Received: 28 April 2023 | Revised: 20 February 2024 | Accepted: 21 February 2024

DOI: 10.1002/mrm.30081

RESEARCH ARTICLE

Magnetic Resonance in Medicine

Reproducibility of arterial spin labeling cerebral blood flow image processing: A report of the ISMRM open science initiative for perfusion imaging (OSIPI) and the ASL MRI challenge

Andre M. Paschoal[1,2] | Joseph G. Woods[3,4] | Joana Pinto[5] | Esther E. Bron[6] | Jan Petr[7] | Flora A. Kennedy McConnell[8,9,10] | Laura Bell[11] | Maria-Eleni Dounavi[12] | Cassandra Gould van Praag[3,13] | Henk J. M. M. Mutsaerts[14,15] | Aaron Oliver Taylor[16] | Moss Y. Zhao[17] | Irène Brumer[18] | Wei Siang Marcus Chan[18] | Jack Toner[9,19] | Jian Hu[9,19] | Logan X. Zhang[5] | Catarina Domingos[20] | Sara P. Monteiro[20] | Patrícia Figueiredo[20] | Alexander G. J. Harms[6] | Beatriz E. Padrela[14,15] | Channelle Tham[21] | Ahmed Abdalle[22] | Paula L. Croal[8,9] | Udunna Anazodo[23]

Correspondence
Udunna Anazodo, Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, QC, Canada.
Email: udunna.anazodo@mcgill.ca

Funding information
Linacre College, University of Oxford, Grant/Award Number: 220204/Z/20/Z; European Union; Fundo de Apoio ao Ensino, à Pesquisa e Extensão, Universidade Estadual de Campinas, Grant/Award Number: 2589/23; DEBBIE, Grant/Award Number: JPND2020-568-106; Netherlands Organisation for health Research and Development; Wellcome Trust, Grant/Award Number: Sir Henry Dale Fellowship 220204/Z/20/Z; ISMRM Perfusion Study Group; Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Number: 2022/06496-7; Netherlands Enterprise

Abstract

Purpose: Arterial spin labeling (ASL) is a widely used contrast-free MRI method for assessing cerebral blood flow (CBF). Despite the generally adopted ASL acquisition guidelines, there is still wide variability in ASL analysis. We explored this variability through the ISMRM-OSIPI ASL-MRI Challenge, aiming to establish best practices for more reproducible ASL analysis.

Methods: Eight teams analyzed the challenge data, which included a high-resolution T1-weighted anatomical image and 10 pseudo-continuous ASL datasets simulated using a digital reference object to generate ground-truth CBF values in normal and pathological states. We compared the accuracy of CBF quantification from each team's analysis to the ground truth across all voxels and within predefined brain regions. Reproducibility of CBF across analysis pipelines was assessed using the intra-class correlation coefficient (ICC), limits of agreement (LOA), and replicability of generating similar CBF estimates from different processing approaches.

Results: Absolute errors in CBF estimates compared to ground-truth synthetic data ranged from 18.36 to 48.12 mL/100 g/min. Realistic motion incorporated

For affiliations refer to page 849

Agency; Heart Foundation, Grant/Award Number: 03-004-2020-T049; EU Joint Program for Neurodegenerative Disease Research; E!, Grant/Award Number: 113701; Fundação para a Ciência e a Tecnologia, Grant/Award Numbers: PTDC/EMD/EMD/29686/2017; SFRH/BD/120006/2016; DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020; 'La Caixa' Foundation, Grant/Award Number: LCF/PR/HP22/0053; ProgramaOperacional Regional de Lisboa, Grant/Award Number: LISBOA-01-0145-FEDER-02968

into three datasets produced the largest absolute error and variability between teams, with the least agreement (ICC and LOA) with ground-truth results. Fifty percent of the submissions were replicated, and one produced three times larger CBF errors (46.59 mL/100 g/min) compared to submitted results.

Conclusions: Variability in CBF measurements, influenced by differences in image processing, especially to compensate for motion, highlights the significance of standardizing ASL analysis workflows. We provide a recommendation for ASL processing based on top-performing approaches as a step toward ASL standardization.

KEYWORDS

ASL, cerebral blood flow, challenges, image analysis, reproducibility

1 | INTRODUCTION

Arterial spin labeling (ASL) is a non-invasive perfusion MRI method for quantitative measurement of perfusion-related physiological parameters by magnetically labeling the arterial blood prior to image acquisition.[1] Although the translation of ASL toward clinical use is continuously increasing, its success relies on the standardization of acquisition protocols to make clinical exams comparable as well as standardization of image processing pipelines for reproducible and reliable quantification of cerebral blood flow (CBF), among other parameters. The consensus guideline for ASL acquisition in 2015 by the International Society of Magnetic Resonance in Medicine (ISMRM) Perfusion Study Group and the European COST-AID Action (BM1103) was an important effort toward standardization of acquisition protocols.[2] The growing use of ASL, driven by availability of acquisition schemes on clinical scanners, have placed a demand for data processing strategies that can meet clinical needs. This includes the need for reproducible, reliable, and accurate, measurements that are compatible with single-subject analysis. However, the proliferation of data processing methods has resulted in heterogeneous image processing pipelines and consequently, ASL measures that are not comparable across studies. Therefore, there is a need to harmonize ASL data processing methods, to increase the wide adoption of ASL in routine clinical use.

A common approach for establishing consensus across imaging processing methods is through focused data analysis Challenges, which draw on community expertise to validate analysis methods, characterize analysis practices, and provide quantitative benchmarks.[3–5] For example, the series of DTI data analysis competitions organized over the past decade by the diffusion MRI community introduced benchmarks for algorithms designed for reconstructing and quantifying white matter fibers.[6] Data Challenges provide a curated dataset with known or pre-defined ground truth that are open to groups from around the world to analyze and submit results for evaluating analysis methods.[3,6–11] Motivated by the lack of standardized approach for analysis of ASL for perfusion quantification, the ISMRM Open Science Initiative for Perfusion Imaging (ISMRM OSIPI) organized the ASL MRI Challenge. OSIPI is an initiative of the ISMRM Perfusion Study Group that aims to improve the reproducibility of perfusion imaging research to speed up translation of perfusion tools for clinical practice. The ISMRM OSIPI ASL MRI Challenge is the primary aim of the OSIPI Task Force 6.1 2-y roadmap. The challenge launched in February 2021[12] was designed to establish ASL image analysis best practices by comparing data processing approaches implemented on ground truth data. Best practices were determined with consideration to the image analysis steps, CBF quantification accuracy, and reproducibility of pipelines according to the completeness of the method documentation. Using a digital reference object (DRO) as the ground-truth data, the results provide a head-to-head comparison of different analysis tools including order of the steps in the workflow, to understand sources of errors and variability in CBF quantification. Capturing analysis approaches of the research community, and subsequently identifying and understanding sources of variability is paramount to establishing consensus on analysis workflow and providing benchmarks to ultimately move ASL closer to routine clinical use.

In this paper, we describe a framework to standardize ASL image processing through an ASL data analysis challenge. We applied this framework to eight submissions to report the consensus on analysis of single post-label delay pseudo-continuous ASL (PCASL) data. Details of the challenge design, including the PCASL DRO, scoring criteria, summary of pipelines provided by the teams, and CBF results along with sources of errors and variability, are outlined below. By analyzing the results, we ultimately

provide a benchmark of ASL pipelines and recommendations for ASL image processing and analysis steps. A comprehensive review of ASL image processing and analysis steps is beyond the scope of this work and has been summarized by other groups.[13]

2 | METHODS

2.1 | Challenge overview

The ISMRM OSIPI ASL MRI Challenge (https://challenge.ismrm.org/forums/topic/osipi-asl-challenge) started with a pre-launched phase, which included development of the data and user manual (https://challenge.ismrm.org/wp-content/uploads/2021/05/OSIPI_ASL_Challenge_Manual.pdf) with instructions. Submissions were accepted until January 2022, and no submission was excluded. The complete design is summarized in the Figure S1. The participating teams were asked to fill a registration form to capture their team composition including level of expertise, download and analyze the data, prepare documentation containing the description of the pipeline used, and submit the outputs for evaluation. Each team was required to submit quantified CBF maps and text files containing their estimate of the mean gray matter (GM) and white matter (WM) CBF, calculated as the mean CBF within the GM and WM masks created by their pipeline. GM and WM partial volume corrected (PVC) CBF maps were an optional result that teams could also include in their submission. The submissions were scored as outlined below to determine the performance of the methods across accuracy, reproducibility, and documentation quality metrics. The optional PVC results were excluded from the performance scores. Best practices and recommendation guidelines were drawn from top performing methods. A maximum of two participants from each submission team were invited to join the manuscript as a co-author after the challenge closed. None of the submission team members including developers of the pipeline participated in design of the challenge data or performed any of the submission analysis. The organizers who performed the error and reproducibility analysis were blinded to the submission teams' identities.

2.2 | Challenge data

The datasets consisted of 10 PCASL[14] datasets in ASL BIDS format[15] comprising of four-dimensional (4D) time-series of control-label pairs, a calibration (M0) image, and a high-resolution T1-weighted anatomical image. The data are archived in the Open Science Framework (OSF) repository[16] and were generated from the following two sources:

1. Healthy Older Population Data: An existing healthy older population DRO generated from a subset of the European Prevention of Alzheimer Dementia (EPAD) study[17] was used to establish variability for a single "real-world" dataset. The population DRO was created by averaging the EPAD dataset of 84 healthy older participants (67.1 ± 7.1 y) acquired on 3T (Philips) MRI scanners. The EPAD dataset comprised of ASL scans acquired using a PCASL sequence with a 2D gradient-echo EPI readout and the following imaging parameters (TE/TR = 10.49/4800 ms, 1650 ms labeling duration, post-label delay (PLD) of 2025 ms, acquisition matrix size of $64 \times 64 \times 36$ covering the whole brain at a voxel size of $3.4 \times 3.4 \times 4.5\ mm^3$, and 30 control-label pairs). The M0 images were acquired with the ASL sequence using TR of 10 s and no label or background suppression pulses. The anatomical T1-weighted images were acquired using 3D T1-weighted turbo field echo (TFE) sequence (TE/TR = 3.09/6.77 ms, flip angel [FA] = 9°, voxel size of $1.20 \times 1.05 \times 1.05\ mm^3$). Conventional image processing was applied to the imaging datasets to co-register the anatomical to the perfusion datasets and a simplified single-compartment model was used to generate voxel-wise CBF maps. The CBF maps were spatially normalized by transformation to the Montreal Neurological Institute (MNI) standard space and then averaged to generate spatially normalized GM and WM CBF maps. The GM and WM CBF were used to synthesize the healthy older population ASL DRO.
2. Synthetic Data: Nine ASL phantom datasets were simulated using an existing ASL-DRO[18] to provide ground-truth data. Each set was simulated in subject-specific anatomical space, using 3D T1-weighted images from healthy young adults[19] (MPRAGE, TE/TR = 2.12/2400 ms, FA = 8°, 1 mm isotropic spatial resolution) and subsequently downsampled to the ASL spatial resolution ($4 \times 4 \times 4\ mm^3$). Input parameters (CBF, arterial transit time, M0, T1, T2 and T2*) were defined per tissue type (GM/WM/cerebrospinal fluid-CSF) for simulation of ASL signal according to the general kinetic model and 3D PCASL timeseries acquired at 3T, with a gradient-echo readout, consistent with acquisition consensus recommendations (PLD: 1800 ms, TE/TR: 10.4/4800 ms, 1800 ms labeling duration, M0 TR = 10s, 30 control-label pairs).[2] For challenge purposes, simulation parameters were withheld while the challenge was open. To capture variability and accuracy in measuring pathological changes, mean tissue perfusion was calculated in six regions of interest (ROIs): namely right primary motor cortex, left hippocampal GM, corpus callosum, frontal GM, cerebellum, bilateral posterior cingulate cortex.

Two of the ROIs were modulated in six of the nine datasets by increasing or decreasing the CBF values by 10%–30% within the regions to simulate pathological CBF changes. To evaluate the robustness of the analysis workflows to subject motion, head motion was simulated across a realistic range of motion in three of the six pathological datasets, mirroring real-world conditions of motion artifacts more common in pathological scans compared to a healthy subject scan. This was achieved by including uniformly distributed random variations (mean = 0, SD = 0.1) in rotation and translation levels across all directions and larger variations in 10% of the volumes between −2 and 2 mm/degrees. The three motion-datasets were simulated for each condition based on T1-weighted images from three healthy young adults randomly selected from the Human Connectome Project database

2.3 | Scoring

For each entrant, a single total score of 100 was given based on three weighted criteria:

1. Accuracy: (weighted 60%) Assessed as the mean absolute error between the ground truth CBF and submitted CBF maps and regional mean values from the synthetic datasets This was done for both GM ROIs and ROIs in which perfusion was altered to simulate pathology (see statistical analysis below). All error measurements were calculated using custom scripts in MATLAB (version 2018b, The MathWorks, Natick, MA). Violin plots of CBF measurements were generated using the ggplot library for R.[20] For each dataset, the absolute difference between the submitted mean GM CBF and the ground truth mean GM CBF was calculated (referred to as GMerror). The mean absolute difference across voxels within the ground truth GM mask between the submitted CBF maps and the ground truth CBF maps was also calculated for each dataset (GM Voxerror). For the six pathological datasets, the mean absolute difference between submitted CBF maps and the ground truth CBF maps was calculated within each of the two ground truth ROI masks (ROI 1 Voxerror and ROI 2 Voxerror, respectively). To generate an overall accuracy score for each submitting team, the errors described above were combined as follows. For each category of error (see Table 1 for the 10 categories included), the absolute errors were normalized between 0 and 1 across the submitting teams (0 being the team with the lowest error for a particular category), then averaged across categories for each team, giving a final normalized error ranking. This error ranking was subtracted from 1 and multiplied by 60, to give an overall accuracy score between 0 and 60, where the team with the highest score had the smallest errors on average. The accuracy scores were assessed by J.G.W., M.E.D., and A.M.P.
2. Reproducibility[21]: (weighted 30%) Assessed to evaluate the ability of the analysis methods to produce reproducible results, based on intra-class correlation coefficient (ICC), Bland–Altman limits of agreement (LOA), and the completeness of the information provided in the documentation of the methods (i.e., the extent to which the information provided was sufficient to reproduce the steps performed and obtain similar results). The LOA between the submissions and the ground-truth were assessed using the Bland–Altman analysis,[22] where the lower the LOA within a confidence interval (95%), the higher the agreement. The two-way random-effects ICC model[23] was used to evaluate the degree of agreement between the CBF maps provided by the teams and the ground truth maps for each simulated synthetic dataset. The ICC was performed voxel wise using the equation below.

$$\mathrm{ICC}\,(2,1) = \frac{\sigma_{r2}}{\sigma_{r2} + \sigma_{c2} + \sigma_{v2}}$$

where σ_{r2} is the between-subject variance, σ_{c2} is the variance between repeated measures or the systematic bias and σ_{v2} is the noise or error in the measurement. An ICC close to 1 indicates a high agreement or similarity between team's results and ground truth and essentially how reproducible the CBF measures are.[23] Three reproducibility categories were scored from 1 to 10 and then combined to a total of 30. All statistical analysis and plots were performed using R.[24] To standardize the documentation and ensure participants describe pertinent steps/approaches and tools used, teams were advised to follow the Committee on Best Practice in Data Analysis and Sharing (COBIDAS) ASL specific recommendations list[25,26] (Table S1 for the suggested items to be reported). The reproducibility scores were assessed by A.P., C.T., A.A., and U.A. The documentation provided was used to rerun the analysis and replicate results for reproducibility assessment. Replication was performed by C.T. and A.A. and neither of them participated in the challenge, as members of submission teams or in the design. The participating teams and their colleagues did not conduct analysis of the submission, documentation evaluation, or scoring of results.
3. Documentation quality: (weighted 10%) Assessed based on a subset of the QUACK (quality, useability,

accuracy, and conciseness) criteria[27]: (1) Quality-clarity, formatting, and structure; (2) Useability-ease of comprehension of text, graphics, or other content for the purposes of reproducing the pipeline; (3) Accuracy-precision of numbers and descriptors; (4) Conciseness-appropriate levels of detail and adherence to the page limit (Table S2). The documentation quality scores were assessed by C.G.P., and its final value was the average score of each QUACK criteria.

3 | RESULTS

3.1 | Challenge's entries

A total of eight entries were submitted and the identity of the teams are withheld. A brief description of the processing pipeline by each team is included below, while a more detailed description is provided in the Supplementary Information. Three teams submitted their COBIDAS checklist, and teams whose method descriptions met the COBIDAS guidelines are summarized in Table S1. All the teams provided a copy of their code or a tutorial with the commands that they used to postprocess the data.

TEAM 1. Used custom MATLAB (2020b) scripts combined with SPM 12 preprocessing tools. The pipeline included brain extraction, spatial smoothing of the M0 image, registration of the T1 structural image to the M0 image, motion correction of the ASL data, and CBF quantification.

TEAM 2. Used the oxford_asl analysis tool within the FSL (version 6.0.4) framework. The pipeline included brain extraction, spatial smoothing of the M0 image and of the ASL images, registration of the ASL data to the M0 image, registration of the T1 image, and CBF quantification following the Buxton model.[28]

TEAM 3. Used the Quantiphyse[29] tool. Image processing included spatial smoothing of ASL images, registration of the T1-weighted image, motion correction of the ASL time series, registration of the ASL data to the M0 image, and CBF quantification.

TEAM 4. Used the BASIL toolbox[30] implemented in FSl (6.0.0). The pipeline consisted of spatial smoothing of ASL images, registration of the T1-weighted image, motion correction, and CBF quantification. Brain extraction used the ANTs toolbox.

TEAM 5. Used ExploreASL with the steps outlined in the ExploreASL documentation and paper,[31] including brain extraction, spatial smoothing, registration of T1-weighted image, motion correction, and CBF quantification.

TEAM 6. Used ASL MRI cloud,[32] including T1 segmentation, motion correction of ASL time series, and CBF quantification.

TEAM 7. Used the Iris pipeline,[33] which consisted of registration to ASL space, registration of the T1-weighted image, motion correction, slice time correction, and CBF quantification. Brain extraction was completed using a multi-atlas segmentation approach.

TEAM 8. Used the LOFT ASL[34] software including the brain extraction, segmentation, motion correction, coregistration of structural and ASL images and CBF quantification.

3.2 | Overall submission results

An overview of the CBF measurements provided by each team for an illustrative healthy condition is shown in Figure 1. The quantitative CBF measures are summarized using violin plots in Figure S2, where DRO 1 to 3 reflected normal condition, DRO 4 to 6 simulated a disease condition, and DRO 7 to 9 simulated disease condition including

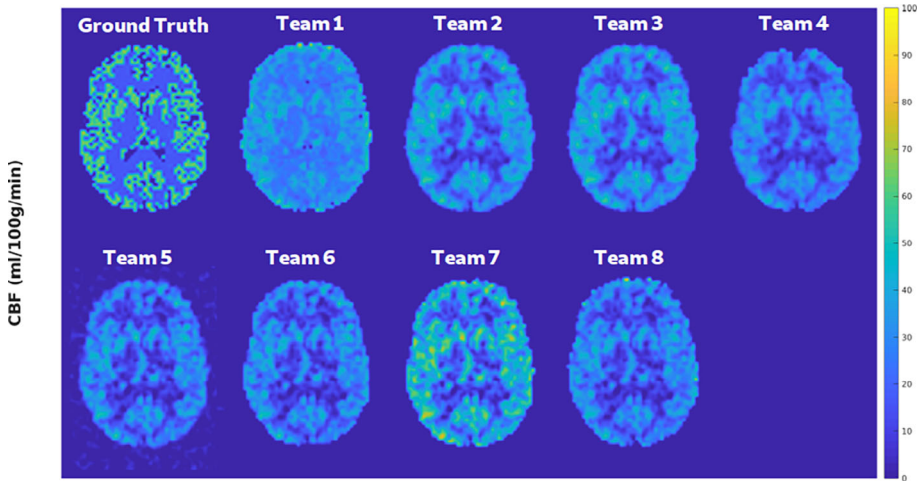

FIGURE 1 Example CBF images (mL/100 g/min) for the synthetic DRO and the CBF maps provided by the teams for the normal condition. The first map is an illustrative slice for the groundtruth, and the next eight maps are an illustrative slice for teams 1–8, respectively.

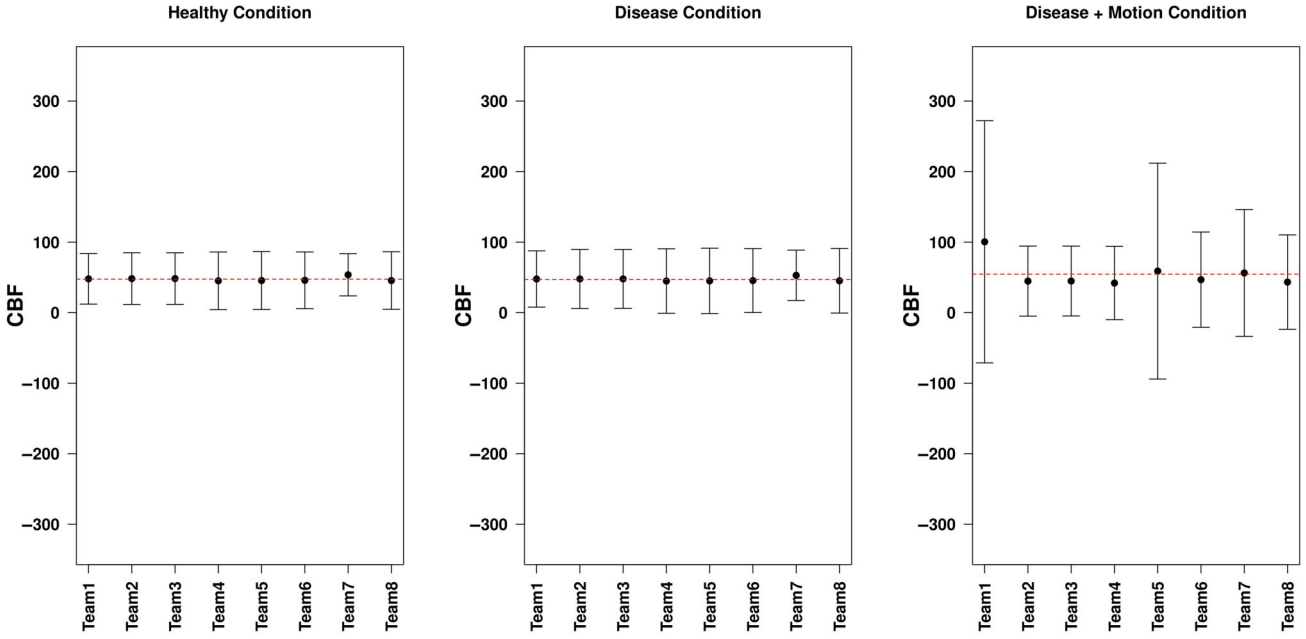

FIGURE 2 Bland–Altman LOA of all the ASL Challenge submissions against the ground truth CBF maps for the normal condition (A), disease condition (B), and disease plus motion condition (C). The horizontal dashed red line is the mean CBF in GM, and the errors bar represents the range of the LOA.

TABLE 1 Mean absolute CBF errors in mL/100 g/min for all teams' submissions and simulated conditions.

Parameter	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
Normal GM error	24.17	18.64	24.59	25.25	26.36	31.07	**14.95**	27.03
Normal GM voxerror	21.01	20.00	20.10	27.72	28.37	25.62	**11.59**	25.62
Disease GM error	24.49	19.01	25.87	25.54	26.89	32.03	**16.22**	28.89
Disease GM voxerror	23.11	23.05	23.55	30.77	28.23	28.52	**14.56**	28.55
Disease ROI 1 Voxerror	23.01	17.71	17.68	23.69	23.46	23.33	**14.06**	23.56
Disease ROI 2 Voxerror	12.64	11.28	11.79	16.53	15.65	15.75	**9.86**	16.36
Disease + motion GM error	123.43	19.06	27.08	26.87	14.16	29.70	**12.02**	53.21
Disease + motion GM Voxerror	102.57	**26.91**	27.02	34.09	63.02	32.28	37.52	32.37
Disease + motion ROI 1 Voxerror	77.21	**22.31**	22.46	27.07	53.27	28.55	29.96	141.21
Disease + motion ROI 2 Voxerror	49.56	**15.87**	15.95	19.18	31.81	19.25	22.88	17.55
Overall errors	48.12	19.38	21.61	25.67	31.12	26.61	**18.36**	39.43
Relative difference (%)	N/A	−2.94	−230.53	−15.42	N/A	N/A	N/A	**−0.80**

Note: The lowest error values are highlighted in bold for each simulated data and condition, as well as for the overall bias estimate. The relative difference is the change in overall errors between replicated CBF values and values submitted for the challenge.

subject motion. Overall, all the submissions showed good agreement with ground truth CBF values for both normal and disease conditions, demonstrated by the 95% LOA as shown in Figures 2A and 2B. For all teams, the LOA centered around the mean CBF with fairly narrow variability or deviation from the mean. When motion was included, the variability between the ground truth and submissions and the absolute errors increased. The mean absolute CBF error for each synthetic dataset for each team is summarized in Table 1 and Figure 3. A visual inspection of Table 1 and Figures 2 and S2 revealed that team 7 had the smallest CBF errors and the lowest 95% LOA for normal conditions and disease conditions without motion modulation, while team 2 and team 3 achieved the smallest errors and 95% LOA when motion was included in the pathological simulations. For all teams, the absolute

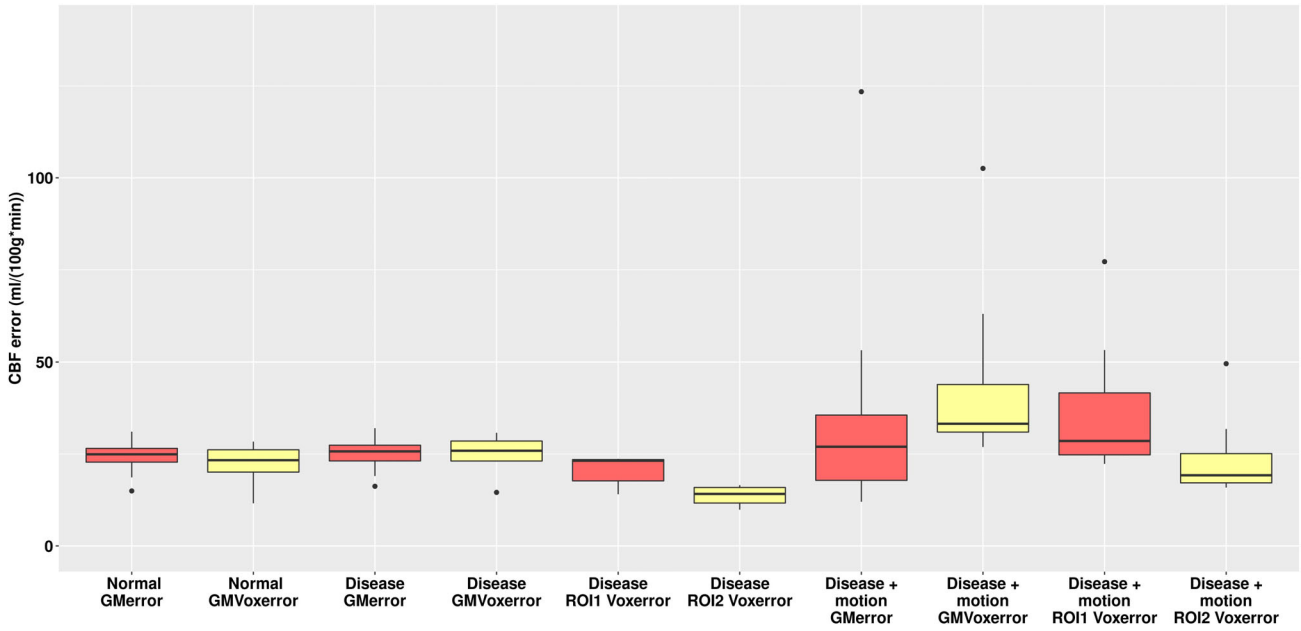

FIGURE 3 Boxplots of mean CBF absolute errors among the eight participating teams for the different conditions. Outliers are identified as black circles.

TABLE 2 ICC for all teams' submissions and simulated conditions compared to the ground truth values.

Team	DRO 1	DRO 2	DRO 3	DRO 4	DRO 5	DRO 6	DRO 7	DRO 8	DRO 9
1	0.68	0.67	0.73	0.66	0.68	0.75	0.06	0.04	0.05
2	0.85	0.83	0.86	0.83	0.83	0.85	**0.71**	**0.68**	**0.77**
3	0.85	0.82	0.86	0.82	0.82	0.85	**0.71**	**0.68**	**0.78**
4	0.76	0.73	0.78	0.72	0.72	0.76	0.65	0.59	0.71
5	0.77	0.75	0.48	0.73	0.74	0.76	0.07	0.13	0.10
6	0.80	0.77	0.81	0.76	0.77	0.78	0.61	0.57	0.67
7	**0.90**	**0.88**	**0.91**	**0.88**	**0.88**	**0.90**	0.16	0.47	0.43
8	0.01	0.25	0.10	0.00	0.09	0.08	0.01	0.00	0.01

Note: DRO 1–3 are simulated normal perfusion conditions. DRO 4–6 are simulated pathological conditions while DRO 7–9 are simulated pathological conditions including motion. The highest ICC for each DRO are shown in bold. ICC close to 1 indicates high agreement with ground truth.

CBF error was higher under simulated motion conditions compared to other conditions and similar for normal and disease conditions without motion (Table 1, Figure 3). The ICC analysis shown in Table 2 mirrored observations of the absolute errors and the LOA analysis, specifically team 7 had the highest ICC values for all conditions without motion, while teams 2 and 3 had highest but moderate ICC values when motion was included. Documentation quality across all teams was reasonably high (mean = 7.97, SD = 0.94). Scores varied most in the usability category (mean = 7.5, SD = 2.83) and least in Accuracy (mean = 10, SD = 0). Team 7 had the highest overall score. The difference in overall errors between the submitted and replicated results are summarized in Table 1 and in Figure S3. Four of the eight methods from team 2, team 3, team 4, and team 8 were replicated. The rest of the teams' results were not reproduced by following the documentation provided. In some teams, the code or script provided was unable to complete the analysis and produced output errors. Only team 2 and 8 had reproduced CBF maps and regional values similar to their submitted results.

3.3 | Variability of submitted results

Figure 4 shows the violin plots of voxelwise measurements of absolute GM and WM CBF errors between each team CBF map submission and the ground truth DRO for a representative data set of a healthy condition and

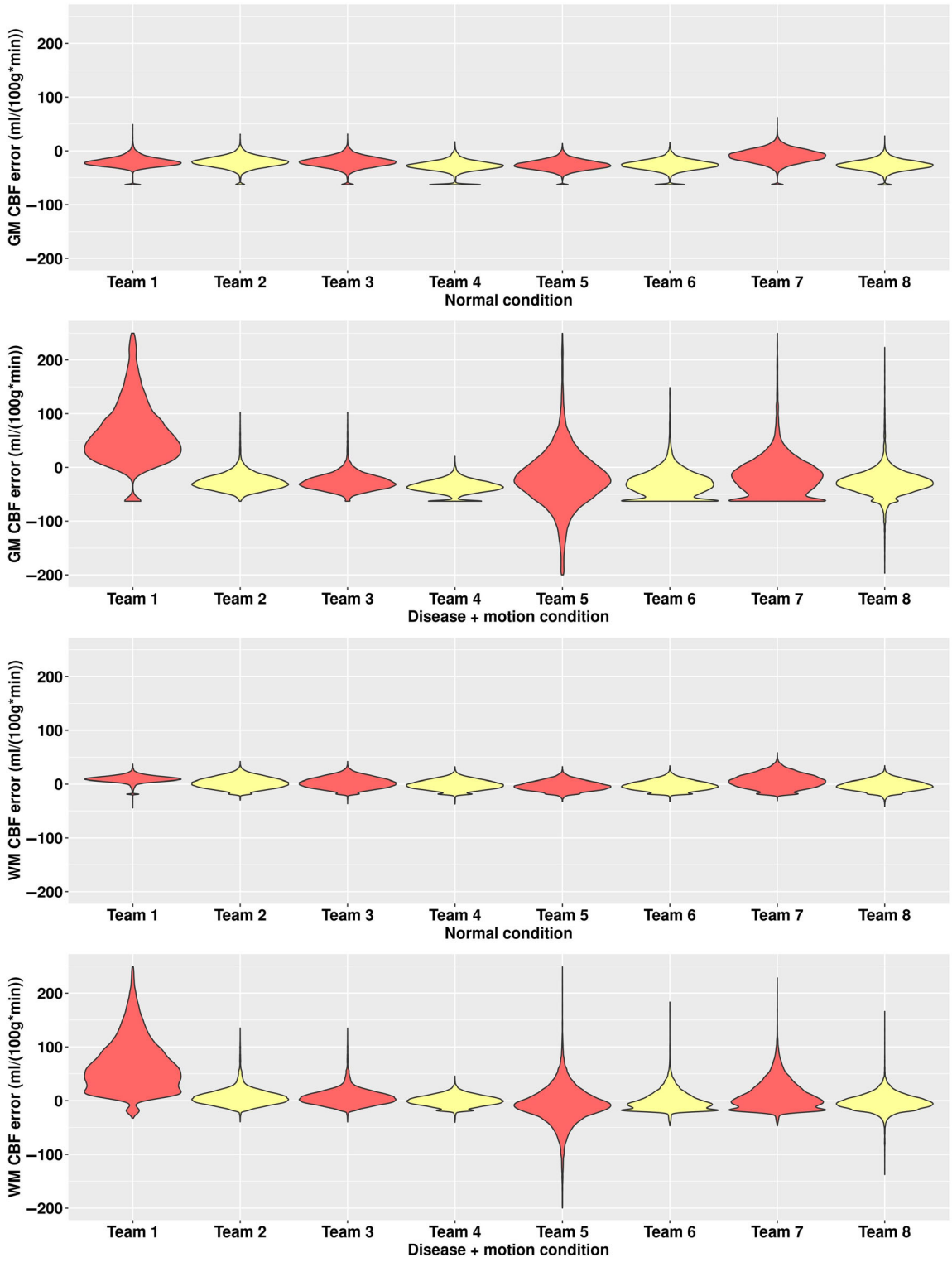

FIGURE 4 Absolute GM and WM CBF errors in mL/100 g/min between teams' submissions and the ground truth for a single normal condition and a single disease + motion condition. The Violin plots represent the range of errors within the regions of interest for each team and one illustrative data set for the normal condition and the disease + motion condition.

for a data set of disease + motion condition, illustrating errors in one of the simulated normal conditions (top two plots) and one DRO simulated for disease plus motion condition (bottom two plots). The mean covariance among the teams is shown in Figure 5 for all the voxels of the predefined ROI simulated to assess for disease effects with and without motion artifact. Although Figure S2 revealed some variability in the voxelwise CBF measures among

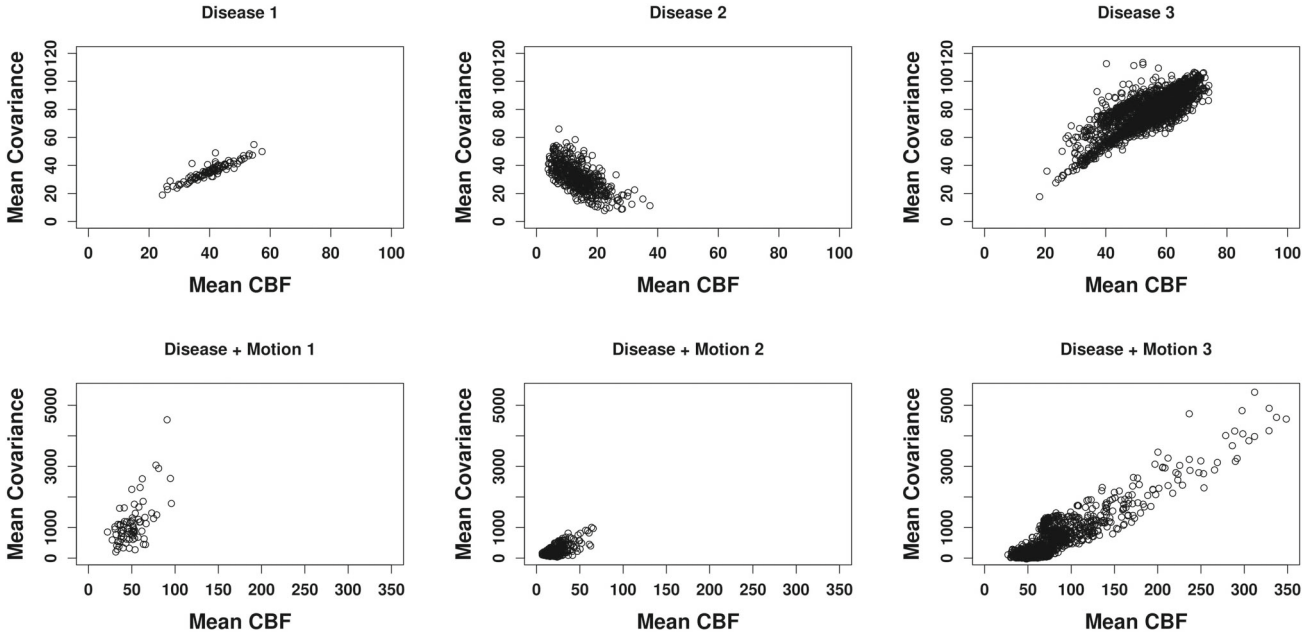

FIGURE 5 Mean covariance plotted against the mean CBF value among all the teams for all the voxels within a disease ROIs, with and without motion.

the teams, it is evident from Figure 4 that, for datasets without motion, all teams had a good performance with small absolute errors. On the other hand, for DRO 7 to 9, where motion was simulated, both the variability in quantitative CBF (Figure S2) and the absolute errors in CBF measures (Figure 4) were enhanced. In these scenarios, teams 2 and 3 reproduced the CBF measures of the ground truth data the best, followed by team 4, while the rest of the team's performance ranged from moderate to poor agreement. For the disease ROI, when there was no motion simulated, the mean covariance among the teams were much smaller compared to conditions where motion was included. Overall, the mean covariance was positively correlated to CBF, despite the second disease condition (no motion) that showed higher covariance for small CBF values.

3.4 | Partial volume correction

Six teams performed PVC and submitted corrected PV CBF maps. In general, the mean absolute CBF errors decreased when PVC was included (Figure 6A,B), and when the mean absolute errors were compared among the teams, the difference was statistically significant ($p < 0.05$) for seven of the nine DRO datasets for the GM ROI and for six of the nine datasets in the WM ROI. Figure 6A shows the boxplots of the absolute CBF errors after PVC between the maps provided by the six teams and the ground truth images for the nine DROs conditions. Comparing the conditions—normal and disease—, similar results between teams were observed with larger errors and presence of outliers for DROs with simulated motion (Figure 6A). Two teams provided PVC CBF maps containing very few non-zero voxels, resulting in increased errors despite their non-PVC GM and WM CBF maps appearing as expected. One team applied very high thresholded GM and WM masks to generate PVC CBF maps, which contributed to high absolute errors. Figure 6B illustrates the voxelwise difference between the absolute GM and WM errors with and without PVC from the difference between each team's submission and the ground truth, where top row plots show differences in one normal condition DRO and the bottom row plots in a representative DRO with motion. The positive values in the violin plots demonstrate the higher absolute CBF error when PVC was not performed compared to when PVC is included in the analysis.

3.5 | ASL challenge scores

The composite score for the accuracy, reproducibility, and documentation quality was tabulated using a ranking system based on the weighted criteria as described in the methods (Table 3). Given that PVC analysis was optional, the results of the analysis and corresponding CBF maps or values were not considered in the tabulation of the composite scores. For completeness, the PVC accuracy scores are also included in (Table 3 and Figure S4).

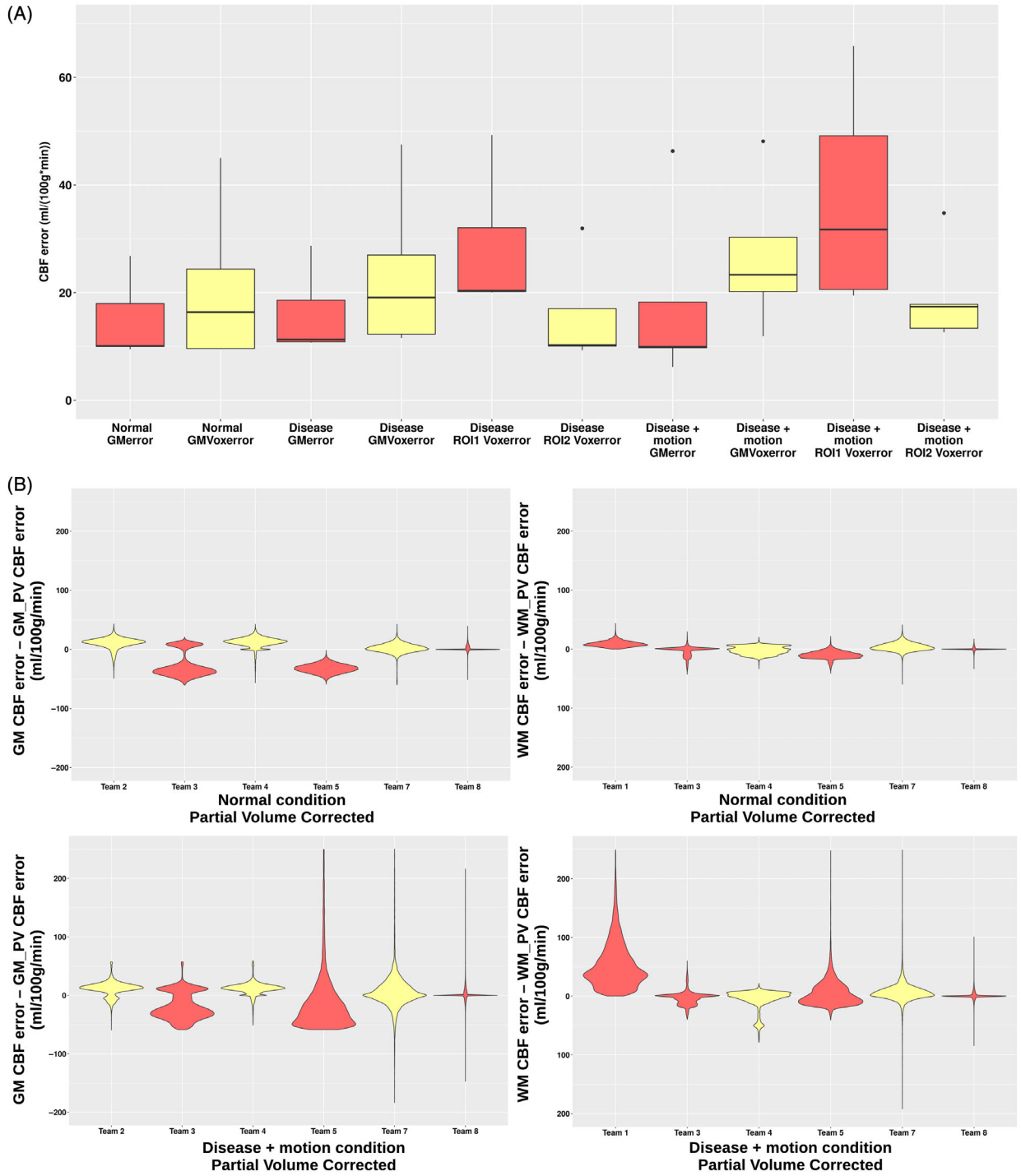

FIGURE 6 (A) Boxplots of absolute errors in partial volume corrected CBF estimates among the teams for the different conditions. Black dots shows the outliers. (B) CBF errors in mL/100 g/min partial volume corrected for GM and WM between teams' submissions and the ground truth for a normal condition and for a disease + motion condition.

3.6 | Benchmark for ASL processing pipelines

Based on the results obtained for accuracy and variability, and by analyzing the main processing steps performed by the top ranked teams, best practices for processing and analyzing single PLD pCASL data was provided, as outlined in Table 4. The Challenge DRO datasets and ground truth values accessible through Open Science Framework (OSF; https://t.co/gi1m80LItL), equally provide a benchmarking tool for ASL analysis pipelines.

TABLE 3 Summary of scores of the ISMRM OSIPI ASL MRI challenge.

Team	Accuracy (0–60)	Reproducibility (0–30)	Documentation quality (0–10)	Composite (0–100)	PVC accuracy (0 out of 100)
1	18.05	19.01	6.75	43.81	NA
2	47.49	**22.39**	8.25	78.13	**98.39**
3	41.54	22.38	8.75	72.67	32.01
4	26.66	21.17	6.5	54.33	80.64
5	22.19	16.56	**9.25**	48.00	16.42
6	24.44	16.72	8.25	49.41	NA
7	**57.52**	19.15	7.75	**84.42**	91.43
8	19.78	10.20	8.25	38.23	36.59

Note: The highest scores for each evaluation criteria are shown in bold. Partial volume correction (PVC) accuracy results were not included in overall score.

4 | DISCUSSION

In this study, we conducted the ASL data challenge to capture variability among different ASL analysis approaches. The analysis pipelines submitted covered the most widely used ASL preprocessing and analysis tools/software,[35] while it may be also explored for other important software available in the literature that did not participate in this challenge, as for example, the ASLprep.[36] The validation of submitted blood flow maps, along with qualitative analysis of the reproducibility of the pipelines, provides a comprehensive empirical assessment of ASL analysis steps. Based on the performance of the submissions in relation to ground-truth, we provide recommendations for ASL analysis for CBF quantification using single delay PCASL data.

4.1 | Variability in CBF results

Despite observed variability in the CBF measurements evident through the distribution pattern of the violin plots (Figures 5 and S2) and the relatively small deviation of the 95% LOA (Figure 2), the GM CBF were in general similar among teams (Figure S1) for all conditions, particularly for motion-free conditions. Individually, team 7 obtained the smallest errors, the smallest 95% LOA deviation, and highest ICC for normal and disease conditions without simulated motion. Visually, the contrast between GM and WM appears similar across teams except for team 1, whose submitted CBF maps showed poor GM/WM contrast, probably due to the larger post-processing smoothing kernel ($6 \times 6 \times 6$ mm^3 FWHM) applied only to M0, as documented.

To account for and assess head motion effects inherent in real-world ASL acquisition, we included three conditions with realistic motion artifacts. Overall, a significant increase in the absolute CBF errors was observed when motion was included. The agreement and variability in CBF measures were also increased for the cases with simulated motion. Team 7, which obtained the smallest errors and higher ICC for normal and disease conditions, showed higher absolute errors, and smaller ICC values when motion was included. Although the reason for this observation is not clear, one possible explanation could be the use of the mean ASL image as reference for motion correction, while other teams used the middle volume of the time-series as reference. In fact, teams 2 and 3, the next top performing teams for all conditions (based on the composite score), outperformed team 7 only for cases where motion was included, employing an approach where motion was corrected using the middle volume as reference. The signal intensity difference between the single middle volume and each image volume in the time-series is smaller than the intensity difference in the time-series and the average mean image. These larger intensity differences could bias estimation of extent of displacement when mean images are used as reference,[37] contributing to the subsequent bias in CBF modeling. It is also possible that the use of a spatial prior approach, performed by the BASIL algorithm during the data fitting process,[38] helped to improve the absolute errors and ICC with the motion DROs for teams 2 and 3. Although team 4 also used a pipeline based on the BASIL toolbox, their pipeline differed from teams 2 and 3 using the Advanced Normalization Tools (ANTs) to generate the whole-brain mask, which was on average 7.95% smaller than the brain masks created by the FSL-based approach. This led to larger errors at the edge of the brain for team 4 because these voxels were masked and so given a value of 0 mL/100 g/min. Similarly, team 1's whole-brain mask generated from summation of GM, WM, and CSF probability maps was also larger, while team 5 did not apply any brain mask in their analysis strategy. The effect of a large or

TABLE 4 Recommended ASL image processing and analysis steps.

Parameter	Steps	Minimal	Recommended	Ideal	Notes
1	Brain extraction/Mask	•		•	Segmentation of individual T1-weighted anatomical scans to obtain a whole-brain mask for extraction of the brain from the skull is recommended to restrict quantification and further analysis to brain tissue. The GM and WM tissue density information can also be obtained for registration and partial volume correction.
2	ASL time series motion compensation	•	•	•	Intra-volume motion correction is recommended as the first step in processing ASL data. The calibration (M0) scan or the middle ASL time point can be used as the reference scan with intra-modal linear (affine) registration using normalized correlation as the image similarity metric and a trilinear interpolation.
3	ASL pairwise subtraction	•	•	•	Simple subtraction of label from control images
4	Voxel-wise calibration	•	•		Recommended by ASL consensus paper to obtain a scaling factor on a voxel-by-voxel basis (2). Using the M0 as reference scan during motion compensation enables alignment of M0 to ASL for voxel-wise calibration.
5	Registration to T1-weighted image	•	•	•	Rigid registration of ASL difference images to the structural image is optimal using the gray matter probability maps rather than the T1-weighted image and a cost function and interpolation of mutual information and cubic b-spline, respectively. If T1-weighted image is used, consider correlation and boundary-based registration cost function and linear interpolation.
6	Quantification	•	•	•	Simple compartmental model and model parameters outlined in the ASL consensus paper (2). For 2D imaging, slice time correction should be taken into account during quantification as an addition to the PLD time.

Note: Recommendation is based on analysis of single post-label delay PCASL data acquired following the acquisition consensus guidelines [2] and preprocessing and analysis approach of the top ranked teams (teams 7 and 2).

no brain mask might have contributed to the significantly larger CBF errors and poor agreement among these teams as well as the higher CBF values at the edge of the brain, especially for the motion corrupted datasets in team 5. Across all teams, team 7 obtained apparent elevated global CBF values, visually observed in submitted CBF maps and in calculated global mean CBF values (Figure S2). Finally, unlike most teams, team 8 did not apply any threshold limit during CBF quantification to remove implausibly high (292 476 mL/100 g/min) and negative CBF estimates, which contributed to high global absolute errors and poor 95% LOA—one of the highest errors and worst agreements (including ICC) for almost all conditions.

Since the voxel dimensions in ASL perfusion imaging are typically relatively large (slice thickness ~ 5 mm), voxels located at the boundary between GM and WM will contain a mixture of both tissues, increasing the apparent CBF values in WM and decreasing it in GM. Although

the inclusion of PVC was not mandatory for participation in this challenge, five teams (teams 2, 3, 4, 7, and 8) did submit PVC CBF maps. While the PVC CBF maps of team 3 were erroneous (the maps mostly contained zeros), the PVC maps of the other five teams resulted in smaller mean absolute errors than when PVC was not applied (Figure 6A). Although the absolute errors are smaller on average, the violin plots of Figure 6B showed that for many voxels (particularly teams 3 and 5), the errors when PVC was employed are larger than when PVC was not performed. PVC was found to reduce CBF errors by an even greater amount in the datasets contaminated by motion, although the errors were still larger than the non-motion-corrupted datasets.

4.2 | Reproducibility assessment

To assess reproducibility, quantitative metrics (ICC and 95% LOA) were supplemented by qualitative scoring of the documentation provided along with the submissions. The overall good agreement (ICC >0.75) of CBF results with ground truth data is encouraging, especially under normal conditions where ICC values surpassed 0.8 for most teams. This finding is consistent with results from test–retest reliability studies in healthy volunteers scanned repeatedly over time.[39,40] In one such study, the reported within session (repeated scans in one imaging session) ICC values ranged from 0.44 to 0.94 across brain regions and between session (repeated scans over a month) ICC values were from 0.34 to 0.74 across the brain.[40] The CBF variability between scans was higher in midbrain regions such as the thalamus—a likely consequence of the area's susceptibility to arterial pulsatility which can produce apparent physiological motion artifacts. It can be expected that differences in CBF results from multiple analysis methods of a single dataset should not exceed the inherent test–retest reliability of ASL CBF measures.

A highly reproducible approach should yield consistent maps and results across repeated acquisitions or repeated analysis of a dataset. In this study, only half of the methods achieved successful result replication using provided analysis guides and scripts. Three of these teams had the highest reproducibility scores, partly due to comprehensive documentation following COBIDAS recommendations. We weighted the scores of a well-described methodology equal to agreement of CBF outputs to ensure that documentation best practices are captured. Comprehensive and clear step-by-step guides facilitate reproducibility, crucial for global ASL adoption. This was demonstrated by the replication exercise, where trainees with no prior ASL analysis experience replicated methods that had the highest documentation scores. While one of the team's methods was reproduced, the results differed substantially and CBF values were three times lower than the results submitted during the challenge. This was largely due to changes in software versions and updates made post-challenge to specifically address analysis of data with non-numeric values (i.e., not a number or NaN; c.f. Supplementary Information team 3). This underscores the importance of documenting software version information and changes as recommended in COBIDAS for data analysis best practices.[25] Hence, to advance ASL's clinical translation and use in multi-center trials, clear analysis documentation is essential.

4.3 | Data analysis reporting

While COBIDAS outlines documentation best practices, it does not provide a mechanism for evaluating the quality of completeness or clarity in documentations. Here, we adapted the QUACK measurement[27] system to further assess the quality of documentation. This assessment demonstrated that the teams were well practiced in delivering high clarity in the details they choose to include in the documentation. However, there was considerable variation in the level of detail provided which stalled reproduction attempts. For example, some teams reported that motion correction was executed without outlining the registration approach/cost function/interpolation used or in other cases failed to outline whether spatial smoothing was applied to the CBF maps. The quality and usability of the documentation submissions varied widely; from inclusion of step-by-step diagrams, to figures of pipeline processes. These visual aids improved clarity of the documentation and enhanced the overall reading experience. However, some of the submissions with highest scores across completeness and conciseness scored the lowest in quality and usability. The value of including images, therefore, should not be minimized. In addition to maximizing reproducibility, complete and well-described procedural documents can further increase adoption and wider dissemination of ASL methods.

Finally, we emphasize the value of reporting analysis methods and results following the COBIDAS guidelines and using a structured (potentially tabular[26]) framework to improve clarity and enable direct comparisons between pipelines. To standardize reporting of ASL analysis methods and foster high reproducible accuracy, the COBIDAS ASL analysis reporting guideline (Table S1) is recommended and could be included as Supplementary Material in ASL publications. The COBIDAS ASL analysis reporting guide introduced here is a starting framework and limited to the analysis of single post-label PCASL such as the data used in this work. We encourage the ASL community

to contribute to the further development of this standardized ASL analysis reporting guideline (c.f. The e-cobidas tool[41]).

4.4 | Limitations

As a limitation, in this work we focused on brain data analysis of single delay PCASL, given its clinical translation promise. Future ASL Challenge efforts should include other ASL acquisition schemes. Recent acquisition and analysis approaches including outlier rejection[42,43] and susceptibility-induced distortion correction schemes using phase-encoded reverse pairs (i.e., blip-up-blip-down) of ASL time series or calibration (M0) scans have been proposed to improve CBF quantification[28] and should be further validated in future ASL challenges. While real-world clinical data may contain more sources of errors than the DRO, using synthetic data enabled accurate assessment of analysis methodological approaches based on a ground truth. Since none of the teams provided a clear report of the spatial smoothing approach applied during ASL analysis or CBF quantification, an empirical recommendation for spatial smoothing was not provided in Table 4. In addition, the reproducibility assessment performed here compared absolute errors between teams analyzing a relatively small dataset, which would have benefited from a larger scaler evaluation of a library of available analysis pipelines.[35] Researchers are encouraged to validate their analysis pipelines using the DRO data and assessment criteria introduced here to improve reproducibility of results. Finally, the critical analysis of the differences in CBF results were limited by the brevity of the analysis method documentation, where detailed information such as smoothing kernel, registration strategy, among others, may have substantially differed and could have contributed to the results variability.[37,44] While analysis of the impact of potential differences in implementation of image processing software packages (SPM and FSL) is beyond the scope of this Challenge, it can be expected that similar moderate bias between packages observed in volumetry and functional MRI (fMRI) analysis,[45–47] could also contribute to the differences in CBF values reported in this Challenge. Nonetheless, variability in CBF estimates among analysis methods can be expected where there is considerable flexibility in parameter selection and analysis strategies, as demonstrated in other image analysis Challenge findings.[6,7,11,48] Foolproof analysis pipelines that minimize variability and maximize reproducibility can be achieved using well-documented self-contained programs that automatically process data following recommended steps and best practices.[25]

5 | CONCLUSIONS

This study is the first ASL MRI challenge, designed to provide an evaluation of ASL image processing and CBF quantification practices, toward a consensus driven recommendation for empirical analysis of ASL data. This challenge framework permitted an extensive summary of widely used ASL image processing tools including FSL schemes (Oxford_asl, BASIL, and Quantiphyse) and MATLAB/SPM-based approaches (ExploreASL, MRI cloud, LOFT ASL Toolbox and the Iris pipeline). The submitted CBF outputs outlined the variability in the CBF measurements even within pipelines based on the same tools, underscoring the impact of different analysis strategies in ASL-based CBF quantification. By using a DRO with ground truth information across conditions including simulation of subtle perfusion changes and effect of motion, we provide head-to-head performance evaluation of eight differing ASL analysis strategies, considering accuracy, reproducibility, and methodological quality. In general, the results of this challenge encourage standardization of ASL analysis pipelines, toward optimization of ASL for use in clinical environments and emphasize the importance of high-quality documentation to support reproducibility.

AFFILIATIONS

[1]Institute of Physics, University of Campinas, Campinas, Brazil

[2]LIM44, Institute of Radiology, Department of Radiology and Oncology of Clinical Hospital, University of Sao Paulo, Sao Paulo, Brazil

[3]Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK

[4]Department of Radiology, Center for Functional Magnetic Resonance Imaging, University of California, San Diego, La Jolla, California, USA

[5]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK

[6]Department of Radiology & Nuclear Medicine, Erasmus MC–University Medical Center Rotterdam, Rotterdam, the Netherlands

[7]Helmholtz-Zentrum Dresden-Rossendorf, Institute of Radiopharmaceutical Cancer Research, Dresden, Germany

[8]Radiological Sciences, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham, UK

[9]Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham, Nottingham, UK

[10]Nottingham Biomedical Research Centre, Queens Medical Centre, Nottingham, UK

[11]Clinical Imaging Group, Genentech, Inc., South San Francisco, California, USA

[12]Department of Psychiatry, University of Cambridge, Cambridge, UK

[13]Department of Psychiatry, University of Oxford, Oxford, UK

[14]Department of Radiology and Nuclear Medicine, Vrije Universiteit Amsterdam, Amsterdam UMC Location VUmc, Amsterdam, the Netherlands

[15]Amsterdam Neuroscience, Brain Imaging, Amsterdam, the Netherlands

[16]Gold Standard Phantoms Limited, London, UK

[17]Department of Radiology, Stanford University, Stanford, California, USA

[18]Computer Assisted Clinical Medicine, Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Heidelberg, Germany

[19]Mental Health & Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, UK

[20]Institute for Systems and Robotics-Lisboa and Department of Bioengineering, Instituto Superior Técnico–Universidade de Lisboa, Lisbon, Portugal

[21]Department of Cognitive Neuroscience, Radboud University Medical Center, Nijmegen, the Netherlands

[22]Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada

[23]Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, Québec, Canada

ACKNOWLEDGMENTS

The authors acknowledge Steven Soubron for his leadership in establishing OSIPI, Yuriko Suzuki for coordinating the MRM paper collection, Sudipto Dolui for providing a manuscription revision, the ISMRM Perfusion Study Group for supporting OSIPI, the challenge and endorsing the manuscript, as well as the ISMRM central office for logistical support with the challenge. The authors are grateful to Zee Wang for vital input on the analysis recommendation. Andre Paschoal is supported by the São Paulo Research Foundation (FAPESP) grant no. 2022/06496-7, and the Fundo De Apoio Ao Ensino, Pesquisa e Extensão (FAEPEX) grant 2589/23. J.G.W. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (220204/Z/20/Z) and Linacre College (University of Oxford). The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). H.M. is supported by the Dutch Heart Foundation (03-004-2020-T049), by the Eurostars-2 joint programme with co-funding from the European Union Horizon 2020 research and innovation programme (ASPIRE E!113701), provided by the Netherlands Enterprise Agency (RvO), and by the EU Joint Program for Neurodegenerative Disease Research, provided by the Netherlands Organization for health Research and Development and Alzheimer Nederland (DEBBIE JPND2020-568-106). J.P. is supported by the Dutch Heart Foundation (03-004-2020-T049), by the Eurostars-2 joint programme with co-funding from the European Union Horizon 2020 research and innovation programme (ASPIRE E!113701), provided by the Netherlands Enterprise Agency (RvO), and by the EU Joint Program for Neurodegenerative Disease Research, provided by the Netherlands Organization for Health Research and Development and Alzheimer Nederland (DEBBIE JPND2020-568-106). J.P. is supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/S021507/1. B.P. is supported by the EU Joint Program for Neurodegenerative Disease Research, provided by the Netherlands Organization for Health Research and Development and Alzheimer Nederland (DEBBIE JPND2020-568-106).

CONFLICT OF INTEREST STATEMENT

Laura Bell receives industry salary from her institution.

ORCID

Andre M. Paschoal https://orcid.org/0000-0001-8269-711X
Joseph G. Woods https://orcid.org/0000-0002-0329-824X
Jan Petr https://orcid.org/0000-0002-3201-6002
Laura Bell https://orcid.org/0000-0001-8164-8324
Henk J. M. M. Mutsaerts https://orcid.org/0000-0003-0894-0307
Moss Y. Zhao https://orcid.org/0000-0002-0210-7739
Irène Brumer https://orcid.org/0000-0002-1936-8687
Jian Hu https://orcid.org/0000-0003-0946-9617
Logan X. Zhang https://orcid.org/0000-0003-1484-3992
Sara P. Monteiro https://orcid.org/0000-0002-2530-5535

TWITTER

Andre M. Paschoal paschoal_am

REFERENCES

1. Detre JA, Leigh JS, Williams DS, Koretsky AP. Perfusion imaging. *Magn Reson med*. 1992;23:37-45.
2. Alsop DC, Detre JA, Golay X, et al. Recommended implementation of arterial spin-labeled perfusion mri for clinical applications: a consensus of the ISMRM perfusion study group and the European consortium for ASL in dementia. *Magn Reson med*. 2015;73:102-116. doi:10.1002/mrm.25197
3. Pujol S, Wells W, Pierpaoli C, et al. The DTI challenge: toward standardized evaluation of diffusion tensor imaging Tractography for neurosurgery. *J Neuroimaging*. 2015;25:875-882. doi:10.1111/jon.12283
4. Grissom WA, Setsompop K, Hurley SA, Tsao J, Velikina JV, Samsonov AA. Advancing RF pulse design using an open-competition format: report from the 2015 ISMRM challenge. *Magn Reson med*. 2017;78:1352-1361. doi:10.1002/mrm.26512
5. Nath V, Schilling KG, Parvathaneni P, et al. Tractography reproducibility challenge with empirical data (TraCED): the 2017 ISMRM diffusion study group challenge. *J Magn Reson Imaging*. 2020;51:234-249. doi:10.1002/jmri.26794

6. Schilling KG, Daducci A, Maier-Hein K, et al. Challenges in diffusion MRI tractography–lessons learned from international benchmark competitions. *Magn Reson Imaging*. 2019;57:194-209. doi:10.1016/j.mri.2018.11.014
7. Veronese M, Rizzo G, Belzunce M, et al. Reproducibility of findings in modern PET neuroimaging: insight from the NRM2018 grand challenge. *J Cereb Blood Flow Metab*. 2021;41:2778-2796. doi:10.1177/0271678X211015101
8. Maffei C, Girard G, Schilling KG, et al. Insights from the IronTract challenge: optimal methods for mapping brain pathways from multi-shell diffusion MRI. *Neuroimage*. 2022;257:119327. doi:10.1016/j.neuroimage.2022.119327
9. Bossier H, Roels SP, Seurinck R, et al. The empirical replicability of task-based fMRI as a function of sample size. *Neuroimage*. 2020;212:116601. doi:10.1016/j.neuroimage.2020.116601
10. Fanelli D. Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci*. 2018;115:2628-2631. doi:10.1073/pnas.1708272114
11. Botvinik-Nezer R, Holzmeister F, Camerer CF, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020;582:84-88. doi:10.1038/s41586-020-2314-9
12. Anazodo U, Pinto J, Mcconnell FAK, et al. The Open Source Initiative for Perfusion Imaging (OSIPI) ASL MRI Challenge. In *Proceedings of the 29th Annual Meeting of the International Society of Magnetic Resonance in Medicine. Vol c*. Virtual Meeting; 2021:1-3. Abstract 2714.
13. Clement P, Petr J, Dijsselhof MBJ, et al. A Beginner's guide to arterial spin labeling (ASL) image processing. *Front Radiol*. 2022;2:929533. doi:10.3389/fradi.2022.929533
14. Dai W, Garcia D, de Bazelaire C, Alsop DC. Continuous flow-driven inversion for arterial spin labeling using pulsed radio frequency and gradient fields. *Magn Reson med*. 2008;60:1488-1497. doi:10.1002/mrm.21790
15. Clement P, Castellaro M, Okell TW, et al. ASL-BIDS, the brain imaging data structure extension for arterial spin labeling. *Sci Data*. 2022;9:543. doi:10.1038/s41597-022-01615-9
16. Anazodo U, Croal P, Paschoal AM. OSIPI ASL MRI Challenge. 2021. doi:10.17605/OSF.IO/6XYU3
17. Lorenzini L, Ingala S, Wink AM, et al. The open-access European prevention of Alzheimer's dementia (EPAD) MRI dataset and processing workflow. *Neuroimage Clin*. 2022;35:103106. doi:10.1016/j.nicl.2022.103106
18. Oliver-Taylor AM, Hampshire T, Stritt M, et al. ASLDRO: digital reference object software for arterial spin labelling. In *Proceedings of the 29th Annual Meeting of the International Society of Magnetic Resonance in Medicine*. Vol 2731. 2021. Virtual https://pypi.org/project/asldro/ Abstract 2731.
19. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn human connectome project: an overview. *Neuroimage*. 2013;80:62-79. doi:10.1016/j.neuroimage.2013.05.041
20. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag; 2016 https://ggplot2.tidyverse.org
21. Schwalbe M, ed. *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results*. National Academies Press; 2016.
22. Gerke O. Reporting standards for a bland-altman agreement analysis: a review of methodological reviews. *Diagnostics*. 2020;10:1-17. doi:10.3390/diagnostics10050334
23. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr med*. 2016;15:155-163. doi:10.1016/j.jcm.2016.02.012
24. R Core Team. R: A Language and Environment for Statistical Computing. Foundation for Statistical Computing https://www.r-project.org/; https://www.r-project.org/ 2020.
25. Nichols TE, das S, Eickhoff SB, et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci*. 2017;20:299-303. doi:10.1038/nn.4500
26. Praag CGV. COBIDAS reporting breakdown.
27. O'Keefe S. Calculating document quality (QUACK). Scriptorium. Accessed November 9, 2023. https://www.scriptorium.com/2010/05/calculating-document-quality-quack/. Published May 14, 2010
28. Buxton RB, Frank LR, Wong EC, Siewert B, Warach S, Edelman RR. A general kinetic model for quantitative perfusion imaging with arterial spin labeling. *Magn Reson med*. 1998;40:383-396.
29. ASL Analysis Tab—Quantiphyse documentation. Accessed November 9, 2023. https://quantiphyse.readthedocs.io/en/latest/asl/asl_analysis.html
30. Chappell MA, Groves AR, Whitcher B, Woolrich MW. Variational Bayesian inference for a nonlinear forward model. *IEEE Trans Signal Process*. 2009;57:223-236. doi:10.1109/TSP.2008.2005752
31. Mutsaerts HJMM, Petr J, Groot P, et al. ExploreASL: an image processing pipeline for multi-center ASL perfusion MRI studies. *Neuroimage*. 2020;219:117031. doi:10.1016/j.neuroimage.2020.117031
32. MRIcloud. Accessed November 9, 2023. https://braingps.mricloud.org/docs/Manual_ASL_processing.v2.pdf.
33. Bron EE, Steketee RME, Houston GC, et al. Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. *Hum Brain Mapp*. 2014;35:4916-4931. doi:10.1002/hbm.22522
34. Zhao C. *LOFT_ASL_toolbox*. GitHub. 2022. Accessed November 9, 2023. https://github.com/chenyang9526/LOFT_ASL_toolbox
35. Fan H, Mutsaerts HJMM, Anazodo U, et al. ISMRM Open Science initiative for perfusion imaging (OSIPI): ASL pipeline inventory. *Magn Reson Med*. 2023;1-16. doi:10.1002/mrm.29869
36. Adebimpe A, Bertolero M, Dolui S, et al. ASLPrep: a platform for processing of arterial spin labeled MRI and quantification of regional brain perfusion. *Nat Methods*. 2022;19:683-686. doi:10.1038/s41592-022-01458-7
37. Wang Z, Aguirre GK, Rao H, et al. Empirical optimization of ASL data analysis using an ASL data processing toolbox: ASLtbx. *Magn Reson Imaging*. 2008;26:261-269. doi:10.1016/j.mri.2007.07.003
38. Groves AR, Chappell MA, Woolrich MW. Combined spatial and non-spatial prior for inference on MRI time-series. *Neuroimage*. 2009;45:795-809. doi:10.1016/j.neuroimage.2008.12.027
39. Almeida JRC, Greenberg T, Lu H, et al. Test-retest reliability of cerebral blood flow in healthy individuals using arterial spin labeling: findings from the EMBARC study. *Magn Reson Imaging*. 2018;45:26-33. doi:10.1016/j.mri.2017.09.004
40. Ssali T, Anazodo UC, Bureau Y, MacIntosh BJ, Günther M, St Lawrence K. Mapping long-term functional changes in cerebral blood flow by arterial spin labeling. *PLoS ONE*. 2016;11:e0164112. doi:10.1371/journal.pone.0164112

41. COBIDAS checklist. July 2019. 10.17605/OSF.IO/ANVQY.
42. Dolui S, Tisdall D, Vidorreta M, et al. Characterizing a perfusion-based periventricular small vessel region of interest. *Neuroimage Clin*. 2019;23:101897. doi:10.1016/j.nicl.2019.101897
43. Shirzadi Z, Stefanovic B, Chappell MA, et al. Enhancement of automated blood flow estimates (ENABLE) from arterial spin-labeled MRI. *J Magn Reson Imaging*. 2018;47:647-655. doi:10.1002/jmri.25807
44. Mutsaerts HJMM, Petr J, Thomas DL, et al. Comparison of arterial spin labeling registration strategies in the multi-center GENetic frontotemporal dementia initiative (GENFI). *J Magn Reson Imaging*. 2018;47:131-140. doi:10.1002/jmri.25751
45. Pauli R, Bowring A, Reynolds R, Chen G, Nichols TE, Maumet C. Exploring fMRI results space: 31 variants of an fMRI analysis in AFNI, FSL, and SPM. *Front Neuroinformatics*. 2016;10:10. doi:10.3389/fninf.2016.00024
46. Kazemi K, Noorizadeh N. Quantitative comparison of SPM, FSL, and Brainsuite for brain MR image segmentation. *J Biomed Phys Eng*. 2014;4:13-26.
47. Seiger R, Ganger S, Kranz GS, Hahn A, Lanzenberger R. Cortical thickness estimations of FreeSurfer and the CAT12 toolbox in patients with Alzheimer's disease and healthy controls. *J Neuroimaging*. 2018;28:515-523. doi:10.1111/jon.12521
48. Marjańska M, Deelchand DK, Kreis R. The 2016 ISMRM MRS study group fitting challenge team. Results and interpretation of a fitting challenge for MR spectroscopy set up by the MRS study group of ISMRM. *Magn Reson med*. 2022;87:11-32. doi:10.1002/mrm.28942

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

Data S1. Supporting Information.

How to cite this article: Paschoal AM, Woods JG, Pinto J, et al. Reproducibility of arterial spin labeling cerebral blood flow image processing: A report of the ISMRM open science initiative for perfusion imaging (OSIPI) and the ASL MRI challenge. *Magn Reson Med*. 2024;92:836-852. doi: 10.1002/mrm.30081