Computers in Biology and Medicine 149 (2022) 106091

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep-learned estimation of uncertainty in measurements of apparent diffusion coefficient from whole-body diffusion-weighted MRI

Konstantinos Zormpas-Petridis [a], Nina Tunariu [a,b], David J. Collins [a], Christina Messiou [a,b], Dow-Mu Koh [a,b], Matthew D. Blackledge [a,*]

[a] *Division of Radiotherapy and Imaging, The Institute of Cancer Research, London, United Kingdom*
[b] *Department of Radiology, The Royal Marsden National Health Service Foundation Trust, Surrey, United Kingdom*

ABSTRACT

Purpose: To use deep learning to calculate the uncertainty in apparent diffusion coefficient (σADC) voxel-wise measurements to clinically impact the monitoring of treatment response and improve the quality of ADC maps.

Materials and methods: We use a uniquely designed diffusion-weighted imaging (DWI) acquisition protocol that provides gold-standard measurements of σADC to train a deep learning model on two separate cohorts: 16 patients with prostate cancer and 28 patients with mesothelioma. Our network was trained with a novel cost function, which incorporates a perception metric and a b-value regularisation term, on ADC maps calculated by combinations of 2 or 3 b-values (e.g. 50/600/900, 50/900, 50/600, 600/900 s/mm^2). We compare the accuracy of the deep-learning based approach for estimation of σADC with gold-standard measurements.

Results: The model accurately predicted the σADC for every b-value combination in both cohorts. Mean values of σADC within areas of active disease deviated from those measured by the gold-standard by 4.3% (range, 2.87–6.13%) for the prostate and 3.7% (range, 3.06–4.54%) for the mesothelioma cohort. We also showed that the model can easily be adapted for a different DWI protocol and field-of-view with only a few images (as little as a single patient) using transfer learning.

Conclusion: Deep learning produces maps of σADC from standard clinical diffusion-weighted images (DWI) when 2 or more b-values are available.

1. Introduction

Whole-body diffusion-weighted MRI (WB-DWI) [1] provides unprecedented disease visualisation for the diagnosis and response assessment of bone metastasis from advanced prostate [2–5] and breast [6] cancers (APC and ABC respectively). Furthermore, it has recently been incorporated into guidelines for the evaluation of myeloma [7–10]. In addition to the high contrast between tumour and normal bone marrow on DWI to facilitate disease detection, the technique is also able to provide a surrogate measurement of tumour cellularity in the form of the Apparent Diffusion Coefficient (ADC), which is derived by sensitising images to water diffusion using two or more ‘b-values’ [2] (units mm^2/s). As DWI is non-invasive, non-ionising and doesn’t involve the use of intravenous contrast agents, clinical evidence suggests that monitoring ADC changes may provide a biomarker of tumour response to novel anti-cancer therapies [11].

However, an important characteristic of any successful tumour response biomarker is that the uncertainty of derived measurements be known if it is to be used for assessing individual patients in personalised treatment paradigms [12]. This is conventionally achieved through dedicated double-baseline measurements [13–16], whereby patients are asked to repeat the same MRI experiment, often in a span of days, twice without medical intervention. Therefore, such studies are all too rare due to increased study costs, hospital capacity limitations and increased patient burden.

To account for these issues, a recently developed methodology has demonstrated that it is possible to statistically model ADC measurement uncertainty, σ_{ADC}, within each imaging exam and for each voxel [17], providing a patient-derived estimate of measurement precision. This is achieved using a minor alteration to currently employed clinical WB-DWI protocols: A conventional diffusion-weighted image consists of taking multiple acquisitions for each b-value at the same anatomical location, and taking an average to produce an image with improved signal-to-noise ratio (SNR) [2] (we denote the number of excitations acquired for a particular b-value as “NEX”, such that for a protocol with three averages and three diffusion-encoding gradient directions per average, NEX = 9). To accurately calculate σ_{ADC}, however, the individual excitation images (NEX = 1) should be used in statistical modelling of ADC with weighted linear least squares fitting. The article demonstrated that maps of σ_{ADC} provide quantitative visualisation of the confidence radiologists should have in ADC estimates within any body-region, which could improve interpretation of imaging changes

[*] Corresponding author.
E-mail address: matthew.blackledge@icr.ac.uk (M.D. Blackledge).

https://doi.org/10.1016/j.compbiomed.2022.106091
Received 12 April 2022; Received in revised form 1 August 2022; Accepted 3 September 2022
Available online 13 September 2022

Table 1
Diffusion-weighted imaging protocol parameters for both study cohorts.

Parameter	Prostate Cancer Cohort	Mesothelioma Cohort
Scanner	1.5T Siemens Aera	1.5T Siemens Avanto
b-values (s/mm^2)	50, 600, 900	100, 500, 800
Gradient Directions (normalised)	(-1,0,0), (0,1,0), (0,0,1)	(-1,0,0), (0,1,0), (0,0,1)
Number of Signal Averages (NSA)*	3 × 1	4 × 1
Echo Time (ms)	79	92
Repetition Time (s)	12.7	6.0
Acquisition matrix (cols x rows)	128 × 104 [256 × 208][a]	128 × 92
Resolution (mm^2)	1.68 × 1.68[b]	3 × 3
Slices per station	40 (4–5 stations)	30 (2 stations)
Slice Thickness (mm)	5	5
Readout Bandwidth (Hz/voxel)	1955	1860
Parallel Imaging	GRAPPA (R = 2)	GRAPPA (R = 2)

[a] Values in square parentheses represent the image dimensions following interpolation by the scanner.

[b] The resolution is presented following image interpolation by the scanner.

after anti-cancer treatments. Additionally, by mathematically combining σ_{ADC} and ADC maps it is possible to produce a novel contrast mechanism (niceDWI) that improves disease visualisation and standardisation of imaging signal between patients.

Unfortunately, a pitfall of this approach is that in the clinical setting individual excitations (NEX = 1) are discarded by the scanner and only the average image is retained at each b-value to reduce data storage overheads. Although calculation of σ_{ADC} is technically possible using only these averaged images, it is greatly corrupted by imaging noise when compared to estimates derived from the approach where each acquisition is retained [17]. Moreover, estimation of σ_{ADC} using averaged images becomes impossible in cases where only two b-values are acquired, which is common in many clinical WB-DWI protocols.

In this paper, we overcome these limitations by:

1. Training a neural network to perform robust estimation of σ_{ADC} at every voxel location directly from averaged images (NEX≥9) using estimates of σ_{ADC} derived form individual acquisitions (NEX = 1) as the target distribution (ground-truth).
2. Demonstrating that the trained network is also able to compute σ_{ADC} when only 2 averaged b-value images are available, thus allowing the use of the technique on virtually any clinical scanner without changes to the clinical acquisition protocol.

We experiment with various well-established deep neural networks based on the U-Net [18] and ResNet [19] architectures using two different loss functions, mean-absolute error (MAE) and feature-wise (perceptual) loss based on a pretrained deep-learning classification network (VGG16) [20]. We have also implemented a novel “b-value regularised” network architecture, which implements a b-value prediction loss function to improve the performance of the network. The final network was evaluated on two separate patient cohorts, the first with metastatic prostate cancer who underwent WB-DWI for surveillance of disease progression, and a second cohort of patients with malignant pleural mesothelioma (MPM) who underwent whole-lung MRI. We further evaluated whether the algorithm trained using the first patient cohort could be used for successful inference of σ_{ADC} in the second patient cohort and evaluated how many patients were required to fine-tune inference using a transfer-learning paradigm. Such experiments are vital to better understand the requirements of clinical adoption of deep-learning algorithms on different site of disease of at different imaging centers.

2. Materials and Methods

2.1. Patient population and imaging protocol

This retrospective study consists of two patient populations: (i) the **prostate cancer cohort** consisting of 16 patients with metastatic prostate cancer and suspected disease in the skeleton, and (ii) the **mesothelioma cohort** consisting of 28 patients with mesothelioma. Both imaging studies were approved by the institutional review board, and the requirement for patient consent in the first study was waived as there was no alteration to the patient clinical pathway and data was fully anonymized before use, whilst in study (ii) patient consent was obtained. In both cases, axial diffusion-weighted imaging (DWI) was performed using 3 b-values across each cohort over 4–5 (prostate) and 1–2 (mesothelioma) sequential imaging stations on a 1.5T scanner (Aera or Avanto, Siemens Healthineers, Germany). Within each cohort, the same protocol was used to ensure consistency of results (see Table 1 for full protocol parameters). In both cohorts, for each b-value a 3-directional orthogonal diffusion encoding scheme was applied using bipolar gradients to mitigate the effects of eddy-current induced distortions. Only a single average was acquired per direction, and this was repeated 3 and 4 times for the prostate cancer and mesothelioma cohort respectively. This led to a total of 9 and 12 acquisitions per b-value and per slice for the prostate cancer and mesothelioma cohorts. For the prostate cancer cohort, patients were randomly split into training/validation/test groups according to 10/3/3; for the mesothelioma cohort, this random split was 20/4/4 [13].

2.2. Data processing

An illustration of the statistical σ_{ADC} data-fitting approach, the results of which were used here as the gold-standard, is presented in Fig. 1(a): ADC maps were calculated assuming a monoexponential decay model [2] as the negative gradient from a linear-least-squares (LLS) approximation of the log-transformed averaged image signals (NEX = 9 for the prostate cancer cohort and NEX = 12 for the mesothelioma cohort). The y-axis intercept was also estimated, $ln(S_0)$, which represents the log-signal expected at b = 0 s/mm ^{2} (no diffusion weighting). From the individually retained acquisitions at each b-value (NEX = 1), maps of σ_{ADC} were calculated using an iterative weighted linear least-squares (IWLS) approach [17]; these σ_{ADC} maps acted as ground truth for our purposes. To make our model robust to different combinations of b-values, this analysis was repeated for each combination available in this dataset: 50/600/900, 50/900, 50/600, and 600/900 s/mm ^{2}. It is clear from Fig. 1(a) that σ_{ADC} is generally much higher when computed using b = 600/900 s/mm^2 due to the inherently low image SNR at these b-values; this is also evident in the apparently noisier estimated ADC map and $ln(S_0)$ image.

We then developed, trained and tested a deep-learning architecture (Fig. 1(b)) to generate maps of σ_{ADC} using as input only the ADC and $ln(S_0)$ maps estimated using the averaged data at each b-value (NEX = 9/12). A total of 6400 training, 2080 validation and 2240 testing images were obtained for the prostate cancer cohort and 4320 training, 960 validation and 960 testing images for the mesothelioma cohort.

2.3. Deep learning

2.3.1. Selecting the deep learning architecture

We experimented with various network architectures to predict the σ_{ADC} maps. Firstly, we trained a neural networks based on the U-Net architecture [18] with the number of filters used in each block being the same as suggested in the original paper [64, 128, 256, 512, 1024, 512, 256, 128, 64], named U-Net$_{heavy}$ (31 million parameters), and then we trained a version with fewer parameters [16, 32, 64, 128, 256, 128, 64, 32, 16], named U-Net$_{light}$ (1.94 million parameters). Subsequently, we implemented residual blocks [19] for the encoder and decoder sections

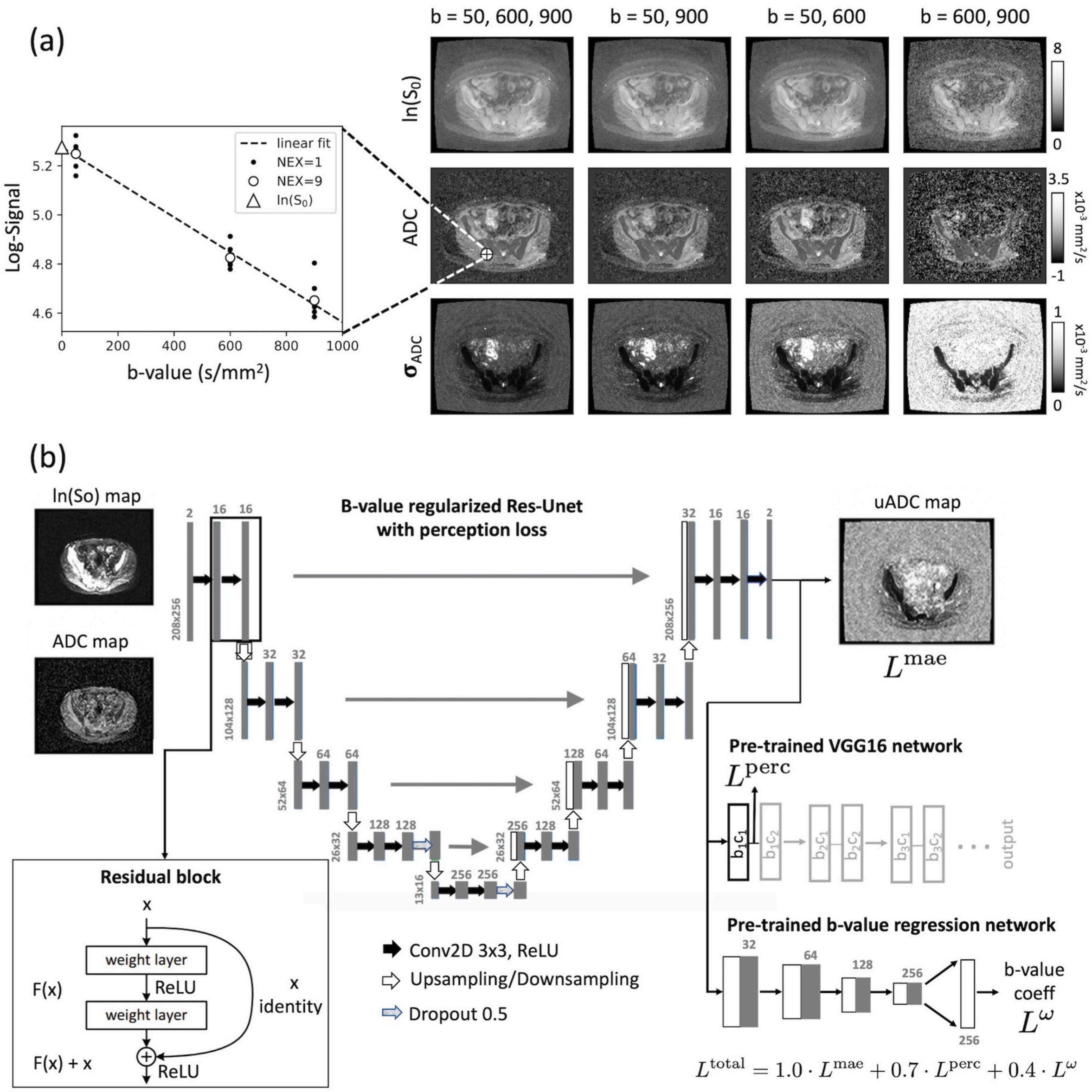

Fig. 1. (a): An illustration of our data processing pipeline. From the nine single acquisition data (NEX = 1, black scatter points) we derive the clinical standard as the average at each b-value (NEX = 9, green scatter points). From the log-transformed averaged data, we obtain estimates of (i) the ADC from the negative gradient of the linear fit, and (ii) the log-transformed S_0 image from the y-axis intercept (red scatter point). σ_{ADC} is calculated using the NEX = 1 data, as its estimation is difficult given only 3 data-points, and impossible if only 2 are provided. It is clear that 2-point estimation of ADC using b = 600 and 900 s/mm^2 results in noisy ADC maps, a fact that is supported by a general increase in σ_{ADC} for these data. **(b)**: Our deep-learning U-Net attempts to reconstruct σ_{ADC} measurements from only input ADC maps and *lS*0 images, making 2-point estimation of σ_{ADC} a possibility.

of the network with the same two combinations of filter numbers as for the U-Net networks; these were similarly named Res-U-Net$_{\text{heavy}}$ (31.2 million parameters) and Res-U-Net$_{\text{light}}$ (1.95 million parameters).

For all networks, a linear activation was used for the last layer and a rectified linear unit (ReLU) activation function in all preceding layers. The weights incident to each hidden unit was constrained to a norm value less than or equal to 3 and random weight initialization with He normal initialization for all weights/biases [21]; 2 layers also used 50% dropout during training (Fig. 1(b)) to reduce overfitting [22]. The networks were trained with a batch size of 25 DWI slices for 100 epochs and optimized using the Adam algorithm [23] with a learning rate of 10^{-3} for the *light* versions and 10^{-4} for the *heavy* versions. All images were standardised to the mean and standard deviation of the training WBDWI set. The networks were trained using a Tesla P100-PCIE-16GB GPU-card and the trained algorithm was applied using a MacBook Pro laptop (2.9 GHz Intel-Core-i7-CPU, 16 GB-2133-MHz-LPDDR3-RAM).

2.3.2. *Selecting the loss function*

We have developed a unique loss function to improve the appearance and quantitative accuracy of derived σ_{ADC} maps. Firstly, the mean-absolute-error (mae) is used as a cost function such that the network minimises the voxel -wise absolute difference between the target (gold-standard) image and the network-derived estimate, $\sigma^{\dagger}_{ADC}$:

$$L^{mae} = \frac{1}{N_c N_r} \sum_{i=1}^{N_c} \sum_{j=1}^{N_r} \left| \sigma_{ADC,i,j} - \sigma^{\dagger}_{ADC,i,j} \right|$$

for each voxel location i,j.

To improve the visual similarity of generated maps to ground-truth, we also incorporated a *perceptual loss* using features derived from a pre-trained classification network (VGG16) with weights previously optimized using the ImageNet database [20]. Features from the first block of the first VGG16 layer were extracted from both the true σ_{ADC} maps and

Table 2
Comparison of image similarities on the test patients of the prostate cancer cohort to select the network architecture. The mean-absolute-error (Mae), structural similarity index (SSIM) and peak-signal-to-noise ratio (PSNR) were calculated for all four b-value combinations: (50, 600, 900), (50, 900), (50, 600) and (600, 900). On the top part the results about the base architecture using a L^{mae} cost function are presented, on the middle part the effects of the different cost functions are shown while keeping a simple base architecture steady and on the bottom the outcome of our proposed architecture. Note that the bold values indicate the best performance achieved for each metric.

Network	(50, 600, 900)			(50, 900)			(50, 600)			(600, 900)		
	Mae (10^{-6} *mm*^{2}*/s*)	*SSIM* (10^{-2})	PSNR	*Mae* (10^{-6} *mm*^{2}*/s*)	*SSIM* (10^{-2})	PSNR	*Mae* (10^{-6} *mm*^{2}*/s*)	*SSIM* (10^{-2})	PSNR	*Mae* (10^{-6} *mm*^{2}*/s*)	*SSIM* (10^{-2})	PSNR
***U-Net**$_{heavy}$* (L^{mae})	13.7	99.4	48.7	17.7	99.1	45.9	25.2	98.7	44.1	40.1	97.9	41.2
***U-Net**$_{light}$* (L^{mae})	13.6	99.5	49	18.1	99.0	45.6	25.2	98.7	43.7	40.4	97.9	41.2
***Res-U-Net**$_{heavy}$* (L^{mae})	13.1	99.5	49.4	17.6	99.1	46	24.7	98.8	44.2	39.9	97.9	41.2
***Res-U-Net**$_{light}$* (L^{mae})	12.8	**99.6**	49.9	17.3	99.2	46.3	24.5	**98.9**	44.3	39.7	**98.0**	41.3
***U-Net**$_{light}$* (L^{mae})	13.6	99.5	49	18.1	99.0	45.6	25.2	98.7	43.7	40.4	97.9	41.2
***U-Net**$_{light}$* ($L^{mae}+L^{perc}$)	12.6	**99.6**	49.9	17	99.2	46.5	24.2	**98.9**	44.4	39.4	**98.0**	41.4
***U-Net**$_{light}$* ($L^{mae}+L^{\omega}$)	12.6	**99.6**	49.7	17.4	99.2	46.2	24.2	**98.9**	44.4	39.6	**98.0**	41.3
***U-Net**$_{light}$* (L^{total})	12.5	**99.6**	49.9	16.9	**99.3**	46.6	23.8	**98.9**	44.6	39.3	**98.0**	41.4
***Res-U-Net**$_{light}$* (L^{total})	**12.3**	**99.6**	**50.2**	**16.7**	**99.3**	**46.7**	**23.6**	**98.9**	**44.7**	**39.1**	**98.0**	**41.5**

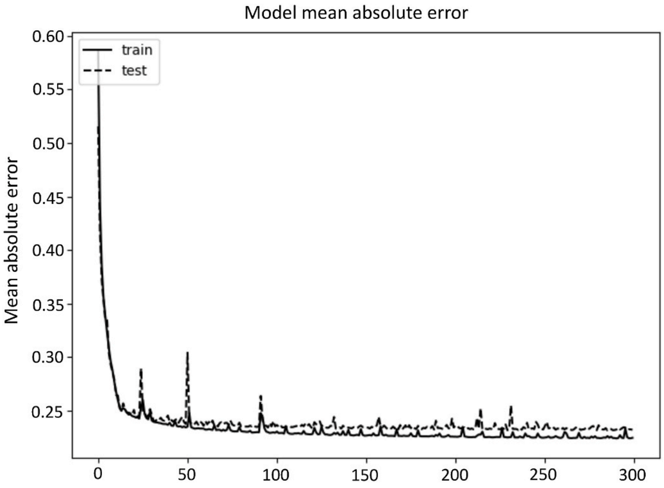

Fig. 2. Training (solid) and validation (dashed) curves depicting the change in mean-absolute-error over the training epochs that we investigated. It is clear that a plateau in the validation curve is observed after epoch 50.

the corresponding deep-learned estimates $\sigma^{\dagger}_{ADC}$, and the mean-squared-error (MSE) between both feature vectors was subsequently minimised in back-propagation:

$$L^{perc} = \frac{1}{N_f}\sum_{k=1}^{N_f}\left[f(\sigma_{ADC})_k - f\left(\sigma^{\dagger}_{ADC}\right)_k\right]^2$$

$f : \mathbb{R}^{N_r N_c} \rightarrow \mathbb{R}^{N_f}$ represents the operation of extracting N_f features from σ_{ADC} maps using the VGG16 network. The first VGG16 layer was selected to capture high-order features within σ_{ADC} maps.

Lastly, we implemented a regularisation parameter. It has been previously shown [17] that the estimated magnitude of ADC uncertainty is mathematically linked to the b-values from which it is derived according to the equation

$$\widehat{\sigma}_{ADC} = \sqrt{\left(B^{\top}WB\right)^{-1}}\cdot\sigma_{\nu}$$

with b-value design matrix $B = \begin{pmatrix} b_1 & b_2 & \dots & b_N \\ 1 & 1 & \dots & 1 \end{pmatrix}^{\top}$ for N not necessarily unique b-values, σ_{ν} representing the standard deviation (noise) of log-transformed data at each b-value (which as an approximation we assume to be constant), and W representing a $N \times N$ square diagonal matrix with diagonal components equal to the desired weight for each b-value when performing ADC fitting. For the purposes of this research, we assume equal weighting for all b-values such that $W_{ij} = 1$ if $i = j$ and 0 otherwise, and subsequently derive

$$\begin{aligned}\widehat{\sigma}_{ADC} &= \sqrt{\frac{N}{N\sum_{l} b_l^2 - \left(\sum_{l} b_l\right)^2}}\cdot\sigma_{\nu} \\ &= \omega\cdot\sigma_{\nu}\end{aligned}$$

$\omega : \mathbb{R}^N \rightarrow \mathbb{R}^1$ encodes the combination of known b-values. We predict from this that estimation of ω from accurate maps of σ_{ADC} should be possible and thus used to help regularise the full $\sigma^{\dagger}_{ADC}$ estimation network during training.

We developed a b-value regression network to estimate ω from ground truth σ_{ADC} maps (illustrated in Fig. 1(b)). The network consists of four VGG-like blocks with 32, 64, 128, and 256 filters respectively (3 × 3 filter size in each case), with each block followed by a max-pooling operation and a 20% dropout layer to reduce overfitting. The final, fully-connected layer consists of 256 neurons (dropout 20%) followed by another dense layer consisting of a single neuron to generate the estimate of ω. A linear activation was used for this last layer whilst a ReLU activation function was used in all preceding layers. The He technique [21] was used to initialize all layers prior to training, and the network was trained using a batch size of 30 for 70 epochs; a MSE loss function and the Adam algorithm with a learning rate of 0.001 for optimisation were employed.

The trained b-value regression network was used to improve the loss function of our σ_{ADC} estimation network: during training the network was used to estimate ω, but this time from the estimated $\sigma^{\dagger}_{ADC}$ maps. This loss may be characterised as

$$L^{\omega} = \left[g\left(\sigma^{\dagger}_{ADC}\right) - \omega\right]^2$$

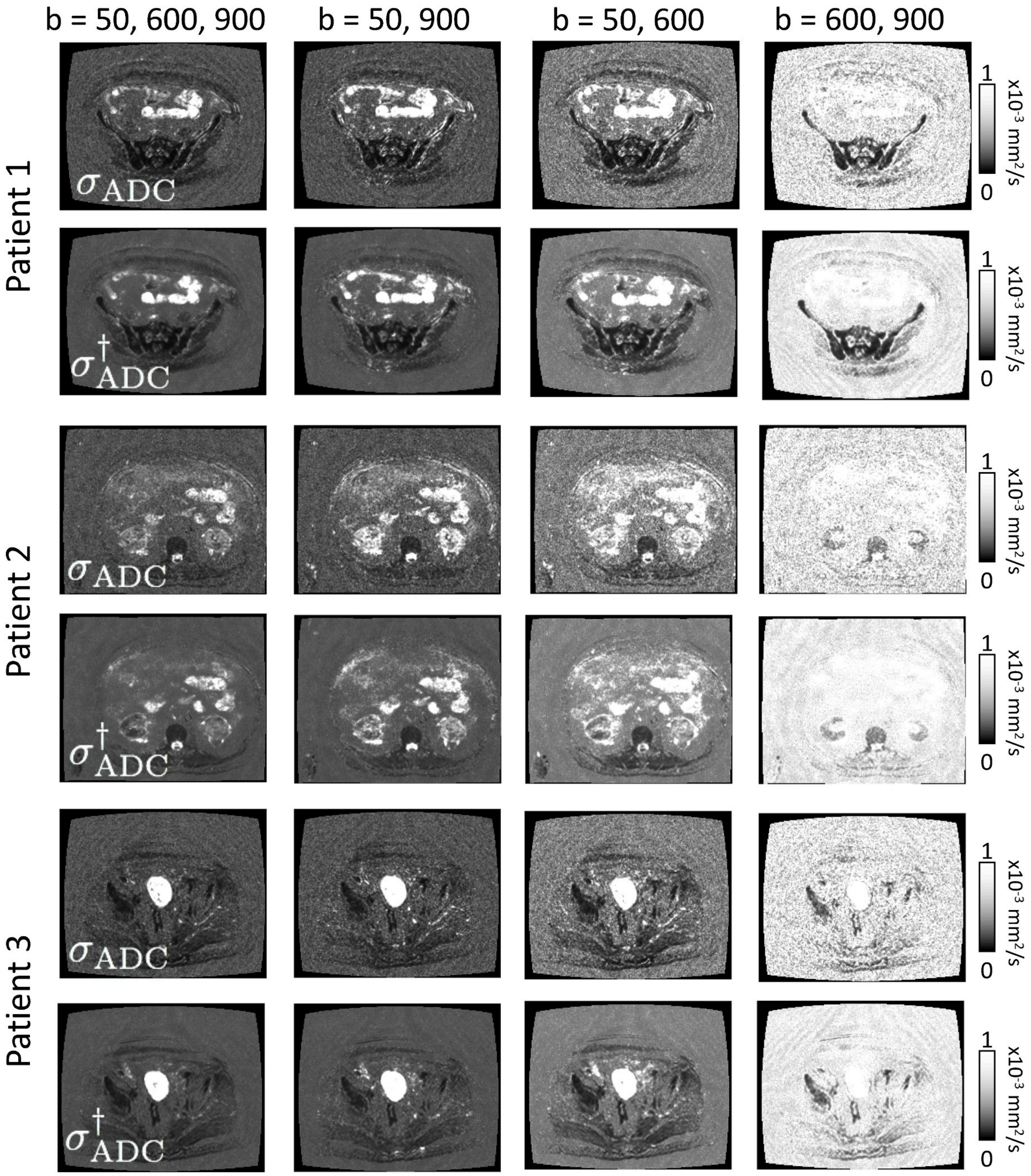

Fig. 3. Example axial slices from each of the test patient datasets collected from the prostate cancer cohort. There is good agreement between gold standard σ_{ADC} and the corresponding deep-learning estimated map, $\sigma^{\dagger}_{ADC}$, in all cases and for all b-value combinations. Surprisingly, the deep-learning approach was also able to reconstruct the characteristic 'ringing' artefact observed in the background of the σ_{ADC} maps. In all images, windowing levels were kept identical.

with $g : \mathbb{R}^{N_r N_c} \to \mathbb{R}^1$ representing the operation of the trained b-value regression network, such that $g(\sigma^{\dagger}_{ADC}) = \omega^{\dagger}$ is the value estimated from the σ_{ADC} map generated by the deep-learning algorithm during training.

Our final cost function for estimating $\sigma^{\dagger}_{ADC}$ was therefore

$$L^{total} = 1.0 \cdot L^{mae} + 0.7 \cdot L^{perc} + 0.4 \cdot L^{\omega}$$

where the weighting for each individual loss was found empirically after experimentation with different values. We also compared the simple U-Net$_{light}$ network trained using the following loss functions: (i) L^{mae}, (ii) $1.0 \cdot L^{mae} + 0.7 \cdot L^{perc}$ and (iii) $1.0 \cdot L^{mae} + 0.4 \cdot L^{\omega}$. The weighting were kept consistent for all experiments.

2.4. Quantitative comparison of image similarity and regions of active disease

To compare the accuracy of deep-learning estimates of σ_{ADC} (denoted henceforth as $\sigma^{\dagger}_{ADC}$) with the gold-standard, a radiologist delineated regions of metastatic bone disease on the prostate cancer cohort, using an in-house developed semi-automated segmentation pipeline [11]. In addition to this, a physicist with more than 10 years' experience in body DWI semi-automatically delineated regions of disease in the mesothelioma cohort using the tools available within 3D Slicer; these regions were subsequently verified and (where needed) corrected by a consultant radiologist (>10 years' experience in body DWI). Resultant regions

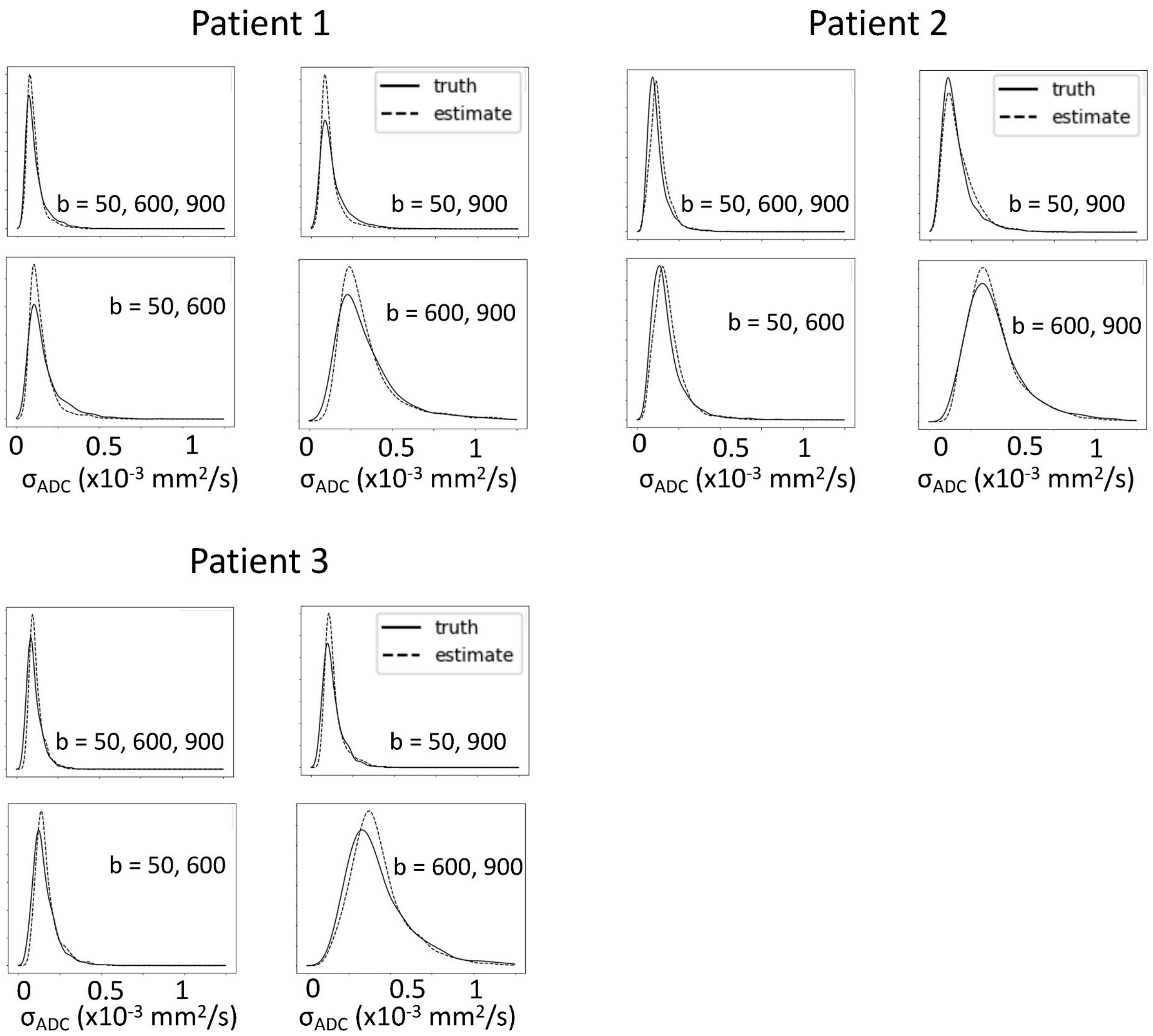

Fig. 4. Kernel density distributions of ground-truth measurements of σ_{ADC} and deep-learned values $\sigma^{\dagger}_{ADC}$ within regions of metastatic prostate disease. Distributions are observed to display similar characteristics indicating that the deep-learned estimation is performing well.

Table 3
Statistics between the gold-standard and neural network predicted values within regions of active disease for the test patients of the prostate cancer cohort. Data are shown as the median values calculated across all three test patients.

Parameter	(50, 600, 900)	(50, 900)	(50, 600)	(600, 900)
Relative difference of means (%)	5.06	3.15	6.13	2.87
Correlation	0.81	0.76	0.78	0.85
Mean absolute error (x10^{-6} mm^2/s)	13.3	18.6	22.1	43.9
Coefficient of Variation	0.08	0.08	0.07	0.06

of interest (ROIs) were transferred onto both the gold standard σ_{ADC} maps, and onto those generated by the deep-learning algorithm. We compared the values within regions of bone disease by calculating the relative difference of means (RDM), the mean absolute voxel-wise difference, the Pearson correlation (r) and the coefficient of variation (CoV). All metrics were calculated using scikit-learn v. 0.14.2. We also calculated the kernel density estimation of the values within bone disease using a gaussian kernel with automatic bandwidth determination [24] to visually assess the similarity of the distributions.

The similarity between the gold-standard σ_{ADC} and the deep learning estimate $\sigma^{\dagger}_{ADC}$ was also assessed by calculating the mean-absolute-error (*mae*), the structural-similarity index (SSIM) and the peak-signal-to-noise ratio (PSNR) for every b-value combination.

2.5. *Transfer learning*

For the four validation mesothelioma patients, we evaluated the MAE between the σ_{ADC} and the $\sigma^{\dagger}_{ADC}$ maps estimated using the Res-U-Net$_{\text{light}}$ network with the L^{total} cost function trained using only the prostate cancer cohort dataset. We used those weights as starting point and subsequently retrained the network by gradually introducing 1 or more training datasets (in increments of one patient) from the mesothelioma cohort. The network was retrained with the same set of parameters and a smaller learning rate of 10^{-5} for 100 epochs. The average *mae* was calculated between network-produced images and ground-truth images across all validation images and across all b-value combinations.

3. Results

3.1. *Selecting the deep learning architecture and loss function*

Our experiments showed that U-Net-like architectures are in general

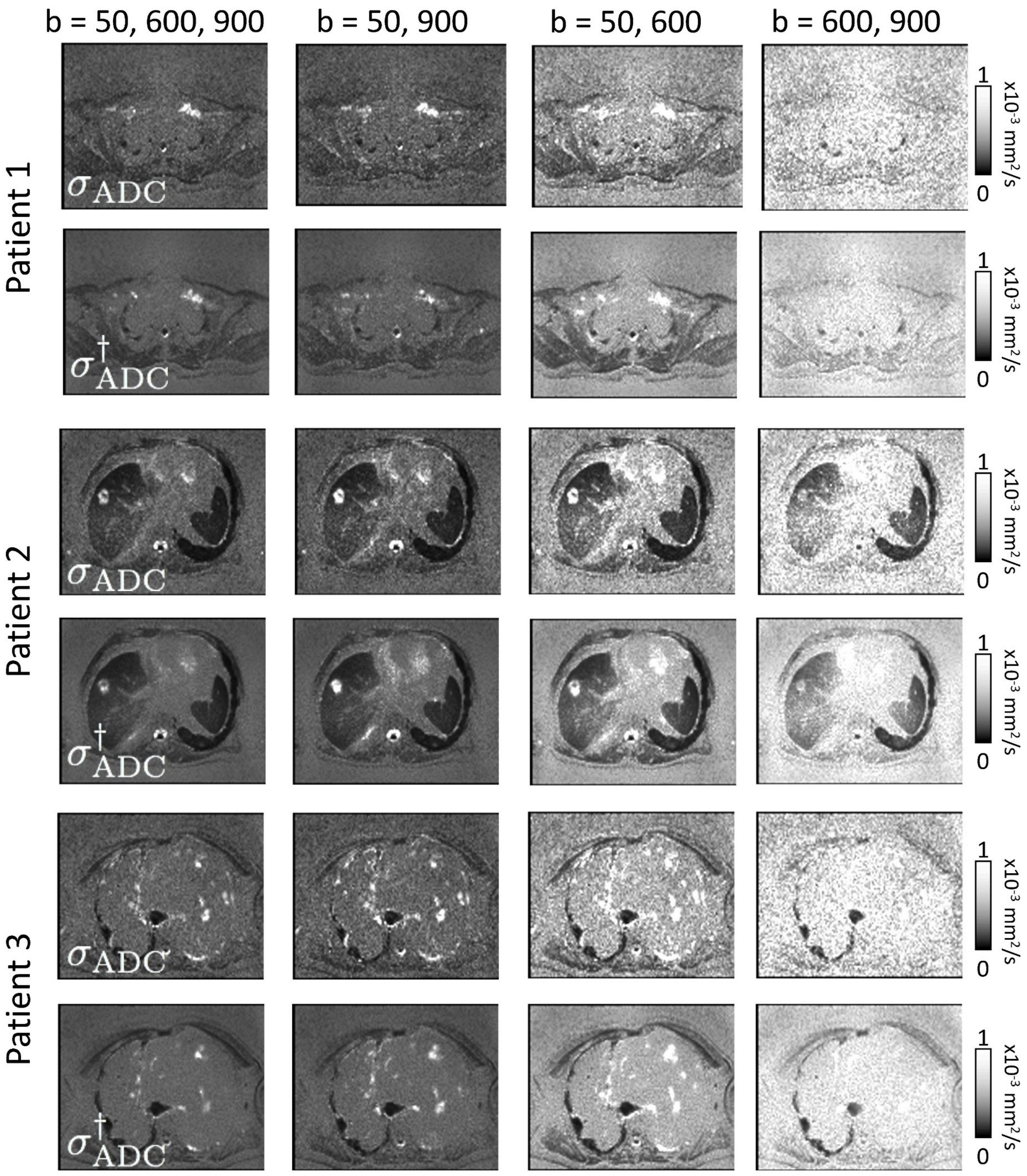

Fig. 5. Example axial slices from each of the test patient datasets collected from the mesothelioma cancer cohort. There is good agreement between gold standard σ_{ADC} and the corresponding deep-learning estimated map, $\sigma^{\dagger}_{ADC}$, in all cases and for all b-value combinations. In all images, windowing levels were kept identical.

appropriate to solve this task. First, we compared the four different base network architectures using the L^{mae} cost function for all four b-value combinations on the test group of the prostate cancer cohort: Res-U-Net$_{light}$ scored the lowest mean-absolute error and highest SSIM and PSNR for all b-value combinations and was therefore selected as our base architecture. Detailed results are presented in Table 2.

Next, to select the *cost function* we used the simple U-Net$_{light}$ architecture and experimented with (i) L^{mae}, (ii) $1.0{\cdot}L^{mae} + 0.7{\cdot}L^{perc}$, (iii) $1.0{\cdot}L^{mae} + 0.4{\cdot}L^{\omega}$ and (iv) L^{total}. L^{total} scored the lowest MAE and highest SSIM and PSNR for all b-value combinations and was therefore selected as our cost function. Also, U-Net$_{light}$ with any enhanced cost function (ii), (iii) and (iv) outperformed Res-U-Net$_{light}$ with L^{mae} in all metrics.

Finally, we applied the Res-U-Net$_{light}$ architecture with the L^{total} cost function which scores the lowest MAE and highest SSIM and PSNR of all previous combinations and was therefore selected as our full network. Detailed results are presented in Table 2. Training and validation curves of the model for the prostate cancer cohort are presented in Fig. 2; clear stabilisation of the validation *mae* loss was observed after approximately 50 epochs.

Based on these results, we selected the Res-Unet^{light} architecture with L^{total} cost function as our final model and we performed our subsequent experiments on the mesothelioma cohort and the delineated regions of metastatic bone disease using only this model.

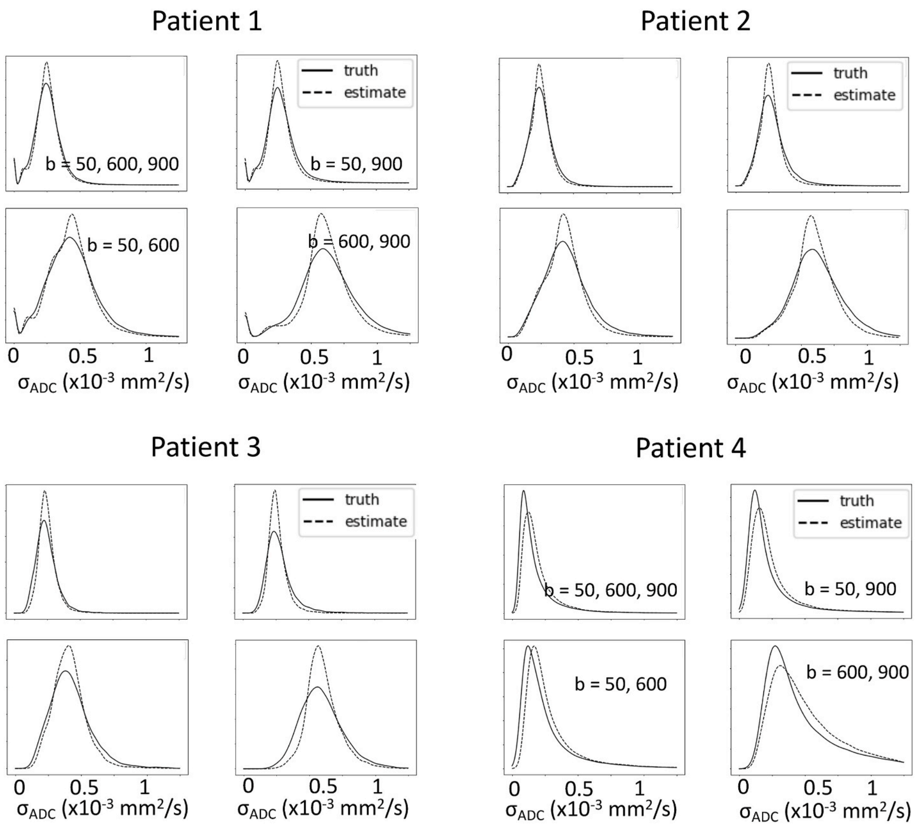

Fig. 6. Kernel density distributions of ground-truth measurements of σ_{ADC} and deep-learned values $\sigma^{\dagger}_{ADC}$ within regions of mesothelioma disease. Distributions are observed to display similar characteristics indicating that the deep-learned estimation is performing well.

Table 4
Image and values within regions of active disease statistics for the validation and test patients of the mesothelioma cohort between the gold-standard and the neural network predicted maps. Data are shown as the median values calculated across all eight validation and test patients.

Image Statistics				
Parameter	(50, 600, 900)	(50, 900)	(50, 600)	(600, 900)
Mean absolute error (x10^{-6} mm^2/s)	11.4	15	23.4	28.1
SSIM (10^{-2})	99.7	99.5	99.2	99.0
PSNR	50.8	47.8	44.9	44.3
Disease Statistics				
Relative difference of means (%)	3.06	4.54	3.48	3.76
Correlation	0.78	0.73	0.78	0.74
Mean absolute error (x10^{-6} mm^2/s)	27.8	34.6	51.2	70.2
Coefficient of Variation	0.06	0.09	0.08	0.06

3.2. *Deep learning can provide estimates of ADC uncertainty*

Fig. 3 demonstrates exemplar slices from each of the three test patient datasets in this cohort. It is clear that good visual agreement is observed between the gold-standard σ_{ADC} and DL-estimated $\sigma^{\dagger}_{ADC}$ maps. Importantly, this has been achieved for all four b-value combinations tested, results that are corroborated in our quantitative comparison (Fig. 4) where distributions of σ_{ADC} and $\sigma^{\dagger}_{ADC}$ values within disease exhibit similar characteristics. Statistics between predicted and gold-standard values within regions of active disease for the three test patients of the cohort are shown in Table 3.

For the mesothelioma cohort our expert radiologists observe the same level of visual agreement (exemplar slices from 3/4 test patients shown in Fig. 5, and histograms of estimates of σ_{ADC} within disease shown in Fig. 6 for all four test patients). This demonstrates the ability of the network to produce good estimates of σ_{ADC} in regions with smaller fields of view. Statistics for the image similarity and values within regions of active disease for the eight validation/test patients the cohort are shown in Table 4.

3.3. *Transfer learning enables retraining with fewer images for application in different clinical protocols*

The network trained on the prostate cancer cohort and applied on the validation/test patients of the mesothelioma cohort directly without retraining produced moderate results (mae = 35 × 10^{-6} mm^2/s). However, after training with only 1 patient the results improved substantially (mae = 14.7 × 10^{-6} mm^2/s), reaching a good quality after 6 patients (mae = 12.3 × 10^{-6} mm^2/s) and remained approximately stable after 10 patients (mae = 11.9 × 10^{-6} mm^2/s - fully-trained model mae = 11.4 × 10^{-6} mm^2/s). We observe that while the network continues to improve by adding more data, it achieves image quality comparable to that of the fully-trained model with data from only a small number of patients (Fig. 7).

4. Discussion

Estimation of ADC measurement reliability is an essential task if WBDWI is to be embraced as a cancer response imaging biomarker in the healthcare community. Whilst it may be possible to acquire specialist datasets that allow calculation of ADC uncertainty, σ_{ADC}, at every anatomical location, the fact remains that most if not all clinical centers will not be able to acquire such data. This will likely be due to (i) an inherent delay by scanner manufacturers to implement such approaches

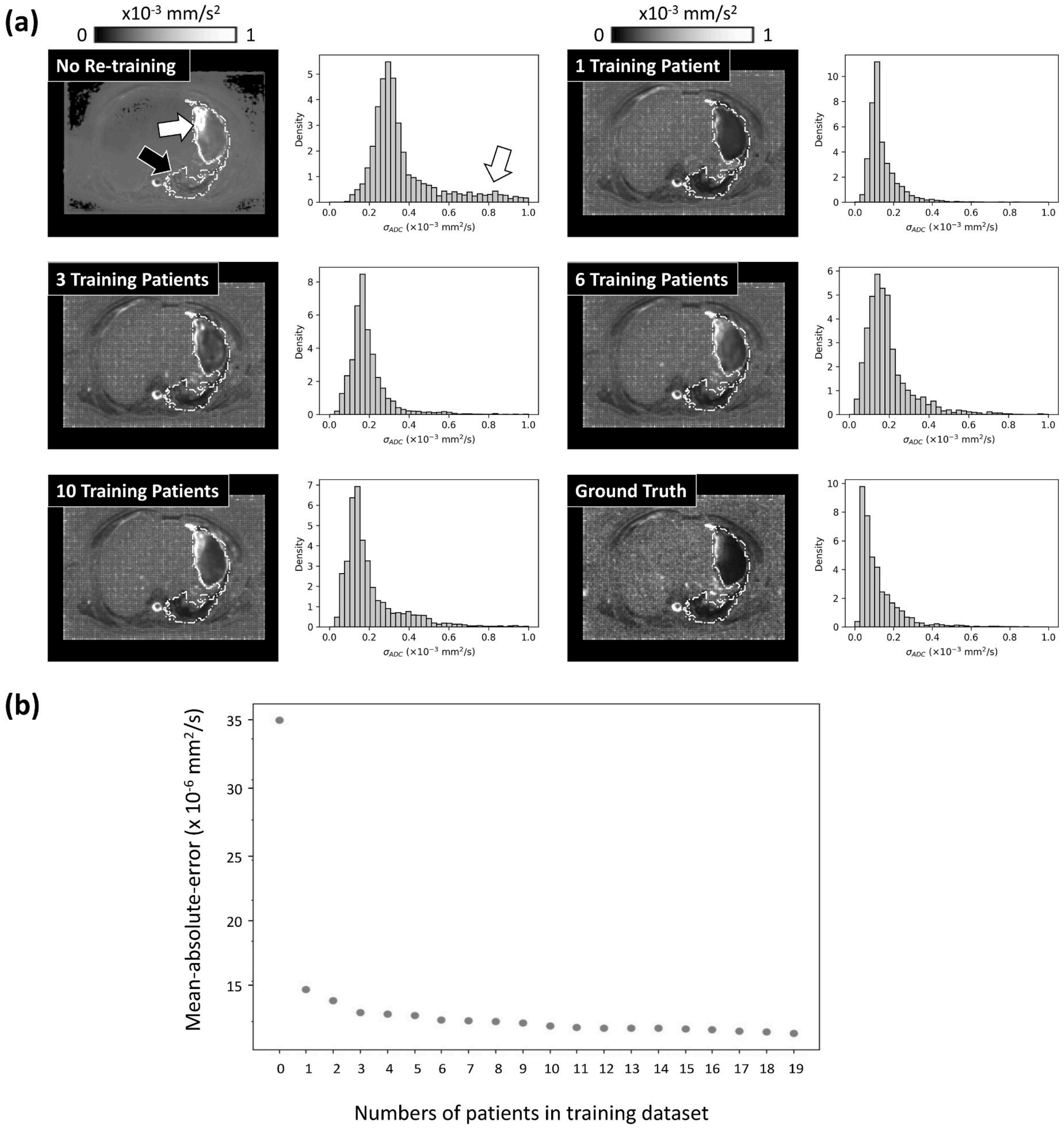

Fig. 7. Our transfer learning approach used the weights of the network trained in the prostate cancer cohort as starting point and applied the network on the mesothelioma dataset first without retraining and subsequently by retraining the network in increments of 1 patient at a time **(a):** Example slice and histograms of the value distributions within regions of fluid pleural effusion (white arrow) and solid disease (black arrow). Note that without re-training the network predicts higher uncertainty within regions of pleural effusion (white arrow), which may be due to the fact that disease of this kind is not present in whole-body MRI of advanced prostate cancer. However by retraining the network, even with images from 1 patient, the accuracy of the prediction was markedly improved. **(b):** The average *mean-absolute-error* was calculated between network-produced images and ground-truth images across all validation images (4 patients) and across all b-value combinations. Note that there is dramatic reduction in *mae* after including just one training patient dataset in the transfer learning approach.

within the clinical pathway, (ii) reluctance of centers for implementing non standard-of-care sequences, (iii) the increased data storage costs required to allow accurate assessment of σ_{ADC} and (iv) the lack of local expertise to process the images and provide results. Using conventional clinical sequences using 3-point measurements produces statistically imprecise σ_{ADC} measurements, whilst 2-point measurements cannot be used.

In this article we demonstrate preliminary evidence that the use of deep-learning with a U-Net-like architecture can break such classical assumptions and provide robust estimation of σ_{ADC} for DWI datasets acquired with only 2 or 3 unique b-values. We hypothesise that this is because the deep-architecture is able to learn complex relationships between a given voxel and its neighbouring regions, in order to arrive at robust estimation of the local noise field. Traditional approaches to this task include spatial filtering and wavelet decomposition, but these techniques tend to perform poorly and can create artefactual edges in the resultant images. The Res-U-net$_{\text{light}}$ network we have employed consists of many trainable parameters (~ 1.95 million) that are able to learn whether voxel differences are due to genuine noise or due to an object feature. Of course, other similar deep learning architectures could potentially be employed in future experiments [25,26].

We also show evidence that fusing the neural network with information inspired by more 'human' concepts, such as the high-order features extracted from the top layer of a pre-trained in ImageNet VGG16 network and forcing the network to predict the b-value combination of the image, can have a significant impact in its performance. A similar concept was presented by *Hagos* et al. [27], where a neural network for cell segmentation increased its detection performance by being forced to also predict the correct number of cells in the image. Here, we show that despite the simple U-Net$_{\text{light}}$ architecture being out-performed by Res-U-Net$_{\text{light}}$, it produced better results with the enhanced cost functions. The best network here being Res-U-Net$_{\text{light}}$ with our unique cost function suggests a need to combine modern complex network architectures with relevant but more abstract concepts to achieve the most accurate and visually similar results. These concepts and loss functions could be fused with virtually any neural network of this type of architecture and potentially improve its performance.

Future studies should include evaluation on larger patient datasets, preferably acquired on a variety of different scanner models. An important finding from our study is that the U-Net could accurately estimate σ_{ADC} independently of the b-value combination that produced the estimated ADC maps and $ln(S_0)$ image used as input. It would be valuable to further assess how robust our algorithm is to other b-value combinations. Our preliminary evidence shows that transfer learning by using a set of trained network weights as a starting point, would easily allow the successful retraining of the network with only a few images, making it suitable for use in potentially any diffusion-weighted imaging clinical protocol. Furthermore, this approach could be used in other quantitative imaging applications (including native T1, T2 and T2* estimation), where knowledge of the voxel-wise uncertainties would be valuable.

5. Conclusion

Deep-learned estimation of ADC uncertainties could provide clinicians with increased confidence when using WBDWI to monitor response of cancer to treatment. Moreover, our technique does not require modification of existing clinical protocols and could therefore be applied to existing datasets for retrospective evaluation of WBDWI.

Declaration of competing interest

Konstantinos Zormpas-Petridis and Matthew D Blackledge have submitted a patent to the Hellenic Industrial Property Organisation directly regarding the work described in this article. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Cancer Research UK (https://www.cancerresearchuk.org/) and Engineering and Physical Sciences Research Council support to the Cancer Imaging Centre at Institute of Cancer Research and Royal Marsden Hospital in association with Medical Research Council and Department of Health C1060/A10334, C1060/A16464 and National Health Service funding to the National Institute for Health Research (https://www.nihr.ac.uk/) Biomedical Research Centre, Clinical Research Facility in Imaging and the Cancer Research Network.

This report was independent research funded by the National Institute for Health Research. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. We thank The Betty Lawes Foundation and Sir Samuel Scott of Yews Trust for funding.

References

[1] A. Barnes, et al., UK quantitative WB-DWI technical workgroup: consensus meeting recommendations on optimisation, quality control, processing and analysis of quantitative whole-body diffusion-weighted imaging for cancer, Br. J. Radiol. 91 (2018), 20170577, https://doi.org/10.1259/bjr.20170577.

[2] D.-M. Koh, et al., Whole-body diffusion-weighted MRI: tips, tricks, and pitfalls, Am. J. Roentgenol. 199 (2012) 252–262, https://doi.org/10.2214/AJR.11.7866.

[3] M. Eiber, et al., Whole-body MRI including diffusion-weighted imaging (DWI) for patients with recurring prostate cancer: technical feasibility and assessment of lesion conspicuity in DWI, J. Magn. Reson. Imag. 33 (2011) 1160–1170, https://doi.org/10.1002/jmri.22542.

[4] A.R. Padhani, et al., Therapy monitoring of skeletal metastases with whole-body diffusion MRI, J. Magn. Reson. Imag. 39 (2014) 1049–1078, https://doi.org/10.1002/jmri.24548.

[5] A.R. Padhani, et al., METastasis reporting and data system for prostate cancer: practical guidelines for acquisition, interpretation, and reporting of whole-body magnetic resonance imaging-based evaluations of multiorgan involvement in advanced prostate cancer, Eur. Urol. 71 (2017) 81–92, https://doi.org/10.1016/j.eururo.2016.05.033.

[6] A.R. Padhani, D.-M. Koh, D.J. Collins, Whole-body diffusion-weighted MR imaging in cancer: current status and research directions, Radiology 261 (2011) 700–718, https://doi.org/10.1148/radiol.11110474.

[7] S.L. Giles, et al., Whole-body diffusion-weighted MR imaging for assessment of treatment response in myeloma, Radiology 271 (2014) 785–794, https://doi.org/10.1148/radiol.13131529.

[8] C. Messiou, et al., Guidelines for acquisition, interpretation, and reporting of whole-body MRI in myeloma: myeloma response assessment and diagnosis system (MY-RADS), Radiology 291 (2019) 5–13, https://doi.org/10.1148/radiol.2019181949.

[9] S.V. Rajkumar, et al., International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma, Lancet Oncol. 15 (2014) e538–e548, https://doi.org/10.1016/S1470-2045(14)70442-5.

[10] Myeloma Diagnosis and Management. NICE (NG35) and Appendices, <*https://www.nice.org.uk/guidance/ng35*> (October 2018).

[11] M.D. Blackledge, et al., Assessment of treatment response by total tumor volume and global apparent diffusion coefficient using diffusion-weighted MRI in patients with metastatic bone disease: a feasibility study, PLoS One 9 (2014), e91779, https://doi.org/10.1371/journal.pone.0091779.

[12] J.P. O'Connor, et al., Imaging biomarker roadmap for cancer studies, Nat. Rev. Clin. Oncol. 14 (2017) 169–186, https://doi.org/10.1038/nrclinonc.2016.162.

[13] J.M. Bland, D. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307–310, https://doi.org/10.1016/S0140-6736(86)90837-8.

[14] D.-M. Koh, et al., Reproducibility and changes in the apparent diffusion coefficients of solid tumours treated with combretastatin A4 phosphate and bevacizumab in a two-centre phase I clinical trial, Eur. Radiol. 19 (2009) 2728–2738, https://doi.org/10.1007/s00330-009-1469-4.

[15] N.P. Jerome, et al., Repeatability of derived parameters from histograms following non-Gaussian diffusion modelling of diffusion-weighted imaging in a paediatric oncological cohort, Eur. Radiol. 27 (2017) 345–353, https://doi.org/10.1007/s00330-016-4318-2.

[16] J.M. Winfield, et al., Extracranial soft-tissue tumors: repeatability of apparent diffusion coefficient estimates from diffusion-weighted MR imaging, Radiology 284 (2017) 88, https://doi.org/10.1148/radiol.2017161965.

[17] M.D. Blackledge, et al., Noise-corrected, exponentially weighted, diffusion-weighted MRI (niceDWI) improves image signal uniformity in whole-body imaging of metastatic prostate cancer, Front. Oncol. 10 (2020) 704, https://doi.org/10.3389/fonc.2020.00704.

[18] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241, https://doi.org/10.48550/arXiv.1505.04597.

[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, https://doi.org/10.48550/arXiv.1512.03385.

[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105, https://doi.org/10.1145/3065386.

[21] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034, https://doi.org/10.48550/arXiv.1502.01852.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958, https://doi.org/10.5555/2627435.2670313.

[23] D.P. Kingma, Ba J. Adam, A Method for Stochastic Optimization, 2014, https://doi.org/10.48550/arXiv.1412.6980 *arXiv preprint arXiv:1412.6980*.

[24] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, 2015, https://doi.org/10.1002/9780470316849.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017, https://doi.org/10.5555/3298023.3298188.

[26] M. Tan, Q.V. Le, Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks, 2019, https://doi.org/10.48550/arXiv.1905.11946 *arXiv preprint arXiv: 1905*.11946.

[27] Y.B. Hagos, P.L. Narayanan, A.U. Akarca, T. Marafioti, Y. Yuan, Concorde-net: cell count regularized convolutional neural network for cell detection in multiplex immunohistochemistry images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 667–675, https://doi.org/10.48550/arXiv.1908.00907.