Received: 1 December 2020 | Revised: 25 January 2021 | Accepted: 8 February 2021

DOI: 10.1002/mrm.28754

FULL PAPER

Magnetic Resonance in Medicine

QSM reconstruction challenge 2.0: Design and report of results

QSM Challenge 2.0 Organization Committee | **Berkin Bilgic**[1,2,3] | **Christian Langkammer**[4] | **José P. Marques**[5] | **Jakob Meineke**[6] | **Carlos Milovic**[7,8,9] | **Ferdinand Schweser**[10,11]

[1]Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, Massachusetts, USA

[2]Department of Radiology, Harvard Medical School, Boston, Massachusetts, USA

[3]Harvard-MIT Health Sciences and Technology, MIT, Cambridge, Massachusetts, USA

[4]Department of Neurology, Medical University of Graz, Graz, Austria

[5]Donders Centre for Cognitive Neuroimaging, Radboud University, Nijmegen, Netherlands

[6]Philips Research, Hamburg, Germany

[7]Department of Electrical Engineering, Pontificia Universidad Catolica de Chile, Santiago, Chile

[8]Biomedical Imaging Center, Pontificia Universidad Catolica de Chile, Santiago, Chile

[9]Department of Medical Physics and Biomedical Engineering, University College London, London, UK

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

[11]Center for Biomedical Imaging, Clinical and Translational Science Institute, University at Buffalo, The State University of New York, Buffalo, New York, USA

Correspondence
Carlos Milovic, Department of Medical Physics and Biomedical Engineering, University College London, London, UK.
Email: c.milovic@ucl.ac.uk

Funding information
National Agency for Research and Development, Millennium Science Initiative Program, Grant/Award Number: NCN17_129; Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: FOM-N-31/16PR1056; Siemens Healthineers; National Institutes of Health, Grant/Award Number: R01 EB028797, U01 EB025162, P41 EB030006 and R01 MH11; Fondo Nacional de Desarrollo Científico y Tecnológico, Grant/Award Number: PIA-ACT192064; National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR001412; Cancer Research UK Multidisciplinary Award, Grant/Award Number: C53545/A24348

Purpose: The aim of the second quantitative susceptibility mapping (QSM) reconstruction challenge (Oct 2019, Seoul, Korea) was to test the accuracy of QSM dipole inversion algorithms in simulated brain data.

Methods: A two-stage design was chosen for this challenge. The participants were provided with datasets of multi-echo gradient echo images synthesized from two realistic *in silico* head phantoms using an MR simulator. At the first stage, participants optimized QSM reconstructions without ground truth data available to mimic the clinical setting. At the second stage, ground truth data were provided for parameter optimization. Submissions were evaluated using eight numerical metrics and visual ratings.

Results: A total of 98 reconstructions were submitted for stage 1 and 47 submissions for stage 2. Iterative methods had the best quantitative metric scores, followed by deep learning and direct inversion methods. Priors derived from magnitude data improved the metric scores. Algorithms based on iterative approaches and total variation (and its derivatives) produced the best overall results. The reported results and analysis pipelines have been made public to allow researchers to compare new methods to the current state of the art.

All committee members contributed equally to this paper, listed here in alphabetical order.

Conclusion: The synthetic data provide a consistent framework to test the accuracy and robustness of QSM algorithms in the presence of noise, calcifications and minor voxel dephasing effects. Total Variation-based algorithms produced the best results among all metrics. Future QSM challenges should assess whether this good performance with synthetic datasets translates to more realistic scenarios, where background fields and dipole-incompatible phase contributions are included.

KEYWORDS

assessment, challenge, dipole inversion, quantitative susceptibility mapping, reconstruction algorithms

1 | INTRODUCTION

Quantitative susceptibility mapping (QSM) is an emerging MRI technique[1] that allows for non-invasive estimation of alterations in tissue iron concentration,[2,3] blood oxygenation,[4] and differentiation of paramagnetic and diamagnetic lesions.[5,6] QSM entails the solution of an ill-posed, ill-conditioned inverse problem that relates the acquired gradient echo (GRE) phase information, reflecting magnetic field inhomogeneities, to the underlying susceptibility distribution that is the cause of the inhomogeneities.[7,8] The QSM community has been active in the development of a wide range of reconstruction algorithms.[9,10] These developments may be categorized as inverse filtering (direct/k-space inversion), image-space (regularized iterative reconstruction), and deep learning (DL)-based approaches. The 2016 QSM Reconstruction Challenge (RC1) provided a common dataset where these algorithms could be compared.[11] This first challenge was highly successful, reflected by 27 submissions from 13 research groups. RC1 used a multi-orientation in vivo dataset. This dataset allowed to perform susceptibility tensor imaging (STI)[12] and Calculation Of Susceptibility through Multiple Orientation Sampling (COSMOS) reconstructions,[13] with the goal of mitigating susceptibility anisotropy effects while providing an adequate ground truth.

RC1 dataset and reconstruction code remained available after the challenge deadline (http://qsm.neuroimaging.at), and found widespread use for benchmarking of new dipole inversion algorithms developed after RC1 had ended.[14,15] Despite its success as a benchmark dataset, the challenge itself had limitations that limited its practical relevance. The limitations were discussed in the report paper and analyzed more quantitatively in a separate manuscript.[16] In brief, it was concluded that the estimated susceptibility tensor component χ_{33} that was used as the ground truth removed anisotropic contributions found in single-orientation phase data, which resulted in an inconsistency between the provided field map and the ground truth susceptibility. While the COSMOS solution had higher consistency, the discrepancy was still relatively large and could not be explained just by the noise and background field remnants. Including the effect of the χ_{13} and χ_{23} anisotropic contributions in addition to the effects of χ_{33} would mitigate, but not eliminate, this discrepancy. Finding an in vivo ground truth susceptibility map that matches acquired single-orientation phase data with high fidelity remains as an open problem, further complicated by the presence of background field remnants and the low signal-to-noise ratio (SNR) of the highly accelerated acquisitions. Absence of reliable ground truth data obstructs quantitative evaluation of the submitted QSM reconstructions.

Our motivation in designing a new reconstruction challenge was the following: (1) understanding the state--of-the-art of QSM algorithms (including the advent of DL techniques[14,15,17-21] since RC1), plus the identification of limitations of existing algorithms; (2) objective comparison of published algorithms, incorporating the lessons learnt from RC1, and (3) providing a new dataset for the future evaluation of dipole inversion algorithms. With these overarching goals, the design of the new challenge, RC2, started during the annual ISMRM meeting in 2018 with a call for ideas. Based on a community driven process, RC2 was designed and dramatically presented to the entire QSM community during the annual ISMRM meeting in 2019. The submissions were evaluated, and results were presented at the 5th International Workshop on MRI Phase Contrast QSM in Seoul.

In the remainder of this manuscript, we describe the challenge design rationale, provided data, evaluation criteria, and results of the challenge.

2 | METHODS

2.1 | Rationale

RC2 used a phantom with known ground truth susceptibility, which was derived from in vivo data through MR physics simulations.[22] While addressing shortcomings of RC1, attention was paid to ensure that using phantom data did not cause

new weaknesses. To avoid promoting piecewise smooth/contrast solutions, the susceptibility ground truth included physiological texture and realistic variations within structures, which was derived from R_1 and R_2^* measurements.[22]

A realistic numerical phantom made it possible to synthesize gradient echo acquisitions, which provided local field map estimates without background field remnants. Inconsistencies between the simulated data and known ground truth were restricted to complex Gaussian noise and intra-voxel dephasing effects, simulated by down-sampling from high-resolution 0.65 mm to lower-resolution 1 mm isotropic voxels[22] via k-space cropping.[7,23] Due to the modular design of the simulations, other effects such as realistic background fields and shimming may be incorporated in a future challenge design. To evaluate the results, global and region of interest (ROI) -based metrics based on the root mean squared error (RMSE) were considered. The suitability of these metrics was corroborated by inspecting their correlations with other metrics (such as the structural similarity index metric [SSIM] and others) and a visual assessment.

The RC2 was designed to take place in two stages. Stage 1 assessed the reconstruction performance of the algorithms in a realistic setting where a ground truth is not available. Participants had to determine optimal algorithmic parameters based on visual or numerical considerations (such as the L-curve) in the absence of a ground truth. In Stage 2, given that ground truth was available, participants could decide to numerically optimize their reconstruction algorithm with respect to one or all the metrics. The difference in performance between both stages was expected to provide insights into the ability to identify the optimal algorithmic parameters in the clinical setting ground truth.

2.2 | Provided data

Both in the first and second stages, two datasets were provided related to two different susceptibility models (namely SIM1 and SIM2).[22] The main difference between both datasets was the presence of an intra-hemispheric calcification (in SIM2), and different levels of susceptibility contrast between tissues, as seen in Figure 1.

As the focus of this challenge was the final dipole inversion step, only brain tissues were used to compute the field perturbations in the simulated MRI data. This ensured that there was no need for introducing a background field removal step.

Each dataset consisted of gradient echo magnitude and phase data simulated with the following parameters: repetition time (TR) = 50 ms; echo time (TE)1/TE2/TE3/TE4 = 4/12/20/28 ms; $\alpha = 15°$, field of view (FOV) = 164 × 205 × 205 mm^3 and 0.65 mm^3 isotropic voxels. Gaussian noise was added to the complex data at the same level for both datasets with peak SNR of 100 in Stage 1. Here, peak SNR was measured relative to the maximum signal of the first echo. A peak SNR of 100 resulted in an SNR of 42 (16) in white

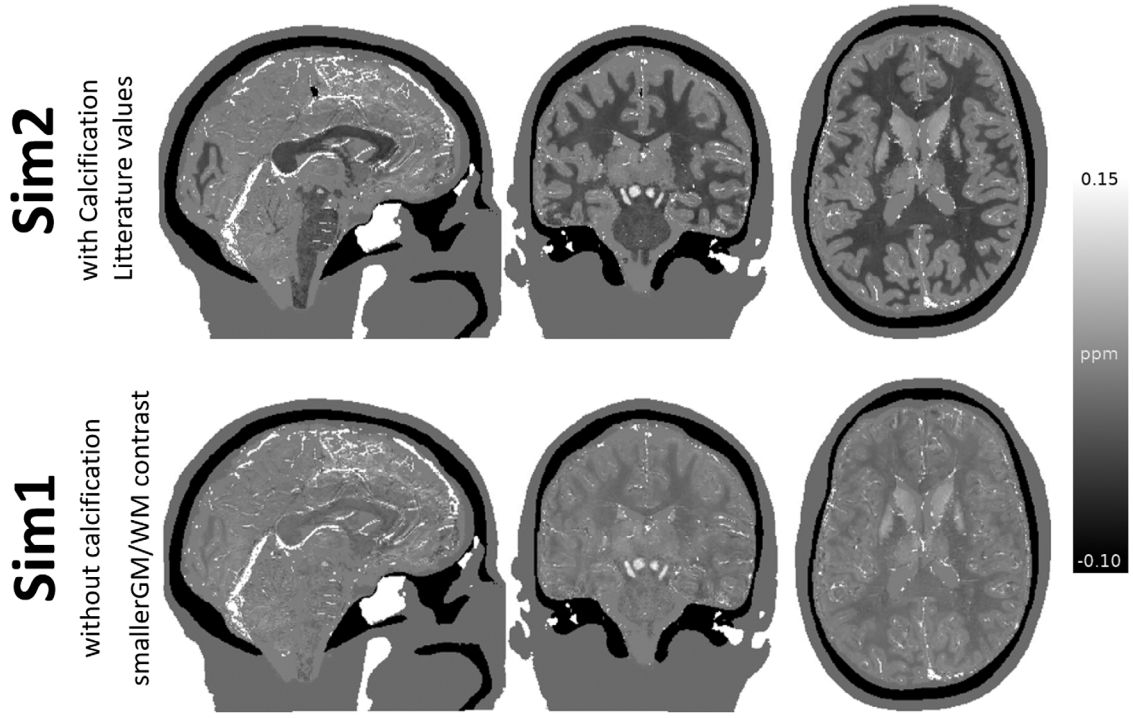

FIGURE 1 Ground truth susceptibility maps used in both steps of the challenge. Sim2 presents a larger contrast between gray and white matter than Sim1, and includes a strong calcification. For RC2, all susceptibility values outside the brain mask were set to zero to remove background fields from the simulations

matter in the first (last) echo. For Stage 2, two datasets were generated, with peak SNRs of 100 and 1000, namely SNR1 and SNR2, respectively, for each susceptibility model (SIM1 and SIM2). The same k-space cropping approach[22] was used to down-sample the high-resolution (0.65 mm to 1 mm) complex signal, ground truth susceptibility maps and segmentation labels. In the case of the susceptibility maps, the sharp edges between structures as well as the orders of magnitude larger susceptibility differences between air/bone and tissue resulted in severe Gibbs ringing artifacts, which were removed using sub-voxel shifts.[22,24] This methodology was repeated in all three spatial directions to ensure no Gibbs ringing remained.

Field maps were provided to be used at the discretion of the participants, by taking a magnitude and TE-based weighted average of three phase differences[23,25] (TE_4-TE_1, TE_3-TE_1, TE_2-TE_1). Additionally, a mask corresponding to the brain region where the QSM reconstructions would be evaluated was provided. Figure 2 shows an overview of the MRI data provided at the two stages.

In Stage 1, participants were asked to use the same processing pipelines for the two datasets (including regularization parameters or regularization optimization). Algorithm parameters for iterative and closed-form algorithms could be re-optimized for Stage 2. For DL algorithms, it was only allowed to modify certain parameters such as epochs, batch size, etc., but it was not allowed to modify the architecture of the network or to incorporate the ground truth into the training set.

2.3 | Announcement and participation

Data and instructions for participation were disseminated on a publicly accessible website (http://qsm.rocks) following the 2019 Annual Meeting of the ISMRM in Montreal.[26] The deadlines for submission of solutions for participation in Stages 1 and 2 were originally set as June 27th and August 27th, 2019, respectively, and extended to August 2nd and September 1st, 2019, respectively.

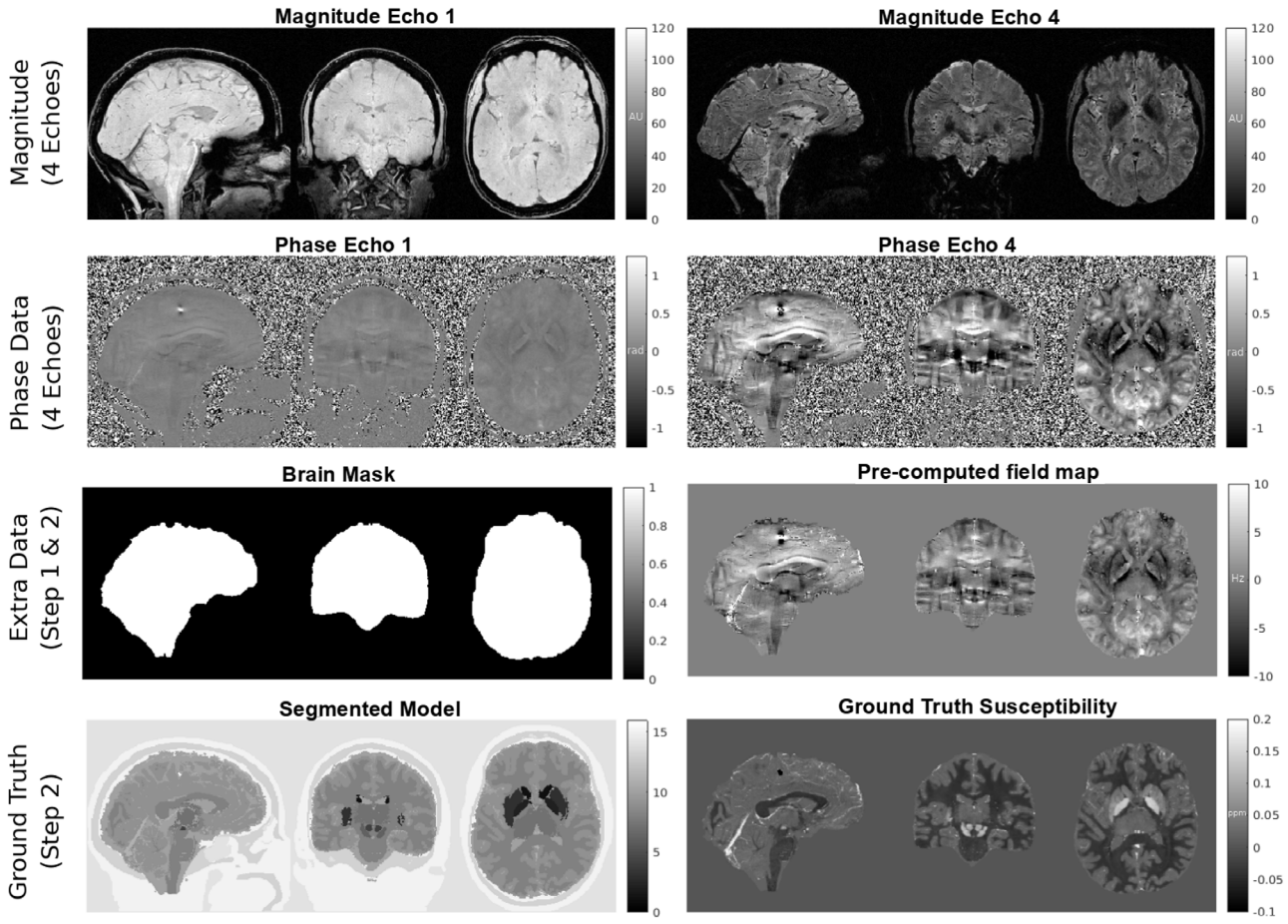

FIGURE 2 Data provided to participants in Step 1 and 2. In the first step, participants were provided with 1 mm isotropic whole brain multi-echo magnitude and phase data (1st and 2nd row), simulated in the absence of background fields. A precomputed brain field map was provided as well as a mask reflecting the region where the scripts would be evaluated (third row). In Step 2, the ground truth maps were given as well as the segmented model of the brain (4th row) that was needed to compute the various specialized metrics

2.4 | Data management and evaluation

The challenge evaluation was designed to allow a fully blinded analysis of submitted solutions. While participants were required to submit personal as well as detailed algorithmic information along with their susceptibility maps at the time of participation, we ensured that personal and algorithmic information was not accessible to the analysis team. Until the announcement of the challenge outcome, only one committee member (F.S.) had access to the identifying information of the participants. Participants were asked to provide a long-form algorithm name as well as an acronym, and a self-chosen random (blinded) 10-characters identifier for the submitted solution. Please see the Supporting Information , which is available online, for a detailed description of the required form fields.

Participants were asked to include their solution in a compressed zip-archive that was named after the self-chosen identifier and upload it through an interface to a file server (Nextcloud; hosted by F.S.'s institution). No further identifying information was supposed to be contained in the archive. Read-only access to the file server containing all submissions (but not the online form data) was given to the analysis team after August 2nd for Stage 1 data and September 1st for Stage 2 data. Algorithmic information was only shared with the analysis team after all metrics had been computed and the winners for each category were already established, on September 16th, 2019.

2.5 | Metrics

If the input phase data contains only information compatible with the magnetic dipole convolutional model, it is to be expected that most of the global metrics tend to produce the same optimal reconstruction parameters for a given algorithm.[16] Phase data inconsistencies or external contributions lead to a disagreement of the optimal parameters,[16] as shown in RC1. Given that RC2 consists of phantom-based forward simulations, to avoid unnecessary complexity in the challenge design (winning categories), only RMSE-based metrics were chosen to evaluate the global performance of the submissions.

In addition to global error metrics, we included three ROI-dependent error measurements (tissue, blood, and deep gray matter), with the aim to provide an assessment more closely related to clinical needs, such as those found in QSM-based oximetry[27,28] or the study of deep brain structures.

The metrics chosen for evaluation were ("evaluation metrics"):

- NRMSE: normalized RMSE, inside the ROI. Normalization is performed by the L2-norm of the respective ground truth.
- dNRMSE: some algorithms are known to produce underestimated results. To address this issue, we included a data demeaned and detrended RMSE score.
- dNRMSE Tissue: dNRMSE specific to White Matter and Gray Matter tissues.
- dNRMSE DeepGM: dNRMSE specific to deep gray matter structures.
- dNRMSE blood: dNRMSE for blood regions (effectively it was a dilated version of the vein mask).
- Deviation from linear slope: absolute error of the slope, derived from the demeaning and detrending process.
- Calcification streaking (CalcStreaking): error metric based on the standard deviation inside a square neighborhood surrounding the calcification, in the difference map.
- Deviation from calcification moment (calcification error): Error in the quantification of the total moment of the calcification, defined as the volume of the reconstructed calcification multiplied by its mean susceptibility.

Further details are described in the Supporting Information. To provide additional validation of the measured metrics and the winning categories of the Challenge, we also calculated SSIM, (both with the standard formulation, and a QSM-specific formulation called XSIM[29]), the high frequency error norm (HFEN),[11] the correlation coefficient (CC), mutual information (MI), mean absolute difference (MAD), and the RMSE of the gradient (first derivatives) domain (GXE),[16] collectively referred to as "additional metrics."

Scores were averaged across SIM1 and SIM2 submissions to reduce the complexity of the analysis.

2.6 | Visual rating

A visual rating scheme was designed to complement the numerical assessment of the submitted solutions. The rating was performed by each of the challenge committee members (authors) individually. A MATLAB graphical user interface was created for this purpose (Supporting Information Figure S1).

The submissions from Stage 1 of the Challenge were presented individually and in random order, different for each rater. After rating for one category, the next category was rated, using the same random order. Images for Sim1 and Sim2 were rated together, and the worse of the two submissions was used to determine the score. Each submission was shown for three fixed orthogonal slices. This set of slices covered: large veins, major deep gray matter region and the calcification region. This restriction was done for three main reasons: (1) minimize the transferred data across raters and memory requirements; (2) speed up the visual rating for the raters, allowing to quickly navigate through various submissions and previously given ratings to ensure consistency; (3) ensure that the evaluation was based

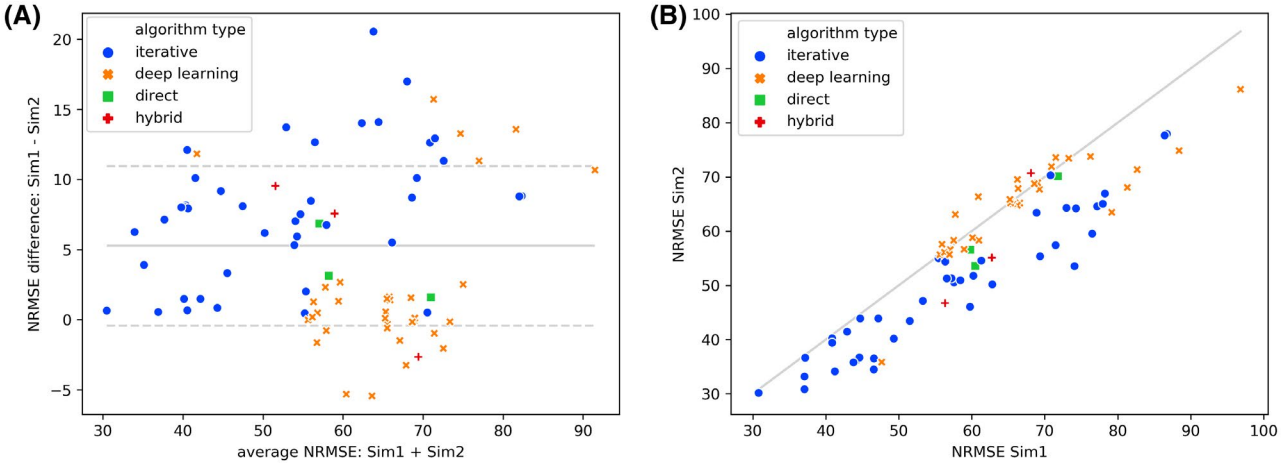

FIGURE 3 Analysis of Sim1 and Sim2 submissions: Bland-Altman plot (A), and Sim1 vs Sim2 NRMSE scores (B). Color codes show different algorithm types. Deep Learning approaches performed worse on average (higher average NRMSE), but more consistent across Sim1 and Sim2 (lower difference in NRMSE) compared to spatial-domain iterative approaches. In general, no systematic differences between Sim1 and Sim2 was noted, and therefore, further analysis was performed on averages across Sim1 and Sim2

on the same aspects of the image reconstruction increasing the consistency across ratings.

The ground truth was not shown alongside the submitted solutions but shown as if it were a submission itself. For the final score, the ratings of all raters were averaged.

Scores from 0 (best) to 3 (worst) were given depending on the artifact level in three distinct categories (streaking, unnaturalness, and noise), which are described in the Supporting Information. It is to note that visual assessment is not a comparison between QSM reconstructions and the ground truth, but a quality assessment of the naturality, and lack of noise or artifacts. A high scoring image (by visual evaluation) may contain significant errors such as wrong morphological structures or misplaced sources. This is of special concern in the case of evaluating DL algorithms. To account for this, in addition to these three categories, a binary rating was performed based on the difference map obtained by subtracting the ground truth from the submission. These are referred as visual discrepancy (white matter/gray matter, deep gray matter, and veins) metrics.

2.7 | Data preprocessing and quality checks

Prior to the evaluation, a check on the dimensions of the reconstruction and global scaling was performed. When clear mismatches existed, authors of the submissions were invited to re-submit their solution. The resubmissions were manually checked to ensure that only rescaling or spatial shifts had been applied to the new solution.

Stage 1 submissions were judged in two categories: (1) NRMSE performance, and (2) "Robustness," which counted how many appearances an algorithm had in the top 5 of any metric. Honorable mentions were awarded to the best performing algorithms in Stage 2.

3 | RESULTS

3.1 | Participation and submission statistics

We received 98 unique submissions for Stage 1. Of those submissions, 47 were submitted to Stage 2 as well (excluding resubmissions). An extended description of the number of downloads and participating countries is provided in the Supporting Information.

The majority (85%) of the submissions used an algorithm that was described either in a published journal paper or conference abstract[14,15,17-21,30-56] (Supporting Information Table S1). 15% of the submissions were not yet published. Further details regarding submitted algorithms are presented in the Supporting Information.

3.2 | Open science

Source code of the QSM algorithms was publicly available for 20% of the submissions (Supporting Information Table S1). Participants agreed to make the algorithm code available after the challenge for 43% of the submissions and they stated they would "maybe" make the code available for 37% of the submissions. None of the participants refused to make the code available on request.

3.3 | Results of stage 1

Figure 3 shows how the NRMSE differed between both simulations, as a function of the average NRMSE, for different algorithm types. Overall, algorithms showed lower errors for SIM2 with the exception of most DL-based methods, which

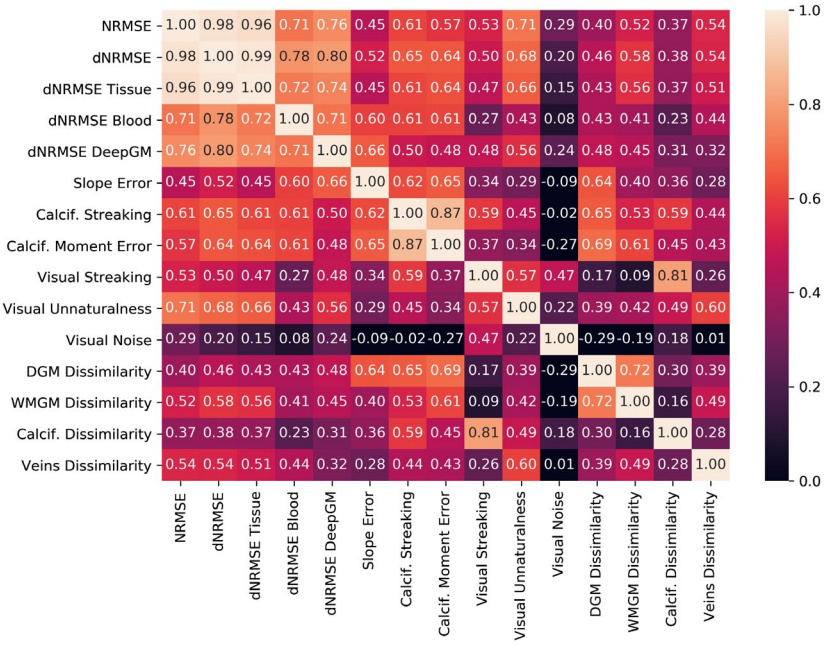

FIGURE 4 Correlation between evaluation metrics and visual metrics, for all Stage 1 submissions with NRMSE < 80

had a similar performance in both simulations. It is important to note that SIM2 has a higher contrast between gray and white matter as well as a diamagnetic calcification, which leads to a higher normalization factor in NRMSE calculation (SIM2:31.4 vs SIM1:17.4), thus, lower NRMSE scores were expected for SIM2. Interestingly the best performing deep learning method (FINE)[21] had a notably different performance on both simulations.

Figure 4 shows a correlation matrix for the evaluation metrics and the visual ratings. Only submissions with NRMSE < 80 were included (83 submissions). RMSE-based metrics were highly correlated ($R \geq 0.96$). RMSE-based metrics were also highly correlated with additional global metrics such as SSIM (XSIM variant), HFEN, CCMI, and MAD (Supporting Information Figures S2 and S3). Correlations between analytical metrics were significantly higher than between visual metrics. Visual metrics also correlated fairly well with RMSE metrics ($r = 0.66$ for the mean visual score). Note that the (Un)naturalness metric had the highest correlation with the global, deep gray matter and tissue RMSE metrics. This is particularly relevant because in a normal scenario, in the absence of a ground truth, (un)naturalness may be the criteria to choose one particular reconstruction pipeline. The visual streaking rating had its highest correlation with the calcification streaking metric. Visual discrepancy correlated fairly well with calcification metrics ($r > 0.45$) and RMSE metrics ($r > 0.40$). Veins discrepancy correlated with $r = 0.44$ with the blood dNRMSE. Visual streaking correlated very well with the calcification dissimilarity ($r = 0.81$). Visual noise was not significantly correlated with most error metrics (except GXE, which correlated poorly with other metrics). This is to be expected as in the definition of the visual noise metric, over regularized solutions (that tend to have higher RMSE) could achieve high ratings (note that the ground truth was not top-ranked in this metric). The ground truth map ranked 1st (together with six other submissions) for "visual streaking," 1st (together with two others) for "unnaturalness" (most natural), and 7th (together with 15 more) for the (lowest) visual noise level. For the visual metrics, inter-rater correlations varied from 0.30 to 0.80 for the "unnaturalness" metric (the most subjective metric) while for the streaking and noise metrics they varied from 0.46 to 0.83. The dissimilarity metrics had broader inter-rater variability (0.06 to 0.81) because of their binary nature.

Plots with the scores of the top 20 scoring submissions for each algorithm type are presented in Figure 5. Overall, iterative algorithms performed better than direct and DL algorithms, in all metrics. The performance of DL algorithms was more similar to iterative algorithms in visual analysis and was considerably worse in the NRMSE. A similar analysis is shown in Supporting Information Figures S4-S6, where the difference in performance is shown depending on whether the supplied frequency map was used as input or the full multi-echo dataset, and similarly for the case of using magnitude information. While using the magnitude information provided a small advantage, submissions using the multi-echo data (either to estimate a new field map, or as part of the algorithm's input/functional) showed on average a clearly better performance across all metrics. Most submissions that did not use the magnitude information used the provided frequency map. Scatter plots for additional metrics are shown in Supporting Information Figures S7 and S8.

Top 5 scoring results in the NRMSE metric and the Honorable Mentions (discussed in the next section) are

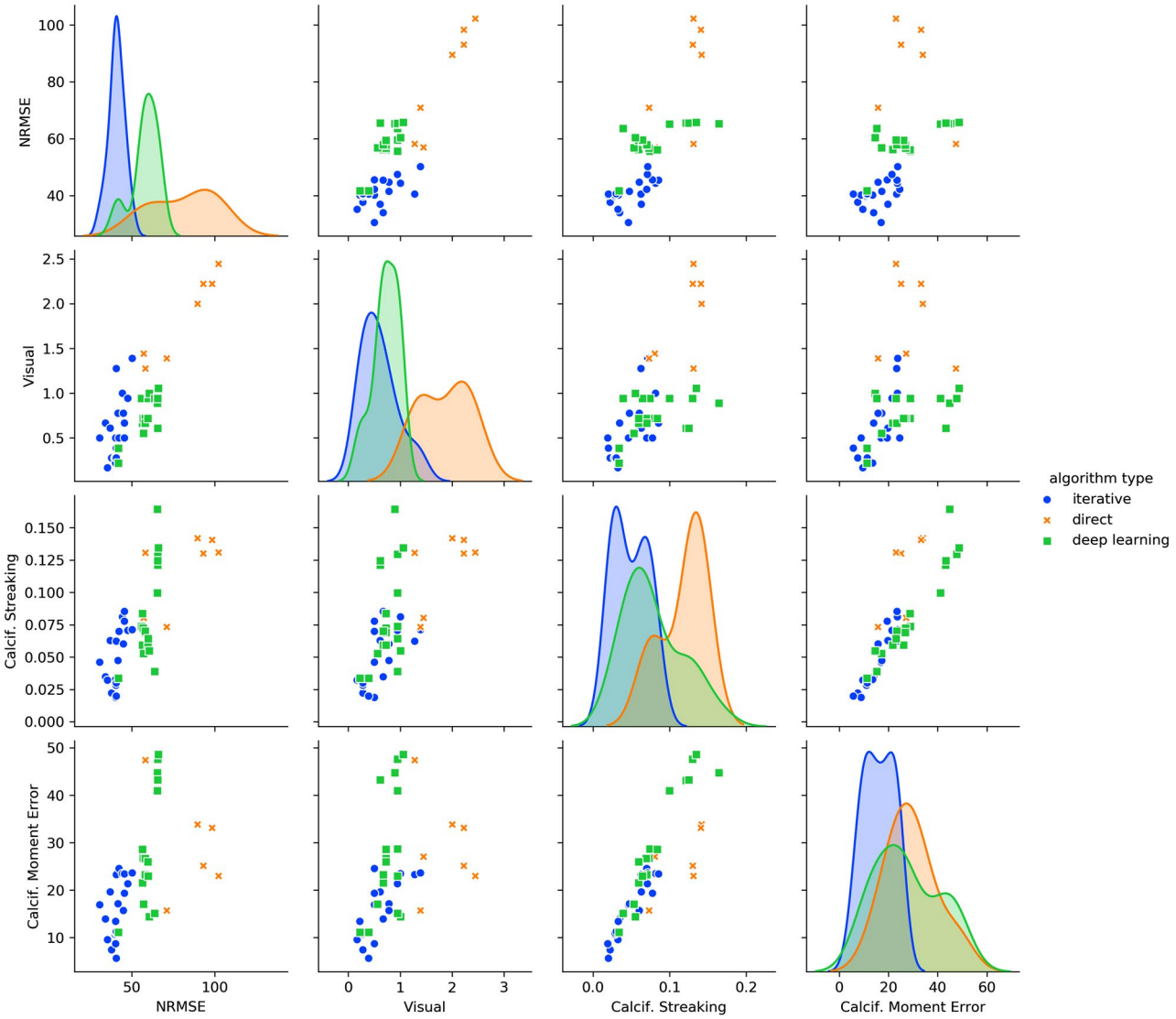

FIGURE 5 Scatterplots between selected pairs of metrics showing the top 20 NRMSE (Stage 1) submissions in each algorithm class (shown as different colors, see legend). The diagonal shows estimated histograms for each metric. In general, the top 20 solutions using iterative methods show lower error metrics than deep learning-based methods, which in turn show lower error metrics than direct inversion methods

shown in Figure 6. Coincidentally, the top five NRMSE were also the top-scoring algorithms in the "Robustness" category. The winner of the Best NRMSE category was submission 5e0YBJKZv1 (wL1L2 algorithm,[48] from Pontificia Universidad Catolica de Chile), which used a combination of L1-norm[49] and L2-norm data fidelity terms and an R_2^* weighted TV regularization (FANSI toolbox[36]). The winner of the Robustness category was submission WgpBiTiZw9 (mcTFI algorithm,[50] from Cornell University), which used an extended functional that included all individual echoes, a preconditioner derived from R_2^* data, and a morphologically enforced TV regularization (MEDI[33] toolbox). Selected metric scores for the top 10 NRMSE submissions shown in Table 1. Extended metric rankings are shown in Supporting Information Tables S2 and S3. The performance of the top-scoring algorithms using the NRMSE, CalcStreaking, and calcification error metrics is visualized comparatively in Figure 7.

3.4 | Results of stage 2

With a few exceptions, rankings for Stage 2 remained similar to those in Stage 1. As shown in Figure 8, a few DL submissions performed worse in RMSE metrics, with slight improvements depicting the calcification. Calcification-related metrics showed little or no improvement for most submissions, with some algorithms performing considerably worse. Similarly, some DL submissions had inferior RMSE performance with improved SNR, whereas most algorithms performed better (Figure 9). The calcification-related metrics

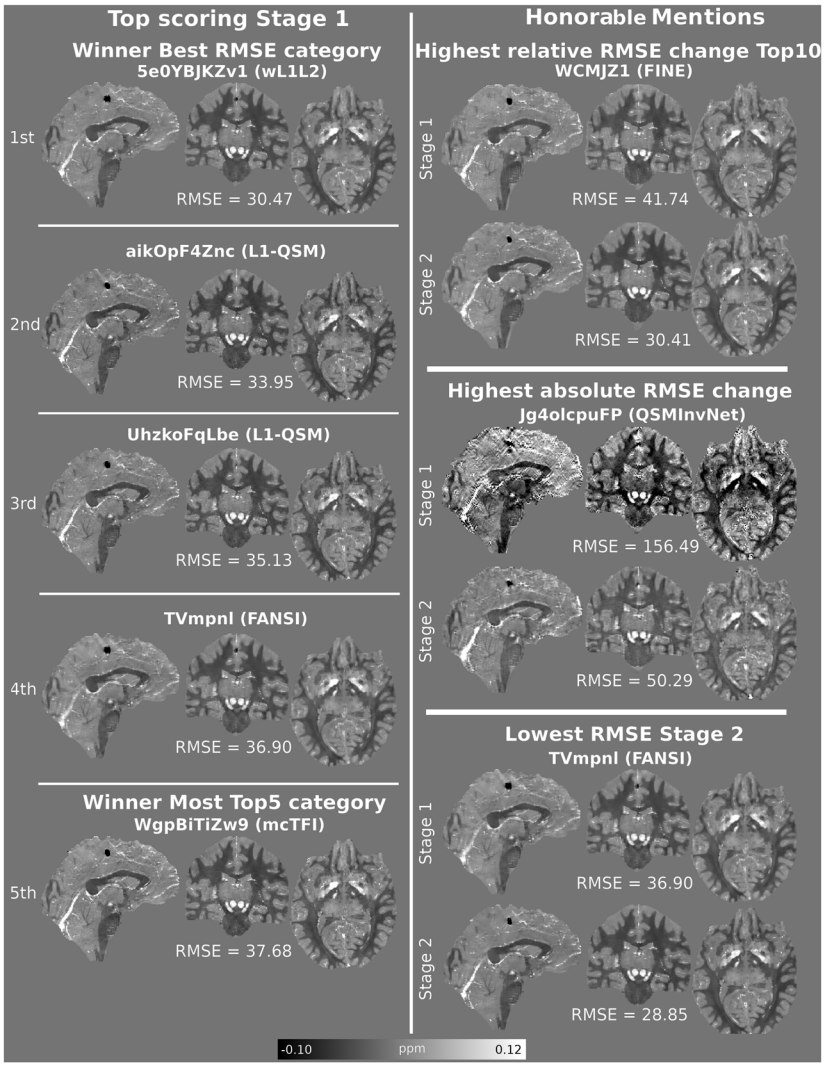

FIGURE 6 Left: Top five NRMSE scoring submissions, in descending order (top to bottom, with wL1L2 the RMSE winner). They were also the top-scoring submissions with most appearances in the top five of all metrics (robustness category). The mcTFI algorithm was the robustness winner, with 7 points. All other shown algorithms tied with 4 points. Right: honorary mentions, based on the NRMSE performance in Stage 2

seemed to be less dependent on the SNR, with mixed results across all algorithm types. Correlation and scatter plots for Stage 2 submissions are shown in Supporting Information Figures S9-S12.

The submission with the best NRMSE score (Honorable Mention) was TVmpnl (FANSI[36] algorithm, from Pontificia Universidad Catolica de Chile), which used a nonlinear data fidelity term and TV regularization. The submission with the best relative NMRSE improvement (for top-scoring algorithms) was WCMJZ1 (FINE[21] algorithm, from University of Cornell), a DL algorithm with an imposed fidelity term that modifies the pre-trained weights. This was also the overall best performing DL algorithm. Finally, the highest absolute NRMSE improvement was performed by submission Jg4olcpuFP (QSMInvNet[30] algorithm, from the Medical College of Wisconsin), which used a nonlocal encoder-decoder convolutional network. Top-scoring Stage 2 results for the NRMSE category is also shown in Table 1. Extended metric rankings are shown in Supporting Information Tables S4 and S5.

4 | DISCUSSION

Compared with the 1st QSM reconstruction challenge, the present RC2 was based on an entirely different concept. The availability of a ground truth and its two-stage-design allowed more in-depth insights and the usage of additional metrical analysis than with the in vivo dataset from RC1.

The analysis of the two datasets with significantly different SNR provided in the stage 2 showed that SNR had little impact on the ordering of the ranking (note the different scaling in Figure 9), suggesting that most algorithms behave in a similar way to different noise levels. Also, the usage of the two susceptibility models, with and without calcification, for which the participants had to submit the reconstructions with the same reconstruction parameters (although it was acceptable to submit multiple times with different parameters) did not have the expected impact. Most methods were able to deal with both datasets with similar performance (see Figure 3).

Overall, taking NRMSE as the main metric for evaluation seemed justified, as this metric was highly correlated with all

TABLE 1 Selected metric scores for top 10 scoring results in NRMSE category

Submission identifier	Preferred acronym	NRMSE	dNRMSE_Blood	CalcStreaking	Calc error	Mean visual	Top fives
Top-scoring NRMSE—Stage 1							
5e0YBJKZv1	wL1L2TV	30.5	79.3	0.046	16.9	0.50	4
aikOpF4Znc	L1-QSM	34.0	73.3	0.035	14.0	0.67	4
UhzkoFqLbe	L1-QSM	35.1	81.3	0.032	9.6	0.17	4
TVmpnl	FANSI	36.9	91.2	0.063	19.7	0.61	4
WgpBiTiZw9	mcTFI	37.7	76.8	0.022	7.5	0.28	7
wgJwSci4bs	TFIPC	39.8	74.6	0.028	10.7	0.28	2
qm2JVMNaV6	WH-QSM	40.1	88.2	0.033	13.4	0.22	3
EODh2MXvXX	WH-nlQSM	40.2	61.1	0.019	8.8	0.50	3
CWfiMHI1ij	WCMJHC1 (TFI)	40.3	83.0	0.028	11.0	0.28	0
wcmrj11111	MEDI	40.5	78.6	0.020	5.7	0.39	2
Submission identifier	**Preferred acronym**	**NRMSE**	**dNRMSE_Blood**	**CalcStreaking**	**Calc error**		**Top fives**
Top-scoring NRMSE—Stage 2							
TVmpnl	FANSI	28.9	62.9	0.013	5.8		5
5e0YBJKZv1	wL1L2TV	29.2	54.1	0.039	16.2		6
aikOpF4Znc	L1-QSM	30.3	70.9	0.028	12.6		4
WCMJZ1	FINE	30.4	50.5	0.015	4.5		7
EODh2MXvXX	WH-nlQSM	30.7	47.9	0.014	8.0		7
UhzkoFqLbe	L1-QSM	31.7	78.6	0.030	10.2		2
mojcYZ0HAA	WH-QSM	32.3	50.1	0.037	19.5		2
qm2JVMNaV6	WH-QSM	32.8	54.9	0.028	8.5		1
wgJwSci4bs	TFIPC	35.5	67.6	0.013	3.9		2
TGVmpnl	FANSI-TGV	36.5	70.7	0.016	10.1		0

Final column shows the number of appearances in the top five scoring results for each measured metric, or "robustness score." Top half shows the results for Stage 1 submissions, and bottom half for Stage 2 submissions.

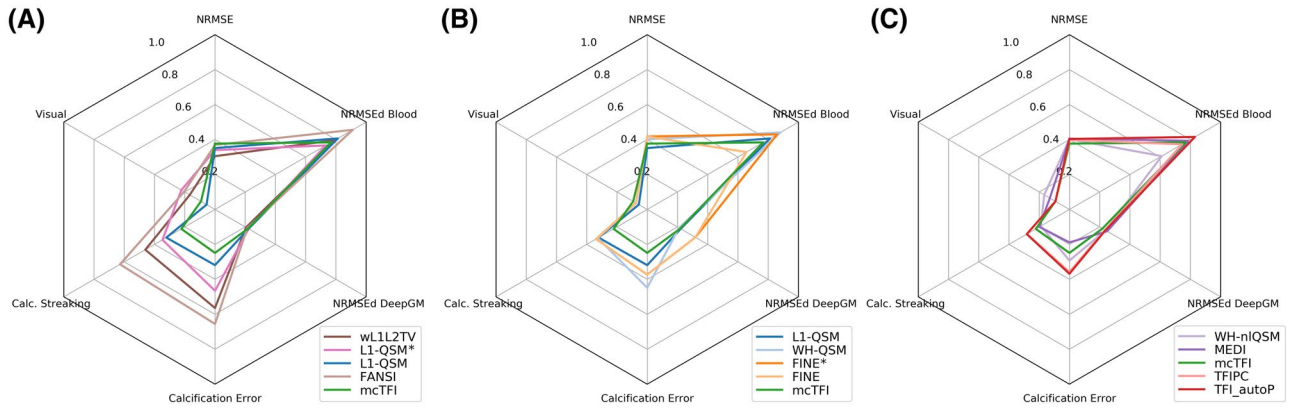

FIGURE 7 Spider plots showing metric scores for top-scoring submissions. (A) Top five submissions sorted by NRMSE. (B) Top five submissions by visual rating (average). (C) Top five submissions sorted by calcification streaking. Scores where normalized by the following factors, for displaying purposes: RMSE-based metrics: 100.0, (mean) visual: 3.0, Calc. streaking: 0.1, calcification error: 30.0. Duplicated algorithm acronyms correspond to multiple submissions with different inputs or parameters

other global metrics. This might be a natural consequence of avoiding the presence of phase incompatibilities, which generate artifacts in the reconstructions. In the presence of such errors, different metrics promote different features or image properties, thus resulting in results with different algorithm parameters when they are used to optimize the fidelity/similarity with a ground truth.[16]

For all metrics, iterative methods performed better in both stages. Metric scores for SIM 2 tended to be better than for SIM 1, mainly because of the increased WM-GM contrast.

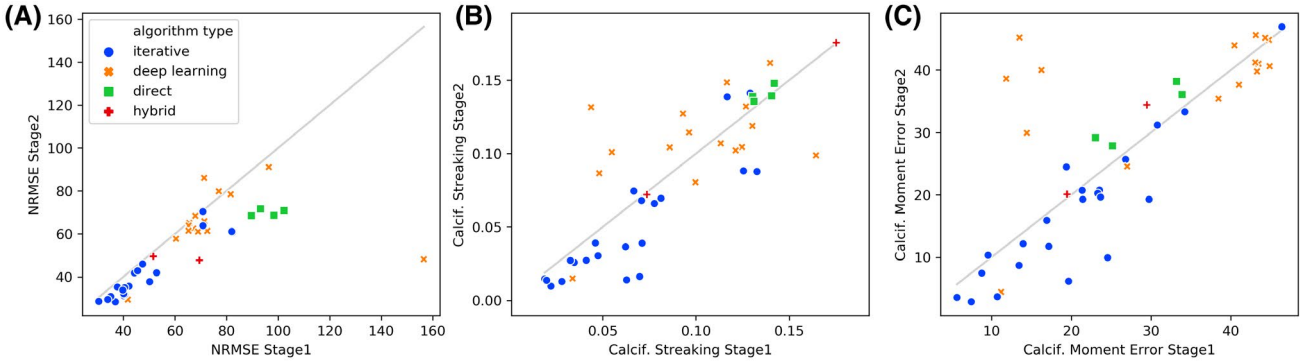

FIGURE 8 NRMSE, Calcification streaking, and calcification moment error change between Stage 1 (horizontal axis) and Stage 2 (vertical axis). The gray lines indicate the "no changes" regime. Solutions over this line worsened their scores, while solutions below this line improved their results

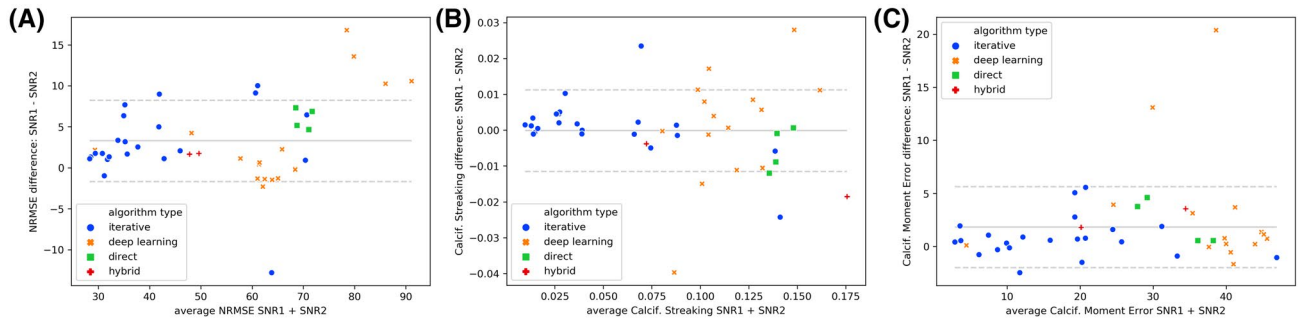

FIGURE 9 Bland-Altman plots comparing SNR1 and SNR2 Stage 2 submissions for the: NRMSE (A), calcification streaking (B), and calcification moment error (C) metrics

In particular, methods based on total variation[15,19,33,36,48-50] (and derivatives such as TGV[35,36]) were the top-scoring algorithms. Direct inversion methods (TKD[55] and Tikhonov[51,52]) had inferior performance. Interestingly and contrary to common expectations, QSM methods based on DL techniques were outperformed by "conventional" iterative methods. DL methods showed a large variance but overall performed significantly worse than iterative methods. An exception was a DL algorithm that used a physical model as the fidelity term, with similar performance to the top-scoring iterative method group.[21] Noteworthy, some DL algorithms performed worse in Stage 2. The underlying reasons for sub-par performance of DL methods could be that the networks failed to generalize well to the RC2 dataset since this fell out of the distribution of susceptibility maps they were trained on, and that the training data might be coming from a model with large data inconsistency such as COSMOS.[13,17] Poor generalization could also be the reason why some DL methods yielded worse results when using higher SNR data, and why they performed considerably worse in the calculation of the calcification moment. Also, the nonlinear nature of noise[57] might have been neglected in the DL methods either in the loss function during training, or in the forward model during inference.[58] Furthermore, DL methods are quite novel compared to the established iterative methods with more than 10 y of development and refinement. However, in addition to the aforementioned factors also other aspects beyond our current intuitive understanding might have contributed and require systematic analysis.

Overall, errors in the estimation of susceptibility values in the vessels were very high, although they correlated very well with other global metrics. It is unclear if this is a cause of intra-voxel dephasing, in such small-scale and high-contrast regions, or an intrinsic problem in QSM methods to estimate large dynamic range data. While previously reported in the literature, most algorithms showed no significant underestimation of susceptibility values in this challenge. Demeaning and detrending did not change the results in a significant way. Whether the issue of underestimation has been resolved in current QSM algorithms or underestimation is a consequence of phase inconsistencies not modeled in the RC2 remains to be investigated.

The five top-ranked algorithms in the NRMSE category also demonstrated similar error metric results in blood and DeepGM dNRMSE (Figure 7), which is expected given the high inter-metric correlations (Figure 4). Especially the error in DeepGM showed a low variance for those algorithms, which provides confidence that the susceptibilities of those iron-rich nuclei can be consistently compared between different algorithms and studies.

Generally, the submitted results allow the conclusion that multi-echo fitting or using all the echoes in the data fidelity term produced better results than using the provided field map, which is calculated from a phase difference method. Using the whole echo train improves the SNR and reduces the presence of artifacts in the vessels and the calcification. It is not clear if this conclusion can be generalizable to the case where a significant initial phase shift exists at TE = 0. Using an extended data fidelity term that incorporates all the echoes in the forward model (as in mcTFI[50]) also provided better solutions to the problem of intra-voxel dephasing in voxels surrounding the calcification.

The calcification introduced in SIM 2 basically split the algorithms in two domains, those yielding the typical streaking artefacts and another group of algorithms handling such discontinuities better by the use of preconditioners or voxel rejection strategies. A related question is whether models including estimations of the background field performed worse in this case. As expected, since no background fields were simulated, additional and unnecessary background field removal yields worse results. One example is the case of FANSI[36] vs WH-QSM,[15] where the latter includes background field remnants in its data model. The calc-streaking metric presented an evaluation of the standard deviation in the surroundings of the calcification. It tended to promote solutions with less streaking, but it was not robust. Calculating the error in the calcification moment seemed to be a more robust metric, but it has a lower correlation with the visual assessment ($r = 0.22$ for the mean visual score).

Using the magnitude information provides advantages, although it is not clear whether this is more relevant as an SNR-based data-fidelity weighting, or as a local constraint in the regularization (ie, morphological consistency between susceptibility maps and the magnitude). It is noteworthy that the edges in the magnitude images are not necessarily identical to edges in the susceptibility distribution, since only the latter is obtained using a segmentation approach. Therefore, no unrealistic advantage is expected from incorporating magnitude information, and it is likely that the conclusions remain valid also for in vivo data.

Rather a conceptual decision than a limitation was the use of synthetic GRE datasets, which were derived from several high-resolution 7T scans.[22] Although the generation of these datasets incorporates a variety of processing steps and emulates conditions close to real-world 3T brain scans, it is acknowledged that in vivo GRE acquisitions might yield different results but would also preclude the availability of a ground truth susceptibility. Although the provided ground truth scored as a natural image in the visual rating when compared to the remaining reconstructions, an experienced viewer might notice that it lacks texture in regions such as the thalamus, ultimately still appearing as a piecewise smooth model. This could have had an impact in the observation that all the top-rated RMSE solutions were TV-based. Researchers wanting to further develop their methods are encouraged to explore new realizations of the digital phantom and simulation data using the toolbox provided in the companion paper[22] of this manuscript, which could have a more natural appearance. Another issue is the inclusion of susceptibility anisotropic and microstructural effects, along with phase inconsistencies arising from flow artifacts and other effects. Further studies or challenges should be able to assess the robustness of inversion algorithm to this type of effects, more closely resembling in vivo clinical settings and not an ideal scenario, such as the one presented here.

5 | CONCLUSIONS

RC2 constituted a 2-stage challenge design based on synthetically generated brain GRE data, and yielded novel insights, which may not be obtained using an in vivo GRE acquisition. It aimed to overcome the shortcomings of the previous challenge, such as background field remnants, low SNR and the absence of a reliable ground truth. Using the RMSE as a fidelity metric in RC2 was successful indicated by a high correlation with other global metrics and the visual assessment. Iterative methods had generally better performance than DL and direct inversion methods. Incorporating the information from all the echoes and magnitude images yielded better metrics. While the synthetic phantom allowed for evaluating the performance of the algorithms in a challenging scenario with calcification, its design remains modular enough to incorporate additional considerations, such as anisotropy and background fields. The data and exemplar code are publicly available, which will facilitate the development and benchmarking of future dipole inversion algorithms.

ACKNOWLEDGMENTS

We are grateful for extensive discussions at the Electro-Magnetic Tissue Properties ISMRM study group meeting. We also thank Dr. Karin Shmueli, Dr. Pinar Ozbay, and Dr. Cristian Tejos for their valuable input in early design stages. C.L. was supported by the Austrian Science Fund (FWF grant numbers: KLI523, P30134) and BioTechMed-Graz, and J.P.M. received support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO—Grant/Award Number: FOM-N-31/16PR1056). C.M. received support from FONDECYT 1191710, PIA-ACT192064 and the Millennium Science Initiative Program—NCN17_129, of the National Agency for Research and Development, ANID. B.B. was supported by the NIH (R01 EB028797, U01 EB025162, P41 EB030006, R01 MH116173, U01 EB026996) and Siemens

Healthineers. F.S. was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR001412. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

CONFLICT OF INTEREST

Jakob Meineke is an employee of Philips Research

DATA AVAILABILITY STATEMENT

The code and data of this QSM reconstruction challenge are openly available in: https://surfdrive.surf.nl/files/index.php/s/uTvrxJ5NELbnGa5. Original anonymized csv files used for Workshop analysis and original submission forms: https://doi.org/10.5281/zenodo.3687196. All original Stage 1 submissions: https://doi.org/10.5281/zenodo.3687341. All original Stage 2 submissions: https://doi.org/10.5281/zenodo.3688702. Figure creation code: https://doi.org/10.5281/zenodo.4117549. The phantom/acquisition generation toolbox[22]: https://data.donders.ru.nl/login/reviewer-113366422/QRFk431i299BX-8bcY6ta6nQPP-MqSzG0DTIhkJrqBs.

ORCID

Berkin Bilgic https://orcid.org/0000-0002-9080-7865
Christian Langkammer https://orcid.org/0000-0002-7097-9707
José P. Marques https://orcid.org/0000-0001-8157-8864
Jakob Meineke https://orcid.org/0000-0001-8663-1468
Carlos Milovic https://orcid.org/0000-0002-1196-6703
Ferdinand Schweser https://orcid.org/0000-0003-0399-9211

REFERENCES

1. Liu C, Wei H, Gong N, Cronin M, Dibb R, Decker K. Quantitative susceptibility mapping: contrast mechanisms and clinical applications. *Tomography*. 2015;1:3-17.
2. Langkammer C, Schweser F, Krebs N, et al. Quantitative susceptibility mapping (QSM) as a means to measure brain iron? A post mortem validation study. *Neuroimage*. 2012;62:1593-1599.
3. Acosta-Cabronero J, Betts MJ, Cardenas-Blanco A, Yang S, Nestor PJ. In vivo MRI mapping of brain iron deposition across the adult lifespan. *J Neurosci*. 2016;36:364-374.
4. Fan AP, Bilgic B, Gagnon L, Witzel T, Bhat H, Rosen BR, Adalsteinsson E. Quantitative oxygenation venography from MRI phase. *Magn Reson Med*. 2014;72:149-159.
5. Deistung A, Schweser F, Wiestler B, et al. Quantitative susceptibility mapping differentiates between blood depositions and calcifications in patients with glioblastoma. *PLoS One*. 2013;8:e57924.
6. Chen W, Zhu W, Kovanlikaya IIhami, et al. Intracranial calcifications and hemorrhages: characterization with quantitative susceptibility mapping. *Radiology*. 2014;270:496-505.
7. Liu C, Li W, Tong KA, Yeom KW, Kuzminski S. Susceptibility-weighted imaging and quantitative susceptibility mapping in the brain. *J Magn Reson Imaging*. 2015;42:23-41.
8. Haacke EM, Liu S, Buch S, Zheng W, Wu D, Ye Y. Quantitative susceptibility mapping: current status and future directions. *Magn Reson Imaging*. 2015;33:1-25.
9. Kee Y, Liu Z, Zhou L, et al. Quantitative susceptibility mapping (QSM) algorithms: mathematical rationale and computational implementations. *IEEE Trans Biomed Eng*. 2017;64:2531-2545.
10. Jung W, Bollmann S, Lee J. Overview of quantitative susceptibility mapping using deep learning: Current status, challenges and opportunities. *NMR Biomed*. 2020:e4292.
11. Langkammer C, Schweser F, Shmueli K, et al. Quantitative susceptibility mapping: report from the 2016 reconstruction challenge. *Magn Reson Med*. 2018;79:1661-1673.
12. Liu C. Susceptibility tensor imaging. *Magn Reson Med*. 2010;63:1471-1477.
13. Liu T, Spincemaille P, De Rochefort L, Kressler B, Wang Y. Calculation of susceptibility through multiple orientation sampling (COSMOS): a method for conditioning the inverse problem from measured magnetic field map to susceptibility source image in MRI. *Magn Reson Med*. 2009;61:196-204.
14. Acosta-Cabronero J, Milovic C, Mattern H, Tejos C, Speck O, Callaghan MF. A multi-scale approach to quantitative susceptibility mapping (MSDI). *NeuroImage*. 2018;183:7-24.
15. Milovic C, Bilgic B, Zhao B, Langkammer C, Tejos C, Acosta-Cabronero J. Weak-harmonic regularization for quantitative susceptibility mapping. *Magn Reson Med*. 2019;81:1399-1411.
16. Milovic C, Tejos C, Acosta-Cabronero J, Özbay PS, Schwesser F, Marques JP, Irarrazaval P, Bilgic B, Langkammer C. The 2016 QSM Challenge: Lessons learned and considerations for a future challenge design. *Magn Reson Med*. 2020;84:1624-1637.
17. Yoon J, Gong E, Chatnuntawech I, et al. Quantitative susceptibility mapping using deep neural network: QSMnet. *Neuroimage*. 2018;179:199-206.
18. Bollmann S, Rasmussen KGB, Kristensen M, et al. DeepQSM—using deep learning to solve the dipole inversion for quantitative susceptibility mapping. *Neuroimage*. 2019;195:373-383.
19. Liu Z, Kee Y, Zhou D, Wang Y, Spincemaille P. Preconditioned total field inversion (TFI) method for quantitative susceptibility mapping. *Magn Reson Med*. 2017;78:303-315.
20. Polak D, Chatnuntawech I, Yoon J, et al. Nonlinear dipole inversion (NDI) enables robust quantitative susceptibility mapping (QSM). *NMR Biomed*. 2020:e4271.
21. Zhang J, Liu Z, Zhang S, et al. Fidelity imposed network edit (FINE) for solving ill-posed image reconstruction. *Neuroimage*. 2020;211:116579.
22. Marques JP, Meineke J, Milovic C, et al. QSM Reconstruction Challenge 2.0: a realistic in silico head phantom for MRI data simulation and evaluation of susceptibility mapping procedures. *BioRxiv*. 2020:316836. doi:10.1101/2020.09.29.316836.
23. Robinson S, Bredies K, Khabipova D, Dymerska B, Marques JP, Schweser F. An illustrated comparison of processing methods for MR phase imaging and QSM: combining array coil signals and phase unwrapping. *NMR Biomed*. 2017;30:e3601.
24. Kellner E, Dhital B, Kiselev VG, Reisert M. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magn Reson Med*. 2016;76:1574-1581.
25. Khabipova D, Wiaux Y, Gruetter R, Marques JP. A modulated closed form solution for quantitative susceptibility mapping—a thorough evaluation and comparison to iterative methods based on edge prior knowledge. *Neuroimage*. 2015;107:163-174.

26. Marques J, Bilgic B, Meineke J, et al. Towards QSM challenge 2.0: creation and evaluation of a realistic magnetic susceptibility phantom. In: Proc. 27th Annual Meeting of the ISMRM, Montreal, Canada. 2019. p. 1122.
27. Balla DZ, Sanchez-Panchuelo RM, Wharton SJ, Hagberg GE, Scheffler K, Francis ST, Bowtell R. Functional quantitative susceptibility mapping (fQSM). *NeuroImage*. 2014;100:112-124.
28. Özbay PS, Warnock G, Rossi C, et al. Probing neuronal activation by functional quantitative susceptibility mapping under a visual paradigm: a group level comparison with BOLD fMRI and PET. *Neuroimage*. 2016;137:52-60.
29. Milovic C, Tejos C, Irarrazaval P. Structural similarity index metric setup for QSM applications (XSIM). 5th International Workshop on MRI Phase Contrast & Quantitative Susceptibility Mapping, Seoul, Korea, 2019.
30. Liu J, Koch KM. Non-locally Encoder-Decoder Convolutional Network for Whole Brain QSM Inversion. *arXiv*. 2019. arXiv: 1904.05493.
31. Tang J, Liu S, Neelavalli J, Cheng YCN, Buch S, Haacke EM. Improving susceptibility mapping using a threshold-based K-space/image domain iterative reconstruction approach. *Magn Reson Med*. 2013;69:1396-1407.
32. Schweser F, Deistung A, Sommer K, Reichenbach JR. Toward online reconstruction of quantitative susceptibility maps: superfast dipole inversion. *Magn Reson Med*. 2013;69:1582-1594.
33. Liu T, Wisnieff C, Lou M, Chen W, Spincemaille P, Wang Y. Nonlinear formulation of the magnetic field to source relationship for robust quantitative susceptibility mapping. *Magn Reson Med*. 2013;69:467-476.
34. Ahn H, Park S, Ye JC. Quantitative susceptibility map reconstruction using annihilating filter-based low-rank Hankel matrix approach. *Magn Reson Med*. 2020;83:858-871.
35. Langkammer C, Bredies K, Poser BA, et al. Fast quantitative susceptibility mapping using 3D EPI and total generalized variation. *Neuroimage*. 2015;111:622-630.
36. Milovic C, Bilgic B, Zhao B, Acosta-Cabronero J, Tejos C. Fast nonlinear susceptibility inversion with variational regularization. *Magn Reson Med*. 2018;80:814-821.
37. Schweser F, Sommer K, Deistung A, Reichenbach JR. Quantitative susceptibility mapping for investigating subtle susceptibility variations in the human brain. *Neuroimage*. 2012;62:2083-2100.
38. Santin MD, Didier M, Valabrègue R, et al. Reproducibility of R2* and quantitative susceptibility mapping (QSM) reconstruction methods in the basal ganglia of healthy subjects. *NMR Biomed*. 2017;30:e3491.
39. Kames C, Wiggermann V, Rauscher A. Rapid two-step dipole inversion for susceptibility mapping with sparsity priors. *Neuroimage*. 2018;167:276-283.
40. Wei H, Dibb R, Zhou Y, et al. Streaking artifact reduction for quantitative susceptibility mapping of sources with large dynamic range. *NMR Biomed*. 2015;28:1294-1303.
41. Chen Y, Jakary A, Avadiappan S, Hess CP, Lupo JM. QSMGAN: Improved Quantitative Susceptibility Mapping using 3D Generative Adversarial Networks with increased receptive field. *NeuroImage*. 2020;207:116389.
42. Heber S, Tinauer C, Bollmann S, Ropele S, Langkammer C. Deep quantitative susceptibility mapping by combined background field removal and dipole inversion. In Proc. Intl. Soc. Mag. Reson. Med. 2019, p. 4028.
43. Jung W, Yoon J, Choi JY, Kim E-Y, Lee J. On the linearity of deep neural network trained QSM. In Proc. Intl. Soc. Mag. Reson. Med. 2019, p. 317.
44. Diefenbach MN, Böhm C, Meineke J, Liu C, Karampinos DC. One-dimensional k-space metrics on cone surfaces for quantitative susceptibility mapping. In Proc. Intl. Soc. Mag. Reson. Med. 2019, p. 322.
45. Cognolato F, Bollmann S, Barth M. QSMResGAN—dipole inversion for quantitative susceptibility mapping using conditional generative adversarial networks. In Proc. Intl. Soc. Mag. Reson. Med. 2020, p. 3542.
46. Bao L, Zhang H. A spatially adaptive cross-modality based three-dimensional reconstruction network for susceptibility imaging. In Proc. Intl. Soc. Mag. Reson. Med. 2020, p. 995.
47. Kames C, Doucette J, Rauscher A. ProxVNET: A proximal gradient descent-based deep learning model for dipole inversion in susceptibility mapping. In Proc. Intl. Soc. Mag. Reson. Med. 2020, p. 3198.
48. Lambert M, Milovic C, Tejos C. Hybrid data fidelity term approach for quantitative susceptibility mapping. In Proc. Intl. Soc. Mag. Reson. Med. 2020, p. 3205.
49. Milovic C, Lambert M, Langkammer C, Bredies K, Tejos C, Irarrazaval P. QSM streaking suppression with L1 data fidelity terms. In Proc. Intl. Soc. Mag. Reson. Med. 2020, p. 3257.
50. Wen Y, Nguyen T, Cho J, Spincemaille P, Wang Y. Improved signal modeling in quantitative susceptibility mapping using multi-echo complex Total Field Inversion (mcTFI). In Proc. Intl. Soc. Mag. Reson. Med. 2020, p. 3200.
51. Bilgic B, Chatnuntawech I, Fan AP, et al. Fast image reconstruction with L2-regularization. *J Magn Reson Imaging*. 2014;40:181-191.
52. Karsa A, Punwani S, Shmueli K. An optimized and highly repeatable MRI acquisition and processing pipeline for quantitative susceptibility mapping in the head-and-neck region. *Magn Reson Med*. 2020;84:3206-3222.
53. Gao Y, Zhu X, Crozier S, Liu F, Sun H. xQSM: quantitative susceptibility mapping with octave convolutional and noise regularized neural networks. 2020;arXiv:2004.06281
54. Chen Y, Jakary A, Avadiappan S, Hess CP, Lupo JM. QSMGAN: improved quantitative susceptibility mapping using 3D generative adversarial networks with increased receptive field. *Neuroimage*. 2019;207:116389.
55. Shmueli K, de Zwart JA, van Gelderen P, Li T-Q, Dodd SJ, Duyn JH. Magnetic susceptibility mapping of brain tissue in vivo using MRI phase data. *Magn Reson Med*. 2009;62:1510-1522.
56. Bao L, Li X, Cai C, Chen Z, van Zijl P. Quantitative susceptibility mapping using structural feature based collaborative reconstruction (SFCR) in the human brain. *IEEE Trans Med Imaging*. 2016;62:1.
57. Gudbjartsson H, Patz S. The rician distribution of noisy MRI data. *Magn Reson Med*. 1995;34:910-914.
58. Jochmann T, Haueisen J, Schweser F. How to train a deep convolutional neural network for quantitative susceptibility mapping (QSM). Proc Intl Soc Mag Reson Med 28 (2020). Sydney, NSW, Australia; 2020, p. 3195.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

FIGURE S1 This shows an example of the simple Matlab GUIs developed for an efficient rating process. Each rater rated one category at a time, while going through all the submissions 103 in the challenge (+ ground truth). The instruction GUI, top row appeared at the start of the rating process and remained in background in case of need to remind the criteria of rating. The Rating GUI showed three orthogonal slices of the two phantoms (SIM1 and SIM2). Raters did not have access to other slices of the reconstruction and the slices were chosen because they allowed a clear visualization of the vein in both the sagittal plane and transverse plane, the calcification in the sagittal and coronal plane and deep grey matter in the coronal and transverse planes. For the first three classes, a mouse press on the desired rating would bring the rater to the next submission. It was possible to go back to the previous rated figure to re-rate in case of mistake, in which case the current attribute ranking to that figure would appear at the top of the GUI

FIGURE S2 Stage 1 correlation between evaluation metrics and visual assessment for A) top 5 submissions, B) submissions with NMRSE<80, and C) all submissions

FIGURE S3 Correlation between evaluation metrics, visual metrics and additional metrics, for all submissions with NMRSE<80 in Stage 1

FIGURE S4 Stage 1 scatterplots between selected pairs of metrics comparing the use of magnitude information as prior or weighting for the reconstruction. The diagonal shows estimated histograms for each metric

FIGURE S5 Stage 1 scatterplots between selected pairs of metrics comparing the use of multi-echo and frequency maps as input for reconstruction. The diagonal shows estimated histograms for each metric. Overall, algorithms using multi-echo data showed lower errors

FIGURE S6 Stage 1 scatterplots between selected pairs of metrics comparing the use of multi-echo and frequency maps as input for reconstruction for top 9 submissions that didn't include the magnitude information as prior or weight

FIGURE S7 Stage 1 scatterplots between NMRE and additional metrics XSIM, HFEN and Correlation Coefficient. Only the best 20 RMSE submissions for each algorithm type are shown

FIGURE S8 Stage 1 scatterplots between NMRE and additional metrics Mutual Information, Mean Absolute Error (MAD) and RMSE measured on the gradient (first derivative) domain (GXE). Only the best 20 RMSE submissions for each algorithm type are shown

FIGURE S9 Correlation matrix for Stage 1 submissions that corresponds to submissions to Stage 2 (top) and all Stage 2 submissions (bottom). Correlations between metrics significantly increased in comparison to Stage 1

FIGURE S10 Stage 2 scatterplots between selected pairs of metrics showing the top 20 submissions in each algorithm class (shown as different colors, see legend). The diagonal shows estimated histograms for each metric. Overall, iterative methods significantly performed better than Deep Learning and Direct approaches, showing larger relative improvements (see main Figure 5)

FIGURE S11 Stage 2 scatterplots between NMRE and additional metrics XSIM, HFEN and Correlation Coefficient. Only the best 20 RMSE submissions for each algorithm type are shown

FIGURE S12 Stage 1 scatterplots between NMRE and additional metrics Mutual Information, Mean Absolute Error (MAD) and RMSE measured on the gradient (first derivative) domain (GXE). Only the best 20 RMSE submissions for each algorithm type are shown

TABLE S1 Publications that describe each submitted algorithm, and code repositories, if available. Multiple submissions with the same algorithm (but significantly different parameters or results) are grouped

TABLE S2 Top 10 best scoring submissions and scores in Stage 1, for each measured metric (official and extended)

TABLE S3 Top 10 best scoring submissions and scores in Stage 2, for each measured metric (official and extended)

How to cite this article: QSM Challenge 2.0 Organization Committee, Bilgic B, Langkammer C, Marques JP, Meineke J, Milovic C, Schweser F. QSM reconstruction challenge 2.0: Design and report of results. *Magn Reson Med*. 2021;00:1-15. https://doi.org/10.1002/mrm.28754