HHS Public Access

Author manuscript

Mult Scler Relat Disord. Author manuscript; available in PMC 2024 May 14.

Published in final edited form as:

Mult Scler Relat Disord. 2023 November ; 79: 104968. doi:10.1016/j.msard.2023.104968.

Reliability of paramagnetic rim lesion classification on quantitative susceptibility mapping (QSM) in people with multiple sclerosis: single-site experience and systematic review

Jack A. Reeves[1], **Maryam Mohebbi**[1], **Robert Zivadinov**[1,2], **Niels Bergsland**[1], **Michael G. Dwyer**[1], **Fahad Salman**[1], **Ferdinand Schweser**[1,2], **Dejan Jakimovski**[1]

[1]Buffalo Neuroimaging Analysis Center, Department of Neurology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA

[2]Center for Biomedical Imaging at the Clinical Translational Science Institute, University at Buffalo, State University of New York, Buffalo, NY, USA

Abstract

Background: Recent developments in iron-sensitive MRI techniques have enabled visualization of chronic active lesions as paramagnetic rim lesions (PRLs) in vivo. Although PRLs have potential as a diagnostic and prognostic tool for multiple sclerosis (MS), limited studies have reported the reliability of PRL assessment. Further evaluation of PRL reliability, through original investigations and review of PRL literature, are warranted.

Methods: A single-center cohort study was conducted to evaluate the inter-rater reliability of PRL identification on QSM in 10 people with MS, 5 people with clinically isolated syndrome, and 5 healthy controls. An additional systematic literature search was then conducted of published PRL reliability data, and these results were synthesized.

Results: In the single-center study, both inter-rater and intra-rater reliability of per-subject PRL number were at an "Excellent" (intraclass correlation coefficient (ICC) of 0.901 for both) level with only 2-years lesion classification experience. Across the reported literature values, reliability of per-lesion rim presence was on average "Near perfect" (for intra-rater; Cohen's $\kappa = 0.833$) and "Substantial" (for inter-rater; Cohens $\kappa = 0.687$), whereas inter-rater reliability of per-subject PRL number was "Good" (ICC = 0.874). Only 4/22 studies reported complete information on rater experience, rater level of training, detailed PRL classification criteria, and reliability cohort size and disease subtypes.

Conclusion: PRLs can be reliably detected both at per-lesion and per-subject level. We recommend that future PRL studies report detailed reliability results, including rater experience level, and use a standardized set of reliability metrics (Cohen's κ or ICC) for improved comparability between studies.

Corresponding Author: Dejan Jakimovski M.D., Ph.D., Research Assistant Professor, Buffalo Neuroimaging Analysis Center, Department of Neurology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, 100 High Street, Buffalo, NY 14203, Fax: 716-859-7066.

Keywords

3T; chronic active lesion; iron rim lesion; multiple sclerosis; paramagnetic rim lesion; quantitative susceptibility mapping; reliability

Introduction

Chronic active lesions (CALs) in multiple sclerosis (MS) are areas of compartmentalized inflammation that occur in the absence of blood-brain barrier breakdown.[1] These could either present as slowly expanding lesions or characterized by a rim of microglia surrounding a demyelinated core, with the microglia containing myelin degradation products and high amounts of iron, giving the lesions an “iron rim” appearance upon iron staining.[2] CAL-associated microglia have been shown to express inflammatory genes and are those that cause active destruction of surrounding white matter.[3]

Recent advances in iron-sensitive MRI techniques such as phase imaging and quantitative susceptibility mapping (QSM) have enabled in-vivo visualization of CALs as paramagnetic rim lesions (PRLs).[2] PRLs appear as hyperintense (iron-rich) rims surrounding diamagnetic (demyelinated) cores.[4] PRLs can be identified on MRI scans with high field strength, such as 7T,[5, 6] as well as on scans obtained using more commonly used lower field strengths, such as 3T or 1.5T.[4, 7, 8]

Previous studies using MRI and neuropathology have demonstrated that PRLs are present in upwards of 50% of people with MS (pwMS) and are associated with clinical disease severity, brain atrophy, relapse rate, and the progression of MS.[9, 10] Furthermore, PRLs have been shown to improve the specificity of MS diagnosis.[11] Comparisons of PRLs in pwMS treated with disease-modifying therapies (DMT) have revealed differences in the longitudinal changes in PRL microstructure, suggesting that PRLs could be a potential marker for monitoring DMT efficacy.[12]

Although PRLs have potential as a diagnostic and prognostic tool for MS, their detection has not been standardized widely, and there have been limited studies reporting the inter- and intra-rater reliability of PRL assessment.[10] As the reliability of PRL detection is crucial for their diagnostic validity and clinical application, a systematic review of the available literature on PRL reliability would be beneficial in determining whether PRLs are reliable enough to be used for the intended purposes.

In this study, we report 3T QSM PRL reliability data from a single-center for raters with different levels of lesion classification experience, and compare our results to previously reported reliability metrics identified via a systematic literature search. Finally, we propose PRL reliability reporting guidelines as a quality control measure for future PRL studies.

Materials and Methods

Single-center PRL reliability study:

A single-center cohort study was conducted to evaluate the inter-rater reliability of the identification of PRLs on QSM in 10 pwMS (5 relapsing-remitting MS and 5 secondary progressive MS), 5 people with clinically isolated syndrome (pwCIS), and 5 healthy controls (HCs). An additional, independent cohort of 90 pwMS or pwCIS was used for training prior to classification on the reliability cohort. All participants within this study were originally part of a larger cardiovascular, environmental and genetic study in MS (CEG-MS).[13] The study subjects utilized in our reproducibility analyses were selected by an independent member of the team and the inclusion criteria consisted of equal distribution of pwCIS, pwRRMS and progressive MS (5 subjects for each group). In the larger CEG-MS study, the inclusion criteria for the pwMS were: 1) age 18-75 years old, and 2) diagnosis of the 2010-revision of the McDonald criteria (which was current at the time of enrollment),[14] 3) On the other hand, the exclusion criteria was: 1) contraindications preventing an MRI examination, 2) pregnant or nursing mothers, 3) presence of clinical relapse or use of intravenous corticosteroid therapy within 30 days of the MRI examination. The inclusion criteria for the HCs was: 1) age 18-75 years old and excluded if 1) had any major neurological disease or 2) had contraindications preventing an MRI examination. Details regarding the inclusion and exclusion criteria are shown elsewhere.[13] The larger study was approved by the Institutional Review Board of the University at Buffalo, and written informed consent was obtained from all participants according to the Declaration of Helsinki.

Imaging methods:

Imaging was performed on the same 3T scanner (Signa Excite HD 12.0; General Electric, Milwaukee, WI, USA) using an eight-channel head-and-neck coil and using the same three-dimensional gradient-echo sequence with first-order flow compensation in read and slice directions (matrix, 512 × 192 × 64; 0.5 × 1 × 2 mm^3; 12° flip; echo time (TE)=22 ms; repetition time (TR)=40 ms; bandwidth, 13.89 kHz). The following additional sequences were acquired during the same imaging session for all subjects: spin-echo T1-weighted (T1w) imaging (matrix, 256 mm x 192 mm; FOV, 256 mm x 192 mm; TE=16 ms; TR=600 ms); FLAIR (matrix, 256 mm x 192 mm; FOV, 256 mm x 192 mm; TE=120 ms; inversion time (TI)=2100 ms; TR=8500 ms; flip angle=90°; echo-train length, 24); dual fast spin-echo proton density- and T2-weighted imaging (matrix, 256 mm x 192 mm; FOV, 256 mm x 192 mm; TE1=9 ms; TE2=98 ms; TR=5300 ms; echo-train length=14); and a 3D high-resolution T1w inversion recovery fast spoiled-gradient echo (IR-FSPGR) (TE=2.8 ms; TI=900 ms; TR=5.9 ms; flip angle, 10°; isotropic 1 mm resolution).

Image analysis:

T2-FLAIR and T1 gadolinium-enhancing lesions were quantified using a semi-automated, deep learning-based lesion segmentation approach with manual correction.[15] Phase and magnitude images were reconstructed from raw k-space data. We applied best-path unwrapping and Laplacian boundary value to the wrapped phase.[16] Subsequently, susceptibility maps for the training and reliability cohorts were reconstructed using HEIDI (Homogeneity Enabled Incremental Dipole Inversion; tolerance parameter = 1E-5,

described in detail elsewhere[17]). Additional, commonly-used inversion algorithms were used to construct susceptibility maps for comparison, including FANSI (Fast Nonlinear Susceptibility Inversion; alpha parameter = 0.0015),[18] LSQR (Least Squares; tolerance parameter = 1E-5),[17] MEDI+0 (Morphology Enabled Dipole Inversion; lambda parameter = 1000),[19] and TKD (Thresholded k-space Division; threshold parameter = 2/3).[20] All algorithms used in this study were implemented with their default parameters, and no personal communications were made with the authors of these algorithms for the purpose of optimization. PRLs were identified on QSM using the proposed criteria determined during the 2022 NAIMS Consensus Statement on Imaging Chronic Active Lesions: 1) a paramagnetic rim continuous with at least 2/3 outer lesion edge that is discernable on at least two image slices, 2) a diamagnetic core relative to surrounding extra-lesional white matter, and 3) non-enhancement on post-contrast T1 short-echo sequence.[21] For confluent lesions, multiple PRLs could be present if they contained separate diamagnetic cores and rims that bordered the lesion edge. PRL ROIs were semi-automatically drawn on QSM images and overlaid on T1 post-contrast images to confirm lack of gadolinium enhancement. The two raters without prior PRL classification experience (MM and JR) received PRL classification training from experienced neuroimaging researchers using an independent set of 90 pwMS. Details on the training are given below (see “PRL classification training procedure” section). Following training, three raters (MM, JR, and DJ) performed PRL classification on the 20-person reliability cohort (see “Inter-rater reliability calculations” section).

PRL classification training procedure:

Following review of relevant PRL literature, the 2 raters without prior PRL classification experience (MM and JR) received two initial hands-on training sessions with 2 researchers with prior PRL and FLAIR lesion classification experience (DJ-MD/PhD with 4 years’ FLAIR lesion experience and NB-PhD with over 15 years’ FLAIR lesion experience). Following this, MM and JR independently classified PRLs in 30 pwMS from the training dataset and compiled presentations visualizing the PRLs on QSM, FLAIR, and T1 images. These were reviewed in multiple meetings where a group of experienced neuroimagers (DJ, NB, RZ-MD/PhD with over 20 years’ FLAIR lesion experience, and MGD – PhD with over 15 years’ FLAIR lesion experience) provided feedback on whether the PRLs were obvious false-positives or likely PRLs. Following this, MM and JR independently classified PRLs in the entire 90 pwMS training dataset then met and created consensus PRL classifications. High-resolution, whole-brain PDFs of the consensus training set classifications (including QSM, QSM + PRL masks, phase images, FLAIR, FLAIR + FLAIR lesion masks, Post-Gad, and T1 images) were then reviewed. Finally, disagreements were reviewed in a consensus meeting (with DJ, NB, RZ, MG, MM, and JR), whole-group consensus classifications were produced, and notes were taken on the causes of false-positive and false-negative classifications.

Inter-rater reliability calculations:

Before reliability analysis, subjects were assigned a randomized ID. Three raters of different levels of formal training and lesion classification experience (i.e. 1, 2, and 4 years’ experience), who were blinded to the subject ID and disease group, conducted PRL analysis twice on each of the 20 scans. The second classification occurred after a minimum gap of

two weeks to minimize recall bias. The initial classifications of the 10 pwMS and 5 pwCIS were used to calculate inter-rater reliability. Although the sample size was limited, the 5 HC scans were used to determine if false-positives were detected.

Inter-rater reliability of per-person PRL number was assessed using Cronbach's alpha and a single-measurement, absolute-agreement, 2-way mixed-effects model. Fleiss κ was used to calculate inter-rater reliability of per-person PRL presence (yes/no). Intraclass correlation coefficient (ICC) for PRL number was calculated using a single-measure, absolute-agreement, 2-way mixed-effects model.

Quantitative reliability values were converted to quality reliability scales for Cohen's κ, Fleiss κ, and ICC based on previously reported scales. Cohen's κ were graded as "No Agreement" (≤ 0), "None to slight" (0.01 – 0.20), "Fair" (0.21 – 0.40), "Moderate" (0.41 – 0.60), "Substantial" (0.61 – 0.80), "Almost Perfect" (0.81 – 1.00).[22] Fleiss κ were graded as "Poor" (< 0.20), "Fair" (0.21 – 0.40), "Moderate" (0.41 – 0.60), "Good" (0.61 – 0.80), and "Very Good" (0.81 – 1.00).[23] ICC values were graded as "Poor" (≤ 0.5), "Moderate" (0.5 – 0.75), "Good" (0.75 – 0.90), and "Excellent" (≥ 0.90).[24]

Systematic Literature Review:

The systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, which outlines guidelines for conducting and reporting systematic reviews. Additionally, the study was pre-registered with the International Prospective Register of Systematic Reviews (PROSPERO) and was assigned a unique record number: CRD42023392683.

Search strategy and inclusion criteria for eligible manuscripts:

The PubMed and EMBASE databases were searched to identify studies investigating PRLs in pwCIS, people with radiologically isolated syndrome (pwRIS), or pwMS. The searches included each of the following categories (with specific terms in parentheses):

1. People with CIS, RIS, or MS ("multiple sclerosis", "MS", "RRMS", "SPMS", "PPMS", "radiologically isolated syndrome", "RIS", "clinically isolated syndrome", "CIS")
2. PwMS undergoing MRI scans ("magnetic resonance imaging", "MRI", "susceptibility", "susceptibility weighted imaging", "SWI", "quantitative susceptibility mapping", "QSM", "phase", "gradient echo", "GE", "GRE")
3. Classification of PRLs ("rim", "ring", "shell", "edge", "smoldering", "smouldering", "chronic active", "paramagnetic", "iron")

An initial search included full-text articles published in peer-reviewed journals between January 1, 2000 and December 5, 2022. Post-hoc review of the manuscripts confirmed that this window captured all relevant manuscripts, with the earliest included manuscript published in 2012.[25] Given the rapid appearance of new PRL publications, a second search (date range: December 6, 2022 to March 24, 2023) was conducted just before final analysis. Only English language articles were included.

Study selection:

The title and abstract of each paper were screened by one author (JR). Full-text articles were retrieved for studies meeting the inclusion criteria and assessed by two authors (JR and DJ) and were included based on consensus agreement by the two authors. Studies were further screened for data on inter- and intra-rater reliability (reliability metric and reliability value) of PRL classification by two authors (JR and DJ). In articles where it is unclear whether these metrics refer specifically to PRL classification the study authors were contacted via email to request the missing information.

Studies were excluded if they met any of the following criteria: 1) Review articles, letters, editorials, opinions, conference abstracts, and case reports with 3 or fewer participants, and 2) Studies where PRLs were only classified on post-mortem MRI or in animals.

Data Extraction:

The primary outcomes for this study were inter- and intra-rater reliability of PRL classification, including, but not limited to: Cohen's κ, Fleiss κ, Lin's concordance coefficient, Kendall's coefficient of concordance, and Krippendorff's alpha. This information was extracted by two authors (JR and DJ). Additionally, one author (JR) extracted data pertaining to (1) background characteristics (author/s and year of publication), (2) subject characteristics (sample size, selection strategy for subjects included in reliability analysis, and type of disease), (3) rater characteristics (number of raters, rater experience level, time between ratings (for intra-rater reliability), and definition of PRL used in the study), and (4) imaging parameters (magnetic field strength and MRI sequence).

Quality Assessment:

The quality of reliability analysis for each study was assessed using modified criteria of the NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies, with each study being score on a scale of 0-5 (for inter-rater reliability) or 0-6 (for intra-rater reliability):

1. Was rater experience (e.g. years) provided?
2. Was rater level of training (e.g. neurology, neuroradiologist, etc.) provided?
3. Was PRL classification criteria defined?
4. Was information on disease type(s) provided for the reliability analysis cohort?
5. Was sample size of the reliability cohort given?
6. (For intra-rater reliability) Was time between classifications reported?

Note that a "PRL classification criterion" was considered to be any additional or more specific criteria beyond simply the presence of a rim (e.g. paramagnetic rim extended 2/3 circumference of lesion, diamagnetic core, etc.).

Data Synthesis:

Extracted data was presented in a table organized by study year. Intra- and inter-rater Cohen's κ and ICC values were synthesized into a separate table, stratified by the PRL

quantity evaluated for reliability (i.e. per-lesion rim presence and per-subject PRL number). Because few studies reported confidence intervals, which are needed to conduct random-effects models, between-study mean Cohen's κ or ICC values were calculated and the between-study ranges were included as a measure of variability. For studies that included multiple reliability values for different scanning parameters (e.g. 3T vs 7T field strength), each reliability value was treated as an independent value and this duplication was noted in the table.

Results

Single-center PRL reliability study:

The HCs consisted of 3 females and 2 males and had a mean age of 53.4 ± 9.9 years (mean ± S.D.). The patient cohort (pwMS and pwCIS) consisted of 5 men and 10 women, mean age of 51.1 ± 9.9 years (mean ± S.D.), mean disease duration of 13.5 ± 11.1 years (mean ± S.D.), Expanded Disability Status Scale (EDSS) of 2.5 [0 – 6.5] (median [range]), and mean T2 lesion volume of 14.0 ± 21.2 mL (mean ± S.D.).

Example consensus PRLs from the training cohort are shown in Figure 2, including their appearance on FLAIR, phase, and QSM. Figure 2 also includes different QSM algorithms including HEIDI (used in our single-center reliability study), FANSI, LSQR, MEDI+0, and TKD. In each case, the PRL rim can be seen on the phase image and all QSM inversion algorithms. Between the QSM inversion algorithms, PRLs were least clear on FANSI due to a blurring-like effect in the reconstructions.

In the final training cohort classifications, 25/90 cases were flagged for consensus review by the experienced neuroimagers. Of these, 12 were confirmed to contain false-positive or false-negative PRL classifications (13.3%) following consensus review. Common causes of false-positive results included paramagnetic vessels incorrectly classified as PRL rims, juxtacortical T2 lesions where the cortical ribbon was incorrectly classified as part of a PRL rim, rim-like QSM artifacts not corresponding to T2 lesions, multiple nodular QSM-hyperintense lesions in close proximity, and incorrect classification of inner edge of lateral ventricular space as T2 lesion due to slight misalignment of the FLAIR and QSM images. False-negative PRLs tended to occur in areas of false-negative T2 lesions co-occurring with T2-hypointense black holes. The results of the intra-rater and inter-rater reliability analyses for the single-center study are shown in Table 1. Reliability of PRL number between the three raters was "Excellent", while the reliability in identifying a pwMS as having one or more PRLs was "Good". In the reliability cohort, 16 consensus PRLs between the three raters were identified from 3 pwRRMS (1, 3, and 5 PRLs) and 2 pwSPMS (1 and 6 PRLs). No consensus PRLs were identified in any pwCIS or HCs. Additionally, none of the raters identified PRLs on any of the HC scans, giving a false positive rate of 0%, a negative predictive value of PRLs of 100%, and an accuracy of 100% in the HC group. The three raters ICCs ranged from "Good" to "Excellent", which increased monotonically with years of lesion classification experience.

Systematic Literature Review and Synthesis

Search results: A total of 8275 studies were identified in the initial database searches and an additional 549 studies were identified in a secondary database search. A flow chart regarding the study classification and study inclusion/exclusion criteria is shown in Figure 1. These were originally reduced to 120 studies following initial screening and removal of duplicates, and finally reduced to 22 studies after removal of studies that did not provide reliability values (n = 82), case reports (n = 4), post-mortem studies (n = 10), studies with overlapping reliability cohorts (n = 1), and one study that included both active lesions and PRLs in reliability analysis.[26]

Study characteristics:

Of the 22 included studies, 12 were conducted on 3T MRI systems, 7 on 7T systems, 1 at 1.5T, 1 included both 1.5T and 3T, and 1 included both 3T and 7T imaging. The paramagnetic rim sign was assessed using phase/susceptibility-weighted imaging (SWI) in 17 studies, QSM in 2 studies, T2*w in 2 studies, and both phase/SWI and QSM in 1 study. 12 studies reported inter-rater reliability only, 3 studies reported intra-rater reliability only, and 7 studies reported both inter- and intra-rater reliability. Of the studies, 12 calculated reliability of per-lesion rim presence (10 reported Cohen's κ, 2 reported percent agreement, 2 reported Fleiss κ, 1 reported Kendall W, and 1 reported Bangdiwala's B-statistics), 2 calculated reliability of per-lesion PRL volume (1 reported ICC and 1 reported Dice coefficients), 5 calculated reliability of per-subject PRL number (4 reported ICC and 1 reported Lin's concordance correlation coefficients), 2 calculated reliability of per-subject PRL categorical classification (i.e. 0, 1-3, or 4+ PRLs; 1 reported ICC and 1 reported Cohen's κ), 2 calculated reliability of per-subject PRL presence (yes/no; both reported as Cohen's κ), and 1 calculated reliability of per-subject PRL volume (reported as ICC).

Details on study year, MRI sequence(s), field strength(s), whether inter- or intra-rater reliability was assessed, quantity measured, reliability metric(s) used, and reliability value(s) are shown in Table 2. Note that two studies reported the same reliability data and were included as a single reference.[27, 28]

Quality assessment:

Details regarding quality assessment are provided in Table 3. All the studies provided sample sizes and most studies defined PRL classification criteria (82%; 18/22). Fewer studies explicitly detailed the MS disease course of the reliability cohort (50%; 11/22), rater training level (45%; 10/22), and rater years' experience (32%; 7/22). Of the 10 studies that reported intra-rater reliability, 4 (40%) reported time elapsed between initial and repeated PRL classification.

Reliability of PRLs:

Quantitative synthesis was conducted on 3 reliability metrics where 3 or more reliability values were available (including from the present single-center study): intra-rater reliability of per-lesion rim presence (Cohen's κ), intra-rater reliability of per-subject PRL presence (Cohen's κ), inter-rater reliability of per-lesion rim presence (Cohen's κ), and inter-rater

reliability of per-subject PRL presence (Cohen's κ), and inter-rater reliability of per-subject PRL number (ICC). These results are shown in Table 4. One study by Hagemeier et al. was not included in quantitative synthesis of per-subject PRL number because it evaluated scan-rescan.[25] Two studies reported two Cohen's κ values for the same data which were averaged before including quantitative synthesis.[8, 29]

Intra-rater reliability of per-lesion rim presence was on average "Near Perfect", with individual studies ranging from "Substantial" to "Near Perfect". Inter-rater reliability of per-lesion rim presence was on average "Substantial", with studies ranging from "Moderate" to "Near Perfect". Inter-rater reliability of per-subject PRL number was on average "Good", ranging from "Good" to "Excellent".

Discussion

The current study reported single-center reliability of PRL detection at 3T and synthesized these results with previously-published literature reliability values. The results indicate that the detection of PRLs on a per-lesion basis is "near perfect" (for intra-rater reliability) and "substantial" (for inter-rater reliability). When evaluating per-subject PRL number, inter-rater reliability was on average "Good" (one category below the maximum category of "Excellent"). These values indicate that PRLs have high within-center reproducibility, despite a variety of reported PRL classification criteria. There were no false-positives in the HC group, leading to a negative predictive value of 100%. The negative predictive value of PRLs is rarely reported in the literature, particularly when contrasting with HCs, but is an important quantity when considering the use of PRLs as a diagnostic marker.

Although not formally quantified, we found that the most noticeable effect of PRL classification training was a decrease in false-positive PRL classifications (e.g., of paramagnetic vessels). Additionally, at least two rounds of training classifications with consensus reviews were necessary to achieve satisfactory performance by the trainees. Therefore, we recommend that future training programs include two or more rounds of training PRL classifications with expert feedback provided after each round.

Consensus evaluations of the training dataset revealed several common causes of false-positive and false-negative classifications. Most misclassifications could potentially be reduced using higher-resolution (i.e., 7T) images, which would provide higher fidelity delineation of paramagnetic vessels and the cortical ribbon. Additionally, we suggest that the accuracy of QSM/FLAIR co-registration be carefully inspected to prevent T2 lesion masks misalignment. Finally, we recommend all PRL raters are trained in T2 lesion analysis prior to PRL training because misclassification of T2 lesions may result in false-positive or false-negative PRL classifications.

The rims of the sample consensus PRLs could be seen on phase imaging and all of the QSM inversion algorithms with particular differences. For example, the FANSI susceptibility maps had a smoothed appearance, which may reduce fidelity of paramagnetic vessels. Misclassification of vessels as PRL rims was a common cause of PRL misclassification, thus FANSI (as currently available) may not be the most suitable reconstruction algorithm.

Although the rims were visual on phase images, previous studies have been noted that phase imaging may cause higher rates of false-positive PRL classifications.[30,31] Note that all QSM inversion algorithms were implemented in their default settings with no personal communications made with the authors of these algorithms for the purpose of optimization. This approach allowed us to evaluate the algorithms in their standard configurations, as they are commonly used in the field. However, it should also be noted that optimization of these algorithms may change the results and should be considered in future studies.

For our own reproducibility results, we found that the reliability of PRL number reached the highest category (“Excellent”), whereas the reliability of evaluating whether a pwMS had at least one PRL was “Good”. Intra-rater reliability of PRL number increased with increasing lesion classification experience, reaching “Excellent” levels in the rater with at least 2 years of experience at the level of a PhD student. Based on these preliminary results, we suggest that PRL rating be conducted by raters with at least 2 years’ lesion classification experience and with formal PRL training. However, reasonable reliability was achieved by a rater with only 1 year of lesion of classification experience. Researchers at this level may be able to perform PRL rating with oversight from a more experienced rater. Given the qualitative nature of these observations, more data on the relationship between rater experience and PRL classification reliability are needed.

Per-subject and per-lesion reliability values may have different utilities in clinical and research settings. Per-subject reliabilities are relevant when determining inclusion and exclusion criteria for future clinical trials. Per-lesion reliabilities are useful in studies where individual PRLs are studied over time, for example if using PRL tissue integrity to measure the effect of disease-modifying therapies.[12] In either case, we showed that the reliability in PRL classification reached the highest or second highest reliability category. This would be improved further by adoption of standardized PRL classification criteria.

Despite identifying a relatively large number of studies that assessed PRLs (120), only a small number of these reported PRL reliability assessments (22). These studies were highly heterogeneous in terms of the reliability metric reported, which included Cohen’s κ, Fleiss κ, Kendall W, Bangdiwala’s B-statistics, ICC, Dice coefficients, and Lin’s concordance correlation coefficients. These metrics measure reliability different types of data (e.g., continuous data for ICC and categorical data for Fleiss κ) and have different standard scales (e.g. “very good” for Fleiss $\kappa > 0.8$ and “excellent” for ICC ≥ 0.9). Therefore, these metrics have different interpretations and, in general, cannot be directly compared. Further, even fewer studies reported standard errors or confidence intervals of the reliability values. These factors made it difficult to synthesize findings across studies, perform meta-analysis, and to compare PRL detection reliability between scanning parameters (i.e., 3T vs 7T and QSM vs SWI). Future research could benefit from standardizing the assessment and reporting of reliability values, such as by following established guidelines and consensus statements.

There are several limitations that affected our systematic analysis and restricted the ability to perform a meta-analysis. Fewer studies provided information about the MS disease course of the reliability cohort, rater training level, and rater years of experience. Knowledge of MS disease course is particularly relevant when interpreting PRL reliabilities because the

prevalence of PRLs has been shown to be different between RRMS and SPMS,[10] which can affect reliability values. Most studies report randomly selecting participants for inclusion in reliability analysis or utilize the entire dataset,[27, 32–35] but others use an enriched dataset,[36] a consideration which can also affect relative prevalence and reliability values. Additionally, out of the studies that reported intra-rater reliability, only a few reported the time elapsed between initial and repeated PRL classification. These findings suggest that while certain aspects of quality assessment were well-documented in the studies, there were gaps in information about other important factors. Future studies should aim to provide a more comprehensive picture of quality assessment measures to improve the rigor and reliability of research in this area.

This study had several limitations. The single-center study had a fairly low sample size, although it was comparable or greater than those used in previous studies (range = 5-25 and median = 9.5 subjects for reliability of per-subject PRL number or presence),[6, 7, 25, 34, 35, 37] which led to relatively large confidence intervals on the reliability values. Power calculations would be useful in future studies, particularly those comparing PRL classification reliability between factors such as MR field strength (e.g., 3T vs 7T) or QSM vs phase imaging. Finally, the single-center study did not calculate reliability of per-lesion rim presence. This choice was made because the semi-automated contouring method did not identify individual lesions, but rather provided a single mask of all T2 lesions.[38] The per-lesion PRL analyses are mostly limited by the inability to properly separate highly confluent T2 lesions. These limitations are more pronounced in long-standing pwMS who have high T2 lesion burden.

The literature review was limited by relatively few studies reporting reliability values and high heterogeneity in reliability metrics reported and the quantities analyzed (e.g., per-lesion rim presence or per-subject PRL number). Additionally, no consensus guidelines have been published for PRL detection, leading to a variety of classification criteria reported between studies. Due to these factors, the effect MR sequence and field strength on PRL classification reliability could not be analyzed in the present study. Greater MR field strength has been shown to be associated with greater detected PRL prevalence,[10] so the effect of these factors on PRL classification reliability will be important to evaluate in future studies.

In conclusion, PRLs can be reliably detected at per-subject level. We recommend that future studies report rater experience (i.e., years), rater training level, specific PRL classification criteria, details on MS disease course of the reliability cohort, cohort size, and, for intra-rater reliability, time between initial and repeated PRL reliability analysis. Additionally, for improved comparability between studies, we recommend future studies report either Cohen's κ (for per-lesion reliabilities) or ICC (for per-subject PRL number reliabilities), along with false-positive values, negative predictive values, and accuracy. These improvements should aim at bring PRL analyses to the real-world use as a neuroimaging marker for pwMS selection and treatment response.

Acknowledgements

The authors would like to thank Daisy R. fur her feedback of this manuscript.

Funding Sources

Research reported in this publication was supported by grants from the National Institutes of Health (R01NS114227 from the National Institute of Neurological Disorders and Stroke and UL1TR001412 from the National Center for Advancing Translational Sciences) and the National MS Society (NMSS RG 5195A4/1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NMSS.

Declaration of Conflicting Interests

R.Z. has received personal compensation from Bristol Myers Squibb, EMD Serono, Sanofi, and Novartis for speaking and consultant fees. He received financial support for research activities from Sanofi, Novartis, Bristol Myers Squibb, Mapi Pharma, Keystone Heart, Protembis, and V-WAVE Medical.

M.D. received personal compensation from Bristol Myers Squibb, Novartis, EMD Serono and Keystone Heart, and financial support for research activities from Bristol Myers Squibb, Novartis, Mapi Pharma, Keystone Heart, Protembis, and V-WAVE Medical.

References

1. Kuhlmann T, Ludwin S, Prat A, et al. An updated histological classification system for multiple sclerosis lesions. Acta Neuropathol 2017; 133: 13–24. 20161217. DOI: 10.1007/s00401-016-1653-y. [PubMed: 27988845]
2. Calvi A, Haider L, Prados F, et al. In vivo imaging of chronic active lesions in multiple sclerosis. Mult Scler 2022; 28: 683–690. 20200923. DOI: 10.1177/1352458520958589. [PubMed: 32965168]
3. Absinta M, Maric D, Gharagozloo M, et al. A lymphocyte-microglia-astrocyte axis in chronic active multiple sclerosis. Nature 2021; 597: 709–714. 20210908. DOI: 10.1038/s41586-021-03892-7. [PubMed: 34497421]
4. Absinta M, Sati P, Fechner A, et al. Identification of Chronic Active Multiple Sclerosis Lesions on 3T MRI. A JNR Am J Neuroradiol 2018; 39: 1233–1238. 20180503. DOI: 10.3174/ajnr.A5660.
5. Dal-Bianco A, Grabner G, Kronnerwetter C, et al. Slow expansion of multiple sclerosis iron rim lesions: pathology and 7 T magnetic resonance imaging. Acta Neuropathol 2017; 133: 25–42. 20161027. DOI: 10.1007/s00401-016-1636-z. [PubMed: 27796537]
6. Kuchling J, Ramien C, Bozin I, et al. Identical lesion morphology in primary progressive and relapsing-remitting MS--an ultrahigh field MRI study. Mult Scler 2014; 20: 1866–1871. 20140429. DOI: 10.1177/1352458514531084. [PubMed: 24781284]
7. Micheletti L, Maldonado FR, Watal P, et al. Utility of paramagnetic rim lesions on 1.5-T susceptibility phase imaging for the diagnosis of pediatric multiple sclerosis. Pediatr Radiol 2022; 52: 97–103. 20211006. DOI: 10.1007/s00247-021-05188-4. [PubMed: 34611736]
8. Hemond CC, Reich DS and Dundamadappa SK. Paramagnetic Rim Lesions in Multiple Sclerosis: Comparison of Visualization at 1.5-T and 3-T MRI. AJR Am J Roentgenol 2022; 219: 120–131. 20211201. DOI: 10.2214/AJR.21.26777. [PubMed: 34851712]
9. Absinta M, Sati P, Masuzzo F, et al. Association of Chronic Active Multiple Sclerosis Lesions With Disability In Vivo. JAMA Neurol 2019; 76: 1474–1483. DOI: 10.1001/jamaneurol.2019.2399. [PubMed: 31403674]
10. Ng Kee Kwong KC, Mollison D, Meijboom R, et al. The prevalence of paramagnetic rim lesions in multiple sclerosis: A systematic review and meta-analysis. PLoS One 2021; 16: e0256845. 20210908. DOI: 10.1371/journal.pone.0256845. [PubMed: 34495999]
11. Maggi P, Sati P, Nair G, et al. Paramagnetic Rim Lesions are Specific to Multiple Sclerosis: An International Multicenter 3T MRI Study. Ann Neurol 2020; 88: 1034–1042. 20200909. DOI: 10.1002/ana.25877. [PubMed: 32799417]
12. Eisele P, Wittayer M, Weber CE, et al. Impact of disease-modifying therapies on evolving tissue damage in iron rim multiple sclerosis lesions. Mult Scler 2022; 28: 2294–2298. 20220701. DOI: 10.1177/13524585221106338. [PubMed: 35778799]
13. Dwyer MG, Bergsland N, Ramasamy DP, et al. Atrophied Brain Lesion Volume: A New Imaging Biomarker in Multiple Sclerosis. J Neuroimaging 2018; 28: 490–495. 20180601. DOI: 10.1111/jon.12527. [PubMed: 29856910]

14. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann Neurol 2011; 69: 292–302. DOI: 10.1002/ana.22366. [PubMed: 21387374]
15. Zivadinov R, Heininen-Brown M, Schirda CV, et al. Abnormal subcortical deep-gray matter susceptibility-weighted imaging filtered phase measurements in patients with multiple sclerosis: a case-control study. Neuroimage 2012; 59: 331–339. 20110727. DOI: 10.1016/j.neuroimage.2011.07.045. [PubMed: 21820063]
16. Zhou D, Liu T, Spincemaille P, et al. Background field removal by solving the Laplacian boundary value problem. NMR Biomed 2014; 27: 312–319. 20140107. DOI: 10.1002/nbm.3064. [PubMed: 24395595]
17. Schweser F, Deistung A, Lehr BW, et al. Quantitative imaging of intrinsic magnetic tissue properties using MRI signal phase: an approach to in vivo brain iron metabolism? Neuroimage 2011; 54: 2789–2807. 20101030. DOI: 10.1016/j.neuroimage.2010.10.070. [PubMed: 21040794]
18. Milovic C, Bilgic B, Zhao B, et al. Fast nonlinear susceptibility inversion with variational regularization. Magn Reson Med 2018; 80: 814–821. 20180110. DOI: 10.1002/mrm.27073. [PubMed: 29322560]
19. Liu Z, Spincemaille P, Yao Y, et al. MEDI+0: Morphology enabled dipole inversion with automatic uniform cerebrospinal fluid zero reference for quantitative susceptibility mapping. Magn Reson Med 2018; 79: 2795–2803. 20171011. DOI: 10.1002/mrm.26946. [PubMed: 29023982]
20. Wharton S, Schafer A and Bowtell R. Susceptibility mapping in the human brain using threshold-based k-space division. Magn Reson Med 2010; 63: 1292–1304. DOI: 10.1002/mrm.22334. [PubMed: 20432300]
21. Bagnato F NAIMS Symposium on Imaging Chronic Active White Matter Lesions. In: ACTRIMS Forum West Palm Beach, FL, United States, 2022.
22. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012; 22: 276–282. [PubMed: 23092060]
23. Altman DG. Practical statistics for medical research. Boca Raton, Fla.: Chapman & Hall/CRC, 1999, p.xii, 611 p.
24. Koo TK and Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016; 15: 155–163. 20160331. DOI: 10.1016/j.jcm.2016.02.012. [PubMed: 27330520]
25. Hagemeier J, Heininen-Brown M, Poloni GU, et al. Iron deposition in multiple sclerosis lesions measured by susceptibility-weighted imaging filtered phase: a case control study. J Magn Reson Imaging 2012; 36: 73–83. 20120307. DOI: 10.1002/jmri.23603. [PubMed: 22407571]
26. Suzuki M, Kudo K, Sasaki M, et al. Detection of active plaques in multiple sclerosis using susceptibility-weighted imaging: comparison with gadolinium-enhanced MR imaging. Magn Reson Med Sci 2011; 10: 185–192. DOI: 10.2463/mrms.10.185. [PubMed: 21960001]
27. Dal-Bianco A, Schranzer R, Grabner G, et al. Iron Rims in Patients With Multiple Sclerosis as Neurodegenerative Marker? A 7-Tesla Magnetic Resonance Study. Front Neurol 2021; 12: 632749. 20211221. DOI: 10.3389/fneur.2021.632749. [PubMed: 34992573]
28. Dal-Bianco A, Grabner G, Kronnerwetter C, et al. Long-term evolution of multiple sclerosis iron rim lesions in 7 T MRI. Brain 2021; 144: 833–847. DOI: 10.1093/brain/awaa436. [PubMed: 33484118]
29. Marcille M, Hurtado Rua S, Tyshkov C, et al. Disease correlates of rim lesions on quantitative susceptibility mapping in multiple sclerosis. Sci Rep 2022; 12: 4411. 20220315. DOI: 10.1038/s41598-022-08477-6. [PubMed: 35292734]
30. Huang W, Sweeney EM, Kaunzner UW, et al. Quantitative susceptibility mapping versus phase imaging to identify multiple sclerosis iron rim lesions with demyelination. J Neuroimaging 2022; 32: 667–675. 20220309. DOI: 10.1111/jon.12987. [PubMed: 35262241]
31. Eskreis-Winkler S, Deh K, Gupta A, et al. Multiple sclerosis lesion geometry in quantitative susceptibility mapping (QSM) and phase imaging. J Magn Reson Imaging 2015; 42: 224–229. 20140830. DOI: 10.1002/jmri.24745. [PubMed: 25174493]

32. Altokhis AI, Hibbert AM, Allen CM, et al. Longitudinal clinical study of patients with iron rim lesions in multiple sclerosis. Mult Scler 2022; 28: 2202–2211. 20220824. DOI: 10.1177/13524585221114750. [PubMed: 36000485]
33. Kolb H, Absinta M, Beck ES, et al. 7T MRI Differentiates Remyelinated from Demyelinated Multiple Sclerosis Lesions. Ann Neurol 2021; 90: 612–626. 20210902. DOI: 10.1002/ana.26194. [PubMed: 34390015]
34. Clarke MA, Pareto D, Pessini-Ferreira L, et al. Value of 3T Susceptibility-Weighted Imaging in the Diagnosis of Multiple Sclerosis. AJNR Am J Neuroradiol 2020; 41: 1001–1008. . DOI: 10.3174/ajnr.A6547. [PubMed: 32439639]
35. Sinnecker T, Schumacher S, Mueller K, et al. MRI phase changes in multiple sclerosis vs neuromyelitis optica lesions at 7T. Neurol Neuroimmunol Neuroinflamm 2016; 3: e259. 20160722. DOI: 10.1212/NXI.0000000000000259. [PubMed: 27489865]
36. Meaton I, Altokhis A, Allen CM, et al. Paramagnetic rims are a promising diagnostic imaging biomarker in multiple sclerosis. Mult Scler 2022; 28: 2212–2220. 20220826. DOI: 10.1177/13524585221118677. [PubMed: 36017870]
37. Treaba CA, Conti A, Klawiter EC, et al. Cortical and phase rim lesions on 7 T MRI as markers of multiple sclerosis disease progression. Brain Commun 2021; 3: fcab134. 20210624. DOI: 10.1093/braincomms/fcab134. [PubMed: 34704024]
38. Zivadinov R, Rudick RA, De Masi R, et al. Effects of IV methylprednisolone on brain atrophy in relapsing-remitting MS. Neurology 2001; 57: 1239–1247. DOI: 10.1212/wnl.57.7.1239. [PubMed: 11591843]
39. Barquero G, La Rosa F, Kebiri H, et al. RimNet: A deep 3D multimodal MRI architecture for paramagnetic rim lesion assessment in multiple sclerosis. Neuroimage Clin 2020; 28: 102412. 20200904. DOI: 10.1016/j.nicl.2020.102412. [PubMed: 32961401]
40. Maggi P, Kuhle J, Schadelin S, et al. Chronic White Matter Inflammation and Serum Neurofilament Levels in Multiple Sclerosis. Neurology 2021; 97: e543–e553. 20210604. DOI: 10.1212/WNL.0000000000012326. [PubMed: 34088875]
41. Krajnc N, Bsteh G, Kasprian G, et al. Peripheral Hemolysis in Relation to Iron Rim Presence and Brain Volume in Multiple Sclerosis. Front Neurol 2022; 13: 928582. 20220629. DOI: 10.3389/fneur.2022.928582. [PubMed: 35865643]
42. Zhang H, Nguyen TD, Zhang J, et al. QSMRim-Net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps. Neuroimage Clin 2022; 34: 102979. 20220301. DOI: 10.1016/j.nicl.2022.102979. [PubMed: 35247730]
43. Calvi A, Clarke MA, Prados F, et al. Relationship between paramagnetic rim lesions and slowly expanding lesions in multiple sclerosis. Mult Scler 2023; 29: 352–362. 20221214. DOI: 10.1177/13524585221141964. [PubMed: 36515487]
44. Krajnc N, Dal-Bianco A, Leutmezer F, et al. Association of paramagnetic rim lesions and retinal layer thickness in patients with multiple sclerosis. Mult Scler 2023; 29: 374–384. 20221220. DOI: 10.1177/13524585221138486. [PubMed: 36537667]

Highlights:

- Reliability of PRL classification was analyzed at a single center, and a systematic literature search of PRL reliability values was conducted
- In the single-center study, inter-rater reliability of per-subject PRL number was “Excellent” (intraclass correlation coefficient = 0.901).
- In meta-analysis of literature, reliability of per-subject PRL number was on average “Good” (ICC = 0.874) and reliability per-lesion rim presence was on average “Near perfect” (for intra-rater; Cohen’s $\kappa = 0.833$) and “Substantial” (for inter-rater; Cohen’s $\kappa = 0.687$).
- PRLs can be reliably detected both at per-lesion level and per-subject level, but more consistent reporting of PRL reliability is needed to ensure PRL validity across studies.

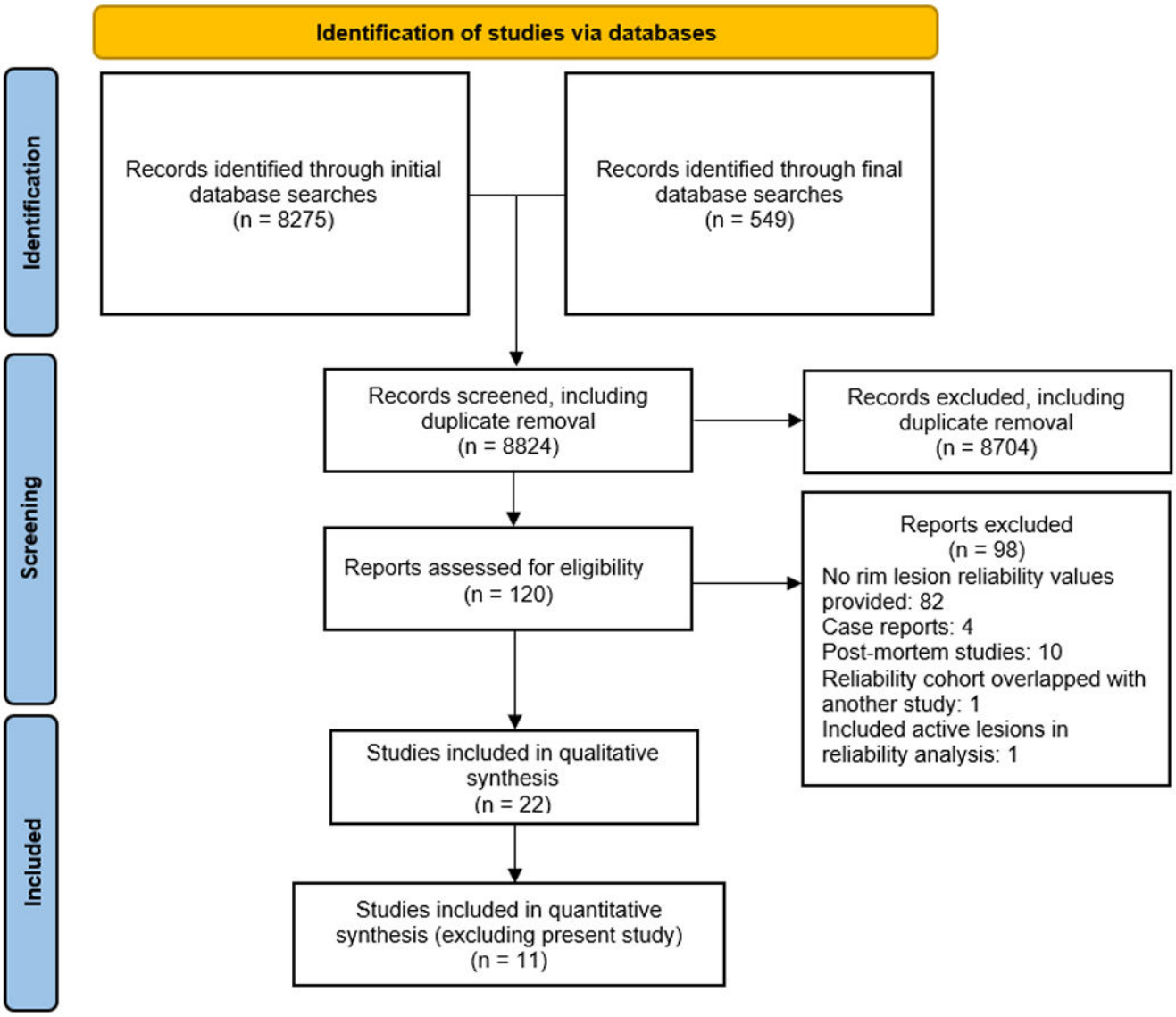

Figure 1.

PRISMA flow diagram

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram showing details on literature review, screening, and study selection.

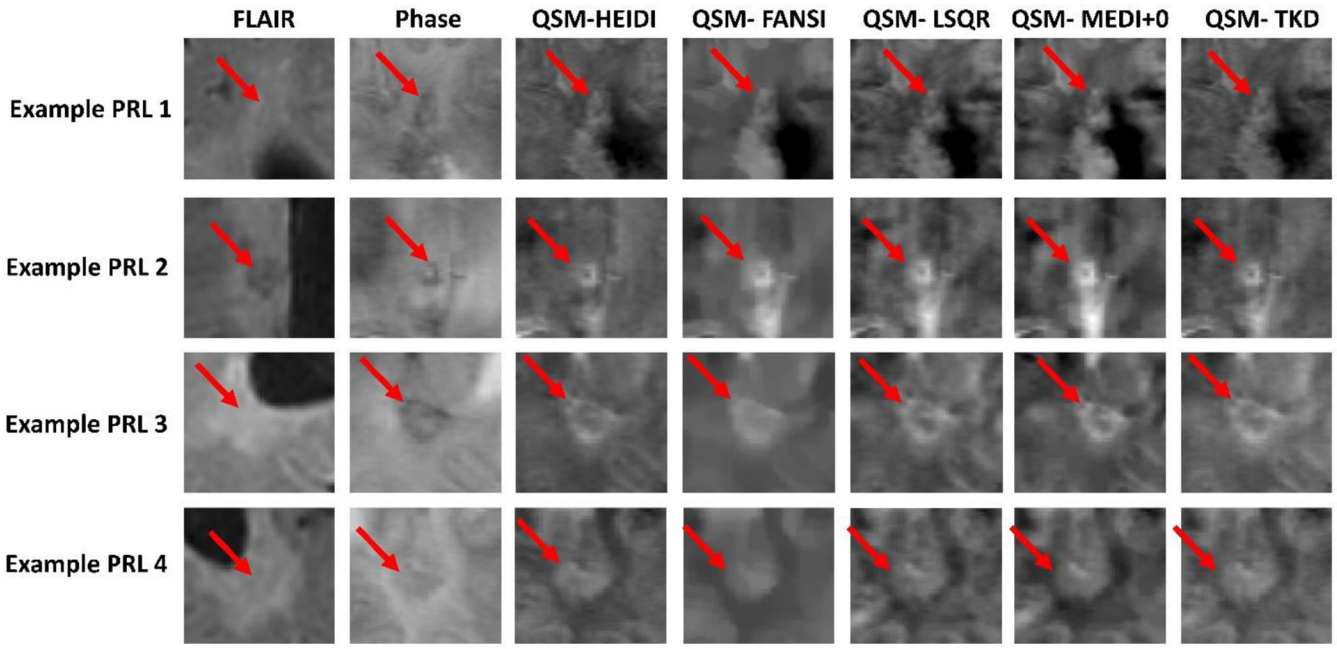

Figure 2.

Exemplar consensus PRLs compared between phase and different QSM inversion algorithms. Phase images have an intensity scale of −1 to 1, and QSM images all have an intensity scale of −0.1 to 0.2.

FANSI – Fast Nonlinear Susceptibility Inversion; FLAIR – Fluid Attenuated Inversion Recovery; HEIDI-Homogeneity Enabled Incremental Dipole Inversion; LSQR – Least Squares; MEDI - Morphology Enabled Dipole Inversion; PRL – Paramagnetic Rim Lesion; TKD - Thresholded k-space Division.

Table 1.

Intra-rater and inter-rater paramagnetic rim lesion reliability in people with multiple sclerosis.

Intra- or inter-rater reliability?	PRL metric	Reliability measure	Value [95% CI]	Rater	Education level	Lesion classification experience (years)
Intra	PRL number	ICC	0.857 [0.626 – 0.949]	1	Master's student	1
			0.901 [0.732 – 0.966]	2	MD/PhD student	2
			0.948 [0.854 – 0.982]	3	MD/PhD	4
Inter	PRL number	ICC	0.901 [0.784 – 0.962]			
	PRL number	Cronbach's Alpha	0.965 [0.916 – 0.987]			
	PRL presence (yes/no)	Fleiss κ	0.644 [0.352 – 0.936]			

CI – confidence interval; ICC – intraclass correlation coefficient; PRL – paramagnetic rim lesion.

Table 2.

Reliability of paramagnetic rim lesions reported in the literature.

Study	Field Strength	Image type	Sample used for reliability	Intra- or inter-rater reliability?	Quantity measured	Reliability metric	Reliability value
Hagemeier et al. (2012)[25]	3T	Phase	5 pwMS and 5 HCs	Intra (scan-rescan)	Per-subject number	ICC	1
					Per-subject volume	ICC	0.999
Kuchling et al. (2014)[6] *	7T	T2*w	10 representative cases (7 pwMS patients and 3 HCs)	Intra	Per-subject number	ICC as a two-way mixed test for average measures	0.854
				Inter	Per-subject number	ICC as a two-way mixed test for average measures	0.796
Sinnecker et al. (2016)[35] *	7T	Phase	10 randomly selected pwMS or pwNMOSD	Inter	Per-subject number	ICC as a 2-way mixed test of average measures	0.96
Dal-Bianco et al. (2017)[27, 28]	7T	FLAIR-SWI	15 rim or non-rim lesions segmented twice	Intra	Per-lesion volume	ICC	0.998
Absinta et al. (2018)[4] *	3T	Phase	100 lesions from 20 pwMS (16 RRMS, 4 PMS), 5 lesions per pwMS	Inter	Per-lesion rim presence	Fleiss κ	0.71
				Intra	Per-lesion rim presence	Cohen's κ	0.77
	7T			Inter	Per-lesion rim presence	Cohen's κ	0.72
				Intra	Per-lesion rim presence	Cohen's κ	0.77
Barquero et al. (2020)[39] *	3T	Phase	124 pwMS (87 RRMS, 21 SPMS, 16 PPMS)	Inter	Per-lesion rim presence	Cohen's κ	0.73
Clarke et al. (2020)[34] *	3T	SWI	25 randomly-chosen scans (from 112 pwCIS and 35 non-MS)	Inter	Per-subject number	ICC (2-way mix model, single measures, absolute agreement)	0.84 (0.64-0.93) (95% CI)
Maggi et al. (2020)[11] *	3T	Phase	83 non-MS cases and 83 pwMS	Inter	Per-lesion rim presence	Cohen's κ	0.79
Dal-Bianco et al. (2021)[27] *	7T	FLAIR-SWI	15 randomly selected PRLs	Intra	Per-lesion volume	Dice	0.92 ± 0.05 (mean ± s.d.)
		MP2RAGE		Intra	Per-lesion volume	Dice	0.90 ± 0.02 (mean ± s.d.)
Kolb et al. (2021)[33] *	7T	T2*w phase	110 randomly selected lesions	Intra	Per-lesion rim presence	Cohen's κ	0.96
				Inter	Per-lesion rim presence	Kendall W	0.88
Maggi et al. (2021)[40]	3T	Unwrapped phase images	118 pwMS (86 RRMS, 32 PMS)	Inter	Per-subject PRL category (0, 1-3, or 4+ PRLs)	Cohen's κ	0.83

Study	Field Strength	Image type	Sample used for reliability	Intra- or inter-rater reliability?	Quantity measured	Reliability metric	Reliability value
Treaba et al. (2021)[37]	7T	Phase	5 pwMS	Intra	Per-subject number	Lin's concordance correlation coefficients	0.99
				Inter	Per-subject number	Lin's concordance correlation coefficients	0.98
Altokhis et al. (2022)[32]	7T	SWI-filtered phase	10 randomly selected pwCIS or pwMS	Intra	Per-subject PRL category (0, 1-3, or 4+ PRLs)	ICC	0.95
				Inter	Per-subject PRL category (0, 1-3, or 4+ PRLs)	ICC	0.81
Hemond et al. (2022) [8]	1.5T	Filtered phase	5 pwRRMS, 3pwSPMS, and 1 pwPPMS	Inter	Per-subject PRL presence	Cohen's κ	0.65
	3T			Inter	Per-subject PRL presence	Cohen's κ	0.72
				Intra	Per-subject PRL presence	Cohen's κ	0.70
Huang et al. (2022)[30] *	3T	High-pass-filtered phase	2062 FLAIR-hyperintense no-enhancing lesions from 80 pwRRMS	Inter	Per-lesion rim presence	Cohen's κ	0.43
				Inter	Per-lesion rim presence	Percent agreement	75.8%
		QSM		Inter	Per-lesion rim presence	Cohen's κ	0.62
				Inter	Per-lesion rim presence	Percent agreement	92.7%
Krajnc et al. (2022)[41]	3T	SWI	75 pwMS	Inter	Per-lesion rim presence	Percent agreement	98.7%
Marcille et al. (2022)[29] *	3T	QSM	159 pwRRMS	Inter	Per-lesion rim presence	Bangdiwala's B-statistics	95.49%
						Cohen's κ	0.706
						Cohen's κ	0.858
						Fleiss κ	0.71
Meaton et al. (2022)[36]	3T	SWI	100 blocks (1/8 brain volume each)	Intra	Per-subject PRL presence	Cohen's κ	0.696
				Inter	Per-subject PRL presence	Cohen's κ	0.827
Micheletti et al. (2022)[7]	1.5T	Phase	29 children (13 pwMS and 16 non-MS). 132 white matter lesions, 63 were in the MS group and 69 were in the non-MS group.	Inter	Per-lesion rim presence	Cohen's κ	0.75
Zhang et al. (2022)[42] *	3T	QSM	4,163 T2-FLAIR lesions from 172 pwMS	Inter	Per-lesion rim presence	Cohen's κ	0.59
Calvi et al. (2023)[43] *	3T	SWI	20 pwMS	Inter	Per-lesion rim presence	Cohen's κ	0.87

Study	Field Strength	Image type	Sample used for reliability	Intra- or inter-rater reliability?	Quantity measured	Reliability metric	Reliability value
Krajnc et al. (2023)[44]	3T	SWI	107 pwMS (97 RRMS, 10 SPMS)	Inter	Per-lesion rim presence	Percent agreement	97.2%

FLAIR – fluid-attenuated inversion recovery; HC – healthy control; ICC – intraclass correlation coefficient; pwNMOSD – people with neuromyelitis optica spectrum disorder; PRL – paramagnetic rim lesion; pwCIS – people with clinically isolated syndrome; pwMS – people with multiple sclerosis; pwPPMS – people with primary progressive multiple sclerosis; pwRRMS – people with relapsing-remitting multiple sclerosis; pwSPMS – people with secondary progressive multiple sclerosis; QSM – quantitative susceptibility mapping; RRMS – relapsing-remitting multiple sclerosis; SPMS – secondary progressive multiple sclerosis; SWI – susceptibility-weighted imaging.

* Included in quantitative synthesis.

Table 3.

Quality assessment of reported literature PRL reliabilities. Studies were assessed for quality of reported reliability data using modified criteria of the NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies.

Question	Was rater experience (i.e. years) provided?	Was rater level of training (e.g. neurologist, neuroradiologist, etc.) provided?	Was PRL classification criteria defined? (Number criteria listed)	Was information on MS disease course (i.e. RRMS, PMS, etc.) provided for the reliability analysis cohort?	Was sample size for the reliability cohort given?	(For intra-rater reliability) Was time between classifications reported?
Hagemeier *et al.* (2012)	X	X	✓ (2)	✓	✓	X
Kuchling et al. (2014)	X	X	X	X	✓	✓
Sinnecker et al. (2016)	X	X	X	X	✓	N/A
Dal-Bianco et al. (2017)	X	✓	✓ (3)	X	✓	X
Absinta et al. (2018)	X	✓	✓ (2)	✓	✓	✓
Barquero et al. (2020)	✓	X	✓ (2)	✓	✓	N/A
Clarke et al. (2020)	✓	✓	✓ (6)	X	✓	N/A
Maggi et al. (2020)	X	X	✓ (2)	X	✓	N/A
Dal-Bianco et al. (2021)	✓	✓	✓ (4)	X	✓	X
Kolb et al. (2021)	X	✓	X	X	✓	✓
Maggi et al. (2021)	X	X	✓ (2)	✓	✓	N/A
Treaba et al. (2021)	X	✓	✓ (2)	X	✓	X
Altokhis et al. (2022)	X	X	✓ (3)	X	✓	X
Hemond et al. (2022)	✓	✓	✓ (3)	✓	✓	✓
Huang et al. (2022)	✓	✓	✓ (2)	✓	✓	N/A
Krajnc et al. (2022)	X	X	✓ (3)	✓	✓	N/A
Marcille et al. (2022)	✓	✓	✓ (3)	✓	✓	N/A
Meaton et al. (2022)	X	X	✓ (4)	X	✓	X
Micheletti et al. (2022)	✓	✓	✓ (3)	✓	✓	N/A
Zhang et al. (2022)	X	X	X	✓	✓	N/A

Question	Was rater experience (i.e. years) provided?	Was rater level of training (e.g. neurologist, neuroradiologist, etc.) provided?	Was PRL classification criteria defined? (Number criteria listed)	Was information on MS disease course (i.e. RRMS, PMS, etc.) provided for the reliability analysis cohort?	Was sample size for the reliability cohort given?	(For intra-rater reliability) Was time between classifications reported?
Calvi et al. (2023)	X	X	✓ (6)	X	✓	N/A
Krajnc et al. (2023)	X	X	✓ (4)	✓	✓	N/A

PRL – paramagnetic rim lesion; RRMS – relapsing-remitting multiple sclerosis; PMS –progressive multiple sclerosis.

Table 4.

Average reliability values calculated from literature values and our single-center study.

Reliability metric	Intra- or inter-rater?	Number of values (studies)	Reliability metric	Average reliability value	Range	Confidence in PRL reliability, average (range)
Per-lesion rim presence	Intra	3 (2)	Cohen's κ	0.833	0.77 – 0.96	"Near perfect" ("Substantial" to "Near Perfect")
	Inter	7 (6)	Cohen's κ	0.687	0.43 – 0.87	"Substantial" ("Moderate" to "Near perfect")
Per-subject PRL number	Inter	4 (4)	ICC	0.874	0.796 – 0.96	"Good"("Good" to "Excellent")

ICC – intraclass correlation coefficient; PRL – paramagnetic rim lesion.