Effects of fitting methods, high b-values and image quality on diffusion and perfusion quantification and reproducibility in the calf

Ying-Hwey Nai[1]*, Xiaomeng Wang[2]*, Julian Gan[3], Cheryl Pei Ling Lian[4], Ryan Fraser Kirwan[5], Forest Su Lim Tan[5], Derek J Hausenloy[2,6-8]

[1] Clinical Imaging Research Centre, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[2] Cardiovascular & Metabolic Disorders Program, Duke-National University of Singapore Medical School, Singapore

[3] Siemens Healthineers, Singapore

[4] Health and Social Sciences Cluster, Singapore Institute of Technology, Singapore

[5] Infocomm Technology Cluster, Singapore Institute of Technology, Singapore

[6] National Heart Research Institute Singapore, National Heart Centre, Singapore

[7] Yong Loo Lin School of Medicine, National University Singapore, Singapore

[8] The Hatter Cardiovascular Institute, University College London, London, UK

* Authors have equal contributions

Corresponding author

Name: Ying-Hwey Nai, Ph.D.

Email: yinghweynai@yahoo.com

Phone: +65- 65167412

Address (Current): MD6, Centre for Translation Medicine, National University of Singapore, 14 Medical Drive, #B1-01, Singapore 117599

ORCID:

Ying-Hwey Nai: 0000-0002-6634-8785

ABSTRACT (250/250 words)

Purposes: The study aimed to optimize diffusion-weighted imaging (DWI) image acquisition and analysis protocols in calf muscles by investigating the effects of different model-fitting methods, image quality, and use of high b-value and constraints on parameters of interest (POIs). The optimized modeling methods were used to select the optimal combinations of b-values, which will allow shorter acquisition time while achieving the same reliability as that obtained using 16 b-values.

Methods: Test-retest baseline and high-quality DWI images of ten healthy volunteers were acquired on a 3T MR scanner, using 16 b-values, including a high b-value of 1200s/mm^2, and structural T1-weighted images for calf muscle delineation. Three and six different fitting methods were used to derive *ADC* from monoexponential (ME) model and D_d, f_p, and D_p from intravoxel incoherent motion (IVIM) model, with or without the high b-value. The optimized ME and IVIM models were then used to determine the optimal combinations of b-values, obtainable with the least number of b-values, using the selection criteria of coefficient of variance (*CV*) ≤10% for all POIs.

Results: The find minimum multivariate algorithm was more flexible and yielded smaller fitting errors. The 2-steps fitting method, with fixed D_d, performed the best for IVIM model. The inclusion of high b-value reduced outliers, while constraints improved 2-steps fitting only.

Conclusions: The optimal numbers of b-values for ME and IVIM models were nine and six b-values respectively. Test-retest reliability analyses showed that only *ADC* and D_d were reliable for calf diffusion evaluation, with *CV*s of 7.22% and 4.09%.

Keywords: (6/5-10 key words)

Diffusion-weighted imaging (DWI); Intravoxel Incoherent Motion (IVIM); Monoexponential (ME); Calf Muscles; Model Fitting; Reliability;

Word Abbreviations

AHC: Age-matched Healthy Control; **AM**: Adductor Magnus; **AN**: Anterior Muscles; **BL**: Baseline; **DM**: Diabetes Mellitus; **DP**: Deep posterior calf muscle compartment; **DWI**: Diffusion-weighted imaging; **EHC**: Elderly Healthy Control; **FMM:** Find minimum multivariable; **GL:** Gastrocnemius lateralis muscle; **GM**: Gastrocnemius medialis muscle; **HC**: Healthy Control; **HQ**: High-quality; **HS**: Hamstrings; **IVIM**: Intravoxel Incoherent Motion; **LA**: Lateral; **LM**: Levenberg-Marquardt algorithm; **LSQ**: Nonlinear least-squares; **ME**: Monoexponential; **MED**: Medial; **MRI**: Magnetic resonance imaging; **PAD**: Peripheral arterial disease; **PER**: Peroneal muscles; **POI**: Parameter of interest; **POS**: Posterior; **QD**: Quadriceps; **RF**: Rectus Femoris; **ROI**: Region of interest; **SOL**: Soleus muscle combined with GL muscle; **TA**: Tibialis anterior; **YHC**: Young HC; **TRR**: Trust-region-reflective algorithm.

Parameter Abbreviations

ADC: Apparent diffusion coefficient from ME; ***b***: b-values of the DWI images in equations; ***CV*:** Coefficients of variance; $\boldsymbol{D_d}$: Diffusion coefficient from IVIM; $\boldsymbol{D_p}$: Pseudodiffusion or perfusion coefficient; $\boldsymbol{f_p}$: Perfusion fraction; ***ICC***: Intraclass correlation coefficient; ***k:*** Number of measurements per subject; $\boldsymbol{MS_B}$: Mean square measurements between subjects; $\boldsymbol{MS_E}$: Mean square error; $\boldsymbol{MS_T}$: Mean square between test-retest measurements; ***n***: Total number of subjects and regions of interest; ***r***: Pearson correlation; ***rpc***: Reproducibility coefficient; ***S***: Measured signal; $\boldsymbol{S_0}$: Measured signal at b = 0 s/mm^2; $\boldsymbol{S_{int}}$: y-intercept or the extrapolated signal of the fitted logarithmic ME equation; ***SNR***: Signal-to-noise ratio; ***SSE***: Sum of squared error;

1 INTRODUCTION

Diffusion-weighted imaging (DWI) is a non-invasive means to measure the diffusion of water molecules in tissue water using magnetic resonance imaging (MRI) [1]. DWI is applied in oncology for the detection of malignant lesions [2], as water diffusion is significantly reduced in malignant tissues, and also in neurology for the detection of ischemia onset or to provide information on white matter structural integrity [1]. Quantitative evaluation of diffusion is achieved by acquiring multiple DWI images at different diffusion gradient strengths, named b-values. The b-values are dependent on the amplitude, duration, and interval between the applied gradients, and higher b-values meant higher gradient amplitude, duration, and interval between gradient pulses, and thus longer acquisition time [2]. At the same time, signal attenuation also increases with b-values, which resulted in a form of monoexponential (ME) decay [3].

Conventionally, the apparent diffusion coefficient (*ADC*) is then estimated by modeling the DWI signals across the different b-values using a ME model, by assuming Gaussian or free diffusion. However, the DWI signal is contaminated by blood perfusion from the capillary networks, which leads to signal attenuation observed at very low b-values. Intravoxel incoherent motion (IVIM) was thus proposed to separate the effect of blood perfusion from water diffusion in tissues [4]. Although IVIM does not measure perfusion directly, the pseudodiffusion or perfusion parameter (D_p), and the perfusion fraction (f_p) have been shown to correlate with micro-perfusion or microvasculature [5]. IVIM thus allowed studies to assess both diffusion and "perfusion" simultaneously in various muscle groups before and after simulation such as exercise to investigate age-related muscle perfusion impairment [6-7] or diseases with microvascular dysfunction, such as peripheral arterial disease (PAD) [7-11]. Under metabolic stress or inflammation, interstitial fluid, and blood flow increases in the muscles and hence the overall DWI signal [9,12]. Still, the estimation of IVIM parameters may be noisy, especially at small b-values [3]. This can be resolved either by acquiring DWI images at many different small b-values, or multiple signals at a few b-values and averaging them out to obtain "high-quality" images.

The acquisition of multiple MR images increases scan time and patient discomfort, which may lead to subject motion and hence poor image quality. Moreover, despite the value of MR imaging in disease diagnosis and management, there is literature showing the limited reproducibility of some MR parameters across centers, scanner parameters, and imaging platforms [13]. DWI, in addition, is affected by the number and choice of b-values, mathematical models, and fitting methods applied to derive the parameters of interest (POIs) [14]. We are interested in using DWI to evaluate diffusion and perfusion in the calf muscles of patients with PAD. Therefore, we are motivated to reduce the image acquisition time for increased patient comfort, but obtain reliable parameters for evaluation, similar to that obtained with more b-values.

Our study objective was to optimize the DWI image acquisition and analysis protocols to evaluate diffusion and perfusion in calf muscles, with a focus to derive reliable diffusion and perfusion parameters, while reducing scan time. This protocol will reduce the DWI scan time and support clinical studies with multiple image acquisition in patients. To achieve this goal, we first evaluated the performance of different fitting methods for both ME and IVIM mathematical models. The results of this evaluation would allow us to determine the most reliable method to derive the POIs, namely *ADC* from the ME model and, diffusion coefficient (D_d), f_p, and D_p from the IVIM model.

Proper estimation of diffusion and perfusion parameters may require high b-values of > 1000 s/mm^2 [1,3]. Thus, we investigated whether the inclusion of one high b-value of 1200 s/mm^2 will help to improve the IVIM modeling. The estimation of f_p and D_p is dependent on image acquisition at the smaller b-values, which can be noisy and lead to great variations in the derived values. The impact of image quality on the POIs was thus investigated by comparing baseline and high-quality (acquired with more averages) images. The selected methods were then used to select the optimal combinations of b-values for deriving reliable POIs. The optimal combinations will allow shorter acquisition time with reduced number of b-values required while achieving the same reliability as that obtained using 16 b-values. Test-retest reproducibility analyses were performed to determine the reliability of the POIs acquired using

16 b-values (ground truths) and the optimal combinations of b-values from both baseline and high-quality DWI images.

2 MATERIALS AND METHODS

2.1 Overview

The study consisted of three main parts (**Fig. 1**): (1) Image acquisition and processing to extract the signals in the DWI images acquired at different b-values, (2) modeling the signals across the b-values to obtain the diffusion and perfusion POIs and (3) determine the minimum combinations of b-values for ME and IVIM model fitting that can obtain reliable POIs. Three and six different model-fitting methods were investigated for ME and IVIM modeling. In addition, the effects of image quality, and the use of high b-value and constraints on modeling were investigated. The best ME and IVIM model-fitting methods were then selected for further investigation.

The optimal b-value combinations for ME and IVIM models were determined in three steps: (1) From 16 b-values, progressively reduce one b-value until only 3 b-values were used for modeling, excluding b-value = 0 s/mm^2, (2) calculate coefficients of variance (*CV*s) for all POIs, and (3) select the minimum combination of b-values that yielded *CV* ≤ 10% for 90% of all regions, subjects and image types. Test-retest reliability analyses were performed using four reliability metrics, as well as statistical analyses to determine the reliability of each POI and the power of the test.

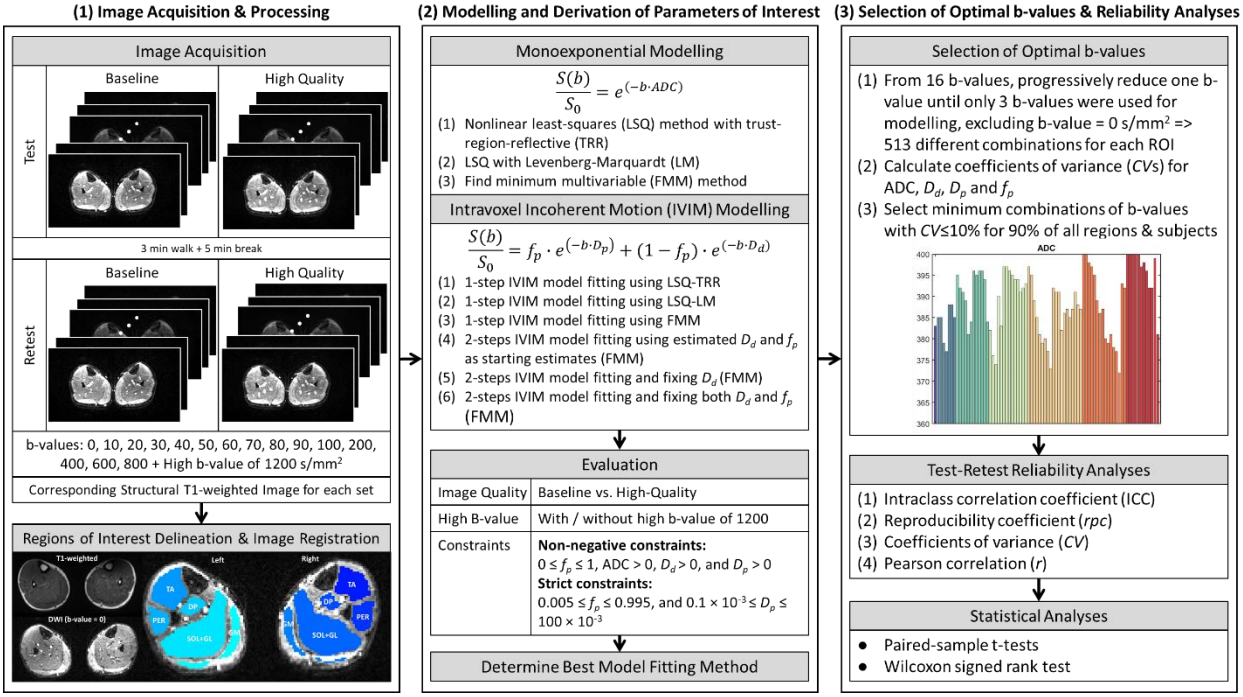

Figure 1: Overview of the analysis workflow: (1) image acquisition and processing, (2) modeling and derivation of parameters of interest, and (3) selection of optional b-values and reliability analysis.

2.2 Image Acquisition

10 healthy volunteers (M/F: 6/4, Age: 30 ± 6 [22 – 44] years, Weight: 61.4 ± 8 [47 - 72] kg) were recruited for this study, with informed consent obtained from the subjects. Subjects were positioned supine with their feet first and covered with a blanket to keep them warm. They were asked to lie on the scanner for 10 minutes before the start of image acquisition and to stay still during image acquisition. Structural T1-weighted images were acquired before a set of baseline and high-quality DWI images. The subject was then asked to come off the scanner and walk around for about 3 minutes before lying down on the scanner bed for 5 minutes for a retest set of baseline and high-quality

DWI images. This led to a total of 4 sets of DWI images, namely baseline, high-quality, baseline-retest, and high-quality-retest (**Fig. 1**). To ensure that the measurements were not due to physiological conditions after getting off the scanner, 5 subjects had their baseline scans first followed by high-quality scans and vice-versa for the other 5 subjects. The total scan times for each set of baseline and high-quality DWI images were 2.5 minutes and 7.5 minutes.

The axial MR images were acquired on the 3T Siemens Biograph mMR (Siemens Healthineers, Germany), with a 12-channel external phased-array coil covering both legs and positioned around the mid-point of the lower leg. The acquisition sequence used was a spin-echo weighted diffusion sequence with single-shot EPI (echo planar imaging) readout, with TR/TE = 3000/80 ms/ms, Flip angle = 90°. A single 10 mm 2D slice was acquired at the largest diameter of the calves, with a field of view of 349 mm sufficient to cover both legs. We chose to scan a single slice to minimize the effects of geometrical distortion, non-linearity of ADC, crosstalk, B1 inhomogeneity, and magnetization transfer, which can affect the signal value. Diffusion sensitization was achieved conventionally using a pair of Stejskal-Tanner gradients around the 180° refocusing pulse to allow shorter TE and the use of very low b-values. Diffusion was applied to 3 orthogonal directions separately and combined to form a trace-weighted image, which is independent of diffusion direction.

16 b-values of 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 400, 600, 800, and 1200 s/mm^2 were used and images were acquired in temporal sequence. For the high-quality scan, multiple averages were acquired for different b-values to increase the signal-to-noise-ratio (*SNR*) for the images to be clinically acceptable. 2 averages were used for the 1st to 11th b-values (b-values of 0 to 100 s/mm^2), 3 averages for the next 3 (b-values of 200, 400, 600 s/mm^2), and 8 averages for the last 2 (b-values of 800, 1200 s/mm^2), with increasing compensation required for the reduced *SNR* at higher b-values. This resulted in an acquisition time of 7min 27s. For the normal/standard image quality, a scan was acquired with the minimum possible 1 average for all b-values, resulting in the fastest acquisition time of 2 min 27s, with all other parameters kept identical. The trace-weighted DWI images were then reconstructed into a final matrix size of 192 × 126 (1.83 × 1.83 mm^2). Structural 3D T1-weighted SPACE was acquired with TR = 520 ms, TE = 27 ms, and Flip angle = 120°, from which a slice that matched the DWI images was selected and reconstructed into a final matrix volume of 484 × 484 (0.651 × 0.651 mm^2).

2.3 Image Processing and Regions of Interest (ROIs) Delineation

5 main regions of interest (ROIs) were delineated on each calf (total of 10 ROIs) on the T1-weighted MR image, specifically, Tibialis anterior (TA), Peroneal muscles (PER), deep posterior calf muscle compartment (DP), Gastrocnemius medialis muscle (GM) and Soleus muscle (SOL) combined with Gastrocnemius lateralis muscle (GL), while avoiding tibia and fibula bones and major blood vessel regions, as shown in **Fig 2**. ROI delineation was performed by an MD-PhD student and reviewed by an experienced senior consultant in cardiology and radiology with over 27 years of experience. Each set of DWI images was realigned to correct for motion using SPM12 (https://www.fil.ion.ucl.ac.uk/spm/). The structural T1-weighted MR images and the corresponding ROIs were then registered to each set of DWI images using a geometric transformation program in MATLAB (R2019a, The MathWorks, Inc., US), with non-reflective similarity transformation consisting of translation, rotation, and scaling. The quality of realignment of the DWI images, and structural T1-weighted MR and DWI image registration was assessed visually and deemed to be good. Good registration was generally obtained as the structural MR image was acquired before each set of baseline and high-quality DWI images in the same MR scanner, and not much subject motion was observed during DWI acquisition (maximum translation of 1.02 mm and rotation of 1.15°). Only one subject moved one leg slightly in the axial direction, which affected the slice acquisition of one DWI image at a b-value of 90 s/mm^2 and may affect modeling, especially for combinations with few b-values.

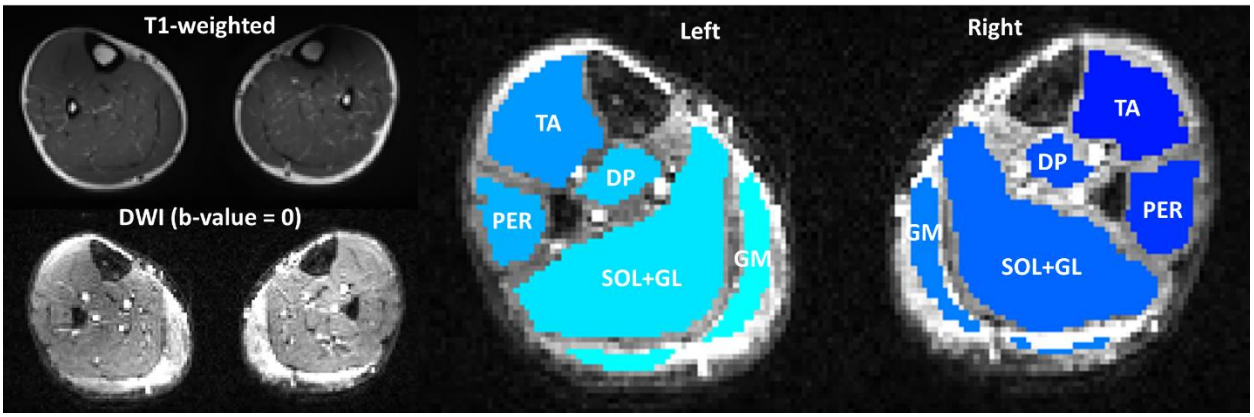

Figure 2: Structural T1-weighted SPACE image (Left-Top) registered to DWI image (b-value = 0 s/mm^2, Left-Bottom) and the corresponding regions of interest (ROIs, Right), drawn on the T1-weighted image, registered to and overlaid on the DWI image. **TA** = Tibialis anterior, **PER =** Peroneal muscles, **DP** = Deep posterior calf muscle compartment, **SOL =** Soleus muscle, **GL =** Gastrocnemius lateralis muscle, **GM** = Gastrocnemius medialis muscle.

2.4 Signal-to-Noise (SNR) Evaluation

The signal-to-noise (*SNR*) was determined using all the signals within all the ROIs across the different b-values for each set of DWI images [15]:

$$SNR(b) = Mean(S(b))/Std(S(b)) \quad (1)$$

where *S* refers to the measured signal, *b* refers to the b-values of the DWI images and Std refers to the standard deviation.

2.5 Monoexponential (ME) and Intravoxel Incoherent Motion (IVIM) models

Water diffusion is commonly modeled using the ME model, from which the apparent diffusion coefficient (*ADC*) of free water diffusion can be estimated from DWI images acquired with different *b*-values:

$$\frac{S(b)}{S_0} = e^{(-b \cdot ADC)} \quad (2)$$

Where *S* refers to the measured signal, *b* refers to the b-values of the DWI images and S_0 is the measured signal at *b* = 0 s/mm^2.

Le Bihan et al. introduced the IVIM model to separate perfusion from the surrounding capillary networks from water molecular diffusion within the tissue space [4]:

$$\frac{S(b)}{S_0} = f_p \cdot e^{(-b \cdot D_p)} + (1 - f_p) \cdot e^{(-b \cdot D_d)} \quad (3)$$

Where *S* refers to the measured signal, *b* refers to the b-values of the DWI images, S_0 is the measured signal at *b* = 0 s/mm^2, f_p is the perfusion fraction, D_p is the pseudo-diffusion or "perfusion" coefficient, and D_d is the molecular diffusion of water.

2.6 ME, 1-Step and 2-Steps IVIM Model Fitting

The ME and 1-step IVIM model fitting were performed by fitting equations (2) and (3) respectively using the nonlinear least-squares (LSQ) method with trust-region-reflective (TRR) or Levenberg-Marquardt (LM) algorithms, as well as the find minimum multivariable (FMM) method with non-negative constraints of $0 \leq f_p \leq 1$, ADC > 0, $D_d >$ 0, and D_p > 0. Non-negative constraints were used to first compare the stability of the different model-fitting methods. The initial starting values for all IVIM modeling were 0.1, 1.1 × 10^{-3} mm^2/s, and 12 × 10^{-3} mm^2/s for f_p, D_d,

and D_p, selected by comparing the average values from literature and values of the first subject. The same initial starting value for D_d was used for *ADC*.

The 2-steps IVIM model fitting was carried out by first estimating D_d (or *ADC*) from the linear fitting of the logarithmic equation (2) for DWI images with b-values ≥ 200 s/mm^2 [10]. f_p is then estimated using S_{int}, the y-intercept, or the extrapolated signal of the fitted logarithmic equation (2) [15].

$$f_p = \frac{S_0 - S_{int}}{S_0} \tag{4}$$

The final D_d, f_p, and D_p were determined by fitting equation (3) with three methods: (1) using the estimated D_d and f_p as starting estimates to determine D_d, f_p, and D_p [15], (2) fixing D_d and using the estimated f_p as a starting estimate to determine both D_p and f_p [7-9,12], or (3) fixing both D_d and f_p and determine D_p only [10]. Model fitting was performed with the FMM algorithm and the same constraints as the 1-step model fitting.

2.7 Impact of High B-value and Constraints

Most centers typically use a maximum b-value of 800 s/mm^2 or 1000 s/mm^2 for general body diffusion, as a compromise between *SNR*, scan time, and sensitivity and specificity of the lesion against background detection. Therefore, the impact of including a high b-value of 1200 s/mm^2 was investigated between baseline and high-quality DWI images to determine if it will improve IVIM modeling to yield more reliable and consistent D_d, f_p, and D_p values. In theory, higher b-values would accentuate the difference between diffusion and perfusion, but noise in measurements at high b-values of > 1000 s/mm^2 may make the measurements unreliable.

Comparison was performed between all 16 b-values and without the high value of 1200 s/mm^2 using the three different 1-step fitting methods, and three different 2-steps IVIM model-fitting methods. Model fitting performed with the FMM function had the same constraints applied as described in section **2.6**. The best method was determined for all subjects and all regions using all 16 b-values, as well as 15 b-values excluding the high b-value of 1200 s/mm^2, by comparing the quality of fit with the sum of squared error (*SSE*), as well as the estimated D_d, f_p, and D_p parameters. All model fitting and analyses were performed using MATLAB (R2019a, The MathWorks Inc., USA). Default MATLAB optimization parameters were applied except for the maximum number of function evaluations, and iterations, which were changed to 1 and 0.2 million respectively.

In addition, voxel analysis was performed using the selected model-fitting methods for the ME model, and 1-step and 2-steps IVIM models, with both lenient and stringent constraints and using all 16 b-values on one subject who had a set of baseline DWI images acquired during the pre-cuffing resting, ischemia, hyperemia and post-cuffing resting phases of 5 minutes per phase. Cuffing was done within 30 s with 50 mmHg pressure exerted on the left thigh. Further details and results can be found in **Supplementary Material: Investigation of IVIM during Ischemia and Hyperemia Phases**.

2.8 Selection of Optimal b-values

Signals measured from the respective DWI images were input into the ME and IVIM models, with the corresponding b-values of the DWI images to derive *ADC* and D_d, f_p, and D_p parameters. The ME and IVIM modeling methods were selected based on results obtained for section 2.7. Different combinations of b-values were used for modeling, starting from all 16 b-values and reduced progressively in step of one until only three b-values, including the b-value of 0 s/mm^2. This led to 513 different combinations for each ROI. In cases where no b-values > 200 s/mm^2 exist for the selected 2-steps IVIM model-fitting method, 1-step model fitting was performed using the selected constraints. The selection of optimal b-values was performed by calculating the coefficients of variance (*CV*s) of *ADC* and D_d between each combination of b-values with the "ground truth" obtained using all 16 b-values. The selection criteria for choosing the optimal combinations of b-values were set according to the selected metric, specifically *CV*, and its requirements for good reliability. Combinations that managed to yield *CV* of ≤ 10% for 90% of all regions and subjects

(10 ROIs, 10 subjects) for both test-retest baseline and high-quality DWI images were selected for further evaluation to select the combinations of b-values that yield *ADC* and D_d values close to the "ground truth" values.

2.9 Test-Retest Reliability Analyses

Test-retest reliability analyses were performed for *ADC*, D_d, D_p, and f_p for baseline and high-quality DWI images, using all 16 b-values and optimal combinations of b-values obtained in section **2.8**. Four reliability metrics, namely intraclass correlation coefficient (*ICC*), reproducibility coefficient (*rpc*), coefficients of variance (*CV*), and Pearson correlation (*r*) were calculated using programs written in MATLAB [16,17]. *ICC* with the two-way random effects, absolute agreement, single measurement model (ICC(2,1)), and α = 0.05, was determined as follows [18-20]:

$$ICC(2,1) = \frac{MS_B - MS_E}{MS_B + (k-1)\cdot MS_E + k\cdot (MS_T - MS_E)/n} \quad (5)$$

where MS_B is the mean square measurements between subjects, MS_E is the mean square error, MS_T is the mean square between test-retest measurements, *k* is the number of measurements per subject (*k* = 2) and *n* is the total number of subjects and regions of interest (*n* = 10 x 10 = 100).

ICC score ≥ 0.90 indicates excellent reliability, 0.75 ≤ *ICC* < 0.90 indicates good reliability, 0.5 ≤ *ICC* < 0.75 indicates moderate reliability and *ICC* < 0.5 indicates poor reliability (Koo, 2016). Bland-Altman analysis was performed using a normal distribution assumption for D_d and f_p, but using a non-parametric distribution assumption for D_p. The *CV* was calculated as the ratio of standard deviation over the average value. For functional MR images, we followed the same *CV* definition of parameter reproducibility as Aronhime et al. [21] where reproducibility or reliability is excellent when *CV* ≤ 10%, good when 10% < *CV* ≤ 20%, acceptable when 20% < *CV* ≤ 30%, and poor when *CV* > 30%. The reproducibility coefficient (*rpc*) was determined as 1.96 times the standard deviation of the differences in measurements, where the smaller the *rpc*, the higher the reliability in measurements. The correlation scatter plots were also plotted, with the Pearson correlation (*r*) displayed.

2.10 Statistical Analyses

Statistical analyses were performed using paired-sample t-tests and Wilcoxon signed rank test for normally-distributed and non-parametric data of *ADC*, D_d, f_p, and D_p parameters, with Gaussian distribution determined using the Kolmogorov-Smirnov test. All statistical analyses were performed with significance defined at $p \leq 0.05$, for all parameters derived from baseline and high-quality DWI images, test and retest datasets using all 16 b-values and without the b-value of 1200 s/mm^2, and using only the optimal number of b-values.

3 RESULTS

3.1 ME, 1-Step and 2-Steps IVIM Model Fitting

The *SSE*s of the ME model fits were relatively similar between the LSQ with LM and FMM methods, but LSQ with TRR method yielded higher *SSE*s (**Supplementary Table 1**). The *ADC* values derived using LSQ with TRR were also significantly smaller (**Fig. 3a**). The *SSE*s of the 1-step IVIM model fits were relatively similar between both LSQ methods but were much smaller using the FMM method (**Supplementary Table 1**). As expected, the *SSE*s were higher using the ME model than the IVIM model except for TA and DP using either LSQ method. See **Supplementary Fig. 1** for examples of how the different fitting methods performed in fitting the measured data.

For the IVIM model, the spread of f_p and D_d values were smaller across subjects and regions using both LSQ methods, with fewer outliers compared to FMM, making LSQ appealing for IVIM modeling (**Fig. 3c-d**). However, the resulting D_d values were generally significantly lower than those derived from other methods (**Fig. 3b**). Moreover, the *SSE*s were highest using the LSQ methods (**Supplementary Table 1**), as shown by the poorer fitting (**Supplementary Fig. 1d-f)**. The LSQ methods might be stuck in local minimum in determining f_p and D_p values, thus yielding small

differences in f_p and D_d values and higher *SSE*s. Thus, the FMM method is more reliable in determining the POIs from the ME and IVIM models.

The *SSE*s were relatively close for 1-step fit and 2-steps fit, using the estimated D_d and f_p as starting points, for all regions and types of DWI images (**Supplementary Table 1**). Moreover, the values of D_d, f_p, and D_p derived from the 1-step IVIM model with FMM were relatively similar to those obtained from the 2-steps IVIM model with the estimated D_d and f_p as starting estimates. These indicated that the selected initial values for model fitting were generally suitable, but as expected, the 2-steps IVIM model yielded slightly smaller *SSE*s. However, none of the three 2-steps fitting methods stood out as the clear winner, though fixing both D_d and f_p yielded much smaller *SSE*s in DP when the other methods all yielded high *SSE*s (**Supplementary Table 1**). The 1-step fitting method with FMM and 2-steps fitting method with the estimated D_d and f_p as starting points generally have larger inter-subject variations in f_p and D_p (**Fig. 3c-d**).

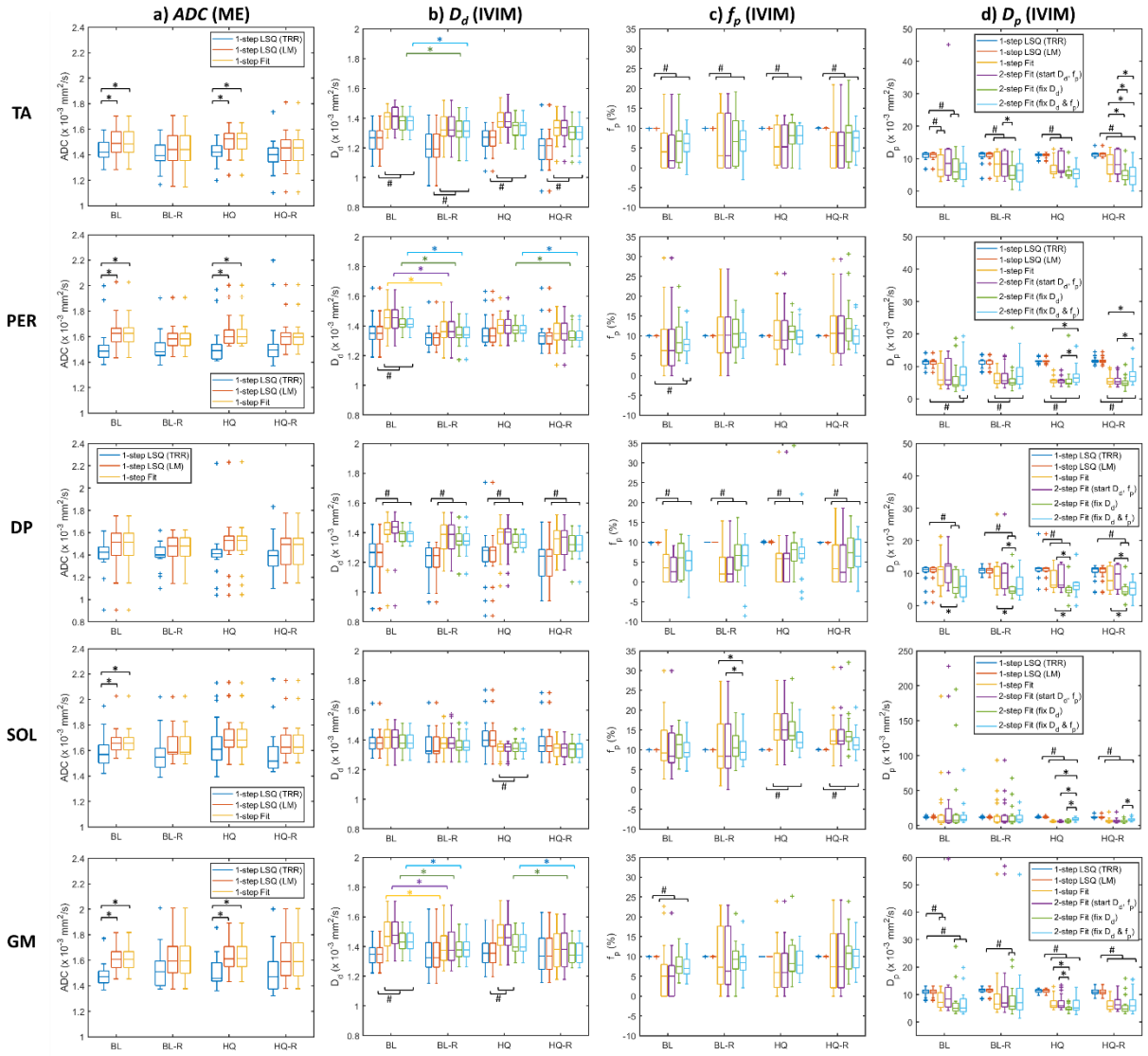

Figure 3: Boxplots showing the distributions of (a) ADC from the monoexponential (ME) model, (b) D_d, (c) f_p, and (d) D_p from the intravoxel incoherent motion (IVIM) model derived from baseline images (BL), baseline-retest images (BL-R), high-quality images (HQ), and high-quality-retest images (HQ-R), obtained using all **16** b-values. Three and six different modeling methods for ME and IVIM models, namely 1-step fitting with least square (LSQ) method with trust-region-reflective (TRR, blue) or Levenberg-Marquardt (LM, red) algorithms and find minimum multivariable

method (FMM, yellow), and 2-steps fitting with the estimated D_d and f_p as starting points (purple), or fixed D_d and used the estimated f_p as starting point (green), or fixed both D_d and f_p and determine D_p only (cyan). **TA** = Tibialis anterior, **PER =** Peroneal muscles, **DP** = Deep posterior calf muscle compartment, **SOL =** Soleus muscle and Gastrocnemius lateralis muscle, **GM** = Gastrocnemius medialis muscle. For each boxplot, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles. + indicates an outlier that lies outside 1.5 times the interquartile range. * Significant differences between specified groups, where colored lines were used between test-retest, while black lines were used between different fitting methods. # indicates significant differences between LSQ (TRR and LM) with the bracketed group(s).

3.2 Impact of Image Quality and High B-value in Model Fitting

High-quality DWI images generally yielded smaller *SSE* for all model-fitting methods, and resulted in smaller variations and fewer outliers in D_p and f_p values, compared to baseline images (**Fig. 3c-d**). Moreover, the average *SNR*s were more consistent across all 10 subjects and were closer between the test and retest scans for high-quality DWI images compared to the baseline DWI images across the 16 b-values (**Supplementary Fig. 2)**. In addition, the test and retest values of all parameters across all regions were closer, with fewer outliers, particularly in D_p, which showed that high-quality DWI images improve IVIM modeling and parameter estimation.

The inclusion of a high b-value for IVIM model fitting led to smaller inter-subject variations of D_d, D_p, and f_p and fewer outliers in f_p and D_p (**Fig. 3 vs. Fig. 4**), particularly for the 2-steps fit method with fixed D_d and f_p. The impact of a high b-value was less obvious for *ADC* from the ME model. Image quality then further reduced the variations and number of outliers in f_p and D_p (**Fig. 3 vs. Fig. 4**). Generally, the fitted curves decayed faster without the high b-value, which led to slightly higher D_d values (**Supplementary Fig. 3**). Similar to that obtained by Cho et al [22], we obtained f_p values of < 0 (**Fig. 2-3**) and D_p values > 1000 mm^2/s (**Fig. 4**), particularly using the 2-steps fit method with fixed D_d and f_p. The 2-steps fitting method of fixing D_d and using the estimated f_p as starting point to determine both D_p and f_p generally worked well with fewer cases of $f_p \leq 0$, compared to the other two 2-steps fitting methods. Inter-subject variations of D_d, D_p, and f_p were also generally smaller for all types of DWI images using this method. Therefore, the 2-steps fitting method with fixed D_d, strict constraints, and the inclusion of high b-value, was selected as the optimal IVIM model-fitting method and FMM was selected as the best algorithm for ME modeling, with non-negative constraints. See **Supplementary Material: Impact of Constraints in Model Fitting** for the selection of constraints.

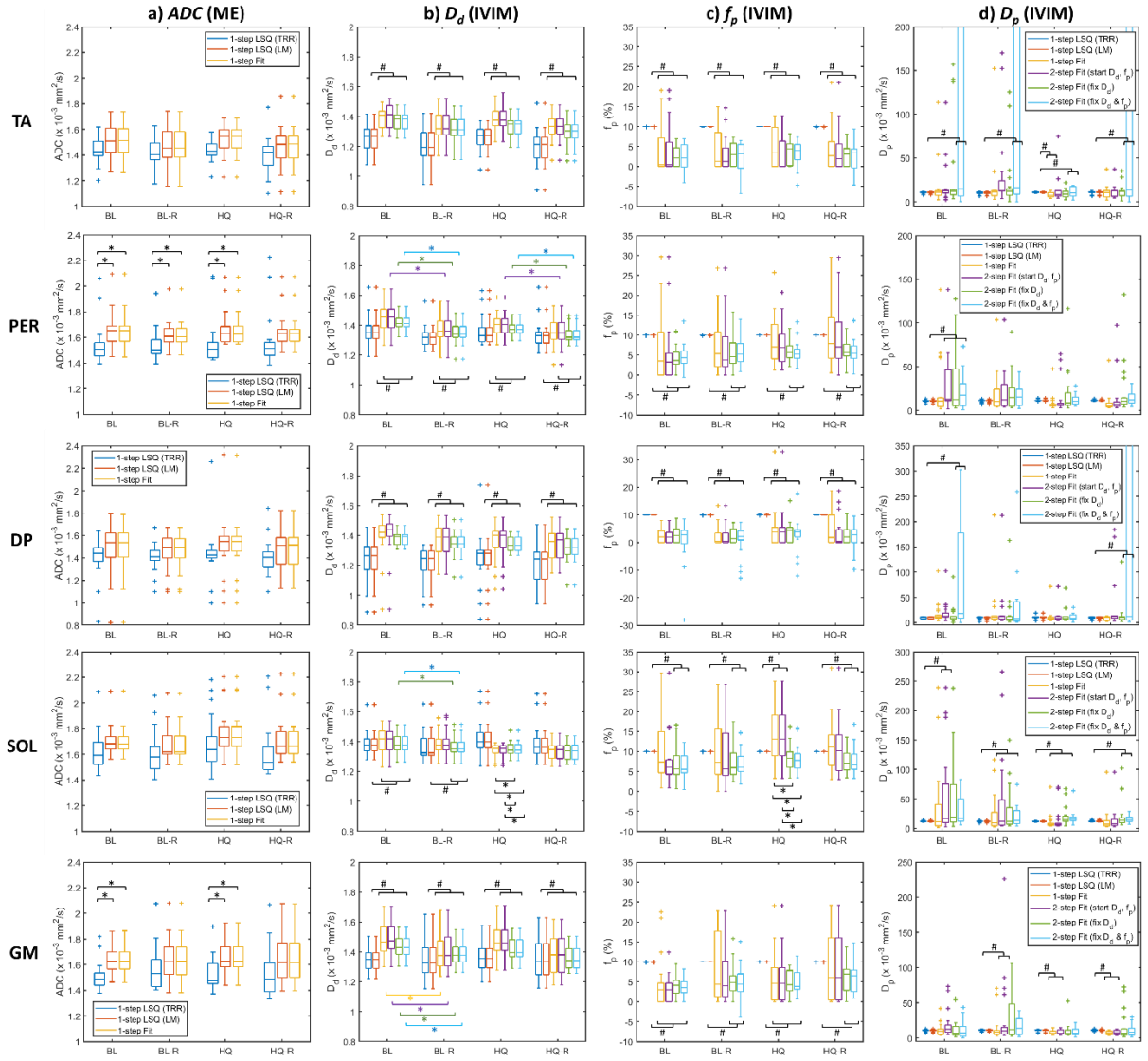

Figure 4: Boxplots showing the distributions of (a) ADC from the monoexponential (ME) model, (b) D_d, (c) f_p, and (d) D_p from the intravoxel incoherent motion (IVIM) model derived from baseline images (BL), baseline-retest images (BL-R), high-quality images (HQ), and high-quality-retest images (HQ-R), obtained using all **15** b-values (excluding high b-value of 1200 s/mm^2). Three and six different modeling methods for ME and IVIM models, namely 1-step fitting with least square (LSQ) method with trust-region-reflective (TRR, blue) or Levenberg-Marquardt (LM, red) algorithms and find minimum multivariable method (FMM, yellow), and 2-steps fitting with the estimated D_d and f_p as starting points (purple), or fixed D_d and used the estimated f_p as starting point (green), or fixed both D_d and f_p and determine D_p only (cyan). **TA** = Tibialis anterior, **PER =** Peroneal muscles, **DP** = Deep posterior calf muscle compartment, **SOL =** Soleus muscle and Gastrocnemius lateralis muscle, **GM** = Gastrocnemius medialis muscle. For each boxplot, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles. + indicates an outlier that lies outside 1.5 times the interquartile range. * indicates significant differences between specified groups, where colored lines were used between test-retest, while black lines were used between different fitting methods. # indicates significant differences between LSQ (TRR and LM) with the bracketed group(s).

3.3 Selection of Optimal B-values

Fig. 5a and **5b** show the combinations (x-axis) and the no of times each combination met the selection criteria (*CV* of ≤ 10% for 90% in all regions, subjects, and image types. Only combinations that met the selection criteria were shown in **Fig. 5**. The colors of each bar represent the number of b-values for each combination as explained in the legend. For example, the combination for D_d, highlighted by the red arrow in **Fig. 5b**, has b-values of 0, 40, 80, 200, 600, and 1200, thus the bar is in dark teal for 6 b-values. **Fig 5c** and **5d** show the corresponding average *CV*s of f_p and D_p for the same combinations in **Fig. 5b**.

The number of combinations of b-values that were able to meet the selection criteria of having *CV* of ≤ 10% for 90% of all regions and subjects for *ADC* and D_d were 82 and 179, down to 9 and 4 b-values respectively as shown in the legends of Fig. 5a and 5b-d for ME and IVIM modeling. The selected 2-steps IVIM fitting with fixed D_d led to a large number of combinations that met the selection criteria with almost all cases having *CV* of ≤ 10%, even with 4 b-values of 0, 200, 800, and 1200 s/mm^2. However, the average *CV*s of f_p and D_p of the corresponding combinations were high (above the dashed line that highlights the selection criteria of *CV* < 10%). The optimal combination of b-values, with the smallest number of b-values but achieved average *CV*s of ≤ 10% for D_d, f_p, and D_p, as highlighted with red arrows, was six b-values of 0, 40, 80, 200, 600, and 1200 s/mm^2. However, for *ADC*, the optimal combination with the smallest number of b-values but met the selection criteria, was nine b-values of 0, 20, 40, 60, 70, 90, 200, 600, and 1200 s/mm^2, as highlighted with a black arrow.

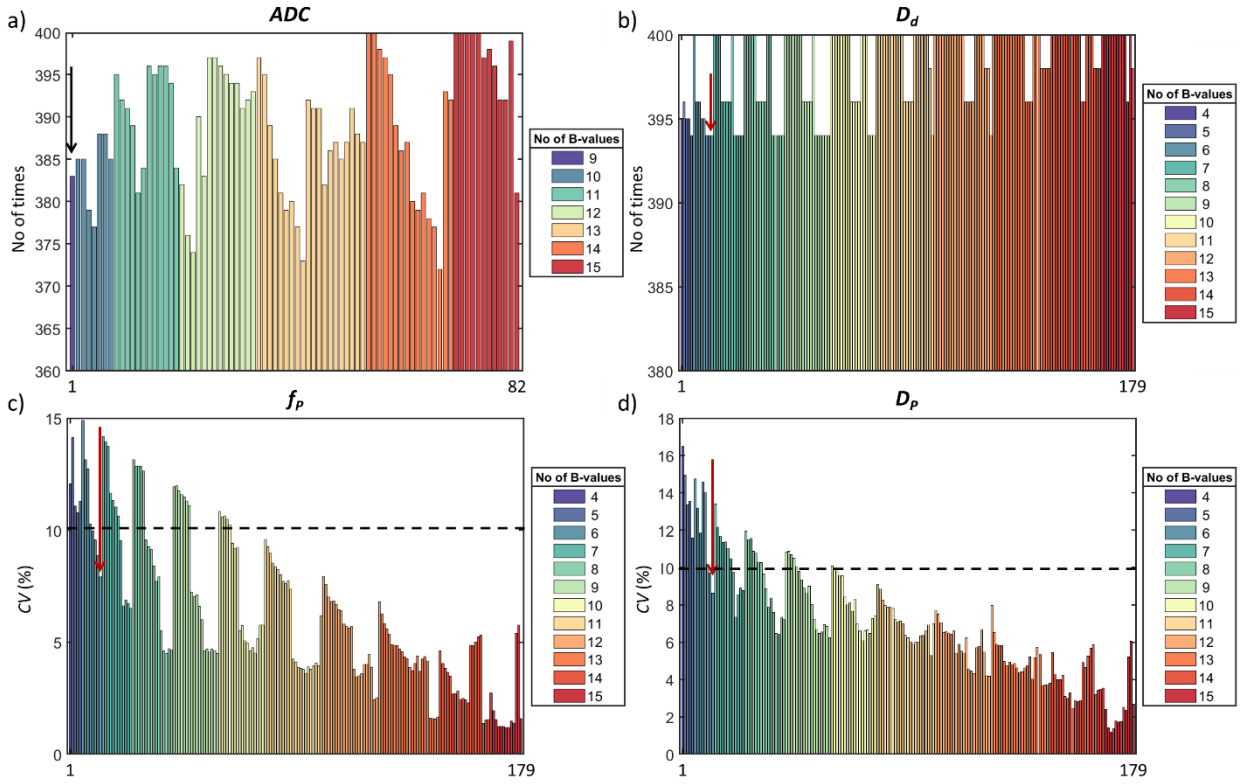

Figure 5: Combinations of b-values that yielded coefficients of variance (*CV*) of ≤ 10% for 90% in all regions and subjects (10 ROIs, 10 subjects) for test-retest of baseline and high-quality DWI images. No. of times (y-axis) each combination met the selection criteria for (a) *ADC* and (b) D_d, where the maximum no is 400 (10 subjects x 10 ROIs x 4 image types and the selection criteria of 90% of these combinations is 360) and the corresponding average *CV*s of (c) f_p and (d) D_p of the same combinations as D_d as shown in (b). The x-axis refers to the different combinations of b-values, with the colors indicating the number of b-values in each combination, as highlighted in the legend, with 82 and 179 combinations that met the selection criteria for ME and IVIM modeling. The dashed lines highlight the selection criteria of *CV* < 10%. The black and red arrows highlight the optimal combinations of b-values that achieved

average *CV*s of ≤ 10% for *ADC* from ME modeling and D_d, f_p, and D_p from IVIM modeling, with the smallest number of b-values.

3.4 Test-Retest Reliability

ADC derived from baseline DWI images was the only parameter that achieved an acceptable *ICC* value of > 0.7 (**Table 1**) [20]. However, the *SSE*s from the ME model were generally higher in baseline DWI images and also compared to all IVIM models, thus *ICC* may not be suitable for evaluating the test-retest reliability of the POIs. Among the four reliability metrics, only *CV* yielded results similar to expectations where D_d performed the best, with the smallest *CV*, followed by *ADC*, f_p then D_p, and the *CV* values were smaller for high-quality images compared to baseline images for D_d, f_p, and D_p (**Table 1**). The results confirmed our choice of using *CV* as the metric for selection criteria in selecting the optimal combinations of b-values in section 3.3. D_d yielded the smallest *CV* of < 5%, followed by *ADC*, with *CV* < 10%, which indicated that D_d and *ADC* have excellent reproducibility. However, f_p and D_p have poor reliability with *CV* > 30%, even when derived from high-quality images, with reduced number of outliers (**Fig. 3-4**), and improved modeling fit (**Supplementary Tables 1-2**). However, high-quality images reduced the variance in the POIs by up to 4 times (**Table 1**). Refer to **Supplementary Fig. 8-9** for Bland-Altman and correlation scatter plots of all POIs of baseline and high-quality DWI images.

All four reliability metrics for *ADC* were generally better using the selected fitting method with FMM than those derived using other model-fitting methods (**Table 1**). The 1-step IVIM model fitting with LSQ methods generally yielded better reliability, with much lower *CV* for f_p and D_p of < 1% and < 20%. However, a comparison of distributions of f_p and D_p values (**Fig. 3-4**), *SSE*s (Supplementary Tables 1-2), and model fits (Supplementary Fig. 1 and 3) revealed that LSQ got stuck in local minima and may not be reliable. Comparing the other IVIM model-fitting methods using FMM, the 2-steps fitting method with fixed D_d and f_p showed good reliabilities for D_d and f_p with all 4 reliability metrics but poor reliability for D_p. The selected 2-steps IVIM modeling method with fixed D_d performed next best for D_d and f_p. It also yielded good reliability for D_p after the 1-step model-fitting methods with LSQ. This supported our choice of optimal ME and IVIM model-fitting methods (**Table 1**).

The values of the four reliability metrics were slightly poorer using the optimal combinations of b-values for *ADC* than those obtained using all 16 b-values. However, for D_d, f_p, and D_p, the four reliability metrics were poorer when derived from baseline DWI images but were relatively similar if not better when derived using high-quality images. These further supported the selection of the optimal combinations of b-values and the use of high-quality images in ensuring reliable parameters for evaluation. Refer to **Supplementary Fig. 10-11** for Bland-Altman and correlation scatter plots of *ADC*, D_d, f_p, and D_p of baseline and high-quality DWI image obtained using the optimal combinations of b-values. The mean values of *ADC* and D_d, derived using all b-values and the optimal combinations of b-values, were relatively similar across the different regions and within each image type. However, that of f_p and D_p differed slightly in most regions but differed greatly in a few regions (**Supplementary Table 3**).

Models	Parameters	Baseline				High-Quality			
		ICC	*rpc*	*CV (%)*	*r*	*ICC*	*rpc*	*CV (%)*	*r*
ME (LSQ-TRR)	*ADC*	0.634	0.245	8.423	0.632	0.552	0.342	11.682	0.552
ME (LSQ-LM)	*ADC*	0.716	0.222	7.239	0.727	0.628	0.294	9.522	0.645
ME (FMM) $^{\alpha}$	*ADC*	0.717	0.221	7.218	0.729	0.629	0.294	9.524	0.645
ME (FMM) $^{\beta}$	*ADC*	0.709	0.234	7.679	0.720	0.618	0.309	10.074	0.637
1-step IVIM (LSQ-TRR)	D_d	0.733	0.166	6.496	0.754	0.642	0.227	8.823	0.661
	f_p	0.291	0.059	0.302	0.303	0.323	0.072	0.369	0.335
	D_p	0.252	3.645	16.762	0.259	0.486	3.037	13.640	0.488
1-step IVIM (LSQ-LM)	D_d	0.733	0.166	6.496	0.754	0.642	0.227	8.823	0.661
	f_p	0.291	0.059	0.302	0.303	0.323	0.072	0.369	0.335
	D_p	0.252	3.645	16.762	0.259	0.486	3.037	13.640	0.488
1-step IVIM (FMM) $^{\#}$	D_d	0.634	0.154	5.633	0.713	0.593	0.167	6.196	0.636

	f_p	0.561	12.885	86.773	0.564	0.617	12.347	69.728	0.616
	D_p	0.500	32.858	154.27	0.582	0.475	5.636	41.853	0.500
1-step IVIM (FMM) *	D_d	0.607	0.203	7.397	0.646	0.582	0.194	7.181	0.614
	f_p	0.597	12.600	76.793	0.597	0.605	12.217	67.758	0.604
	D_p	0.090	12.909	131.33	0.090	0.391	5.458	57.183	0.410
2-step IVIM (start D_d, f_p)#	D_d	0.620	0.157	5.704	0.709	0.591	0.168	6.241	0.633
	f_p	0.543	12.647	89.740	0.557	0.618	12.396	70.246	0.617
	D_p	0.299	56.159	204.66	0.450	0.517	6.011	42.604	0.533
2-step IVIM (start D_d, f_p)*	D_d	0.623	0.156	5.620	0.699	0.551	0.175	6.507	0.586
	f_p	0.310	12.669	99.551	0.318	0.570	12.439	71.501	0.569
	D_p	0.280	57.133	203.29	0.282	-0.002	27.878	221.17	-0.003
2-step IVIM (fixed D_d & f_p)#	D_d	0.516	0.121	4.489	0.640	0.597	0.108	4.090	0.684
	f_p	0.622	7.329	49.471	0.628	0.647	6.864	38.561	0.645
	D_p	0.653	1659.1	372.839	0.655	0.387	1658.8	659.82	0.393
2-step IVIM (fixed D_d & f_p)*	D_d	0.516	0.121	4.489	0.640	0.597	0.108	4.090	0.684
	f_p	0.622	7.329	49.471	0.628	0.647	6.864	38.561	0.645
	D_p	0.573	37.520	155.87	0.572	0.344	33.625	191.54	0.348
2-step IVIM (fixed D_d)#	D_d	0.516	0.121	4.489	0.640	0.597	0.108	4.090	0.684
	f_p	0.540	9.180	55.528	0.547	0.628	10.007	48.896	0.628
	D_p	0.427	39.008	205.779	0.712	0.163	6.432	58.795	0.163
2-step IVIM (fixed D_d) *, α	D_d	0.516	0.121	4.489	0.640	0.597	0.108	4.090	0.684
	f_p	0.510	8.968	53.146	0.517	0.613	9.945	48.188	0.613
	D_p	0.633	21.968	139.17	0.780	0.334	5.346	54.882	0.335
2-step IVIM (fixed D_d) β	D_d	0.515	0.129	4.783	0.625	0.600	0.109	4.159	0.683
	f_p	0.487	8.285	49.247	0.488	0.647	7.521	36.759	0.645
	D_p	0.281	31.984	197.51	0.280	0.453	5.590	53.751	0.455

Table 1: Test-retest reliability analysis results: Intraclass correlation coefficient (*ICC*), reproducibility coefficient (*rpc*), coefficients of variance (*CV*), and Pearson correlation (*r*) of 4 parameters from baseline (left) and high-quality (right) DWI images, using all **16** b-values: (1) *ADC* derived from monoexponential (ME) model using 1-step least square (LSQ) method with trust-region-reflective (TRR) or Levenberg-Marquardt (LM) algorithms and find minimum multivariable method (FMM), as well as (2) D_d, (3) f_p and (4) D_p derived from intravoxel incoherent motion (IVIM) model using 1-step and 2-steps model-fitting methods, with the estimated D_d and f_p as starting estimates, or fixed D_d and used the estimated f_p as starting estimate, or fixed both D_d and f_p and determine D_p only. # Lenient constraints applied. * Strict constraints applied. α Selected as optimal modeling method. β Modeling was performed using the selected optimal modeling methods and optimal combinations of b-values instead of all 16 b-values.

4 DISCUSSION

Our study investigated the effects of different models and fitting methods, image quality, and the use of high b-value and constraints to optimize the modeling process. From the optimized fitting methods, we determine the minimum number and combinations of b-values required for ME and IVIM modeling to obtain *ADC*, D_d, f_p, and D_p, while ensuring the reliability of these POIs were similar to that obtained using all 16 b-values in calf muscles. To the best of our knowledge, this is the first study where an inter-scan reproducibility study of ADC, D_d, f_p, and D_p was performed in the calf, and the evaluation of the effects of high-quality DWI images was investigated using high-quality DWI images generated directly from the scans with increased averages, and not from using image denoising tools [23].

4.1 Comparison with Previous Studies

Most previous works only considered only one or a few modeling methods (**Table 2**, [15,24-27]) and few investigated the impact of factors like image quality and high b-value on modeling [28]. Moreover, optimal b-values were

obtained from simulation [6,15] and not using actual clinical data. We obtained nearly similar *ADC* values but our D_d values were generally slightly lower due to the inclusion of the high b-value of 1200 s/mm^2. Still, their D_d values were similar to our results obtained without the high b-value (**Supplementary Table 3** vs. **Table 2**). The f_p and D_p values differed greatly across different studies even at resting states, which may be due to different numbers and combinations of b-values, model-fitting methods and regions of interest used, and the subjects' ages and physical conditions. Still, we obtained fairly similar ranges in these values to that obtained by Ohno et al [9] and Jungmann et al [12] using the same 2-steps fitting method with fixed D_d. Most reproducibility studies, especially cancer-related studies, investigated the reliabilities of the POIs from the ROIs delineated by multiple readers, with fewer b-values in different regions and using different models or fitting methods (**Supplementary Table 4**). These studies did not investigate the reliability of the output POIs from different models or acquired from different image qualities for clinical evaluation. This made comparison difficult but, we obtained similar trends where the inclusion of higher b-values (highest included: 1000 s/mm^2) led to more reliable POIs [29] and *CV* (%) of D_d was smaller than f_p, which was smaller than that of D_p [29,30].

Exercise Type	No of b-values	b-values used	1-step / 2-steps	Condition	Subjects (Age)	Regions	†*ADC* / D_d (× 10^{-3} mm^2/s)	f_p (%)	D_p (× 10^{-3} mm^2/s)	Ref
None	8	0, 10, 30, 70, 150, 300, 500, 800	2-steps	Resting-state	22 AHC	QD	1.46 ± 0.15	9.01 ± 3.36	39.21 ± 18.66	[10]
						HS	1.57 ± 0.10	8.23 ± 3.29	37.14 ± 19.37	
					27 DM	QD	1.43 ± 0.18	8.75 ± 2.93	31.38 ± 14.90	
						HS	1.58 ± 0.07	8.46 ± 3.08	36.34 ± 15.71	
None	9	0, 20, 50, 100, 150, 200, 300, 500, 800	1-step	Resting-state	10 AMH	AN	1.53 (1.43-1.56)	5.5 (4.5-6.0)	15.5 (11.0-20.9)	[11]
						LA	1.62 (1.52-1.70)	5.6 (3.8-6.1)	19.0 (16.2-21.4)	
						SOL	1.52 (1.38-1.67)	5.2 (4.6-5.5)	18.7 (15.6-21.0)	
						GM	1.61 (1.50-1.70)	4.7 (3.6-5.3)	19.8 (17.1-21.3)	
					14 PAD	AN	1.48 (1.24-2.02)	5.7 (4.5-12.0)	17.2 (11.0-20.5)	
						LA	1.51 (0.97-1.72)	5.6 (4.4-14.5)	14.8 (9.1-18.1)	
						SOL	1.44 (1.28-1.99)	5.2 (3.7-7.7)	17.2 (9.4-27.5)	
						GM	1.51 (1.33-1.88)	5.6 (3.7-7.7)	17.3 (12.5-21.9)	
Isometric intermittent plantar-flexion exercise	8	0, 10, 20, 50, 100, 200, 350, 500	*2-steps	Resting-state	12 HC	SOL	1.59 ± 0.038	8.5 ± 1.9	32.6 ± 12.8	[8]
						GM	1.54 ± 0.047	6.8 ± 1.3	21.0 ± 6.7	
				Exercise		SOL	1.72 ± 0.353	16 ± 6.2	43.0 ± 25.8	
						GM	1.69 ± 0.348	13 ± 6.0	47.2 ± 22.3	
				Rest		SOL	1.61 ± 0.046	8.5 ± 1.2	33.6 ± 8.59	
						GM	1.58 ± 0.041	6.8 ± 1.1	25.8 ± 6.61	
None	8	0, 25, 50, 75, 100, 200, 500, 800	†ME *2-steps	Resting-state	95 HC	AN	†1.57 ± 0.05 1.56 ± 0.048	8.70 ± 1.80	26.26 ± 4.71	[7]
						MED	†1.51 ± 0.05 1.51 (1.47-1.54)	9.94 (8.88-11.1)	26.21 ± 3.36	
						POS	†1.52 ± 0.09 1.51 ± 0.048	11.23 ± 2.51	27.3 (24.7-29.7)	
Pedal Cycling	16	0, 3, 7, 10, 15, 20, 25, 30, 40, 50, 70, 100, 200, 400, 600, 800	1-step	At baseline	4 YHC (21-30)	AM	1.81 ± 0.01	2.09 ± 0.32	119 ± 18	[6]
						RF	2.19 ± 0.03	4.63 ± 0.62	214.2 ± 15.5	
					4 EHC (60-90)	AM	1.87 ± 0.03	1.76 ± 0.11	149.1 ± 17.1	
						RF	2.22 ± 0.04	4.33 ± 0.56	204.7 ± 13.5	
				After 1st exercise [after the	4 YHC (21-30)	AM	1.76 ± 0.03 [1.81 ± 0.01]	1.44 ± 0.17 [2.10 ± 0.28]	110.4 ± 24.4 [148.7 ± 25.7]	
						RF	2.43 ± 0.03 [2.30 ± 0.03]	6.26 ± 0.89 [5.25 ± 0.48]	175.4 ± 15.4 [221.3 ± 12]	

				last exercise]	4 EHC (60-90)	AM	1.83 ± 0.02 [1.83 ± 0.01]	1.48 ± 0.27 [1.68 ± 0.24]	124.8 ± 22.9 [123.5 ± 19.0]	
						RF	2.35 ± 0.03 [2.30 ± 0.02]	5.1 ± 0.98 [5.05 ± 0.46]	241.9 ± 55.2 [229.5 ± 19.2]	
Walking and running task	9	0, 50, 100, 150, 200, 400, 600, 800, 1000	*2-steps	Rest	8 HC	Lower leg	1.33 ± 0.09	9.1 ± 3.1	11.9 ± 4.8	[12]
				Walking			1.40 ± 0.09	12.0 ± 3.8	13.3 ± 5.0	
				Running			1.42 ± 0.08	15.3 ± 4.6	15.9 ± 4.8	
Full dorsiflexion	10	0, 10, 20, 40, 80, 160, 240, 320, 500, 750	ME *2-steps	Rest	13 HC (22-27)	TA	†1.67 ± 0.18 1.55 ± 0.19	6.5 ± 3.0	8.9 ± 4.8	[9]
				After exercise			†2.06 ± 0.19	13.1 ± 3.3	20.6 ± 11.3	

Table 2: Compilation of IVIM studies in the lower leg. Data are presented as Mean ± Stdev or Median (Min-Max). **AM** = Adductor Magnus, **AN** = Anterior Muscles, **GM** = Gastrocnemius medialis muscle, **HS** = Hamstrings, **LA** = Lateral, **MED** = Medial, **POS** = Posterior, **QD** = Quadriceps, **RF** = Rectus Femoris, **SOL** = Soleus Muscles, **TA** = Tibialis Anterior, **HC** = Healthy Control, **YHC** = Young HC, **EHC** = Elderly HC, **AHC** = Age-matched HC, **DM** = Diabetes Mellitus, **PAD** = Peripheral arterial disease, **ME** = Monoexponential model. † ADC from ME modeling. * 2-steps fitting with fixed D_d.

4.2 Selection of Model-Fitting Methods

Three fitting methods for deriving *ADC* from the ME model and six fitting methods for IVIM modeling were compared. The LSQ method was commonly applied, especially with the LM algorithm. However, our results showed that the FMM method was more flexible and performed as well as, if not better, with smaller *SSE*s obtained in determining all POIs than the LSQ method (**Fig. 3**, **Supplementary Table 1**). Among the six methods for IVIM modeling, the 2-steps fitting method with fixed D_d generally worked well with few cases of $f_p \leq 0$ and D_p values > 1000 mm^2/s, and smaller inter-subject variations (**Fig. 3**). The inclusion of a high b-value reduced the number of cases of $f_p \leq 0$ and D_p values > 1000 mm^2/s, though its effect was less than the choice of fitting methods (**Fig. 3** vs. **Fig. 4**). The use of constraints improved the 2-steps fitting, with smaller *SSE*s, but was worse for 1-step fitting method (**Supplementary Material: Impact of Constraints in Model Fitting**). Therefore, we recommend the 2-steps fit with fixed D_d, with more stringent constraints and inclusion of a high b-value for IVIM modeling, but the FMM method with non-negative constraints for ME modeling.

4.3 Optimal Combinations of B-values & Reliability of Parameters

The optimal combination of b-values, with the smallest number of b-values but achieved average *CV*s ≤ 10% for D_d, f_p, and D_p, was six b-values of 0, 40, 80, 200, 600, and 1200 s/mm^2, while that for *ADC*, was nine b-values of 0, 20, 40, 60, 70, 90, 200, 600, and 1200 s/mm^2 (**Fig. 5**). Interestingly, more b-values were required for ME modeling to derive *ADC* values, even though three b-values were commonly used to derive *ADC*. On the other hand, only six b-values were required for IVIM modeling, which was different from the common view that since the perfusion effect is small, more images are required at low b-values (Iima and Le Bihan, 2016). This may however be due to the choice of the 2-steps modeling method with fixed D_d. Moreover, the use of a high b-value helped improve model fitting and the removal of outliers in f_p and D_p. Perucho et al. [31] also obtained an optimal IVIM protocol with six b-values using adaptive IVIM with fixed D_d and f_p in cervical cancer patients (**Supplementary Table 4**). Test-retest reliability analysis results between baseline and high-quality DWI images showed that only *ADC* and D_d were reliable enough to measure diffusion in the calf muscles, with *CV* < 10%, while perfusion measurements varied too much (**Table 1**) and may be subjected to noisy measurements even though there was no detected motion in images acquired with smaller b-values (**Supplementary Fig. 1c, 1f, 1i, 3c, 3f**). *ADC* was slightly less reliable using the optimal combinations of b-values than that obtained using all 16 b-values, while D_d, f_p, and D_p, were relatively similar if not better especially if derived from high-quality images (**Fig. 3**). This further supported that high-quality images not only improved model fitting but also improved reproducibility of POIs (**Table 1, Supplementary Table 3**).

4.4 Voxel-based Evaluation

IVIM model is a better model in fitting all the measured points to accurately measure diffusion in the calf. However, voxel-based analysis of *ADC*, D_d, f_p, and D_p, images, derived using the 1-step fitting method, showed clearer changes during ischemia and hyperemia phases of cuffed muscles than that obtained using the selected 2-steps fitting method (**Supplementary Fig. 6 vs. 7**). Similarly, Adelinia et al. [6] also obtained voxel map images that showed clear differences using the 1-step fitting method. However, the average values from the voxel maps of D_d, f_p, and D_p showed that the 2-steps fitting method was more reliable with more consistent parameter values during pre-cuffing and post-cuffing resting phases on the uncuffed leg. In addition, D_d showed little differences across the resting, ischemia, and hyperemia phases, while f_p and D_p showed clear changes.

4.5 Study Limitations

The impact of a high b-value was investigated by fitting 16 and 15 points, which may not be a fair comparison. However, *SSE*s were calculated using all 16 points and the fitted curves using 15 b-values decayed faster at the end compared to that using 16 b-values, especially for the 2-steps fitting methods (**Supplementary Fig. 3**). A high b-value of 1200 s/mm^2 was selected instead of even higher values due to the long scan time, particularly for high-quality images. Moreover, due to the nature of the MR imaging signal where a magnitude signal cannot be negative, there is always some background noise signal left which causes the diffusion signal to remain above a threshold, instead of asymptotically approaching 0. Thus, a curvature effect was obtained at high b-values [3], which should be evaded, to avoid over- or underestimation of the model outputs [1]. Our results also showed that *ADC* and D_d tended to be slightly lower by including high b-values. However, contrary to it, our results revealed that the inclusion of a high b-value helped to stabilize the model fitting to obtain more reliable parameters as mentioned by Iima and Le Bihan [1].

We did not investigate the impact of scanners, scan parameters, and field strength [13] or evaluate other regions of interest, such as the brain, liver, etc. [26,27,32], or the impact of subject motion on modeling as it is beyond the scope of our work. Newer methods for optimizing the model fitting were not performed in this study, which may further improve the reliability of the measurements [15,24-27]. Simulation was not performed to determine the optimal b-values. Instead, the optimal b-values were selected purely from acquired data of both baseline and high-quality DWI images. This was due to the ease of setting up the imaging protocol to acquire both baseline and high-quality DWI images within a reasonable scan time. Moreover, we had sufficient subjects and regions of interest to provide enough power and variations in the data for modeling. Interrater reliability was not performed as this study focused on the reliability of the modeling methods. Test-retest reliability assessment was performed on the same day and not over an interval of 1 or 2 weeks [20]. However, the subjects were given time to hydrate and walk around and rest before their second scans to replicate nearly similar scanning conditions for determining the inter-scan reproducibility. Moreover, baseline and high-quality images were acquired separately, instead of generating baseline DWI images by extracting one average each out of the high-quality DWI images, as this would have removed confounding variables introduced by rescanning, such as subject movements.

5 CONCLUSIONS

We investigated the impact of image quality, high b-value, and constraints, as well as three and six different fitting methods for ME and IVIM models on the outcome POIs to determine the suitable DWI analysis protocol in the calf muscles. Using the optimized ME and IVIM models, we determine the minimum optimal b-values combinations that could obtain nearly similar reliability as that obtained using 16 b-values. Compared to the nonlinear LSQ fitting methods with TRR or LM algorithms, the FMM method was more flexible and yielded smaller fitting errors for both ME and IVIM models. For IVIM modeling, the 2-steps fitting method, with fixed D_d, yielded more reliable estimates of f_p and D_p. The inclusion of a high b-value reduces outliers, while the use of constraints improved 2-steps fitting, but not 1-step fitting. The optimal combination of b-values in calf muscles for the IVIM model was six b-values of 0, 40, 80, 200, 600, and 1200 s/mm^2, while that for the ME model was nine b-values of 0, 20, 40, 60, 70, 90, 200, 600, and 1200 s/mm^2. Fewer DWI images can be acquired for calf evaluation in subjects with the optimal b-values

combinations, which will help to reduce the scan time and increase patient comfort. High-quality images improved model fitting, yielding smaller fitting errors, fewer outliers, and improved reproducibility for IVIM modeling. The *CV* values between test and retest baseline and high-quality DWI images for *ADC*, D_d, f_p, and D_p from the selected models were 7.22%, 4.09%, 36.76%, and 54.88%. This indicated that only *ADC* and D_d were reliable enough for diffusion evaluation in calf muscles, with *CV* <10%. Although IVIM could assess microvascular perfusion in calf muscles, the f_p and D_p parameters may not reliable enough for evaluation. Better models are required to improve the reliability of the parameters.

ACKNOWLEDGEMENTS

We acknowledge the nurse and radiographer at the Clinical Imaging Research Centre of the National University of Singapore for their contributions to subject preparation and data acquisition.

Funding: Derek J Hausenloy is supported by the Duke-NUS Signature Research Programme funded by the Singapore Ministry of Health's National Medical Research Council under its Singapore Translational Research Investigator Award [MOH-STaR21jun-0003], Centre Grant scheme [NMRC CG21APR1006], and Collaborative Centre Grant scheme [NMRC/CG21APRC006]. This study is supported by the National Research Foundation Competitive Research Program [NRF CRP25-2020RS-0001] and the Singapore Institute of Technology (SIT) Academic Research Fund Tier 1 grant [T1 1/2022 – 28].

Declarations of interest: None Declared

Ethical approval: This study was approved by the Domain Specific Review Board (DSRB) of the National University Hospital Singapore and was conducted following the 1964 Helsinki declaration.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Data Accessibility Statement: The authors do not have permission to share the data.

AUTHORS' CONTRIBUTIONS

Ying-Hwey Nai: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Visualization, Project administration. **Xiaomeng Wang**: Investigation, Resources, Writing - Review & Editing, Project administration. **Julian Gan**: Methodology, Writing - Review & Editing. **Cheryl Pei Ling Lian**: Writing - Review & Editing, Project administration, Funding acquisition. **Ryan Fraser Kirwan**: Funding acquisition. **Forest Su Lim Tan**: Funding acquisition. **Derek J Hausenloy**: Resources, Writing - Review & Editing, Supervision, Project administration.

References

[1] Iima, M., Le Bihan, D., 2016. Clinical intravoxel incoherent motion and diffusion MR imaging: Past, present, and future. Radiology 278, 13–32. https://doi.org/10.1148/radiol.2015150244

[2] Abou khadrah, R.S., Imam, H.H., 2019. Multiple b values of diffusion-weighted magnetic resonance imaging in evaluation of solid head and neck masses. Egyptian Journal of Radiology and Nuclear Medicine 50, 54. https://doi.org/10.1186/s43055-019-0054-3

[3] Le Bihan, D., 2019. What can we see with IVIM MRI? Neuroimage 187, 56–67. https://doi.org/10.1016/j.neuroimage.2017.12.062

[4] Le Bihan, D., Breton, E., Lallemand, D., Aubin, M.L., Vignaud, J., Laval-Jeantet, M., 1988. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. 168, 497–505. https://doi.org/10.1148/RADIOLOGY.168.2.3393671

[5] Chabert, S., Verdu, J., Huerta, G., Montalba, C., Cox, P., Riveros, R., Uribe, S., Salas, R., Veloz, A., 2020. Impact of b-value sampling scheme on brain ivim parameter estimation in healthy subjects. Magn. Reson. Med. Sci. 19, 216–226. https://doi.org/10.2463/mrms.mp.2019-0061

[6] Adelnia, F., Shardell, M., Bergeron, C.M., Fishbein, K.W., Spencer, R.G., Ferrucci, L., Reiter, D.A., 2019. Diffusion-weighted MRI with intravoxel incoherent motion modeling for assessment of muscle perfusion in the thigh during post-exercise hyperemia in younger and older adults. NMR Biomed. 32, 1–24. https://doi.org/10.1002/nbm.4072

[7] Yoon, M.A., Hong, S.J., Ku, M.C., Kang, C.H., Ahn, K.S., Kim, B.H., 2018. Multiparametric MR imaging of age-related changes in healthy thigh muscles. Radiology 287, 235–246. https://doi.org/10.1148/radiol.2017171316

[8] Mastropietro, A., Porcelli, S., Cadioli, M., Rasica, L., Scalco, E., Gerevini, S., Marzorati, M., Rizzo, G., 2018. Triggered intravoxel incoherent motion MRI for the assessment of calf muscle perfusion during isometric intermittent exercise. NMR Biomed. 31, 1–13. https://doi.org/10.1002/nbm.3922

[9] Ohno, N., Miyati, T., Fujihara, S., Gabata, T., Kobayashi, S., 2020. Biexponential analysis of intravoxel incoherent motion in calf muscle before and after exercise: Comparisons with arterial spin labeling perfusion and T2. Magn. Reson. Imaging 72, 42–48. https://doi.org/10.1016/j.mri.2020.06.003

[10] Sigmund, E.E., Baete, S.H., Luo, T., Patel, K., Wang, D., Rossi, I., Duarte, A., Bruno, M., Mossa, D., Femia, A., Ramachandran, S., Stoffel, D., Babb, J.S., Franks, A.G., Bencardino, J., 2018. MRI assessment of the thigh musculature in dermatomyositis and healthy subjects using diffusion tensor imaging, intravoxel incoherent motion and dynamic DTI. Eur. Radiol. 28, 5304–5315. https://doi.org/10.1007/S00330-018-5458-3

[11] Suo, S., Zhang, L., Tang, H., Ni, Q., Li, S., Mao, H., Liu, X., He, S., Qu, J., Lu, Q., Xu, J., 2018. Evaluation of skeletal muscle microvascular perfusion of lower extremities by cardiovascular magnetic resonance arterial spin labeling, blood oxygenation level-dependent, and intravoxel incoherent motion techniques. J. Cardiovasc. Magn. Reson. 20, 1–14. https://doi.org/10.1186/s12968-018-0441-3

[12] Jungmann, P.M., Pfirrmann, C., Federau, C., 2019. Characterization of lower limb muscle activation patterns during walking and running with Intravoxel Incoherent Motion (IVIM) MR perfusion imaging. Magn. Reson. Imaging 63, 12–20. https://doi.org/10.1016/j.mri.2019.07.016

[13] Schmeel, F.C., 2019. Variability in quantitative diffusion-weighted MR imaging (DWI) across different scanners and imaging sites: is there a potential consensus that can help reducing the limits of expected bias? Eur. Radiol. 29, 2243–2245. https://doi.org/10.1007/s00330-018-5866-4

[14] Feng, Z., Min, X., Margolis, D.J.A., Duan, C., Chen, Y., Sah, V.K., Chaudhary, N., Li, B., Ke, Z., Zhang, P., Wang, L., 2017. Evaluation of different mathematical models and different b-value ranges of diffusionweighted imaging in peripheral zone prostate cancer detection using b-value up to 4500 s/mm2. PLoS One 12, 1–15. https://doi.org/10.1371/journal.pone.0172127

[15] Ye, C., Xu, D., Qin, Y., Wang, L., Wang, R., Li, W., Kuai, Z., Zhu, Y., 2019. Estimation of intravoxel incoherent motion parameters using low b-values. PLoS One 14, 1–16. https://doi.org/10.1371/journal.pone.0211911

[16] Klein, R. 2022. Bland-Altman and Correlation Plot: (https://www.mathworks.com/matlabcentral/fileexchange/45049-bland-altman-and-correlation-plot), MATLAB Central File Exchange. Retrieved August 1, 2022.

[17] Zoeller, T. 2022. Intraclass correlation coefficient with confidence intervals: (https://www.mathworks.com/matlabcentral/fileexchange/26885-intraclass-correlation-coefficient-with-confidence-intervals), MATLAB Central File Exchange. Retrieved August 2, 2022.
[18] Koo, T.K., Li, M.Y., 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
[19] Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability.1. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979, 86:420–8. Psychol. Bull. 86, 420–8.
[20] Vilagut, G., 2014. Encyclopedia of Quality of Life and Well-Being Research, Encyclopedia of Quality of Life and Well-Being Research. https://doi.org/10.1007/978-94-007-0753-5
[21] Aronhime, S., Calcagno, C., Jajamovich, G.H., Dyvorne, H.A., Robson, P., Dieterich, D., Isabel Fiel, M., Martel-Laferriere, V., Chatterji, M., Rusinek, H., Taouli, B., 2014. DCE-MRI of the liver: Effect of linear and nonlinear conversions on hepatic perfusion quantification and reproducibility. J. Magn. Reson. Imaging 40, 90–98. https://doi.org/10.1002/jmri.24341
[22] Cho, G.Y., Moy, L., Zhang, J.L., Baete, S., Lattanzi, R., Moccaldi, M., Babb, J.S., Kim, S., Sodickson, D.K., Sigmund, E.E., 2015. Comparison of fitting methods and b-value sampling strategies for intravoxel incoherent motion in breast cancer. Magn. Reson. Med. 74, 1077–1085. https://doi.org/10.1002/mrm.25484
[23] Huang, H.M., 2020. Reliable estimation of brain intravoxel incoherent motion parameters using denoised diffusion-weighted MRI. NMR Biomed. 33, 1–8. https://doi.org/10.1002/nbm.4249
[24] Barbieri, S., Donati, O.F., Froehlich, J.M., Thoeny, H.C., 2016. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs. Magn. Reson. Med. 75, 2175–2184. https://doi.org/10.1002/MRM.25765
[25] Chevallier, O., Zhou, N., He, J., Loffroy, R., Wáng, Y.X.J., 2018. Removal of evidential motion-contaminated and poorly fitted image data improves IVIM diffusion MRI parameter scan-rescan reproducibility. Acta Radiol. 59, 1157–1167. https://doi.org/10.1177/0284185118756949
[26] Koopman, T., Martens, R., Gurney-Champion, O.J., Yaqub, M., Lavini, C., de Graaf, P., Castelijns, J., Boellaard, R., Marcus, J.T., 2021. Repeatability of IVIM biomarkers from diffusion-weighted MRI in head and neck: Bayesian probability versus neural network. Magn. Reson. Med. 85, 3394–3402. https://doi.org/10.1002/MRM.28671
[27] Neil, J.J., Bretthorst, G.L., 1993. On the use of Bayesian probability theory for analysis of exponential decay data: an example taken from intravoxel incoherent motion experiments. Magn. Reson. Med. 29, 642–647. https://doi.org/10.1002/MRM.1910290510
[28] Ahn, H.S., Kim, S.H., Kim, J.Y., Park, C.S., Grimm, R., and Son, Y. 2021. Image quality and diagnostic value of diffusion-weighted breast magnetic resonance imaging: Comparison of acquired and computed images. PLoS One. 2021 Feb 22;16(2):e0247379. https://doi.org/10.1371/journal.pone.0247379
[29] Kim, S.Y., Soo Lee, S., Bumwoo Park, Kim, N., Kim, J.K., Park, S.H., Byun, J.H., Song, K.J., Koo, J.H., Kyung Choi, E., Lee, M.G., 2012. Reproducibility of measurement of apparent diffusion coefficients of malignant hepatic tumors: Effect of DWI techniques and calculation methods. J. Magn. Reson. Imaging 36, 1131–1138. https://doi.org/10.1002/jmri.23744
[30] Wong, S.M., Backes, W.H., Zhang, C.E., Staals, J., Van Oostenbrugge, R.J., Jeukens, C.R.L.P.N., Jansen, J.F.A., 2018. On the Reproducibility of Inversion Recovery Intravoxel Incoherent Motion Imaging in Cerebrovascular Disease. AJNR. Am. J. Neuroradiol. 39, 226–231. https://doi.org/10.3174/AJNR.A5474
[31] Perucho, J.A.U., Chang, H.C.C., Vardhanabhuti, V., Wang, M., Becker, A.S., Wurnig, M.C., Lee, E.Y.P., 2020. B-value optimization in the estimation of intravoxel incoherent motion parameters in patients with cervical cancer. Korean J. Radiol. 21, 218–227. https://doi.org/10.3348/kjr.2019.0232
[32] Wurnig, M.C., Donati, O.F., Ulbrich, E., Filli, L., Kenkel, D., Thoeny, H.C., Boss, A., 2015. Systematic analysis of the intravoxel incoherent motion threshold separating perfusion and diffusion effects: Proposal of a standardized algorithm. Magn. Reson. Med. 74, 1414–1422. https://doi.org/10.1002/mrm.25506