Received: 14 April 2023 | Revised: 22 August 2023 | Accepted: 16 October 2023

DOI: 10.1002/mrm.29909

Magnetic Resonance in Medicine

RESEARCH ARTICLE

OSIPI Special Section

# The ISMRM Open Science Initiative for Perfusion Imaging (OSIPI): Results from the OSIPI–Dynamic Contrast-Enhanced challenge

**Eve S. Shalom[1,2] | Harrison Kim[3] | Rianne A. van der Heijden[4,5] | Zaki Ahmed[6] | Reyna Patel[7] | David A. Hormuth II[8] | Julie C. DiCarlo[9] | Thomas E. Yankeelov[10,11] | Nicholas J. Sisco[12] | Richard D. Dortch[12] | Ashley M. Stokes[12] | Marianna Inglese[13,14] | Matthew Grech-Sollars[14,15,16] | Nicola Toschi[13,17] | Prativa Sahoo[18] | Anup Singh[19] | Sanjay K. Verma[20] | Divya K. Rathore[21] | Anum S. Kazerouni[22] | Savannah C. Partridge[22] | Eve LoCastro[23] | Ramesh Paudyal[23] | Ivan A. Wolansky[23] | Amita Shukla-Dave[23,24] | Pepijn Schouten[25] | Oliver J. Gurney-Champion[25,26] | Radovan Jiřík[27] | Ondřej Macíček[27] | Michal Bartoš[28] | Jiří Vitouš[27] | Ayesha Bharadwaj Das[29] | S. Gene Kim[29] | Louisa Bokacheva[30] | Artem Mikheev[30] | Henry Rusinek[30] | Michael Berks[31] | Penny L. Hubbard Cristinacce[31] | Ross A. Little[31] | Susan Cheung[31] | James P. B. O'Connor[31,32,33] | Geoff J. M. Parker[34,35] | Brendan Moloney[36] | Peter S. LaViolette[37] | Samuel Bobholz[37] | Savannah Duenweg[37] | John Virostko[38] | Hendrik O. Laue[39] | Kyunghyun Sung[40] | Ali Nabavizadeh[41,42] | Hamidreza Saligheh Rad[43,44] | Leland S. Hu[45] | Steven Sourbron[2] | Laura C. Bell[46] | Anahita Fathi Kazerooni[43,47]**

**Correspondence**
Anahita Fathi Kazerooni, CHOP Roberts Center for Pediatric Research, 734 Schuylkill Ave, Floor 19, Office 19282, Philadelphia, PA 19146, USA.
Email: fathikazea@chop.edu

**Funding information**
EPSRC-CASE, Grant/Award Number: 2282622; Bracco Diagnostics

## Abstract

**Purpose:** $K^{trans}$ has often been proposed as a quantitative imaging biomarker for diagnosis, prognosis, and treatment response assessment for various tumors. None of the many software tools for $K^{trans}$ quantification are standardized. The ISMRM Open Science Initiative for Perfusion Imaging–Dynamic Contrast-Enhanced (OSIPI-DCE) challenge was designed to benchmark methods to better help the efforts to standardize $K^{trans}$ measurement.

**Methods:** A framework was created to evaluate $K^{trans}$ values produced by DCE-MRI analysis pipelines to enable benchmarking. The perfusion MRI community was invited to apply their pipelines for $K^{trans}$ quantification in glioblastoma from clinical and synthetic patients. Submissions were required to include the entrants' $K^{trans}$ values, the applied software, and a standard operating procedure. These were

For affiliations refer to page 1817

evaluated using the proposed $OSIPI_{gold}$ score defined with accuracy, repeatability, and reproducibility components.

**Results:** Across the 10 received submissions, the $OSIPI_{gold}$ score ranged from 28% to 78% with a 59% median. The accuracy, repeatability, and reproducibility scores ranged from 0.54 to 0.92, 0.64 to 0.86, and 0.65 to 1.00, respectively (0–1 = lowest–highest). Manual arterial input function selection markedly affected the reproducibility and showed greater variability in $K^{trans}$ analysis than automated methods. Furthermore, provision of a detailed standard operating procedure was critical for higher reproducibility.

**Conclusions:** This study reports results from the OSIPI-DCE challenge and highlights the high inter-software variability within $K^{trans}$ estimation, providing a framework for ongoing benchmarking against the scores presented. Through this challenge, the participating teams were ranked based on the performance of their software tools in the particular setting of this challenge. In a real-world clinical setting, many of these tools may perform differently with different benchmarking methodology.

**KEYWORDS**

challenge, data analysis, DCE-MRI, glioblastoma, open-science, perfusion

## 1 | INTRODUCTION

Dynamic contrast-enhanced (DCE) MRI provides physiological parameters associated with the exchange of a contrast agent between intravascular and extravascular spaces.[1] In patients with glioblastoma, the volume transfer constant ($K^{trans}$) has been proposed as a marker for characterizing tumor pathophysiology, which can aid in grading,[2] assessment of treatment response,[3] and differentiation of recurrence from radiation necrosis.[4]

Quantitative DCE-MRI through pharmacokinetic (PK) modeling is intended to yield reproducible parameters across different studies and institutions.[5] However, the variation in the arterial input function (AIF), chosen PK models, and stability in model fitting adversely affect the quantification of $K^{trans}$ values.[6] Therefore, the reported $K^{trans}$ values differ among studies, making it currently unsuitable as a marker in multi-institutional clinical trials. Furthermore, a small number of studies have measured repeatability. Based on this limited literature, current Quantitative Imaging Biomarkers Alliance (QIBA) guidelines state that a change of $K^{trans}$ above 21.3% may indicate true $K^{trans}$ change with 95% confidence in gliblastoma.[7,8] Therefore, methods with a repeatability coefficient (%RC)[8] above this threshold cannot reliably detect tumor progression in longitudinal studies, further contributing to the limitations of quantitative DCE-MRI in clinical trials.

Over the past decade, more attention has been brought to the replication of research studies, the so-called reproducible research.[9,10] While researchers make their best effort to report accurate data, the choices they have to make about different aspects of data collection and analysis methods could influence the outcome of their significance tests and, therefore, the derived conclusions.[11,12] This "researcher degrees of freedom" issue imposes challenges for the reproducibility of the results when reanalyzing the same data, or generalizability of the findings to independent data.[11,12]

For quantification of $K^{trans}$ from DCE-MRI, there is an extensive list of available tools[13] from which to choose, although no "gold standard" analysis technique exists for clinical data. Evaluation and validation of these tools in the reported literature are based on data sets collected by authors, rendering it difficult to perform a fair comparison between them.[14] When provided with a wide range of possible (well-grounded) options for analysis methods, researchers may select or report methods that yield more favorable results for their data.[11,12] Despite researchers' best intentions, the inclination to show statistically significant results, referred to as "selective analysis reporting," could accompany false positive errors.[14] To avoid such errors, which hinder reproducibility, it is critical to provide comprehensive and open/transparent details about the study design and analysis approaches.

The ISMRM Open-Science Initiative for Perfusion Imaging (OSIPI), an ISMRM perfusion study group initiative, was founded to promote reproducible research and open science in perfusion imaging and to facilitate the translation of software tools into clinical practice.

The OSIPI task force on DCE and dynamic susceptibility contrast challenges (Task Force 6.2) was formed in February 2020 with a group of medical physicists, radiologists, and biomedical engineers to address the current issues of benchmarking perfusion quantification methods by organizing community challenges.

The OSIPI-DCE, as an ISMRM challenge, was the first of such challenges. OSIPI-DCE aims to design and build a systematic and controlled framework to benchmark the quantification of $K^{trans}$ as a diagnostic or prognostic biomarker in brain tumors, and to apply this framework to submissions from the community. This setup allowed the evaluation and validation of software packages in a single setting with synthetic and real-world clinical data. For the first time in a challenge setting, the accuracy, repeatability, and reproducibility of various methods were assessed for $K^{trans}$ quantification in glioblastomas. This article describes the challenge data, design, results of evaluating different analysis methods, and obstacles in the assessment of reproducibility.

## 2 | METHODS

The OSIPI-DCE challenge was launched at the ISMRM Annual Meeting on May 15, 2021, upon presentation of the abstract[15] on the outline of the challenge during this annual meeting.

### 2.1 | Challenge setup

This challenge aimed to assess the results and analysis methods submitted by the participating teams according to the $OSIPI_{gold}$ score (Table 1) for their (1) accuracy in the quantification of $K^{trans}$ using a set of synthetic data designed for this challenge, (2) repeatability using open-access test–retest scans of 8 patients with glioblastoma,[16,17] and (3) reproducibility based on an independent re-analysis of the data by a neutral evaluator team.

The challenge design was submitted as an abstract[15] for peer review at the 2020 ISMRM Annual Meeting. The researchers in the perfusion MRI community were invited by email to participate through the ISMRM Perfusion Study Group, LinkedIn, Twitter, or via direct contact. Interested teams registered on the ISMRM Challenges website and received submission guidelines via an automated email. The participants were asked to submit their results along with a report about the analysis approach, as described below:

- Matrices of voxelwise $K^{trans}$ maps (in the original space) for all slices in the synthetic and clinical DCE-MRI (two visits per subject) in NIfTI format.
- Standard operating procedures (SOPs) with sufficient detail to allow a neutral evaluator team to reproduce the results without interaction with the challenge participants. The SOPs should explain software access and installation and provide a step-by-step guide to reproduce the analysis. It is essential that the synthetic and patient data are analyzed with the same approach. Copies of each submission SOP are contained in Supporting Information Data S7.

No requirement was placed on the challenge participants to release their source codes or to base their submissions on open-source or open-access software. However, for commercial or in-house software that was not freely available, the participants were asked to provide a trial license or an executable file for the independent replication of the results. The license could be temporary, allowing sufficient time for re-analysis. Instructions on how to obtain the license should have been included in the SOPs without requiring interactions between submitters and neutral evaluator teams.

The challenge was open for submissions through the end of the year 2021. The task force reached out to experienced DCE scientists to help in evaluating the submissions in terms of procedural reproducibility and the reported $K^{trans}$ maps, after the challenge was closed. The evaluators had either more than 1 (I.W.), 5 (Z.A., S.B.), or 10 (P.S.L., J.V., H.O.L., K.S.) years of experience in DCE analysis. At the end of the challenge, the SOPs and software tools for each submission were provided to the independent evaluators.

### 2.2 | Data description

Two sets of data were provided in our challenge repository.[18]

#### 2.2.1 | Clinical data

A set of repeat DCE-MRI and $T_1$-mapping scans with accompanying $T_1$ contrast-enhanced (CE) FLASH and $T_2$ CE fluid-attenuated inversion recovery from 8 patients with glioblastoma, selected from the RIDER Neuro MRI database[16,17] and renamed, were acquired on a 1.5T Siemens scanner at two scan dates, typically 1–2 days apart. Sequence details are provided in Supporting Information Data S1.

#### 2.2.2 | Synthetic data

Two synthetic DCE-MRI patient data sets were generated from RIDER subjects[16,17] to be analyzed with the same

**TABLE 1** Summary of Open Science Initiative for Perfusion Imaging (OSIPI) scoring metric definitions. Here, $OSIPI_{gold}$ defines the full proposed metric, whereas $OSIPI_{silver}$ defines the score applied in cases in which reproduction of submissions was not possible.

| | |
|---|---|
| **Global OSIPI scoring metric** | |
| $OSIPI_{gold} = 100 \cdot Score_{accuracy} \cdot Score_{repeat} \cdot Score_{reproduce}$ $OSIPI_{silver} = 100 \cdot Score_{accuracy} \cdot Score_{repeat}$ | |
| **Component scoring metrics** | |
| $Score_{accuracy} = exp\left( -\frac{1}{4}\sum_{i=1}^{4} \frac{\sigma\left(K_i^{trans}, K_{i,exact}^{trans}\right)}{\mu\left(K_i^{trans}, K_{i,exact}^{trans}\right)} \right)$ | Accuracy score: a measure for the accuracy of the $K^{trans}$ values of submissions ($K_i^{trans}$) by comparison with exact values ($K_{exact}^{trans}$) in the synthetic data. The bar here denotes a mean $K^{trans}$ value over the tumor mask in each scan, with averaging over the four synthetic data sets ($\sigma$ and $\mu$ represent sample SD and mean, respectively, between the two scans). |
| $Score_{repeat} = exp\left( -\frac{1}{8}\sum_{i=1}^{8} \frac{\sigma\left(K_{i,v1}^{trans}, K_{i,v2}^{trans}\right)}{\mu\left(K_{i,v1}^{trans}, K_{i,v2}^{trans}\right)} \right)$ | Repeatability score: a measure for the repeatability of the $K^{trans}$ values of submissions. This score compares the submitted $K^{trans}$ values for the test ($K_{i,v1}^{trans}$) and retest ($K_{i,v2}^{trans}$). The bar here denotes a mean $K^{trans}$ value over the tumor mask in each scan, with averaging over the eight clinical patient data sets ($\sigma$ and $\mu$ represent sample SD and mean, respectively, between the two scans). |
| $Score_{reproduce} = exp\left( -\frac{1}{20}\sum_{i=1}^{20} \frac{\sigma\left(K_i^{trans}, K_{i,neutral}^{trans}\right)}{\mu\left(K_i^{trans}, K_{i,neutral}^{trans}\right)} \right)$ | Reproducibility score: a measure that quantifies to what extent the submitted $K^{trans}$ values are independently reproducible. The metric compares the submitted $K^{trans}$ values ($K_i^{trans}$) calculated for the two visits of each of the 10 cases (i.e., two synthetic and eight patient data) against the same values reproduced independently ($K_{i,\ neutral}^{trans}$) by experienced neutral evaluators, based on the standard operating procedures (SOPs) provided. The bar here denotes a mean $K^{trans}$ value over the tumor mask in each scan with averaging over the 20 visit data sets ($\sigma$ and $\mu$ represent sample SD and mean, respectively, between the two scans). |

processing pipeline as the clinical data. For this reason, the synthetic DCE-MRI data were integrated into an original DICOM study, also including the anatomical reference data from the same RIDER subjects.

Synthetic data were created following two steps: First, an inverse model was applied to the DCE-MRI and variable flip angle (VFA) data set to obtain parameter maps for precontrast relaxation rate ($R_{10}$), rate constants ($K^{trans}$, $k_{ep}$), capillary plasma volume per unit volume of tissue $(v_p)$, and an AIF; subsequently, these parameter maps had thresholds and filters applied to produce an unknown ground truth. The forward model was applied to produce synthetic DCE-MRI and VFA signal intensity curves. All details of the inverse modeling remained undisclosed during the submission period. The challenge guideline detailed the PK model used for the forward modeling as well as the assumed concentration and relaxation rate relationships. In addition, it defined the creation of VFA data using $R_{10}$ maps and a constant $R_{20}^*$ throughout.

For the inverse approach, initial parameter values were recovered from the RIDER data using matrix form.[19] A partial volume correction was applied using the sagittal sinus signal. Thresholds were then applied to the output to discard negative values produced during the least-squares fitting process and limit maximal volume fraction values to 1. Smoothing of the fitted values was carried out using a $3 \times 3$ median filter. The AIF signal (corrected with hematocrit 0.45)[20] was selected from the middle cerebral arteries and scaled to have a realistic peak value of 6 mM.[21]

For the forward model, the extended Tofts model was applied with the AIF and parameter maps ($K^{trans}$, $k_{ep}$, $v_p$). The resulting tissue concentration–time curves were converted into $R_1(t)(= 1/T_1(t))$ and $R_2^*(t)\left(= 1/T_2^*(t)\right)$, assuming a linear relationship between concentration and the relaxation rates according to the $r_1$ and $r_2^*$ relaxivities of gadolinium–diethylenetriaminepentaacetic acid, respectively (3.9 and 10 Hz/mM),[22,23] and by making use of the $R_{10}$ and $R_{20}^*$ maps. To deduce the precontrast relaxation rates, the signal evolution was modeled as a spoiled gradient-echo sequence in steady state. This was applied to express the signal everywhere at the initial time and the initial sagittal sinus signal. These relations were combined in order to give a calculated $R_{10}$ map using the reference $T_1$ of 1.48 s in the sagittal sinus at 1.5 T.[24] The constant precontrast $R_{20}^*$ applied was 17.24 Hz.[23] Scan-specific constants match the original scan values with flip angle = 25°, TR = 3.8 ms, TE = 1.8 ms, and using a $1 \times 1 \times 5\ mm^3$ voxel size.

Subsequently, these relaxation rates were converted into DCE-MRI signal time curves, again modeling signal evolution as a spoiled gradient-echo sequence. A multiplicative constant was defined to give the synthetic data similar maximal signal values to the original RIDER data

set. This produced synthetic data with a 16-slice volume captured with a temporal resolution of 4.8 s.

Finally, Rician noise was added to the signal–time data by assessing the SD across the precontrast time steps within each voxel from the original RIDER data set. Voxelwise noise values were then applied from randomly sampled Gaussian distributions with the voxelwise SD; the absolute resulting signal was taken. Across all voxels, the signal noise applied had a mean and SD ($\mu \pm \sigma$) of $5.65 \pm 3.21$ and $5.51 \pm 3.14$ for synthetic Patients 1 and 2, respectively. The VFA data were recreated using the same signal model and $R_{10}$ maps with flip angles of 5°, 10°, 15°, 20°, 25°, and 30°.

Synthetic signal intensity–time data were exported in DICOM file format, and the original DICOM DCE-MRI data were replaced by the synthetic data. The patient identifiers were overwritten to avoid confusion with the original RIDER data from which the synthetic data were derived. The parameter maps ($K^{trans}$, $k_{ep}$, $v_p$) were then changed to create the second visit data through the same process with identical AIF. No guarantee was offered that the second visit data were identical to the first, and some substantial differences were deliberately introduced. The differences between visits helped to assess the accuracy while also penalizing methods that overconstrained visits to have repeatable values.

#### 2.2.3 | Tumor segmentation

Segmentation of brain tumors for each visit in the clinical and synthetic data was performed on the last time series of DCE-MRI scans by comparing to the anatomical postcontrast $T_1$ and fluid-attenuated inversion-recovery images to delineate the enhancing tumor region. The regions of interest (ROIs) were not released to the challengers. The segmentations were carried out using the *ITK-SNAP* software[25] by an experienced neuroimaging researcher (R.P.) under the supervision of a senior neuroradiologist (L.H.). These mask data were output in NIfTI format to be overlaid on the submitted $K^{trans}$ NIfTI matrices.

### 2.3 | Leaderboard evaluation

The entry submissions were evaluated using the OSIPI scoring metrics as defined in Table 1.

#### 2.3.1 | Segmentation overlay

The segmentation masks were overlaid in *Python* using the Nibabel library[26] onto the ground-truth data and submitted $K^{trans}$ maps for all data sets. The extracted arrays from all submissions were visualized within *Python* to ensure correct alignment with the segmentation masks. It was found that the submitted NIfTI files had varying alignment quality due to the nature of the analysis techniques in stripping array data from DICOM files. Any submissions with alignment issues were transformed without interpolation using NumPy 90° rotations or axis reflection, to ensure full overlap with the correct ROIs within the tumor segmentation masks.[27] Mean $K^{trans}$ values were calculated by considering the average values—including negative and zero values but excluding NaN values—within the tumor-mask ROIs (TM-ROIs) and used within the scoring metrics (Table 1).

#### 2.3.2 | OSIPI scoring

The entries were planned to be scored over three main scoring metrics (Table 1): accuracy, repeatability, and reproducibility. The three metrics were multiplied to produce a single final score, which implies that a method needs to score well against all three criteria in order to score well overall. Methods should return values in a repeatable way to allow tracking of any changes that occur and give accurate values for this. Reproducible methods are of pivotal importance to allow dependable use across centers or collaborators.

These scoring metrics (Table 1) are defined similarly to conventional definitions of accuracy and repeatability. It is worthwhile to mention that we opted for a novel definition of accuracy in this challenge as an alternative to the conventional definition (1-bias), to overcome the limitations such as negative scores. These new metric definitions enable the separate metrics to be combined into one overall score that is equally influenced by the three criteria.

### 2.4 | Preliminary evaluation of the challenge

An independent team of two scientists, D.A.H. and J.D.C., were invited to perform a test run for the whole challenge process. One of these scientists (D.A.H.) used their in-house $K^{trans}$ quantification software on the DCE-MRI scans of all subjects and visits in the synthetic and clinical cohorts, according to the challenge guidelines. They provided the $K^{trans}$ maps along with the SOP of the analysis approach. The second scientist (J.D.C.) from the same institution followed the SOP to reproduce the results. The final results by the two scientists were used to test our scoring metrics and revise the challenge guidelines where

necessary. As these results may be biased, they are not reported.

### 2.5 | Statistical analysis

To take full advantage of the data submitted for the challenge, submissions were evaluated by several voxelwise approaches complementing the $\text{Score}_{\text{accuracy}}$ and $\text{Score}_{\text{reproduce}}$. This provided vital information, as similar TM-ROI mean $K^{\text{trans}}$ values may stem from vastly different distributions.[28]

For accuracy, a voxelwise Bland–Altman analysis was applied to assess the differences between each entry and the ground-truth $K^{\text{trans}}$ values. The mean difference and SD were calculated for all the synthetic visits combined. Bland–Altman plots can be found in the Supporting Information Data S5. Additionally, the proportional change in $K^{\text{trans}}$ values within the TM-ROI ($dK^{\text{trans}}_{\text{prop}} = \left(K^{\text{trans}}_{\text{v1}} - K^{\text{trans}}_{\text{v2}}\right)/K^{\text{trans}}_{\text{v1}}$) between the two visits ($K^{\text{trans}}_{\text{v1}}$ and $K^{\text{trans}}_{\text{v2}}$) in synthetic patients was computed.

For reproducibility, a voxelwise analysis was applied to deduce the differences between the submission and neutral teams' $K^{\text{trans}}$ values. The mean difference and SD details were calculated for all patient visits combined.

To allow a more detailed conclusion about the repeatability outcomes from the submissions, two metrics were extracted: (1) TM-ROI mean $K^{trans}$ difference between clinical patient visits calculated for each of the 8 clinical patients and (2) repeatability coefficient (RC) for TM-ROI mean $K^{\text{trans}}$ ($\%RC = 2.77 \times wCV$, where $wCV$ denotes the within-subject coefficient of variation, defined by the RMS of $\sigma/\mu$)[8,29,30] to measure repeatability between clinical visits within the same submission.

## 3 | RESULTS

Ten submissions, identified by a team name, were received from May 15, 2021, through December 30, 2021. SOPs of four submissions could not be reproduced due to runtime errors or extensively long computational time. Therefore, the reproducibility score was calculated only for six submissions and $\text{OSIPI}_{\text{gold}}$ was only reported for them. To compare the methods in terms of accuracy and repeatability, $\text{OSIPI}_{\text{silver}}$ was used for the remaining submissions, ranked below those with $\text{OSIPI}_{\text{gold}}$ score.

### 3.1 | Overview of challenge entries

An overall summary of the procedures used for each submission is provided in Table 2, which includes preprocessing methods (brain masking, denoising, co-registration), PK model, AIF selection method, and the DCE-MRI image quantification tool applied. There was a wide variety of AIF selection methods ranging from manual to fully automatic, but most teams opted to apply the extended Tofts PK model.

### 3.2 | OSIPI scores

The $\text{OSIPI}_{\text{gold}}$ scores as defined in Table 1 for each received entry are given in Table 3 (see Supporting Information Table S4 for component-score confidence intervals). The highest overall score was obtained by the *DCE-NET* submission with $\text{OSIPI}_{\text{gold}} = 78\%$, followed by *Maydm* and *PerfLab* with 73% and 61%, respectively. The $\text{OSIPI}_{\text{gold}}$ scores ranged from 28% to 78% across submissions with a 59% median score. The $\text{Score}_{\text{accuracy}}$, $\text{Score}_{\text{repeat}}$, and $\text{Score}_{\text{reproduce}}$ ranged from 0.54 to 0.92, 0.64 to 0.86, and 0.65 to 1.00, with median values of 0.69, 0.81 and 0.95, respectively.

### 3.3 | Further evaluation

A summary of the mean $K^{\text{trans}}$ values extracted from the TM-ROIs for the clinical patient data sets are included in Supporting Information Table S1, with the values for the synthetic patient sets reported alongside the ground truth in Supporting Information Table S2. Figure 1 shows the distributions of $K^{\text{trans}}$ values within the TM-ROIs for all patients across the submissions. Different methods lead to vastly different mean values and distributions.

An example of the clinical data received from each submission is shown in Figure 2, which provides details on the $K^{\text{trans}}$ ($\text{min}^{-1}$) map for both visits in the same slice. The estimations within the TM-ROI and the rest of the brain are highly variable between submissions. A similar plot is detailed in Figure 3, where the $K^{\text{trans}}$ ($\text{min}^{-1}$) maps are displayed for Synthetic Patient 2, indicating the variability among different tools. Small visual differences that were picked up by each software between the two visits can be observed.

### 3.4 | Accuracy

Figure 4 summarizes the voxelwise differences between ground-truth and submitted $K^{\text{trans}}$ values for the TM-ROI in the synthetic patients. This figure illustrates the wide range of accuracy of the tools in the synthetic data. Some entries (*DCE-NET, Madym, ROCKETSHIP*, and *PerfLab*) have largely symmetric difference distributions, whereas

**TABLE 2** Overview of methods used by the challenge entries. Submissions are shown using the preferred team name with the institution below.

| Submissions | Brain mask for background removal | Denoising | DCE image coregistration | DCE quantification tool/language | AIF selection method | PK model |
|---|---|---|---|---|---|---|
| *DCE-NET* (Amsterdam UMC) | Yes | No | No | In-house *Python*[31,32] | Population-based[a] | Extended Tofts |
| *Madym* (QBI Manchester) | Yes | No | No | Open-source C++ toolkit DCE-MRI *Madym*, with its integrated *Python* wrappers[33,34] | Population-based[b] | Extended Tofts |
| *PerfLab* (ISI Brno) | Yes | No | No | *PerfLab*[35] | Automatic[36,37] | Extended Tofts |
| *MRI-QAMPER* (MSKCC) | Yes | No | No | *MRI-QAMPER MATLAB* toolkit | Semi-automatic | Extended Tofts |
| *FireVoxel* (Cornell/NYU) | Yes | Yes, only for VFA data set | No | *FireVoxel*/C++[38] | Semi-automatic (sagittal vein) | Extended Tofts |
| *ROCKETSHIP* (Barrow) | Yes | Yes | No | *ROCKETSHIP* (open source)/*MATLAB*[39] | Manual | Extended Tofts |
| *ImageJ*/*MRIcron* (ADISL) | Yes | Yes | No | *ImageJ*[40] DCE module (open source)/Java | Automatic[41] | Extended Tofts |
| *OHSU* (OHSU) | Yes | No | No | In-house *Python* | Semi-automatic | Extended Tofts |
| *UW QBI Lab* (Washington) | Yes | No | No | In-house *MATLAB* | Population-based[c] | Extended Tofts |
| *ALICE* (ICL) | Yes | No | Yes | In-house *MATLAB* | Manual | Tofts, extended Tofts, Shutterspeed, no-exchange model |

Abbreviations: AIF, arterial input function; DCE, dynamic contrast-enhanced; PK, pharmacokinetic; VFA, variable flip angle.

[a] Parameterized population AIF.[42]

[b] Parameterized population AIF.[21]

[c] Population AIF extracted from challenge data via manual AIF selection in each challenge patient.

**TABLE 3** Summary of Open Science Initiative for Perfusion Imaging (OSIPI) scores for all entries.

| Submission | Rank | $\text{Score}_{\text{accuracy}}$ | $\text{Score}_{\text{repeat}}$ | $\text{Score}_{\text{reproduce}}$ | $\text{OSIPI}_{\text{silver}}$ (%) | $\text{OSIPI}_{\text{gold}}$ (%) |
|---|---|---|---|---|---|---|
| *DCE-NET* | 1 | 0.92 | 0.85 | 1.00 | 78 | 78 |
| *Madym* | 2 | 0.85 | 0.85 | 1.00 | 73 | 73 |
| *PerfLab* | 3 | 0.78 | 0.80 | 0.98 | 62 | 61 |
| *MRI-QAMPER* | 4 | 0.72 | 0.86 | 0.93 | 62 | 57 |
| *FireVoxel* | 5 | 0.57 | 0.78 | 0.65 | 45 | 29 |
| *ROCKETSHIP* | 6 | 0.59 | 0.64 | 0.74 | 37 | 28 |
| *ImageJ/MRIcron* | 7 | 0.85 | 0.68 | N/R | 58 | N/R |
| *OHSU* | 8 | 0.67 | 0.79 | N/R | 53 | N/R |
| *UW QBI Lab* | 9 | 0.61 | 0.81 | N/R | 50 | N/R |
| *ALICE* | 10 | 0.54 | 0.86 | N/R | 46 | N/R |

Abbreviation: DCE, dynamic contrast-enhanced; N/R, not ranked.

others (*MRI-QAMPER, OHSU, FireVoxel*, and *UW QBI Lab*) show a tendency to underestimate or overestimate the ground-truth values. Bland–Altman plots of both the TM-ROI mean (Supporting Information Figure S1) and voxelwise $K^{\text{trans}}$ values (Supporting Information Figure S2) reflect trends seen in $\text{Score}_{\text{accuracy}}$ and Figure 4. It should be noted that, as Figure 4 illustrates, while the $\text{Score}_{\text{accuracy}}$ from *MRI-QAMPER* and *PerfLab* are comparable, the voxelwise $K^{\text{trans}}$ values are more variable in *MRI-QAMPER*, suggesting that the voxels with high values in *MRI-QAMPER* are averaged out in the calculation of $\text{Score}_{\text{accuracy}}$ (Supporting Information Figure S5).

A summary of $dK^{\text{trans}}_{\text{prop}}$ for Synthetic Patient 2 can be found in Supporting Information Table S5. This also includes the absolute difference from the $dK^{\text{trans}}_{\text{prop}}$ of the synthetic ground truth, ranging from 0.025 to 0.453.

## 3.5 | Repeatability

Figure 5 shows the distribution of relative changes in $K^{\text{trans}}$ values between patient visits, which correspond well with the repeatability score (Table 3). A higher mean relative difference of $K^{\text{trans}}$ values between the clinical visits corresponds to a lower repeatability score. Within Figure 1, a summary of the clinical data analysis provides patient-wise insight into the overall submission distributions (Figure 5). A Bland–Altman analysis of test–retest TM-ROI mean $K^{\text{trans}}$ values (Supporting Information Figure S3) reports test–retest variability within submissions.

Table 4 lists the %RC values with a range of 0.56% to 1.45% in the clinical patients, comparing the mean TM-ROI $K^{\text{trans}}$ between the test and retest visits. The OSIPI repeatability score (Table 3) shows strong negative correlation with %RC values between clinical visits, with a Pearson correlation coefficient of $-0.986$ ($p < 0.001$), thereby supporting the validity of the defined repeatability metric (Table 1) across these submissions.

## 3.6 | Reproducibility

To complement the overall reproducibility scores, as indicated in Table 3, the TM-ROI voxelwise differences between the original and reproduced entries are summarized in Table 5 in order of magnitude. The values differ by several orders of magnitude. The ranking of these data largely follows the $\text{Score}_{\text{reproduce}}$, but for some submissions their rank is slightly improved or worsened. A Bland–Altman plot comparing the TM-ROI mean $K^{\text{trans}}$ in each patient visit between the reproduced and original entry (Supporting Information Figure S4) follows similar trends as the voxelwise analysis (Table 5).

# 4 | DISCUSSION

In this work, we systematically evaluated variability in quantification of $K^{\text{trans}}$ obtained by different analysis pipelines using a standardized benchmark. The submissions were assessed in terms of a scoring model that measured accuracy, repeatability, and the ability to reproduce results independently.

## 4.1 | Accuracy

Submissions using population-based AIF (*DCE-NET* and *Madym*) scored highly for accuracy. This is interesting, as

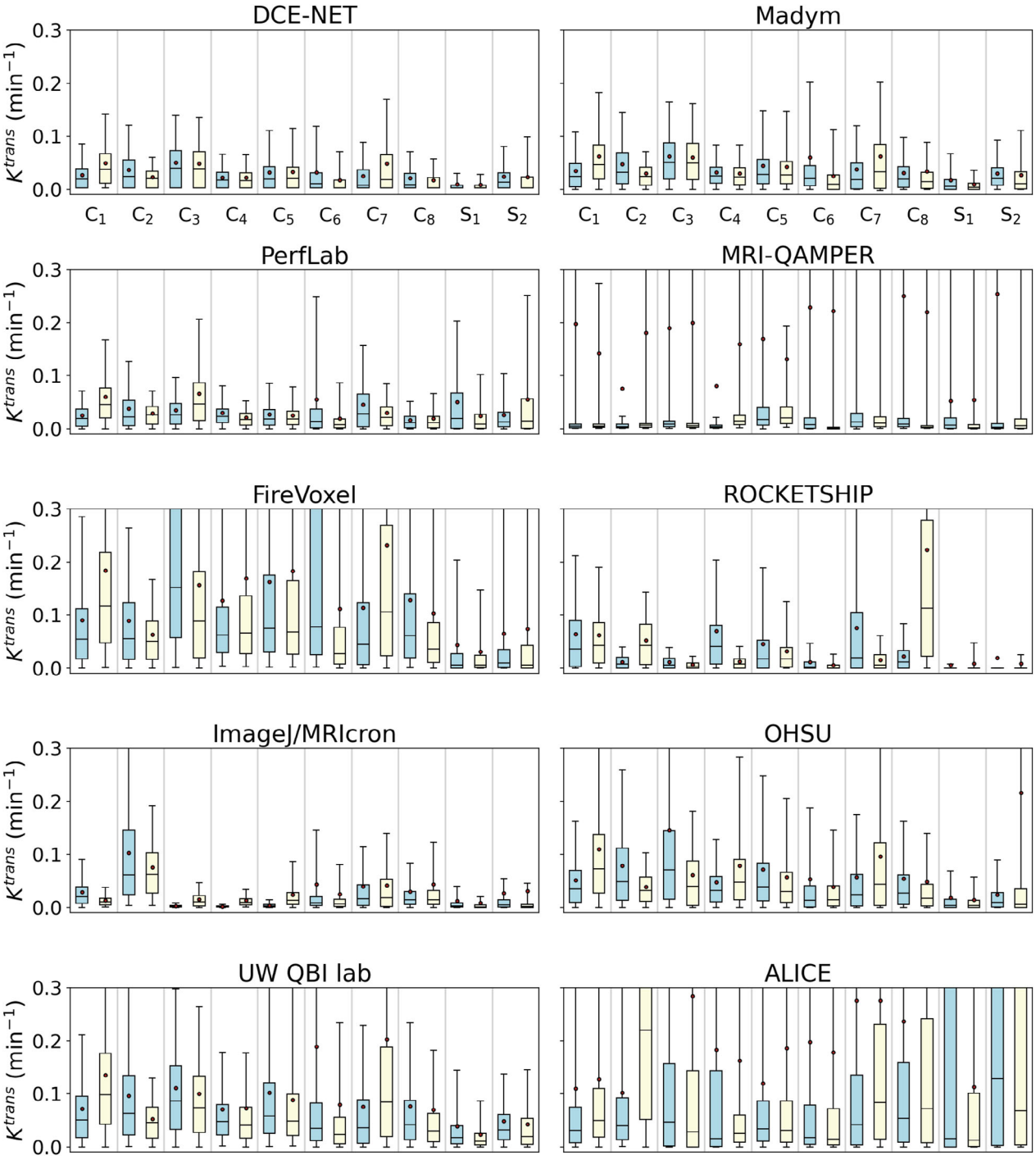


**FIGURE 1** Boxplots showing the distribution of voxelwise $K^{trans}$ values within the tumor mask for each patient visit in each challenge submission. The filled dots denote the mean $K^{trans}$ values over the tumor region, with the boxed region and central line showing the interquartile range and median, respectively. Whiskers show the 5th to 95th percentile values. The panels are arranged by the submission team, with gray lines separating each of the clinical patients and Visit 1 (blue) and Visit 2 (yellow) both shown. The $K^{trans}$ axis is limited to 0.3 min$^{-1}$ for clarity of comparison.

population-based AIFs do not account for between-subject differences and are typically seen as a means of trading off accuracy against precision.[21,43] The results indicate that this trade-off is favorable even in terms of accuracy—possibly indicating that AIF measurement biases in this application area are larger than typical between-subject differences. On the other hand, the results may have been biased by the synthetic data generation, for which the maximal value of the widely used population-based AIF[21]—although not its functional

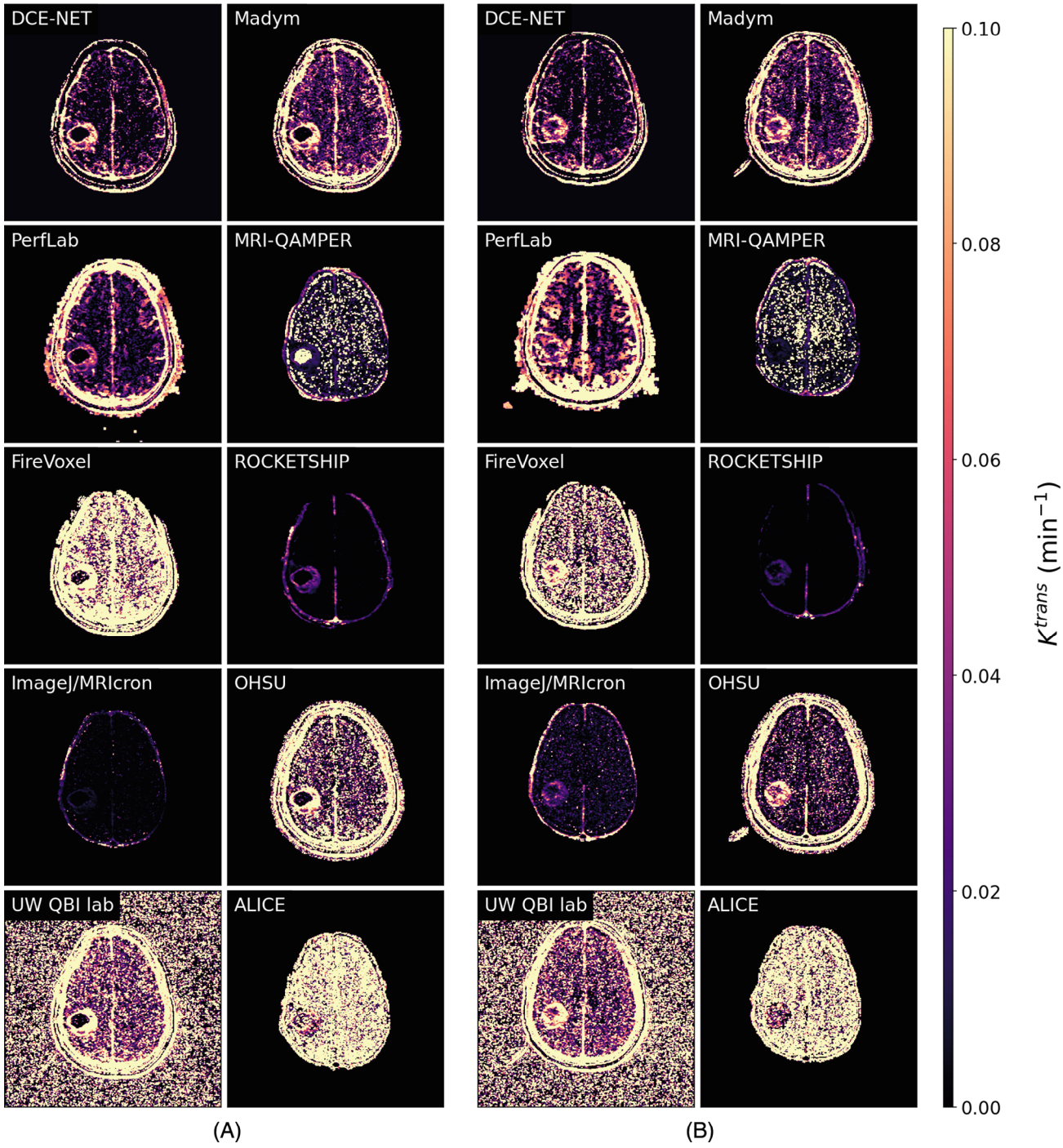


**FIGURE 2** The $K^{\text{trans}}$ values for Clinical Patient 3 over all submissions. Sets A and B correspond to Visits 1 and 2, respectively. The maximum $K^{\text{trans}}$ value is restricted to 0.1 min$^{-1}$ for comparison. NaN values are set to 0 in this figure for visualization purposes.

form—was applied to scale the selected AIF. The other highest-scoring methods for accuracy after these submissions were from *ImageJ/MRIcron* and *PerfLab*, which used fully automatic AIF methods. Due to the data-driven approach, these software tools should be more robust compared with the methods that use population-based AIF in synthetic data that have been developed with a different AIF.

The synthetic data were purposefully different between the "test" and "retest" visits, to enable detection of any method that attempted to enforce the repeatability between visits if distinct differences in values were present. It should be noted that the synthetic data were not evaluated in the repeatability scoring. This was potentially observed in the *ALICE* submission, which had a high $\text{Score}_{\text{reproduce}}$ but a low $\text{Score}_{\text{accuracy}}$. The ability to detect

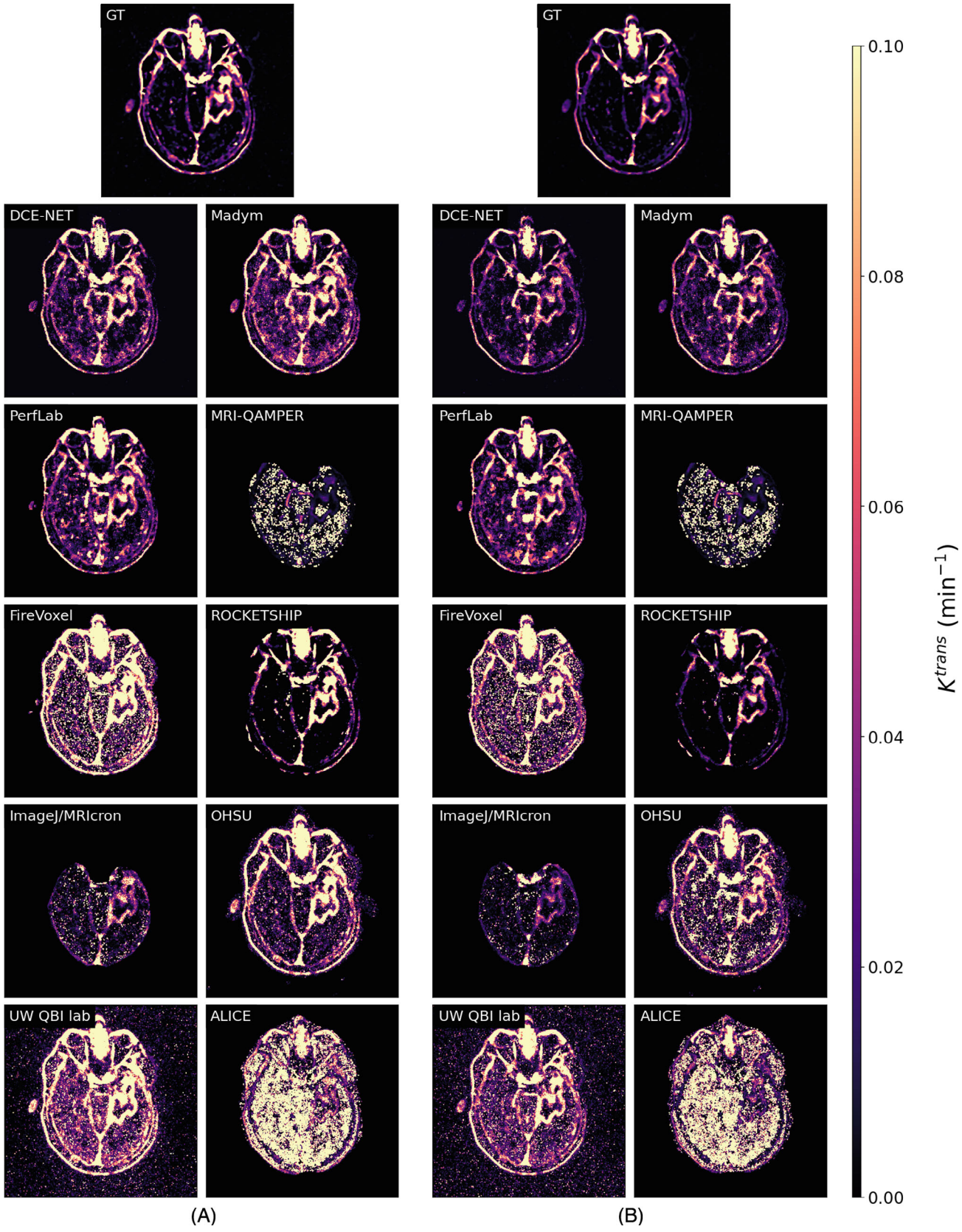


**FIGURE 3** The $K^{\text{trans}}$ maps for Synthetic Patient 2 for both Visit 1 (A) and Visit 2 (B) over all submission teams and the ground truth (GT). The maximum $K^{\text{trans}}$ value is restricted to 0.1 min$^{-1}$ for comparison. NaN values are set to 0 in this figure for visualization purposes.

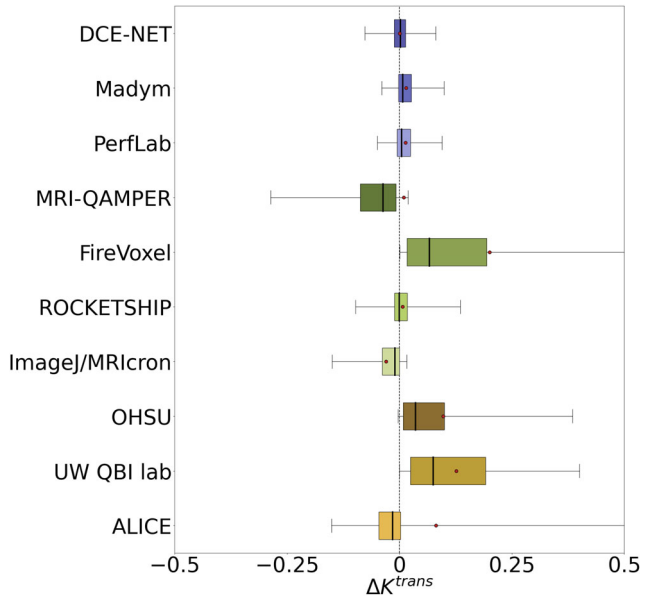


FIGURE 4 Boxplot of voxelwise differences in $K^{trans}$ ($min^{-1}$) values within tumor regions of interest between the ground-truth and entry values. Filled point shows the value bias (mean voxelwise difference between entry and ground-truth values), with the box and central line showing the interquartile range and median of the distribution. Whiskers show the 5th to 95th percentile values.

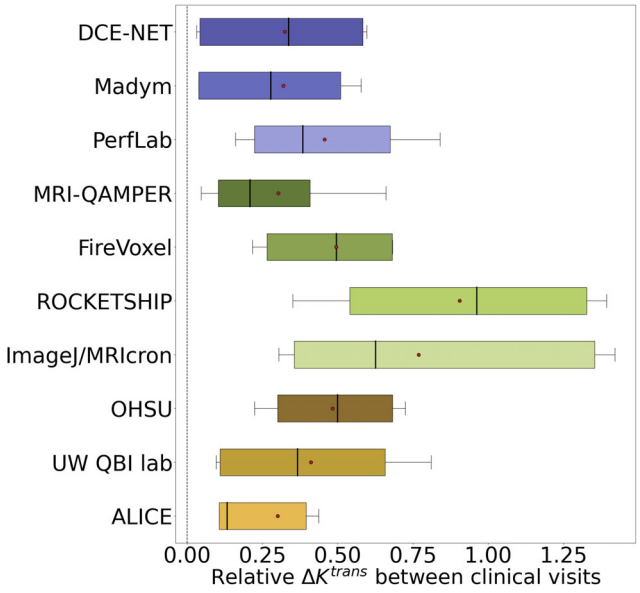


FIGURE 5 Boxplots showing the distribution of absolute relative changes in tumor mask region of interest (TM-ROI) mean $K^{trans}$ ($min^{-1}$) values between Visit 1 and Visit 2 for each submission team. The filled dots denote the mean $K^{trans}$ values over the tumor region with the boxed region and central line showing the interquartile range and median, respectively. Whiskers show the 5th to 95th percentile values.

these changes in $K^{trans}$ was further investigated using $dK^{trans}_{prop}$, to compare to the synthetic data. Based on percentage difference from $dK^{trans}_{prop}$ in the synthetic data, Madym showed the lowest deviation from the known change. This metric aimed to provide an overview of accuracy less biased to systematic offsets in $K^{trans}$ values that may have resulted from differences in concentration–signal conversion parameters and transit time handling.

Synthetic Patient 1 was based on a RIDER patient with no obvious enhancing tumor region, although this information was not revealed to participants; therefore, the ROI was placed within the normal-appearing white matter (NAWM). The $K^{trans}$ values in NAWM have been shown to be small but non-zero in several studies at high field strengths,[44–46] with a distinguishable difference in values also reported between some patient groups with healthy controls. As $K^{trans}$ is expected to be minimal in NAWM, this selection of the synthetic data was meant to evaluate how the analysis tools perform in the absence of blood–brain barrier disruption or the regime of low $K^{trans}$. This choice could have biased the score against methods that are not optimized to return $K^{trans}$ in NAWM, as some of the tools are developed and applied to the enhancing tumor regions. Additionally, methods that cover a wide range of $K^{trans}$ values by estimating more free parameters may overfit a scenario in which a priori knowledge exists. However, this evaluation was considered necessary for absolute quantification of $K^{trans}$ and standardization of values across different studies. Closer quantification of low $K^{trans}$ values was observed in Synthetic Patient 1 in *ImageJ/MRIcron*, *Madym*, *DCE-NET*, and *PerfLab* compared with other submissions (Supporting Information Table S2); these packages also achieved closer quantification of the tumor region in Synthetic Patient 2.

TABLE 4 Repeatability coefficient (%RC) values for the clinical test–retest visits, applied to tumor region-of-interest mean values for each of the submissions.

| Submission | %RC | ± 95% CI |
|---|---|---|
| *MRI-QAMPER* | 0.56 | 0.26 |
| *DCE-NET* | 0.57 | 0.24 |
| *Madym* | 0.59 | 0.27 |
| *ALICE* | 0.62 | 0.32 |
| *UW QBI Lab* | 0.73 | 0.32 |
| *OHSU* | 0.75 | 0.23 |
| *PerfLab* | 0.76 | 0.29 |
| *FireVoxel* | 0.79 | 0.27 |
| *ImageJ/MRIcron* | 1.29 | 0.50 |
| *ROCKETSHIP* | 1.45 | 0.51 |

Abbreviations: AIF, arterial input function; CI, confidence interval; DCE, dynamic contrast-enhanced.

**TABLE 5** Comparison of original submissions and reproduced $K^{trans}$ values. Columns 1 and 2 shown the mean $K^{trans}$ values across every voxel within masked regions for the original and reproduced results. Columns 3–5 give summary statistics of the differences between the submission and reproduced values calculated within each voxel.

| | Mean $K^{trans}$ ($min^{-1}$) | | Voxelwise differences $K^{trans}$ ($min^{-1}$) | | |
|---|---|---|---|---|---|
| **Submission** | **Original** | **Reproduced** | **Mean** | **SD** | **± 95% CI** |
| *Madym* | 5.4E−2 | 5.4E−2 | −3.9E−13 | 1.9E−9 | 9.4E−12 |
| *ROCKETSHIP* | 5.2E−2 | 5.3E−2 | −1.8E−4 | 1.0E−1 | 5.2E−4 |
| *DCE-NET* | 4.2E−2 | 4.2E−2 | −1.9E−4 | 3.2E−3 | 1.6E−5 |
| *PerfLab* | 4.5E−2 | 4.4E-2 | 3.6E−4 | 2.5E−2 | 1.2E−4 |
| *MRI-QAMPER* | 1.7E−1 | 1.7E−1 | −5.1E−4 | 4.7E−1 | 2.3E−3 |
| *FireVoxel* | 1.6E−1 | 8.7E−2 | 6.9E−2 | 3.1E−1 | 1.6E−3 |

Abbreviation: CI, confidence interval.

## 4.2 | Repeatability

In all submissions, %RC values for clinical patients' test–retest visits were below the 21.3% threshold[7] suggested currently by QIBA as an estimate of true change in assessment of glioblastoma. The following discussion will have to be reassessed subject to threshold changes in future QIBA guidelines. Therefore, with any of these packages, in follow-up studies on treatment response assessment in glioblastoma, any measured changes of $K^{trans}$ that exceed this threshold can be attributed to treatment response with 95% confidence. High repeatability of $K^{trans}$ is essential for longitudinal monitoring of the tumor's response to treatment or its progression,[8] as the reconstruction of the same conditions could reliably help show detection of changes, should the values deviate. Although high repeatability should not come at the expense of sensitivity to actual changes, this balancing act is crucial for longitudinal studies.

Some or all of the mean TM-ROI $K^{trans}$ values in the *MRI-QAMPER, ALICE*, and *FireVoxel* submissions are outside the interquartile range of the TM-ROI distribution, indicating heavy influence by voxels with outlier values. This may be due to the choice of masks for analysis by these submissions, as some methods are solely designed to return values within tumor regions. If outlier thresholding was applied, these entries may have performed better on this metric, as the central interquartile-range distributions appear much more consistent between visits. Regardless, the repeatability and %RC scores for *MRI-QAMPER* and *ALICE* were not overly affected, although this may not hold if the methods were applied to different data sets with more prevalent outliers.

Lower repeatability scores in *ROCKETSHIP* and *ImageJ/MRIcron* tools may have been associated with their denoising routines within the methodology. Application of this preprocessing step may have influenced differences that have been reported between the visits, potentially applying different amounts of denoising and affecting the resulting parameter retrieval.

## 4.3 | Reproducibility

The evaluators were able to reproduce six entries with no or limited interactions with the participants (limited interactions were defined as a small number of interactions on simple issues, such as resolution of installation problems). For the remaining entries, interactions consisted of concerns and issues of software malfunction or image processing time. In three out of four teams that remained unreproduced, there were some manual steps involved, namely, fully manual (*UW QBI Lab* and *ALICE*) or semi-automatic (*OHSU*) AIF selection.

Specifically, the evaluators encountered software malfunction for *ImageJ/MRIcron* and *OHSU*, and incomplete software in *UW QBI Lab*. Finally, due to fitting several PK modeling strategies, long computational time (an estimated ∼100 days) was an issue for the *ALICE* submission, as the timeline available to the evaluators was insufficient to reproduce the results.

The issues highlighted here suggest the need for clear guidelines about the level of detail in the SOPs that is required to allow the straightforward replication of the methods for widespread use. SOPs may contain video tutorials, walking through each step and clarifying from where the installation or runtime errors may stem. In addition, it would be helpful to combine all software modules into a single executable file so that future users would download the entire software package at once. Moreover, it is necessary to minimize any manual decisions that can vary across different operators for higher reproducibility. For example, the software packages requiring manual interaction within the AIF present lower reproducibility in this study, with

*ROCKETSHIP* and *FireVoxel* receiving the lowest reproducibility scores. In addition, the submissions with the highest reproducibility scores (*Madym*, *DCE-NET*, and *PerfLab*) used population-based or fully automatic AIF selection processes, thereby eliminating most user-specific interactions.

In general, software packages should ideally be developed with community distribution in mind using best practice guidelines[47–49] concerning use guidance, documentation, and issue logging. To ensure this, testing with users equipped with a range of expertise and operating systems as well as between institutes is essential. If a software package only runs under certain system requirements, it should be clearly mentioned in the SOP, so that the users can address this before installation. Then, the software may be used regardless of the user's depth of experience in quantitative DCE-MRI.

## 4.4 | Implication on future challenge design

The design of the metric based on the mean $K^{trans}$ values could be biased toward methods with outlier handling, suggesting that a score using the median could be more representative of most values produced via these methods. The analysis was rerun using the median $K^{trans}$ values, but this did little to change the ranking order; it caused the greatest improvement in accuracy scores from *ALICE* and *FireVoxel*.

A second issue encountered was the encoding of NaN values. After preliminary analysis of all submissions, it was discovered that *FireVoxel* presented values of $1 \times 10^{60}$ following extraction with the Nibabel package in *Python*.[26] After discussion with the authors of this submission, it became clear that these were intended to encode NaN values. Therefore, it was decided that these values were treated as NaN and excluded from further analysis. Before this correction, *FireVoxel* produced substantially different scores of 0.40, 0.54, and 0.56 for accuracy, repeatability, and reproducibility, respectively demonstrating the importance of proper attention to NaN handling in the challenge design. To avoid this in the future, inclusion of a specific section within the SOP outlining the NaN handling processes is recommended, to exclude unphysical or missing values from score calculation. Particularly, masked zero values would artificially lower the TM-ROI mean and may improve $OSIPI_{gold}$ and %RC values.

Our synthetic data were produced using a single-voxel approach. Although this approach provides value in terms of benchmarking, giving an equal comparison from known parameters to score the entries, for future work, the use of an interacting voxel simulation would be beneficial.[50] This approach would only have been of concern if the entries were hitting perfect accuracy levels, suggesting a bias created by the production method. In addition, values used for concentration–signal conversions may be beneficial to provide, to avoid systematic effects in $K^{trans}$ estimation. Moreover, $r_2^*$ has been shown to vary[51] but was assumed constant for synthetic data production due to short TE. An interesting extension of the modeling for future challenges would reduce such assumptions.

For future challenges, scoring should ideally include a reproducibility score for all submitted entries. An alternative setup for reproducibility could be more efficient, perhaps requiring submission of an independent reproduction along with the entry method. However, this may prevent the number of received submissions. A checklist for inclusion in the SOP could be of value, including a detailed summary of pipeline components, total run time, and any licensing requirements.

Although our proposed scoring metric for assessment of accuracy measures the bias of the submitted methods in estimating $K^{trans}$ reliably, as noted in Section 3, the voxelwise variability in $K^{trans}$ quantification may be averaged out when calculating the TM-ROI mean. Future challenges may account for the voxel-by-voxel differences (instead of TM-ROI mean) between the submitted and ground-truth $K^{trans}$ values.

## 4.5 | Study limitations

The scope of the challenge is limited to $K^{trans}$ and does not necessarily present a full report on the state of the tissue under study. The participants were asked to submit $K^{trans}$ values with no requirement to fit any specific PK model, although the extended Tofts model was almost universally applied in the submissions. In future challenges, definition of a PK model to use and requiring submission of all parameter maps ($K^{trans}$, $v_p$, $k_{ep}$) would allow for an improved analysis, focusing particularly on covariance of model parameters. Additionally, requesting AIF details would be recommended to compare the effect of AIF amplitude on variances in parameter estimation between the different approaches. This has been shown to be a factor for $K^{trans}$ estimation[52] and would inform discussion on the influence of AIF type on parameter accuracy.

The results of this study highlight the variability of pipeline choices in the submissions received. The presented results and discussions are not able to fully untangle the relative impact of each methodology choice on the resulting $K^{trans}$ values. To address this, future challenges should design methodology to investigate specific pipeline choices. For example, a challenge might supply a smaller data set but ask for several pipeline options.

In this challenge, no commercial software was submitted, which was unfortunate and may potentially reflect difficulties in the provision of licenses for the evaluators. Even with the disclaimer that there was no expectation of making code packages or software freely available beyond the evaluators, potentially the open-science aspect dissuaded interested parties from the outset. Perhaps including this disclaimer in the advertisement could bring in more commercial packages. Interested parties are encouraged to analyze the OSIPI-DCE challenge data with any commercial packages they hold licenses, to enable benchmarking among all software types.

In this article, variabilities in quantitating $K^{trans}$ using different tools were reported. Although the submitted tools were ranked using $OSIPI_{gold}$ and $OSIPI_{silver}$, the OSIPI-DCE challenge did not aim to find the "best" tool for analysis of DCE-MRI, rather to provide a platform for comparing the methods. The submitted tools may not have been specifically designed and validated for quantification of $K^{trans}$ in brain gliomas, nor were they necessarily tailored to the design of the challenge or the specific scoring metrics used. Nevertheless, the proposed $OSIPI_{gold}$ score remains beneficial as a benchmark. Other research groups working on DCE-MRI analysis tools are encouraged to apply their methods to our provided data set and evaluate their results using our scoring metrics.

## 5 | CONCLUSIONS

The OSIPI-DCE challenge highlighted the variability in $K^{trans}$ quantification between submissions and how the choice of methods in analysis pipelines affect $K^{trans}$ estimations. Further developments and consensus are needed within the community to standardize pipeline selection in different clinical settings to estimate $K^{trans}$ at a standard biomarker level. Some aspects that can be improved were identified as greater detail in description of analysis methodology to enable dissemination of approaches beyond the immediate developers, outlier handling, and the level of manual interactions as in the AIF and brain-tissue mask selection. Benchmarking efforts, such as the presented challenge, aid the translation of $K^{trans}$ from research to clinical application.[53] Moreover, as the field moves toward increasingly complex PK and signal modeling, and application of deep learning to replace model-based approaches, benchmarking the tools that produce reliable $K^{trans}$ estimations can provide a base for comparison of other advanced markers. To this end, the challenge data and assessment methodology will persist, providing an ongoing benchmarking tool for software development and pipeline selection.

## AFFILIATIONS

[1]School of Physics and Astronomy, University of Leeds, Leeds, UK

[2]Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK

[3]Department of Radiology, University of Alabama, Birmingham, Alabama, USA

[4]Department of Radiology & Nuclear Medicine, Erasmus MC University Medical Center, Rotterdam, The Netherlands

[5]Department of Radiology, University of Wisconsin–Madison, Madison, Wisconsin, USA

[6]Corewell Health William Beaumont University Hospital, Royal Oak, Michigan, USA

[7]Department of Radiology, Neuroradiology Division, Mayo Clinic, Scottsdale, Arizona, USA

[8]Oden Institute for Computational Engineering and Sciences, The University of Texas, Austin, Texas, USA

[9]Biomedical Imaging Center, Livestrong Cancer Institutes, University of Texas at Austin, Austin, Texas, USA

[10]Departments of Biomedical Engineering, Diagnostic Medicine, Oncology, Livestrong Cancer Institutes, Oden Institute for Computational Engineering and Sciences, The University of Texas, Austin, Texas, USA

[11]Department of Imaging Physics, MD Anderson Cancer Center, Houston, Texas, USA

[12]Department of Translational Neuroscience, Barrow Neurological Institute, Phoenix, Arizona, USA

[13]Department of Biomedicine and Prevention, University of Rome, Tor Vergata, Italy

[14]Department of Surgery and Cancer, Imperial College, London, UK

[15]Department of Computer Science, University College London, London, UK

[16]Lysholm Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK

[17]Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, Boston, Massachusetts, USA

[18]University Medical Center Göttingen, Göttingen, Germany

[19]Center for Biomedical Engineering, Indian Institute of Technology Delhi, New Delhi, India

[20]Institute of Bioengineering and Bioimaging, Singapore, Singapore

[21]Institute of Psychiatry, Psychology & Neuroscience, King's College, London, UK

[22]Department of Radiology, University of Washington, Seattle, Washington, USA

[23]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[24]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[25]Department of Radiology and Nuclear Medicine, University of Amsterdam, Amsterdam, The Netherlands

[26]Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, The Netherlands

[27]Czech Academy of Sciences, Institute of Scientific Instruments, Brno, Czech Republic

[28]Czech Academy of Sciences, Institute of Information Theory and Automation, Praha, Czech Republic

[29]Department of Radiology, Weill Cornell Medical College, New York, New York, USA

[30]Department of Radiology, Grossman School of Medicine, New York University, New York, New York, USA

[31]Division of Cancer Sciences, University of Manchester, Manchester, UK

[32]Department of Radiology, The Christie Hospital NHS Trust, Manchester, UK

[33]Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, UK

[34]Center for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, UK

[35]Bioxydyn Ltd, Manchester, UK

[36]Advanced Imaging Research Center, Oregon Health & Science Institute, Portland, Oregon, USA

[37]Department of Radiology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

[38]Department of Diagnostic Medicine, University of Texas, Austin, Texas, USA

[39]Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany

[40]Department of Radiological Sciences, University of California, Los Angeles, California, USA

[41]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[42]Center for Data-Driven Discovery, Division of Neurosurgery, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[43]Quantitative MR Imaging and Spectroscopy Group, Research Center for Molecular and Cellular Imaging, Tehran University of Medical Sciences, Tehran, Iran

[44]Center for Computational Imaging & Simulation Technologies in Biomedicine, School of Computing/School of Medicine, University of Leeds, Leeds, UK

[45]Neuroradiology Division, Department of Radiology, Mayo Clinic, Phoenix, Arizona, USA

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

[47]Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, USA

## ACKNOWLEDGMENTS

The authors thank Dr. Keyvan Farahani from the National Cancer Institute/National Institute of Health and Dr. Natenael Semmineh from the Barrow Neurological Institute for their help in providing information for the design of this challenge.

## FUNDING INFORMATION

E.S.S receives funding from an EPSRC-CASE Studentship (Ref: 2282622). The Department of Radiology of the University of Wisconsin-Madison receives research support from Bracco Diagnostics.

## CONFLICT OF INTEREST

E.S.S. receives partial research support from Bayer AG via EPSRC-CASE studentship. S.P.S. receives partial research support from Bayer AG. G.J.M.P. is a director and shareholder in Bioxydyn Limited, a company with an interest in DCEMRI. L.C.B. is an employee at Genentech Inc. R.v.d.H. is employed by the University of Wisconsin, which receives research support from Bracco Diagnostics.

## DATA AVAILABILITY STATEMENT

The clinical and synthetic data sets provided during the challenge can be obtained from the OSIPI GitHub.[54] This repository also contains the associated ground-truth $K^{trans}$ maps, the script used for scoring the entries, and the code used for DRO production.[18]

## ORCID

*Eve S. Shalom* https://orcid.org/0000-0001-8762-3726
*Rianne A. van der Heijden* https://orcid.org/0000-0001-6964-2536
*Zaki Ahmed* https://orcid.org/0000-0001-5648-0590
*David A. Hormuth II* https://orcid.org/0000-0002-9643-1694
*Julie C. DiCarlo* https://orcid.org/0000-0002-6965-3611
*Thomas E. Yankeelov* https://orcid.org/0000-0002-7022-0565
*Richard D. Dortch* https://orcid.org/0000-0002-9978-2203
*Matthew Grech-Sollars* https://orcid.org/0000-0003-3881-4870
*Anum S. Kazerouni* https://orcid.org/0000-0002-4200-534X
*Savannah C. Partridge* https://orcid.org/0000-0001-6370-9111
*Eve LoCastro* https://orcid.org/0000-0003-3975-8153
*Ivan A. Wolansky* https://orcid.org/0000-0002-2109-3942
*Radovan Jiřík* https://orcid.org/0000-0003-2555-9428
*Ondřej Macíček* https://orcid.org/0000-0002-0179-5779
*Michal Bartoš* https://orcid.org/0000-0003-4389-7703
*Jiří Vitouš* https://orcid.org/0000-0002-9183-8794
*Ayesha Bharadwaj Das* https://orcid.org/0000-0001-5782-8894
*S. Gene Kim* https://orcid.org/0000-0002-6288-0678
*Henry Rusinek* https://orcid.org/0000-0001-6035-6314
*Michael Berks* https://orcid.org/0000-0003-4727-2006
*Penny L. Hubbard Cristinacce* https://orcid.org/0000-0003-4213-3234
*James P. B. O'Connor* https://orcid.org/0000-0002-4044-8497

*Geoff J. M. Parker* https://orcid.org/0000-0003-2934-2234
*Peter S. LaViolette* https://orcid.org/0000-0002-9602-6891
*Samuel Bobholz* https://orcid.org/0000-0003-1525-7418
*Savannah Duenweg* https://orcid.org/0000-0003-4010-7737
*John Virostko* https://orcid.org/0000-0003-3413-8801
*Kyunghyun Sung* https://orcid.org/0000-0003-4175-5322
*Hamidreza Saligheh Rad* https://orcid.org/0000-0001-9065-2149
*Steven Sourbron* https://orcid.org/0000-0002-3374-3973
*Laura C. Bell* https://orcid.org/0000-0001-8164-8324
*Anahita Fathi Kazerooni* https://orcid.org/0000-0001-7131-2261

## TWITTER

*Anahita Fathi Kazerooni* anahita_fathi

## REFERENCES

1. Bell L, Ragunathan S, Kazerooni F. Contrast agent-based perfusion MRI methods. In: Choi IY, Jezzard P, eds. *Advanced Neuro MR Techniques and Applications. Vol 4*. Academic Press; 2021:195-209. doi:10.1016/B978-0-12-822479-3.00024-5
2. Zhang N, Zhang L, Qiu B, Meng L, Wang X, Hou B. Correlation of volume transfer coefficient $K_{trans}$ with histopathologic grades of gliomas. *J Magn Reson Imaging*. 2012;36:355-363. doi:10.1002/jmri.23675
3. Kickingereder P, Wiestler B, Graf M, et al. Evaluation of dynamic contrast-enhanced MRI derived microvascular permeability in recurrent glioblastoma treated with bevacizumab. *J Neurooncol*. 2015;121:373-380. doi:10.1007/s11060-014-1644-6
4. Thomas A, Arevalo-Perez J, Kaley T, et al. Dynamic contrast enhanced T1 MRI perfusion differentiates pseudoprogression from recurrent glioblastoma. *J Neurooncol*. 2015;125:183-190. doi:10.1007/s11060-015-1893-z
5. Zhu X, Li K, Jackson A. Dynamic contrast-enhanced MRI in cerebral tumours. In: Alan J, Buckley DL, Parker GJ, eds. *Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Oncology*. Springer; 2005:117-143. doi:10.1007/3-540-26420-5_9
6. O'Connor J, Jackson A, Parker G, Jayson G. DCE-MRI biomarkers in the clinical evaluation of antiangiogenic and vascular disrupting agents. *Br J Cancer*. 2007;96:189-195. doi:10.1038/sj.bjc.6603515
7. QIBA MR Biomarker Committee. MR DCE-MRI Quantification (DCEMRI-Q), Quantitative Imaging Biomarkers Alliance. Profile Stage: Public Comment. http://qibawiki.rsna.org/index.php/Profiles (accessed June 2022)
8. Shukla-Dave A, Obuchowski NA, Chenevert TL, et al. Quantitative Imaging Biomarkers Alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *J Magn Reson Imaging*. 2019;49:e101-e121. doi:10.1002/jmri.26518
9. Lindquist M. Neuroimaging results altered by varying analysis pipelines. *Nature*. 2020;582:36-37. doi:10.1038/d41586-020-01282-z
10. Stikov N, Trzasko J, Bernstein M. Reproducibility and the future of MRI research. *Magn Reson Med*. 2019;82:1981-1983. doi:10.1002/mrm.27939
11. Simmons J, Nelson L, Simonsohn U. False-positive psychology. *Psychol Sci*. 2011;22:1359-1366. doi:10.1177/0956797611417632
12. Wicherts J, Veldkamp C, Augusteijn H, Bakker M, van Aert R, van Assen M. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front Psychol*. 2016;7:1832. doi:10.3389/fpsyg.2016.01832
13. OSIPI TF 1.2. Public Version of OSIPI_DSCDCEinventory: DSC+DCE Software. https://docs.google.com/spreadsheets/d/e/2PACX-1vSOHrNIiWcwDD5BoHij1dpXKgEejtohqKF6KZQMZi3G6GzM_BP8xpuwRbjFFvDW9Q/pubhtml?gid=1031101549&single=true (accessed June 2023).
14. Carp J. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*. 2012;63:289-300. doi:10.1016/j.neuroimage.2012.07.004
15. Kazerooni A, Bell L, Van den Abeele F, et al. The open source initiative for perfusion imaging (OSIPI): DCE-MRI challenge. In: *Proceedings of the 10th Annual Meeting of ISMRM*. 2021:1094.
16. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045-1057. doi:10.1007/s10278-013-9622-7
17. Barboriak D. Data From RIDER NEURO MRI. *Cancer Imag Arch*. 2015. https://doi.org/10.7937/K9/TCIA.2015.VOSN3HN1
18. Kazerooni A, Shalom E, Ahmed Z, et al. OSIPI DCE Challenge. Accessed August 2023. https://osf.io/u7a6f/
19. Murase K. Efficient method for calculating kinetic parameters usingT1-weighted dynamic contrast-enhanced magnetic resonance imaging. *Magn Reson Med*. 2004;51:858-862. doi:10.1002/mrm.20022
20. Brix G, Kiessling F, Lucht R, et al. Microcirculation and microvasculature in breast tumors: pharmacokinetic analysis of dynamic MR image series. *Magn Reson Med*. 2004;52:420-429. doi:10.1002/mrm.20161
21. Parker G, Roberts C, Macdonald A, et al. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. *Magn Reson Med*. 2006;56:993-1000. doi:10.1002/mrm.21066
22. Pintaske J, Martirosian P, Graf H, et al. Relaxivity of Gadopentetate Dimeglumine (Magnevist), Gadobutrol (Gadovist), and Gadobenate Dimeglumine (MultiHance) in human blood plasma at 0.2, 1.5, and 3 Tesla. *Invest Radiol*. 2006;41:213-221. doi:10.1097/01.rli.0000197668.44926.f7
23. Siemonsen S, Finsterbusch J, Matschke J, Lorenzen A, Ding X, Fiehler J. Age-dependent normal values of T2* and T2′ in brain parenchyma. *Am J Neuroradiol*. 2008;29:950-955. doi:10.3174/ajnr.A0951
24. Zhang X, Petersen ET, Ghariq E, et al. In vivo blood T1 measurements at 1.5 T, 3 T, and 7 T. *Magn Reson Med*. 2013;70:1082-1086. doi:10.1002/mrm.24550
25. Yushkevich P, Gao Y, Gerig G. ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. *Proceedings of the 38th annual International Conference*

*of the IEEE Engineering in Medicine and Biology Society*; IEEE; 2016:3342-3345. doi:10.1109/EMBC.2016.7591443
26. Brett M, Markiewicz C, Hanke M, et al. nipy/nibabel: 3.2.1. 2020. Accessed March 2023. https://zenodo.org/records/4295521
27. Harris C, Millman K, van der Walt S, et al. Array programming with NumPy. *Nature*. 2020;585:357-362. doi:10.1038/s41586-020-2649-2
28. Matejka J, Fitzmaurice G. Same stats, different graphs. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM; 2017:1290-1294. doi:10.1145/3025453.3025912
29. Obuchowski N, Bullen J. Quantitative imaging biomarkers: effect of sample size and bias on confidence interval coverage. *Stat Methods Med Res*. 2018;27:3139-3150. doi:10.1177/0962280217693662
30. Obuchowski NA, Huang E, de Souza NM, et al. A framework for evaluating the technical performance of multiparameter quantitative imaging biomarkers (mp-QIBs). *Acad Radiol*. 2023;30:147-158. doi:10.1016/j.acra.2022.08.031
31. Ottens T, Barbieri S, Orton MR, et al. Deep learning DCE-MRI parameter estimation: application in pancreatic cancer. *Med Image Anal*. 2022;80:102512. doi:10.1016/j.media.2022.102512
32. Gurney-Champion OJ, Ottens T. DCE-NET, GitHu. https://github.com/oliverchampion/DCENET (accessed July 2022)
33. Berks M, Parker G, Little R, Cheung S. Madym: a C++ toolkit for quantitative DCE-MRI analysis. *J Open Source Softw*. 2021;6:3523. doi:10.21105/joss.03523
34. Berks M. madym_cxx. GitLab. https://gitlab.com/manchester_qbi/manchester_qbi_public/osipi-dce-challenge (accessed July 2022)
35. Jiřík R. PerfLab. CERIT Scientific Cloud. http://perflab.cerit-sc.cz/ (accessed July 14, 2022).
36. Jiřík R, Taxt T, Macíček O, et al. Blind deconvolution estimation of an arterial input function for small animal DCE-MRI. *Magn Reson Imaging*. 2019;62:46-56. doi:10.1016/j.mri.2019.05.024
37. Jiřík R, Bartoš M, Macíček O, Starčuk Z. Arterial input function estimation using all-channel blind deconvolution with spatial regularization in DCE-MRI. In: *Proceedings of the 31st Annual Meeting of ISMRM*. London, UK 2022 p. 1582.
38. Mikheev A, Rusinek H. FireVoxel. https://firevoxel.org/ (accessed July 2022).
39. Barnes SR, Ng TSC, Santa-Maria N, Montagne A, Zlokovic BV, Jacobs RE. ROCKETSHIP: a flexible and modular software tool for the planning, processing and analysis of dynamic MRI studies. *BMC Med Imaging*. 2015;15:19. doi:10.1186/s12880-015-0062-3
40. Schneider CA, Rasband WS, Eliceiri KW. NIH image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9:671-675. doi:10.1038/nmeth.2089
41. Singh A, Rathore RKS, Haris M, Verma SK, Husain N, Gupta RK. Improved bolus arrival time and arterial input function estimation for tracer kinetic analysis in DCE-MRI. *J Magn Reson Imaging*. 2009;29:166-176. doi:10.1002/jmri.21624
42. Rata M, Collins DJ, Darcy J, et al. Assessment of repeatability and treatment response in early phase clinical trials using DCE-MRI: comparison of parametric analysis using MR- and CT-derived arterial input functions. *Eur Radiol*. 2016;26:1991-1998. doi:10.1007/s00330-015-4012-9
43. Port RE, Knopp MV, Brix G. Dynamic contrast-enhanced MRI using Gd-DTPA: interindividual variability of the arterial input function and consequences for the assessment of kinetics in tumors. *Magn Reson Med*. 2001;45:1030-1038. doi:10.1002/mrm.1137
44. Larsson H, Courivaud F, Rostrup E, Hansen A. Measurement of brain perfusion, blood volume, and blood-brain barrier permeability, using dynamic contrast-enhanced T1 weighted MRI at 3 Tesla. *Magn Reson Med*. 2009;62:1270-1281. doi:10.1002/mrm.22136
45. Taheri S, Gasparovic C, Huisa B, et al. Blood–brain barrier permeability abnormalities in vascular cognitive impairment. *Stroke*. 2011;42:2158-2163. doi:10.1161/STROKEAHA.110.611731
46. Cramer S, Simonsen H, Frederiksen J, Rostrup E, Larsson H. Abnormal blood–brain barrier permeability in normal appearing white matter in multiple sclerosis investigated by MRI. *NeuroImage Clin*. 2014;4:182-189. doi:10.1016/j.nicl.2013.12.001
47. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal T. Good enough practices in scientific computing. *PLoS Comput Biol*. 2017;13:e1005510. doi:10.1371/journal.pcbi.1005510
48. Wilson G, Aruliah D, Brown C, et al. Best practices for scientific computing. *PLoS Biol*. 2014;12:e1001745. doi:10.1371/journal.pbio.1001745
49. van Houdt PJ, Ragunathan S, Berk M, et al. Open Science initiative for perfusion imaging (OSIPI): a community-led, open-source code library for analysis of DCE/DSC-MRI. In: *Proceedings of the 31st Annual Meeting of ISMRM*. London, UK 2022 p. 2691.
50. Hanson E, Sandmann C, Malyshev A, Lundervold A, Modersitzki J, Hodneland E. Estimating the discretization dependent accuracy of perfusion in coupled capillary flow measurements. *PloS One*. 2018;13:e0200521. doi:10.1371/journal.pone.0200521
51. Blockley NP, Jiang L, Gardener AG, Ludman CN, Francis ST, Gowland PA. Field strength dependence of R1 and R 2* relaxivities of human whole blood to prohance, vasovist, and deoxyhemoglobin. *Magn Reson Med*. 2008;60:1313-1320. doi:10.1002/mrm.21792
52. Cheng HLM. Investigation and optimization of parameter accuracy in dynamic contrast-enhanced MRI. *J Magn Reson Imaging*. 2008;28:736-743. doi:10.1002/jmri.21489
53. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. 2017;14:169-186. doi:10.1038/nrclinonc.2016.162
54. OSIPI,Shalom E, Kazerooni AF. TF6.2_DCE-DSC-MRI_Challenges. GitHub. https://github.com/OSIPI/TF6.2_DCE-DSC-MRI_Challenges (accessed June 2022).

## SUPPORTING INFORMATION

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

**FIGURE S1.** Bland–Altman plots showing comparison between the submissions and the ground truth for all synthetic patient visits for mean tumor region-of-interest (ROI) $K^{trans}$ values. Any zero values within the tumor ROI were included in the analysis. Colored dashed line and black dashed lines in each panel show the mean difference (bias) and upper/lower limits of agreement (bias $\pm$ 1.96$\sigma$), respectively.

**FIGURE S2.** Bland–Altman plots showing the comparison between the submissions and the ground truth for all synthetic patient visits for voxelwise $K^{trans}$ values. Any zero values within the tumor region of interest (ROI) were included in the analysis. Red dashed lines show a general linear fit bias with black dashed lines giving the upper/lower limits of agreement (bias $\pm$ 1.96$\sigma$).

**FIGURE S3.** Bland–Altman plots showing the comparison between the submissions Visit 1 and Visit 2 mean tumor region-of-interest (ROI) $K^{trans}$ values across all clinical patient cases. Any zero values within the tumor ROI were included in the analysis. Colored dashed line and black dashed lines in each panel show the mean difference (bias) and upper/lower limits of agreement (bias $\pm$ 1.96$\sigma$), respectively.

**FIGURE S4.** Bland–Altman plots showing the comparison between the submissions and the neutral evaluators for all visits mean tumor region-of-interest (ROI) $K^{trans}$ values. This is displayed only for submissions that were reproduced. Any zero values within the tumor ROI were included in the analysis. Colored dashed line and black dashed lines in each panel show the mean difference (bias) and upper/lower limits of agreement (bias $\pm$ 1.96$\sigma$), respectively.

**FIGURE S5.** Voxelwise $K^{trans}$ values for the synthetic patients across the *MRI-QAMPER* and *PerfLab* submissions. These submissions have a very similar accuracy score (Table 3, main manuscript), but the voxelwise differences (Figure 4, main manuscript) show larger variation in *MRI-QAMPER*. In each panel, the black line shows the visit tumor mask mean, and the red dashed line at 0.1 shows the Figure 3 cutoff value. Different $K^{trans}$ ranges are reported for Synthetic Patient 1 (top four panels) and Patient 2 (bottom four panels) to match the range of $K^{trans}$ values present in each patient across both submissions.

**TABLE S1.** Values of the recovered $K^{trans}$ values (min$^{-1}$) over the tumor mask for each clinical visit in all entries. Here, $C_1$ $v_1$ denotes clinical patient Set 1 at Visit 1, and the naming system follows directly for the remaining sets.

**TABLE S2.** Values of the recovered $K^{trans}$ values (min$^{-1}$) over the tumor mask for each synthetic visit in all entries and the ground-truth Digital Reference Object (DRO). Here $S_1$ $v_1$ denotes Synthetic Patient 1 at Visit 1, and the naming system follows directly for the remaining sets.

**TABLE S3.** Values of the recovered $K^{trans}$ values (min$^{-1}$) over the tumor mask for each visit from the neutral evaluators in all reproduced submissions. Here, $C_1$ $v_1$ and $S_1$ $v_1$ denote clinical and Synthetic Patient 1 at Visit 1, respectively. The naming system follows directly for the remaining sets.

**TABLE S4.** A summary of Open Science Initiative for Perfusion Imaging (OSIPI) scores for all entries. The 95% confidence intervals for $\text{Score}_{\text{accuracy}}$, $\text{Score}_{\text{repeat}}$, and $\text{Score}_{\text{reproduce}}$ are shown with $\pm$ notation. Confidence intervals are generated using scores for each $\sigma/\mu$ term in the summations (Table 1) separately.

**TABLE S5.** A summary of the proportional change in mean tumor mask region-of-interest (TM-ROI) $K^{trans}$ value ($dK^{trans}_{prop}$) in Synthetic Patient 2 for all submissions and the ground truth (GT). To compare the $dK^{trans}_{prop}$ values from each submission, the absolute difference from the GT $dK^{trans}_{prop}$ was calculated.

**How to cite this article:** Shalom ES, Kim H, van der Heijden RA, et al. The ISMRM Open Science Initiative for Perfusion Imaging (OSIPI): Results from the OSIPI–Dynamic Contrast-Enhanced challenge. *Magn Reson Med*. 2024;91:1803-1821. doi: 10.1002/mrm.29909