Received: 13 April 2021 | Revised: 30 August 2021 | Accepted: 30 August 2021

DOI: 10.1002/mrm.29014

RESEARCH ARTICLE

Magnetic Resonance in Medicine

Training data distribution significantly impacts the estimation of tissue microstructure with machine learning

Noemi G. Gyori[1,2] | **Marco Palombo**[1] | **Christopher A. Clark**[2] | **Hui Zhang**[1] | **Daniel C. Alexander**[1]

[1]Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK

[2]Great Ormond Street Institute of Child Health, University College London, London, UK

Correspondence
Noemi G. Gyori, Centre for Medical Image Computing, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: noemi.gyori.17@ucl.ac.uk

Funding information
Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/M009513/1; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/M020533/1 and EP/N018702/1; UK Research and Innovation, Grant/Award Number: MR/T020296/1; NIHR GOSH Biomedical Research Centre; NIHR UCLH Biomedical Research Centre

Purpose: Supervised machine learning (ML) provides a compelling alternative to traditional model fitting for parameter mapping in quantitative MRI. The aim of this work is to demonstrate and quantify the effect of different training data distributions on the accuracy and precision of parameter estimates when supervised ML is used for fitting.

Methods: We fit a two- and three-compartment biophysical model to diffusion measurements from in-vivo human brain, as well as simulated diffusion data, using both traditional model fitting and supervised ML. For supervised ML, we train several artificial neural networks, as well as random forest regressors, on different distributions of ground truth parameters. We compare the accuracy and precision of parameter estimates obtained from the different estimation approaches using synthetic test data.

Results: When the distribution of parameter combinations in the training set matches those observed in healthy human data sets, we observe high precision, but inaccurate estimates for atypical parameter combinations. In contrast, when training data is sampled uniformly from the entire plausible parameter space, estimates tend to be more accurate for atypical parameter combinations but may have lower precision for typical parameter combinations.

Conclusion: This work highlights that estimation of model parameters using supervised ML depends strongly on the training-set distribution. We show that high precision obtained using ML may mask strong bias, and visual assessment of the parameter maps is not sufficient for evaluating the quality of the estimates.

KEYWORDS
machine learning, microstructure imaging, model fitting, quantitative MRI, training data distribution

Abbreviations: CSF, cerebrospinal fluid; dMRI, diffusion MRI; GM, grey matter; ML, machine learning; qMRI, quantitative MRI; ROI, region of interest; SMT, spherical mean technique; SNR, signal-to-noise ratio; WM, white matter.

1 | INTRODUCTION

Quantitative MRI (qMRI) aims to quantify inherent tissue properties, such as T1- and T2-relaxation times, proton density, magnetization transfer, susceptibility and diffusivity. Quantifying physical tissue features has many potential benefits, such as ease of interpretation, reproducibility across imaging setup, and straightforward comparisons between measurements made at different times or across different populations.[1] However, to quantify the tissue features of interest, it is necessary to define a model linking those features to the measured MRI signal and fit it to appropriately collected data. For example, in diffusion MRI (dMRI), a rich arsenal of biophysical models, signal representations and acquisition strategies have been proposed to quantify several tissue properties, such as mean diffusivity, microscopic anisotropy, neurite density, and dispersion.[2,3] One of the key challenges in qMRI is estimating tissue features accurately, precisely and in a reproducible way, given a model and MRI data.

Conventionally, model fitting is performed voxel-by-voxel using optimization techniques, often based on minimizing a non-linear objective function. However, as models become more complex, conventional fitting approaches become slow and prone to local minima, and the estimation performance degrades with decreasing amount of available data and signal-to-noise ratio (SNR). These drawbacks can hamper the widespread use of qMRI in clinically relevant applications.

Recently, machine learning (ML) has emerged as a promising tool for improving model fitting for qMRI. For example, ML methods based on artificial neural networks have been used to reduce estimation time of myelin water fraction in the brain[4] and to estimate T1 and T2 in a fast and robust way using sparse data from magnetic resonance fingerprinting[5]; whereas ML methods based on convolutional neural network approaches have been developed to estimate susceptibility using a single subject orientation.[6] In dMRI, ML has been used, for example, to bridge the gap between data-hungry imaging techniques and clinically feasible scans, for example by reconstructing super-resolved maps from low spatial resolution data,[7,8] or by estimating advanced diffusion-based metrics from sparse q-space acquisitions.[9–11]

Most ML methods used in qMRI are "supervised," i.e., rely on learning patterns from large training data sets of known corresponding inputs and outputs. A key issue with supervised ML is that in the absence of balanced training data, ML models may learn biased mappings. There are compelling examples of this in healthcare technology, where racial[12] and gender[13] biases arise from imbalances in training data. The performance of supervised ML tools is therefore only as good as the data used to train them.

Recent works that leverage supervised ML for model parameter estimation in qMRI typically employ one of two training data distributions: (1) parameter combinations obtained from traditional model fitting and the corresponding measured qMRI signals,[4,6,9,11,14–17] or (2) parameters sampled uniformly from the entire plausible parameter space with simulated qMRI signals.[5,18–24] While (1) uses parameter combinations directly estimated from the data so likely quantifies the model parameters with higher accuracy and precision for a given specific dataset, (2) supports choice of training data distribution, which may help improve generalizability and avoid problems arising from imbalance.

Here, extending preliminary results in Ref. 25, we focus on dMRI as an exemplar case representing wider model-based qMRI and investigate the effect of training data distribution on microstructural parameter estimates. To this end, we quantify bias and variance in estimates throughout the parameter space of two simple dMRI models where the complexity of the estimation task and the dimensionality of the parameter space are low. Specifically, we use a simple two-compartment model based on the spherical mean technique (SMT),[26,27] which has only two independent parameters, and a three-compartment extension of SMT that includes a free water compartment and has three independent parameters. We estimate the parameters of these models using both traditional non-linear optimization and supervised ML trained on different distributions of ground truth parameters. We visualize how bias and variance manifest throughout the parameter space, and how regions of high and low estimation performance depend on the distribution and noise level of the training data.

2 | METHODS

2.1 | Data acquisition and pre-processing

After informed written consent, six healthy volunteers (ages = 26.3 ± 1.5 y) were scanned on a 3T Siemens Prisma scanner using a 64-channel head coil. Ethical approval for the study was obtained from the UCL Research Ethics Committee. We acquired diffusion weighted images with b-values of [1000, 2000, 3500, 5000] s/mm^2 and a total of 128 uniformly distributed gradient directions[28] with 32 gradient directions per b-value. We acquired 13 b0 images with no diffusion weighting, including one b0 image with reversed phase encoding. Measurement parameters include isotropic 2 mm resolution with acquisition matrix 128 × 128 × 70, partial Fourier imaging 0.75, TE = 94 ms, TR = 9.2 s and GRAPPA parallel imaging with acceleration factor 2. The SNR of the diffusion images was approximately 25 based on the b0 images and averaged across white and gray matter. Additionally, a 3D T1-weighted MPRAGE with 1 mm isotropic resolution

was acquired and segmented using FreeSurfer[29] to identify white and gray matter regions in the brain.

To pre-process the diffusion data, we first removed Gibbs ringing artefacts.[30] Using the FSL toolbox,[31] we estimated the susceptibility-induced off-resonance field with two b0 images with reversed phase encoding polarities[32] and corrected for susceptibility and eddy-current induced geometric distortions and subject motion.[33] We created a binary mask to remove non-brain regions.[34] Finally, we estimated the noise level in the diffusion data[35] and corrected for Rician noise bias.[36]

2.2 | Biophysical models

2.2.1 | 2-compartment SMT

We use the two-compartment SMT model[26,27] as a convenient example that consists of only two independent parameters, which makes visualization of the parameter space straightforward. This model assumes brain tissue consists of heterogeneously oriented cylindrical compartments and the surrounding extra-cellular volume giving normalized signal

$$\frac{\overline{S}(b)}{S_0} = v_{\text{cyl}} \frac{\sqrt{\pi}\,\text{erf}\left(\sqrt{b\lambda_{\text{cyl}}}\right)}{2\sqrt{b\lambda_{\text{cyl}}}} + v_{\text{ext}} \frac{\sqrt{\pi}\,\text{erf}\left(\sqrt{b(\lambda^{\parallel}_{\text{ext}} - \lambda^{\perp}_{\text{ext}})}\right)}{2\sqrt{b(\lambda^{\parallel}_{\text{ext}} - \lambda^{\perp}_{\text{ext}})}} \exp\left(-b\lambda^{\perp}_{\text{ext}}\right) \tag{1}$$

where erf is the error function such that $\lim_{x\to 0} \text{erf}(x)/x = 2/\pi$, $\overline{S}$ is the direction-averaged diffusion signal at a specific b-value (b), S_0 is the signal with no diffusion weighting, v_{cyl} and v_{ext} are the cylindrical and extra-cellular volume fractions, respectively, λ_{cyl} is the diffusivity parallel to cylindrical compartments, and $\lambda^{\parallel}_{\text{ext}}$ and $\lambda^{\perp}_{\text{ext}}$ are the parallel and perpendicular extra-cellular diffusivities, respectively. The model assumes that within cylindrical compartments, perpendicular diffusivity is negligible, i.e., $\lambda^{\perp}_{\text{cyl}} = 0$, that $v_{\text{cyl}} + v_{\text{ext}} = 1$, and that the extra-cellular diffusivities may be approximated by a tortuosity approximation,[37] whereby $\lambda^{\parallel}_{\text{ext}} = \lambda_{\text{cyl}}$ and $\lambda^{\perp}_{\text{ext}} = (1 - v_{\text{cyl}})\,\lambda_{\text{cyl}}$. Thus, the model has two independent parameters: v_{cyl} and λ_{cyl}.

2.2.2 | Three-compartment SMT

To investigate a more complex estimation task, we extend the SMT model to include a free water compartment representing cerebrospinal fluid (CSF):

$$\frac{\overline{S}(b)}{S_0} = v_{\text{cyl}} \frac{\sqrt{\pi}\,\text{erf}\left(\sqrt{b\lambda_{\text{cyl}}}\right)}{2\sqrt{b\lambda_{\text{cyl}}}} + v_{\text{ext}} \frac{\sqrt{\pi}\,\text{erf}\left(\sqrt{b(\lambda^{\parallel}_{\text{ext}} - \lambda^{\perp}_{\text{ext}})}\right)}{2\sqrt{b(\lambda^{\parallel}_{\text{ext}} - \lambda^{\perp}_{\text{ext}})}} \exp\left(-b\lambda^{\perp}_{\text{ext}}\right) + v_{\text{csf}} \exp\left(-b\lambda_{\text{free}}\right) \tag{2}$$

where v_{csf} is the volume fraction of CSF and λ_{free} is the diffusivity of free water, which is approximately 3 μm^2/ms at body temperature. The model has three compartments, which satisfy $v_{\text{cyl}} + v_{\text{ext}} + v_{\text{csf}} = 1$. As λ_{free} is fixed and $\lambda^{\parallel}_{\text{ext}}$, $\lambda^{\perp}_{\text{ext}}$ are estimated as in the SMT model above, this model has three independent parameters, v_{cyl}, v_{csf} and λ_{cyl}. We refer to this three-compartment model as 3-SMT and refer to the two-compartment model in Section 2.2.1. as 2-SMT.

2.3 | Parameter estimation

We estimate the parameters of the biophysical models using two methods: (1) traditional model fitting via nonlinear least squares optimization (software available at https://github.com/ekaden/smt) and (2) supervised ML consisting of artificial neural networks implemented using TensorFlow 2.0 (https://www.tensorflow.org) and random forest regressors implemented in Scikit-learn.[38] The following subsections detail the properties of the ML models, for which code is available upon request, and the training data.

2.3.1 | Artificial neural network architecture

The inputs to the artificial neural networks are the direction-averaged and T2-normalized diffusion signals for the four b-values used: $[\overline{S}(b=1000)/S_0, \overline{S}(b=2000)/S_0, \overline{S}(b=3500)/S_0, \overline{S}(b=5000)/S_0]$. The networks consist of fully connected layers with rectified linear unit (ReLU) activation functions. We include three fully connected layers (two hidden layers, output layer) for the artificial neural networks trained with noise and nine layers (eight hidden layers, output layer) for the artificial neural networks trained without noise (ie, infinite SNR), as more learning capacity is needed to map parameters to noise-free data, as we demonstrate in Supporting Information Figure S1, which is available online. Each hidden layer contains 280 nodes to ensure that the model fitting is as good as possible and the comparison to traditional model fitting is fair. For training, we use a stochastic gradient descent optimiser with learning rate = 0.001, momentum

0.9 and the mean squared error loss between the predicted and ground truth model parameter values. We used two batches with 2^{18} samples per batch. To facilitate fair comparison between the different neural networks, each network was trained over 100 000 epochs, for which the neural networks show good convergence (Supporting Information Figure S2). The training time on a GPU was approximately 7 h for the three-layer networks and approximately 16 h for the nine-layer networks. Estimation performance is stable for different network initializations, as shown in Supporting Information Figure S3.

To train the neural networks, we simulated the direction-averaged and T2-normalized diffusion signal, $\overline{S}/S_0$, using Equation (1) or (2) for each *b*-value used in this work. Equations (1) and (2) provide one signal per *b*-value, whereas the in-vivo data has 32, one for each gradient direction. Here, we set all 32 measurements in the same b-shell to the same value. We then added noise from a Gaussian distribution with a fixed standard deviation (SD) corresponding to a specific SNR and computed the average of the noised signals for each *b*-value. We implemented noise addition and direction-averaging as pre-processing layers in the neural network, as this ensures that a different instance of Gaussian noise is added at each epoch, which avoids overfitting to the noise. We trained neural networks with three different noise levels corresponding to SNR = [5, 25, ∞].

For the 2-SMT model, the neural network outputs are $\text{logit}(v_{\text{cyl}})$ and $\text{logit}(\lambda_{\text{cyl}}/\lambda_{\text{free}})$, where $\text{logit}(x) = \log(x) - \log(1-x)$ and λ_{free} is the diffusivity of free water, set to 3 μm^2/ms. The form of the outputs ensures that the parameter estimates lie within a biophysically plausible range, such that $0 \le v_{\text{cyl}} \le 1$ and $0 \le \lambda_{\text{cyl}} \le \lambda_{\text{free}}$. For the 3-SMT model, the network outputs are $\text{logit}(v_{\text{cyl}} + v_{\text{csf}})$, $\text{logit}(v_{\text{cyl}}/(v_{\text{cyl}} + v_{\text{csf}}))$ and $\text{logit}(\lambda_{\text{cyl}}/\lambda_{\text{free}})$. This also ensures that $v_{\text{cyl}} + v_{\text{csf}} \le 1$.

2.3.2 | Random forest regressor

We used the random forest regressors implemented in Scikit-learn[38] with 200 trees and a maximum tree depth of 20, similarly to previous works.[18,21] We added noise to the training data and computed the direction-average explicitly before training each random forest regressor. The inputs are the direction-averaged, T2-normalized signals, $[\overline{S}(b=1000)/S_0,\ \overline{S}(b=2000)/S_0,\ \overline{S}(b=3500)/S_0,\ \overline{S}(b=5000)/S_0]$, whereas the outputs are $\text{logit}(v_{\text{cyl}})$ and $\text{logit}(\lambda_{\text{cyl}}/\lambda_{\text{free}})$ for the 2-SMT model and $\text{logit}(v_{\text{cyl}} + v_{\text{csf}})$, $\text{logit}(v_{\text{cyl}}/(v_{\text{cyl}} + v_{\text{csf}}))$ and $\text{logit}(\lambda_{\text{cyl}}/\lambda_{\text{free}})$ for the 3-SMT model.

2.3.3 | Training data distributions

The ML models were trained on synthetic data simulated using the same set of *b*-values as in the in-vivo data described in Section 2.1. For the 2-SMT model, 2^{19} parameter combinations were drawn from the parameter space bounded by $0 \le v_{\text{cyl}} \le 1$ and $0 \le \lambda_{\text{cyl}} \le 3$ μm^2/ms, of which 75% were used for training and 25% for validation. We drew samples from the following distributions for training:

(i) *Uniform distribution:* v_{cyl} drawn uniformly between [0, 1], and λ_{cyl} drawn uniformly between [0, 3] μm^2/ms. This distribution corresponds to one of the two approaches used in recent works that estimate tissue microstructure with supervised ML.

(ii) *Healthy brain distribution:* v_{cyl} and λ_{cyl} sampled using parameter combinations obtained from traditional model fitting in five healthy subjects. We fit each of the five healthy data sets with traditional model fitting and pooled the resulting parameter combinations. The total number of parameter combinations was approximately 135 000, fewer than the 2^{19} training data samples used in this work. To ensure that there were sufficient unique samples for training, we sampled proportionally to the density of the pooled parameter combinations. First, we computed the 2D histogram of available parameter combinations using 500 bins in both dimensions and used cubic interpolation to approximate the continuous density function $d(v_{\text{cyl}}, \lambda_{\text{cyl}})$ throughout the $v_{\text{cyl}} - \lambda_{\text{cyl}}$ parameter space. We then performed rejection sampling by selecting a random sample d' between the minimum and maximum of the density, as well as a random parameter combination v_{cyl}' and λ_{cyl}'. We computed $d(v_{\text{cyl}}', \lambda_{\text{cyl}}')$, and if $d' < d(v_{\text{cyl}}', \lambda_{\text{cyl}}')$, the parameter combination was accepted, otherwise it was rejected. This distribution is an approximation of the second approach used in recent works, whereby ML models are trained on parameter combinations estimated via traditional model fitting and the corresponding measured signals. We make one necessary change which is to simulate the diffusion signals using Equation (1) instead of using the measured signals. This allows for increased flexibility in injecting noise into the training data.

(iii) *Mixed uniform and healthy brain distribution:* half the samples drawn from (i) and half drawn from (ii).

To investigate extreme cases where we train on only white or gray matter parameter combinations, we test two further training data distributions:

(iv) *Healthy WM distribution:* v_{cyl} and λ_{cyl} sampled similarly as in (ii), but for white matter (WM) voxels only, determined from the FreeSurfer[29] segmentations.
(v) *Healthy GM distribution:* v_{cyl} and λ_{cyl} sampled similarly as in (ii), but for gray matter (GM) voxels only, determined from the FreeSurfer[29] segmentations.

For the 3-SMT model, we drew 2^{19} samples (of which 75% were used for training and 25% for validation) from the plausible parameter space of this model, given by $0 \leq \lambda_{cyl} \leq 3\ \mu m^2/ms$, $0 \leq v_{cyl} \leq 1$ and $0 \leq v_{csf} \leq 1$, such that $v_{cyl} + v_{csf} \leq 1$. For this model, we drew samples from two different distributions:

(i) *Uniform distribution:* λ_{cyl} was drawn uniformly between [0, 3] $\mu m^2/ms$, whereas v_{cyl} and v_{csf} were drawn uniformly on a two-simplex using a Dirichlet distribution.
(ii) *Healthy brain distribution:* $v_{cyl,}$ v_{csf} and λ_{cyl} sampled using parameter combinations obtained from traditional model fitting in five healthy adult subjects, sampled similarly as the healthy brain distribution in the 2-SMT model.

Table 1 summarizes the ML estimators trained in this work and the names we use to refer to each estimator.

2.4 | Test data

We tested the impact of the training strategy on four sets of test data. First, we use an in-vivo brain scan to compare estimates from traditional model fitting and ML to test whether estimates are impacted by training data distribution. Second, to unpick estimation performance from the different estimators at different parameter combinations, we map bias and variance in estimates across the entire parameter space using synthetic data. Third, we test performance on various example abnormal parameter combinations, motivated by potential pathological scenarios, to probe estimation accuracy for examples that are not well-represented in the training data. Finally, we simulate a lesion in a brain-like data set to investigate whether small abnormalities can be visually detected with the different estimators. We outline the data used for these four test cases in the following subsections.

2.4.1 | In-vivo test data

We used the diffusion measurements of the sixth healthy volunteer that was not included in the training parameter pool used in distributions (ii)–(v) described in Section 2.3.3 as a test set. The SNR of this data set was approximately 25, and the images were pre-processed as described in Section 2.1.

2.4.2 | Simulated data for different parameter combinations

We synthesized test data using Equations (1) and (2) for the 2-SMT and 3-SMT models, respectively, using the same set of *b*-values as in the in-vivo data described in Section 2.1. For the 2-SMT model, we chose 441 points on a 21×21 grid covering the parameter space, such that v_{cyl} ranged from 0 to 1 at increments of 0.05, and λ_{cyl} ranged from 0 to 3 $\mu m^2/ms$ at increments of 0.15 $\mu m^2/ms$. For each point on this grid, we synthesized 10 000 samples of the diffusion signals and added Gaussian noise. We created three such data sets with SNR = [5, 25, ∞]. For each test set we used ML models trained with the corresponding noise level to estimate parameters. For the 3-SMT model, we chose a total of 1617 points, such that v_{cyl} and v_{csf} ranged from 0 to 1 at increments of 0.05 with $v_{cyl} + v_{csf} \leq 1$, and λ_{cyl} ranged from 0.5 to 3 $\mu m^2/ms$ at increments of 0.5 $\mu m^2/ms$. We synthesized 10 000 samples of the diffusion signals and added Gaussian noise with SNR = 25 for each test point.

2.4.3 | Simulated abnormal parameter combinations

For the 2-SMT model, we synthesized the signals for four further parameter combinations representing different types of tissue abnormalities (Table 2). Abnormality 1 is an extreme example where λ_{cyl} is very low, such as we might expect when macromolecules accumulate in tissue. Abnormality 2 is an example where v_{cyl} is very low, such as in extreme cases of chronic black holes in multiple sclerosis.[39] Abnormality 3 has slightly lower v_{cyl} and λ_{cyl} than average WM in our data set (ages = 26.3 ± 1.5 y), similar to normal white matter reported in a cohort of older subjects (ages = 41.7 ± 10 y).[39] Abnormality 4 also has lower v_{cyl} and λ_{cyl} than average WM and exemplifies typical white matter lesions in multiple sclerosis reported in Ref. 40. We highlight that these abnormalities are not exact representations of specific pathologies, which vary widely, but serve to demonstrate estimation performance in abnormal tissue configurations likely to arise in practice. For each combination, we synthesized 10 000 samples and added Gaussian noise corresponding to SNR = [5, 25].

TABLE 1 Summary of the ML models trained in this work indicating whether we used the artificial neural network or the random forest regressor, the training data distribution and noise levels used in each trained model

Estimator name	ML model	Training data distribution	SNR of training data	Model
Net-uniform-SNRINF	Artificial neural network	Uniform distribution	∞	2-SMT
Net-uniform-SNR25	Artificial neural network	Uniform distribution	25	2-SMT
Net-uniform-SNR5	Artificial neural network	Uniform distribution	5	2-SMT
Net-healthy-brain-SNRINF	Artificial neural network	Healthy brain distribution	∞	2-SMT
Net-healthy-brain-SNR25	Artificial neural network	Healthy brain distribution	25	2-SMT
Net-healthy-brain-SNR5	Artificial neural network	Healthy brain distribution	5	2-SMT
Net-healthy-WM-SNR25	Artificial neural network	Healthy WM distribution	25	2-SMT
Net-healthy-GM-SNR25	Artificial neural network	Healthy GM distribution	25	2-SMT
Net-mixed-SNR25	Artificial neural network	Mixed uniform and healthy brain distribution	25	2-SMT
Net-mixed-SNR5	Artificial neural network	Mixed uniform and healthy brain distribution	5	2-SMT
RF-uniform-SNRINF	Random forest regressor	Uniform distribution	∞	2-SMT
RF-uniform-SNR25	Random forest regressor	Uniform distribution	25	2-SMT
RF-uniform-SNR5	Random forest regressor	Uniform distribution	5	2-SMT
RF-healthy-brain-SNRINF	Random forest regressor	Healthy brain distribution	∞	2-SMT
RF-healthy-brain-SNR25	Random forest regressor	Healthy brain distribution	25	2-SMT
RF-healthy-brain-SNR5	Random forest regressor	Healthy brain distribution	5	2-SMT
RF-mixed-SNR25	Random forest regressor	Mixed uniform and healthy brain distribution	25	2-SMT
Net-3SMT-uniform-SNR25	Artificial neural network	Uniform distribution	25	3-SMT
Net-3SMT-healthy-brain-SNR25	Artificial neural network	Healthy brain distribution	25	3-SMT
RF-3SMT-uniform-SNR25	Random forest regressor	Uniform distribution	25	3-SMT
RF-3SMT-healthy-brain-SNR25	Random forest regressor	Healthy brain distribution	25	3-SMT

Abbreviations: GM, gray matter; WM, white matter.

TABLE 2 Specific parameter combinations chosen to illustrate performance in abnormal parameter combinations for the 2-SMT model

Parameter combination name	v_{cyl}	λ_{cyl} (μm^2/ms)
Abnormality 1	0.67	0.5
Abnormality 2	0.05	1.5
Abnormality 3	0.60	1.8
Abnormality 4	0.47	1.9

2.4.4 | Simulated brain data with abnormality

We replaced a region of interest (ROI) in white matter of the 2-SMT parameter maps by the parameter combination of Abnormality 3. We then simulated diffusion signals to create a full synthetic brain-like data set. We added noise to the simulated signals corresponding to SNR = [5, 25].

3 | RESULTS

3.1 | In-vivo parameter maps

In Figure 1, we map in-vivo parameter estimates for the 2-SMT model from traditional model fitting and differences for parameter maps obtained using Net-uniform-SNR25 and Net-healthy-brain-SNR25. The maps demonstrate that different parameters are estimated with each model, suggesting variation in the performance across the different methods. We show similar maps for estimates using Net-mixed-SNR25, Net-healthy-WM-SNR25 and Net-healthy-GM-SNR25 in Supporting Information Figure S4, which demonstrates that when we train only on parameter combinations typical in white matter, estimates in gray matter are substantially different from those obtained from traditional model fitting, and similarly vice versa. We show similar effects in estimates obtained using random forest regressors in Supporting Information Figure S5.

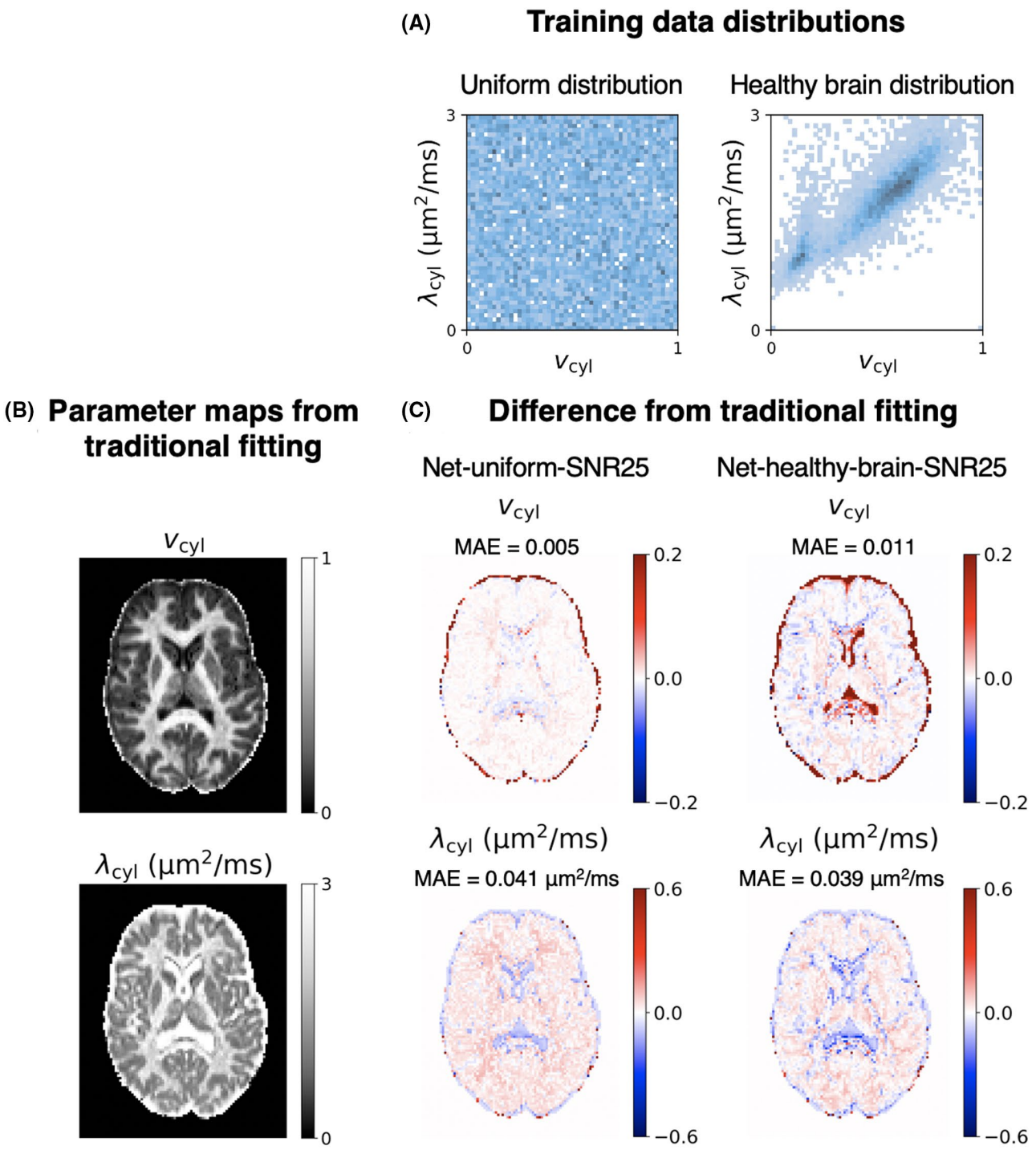

FIGURE 1 (A) Uniform and healthy brain parameter distributions used to train the ML models for the two-compartment SMT model. (B) v_{cyl} and λ_{cyl} parameter maps obtained from traditional model fitting. (C) The difference between parameter maps obtained from the neural networks and traditional model fitting. We indicate the mean absolute error (MAE) between estimates from the neural networks and traditional model fitting for white and gray matter regions

3.2 | Accuracy and precision using synthetic test data

Figure 2 maps bias in parameter estimation for different combinations of v_{cyl} and λ_{cyl} for the 2-SMT model for SNR = [5, 25] using the uniform and healthy brain training distributions. As SNR is reduced, bias in the parameter estimates increases for each estimation method, with traditional model fitting providing the lowest overall bias. Estimates obtained from the neural networks

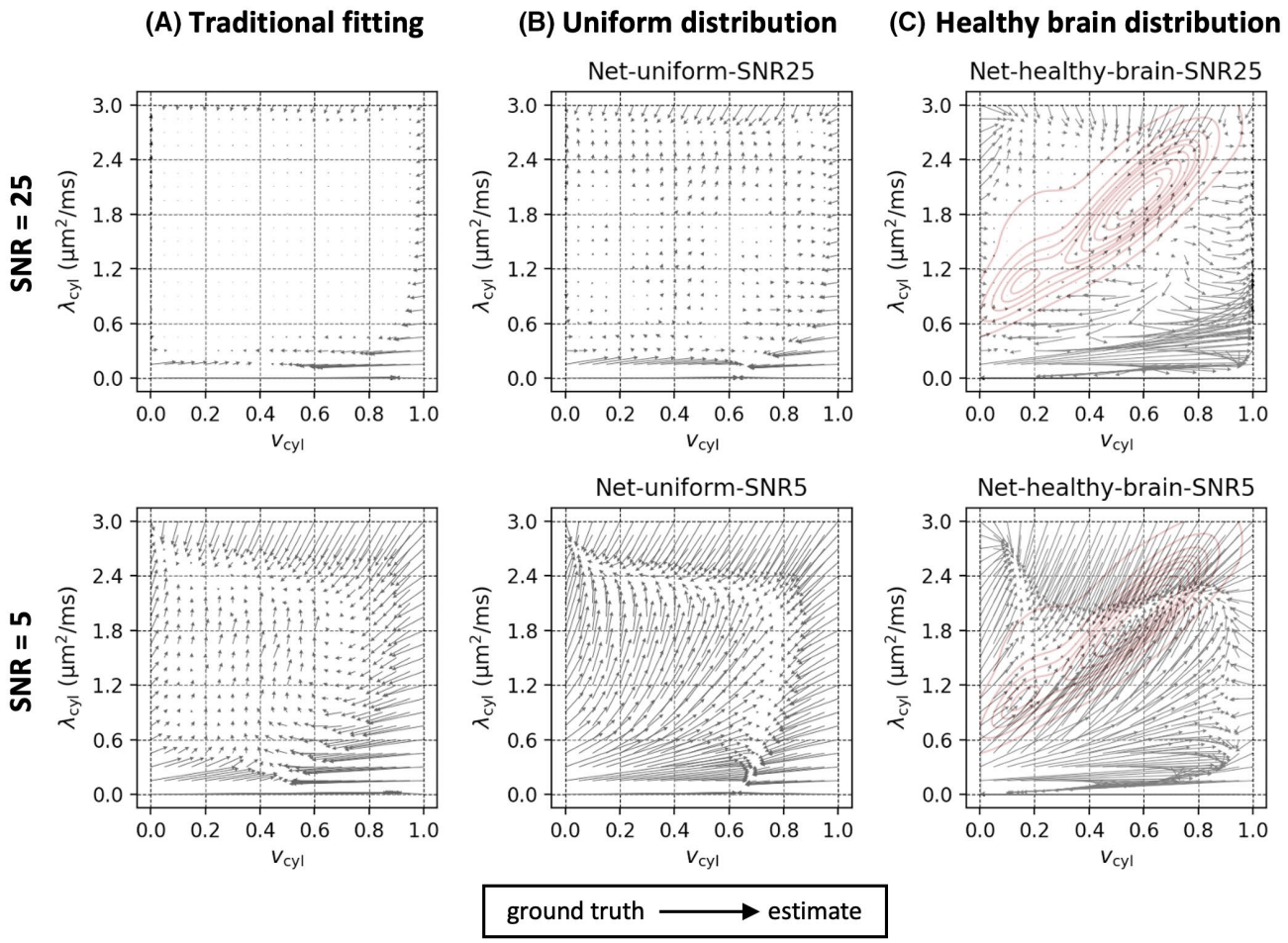

FIGURE 2 Bias mapped using quiver plots for traditional model fitting (A), neural networks trained using the uniform distribution (B), and neural networks trained using the healthy brain distribution (C). The arrows point from the ground truth values to the mean of the estimated values. In column (C), the red contours show the training data density. Each row shows the biases at different values of SNR, according to which Gaussian noise was added to both the training data and test data

trained on the healthy brain distribution have higher overall bias compared to the neural networks trained on the uniform distribution, and bias is consistently high in the low v_{cyl} and high λ_{cyl} region where the training data has low density. Interestingly, certain regions of the parameter space act as 'sinks', towards which estimates of nearby parameters are biased. The location of these sinks depends on both the training data distribution and the noise level. For example, in the networks trained on in-vivo parameter combinations a sink forms near the highest training data density region. For each fitting approach, biases are high when $\lambda_{cyl} = 0$, as the biophysical model is degenerate when there is no diffusion. The pull of the sinks becomes stronger as the SNR is reduced, but interestingly for the healthy brain distribution, sinks appear even when the training and testing data is noise-free (Supporting Information Figure S6). For estimators trained on healthy white and gray matter parameter combinations, even stronger biases manifest (Supporting Information Figure S7). We obtained similar results using random forest regressors (Supporting Information Figure S8).

In Figure 3, we illustrate biases for the 3-SMT model trained and tested on data with SNR = 25 using traditional model fitting, Net-3SMT-uniform-SNR25 and Net-3SMT-healthy-brain-SNR25 across the $v_{cyl} - v_{csf}$ plane for two different values of λ_{cyl} (the full set of tested λ_{cyl} are shown in Supporting Information Figure S9 and S10 for the neural networks and random forest regressors, respectively). Biases in v_{cyl} and v_{csf} are marked with arrows that point from the ground truth to the estimates, whereas the colour of the arrows marks biases in λ_{cyl}. Figure 3 demonstrates that for a more complex model, biases are substantially higher throughout the parameter space, even for data with reasonable SNR. In particular, complex biases manifest for Net-3SMT-healthy-brain-SNR25, where the arrows marking biases in $v_{cyl} - v_{csf}$ may overlap.

Figure 4 shows SDs in the 2-SMT v_{cyl} and λ_{cyl} estimates obtained from traditional model fitting and from the neural networks for SNR = [5, 25] (for SDs for infinite SNR

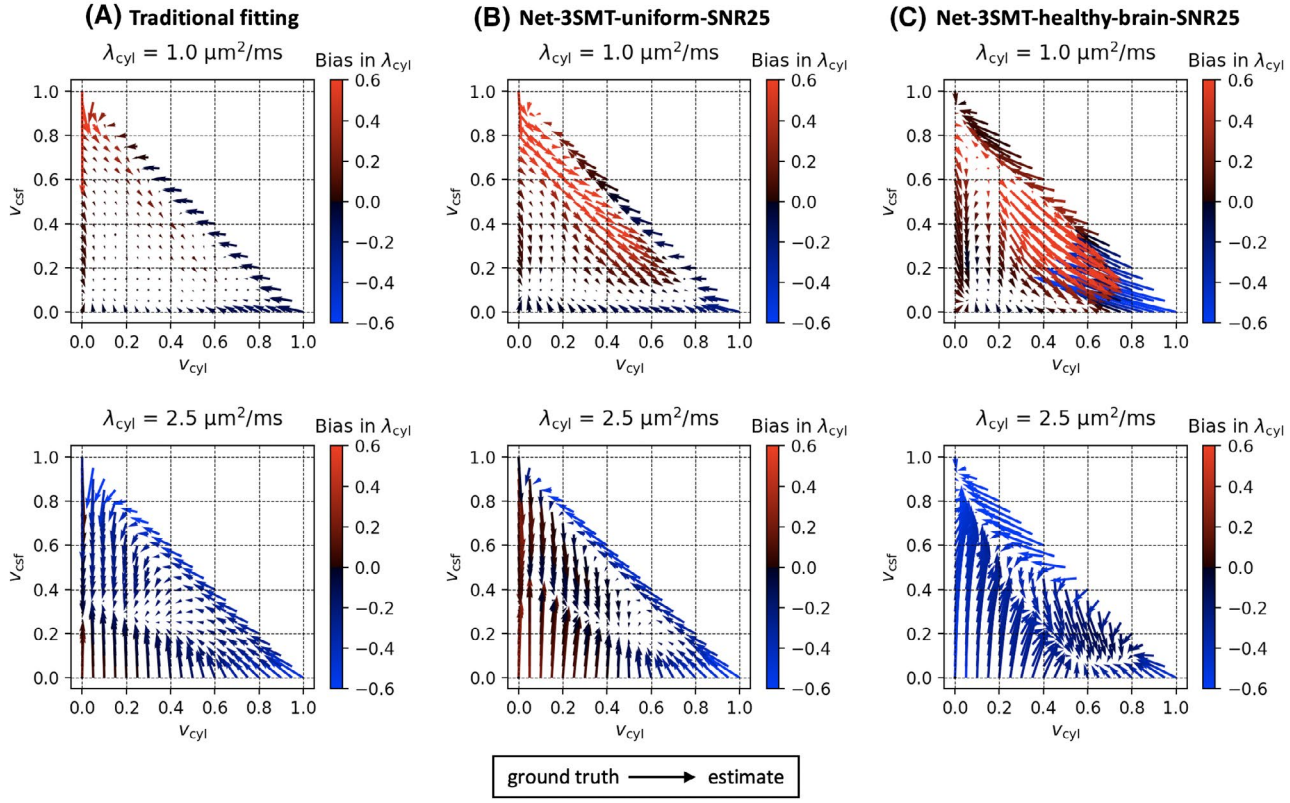

FIGURE 3 Bias mapped using quiver plots for the 3-SMT model. We compare biases for traditional model fitting (A), neural networks trained using the uniform distribution (B), and neural networks trained using the healthy brain distribution (C). Arrows indicate biases in v_{cyl} and v_{csf}, whereas the colours indicate bias in λ_{cyl}. The two rows show biases at two different diffusivities (further diffusivities shown in Figure S8). The SNR in both the training and test data was 25

see Supporting Information Figure S11). Parameters are estimated precisely using all three methods when the training and test data are noise-free. As SNR is reduced, the precision of the parameter estimates from traditional model fitting degrades more than using the artificial neural networks. We obtained similar results using random forest regressors (see Supporting Information Figure S12) and we summarize the overall RMSE, bias, and SD using the different 2-SMT parameter estimation methods when SNR = 25 in Table 3. For the 3-SMT model, Supporting Information Figures S13 (neural networks) and S14 (random forest) reveal that precision is high for the ML estimators, even for the more complex model.

In Figure 5, we probe estimation performance in the 2-SMT model for the specific parameter combinations representing tissue abnormalities. When SNR = 25, Abnormality 1 is estimated inaccurately with the ML models and in particular with Net-healthy-brain-SNR25, and with low precision for the other methods. Abnormalities 2–4 are estimated with high precision with all the methods, but for Abnormality 3, estimates are slightly inaccurate for Net-mixed-SNR25 and Net-healthy-brain-SNR25, whereas for Abnormality 4, estimates are slightly inaccurate for Net-uniform-SNR25. When SNR = 5, all three ML estimators substantially over-estimate λ_{cyl} in Abnormalities 2 and 3. Importantly, in Abnormality 3, this over-estimation pushes parameter estimates closer to typical white matter parameter combinations, making it difficult to distinguish the abnormality from healthy white matter. In Abnormality 4, estimates are more accurate, and strong bias is only apparent in λ_{cyl} as estimated by Net-uniform-SNR5, suggesting that not all abnormalities are equally affected. Parameter estimates obtained from traditional fitting remain remarkably accurate across the abnormalities but have substantially lower precision compared to ML estimates.

Figure 6 highlights visually the potential for high bias and low variance in parameter estimates to obscure lesions such as Abnormality 3. When SNR = 25, estimates are accurate and precise for all estimators. Contrarily, when SNR = 5, estimates from traditional model fitting are noisy throughout the brain, whereas estimates from ML are strongly biased but smooth, particularly for Net-mixed-SNR5 and Net-healthy-brain-SNR5. The lesion is obscured for both traditional fitting and ML at low SNR, but for ML, poor estimation performance is not visually

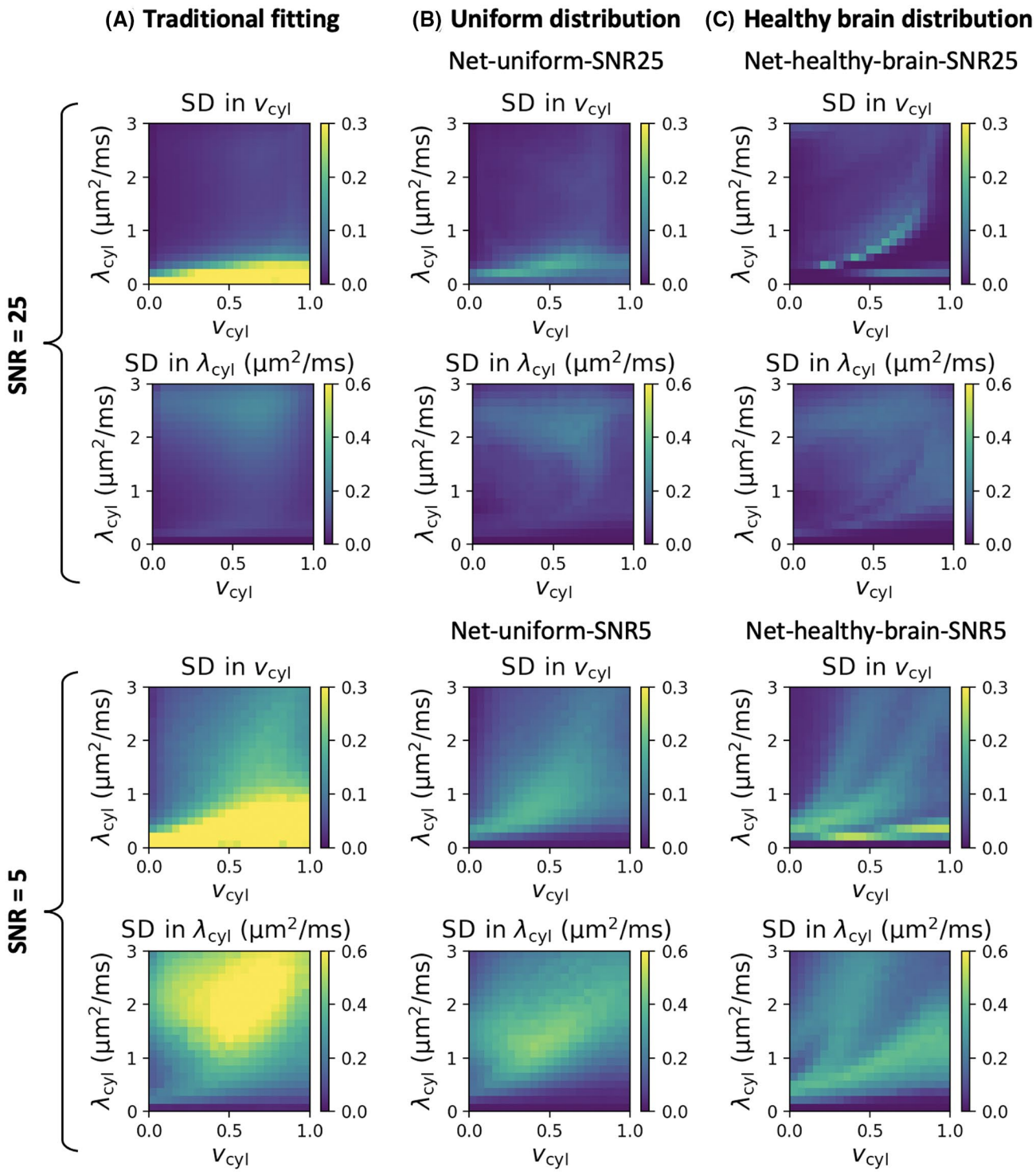

FIGURE 4 Standard deviation (SD) in v_{cyl} and λ_{cyl} estimates using traditional model fitting (A), neural networks trained using the uniform distribution (B), and neural networks trained using the healthy brain distribution (C). SDs are shown for two different noise levels in both the training and test data sets

obvious unlike in traditional model fitting, making the inaccurate parameter estimates appear convincing. We show similar results to Figures 4 and 5 for random forest regressors in Supporting Information Figures S15 and S16, respectively.

4 | DISCUSSION

This work highlights two key properties of supervised ML-based fitting techniques, which differ from traditional model fitting. First, we show that parameter estimates are

TABLE 3 The mean RMSE, bias and standard deviation (SD) in v_{cyl} and λ_{cyl} over the entire parameter space for the 2-compartment SMT estimation methods using SNR = 25

Estimation method	Mean v_{cyl} RMSE	Mean λ_{cyl} RMSE ($\mu m^2/ms$)	Mean v_{cyl} bias	Mean λ_{cyl} bias ($\mu m^2/ms$)	Mean v_{cyl} SD	Mean λ_{cyl} SD ($\mu m^2/ms$)
Traditional fitting	0.0883	0.1106	**0.0313**	**0.0130**	0.0742	0.1087
Net-uniform-SNR25	**0.0642**	**0.1082**	0.0338	0.0386	0.0449	0.0965
Net-healthy-brain-SNR25	0.1191	0.1386	0.1017	0.0724	**0.0324**	0.1030
Net-mixed-SNR25	0.0663	0.1116	0.0367	0.0404	0.0476	0.0975
RF-uniform-SNR25	0.0670	0.1109	0.0342	0.0400	0.0490	0.0980
RF-healthy-brain-SNR25	0.1209	0.1390	0.1008	0.0806	0.0358	**0.0930**
RF-mixed-SNR25	0.0683	0.1114	0.0340	0.0405	0.0529	0.0978

Note: Bold values highlight the lowest value in each column.

significantly affected by the distribution of training data. Second, we demonstrate that smooth parameter maps obtained via ML may be deceptive, as high precision may hide strong biases. This is in contrast with traditional fitting, where low reliability in estimates is typically reflected by noisy parameter maps. Although here we focus on dMRI, we expect similar results for other qMRI techniques that use supervised ML methods for model fitting.

In Section 3.2. we focus on three different training data distributions: healthy parameter combinations obtained using traditional model fitting, uniformly distributed parameter combinations, and healthy parameter combinations augmented with uniformly distributed parameter combinations. Recently, authors in Ref. 41 compared the fitting performance of the first two training strategies, and authors in Ref. 42 assessed the trade-off between accuracy and generalizability when combining them to analyze diffusion-relaxation data. Our results show that training on healthy parameter combinations facilitates precise estimates in healthy tissue but may yield strong biases in atypical parameter combinations not represented in training. This bias is mitigated when healthy data are combined with atypical parameter combinations in training, in line with recent findings in Ref. 42. However, here we show that even when healthy training data is combined with atypical parameter combinations, and in fact even when the full parameter space is uniformly represented in the training data, supervised ML may still introduce substantial biases that can hamper the clinical utility of qMRI techniques.

Parameter estimates obtained from traditional model fitting are overall more accurate than the estimates obtained from the ML models at each noise level tested in this work. However, at low SNR, traditional fitting suffers from high variance, which manifests as noisy appearing parameter maps. Maps obtained using the ML estimators appear less noisy, which may mistakenly convince users that the estimates are reliable even at low SNR. In Figure 6 we show that this apparent improvement can be misleading. Specifically, a small abnormality, linked for example to aging, may be obscured in ML estimates when SNR is low, even for the simple 2-SMT model. This issue is particularly pronounced for ML models trained on healthy parameter combinations, but maps obtained using the uniform distribution may also mislead users. We emphasize that the abnormal parameter combinations we use to show these effects are not exact representations of any particular pathology but are designed to highlight that different bias effects may arise in different types of atypical tissue. We also show that biases may be exacerbated with more complex models such as the 3-SMT model, even though precision remains high with ML fitting.

We show results for two different ML estimators: artificial neural networks and random forest regressors. While we observe similar effects with both ML models, several differences also arise. For example, in regions of the parameter space that are not well represented in the training data, biases tend to be high in both cases, but the direction of the bias may be different. Additionally, in some regions of the parameter space, variance is higher for random forest than for neural network estimates. A possible reason for these discrepancies is the different noise handling in the two ML approaches. In the neural networks, noise is injected at every epoch, and hence each noise instance is unique, whereas for the random forest regressors noise is injected only once.

The analysis and visualization approaches proposed here (Figures 2–6) provide tools to quantify the expected impact of a chosen estimation strategy and to aid the interpretation of resulting parameter estimates. For example, parameter estimates near 'sinks' in the bias quiver plots should be interpreted with caution, as these parameter combinations may mask substantial biases. The location and evolution of these sinks can inform future experimental design and training strategies optimized to mitigate their impact.

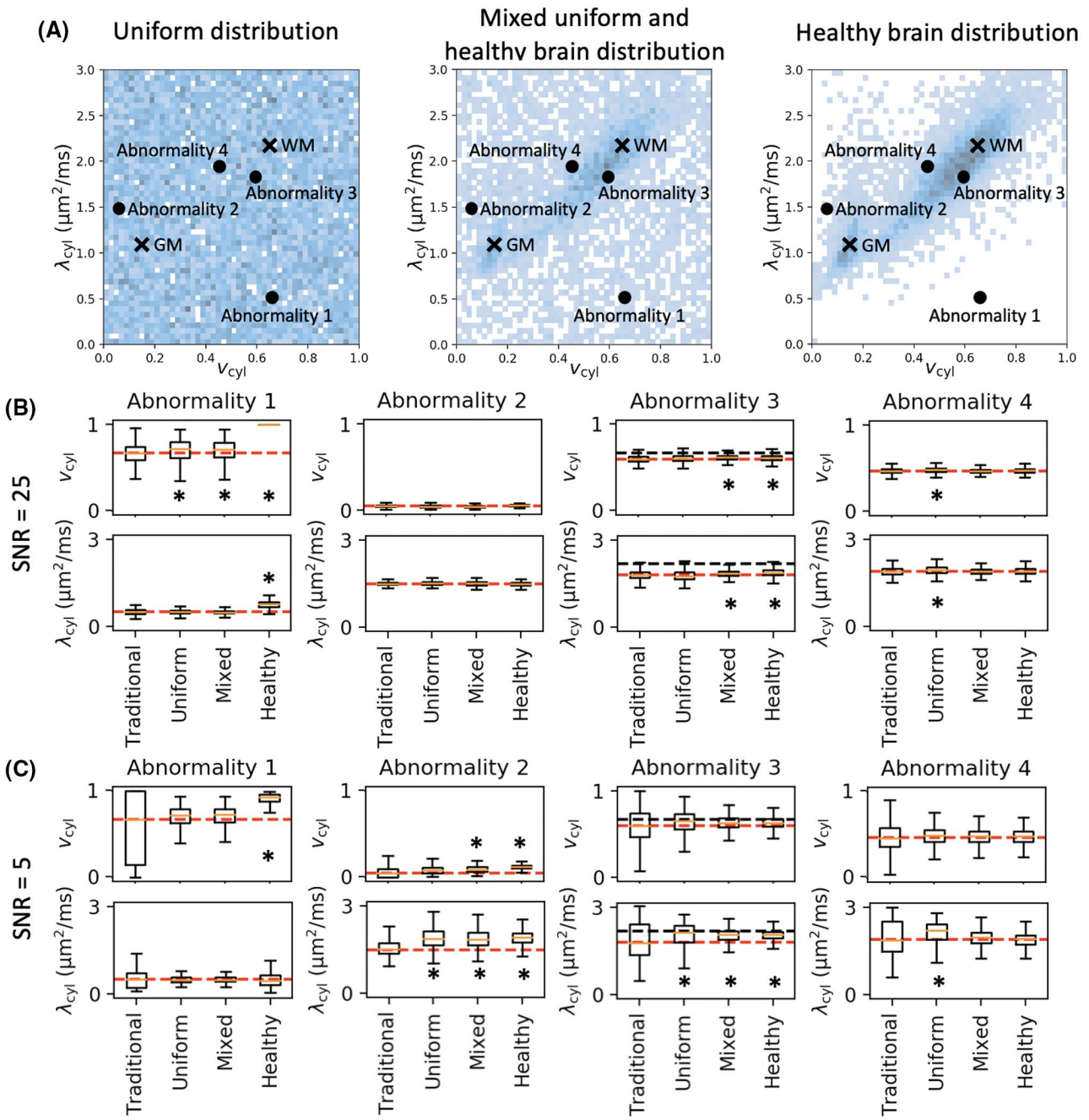

FIGURE 5 (A) Different training data distributions: uniform data distribution, healthy brain distribution, and a mixed distribution where 50% of the samples are from the uniform distribution, and 50% of the samples are from the healthy brain distribution. We mark four atypical parameter combinations: Abnormality 1 with low λ_{cyl} exemplifying accumulation of macromolecules, Abnormality 2 exemplifying extreme cases of chronic black holes, Abnormality 3 exemplifying normal white matter in an older cohort, and Abnormality 4 exemplifying white matter lesions in multiple sclerosis. We also mark typical white matter (WM) and grey matter (GM) parameter combinations for reference. (B,C) Box plots of the estimates for the four using synthetic data with SNR = 25 and with SNR = 5, respectively. The dashed red line marks the ground truth, and for Abnormality 3, the black line marks average WM in our data set. Stars indicate where effect size comparing the distribution of estimates to the distribution of noised ground truth is medium or large, ie, the magnitude of Cohen's $d > 0.5$

Our findings highlight that training strategies for parameter estimation should not be used blindly and suggest that further consideration and development is required. For example, certain applications might be tolerant of bias as long as it is well-characterized as in Figures 2 and 3, but where accuracy is important perhaps the role of supervised ML may simply be to provide close starting points for iterative search that reduces overall computation time. Computing uncertainty in ML-based estimation, cf.,[43] may also help assess estimation reliability, particularly

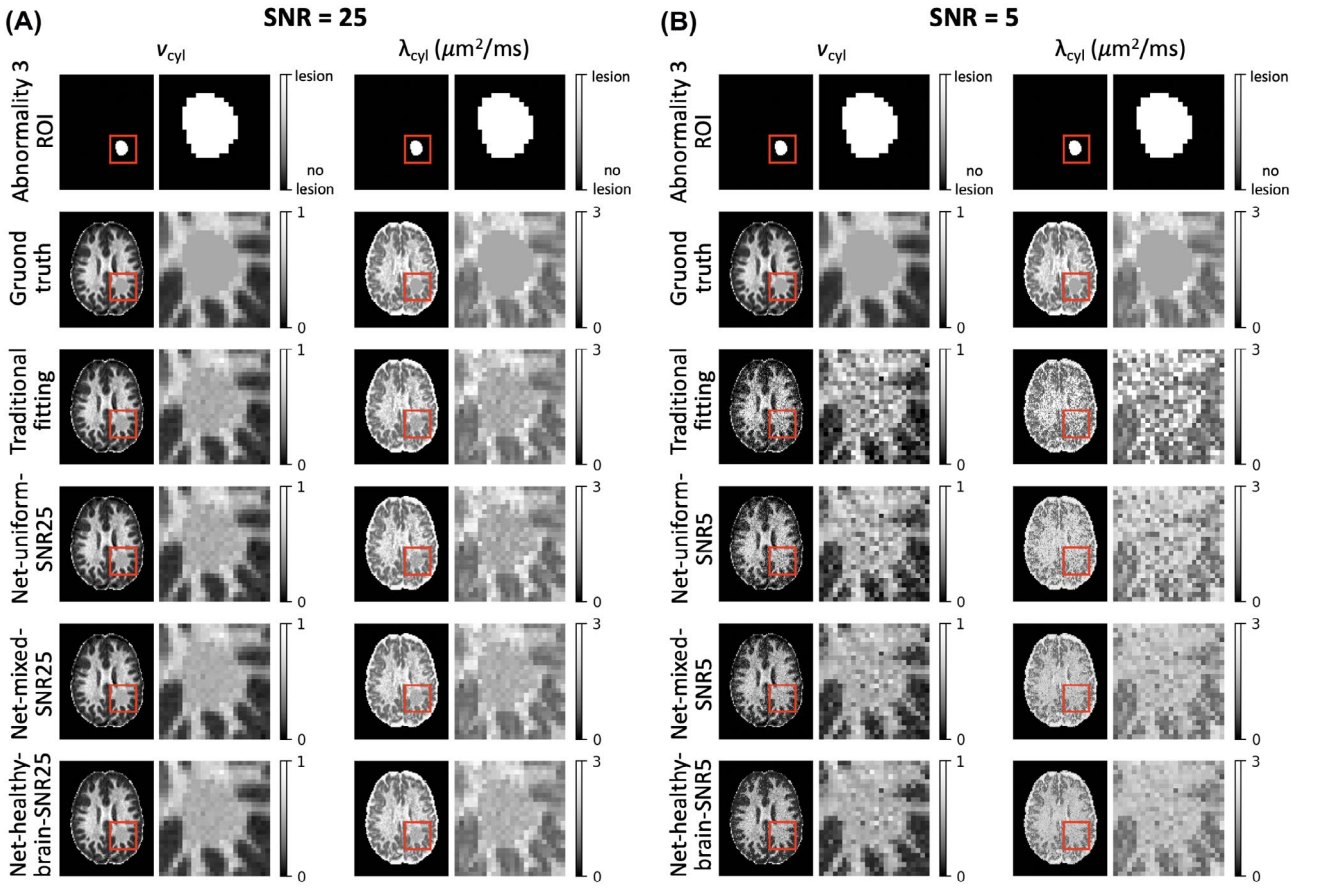

FIGURE 6 Parameter estimates for SNR = 25 (A) and SNR = 5 (B). The data sets used here were simulated using parameter values obtained from traditional fitting with Abnormality 3 applied to an ROI shown in the top row. Abnormality 3 is highlighted in the red box and shown in adjacent zoomed plots

when ML is used to compensate for lower quality data. Training strategies that iteratively augment the training data distributions in regions of high bias may ameliorate the issues we raise. To avoid the impact of training data distribution altogether, unsupervised learning[44] may be used as an alternative to supervised learning, but this may come at the expense of lower estimation accuracy, particularly in the presence of Rician noise.[41] Future work will explore these ideas.

This work presents a limited case study of parameter estimation in dMRI. We focus on two simple models, which represent a broader set of strategies that use direction-averaged signals and, hence, benefit from simpler models that are less dependent on specific acquisition protocols than strategies that use the raw signal. While such strategies are commonly used in the dMRI community, direction-averaging the diffusion signals likely results in loss of information. Future work might investigate whether richer, direction-sensitive data may reduce biases observed in this work. Our analysis was also limited to a single set of *b*-values, and different numbers and combinations of *b*-values would likely affect both the overall accuracy and the position of 'sinks' in the parameter space towards which nearby parameter combinations are biased. We chose simple ML architectures similar ones used in previous works. We investigated how estimation performance depends on, for example, network depth, but a detailed investigation of the impact of architecture and hyperparameter choice on parameter estimation remains future work. We removed the Rician noise floor from the in-vivo data and added Gaussian noise to the synthetic training and test data, as this simple strategy best avoids any concern that the issues we raise are specific to a particular noise profile. Rician noise and other noise-like behavior arising for example from residual motion artefacts likely exacerbate the effects we observe here.

ML is a promising tool for enhancing medical imaging technology, where resources are often limited, and the potential impact may be life changing. qMRI may benefit in particular, as advanced MRI acquisitions and subsequent model fitting may be time-consuming. However, work still needs to be done to mitigate biases and assess estimation reliability in order to use ML effectively.

ACKNOWLEDGMENTS

NGG thanks the London Interdisciplinary Bioscience PhD Consortium and is funded by BBSRC grant BB/M009513/1. MP is supported by UKRI Future Leaders Fellowship (MR/T020296/1). EPSRC grants EP/M020533/1 and EP/N018702/1 and the NIHR UCLH Biomedical Research Centre and the NIHR GOSH Biomedical Research Centre support our work on this topic. We acknowledge Enrico Kaden's background contributions in conceptualization, design and set-up of the diffusion experiment, and initial discussions on studying the effects of training data distribution. We also thank Filip Szczepankiewicz and Markus Nilsson for sharing their diffusion EPI sequence.

DATA AVAILABILITY STATEMENT

Data may be available upon reasonable request.

ORCID

Noemi G. Gyori https://orcid.org/0000-0001-9021-0957
Marco Palombo https://orcid.org/0000-0003-4892-7967

TWITTER

Noemi G. Gyori @GyoriNoemi

REFERENCES

1. Cercignani M, Dowell NG, Tofts P. *Quantitative MRI of the Brain: Principles of Physical Measurement*. CRC Press Taylor & Francis Gruop; 2018.
2. Alexander DC, Dyrby TB, Nilsson M, Zhang H. Imaging brain microstructure with diffusion MRI: practicality and applications. *NMR Biomed*. 2019;32:e3841.
3. Novikov DS, Kiselev VG, Jespersen SN. On modeling. *Magn Reson Med*. 2018;79(6):3172-3193.
4. Liu H, Xiang Q-S, Tam R, et al. Myelin water imaging data analysis in less than one minute. *Neuroimage*. 2020;210:116551.
5. Cohen O, Zhu B, Rosen MS. MR fingerprinting Deep RecOnsrtuction Network (DRONE). *Magn Reson Med*. 2018;80(3):885-894.
6. Yoon J, Gong E, Chatnuntawech I, et al. Quantitative susceptibility mapping using deep neural networks: QSMnet. *Neuroimage*. 2018;179:199-206.
7. Alexander DC, Zikic D, Ghosh A, et al. Image quality transfer and applications in diffusion MRI. *Neuroimage*. 2017;152:283-298.
8. Hong Y, Chen G, Yap P-T, Shen D. Multifold acceleration of diffusion MRI via deep learning reconstruction from slice-undersampled data. *Inf Process Med Imaging*. 2019;11492:530-541.
9. Golkov V, Dosovitskiy A, Sperl JI, et al. q-Space deep learning: twelve-fold shorter and model-free diffusion MRI scans. *IEEE Trans Med Imaging*. 2016;35(5):1344-1351.
10. Tian Q, Bilgic B, Fan Q, et al. DeepDTI: high-fidelity six-direction diffusion tensor imaging using deep learning. *Neuroimage*. 2020;219:117017.
11. Aliotta E, Nourzadeh H, Patel SH. Extracting diffusion tensor fractional anisotropy and mean diffusivity from 3-direction DWI scans using deep learning. *Magn Reson Med*. 2021;85:845-854.
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage health populations. *Science*. 2019;366(6464):447-453.
13. Cirillo D, Catuara-Solarz S, Morey C, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med*. 2020;3:81.
14. Hubertus S, Thomas S, Cho J, Zhang S, Wang Y, Schad LR. Using an artificial neural network for fast mapping of the oxygen extraction fraction with combind QSM and quantitative BOLD. *Magn Reson Med*. 2019;82:2199-2211.
15. Koppers S, Merhof D. Direct estimation of fibre orientations using deep learning in diffusion imaging. *International Workshop on Machine Learning in Medical Imaging*; 2016:53-60. https://doi.org/10.1007/978-3-319-47157-0_7
16. Chen G, Hong Y, Zhang Y, et al. Estimating tissue microsctructure with undersampled diffusion data via graph convolutional neural networks. *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020*; 2020:280-290.
17. Fan H, Su P, Huang J, Liu P, Lu H. Multi-band MR fingerprinting (MRF) ASL imaging usnig artificial-neural-network trained with high-fidelity experimental data. *Magn Reson Med*. 2021;85(4):1974-1985.
18. Nedjati-Gilani GL, Schneider T, Hall MC, et al. Machine learning based compartment models with permeability for white matter microstructure imaging. *Neuroimage*. 2017;150:119-135.
19. Palombo M, Ianus A, Guerreri M, et al. SANDI: a compartment-based model for non-invasive apparent soma and neurite imaging by diffusion MRI. *Neuroimage*. 2020;15:116835.
20. Bollmann S, Rasmussen KGB, Kristensen M, et al. DeepQSM—using deep learning to solve the dipole inversion for quantitative susceptibility mapping. *Neuroimage*. 2019;195:373-383.
21. Hill I, Palombo M, Santin M, et al. Machine learning based white matter models with permeability: an experimental study in cuprizone treated in-vivo mouse model of axonal demyelination. *Neuroimage*. 2020;224:117425.
22. Gyori NG, Clark CA, Alexander DC, Kaden E. In-vivo neural smoa imaging using B-tensor encoding and deep learning. *Neuroimage*. 2021;239:118303.
23. Yu T, Canales-Rodriguez EJ, Pizzolato M, et al. Model-informed machine learning for multi-component T2 relaxometry. *Med Image Anal*. 2021;69:101940.
24. Kim B, Schar M, Park H, Heo H-Y. A deep learning approach for magnetization transfer contrast MR fingerprinting and chemical exchange saturation transfer imaging. *Neuroimage*. 2020;221:117165.
25. Gyori NG, Palombo M, Clark CA, Zhang H, Alexander DC. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Proceedings of the ISMRM*; 2021:400.
26. Kaden E, Kelm ND, Carson RP, Does MD, Alexander DC. Multi-compartment microscopic diffusion imaging. *Neuroimage*. 2016;139:346-359.
27. Kaden E, Kruggel F, Alexander DC. Quantitative mapping of the per-axon diffusion coefficients in brain white matter. *Magn Reson Med*. 2016;75:1752-1763.

28. Caruyer E, Lenglet C, Sapiro G, Deriche R. Design of multishell samplling schemes with uniform coverage in diffusion MRI. *Magn Reson Med*. 2013;69:1524-1540.
29. Fischl B. FreeSurfer. *Neuroimage*. 2012;62:774-781.
30. Kellner E, Dhital B, Kiselev VG, Reisert M. Gibbs-ringing artifact removal based on local subvocel-shifts. *Magn Reson Med*. 2016;76(5):1574-1581.
31. Smith SM, Jenkinson M, Woolrich MW, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*. 2004;23:S208-S219.
32. Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: applications to diffusion tensor imaging. *Neuroimage*. 2003;20:870-888.
33. Andersson JLR, Sotiropoulos SN. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage*. 2016;125:1063-1078.
34. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp*. 2002;17:143-155.
35. Veraart J, Novikov DS, Christiaens D, Ades-aron B, Sijbers A, Fieremans E. Denoising of diffusion MRI using random matrix theory. *Neuroimage*. 2016;142:394-406.
36. Koay CG, Basser PJ. Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *J Magn Reson*. 2006;179:317-322.
37. Szafer A, Zhong J, Gore JC. Theoretical model for water diffusion in tissues. *Magn Reson Med*. 1995;33:697-712.
38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
39. Bagnato F, Franco G, Li H, et al. Probing axons using multi-compartmental diffusion in multiple sclerosis. *Ann Clin Transl Neurol*. 2019;6:1595-1605.
40. Johnson D, Ricciardi A, Brownlee W, et al. Comparison of neurite orientation dispersion and density imaging and two-compartment spherical mean technique parameter maps in multiple sclerosis. *Front Neurol*. 2021;12:662855.
41. Grussu F, Battiston M, Palombo M, Schneider T, Gandini Wheeler-Kingshott CAM, Alexander DC. Deep learning model fitting for diffusion-relaxometry: a comparative study. *bioRxiv*. 2020.
42. de Almeida Martins JP, Nilsson M, Lampinen B, et al. On the use of neural networks to fit high-dimensional microstructure models. *Proceedings of the ISMRM*; 2021:401.
43. Tanno R, Worrall DE, Kaden E, et al. Uncertainty modelling in deep learning for safer neuroimage enhancement: demonstration in diffusion MRI. *Neuroimage*. 2021;225:117366.
44. Barbieri S, Gurney-Champion OJ, Klaassen R, Thoeny HC. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magn Reson Med*. 2020;83:312-321.

SUPPORTING INFORMATION

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

FIGURE S1 RMS errors in v_{cyl} and λ_{cyl} estimates using different numbers of network layers when there is no noise in the training data (top) and when noise is added with SNR = 25 (bottom). In both cases, we used uniformly distributed parameter combinations to train the neural networks

FIGURE S2 Example loss curves for Net-uniform-SNRINF showing that MSE in both the training and validation sets is decreasing, and hence the neural network is not overfitting within 100 000 epochs even when no noise is injected into the training data

FIGURE S3 Biases mapped for estimates obtained from three different versions of Net-uniform-SNR25. In each case the network was trained on the same data, but with different initialisation and batch-shuffling. Estimation performance appears largely stable across the networks

FIGURE S4 Panel (A): mixed, healthy WM and healthy GM parameter distributions for the 2-SMT model. Panel (B): maps from traditional model fitting. Panel (C): Difference between estimates obtained from the neural networks trained on parameter combinations shown in panel (A) and the parameter maps from traditional model fitting in panel (B)

FIGURE S5 Training data distributions for the 2-SMT model (panel A), maps from traditional fitting (panel B) and differences between estimates from random forest regressors trained on the different parameter distributions (panel C)

FIGURE S6 Biases for the 2-SMT model when no noise is added to the training or testing data

FIGURE S7 Biases for the 2-SMT model for mixed, healthy WM and healthy GM distributions at different noise levels. The red contours show the data density for each of the underlying training data distributions

FIGURE S8 Biases mapped for the 2-SMT model obtained from the random forest regressors for all the tested training data distributions

FIGURE S9 Bias maps obtained for the 3-SMT model at all the tested values of λ_{cyl} using traditional model fitting (A) and neural networks trained on uniform (B) and healthy brain (C) data distributions. Noise corresponding to SNR = 25 was added to both the training and test data

FIGURE S10 Bias maps obtained for the 3-SMT model at all the tested values of λ_{cyl} using random forest regressors trained on uniform (A) and healthy brain (B) data distributions. Noise corresponding to SNR = 25 was added to both the training and test data

FIGURE S11 Standard deviation in v_{cyl} and λ_{cyl} estimates of the 2-SMT model for traditional fitting (A) and neural networks trained on uniform (B) and healthy brain (C) distributions when no noise is added to the training or testing data. We highlight the different scale in the colour bar compared to similar plots in Figure 4 for noisy data

FIGURE S12 Standard deviation in v_{cyl} and λ_{cyl} estimates of the 2-SMT model for random forest regressors trained on uniform (A) and healthy brain (B) distributions at different noise levels. We highlight the different scale in the colour bar for SNR = ∞ compared to SNR = [5, 25]

FIGURE S13 Standard deviation in v_{cyl}, v_{csf} and λ_{cyl} estimates of the 3- SMT model for traditional fitting (A) and neural networks trained on uniform (B) and healthy brain (C) distributions when both the training and test data sets have SNR = 25

FIGURE S14 Standard deviation in v_{cyl}, v_{csf} and λ_{cyl} estimates of the 3- SMT model for random forest regressors trained on uniform (A) and healthy brain (B) distributions when both the training and test data sets have SNR = 25

FIGURE S15 Equivalent to results in Figure 5, but for random forest regressors instead of neural networks

FIGURE S16 Equivalent to results in Figure 6, but for random forest regressors instead of neural networks

How to cite this article: Gyori NG, Palombo M, Clark CA, Zhang H, Alexander DC. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. *Magn Reson Med*. 2022;87:932–947. doi:10.1002/mrm.29014