Tina Memo No. 2014-007
Internal, IMI preliminary work

Multi-site Liver Tumour ADC Reproducibility at 1.5 T.

Hossein Ragheb, Neil A. Thacker, Ryan Pathak, David M. Morris and Alan Jackson

Last updated
16 / 01 / 2015

Centre for Imaging Sciences,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Multi-site Liver Tumour ADC Reproducibility at 1.5 T.

Abstract

This work assesses the reproducibility of ADC measurement for data acquired across several clinical sites within IMI QuIC-ConCePT project. We present a model for expected ADC reproducibility which takes account of the initial ADC distribution within tumours and the volume of measurement. We show that the accuracies of ADC currently achieved are on average 7.5% but that better measurements are generally associated with increasing the measurement volume. Our analysis generates methods which are capable of predicting the reproducibility of individual tumours, and is therefore suitable for guiding the region of interest selection process. Overall performance of reproducibility for the 'on scanner' averaged acquisition (protocol A) is found to be currently insufficient for detection of change in individuals, but recent results regarding the expected improvement using 'off scanner' averaging (protocol B) would suggest the possibility of patient specific adjustment of therapy. Real tumour data for protocol B data is now required in order to confirm this expected benefit and such data is planned to be acquired as part of the upcoming BOS2 trial.

1 Introduction

Following successful treatment, dense tumour cells are expected to die and become necrotic, increasing ADC from approximately 80-100 10^{-5} mm^2/s to near that of water (180-220 10^{-5} mm^2/s). The rational for using ADC to assess the affects of drug treatment is to be able to detect early responses (of the order of days rather than weeks). In such time periods we do not expect a whole tumour to respond, nor do we expect a responding region to become entirely necrotic. In order to predict a typical level of change in ADC we assume that only 20% of the tumour will respond and this region will have an increase in ADC of around 50%. Thus the average change of ADC within the entire region is expected to be around 10%. This level of change is consistent with published figures [1, 2]. Clearly, these approximations have potential for high levels of variability and in specific cases we may see more or less change. If we were to wait longer between scans, in order to increase the size of effect, it may be easier to see an effect. But if we wait too long (so that the structure of the tumour has also begun to change) we obviate the usefulness of ADC entirely, as this contradicts the idea of detecting an early response. For this work, we will therefore continue to assume that we wish to detect a 10% change in mean ADC and to do this with a reliability at 90% or better[1], we will need a target reproducibility around 3-4 %.

1.1 Data

The data-sets available for this study were the IMI QuIC-ConCePT technical validation data. There were 20 liver tumour patient data-sets acquired using Siemens, GE and Phillips scanners at four sites A, B, C and E with 5 patients per site. In order to test reproducibility of ADC in the liver tumour, each patient was scanned twice within a week.

Two protocols are considered in the IMI project: A and B. For each scan, protocol A data consist of 40 ('on scanner') averaged image slices corresponding to each of the three b-values, namely 100, 500 and 900 s/mm^2. Protocol B, on the other hand, provides 12 sets of these 40 image slices per b-value, which correspond to 4 repeated acquisitions with 3 separate diffusion gradient directions. These repeat data and gradient directions can be aligned and combined 'off scanner'. Simple averaging of the raw data without alignment is expected to generate data equivalent to protocol A. In this paper, we only analyse protocol A data. The study however predicts how a combination of protocol B data and retrospective motion correction could be expected to improve the ADC reproducibility.

2 Methods

2.1 ROI Annotation

A clinician annotated various definitions of region of interest (ROI) on the 20 patient data-sets under study:

[1] A 2.5-3.0 standard deviation effect

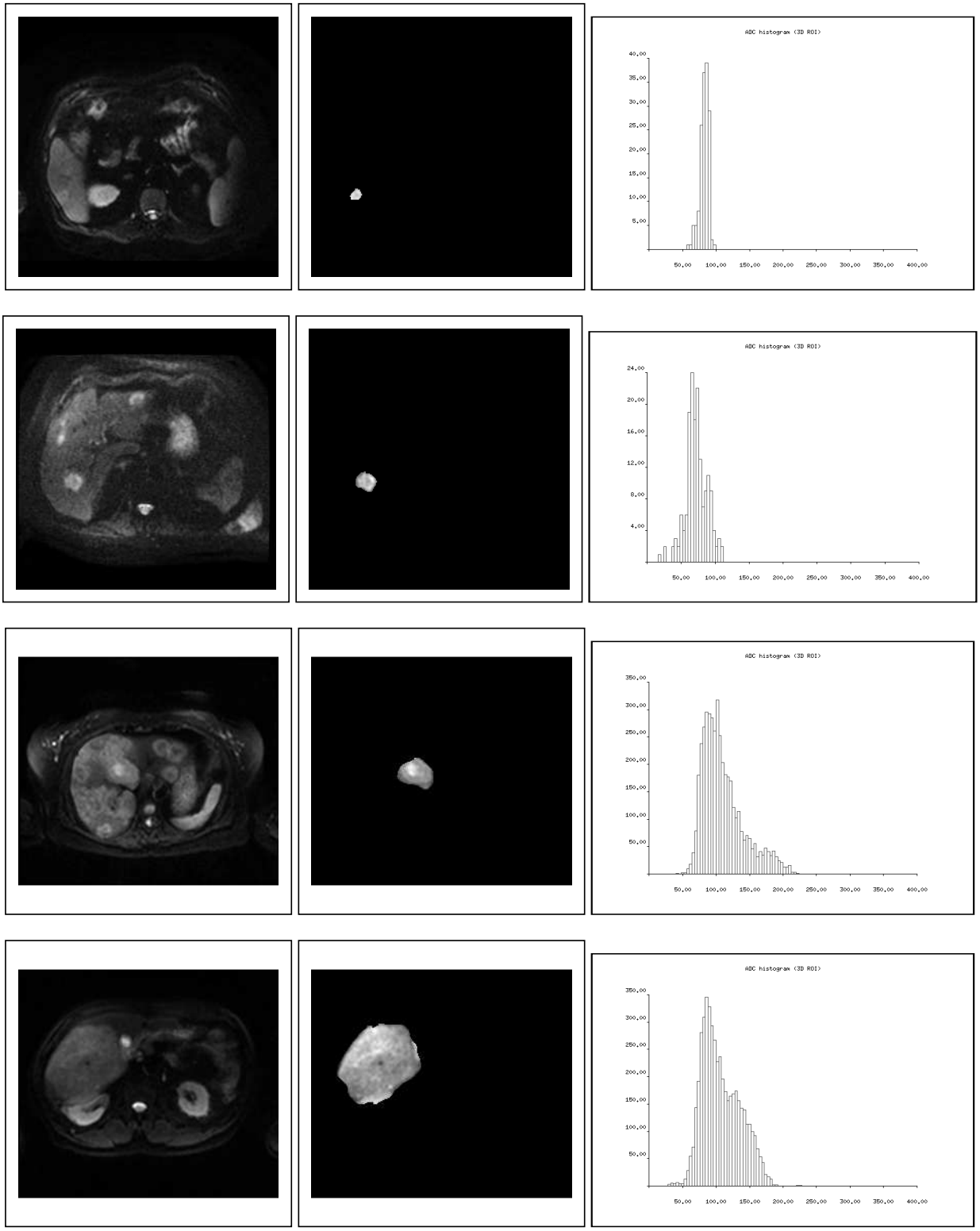

Figure 1: Sample image slice from the b=100 s/mm^2 data (left), selected 2D tumour ROI on the b=100 s/mm^2 data (middle) and ADC histogram corresponding to the 3D tumour ROI (right); data-sets from top to the bottom: B-V2 (baseline 2), A-V1 (baseline 1), E-V6 (baseline 2) and C-V2 (baseline 1); see Table 1 for details.

- the 3D tumour ROI corresponding to tumour tissue (including non-necrotic and necrotic regions) observed on several adjacent slices (3D);
- the largest 2D tumour ROI (on a single slice) was identified from the 3D tumour ROI (2D-L);
- the most representative solid tumour ROI (on a single slice) was identified from the 3D tumour ROI (2D-T);

- the normal tissue 2D ROI on the liver from a single slice location where good image quality was found, as a fixed size circular region (2D-N).

During the process of ROI annotation the clinician also identified the data-sets which were adversely affected by apparent motion. These data could have been rejected earlier during QC, but were included here in order to quantify the consequences of motion on measurement. In most cases only one of the two baseline data-sets was affected, but this is sufficient to result in poor reproducibility.

In Figure 1, we show sample abdominal image slices together with the corresponding annotated 2D ROIs and the ADC histogram for the whole tumour for patients from different sites (scanners). This figure illustrates some of the possible differences in the shape and size of liver tumours and corresponding image quality. The histograms corresponding to the ROI's show a variety of distribution widths and shapes, however, (as we will explain below) the accuracy of the ability to estimate a mean or median is also determined by the number of histogram entries (total voxel volume).

2.2 ADC Fitting

For a given image slice location, at each image pixel we fit the signal values from the corresponding b-value image slices to an exponential curve. The decay parameter of the resulting fit is the ADC for that pixel [2]. Specifically in clinical data, as the noise distribution is skewed, we find that a first order bias correction factor improves the quality of fit [6] as it removes the SNR dependent bias.

We estimate ADC, referred to as D, through a likelihood-based parameter optimisation $\log P(I|D,S0)$ (the probability of the image data given the assumed parameters).

$$\log P(I|D,S0) = \frac{-1}{2\sigma^2}\sum_{b_k}[I(b_k) - f(b_k,D,S0)]^2 \;+\; const \tag{1}$$

where $f(b_k,D,S0)$ is the theoretical value of the bias corrected exponential function and $I(b_k)$ is the signal value from the b-value image pixel. f is a function of b-value b_k and the current estimates of ADC D and no-diffusion signal $S0$ (at $b_k = 0$), and is computed using

$$f^2(b_k,D,S0) = S0^2\exp(-2b_kD) + \alpha\sigma_I^2 \quad ; \quad k \in [1,2,3] \tag{2}$$

where $k \in [1,2,3]$ refers to the three b-values b_k used, e.g. 100, 500 and 900 s/mm^2. The signal value for no diffusion (b=0) $S0$ is the second parameter which is estimated. α is a fixed value defining the amount of bias correction applied and it may be adjusted depending on the amount of image smoothing corresponding to the specific imaging protocol used by the scanner (for our data α was set to the theoretical value of one). An estimate of the standard deviation (SD) of noise in the image σ_I is computed from the distribution of second derivatives (for x and y) around zero [7], in a central rectangular region on the tissue.

Our ADC measurement software has been implemented on the QuIC-ConCePT platform and has become available to all sites within the project to use. Once ADC values are estimated for individual voxels in the region of interest, we construct the ADC histogram corresponding to a 2D ROI (single slice location) or a 3D ROI (several 2D ROIs belonging to the same tissue definition, e.g. tumour). We then extract from each ADC histogram, the mean ADC, the median ADC, the standard deviation (width of the ADC distribution) and the number of histogram entries.

2.3 Measuring Change in ADC

The ADC reproducibility study is performed to quantify how close we can get to the ideal condition where no change is seen between an ADC metric (such as mean or median) between baseline 1 and 2. To study change in ADC, one method is to plot the absolute difference in the ADC metric $(D_1 - D_2)$ between baselines 1 and 2 against their average $(D_1 + D_2)/2$. This is a common approach which appears in recent publications such as [9].

Here, we propose an alternative method for studying reproducibility using percentage change in the ADC metric given by

$$R_{12} = 2\frac{(D_1 - D_2)}{(D_1 + D_2)} \times 100 \tag{3}$$

For a proportional error dependency on D, we expect the variable R_{12} to be a more consistent statistical summary of difference than $(D_1 - D_2)$. However, when attempting to understand measurement processes the reproducibility may also be affected more by other parameters. We expect effects due to signal to noise, imaging artefacts and

subject motion, but we also expect factors due to our analysis strategy, such as region of interest selection and the measurement volume. Although group level analysis (such as a t-test) can be performed without explicitly identifying sources of measurement error, for a single subject, understanding any observed changes between two measurements requires that we take these factors into account, if we do not wish to interpret instabilities in measurement as a significant biological change.

Contributions to errors on R_{12} ($\epsilon_{R_{12}}$) can be estimated from error propagation (see Appendix A for details), based upon the expected errors on D_1 and D_2 (σ_{D_1} and σ_{D_2}) as

$$\epsilon_{R_{12}}(\sigma_{D_1}, \sigma_{D_2}) = \frac{400\sqrt{D_1^2\sigma_{D_2}^2 + D_2^2\sigma_{D_1}^2}}{(D_1 + D_2)^2} \tag{4}$$

Estimates of σ_D are in turn dependent upon the signal to noise and the way summary statistics (e.g. ADC mean or median) are obtained from a region of interest. Consequently, we expect the usual $\sqrt{N}$ dependency on the region size N. Otherwise, for approximately fixed fitting accuracies (i.e. the signal to noise and fitted parameters are all similar) these values would be constant σ_{fix}. However, for varying signal to noise or tissue heterogeneity the width of the ADC distribution will vary. These inaccuracies will be further increased by subject motion in heterogeneous regions, with the obvious consequences for the accuracy of the mean or median ($\sigma_D \propto SD(D)/\sqrt{N}$).

We wish to understand the overall behaviour of reproducibility and to do this we fit a general model to our data to find the approach which is most predictive of reproducibility measurements. As measurement error is expected to contain significant variations caused by region size, for visualisation and quantitation of measurement behaviour it is better to plot reproducibility of measurement R against region size N or $\log(N)$ than to use the conventional Baland-Altman plot. We assume that $\sigma_{D_i} = SD(D_i)/\sqrt{N'}$ where $\sigma(D_i)$ is a measure of the width of the ADC distribution and N' is the number of independent measurements in the region[2]. In order to help interpret the effects of motion we identify those points associated with datasets which exhibit large amount of apparent motion. In addition we expect systematic effects (such as measurement inhomogeneity [8] or calibration drift effects) to contribute a fixed proportional error due to an inability to exactly replicate equivalent image data on repeated attempts (ϵ_{sys}).

We are left with three parameters with which to describe observed reproducibility; a fixed fitting error σ_{fix} , terms proportional to ADC width σ_D and the systematic error ϵ_{sys}. This is done by fitting the data seen in the reproducibility size plot (excluding known movement outliers) so that the error formula

$$\epsilon_{R_{12}}^2 = \beta^2\epsilon_{R_{12}}^2(\sigma_{D_1}, \sigma_{D_2}) + \epsilon_{R_{12}}^2(\sigma_{fix}, \sigma_{fix}) + \epsilon_{sys}^2 \tag{5}$$

has a pull distribution ($\frac{R_{12}[j]}{\epsilon_{R_{12}[j]}}$) with unit standard deviation. In practice, we can minimise the likelihood cost function (see Appendix B)

$$-\log L = \sum_{j=1}^{J}\left[\frac{1}{2}\left(\frac{R_{12}[j]}{\epsilon_{R_{12}[j]}}\right)^2 + log(\sqrt{2\pi}\ \epsilon_{R_{12}[j]})\right] \tag{6}$$

in order to estimate the three parameters β, σ_{fix} and ϵ_{sys}, where j is the sample number for the whole population J (in our case $J = 60$, i.e. 15 samples with insignificant motion from each of the 4 sets of ROIs annotated).

For the regions identified in this study, this error model is plotted against $\log(N)$ in Figure 5 (for mean or median ADC). Using this analysis we can calculate the expected error on percentage change in ADC for any region of a given size and ADC distribution (i.e. for each test).

3 Results

In Tables 1-4, we tabulate the mean and median ADC measurements extracted from the ADC histograms corresponding to various ROI definitions. In each table, we also list the width of the ADC distribution and the number of ADC histogram entries (this number is equivalent to the number of voxels in the ROI if there is no fit failures). Finally, in each table, we have computed the ADC percentage change R_{12} for the mean and median ADC.

[2]This estimate of error is strictly true only for random samples from Gaussian distributions. Figure 1 shows that typical ADC distributions are significantly non-Gaussian. However, the same approach works for any distribution for very large samples in accordance with the central limit theorem. N' will not be the number of voxels N due to the way image data is generated from k-space data during image acquisition, but can be assumed to be proportional ($N' = N/\beta$).

Using the mean ADC measurements tabulated in Tables 1-4 for baselines 1 and 2, we plot the Bland-Altman plot in Figure 2. Bland-Altman plots assess the reproducibility of a parameter over a large range of values for that parameter. The trend we expect to see in the Bland-Altman plot is a linear dependency of reproducibility of measurement with ADC, this is because the individual mechanisms which generate errors on ADC estimation are generally proportional to ADC. However, for this data we see only a 20% range of ADC values, far too small to empirically determine an error dependency. A similar pattern is seen in the Bland-Altman plot for the median ADC measurements in Figure 3.

We used 60 samples as data to fit the error model (Eq. 5) by minimising the log-likelihood cost function (Eq. 6). These are 15 samples from 4 groups of ROI's annotated on baseline 1 and 2 of 15 patients. The 5 patient data-sets which were not used for fitting the error model had considerable motion in one or both of their baseline data. We can put the groups together and use as independent samples because at the level of percentage change in ADC in specific ROI R_{12}, they can be treated as independent samples. After minimisation of the cost function, the parameters in the error model are found to be: $\beta = 4.87$, $\sigma_{fix} = 69.35$ and $\epsilon_{sys} = 2.65$ (giving a value about 59.99 for the corresponding χ^2). In Figure 11, the contribution of the first term of the error model is plotted against that of the second term, i.e. $\beta\epsilon_{R_{12}}(\sigma_{D_1}, \sigma_{D_2})$ against $\epsilon_{R_{12}}(\sigma_{fix}, \sigma_{fix})$. The plot shows that in most cases, the first term has a larger contribution to the total error compared to the second term. Also, the value of the third term $\epsilon_{sys} = 2.65$, i.e. the systematic error, is quite small for the majority of 60 samples shown in the scatter plot of this figure. This is expected because in ideal conditions any systematic errors in the original measurements should have cancelled following the calculation of the percentage difference. Also, if we remove the second term from the error model (i.e. setting σ_{fix} to zero), we find that $\beta = 5.48$ and $\epsilon_{sys} = 3.89$.

We cannot use the χ^2 from these likelihood fits directly to test the models, as the fitting process guarantees a value in each case of 60. Instead we test the suitability of the error model for each of the 4 ROI groups using the alternative χ^2 (by omitting the second term, as described in Appendix B). Results for the 3-parameter error model are as follows. For the 3D tumour ROI $\chi^2 = 11.33$, for the 2D-L (largest 2D tumour ROI) $\chi^2 = 10.55$, for the 2D-T (most representative 2D solid tumour ROI) $\chi^2 = 24.30$, and for 2D-N (2D normal liver ROI) $\chi^2 = 13.83$. These are expected to correspond to a χ^2 with 15 degrees of freedom (DoF). From these numbers we can argue that this is an acceptable error model within an approximate 10% deviation, i.e. clinical management decisions made using this estimate of measurement error would not be signifcantly affected at this level of conformity. When a 2-parameter error model is fitted, the results are as follows. For the 3D ROI $\chi^2 = 8.79$, for the 2D-L $\chi^2 = 9.28$, for the 2D-T $\chi^2 = 21.94$, and for the 2D-N $\chi^2 = 19.99$. This is marginally worse than the 3-parameter model, but may still be considered adequate for this data.

An alternative approach is to assume the conventional form of measurement error, i.e. simply constant across all samples. For our model this is equivalent to using only the systematic error and ignoring the two statistical terms in Eq. (5). For such an approach, optimisation gives $\epsilon_{sys} = 8.93$ (with a value about 59.93 for the corresponding χ^2 with 59 DoF). Again, for individual groups of ROIs, the 3D ROI gives $\chi^2 = 4.86$, the 2D-L gives $\chi^2 = 11.82$, the 2D-T gives $\chi^2 = 26.76$, and the 2D-N gives $\chi^2 = 16.49$ (corresponding to a χ^2 with 14 DoF). Thus while both the first two models which account for area and ADC distribution are acceptable descriptions, the hypothesis for a more conventional single value error model is rejected.

We see in Figure 5 that the effects of size from largest to smallest region induce a factor of 1000% difference in estimated error. As the effects of ADC seen in the original Bland-Altman plot are expected to be less than 20% over the range of ADC values seen, this confirms our initial assertion that the effects of size on reproducibility are much more pronounced than any dependency on ADC. In Figures 6 and 9, the percentage change scaled to the estimated error is plotted against ROI volume. Although it is difficult to attribute the outliers in Figures 4 and 8 to motion, once the effects of measurement accuracy have been properly accounted for the majority of outliers are indeed associated with motion (Figure 6 for mean ADC and Figure 9 for median ADC). The extra instability in ADC measurement due to motion is evident in all data except for the homogeneous normal tissue. This illustrates the benefits of better understanding the sources of error as poor repeat measurements can be correctly attributed to motion as opposed to (for example) region size. Further, after scaling each percentage change metric by its estimated error, data points in the corresponding plot are directly comparable and expected to follow a t-distribution. Excluding motion outliers (squares), 95% of data points fall within $t = 2.145$, consistent with a t-distribution with 14 DoF. This is shown in Figures 6 and 9 using two dashed lines at ± 2.145.

To investigate whether using 3D ROIs results in improved ADC reproducibility (compared to the 2D ROIs), in Figures 7 we plot the scaled percentage change in the mean ADC (and in Figure 10, in the median ADC), where the vertical axis corresponds to the 3D ROI and the horizontal axis to the 2D ROIs. Excluding motion outliers (squares), data points are scattered almost uniformly (with equal variance) around the diagonal line, suggesting that beyond the obvious effects of sample size there is no significant advantage to be gained by using 3D rather than 2D ROIs.

4 Discussion

This is a multi-centre project which aims to demonstrate the feasibility of making ADC measurements across a variety of vendor platforms. Understanding the reproducibility of measurements therefore requires us to make some statements regarding the consistency of accuracy for different machines. Basic results regarding between scanner consistency for normal liver tissue have been published previously [3, 4]. Unfortunately, as we will see, the quantity of data available here is not sufficiently large to perform a comparison with any great accuracy. The various factors affecting reproducibility are too large to allow such comparison, particularly for a small sample size (≤ 5). Instead, we fit an average theoretical model of percentage ADC error based upon statistical sampling theory, which takes into account the specific circumstances of ADC distribution and volume in each tumour, in order to predict expected accuracy for individual samples. These predictions of accuracy then provide a more accurate basis for the assessment of the measurements which were made on different machines by allowing us to factor out some of the differences between measurements which arise due to predictable measurement effects (e.g. SNR and region size). By taking into account the processes known to affect reproducibility we can then ask if the reproducibility data corresponding to individual scanners are better or worse than that predicted by the average model.

In the ideal case (for independent ADC samples) we believe the value of β should be 1.0. The value of β seen implies that we need of the order of 25 ADC fits (β^2) from these data to obtain one independent measurement. This is presumably a consequence of the processing used during image reconstruction to suppress the visual appearance of noise (e.g. up interpolation and k-space filtering)[3].

Figure 12 shows that data from each site is well described by the current average error model, while Figure 13 shows that data from site E was marginally more accurate than the other sites, but this factor is no more than 20% and could therefore be due to the small sample size (<=20).

With respect to other work in this field, generally reproducibility figures would be assessed using a simple coefficient of variation, which in this case is about 7.5%. however we can see here that such an approach would fail to identify the genuine behaviour of measurement error which can vary by a factor of two on a case by case basis around this average. We would predict improvements in statistical efficiency for any analysis if a case by case estimate of measurement reproducibility were properly accounted for. In addition, 'region-wise' analysis is often performed by taking signal averages across ROIs rather than constructing a histogram of multiple 'voxel-wise' ADC fits. The former approach would be incapable of supporting a data specific error estimate.

We can now use our model of accuracy to assess the degree of correlation seen between the alternative region selection methods. We see in Figure 7 for mean ADC and Figure 10 for median ADC that in all cases a strong correlation in reproducibility differences is present (i.e. all methods seem to summarise the same information within expected errors). We can thus say that beyond the increase in accuracy seen with using larger numbers of voxels, and the exclusion of datasets strongly affected by motion, there seems to be no inherent advantage to using any specific strategy for selecting a region within a tumour. Provided that equivalent regions can be reliably identified between two scans, we only need to identify a sufficient number of voxels with which to measure expected change.

In terms of the general (average) level of reproducibility across sites, Protocol A data fails to meet our target reproducibility of 3-4% for the majority of measurements made in this study. Currently only the largest region sizes (whole liver tumours) approach this figure. In order to obtain this level of accuracy for all data we need to generally improve performance (even for data not affected by motion artefact). For a factor of two increase in accuracy (see below) the target accuracy will be reached for data sets with a region size roughly greater than $e^{7.6} = 2000$ (Figure 5). This suggests a strategy for measuring tumour data which achieves a minimum number of samples in order to obtain a specified level of accuracy. For instance, one might combine several metastatic regions to obtain enough entries for the single corresponding ADC histogram. In addition, the observation of a significant variation in the stability of measurements with ROI volume has implications for the designs of clinical studies, in terms of both statistical efficiency of analyses and also reliability of biological conclusions.

For larger tumour regions we can say that we are within a factor of two of the required performance. However, motion was always predicted to be a large source of unwanted uncertainty for ADC measurement in the region of the liver, and for this reason we have specified an alternative protocol B, in order to support retrospective motion correction. The main benefits of motion correction are to reduce the frequent elimination of poor data during quality control, however there is the possibility that reproducibility will also show general improvement. Our model for error estimation in regional tumour ADC measurements contains terms relating to the standard deviation of regional ADC value, on average this value was approximately 32 10^{-5} mm^2/s for this data. In previous

[3]This is a consequence of a poorly understood statistical fact; we cannot increase the information quantity of a region in an image simply by smoothing it. It may appear that we have reduced the noise, but we have simply converted independent noise into correlated noise.

work we have shown that motion correction will reduce this distribution width by 23 10^{-5} mm^2/s. The fractional improvement in ADC reproducibility is therefore predicted to be of the order of 30%. Unfortunately, the only sure way to assess this approach is to obtain protocol B data in a multi-centre study.

5 Conclusions

This study has generated a mathematical model which can be used to estimate the reproducibility of regional tumour ADC measurements in individual data sets. The same approach seems applicable across all region of interest selection methods and scanners from different vendors. These calculations show that the variations in expected reproducibility are too significant to ignore, particularly when assessing change in a single tumour. The Results also show that motion effects are currently significant, with as much as 20% of all data failing to be of sufficient quality to support a reliable measurement. The overall accuracy of the remaining data is then within a factor of two of the level which would be required to reliably detect a 10% change in mean or median ADC. We believe that this level of reproducibility might be attained by using 'off scanner' (retrospective) motion correction combined with ROI's chosen to be roughly greater than 2000 voxels.

References

[1] B. Turkbey, O. Aras, N. Karabulut, et al, "Diffusion-Weighted MRI for Detecting and Monitoring Cancer: A Review of Current Applications in Body Imaging", *Diagnostic Interventional Radiology, Turkish Society of Raiology*, 18(1):46-59, 2012.

[2] A.R. Padhani, G. Liu, D. Mu-Koh, T.L. Chenevert, H.C. Thoeny, T. Takahara, A. Dzik-Jurasz, B.D. Ross, M. Van Cauteren, D. Collins, D.A. Hammoud, G.J.S. Rustin, B. Taouli and P.L. Choyke, "Diffusion-Weighted Magnetic Resonance Imaging as a Cancer Biomarker: Consensus and Recommendations", *Neoplasia*, 11(2):102-125, 2009.

[3] H. Ragheb, N.A. Thacker, D.M. Morris, N.H.M. Douglas and A. Jackson, "Predicting the Quantitative Accuracy of In-Vivo ADC using an Ice-Water Phantom", *Proc. Int'l Soc. Magnetic Resonance in Medicine*, pp. 2650, 2014.

[4] H. Ragheb, N.A. Thacker, D.M. Morris, and A. Jackson, "Interpreting Ice-Water Phantom Data for Prediction of Clinical ADC Measurement", *Tina memo* 2013-005, 2013, *http://www.tina-vision.net/docs/memos/2013-005.pdf*.

[5] N.A. Thacker, J.V. Manjon and P.A. Bromiley, "A Statistical Interpretation of Non-Local Means", *Proc. VIE 2008*, pp. 250-255, 2008.

[6] H. Gudbjartsson and S. Patz, "The Rician Distribution of Noisy MRI Data", *Magnetic Resonance in Medicine*, 34(6):910-4, 1995.

[7] N.A. Thacker, "Useful Image Processing Methods", *Tina memo* 2008-010, 2008, *http://www.tina-vision.net/docs/memos/2008-010.pdf*.

[8] D. I. Malyarenko, B. D. Ross and T. L. Chenevert, "Analysis and Correction of Gradient Nonlinearity Bias in Apparent Diffusion Coefficient Measurements", *Magnetic Resonance in Medicine*, 71:13121323, 2014.

[9] F. Deckers, B. De Foer, F. Van Mieghem, T. Botelberge, R. Weytjens, A. Padhani and M. Pouillon, "Apparent diffusion coefficient measurements as very early predictive markers of response to chemotherapy in hepatic metastasis: A preliminary investigation of reproducibility and diagnostic value", *J Magn Reson Imaging*, 40(2):448-56, 2014.

Appendix A

For mean parameters x and y with accuracies here in each case $\sigma_i = SD/\sqrt{N_i}$ (with SD being the corresponding standard deviation and N_i being the corresponding number of samples), we wish to find the accuracy on the parameter z

$$z = 200\frac{(x-y)}{(x+y)} \tag{7}$$

Using error propagation, the accuracy on z is given by

$$\epsilon_z^2 = \sigma_x^2 \left(\frac{dz}{dx}\right)^2 + \sigma_y^2 \left(\frac{dz}{dy}\right)^2 \tag{8}$$

The derivatives of z with respect to x and y are

$$\frac{dz}{dx} = \frac{400y}{(x+y)^2} \qquad ; \qquad \frac{dz}{dy} = \frac{-400x}{(x+y)^2} \tag{9}$$

Hence, we can write

$$\epsilon_z = \frac{400}{(x+y)^2}\sqrt{y^2\sigma_x^2 + x^2\sigma_y^2} \tag{10}$$

Appendix B

In order to construct a likelihood for multiple repeat samples (N), based on the difference between the expected and observed variances, we assume an approximate Gausssian distribution for the difference $y_{1i} - y_{2i}$

$$log\ L\ =\ \sum_i^N (y_{1i}\ -\ y_{2i})^2/\sigma_{yi}^2\ +\ log(\sqrt{2\pi}\sigma_{y_i}) \tag{11}$$

The last term is normally ommitted from fitting routines (on the basis that is is constant), but is needed if the value of σ_{y_i} is allowed to vary as part of the fit (as here).

The resulting log likelihood cannot be treated as a χ^2 statistic as the fitting process gaurantees that $\chi^2 = N$. However, once the error model is determined we can assess the adequacy of the model for describing sub groups g, with

$$\chi^2\ =\ \sum_{i\in g}^N (y_{1i}\ -\ y_{2i})^2/\sigma_{yi}^2$$

Notes

Some may say that the approach we have taken to error analysis (modelling individual contributions to errors and using transformations of variables to control distributions and construct homoscedastic variables), is overblown and unnecessary. Our response would be as follows. Once we have decided that there is a more stable statistical analysis which can be supported by the available quantity of data, the only way to assess the consequences is to apply it. Any differences between approaches then naturally leads to suspicions regarding the simpler method, i.e. we must abandon these conclusions. When working on many data analysis tasks involving these options, we realise that there is little point wasting time with the simple method if a better one is available, even knowledge of the answers we would obtain become irrelevant, except for purposes of issuing a caution to others. In this case the caution would be that the variability of reproducibility for ADC measurements is too complicated to be summarised by one average figure per scanner.

data-set	Mn-B1	Md-B1	Mn-B2	Md-B2	SD-B1	SD-B2	N-B1	N-B2	Mn-R%	Md-R%
A-V1	83.61	83.58	69.41	71.21	16.55	21.51	1143	1138	18.55	15.98
A-V2	90.50	90.98	113.13	108.01	26.07	29.24	3349	3077	-22.22	-17.11
A-V3	97.16	94.29	96.03	96.21	31.16	34.67	2848	2842	1.16	-2.01
A-V5	82.86	84.82	71.52	72.25	21.98	22.47	1298	1296	14.69	16.01
A-V6	96.68	96.41	100.01	99.18	25.37	35.59	618	589	-3.39	-2.83
B-V2	85.92	85.60	88.11	86.97	16.19	10.69	580	566	-2.51	-1.58
B-V3	124.74	127.28	121.97	126.20	41.68	45.99	1527	1768	2.24	0.85
B-V4	106.75	102.76	98.93	97.96	26.42	19.55	3154	3201	7.60	4.78
B-V5	93.20	90.35	97.34	94.48	37.18	42.57	4622	4556	-4.34	-4.46
B-V6	104.26	90.25	102.51	91.56	40.28	49.06	3767	3694	1.69	-1.44
C-V1	145.31	141.73	136.19	126.04	49.28	46.91	5967	5948	6.47	11.71
C-V2	96.15	90.30	108.57	101.21	30.00	31.64	72321	76823	-12.13	-11.39
C-V3	93.54	86.56	92.41	88.75	29.06	20.76	5671	5908	1.21	-2.49
C-V4	89.38	91.76	96.21	95.42	17.68	17.20	262	277	-7.36	-3.91
C-V5	117.24	111.53	118.55	111.57	32.06	34.76	60573	63686	-1.11	-0.03
E-V2	123.83	118.31	130.57	125.30	44.38	45.86	8998	8577	-5.29	-5.73
E-V3	98.50	94.26	97.46	94.88	20.01	14.79	4211	4419	1.06	-0.65
E-V4	129.87	122.57	128.53	121.57	32.42	28.72	2114	2165	1.03	0.81
E-V5	201.21	208.61	195.57	201.27	53.46	52.48	8245	7582	2.84	3.58
E-V6	109.90	101.56	110.24	102.77	30.37	31.03	4228	4380	-0.30	-1.18

Table 1: 3D ROIs on liver tumour (3D): mean ADC (Mn); median ADC (Md); baseline (B); standard deviation (SD); number of ADC fitted voxels (N); percentage change in ADC (R%); ADC units: 10^{-5} mm^2/s; b-values: 100, 500 and 900 s/mm^2.

data-set	Mn-B1	Md-B1	Mn-B2	Md-B2	SD-B1	SD-B2	N-B1	N-B2	Mn-R%	Md-R%
A-V1	72.15	71.44	59.50	58.94	15.95	25.60	169	341	19.21	19.17
A-V2	89.49	90.76	101.20	100.74	26.42	22.28	901	830	-12.28	-10.42
A-V3	104.43	99.38	108.07	98.55	30.93	38.93	478	510	-3.42	0.83
A-V5	84.17	85.66	71.08	73.48	22.92	24.62	339	369	16.86	15.31
A-V6	103.70	104.75	105.84	106.71	31.43	29.48	227	203	-2.04	-1.85
B-V2	83.88	85.07	82.61	83.35	13.00	6.62	155	154	1.52	2.04
B-V3	120.63	129.24	127.80	134.77	41.02	51.14	477	528	-5.77	-4.18
B-V4	115.92	108.51	98.87	96.67	30.99	23.15	986	947	15.87	11.54
B-V5	95.18	89.03	99.30	98.40	48.45	48.86	1078	1038	-4.23	-9.99
B-V6	101.04	87.81	94.75	85.42	41.57	47.16	760	810	6.42	2.75
C-V1	155.32	158.09	142.15	130.36	48.53	46.30	722	775	8.85	19.22
C-V2	107.38	101.93	114.11	108.63	28.25	29.78	4921	4873	-6.07	-6.36
C-V3	96.58	83.88	94.21	88.64	42.45	26.87	1108	1069	2.48	-5.51
C-V4	94.00	94.11	106.93	108.22	11.90	14.11	77	83	-12.87	-13.94
C-V5	124.99	120.45	132.44	124.21	30.93	39.17	4527	4564	-5.78	-3.07
E-V2	134.91	122.25	124.79	107.24	51.43	63.10	1247	1190	7.79	13.08
E-V3	98.33	92.61	93.48	91.09	20.27	13.33	621	625	5.05	1.65
E-V4	134.39	119.87	127.96	120.32	38.99	30.13	398	369	4.90	-0.37
E-V5	192.81	204.67	201.39	212.90	61.68	50.57	1094	1042	-4.35	-3.94
E-V6	122.49	112.96	108.86	97.95	34.63	34.09	995	974	11.78	14.23

Table 2: Largest 2D ROIs on liver tumour (2D-L): mean ADC (Mn); median ADC (Md); baseline (B); standard deviation (SD); number of ADC fitted voxels (N); percentage change in ADC (R%); ADC units: 10^{-5} mm^2/s; b-values: 100, 500 and 900 s/mm^2.

data-set	Mn-B1	Md-B1	Mn-B2	Md-B2	SD-B1	SD-B2	N-B1	N-B2	Mn-R%	Md-R%
A-V1	83.87	83.82	76.14	76.22	13.48	16.79	186	246	9.66	9.49
A-V2	87.46	88.10	110.92	108.79	21.13	23.72	980	806	-23.65	-21.01
A-V3	112.71	112.97	101.72	98.96	21.79	19.00	436	423	10.25	13.22
A-V5	83.10	85.76	70.43	72.73	20.74	19.26	339	337	16.50	16.44
A-V6	82.18	83.26	117.36	121.20	18.79	37.95	189	193	-35.26	-37.11
B-V2	83.85	85.11	81.83	82.92	12.37	7.54	170	167	2.43	2.60
B-V3	120.63	129.24	127.80	134.77	41.02	51.14	477	528	-5.77	-4.18
B-V4	118.96	110.60	105.09	104.67	33.76	16.73	818	826	12.38	5.50
B-V5	88.47	86.39	103.37	96.95	27.20	41.55	954	933	-15.53	-11.51
B-V6	85.96	81.62	68.13	64.17	22.34	32.21	651	570	23.14	23.93
C-V1	152.30	156.00	125.47	121.70	50.61	38.58	744	659	19.31	24.70
C-V2	88.20	85.08	93.98	90.06	22.11	24.52	4639	4847	-6.34	-5.68
C-V3	86.33	81.60	96.01	94.90	18.41	16.55	1007	918	-10.61	-15.07
C-V4	90.84	93.00	106.93	108.22	17.94	14.11	63	83	-16.27	-15.12
C-V5	115.91	109.00	108.33	102.66	25.88	29.83	3966	3931	6.76	5.99
E-V2	112.86	107.21	118.16	121.36	36.30	30.47	1201	922	-4.58	-12.38
E-V3	100.27	97.35	98.81	95.93	18.08	10.46	597	571	1.46	1.46
E-V4	121.14	119.14	121.29	112.57	24.18	26.10	348	307	-0.12	5.67
E-V5	192.81	204.67	201.39	212.90	61.68	50.57	1094	1042	-4.35	-3.94
E-V6	95.61	89.82	105.76	102.84	21.69	21.27	740	782	-10.08	-13.51

Table 3: Most representative solid 2D ROIs on liver tumour (2D-T): mean ADC (Mn); median ADC (Md); baseline (B); standard deviation (SD); number of ADC fitted voxels (N); percentage change in ADC (R%); ADC units: 10^{-5} mm^2/s; b-values: 100, 500 and 900 s/mm^2.

data-set	Mn-B1	Md-B1	Mn-B2	Md-B2	SD-B1	SD-B2	N-B1	N-B2	Mn-R%	Md-R%
A-V1	71.97	73.74	63.06	63.71	17.85	23.55	208	206	13.19	14.59
A-V2	89.23	86.78	100.28	101.04	19.65	15.26	208	208	-11.66	-15.18
A-V3	51.95	52.12	46.27	46.12	15.80	17.70	207	173	11.56	12.21
A-V5	66.88	65.07	67.64	69.30	24.46	18.78	206	203	-1.13	-6.30
A-V6	96.05	95.50	85.66	86.67	22.53	15.83	208	208	11.44	9.69
B-V2	95.13	95.33	100.29	99.51	10.07	9.21	208	208	-5.28	-4.29
B-V3	169.22	168.40	163.19	164.67	26.72	25.05	208	208	3.62	2.23
B-V4	95.96	96.14	94.12	94.12	12.63	10.48	208	208	1.93	2.12
B-V5	83.96	82.95	94.22	95.20	15.08	16.37	208	208	-11.51	-13.75
B-V6	95.32	96.44	80.75	80.20	15.44	18.94	208	208	16.55	18.38
C-V1	81.40	84.53	92.99	92.59	19.96	11.47	208	208	-13.29	-9.10
C-V2	109.41	109.33	108.87	109.33	9.96	12.47	208	208	0.49	0.00
C-V3	113.15	110.18	117.51	117.20	22.92	22.33	208	208	-3.78	-6.17
C-V4	110.74	111.20	105.07	103.30	11.04	10.09	208	208	5.25	7.36
C-V5	124.19	123.36	110.79	111.31	11.20	9.52	208	208	11.40	10.26
E-V2	109.34	108.24	113.98	117.00	21.85	22.34	208	208	-4.15	-7.77
E-V3	106.08	103.00	118.76	118.95	13.04	10.61	208	208	-11.27	-14.37
E-V4	107.08	106.18	105.83	105.45	13.19	11.84	208	208	1.17	0.68
E-V5	106.44	106.50	102.30	101.76	8.33	8.85	208	208	3.96	4.55
E-V6	132.11	132.62	113.78	113.78	12.40	7.43	208	208	14.90	15.29

Table 4: Fixed size 2D ROIs (circular) on liver normal tissue (2D-N): mean ADC (Mn); median ADC (Md); baseline (B); standard deviation (SD); number of ADC fitted voxels (N); percentage change in ADC (R%); ADC units: 10^{-5} mm^2/s; b-values: 100, 500 and 900 s/mm^2.

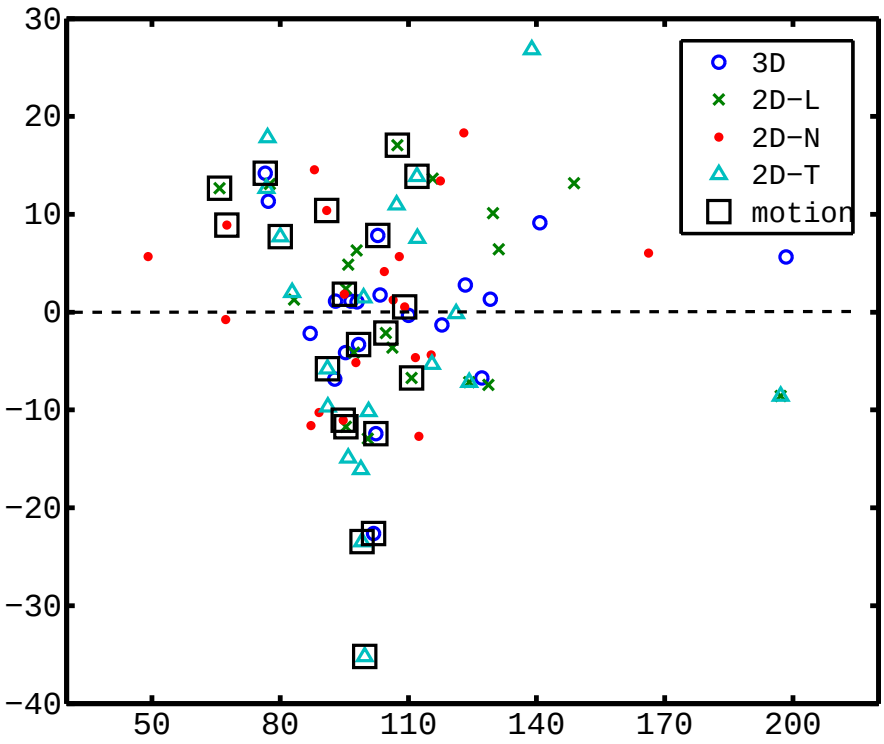

Figure 2: Bland-Altman plot: difference in the mean ADC against the average mean ADC; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion; ADC units: 10^{-5} mm^2/s.

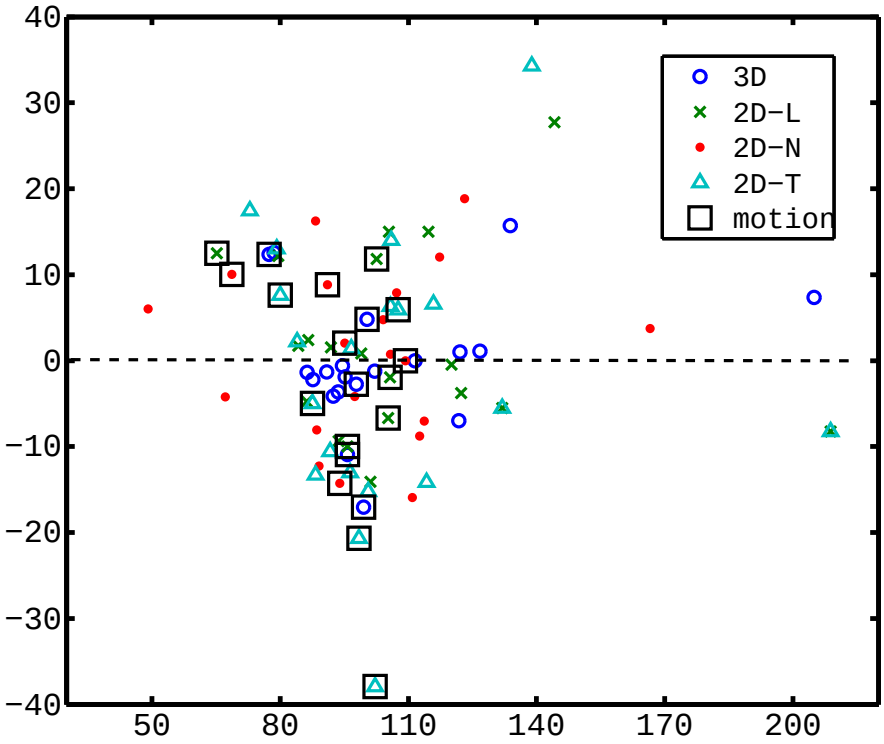

Figure 3: Bland-Altman plot: difference in the median ADC against the average median ADC; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion; ADC units: 10^{-5} mm^2/s.

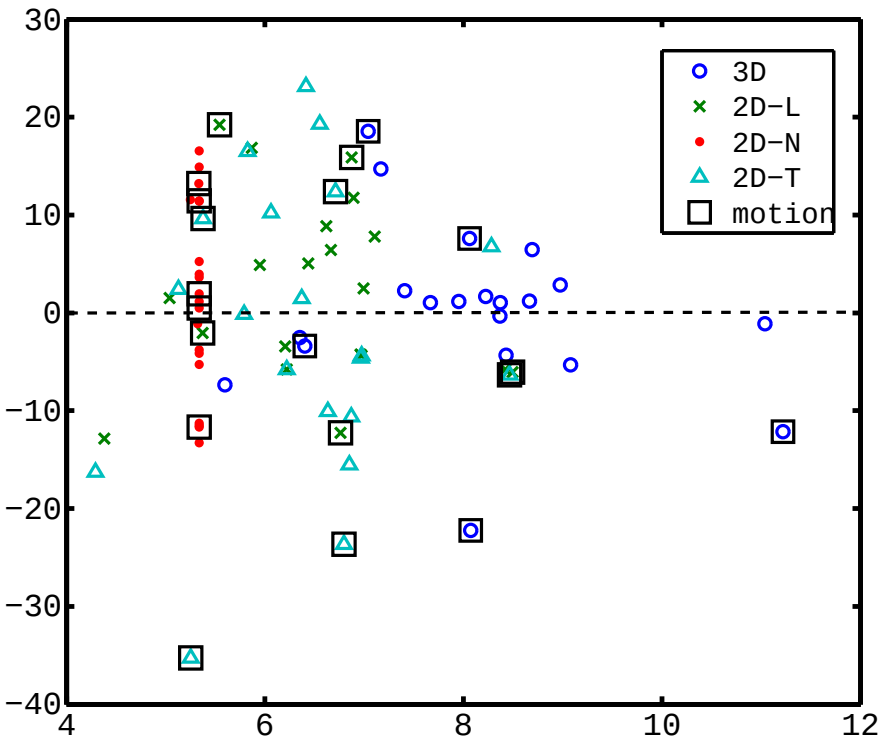

Figure 4: Percentage change in mean ADC against log of the average number of voxels in the ROI; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

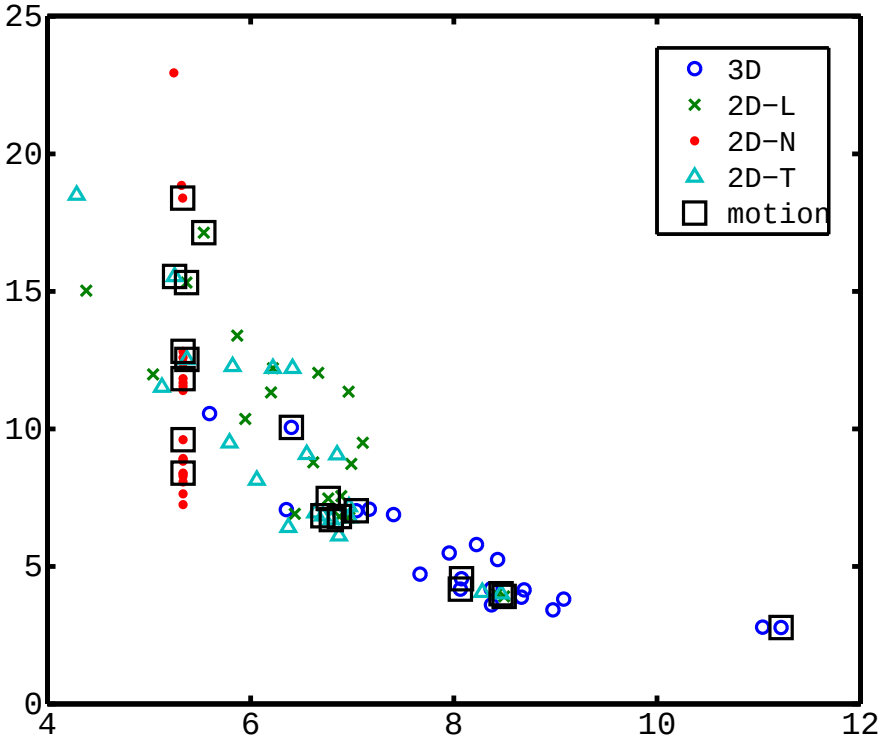

Figure 5: Fitted percentage reproducibility of the mean ADC (based on sample size and distribution width) against log of the average number of voxels; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

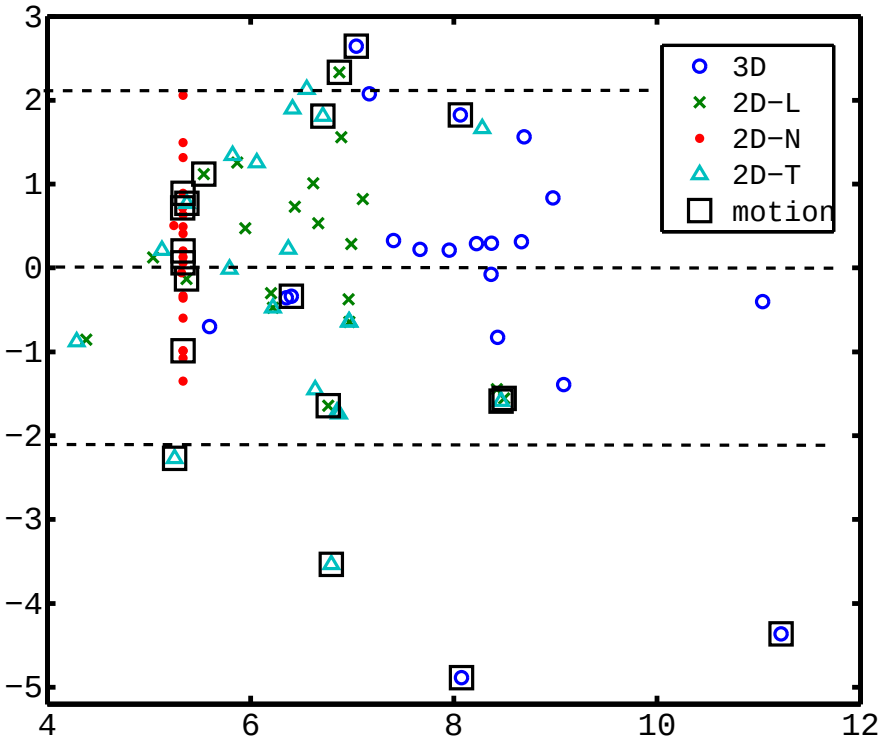

Figure 6: Percentage change in mean ADC scaled by error estimates against log of the average number of voxels; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

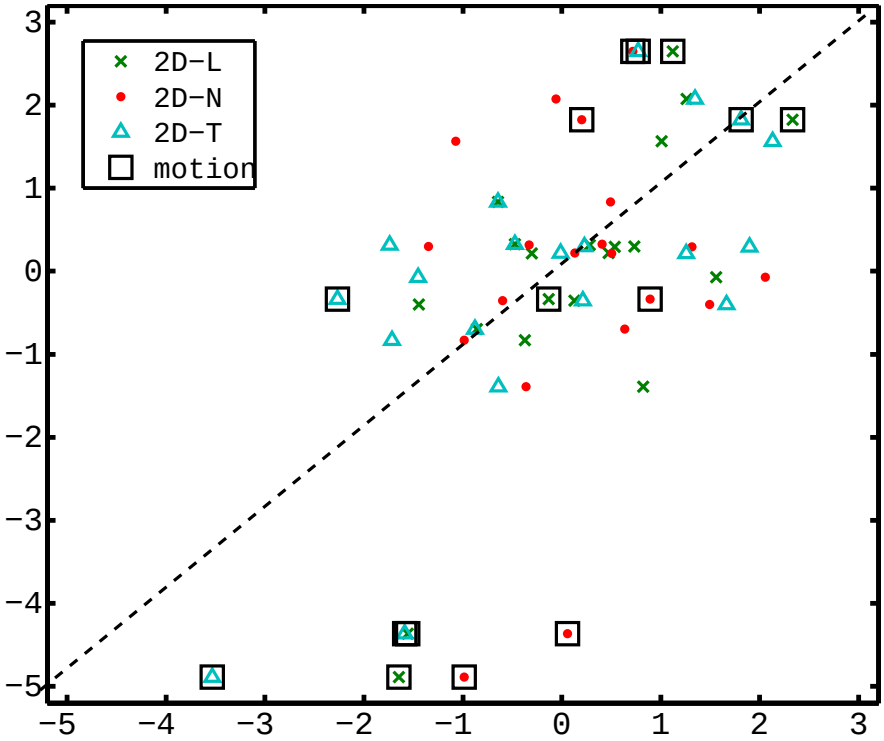

Figure 7: Percentage change in mean ADC scaled by error: the tumour 3D ROI against several 2D ROI definitions; markers: cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

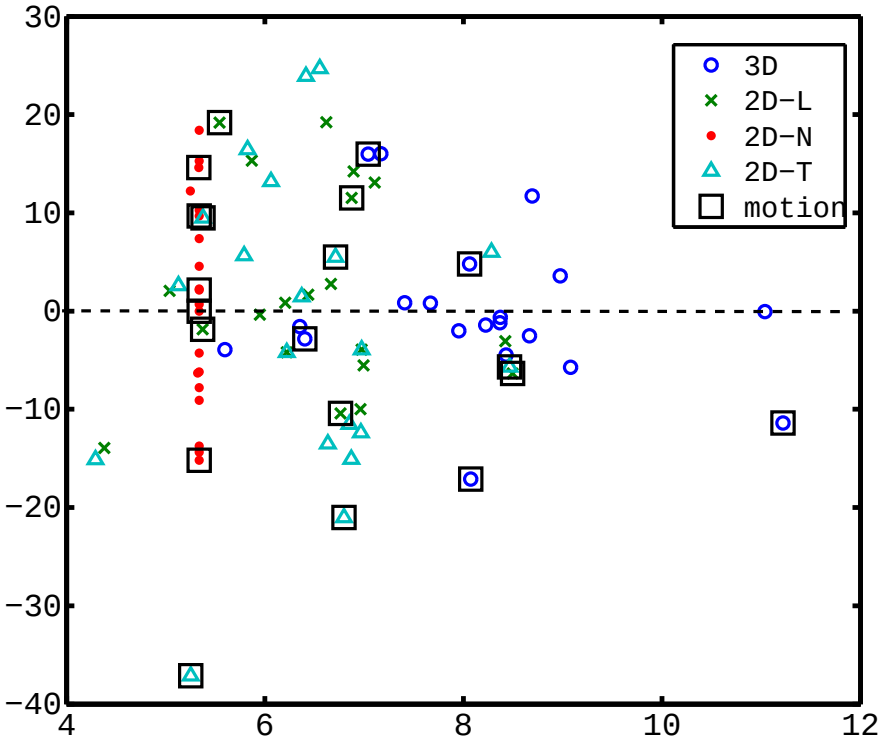

Figure 8: Percentage change in median ADC against log of the average number of voxels; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

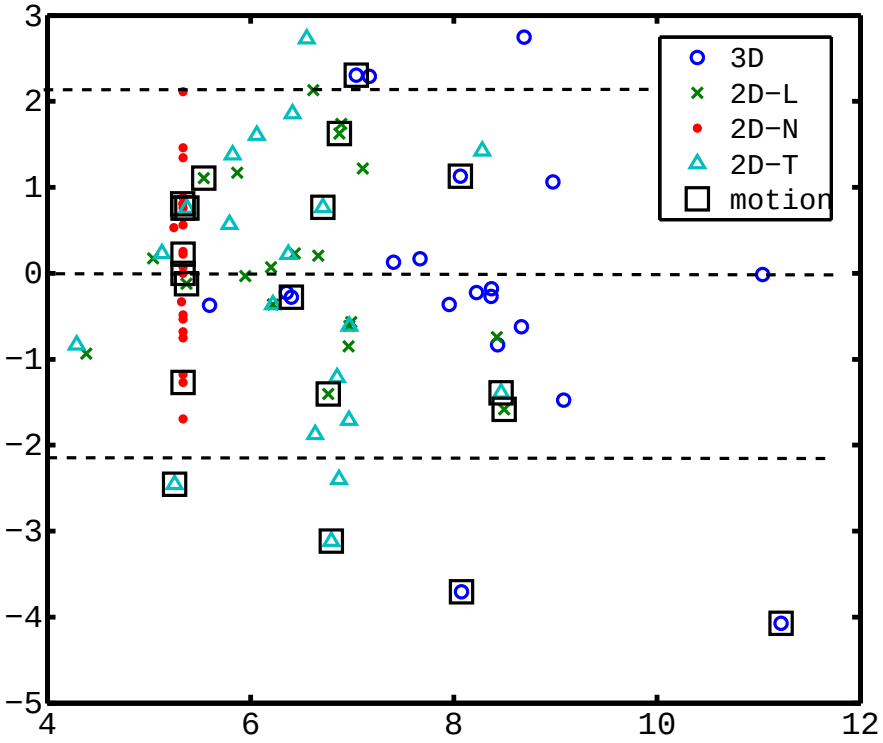

Figure 9: Percentage change in median ADC scaled by error estimates against log of the average number of voxels; markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

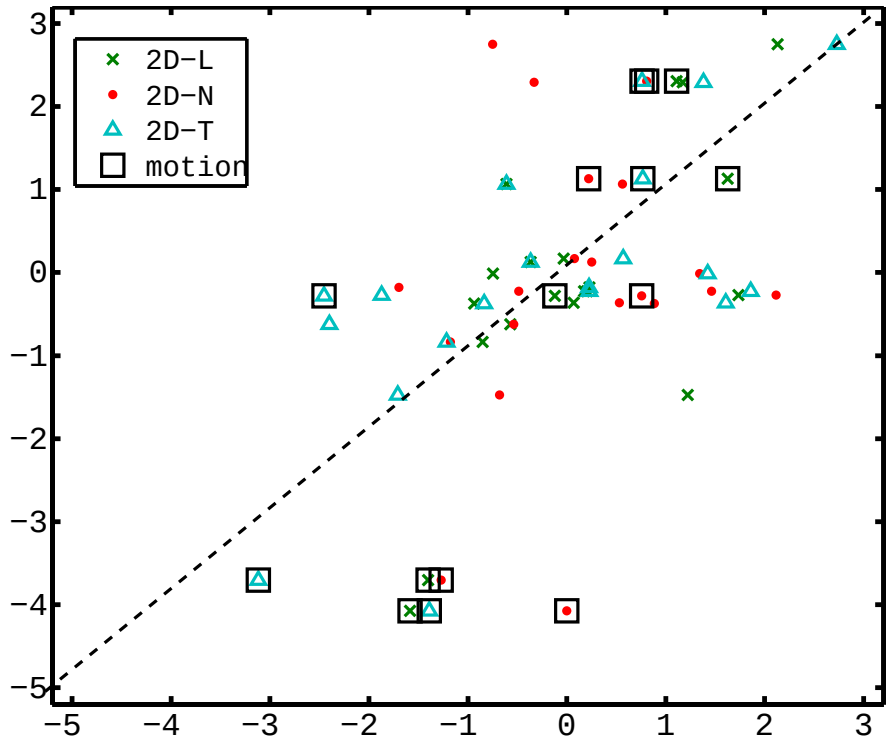

Figure 10: Percentage change in median ADC scaled by error: the tumour 3D ROI against several 2D ROI definitions; markers: cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

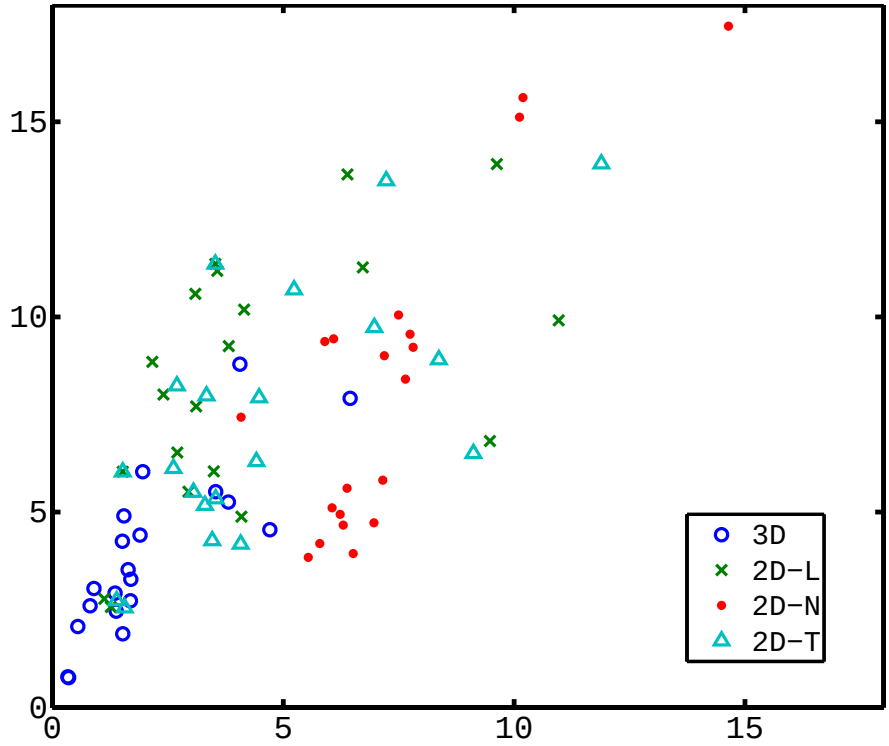

Figure 11: Statistical term with variable width against that with fixed width of ADC distributions, showing the relative contributions from the two statistical terms in the total reproducibility error (3-parameter model) i.e. $\beta\epsilon_{R_{12}}(\sigma_{D_1}, \sigma_{D_2})$ against $\epsilon_{R_{12}}(\sigma_{fix}, \sigma_{fix})$ (Eq. 5); markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI.

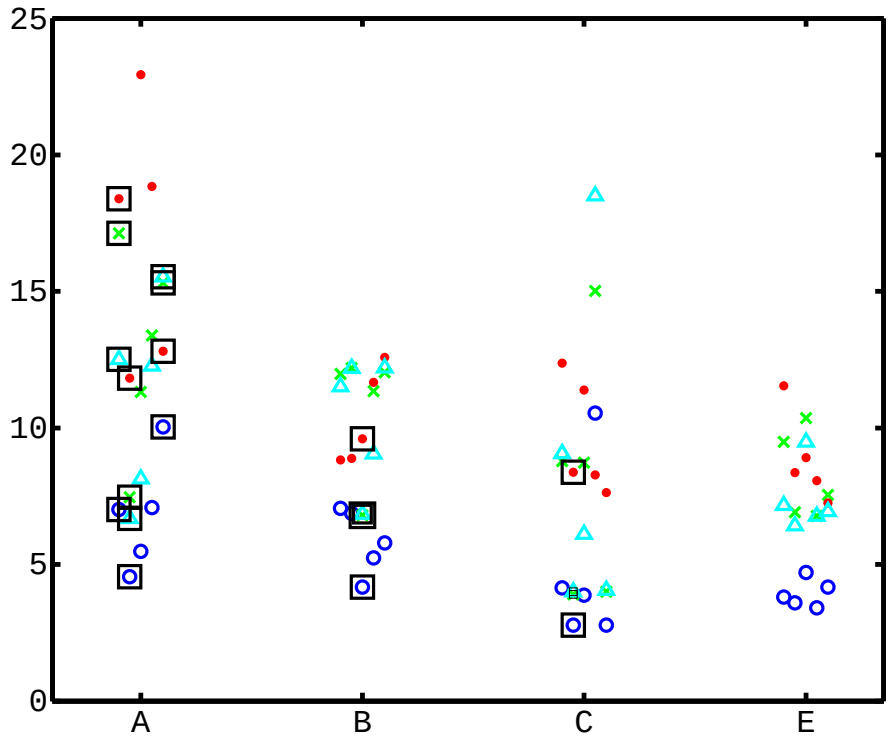

Figure 12: Percentage reproducibility of the mean ADC against the corresponding site (A, B, C, and E); markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.

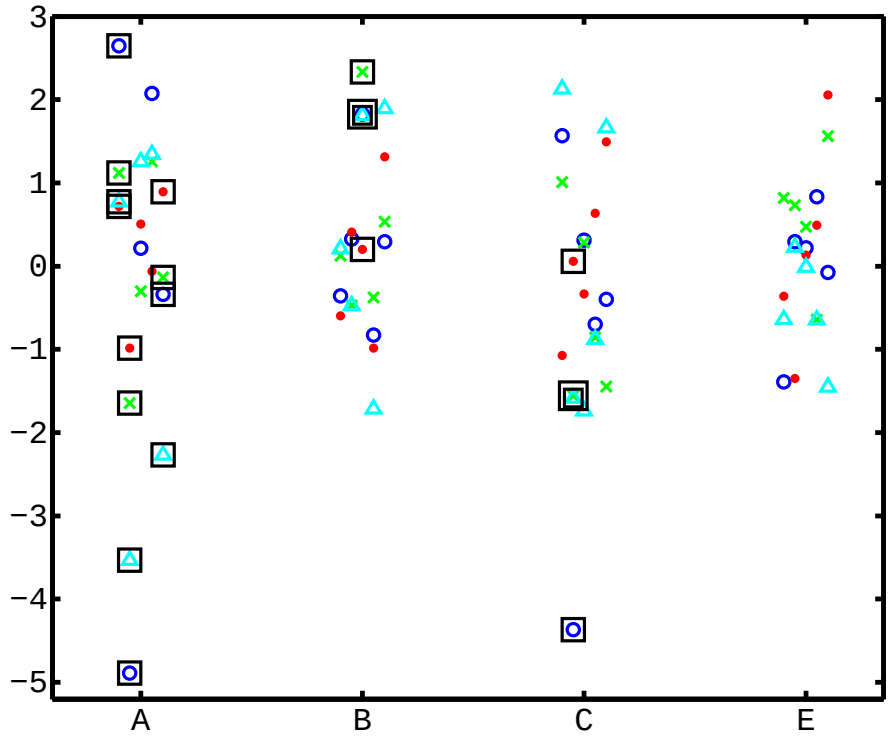

Figure 13: Percentage change in mean ADC scaled by error estimates against the corresponding site (A, B, C, and E);markers: circle: 3D tumour ROI; cross: 2D largest tumour ROI; triangle: 2D most representative solid tumour ROI; dot: 2D normal tissue ROI; square on top of a marker: any 2D or 3D ROI affected by high motion.