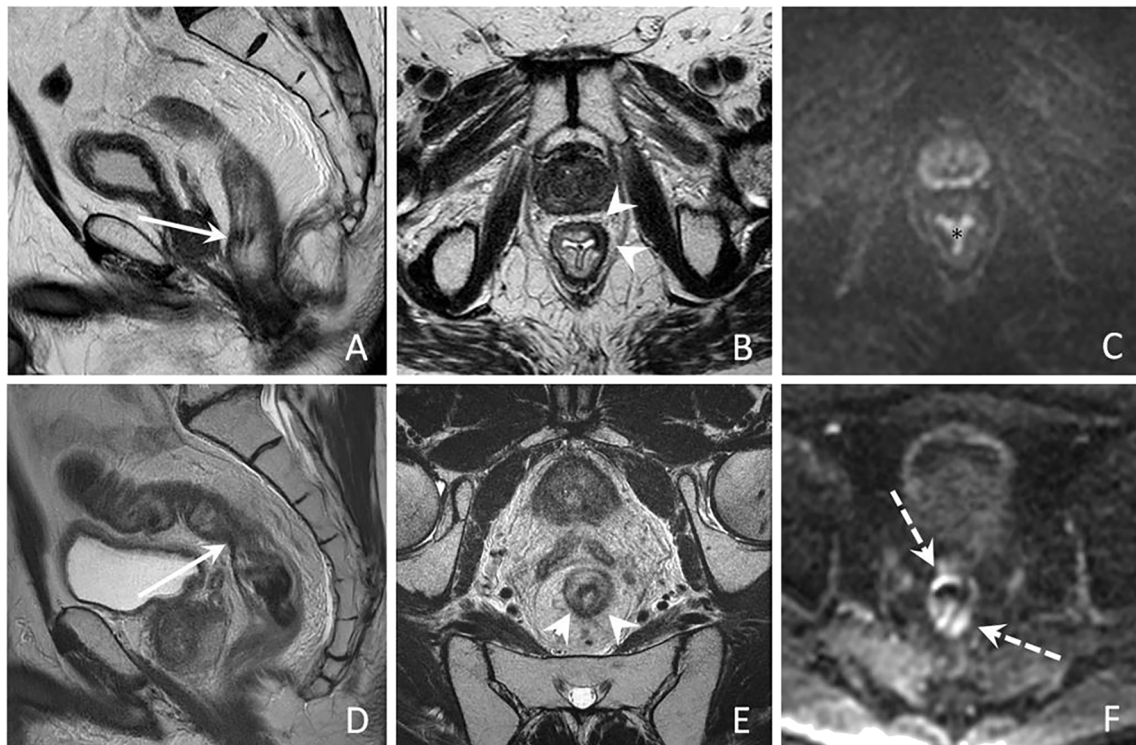


Fig. 2 Impact of image quality. Upper row shows the sagittal (A) and axial (B) T2-weighted images and DWI scan (C) of a male patient with an MRI that was graded as good quality. The fibrotic tumor bed is indicated by the white arrow in A and arrowheads in B. * indicates shine through of luminal fluid with a typical star shape on DWI (C). The majority of readers scored this case as a complete response with an mrTRG score of 1–2, no suspicious signal on DWI (modified TRG score 0; DWI

pattern C-) and with a positive split scar sign. The bottom row shows the T2-weighted (D; E) and DWI (F) images of a male patient where image quality was graded as below average. Much variation was observed between readers: mrTRG scores ranged from 2 to 5, modified TRG scores ranged from 0 to 2, DWI pattern scores included A+, B-, B+, and C+, and a split scar sign was detected by 1 out of 22 readers

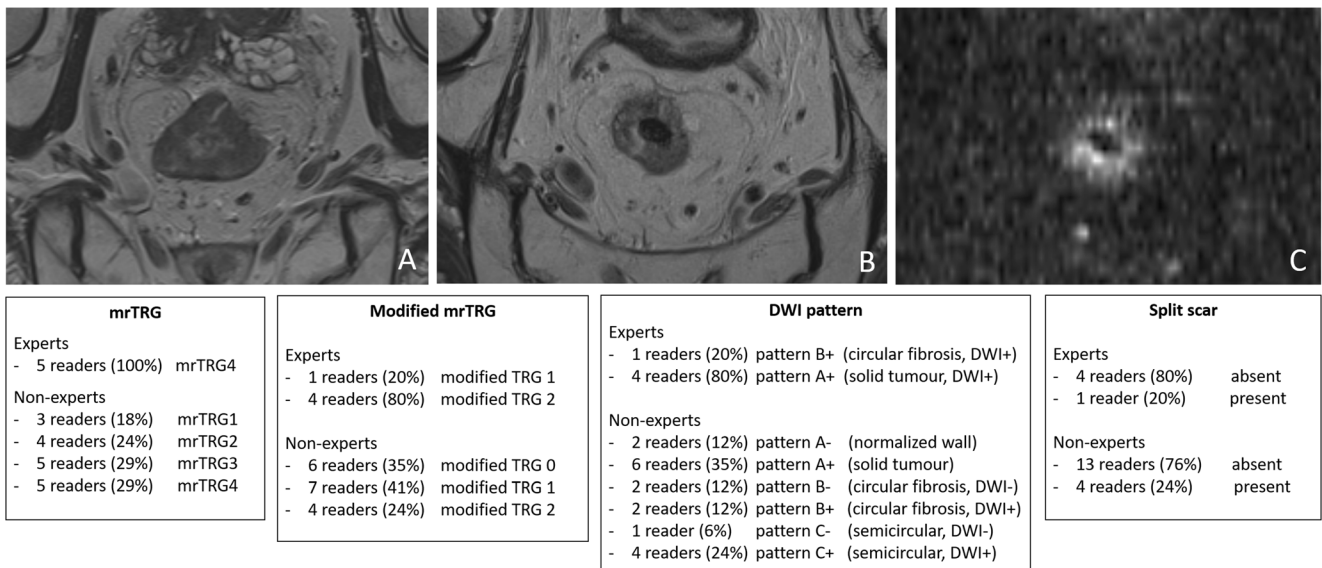


Fig. 3 Impact of reader experience. Pre-treatment axial T2-weighted (A) and post-CRT axial T2-weighted (B) and DWI (C) images of a male patient with a scan graded as good quality. On pre-treatment MRI, a circular tumor lesion is seen. Post-CRT, predominant tumor signal remains with a persistent high signal on DWI. As detailed in the boxes

below the images, the expert readers reached 80–100% agreement for the different scoring methods; for the less-experienced general and abdominal radiologists, agreement for the different scoring methods was much lower, ranging from 29 to 76%

Table 4 Interobserver agreement and reader preference

	mrTRG	modTRG	DWI pattern	Split scar
IOA (κ ; median with ranges in parentheses)				
All readers ($n = 22$)	0.41 (0.15–0.66)	0.42 (0.09–0.68)	0.48 (0.1–0.77)	0.17 (–0.07 to 0.6)
Expert readers ($n = 5$)	0.55 (0.45–0.66)	0.54 (0.42–0.64)	0.60 (0.54–0.77)	0.18 (0.02–0.33)
Non-expert readers ($n = 17$)	0.41 (0.15–0.63)	0.40 (0.09–0.68)	0.47 (0.1–0.71)	0.17 (–0.07 to 0.6)
Difficulty to apply response method (%)				
Easy	42%	49%	55%	43%
Moderate	45%	42%	36%	37%
Difficult	13%	9%	9%	20%
Preferred response method (%)				
	18%	68%	73%	5%

higher specificity compared to the two methods based solely on T2W-MRI (mrTRG and split scar). This implies a better performance for DWI-MRI to detect residual tumor within the fibrotic tumor bed, which is known to be one of the key strengths of DWI in the restaging setting and an important issue when aiming to safely select patients for W&W [16]. Specificity was particularly high (up to 90%) for the expert readers, with results comparable to the initial study publications [10, 13]. Sensitivity for both DWI-based scoring methods ($\pm 40\%$) was however lower than in the initial reports. This indicates a risk that complete responders are wrongly classified as having residual tumor due to the presence of non-tumor (“false positive”) high signal on DWI, which is a known limitation of DWI [3, 12]. When relying on DWI for clinical decision-making, steps should be taken to optimize DWI image quality, such as giving patients a preparatory micro-enema or adapting acquisition protocols to make the DWI sequence less susceptible to artefacts [17, 23–25].

Out of the four investigated methods, the mrTRG has been studied the most in previous literature. In a recent meta-analysis including six studies and a total of 916 patients, pooled sensitivity to diagnose a complete response using a mrTRG score of 1–2 was somewhat higher than in our current report (70% vs. 57%) [7]. Interestingly, sensitivity for mrTRG in our study was higher than for the other three methods (57% vs. 36–40%), suggesting a better performance for mrTRG in identifying complete responders with a lower risk of overcalling the presence of a residual tumor. The specificity of 62% for mrTRG in our study was comparable to that reported in the previous meta-analysis (64%) [7], but lower compared to the other three methods under evaluation (specificity 79–82%), indicating a higher risk of missing residual tumor. Notably, the mrTRG—despite being probably the most well-known method out of the four—was selected as the preferred response method by only 18% of our study readers.

The fourth method under evaluation was the split scar sign, proposed by Santiago et al [14]. The split scar sign describes a particular morphologic appearance of the tumor bed (scar) after CRT which gives the rectal wall a characteristic layered appearance. In the original publication with two readers, a higher sensitivity of 52–64% was reported compared to the average sensitivity of only 36% for the 22 readers in our current study. The average specificity in our current study was 79%, versus 97% in the original publication. Overall accuracy for the split scar sign in our current study was similar to that of the other three methods. However, it was clearly the least preferred scoring method amongst the study readers. In up to 20% of cases, our readers experienced difficulties in assessing the split scar sign, and a positive split scar sign was recognized in only a very small minority of the cases. Several of our readers furthermore noted that the split scar sign was not applicable in cases with a complete response without any visible fibrosis. Santiago et al stated explicitly in their publication that high-resolution T2W imaging is required for the evaluation of the split scar sign. A substantial number of scans in our cohort were acquired with a slice thickness of > 3 mm and/or limited in-plane resolution. This suggests that out of the four response methods, the split scar sign may be the most influenced by T2W scan quality and therefore more challenging to reproduce in a heterogeneous clinical dataset with less optimized acquisition protocols.

With respect to the interobserver agreement, results were comparable for the mrTRG, modified TRG, and DWI pattern approach, with median kappa’s ranging between 0.41 and 0.48 (with the highest scores for the DWI pattern score). Agreement for the split scar sign was considerably lower with a median kappa of 0.17, which is also much lower than the $\kappa 0.69$ reported in the initial paper by Santiago et al. This will likely again be related to difficulties in applying this method in a heterogeneous dataset, but perhaps also to the fact that out of all methods, readers may be least familiar with the split scar

sign. Compared to previous publications, IOA for the other 3 methods was similar or also somewhat lower. For example, Siddiqi et al reported a median IOA of $\kappa 0.57$ for 35 radiologists in applying the mrTRG in a small group of 12 patient cases [6], compared to a median $\kappa 0.41$ in our current report with a considerably larger number of patient cases. Previously reported IOAs for the modified TRG and DWI pattern scores ranged between $\kappa 0.58$ and 0.75 [10, 13]. Results for the more experienced readers in our current study were in the same range, with kappa's varying between 0.42 and 0.77.

Since the MRIs in our dataset date back as far as 2010, several scans did not meet current state-of-the-art recommendations for image acquisition. These “below-average” quality scans had a negative impact on our study results, and also offered us valuable insights into the importance of standardized scan quality. There are some other limitations to our study design. First, selection bias may have occurred as scans were semi-randomly selected from a larger dataset as detailed in the methods section. For the sake of feasibility, the number of cases was kept < 100 , which is low compared to the number of study readers. Second, the four methods addressed in this study focus specifically on luminal response assessment. From a clinical perspective, MRI mainly has a supporting role (in addition to endoscopy) for luminal response assessment when selecting patients for and monitoring them during organ preservation [11, 26]. Though we acknowledge that one of the main strengths of MRI is the assessment of extraluminal disease (e.g. lymph nodes), assessing its value in this setting was outside the scope of our study, as was the assessment of MRI for follow-up during organ preservation. Third, the comparison of the four scoring methods may be somewhat biased in the sense that some (DWI patterns, split scar) are designed specifically for the differentiation between a complete response and residual tumor, while others are intended to grade the overall response and were dichotomized for the purpose of this study. Moreover, the number of response categories differs between the different methods. The degree to which readers were already accustomed to using the respective methods prior to the study will also likely have varied, though this is also reflective of variations between countries and centers in daily reporting practice. Fourth, the readers had access to all available images while performing their scorings. Though readers were instructed to only review the T2W images when evaluating the mrTRG and split scar, we cannot rule out that readers were biased by the findings of DWI. Finally, all MRI exams included in this study originate from the Netherlands. Considering that Dutch guidelines for image acquisition are in line with international guidelines, we are confident that the dataset offers a representative sample including a representation of all commonly used MRI vendors and protocol variations reflective of everyday clinical practice in the Netherlands as well as worldwide.

In conclusion, this multireader and multicentre validation study has shown that out of four previously published methods for rectal tumor response evaluation after CRT (mrTRG, modified mrTRG, DWI patterns, and the split scar sign), the methods incorporating DWI showed the most favorable overall results taking into account its high specificity, interobserver agreement, and strong reader preference. Both reader experience and image quality had a clear impact on diagnostic performance and interobserver agreement, which emphasizes the need for good quality imaging using state-of-the-art MRI protocols, and the importance of dedicated radiologist training to gain sufficient expertise.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00330-022-09342-w>.

Acknowledgments

Other authors in the study group:

Frans C. H. Bakers¹; Perla Barros²; Ferdinand Bauer³; Shira H. de Bie⁴; Stuart Ballantyne⁵; Joanna Brayner Dutra^{6,7}; Laura Buskov⁸; Nino Bogveradze^{9,10,11}; Gerlof P. T. Bosma¹²; Vincent C. Cappendijk¹³; Francesca Castagnoli¹⁴; Sotiriadis Charalampos¹⁵; Andrea Delli Pizzi¹⁶; Michael Digby¹⁷; Remy W. F. Geenen¹⁸; Joost J. M. van Griethuysen^{9,19}; Julie Lafrance²⁰; Vandana Mahajan²¹; Sonaz Malekzadeh²²; Peter A. Neijenhuis²³; Gerald M. Peterson²⁴; Indra Pieters²⁵; Niels W. Schurink^{9,10}; Ruth Smit²⁶; Cornelis J. Veeken²⁷; Roy F. A. Vliegen²⁸; Andrew Wray²⁹; Abdel-Rauf Zeina³⁰

¹Department of Radiology and Nuclear Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands

²Department of Radiology, Instituto Oncológico Fundación Arturo López Pérez, Santiago, Chile

³Radiologie Zentrum, Kaufbeuren, Germany

⁴Department of Radiology, Deventer Ziekenhuis, The Netherlands

⁵Department of Radiology, Queen Elizabeth University Hospital, Glasgow, United Kingdom

⁶Department of Radiology, Real Hospital Portugues (RHP), Pernambuco, Brazil

⁷Department of Radiology, Instituto de Medicina Integral Professor Fernando Figueira (IMIP), Recife, Brazil

⁸Department of Radiology, Bispebjerg Hospital, Copenhagen, Denmark

⁹Department of Radiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

¹⁰GROW School for Oncology & Developmental Biology, University of Maastricht, Maastricht, The Netherlands

¹¹Department of Radiology, American Hospital Tbilisi, Tbilisi, Georgia

¹²Department of Radiologie Elisabeth Tweesteden Hospital, Tilburg, The Netherlands

¹³Department of Radiology, Jeroen Bosch Hospital, 's Hertogenbosch, The Netherlands

¹⁴Department of Radiology, University of Brescia, Brescia, Italy

¹⁵Department of Radiology, Hôpital Riviera Chablais, Rennaz, Switzerland

¹⁶Department of Innovative Technologies in Medicine & Dentistry, Gabriele d'Annunzio University of Chieti, Chieti, Italy

¹⁷Department of Radiology, Glasgow Royal Infirmary, Glasgow, United Kingdom.

¹⁸Department of Radiology, Northwest Clinics, Alkmaar, The Netherlands

¹⁹Department of Radiology, Gelre Hospital, Apeldoorn, The Netherlands

²⁰Department of Radiology, Maisonneuve-Rosemont Hospital, Montreal, Canada

²¹Department of Radiology, Apollo Cancer Hospital, Chennai, India

²²Department of Radiology, Sion Hospital, Sion, Switzerland

²³Department of Surgery, Alrijne Hospital, Leiderdorp, The Netherlands

²⁴Department of Radiology, Spaarne Gasthuis, Haarlem, The Netherlands

²⁵Department of Radiology, Telemedicine Clinic, United Kingdom

²⁶Department of Radiology, Amsterdams UMC, Amsterdam, The Netherlands

²⁷Department of Radiology, IJsselland Hospital, Capelle aan den IJssel, The Netherlands

²⁸Department of Radiology, Zuyderland Medical Center, Heerlen, The Netherlands

²⁹Department of Radiology, Ulster Hospital, Belfast, United Kingdom

³⁰Department of Radiology, Hillel Yaffe Medical Center, Hadera, Israel

Funding The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Dr Doenja MJ Lambregts.

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, Mr Renaud Tissier, has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Some study subjects included in the current cohort have been previously reported on the following:

$n = 90$ in a study focused on retrospectively evaluating staging trends in the Netherlands following guidelines updates (Bogverdze et al *Abdom Radiol (New York)*. 2022;47(1):38–47).

$n = 11$ in a study focused on common interpretation pitfalls in rectal DWI and their use for teaching (Lambregts et al *Eur Radiol* 2017; 27, 4445–4454)

$n = 80$ in a technical study focused on assessing the reproducibility of quantitative imaging features in multicentre study cohorts (Schurink et al *Eur Radiol*. 2022;32(3):1506–1516).

$n = 6$ in a study focused on assessing the sigmoid take-off as a landmark to distinguish rectal from sigmoid tumors on MRI (Bogverdze et al *Eur J Surg Oncol* 2022;48:237–244)

$n = 16$ in a single-center pilot study investigating the DWI pattern method (Lambregts et al *Dis Colon Rectum* 2018;61(3):328–337).

Methodology

- retrospective
- observational
- multicentre study

References

1. Oronsby B, Reid T, Larson C, Knox SJ (2020) Locally advanced rectal cancer: the past, present, and future. *Semin Oncol* 47(1):85–92
2. van der Valk MJM, Hilling DE, Bastiaannet E et al (2018) Long-term outcomes of clinical complete responders after neoadjuvant treatment for rectal cancer in the International Watch & Wait Database (IWWD): an international multicentre registry study. *Lancet* 391(10139):2537–2545
3. Lambregts D, Boellaard TN, Beets-Tan R (2019) Response evaluation after neoadjuvant treatment for rectal cancer using modern MR imaging: a pictorial review. *Insights Imaging* 10(1):15
4. Barbaro B, Vitale R, Leccisotti L et al (2010) Restaging locally advanced rectal cancer with MR imaging after chemoradiation therapy. *Radiographics* 30(3):699–716
5. Mandard AM, Dalibard F, Mandard JC et al (1994) Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. *Cancer* 73(11):2680–2686.tumor
6. Siddiqui MR, Gornly KL, Bhoday J et al (2016) Interobserver agreement of radiologists assessing the response of rectal cancers to preoperative chemoradiation using the MRI tumor regression grading (mrTRG). *Clin Radiol* 71(9):854–862
7. Jang JK, Choi SH, Park SH et al (2020) MR tumor regression grade for pathological complete response in rectal cancer post neoadjuvant chemoradiotherapy: a systematic review and meta-analysis for accuracy. *Eur Radiol* 30(4):2312–2323.tumor
8. Bhoday J, Smith F, Siddiqui MR et al (2016) Magnetic resonance tumor regression grade and residual mucosal abnormality as predictors for pathological complete response in rectal cancer postneoadjuvant chemoradiotherapy. *Dis Colon Rectum* 59(10):925–933.tumor
9. Patel UB, Brown G, Rutten H et al (2012) Comparison of magnetic resonance imaging and histopathological response to chemoradiotherapy in locally advanced rectal cancer. *Ann Surg Oncol* 19(9):2842–2852
10. Lee MA, Cho SH, Seo AN et al (2017) Modified 3-point MRI-based tumor regression grade incorporating DWI for locally advanced rectal cancer. *AJR Am J Roentgenol* 209(6):1247–1255.tumor
11. Haak HE, Maas M, Lahaye MJ et al (2020) Selection of patients for organ preservation after chemoradiotherapy: MRI identifies poor responders who can go straight to surgery. *Ann Surg Oncol* 27(8):2732–2739
12. Schurink NW, Lambregts D, Beets-Tan R (2019) Diffusion-weighted imaging in rectal cancer: current applications and future perspectives. *Br J Radiol* 92(1096):20180655
13. Lambregts DMJ, Delli Pizzi A, Lahaye MJ et al (2018) A pattern-based approach combining tumor morphology on MRI with distinct signal patterns on diffusion-weighted imaging to assess response of rectal tumors after chemoradiotherapy. *Dis Colon Rectum* 61(3):328–333.tumortumor
14. Santiago I, Barata M, Figueiredo N et al (2020) The split scar sign as an indicator of sustained complete response after neoadjuvant therapy in rectal cancer. *Eur Radiol* 30(1):224–238

15. Beets-Tan RGH, Lambregts DMJ, Maas M et al (2018) Magnetic resonance imaging for clinical management of rectal cancer: updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. *Eur Radiol* 28(4):1465–1475
16. Gormly K (2021) Rectal MRI: the importance of high resolution T2 technique. *Abdom Radiol (NY)* 46(9):4090–4095
17. van Griethuysen JJM, Bus EM, Hauptmann M et al (2018) Gas-induced susceptibility artefacts on diffusion-weighted MRI of the rectum at 1.5 T - effect of applying a micro-enema to improve image quality. *Eur J Radiol* 99:131–137
18. Van Griethuysen JJM, Schurink NW, Lahaye MJ et al (2020) Deep learning for fully automated segmentation of rectal tumors on MRI in a multicentre setting. *Insights into imaging – ESGAR book of abstracts* 11(Suppl 3):SS5.4
19. Ziegler E, Urban T, Brown D et al (2020) Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research. *JCO Clin Cancer Inform* 4:336–345
20. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20(1):37–46
21. Antoine, J. Y., Villaneau, J., & Lefevre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. *EACL 2014* (pp. 10-p).
22. Hayes AF, Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 1(1): 77–89
23. Nasu K, Kuroki Y, Kuroki S, Murakami K, Nawano S, Moriyama N (2004) Diffusion-weighted single shot echo planar imaging of colorectal cancer using a sensitivity-encoding technique. *Jpn J Clin Oncol* 34(10):620–626
24. Korn N, Kurhanewicz J, Banerjee S, Starobinets O, Saritas E, Noworolski S (2015) Reduced-FOV excitation decreases susceptibility artifact in diffusion-weighted MRI with endorectal coil for prostate cancer detection. *Magn Reson Imaging* 33(1):56–62
25. Thian YL, Xie W, Porter DA, Weileng AB (2014) Readout-segmented echo-planar imaging for diffusion-weighted imaging in the pelvis at 3T-a feasibility study. *Acad Radiol* 21(4):531–537
26. Maas M, Lambregts DM, Nelemans PJ et al (2015) Assessment of clinical complete response after chemoradiation for rectal cancer with digital rectal examination, endoscopy, and MRI: selection for organ-saving treatment. *Ann Surg Oncol* 22(12):3873–3880

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.