Tina Memo No. 2013-005
Internal, IMI preliminary work

# Interpreting Ice-Water Phantom Data for Prediction of Clinical ADC Measurement.

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

Last updated
29 / 10 / 2013


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

# Interpreting Ice-Water Phantom Data for Prediction of Clinical ADC Measurement.

**Abstract**

The aim of this study was to investigate the extent to which measurements made in an ADC (apparent diffusion coefficient) calibration phantom could be shown to correlate with accuracies of measurements made in normal liver. Ideally, a strong correlation might be used in the future in order to perform an assessment of scanner performance based entirely on the phantom.

Several metrics of image quality were automatically extracted from the phantom and ROI's (regions of interest) were identified manually for human livers. Analysis was performed in two stages, firstly looking at correlations between metrics measured in the phantom in order to gain an understanding of reliability, and secondly looking at the reproducibility of human data.

Correlations in phantom measurements help to identify which of the possible summary statistics are likely to be best measured and meaningful. This has allowed us to refine our definitions of the parameters which are most useful for summarising phantom data. It has also allowed us to identify phantom design weaknesses which might be improved. Repeatability in humans sets a limit for realistic clinical performance and also highlights potential problems in current methodology.

## Introduction

The IMI QuIC ConCePT project is a multi-centre proof of concept study to investigate and develop methodologies which support measurement of clinical bio-markers. One of our focuses is in the use of apparent diffusion coefficient (ADC). There is potential to use these techniques either in drug trials or for oncological decision support. Key to use of ADC in this project, and the clinical use of this parameter in general [6] is the development of suitable calibration and Quality Assurance (QA) methods. We have been investigating the use of an ice-water phantom [7] such as suggested in [2] in the context of a clinical study (Figure 1). The data used here was obtained as part of a one-off analysis of scanner performance on 4 separate platforms, followed later by scans from 5 normal volunteers on each.

Earlier, in a related study on the phantom data [1], we noticed that inhomogeneity of the ADC measured in the ice water was an important source of inconsistency. Specifically, we performed an experiment where the phantom was rotated and imaged 4 times, roughly 90° each time. We expected that ADC measurement from each tube should be consistent between different rotations. However the results proved to be slightly inconsistent, apparently due to the difference between the nominal b-value set in the scanner prior to the scan and the effective b-value. The typical magnitude of this issue has also been replicated in [3].

While understanding absolute measurements made using diffusion based protocols is important, what is crucially needed is a phantom which delivers values which correlate with in-vivo performance. We also need to know the accuracy of these measurements and how to interpret them. We aim to develop a simple process for the assessment of machine performance for use in in-vivo studies involving a single acquisition of a multi-purpose phantom. Analysis of this data includes automated estimation of several variables which are intended to characterise important aspects of performance [9]. In addition to mean ADC and mean S0 signal in each tube, we also compute the $\chi^2$ which can be used after normalisation as a reliable measure for goodness of fit. Further, an estimate of the standard deviation of noise in the image is computed from the distribution of second derivatives (for x and y) around zero [4], in a central rectangular region including all five tubes.

### Data

Four sites A, B, C and E participated in this study by providing DW-MRI data of the phantom and different human volunteers using their own scanners. The phantom data used here consist of three scans of the same phantom acquired based on protocol A and more than one week apart (e.g. referred to as A[1], A[2] and A[3] for site A data in Tables 1-2). Each phantom data-set contains five image slices for which there are four b-values, namely 0, 100, 500 and 900. There are five tubes (T0, T1, T2, T3 and T4) in each phantom containing specific fluid. Here, we use the 3 middle slices 2, 3 and 4 to obtain higher measurement accuracies (e.g. referred to as T0(2), T0(3) and T0(4) for tube 0 data in Tables 1-2). The space between the phantom wall and the tubes are filled with ice water with each site following what they understood to be the specified protocol to achieve thermal equilibrium.

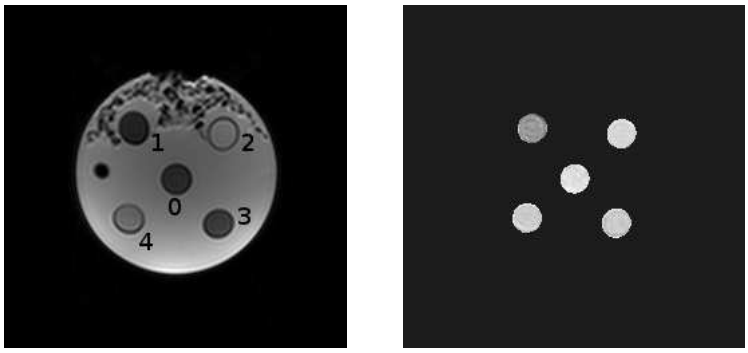


Figure 1: A sample image slice (b=0) from the phantom data used in our study with tube numbers superimposed (tubes 0 to 4 are referred to as T0 to T4); the corresponding ADC map we have generated.

The clinical data used here correspond to five different healthy volunteers per site (except site C which had only four volunteers). Each volunteer has been scanned twice (several days apart) based on protocol A (e.g. referred to as V1[1] and V1[2] for volunteer 1 data in Tables 3-5). The data is from a larger study which investigated the effect of fasting on ADC repeatability. Here we only use the fasted data-sets. Each data-set corresponds to the abdominal slice of the volunteer and contains 40 image slices for which three b-values are available, namely 100, 500 and 900. An expert has annotated three adjacent image slices on each liver using one circular region of interest (ROI) on the top area and one on the bottom area of the image slice (e.g. referred to as A[L58] and A[L67] for site A data in Tables 3-4). The ROI's are expected to be roughly in similar locations of the liver with similar tissue characteristics.

Ideally the phantom should be scanned on the same day as the volunteer. However, it is clear from the scan dates for the phantom and clinical data (see Tables 1 and 5) that although repeat scans for each case have been performed within few weeks from the first scan, except for site-E, there are large gaps from 3 to 13 months between the phantom and clinical scans.

# Methods

In this study we use phantom data (without any rotations) from 4 different scanners and attempt to estimate inhomogeneity using in-house software [8]. This software is capable (on the basis of moderate assumptions) of estimating a general multiplicative trend across an image, independent of the contents, and was applied to whole image ADC maps.

One further possible source of ADC inconsistency was suggested to be lack of sufficient temperature control [2]. It is possible that some parts of the phantom may get warmer than others. Consequently, the amount of ice present in the ice-water regions of the phantom may have an impact on the reliability of measurements. To investigate this, we have visually estimated the amount of ice in each phantom data-set.

For an image slice location given, at each image pixel we fit the corresponding 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. Specifically in clinical data, as the noise distribution is skewed, we find that a first order Rician correction factor improves the quality of fit [5]. We store an average goodness of fit per ROI (individual tubes in the phantom) together with a mean ADC ($\bar{D}$), a mean S0 signal ($\bar{S}0$) and an estimate of the standard deviation of the image noise $\sigma_I$. The goodness of fit is computed using

$$\eta = \kappa\sqrt{\chi^2}\sigma_I/\bar{S}0 \tag{1}$$

where $\kappa = 84.5$ is a fixed coefficient which normalises these values to a level comparable to $\sqrt{\chi^2}$ for which we draw error bars (see Figures 3 and 6). $\kappa$ is the ratio of the mean of all 12 mean signals to the mean of all 12 noise standard deviations from all 4 sites.

In Table 1 we tabulate the mean ADC measurements on five tubes and three image slices for the scans from the four sites. The last column on the table provides the mean $\bar{D}(T)$ on each tube across all scans (multiple data-sets). Meanwhile, we can compute the mean and standard deviation for each scan (single data-set) using the ADC measurements from all five tubes in each phantom image. These are referred to as $\bar{D}(P)$ and $\sigma_D(P)$.

| date | 8/01 | 8/08 | 7/18 | 4/17 | 4/26 | 3/29 | 9/07 | 9/21 | 4/13 | 5/03 | 5/15 | 4/12 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| tube(slc) | A[1] | A[2] | A[3] | B[1] | B[2] | B[3] | C[1] | C[2] | C[3] | E[1] | E[2] | E[3] | $\bar{D}(T)$ |
| T0(2) | 109 | 104 | 108 | 115 | 112 | 117 | 110 | 109 | 116 | 112 | 112 | 113 | 111.7 |
| T0(3) | 108 | 104 | 112 | 115 | 112 | 117 | 110 | 108 | 115 | 111 | 112 | 114 | 112.1 |
| T0(4) | 110 | 109 | 113 | 115 | 112 | 117 | 111 | 109 | 117 | 113 | 114 | 115 | 113.3 |
| T1(2) | 74.7 | 76.1 | 72.8 | 74.0 | 73.6 | 74.5 | 71.1 | 71.6 | 72.9 | 74.5 | 75.7 | 75.7 | 73.98 |
| T1(3) | 75.6 | 73.6 | 73.8 | 71.6 | 74.3 | 75.7 | 71.3 | 71.2 | 72.9 | 76.8 | 74.8 | 75.6 | 73.98 |
| T1(4) | 76.8 | 71.6 | 74.2 | 73.4 | 73.1 | 76.1 | 71.2 | 71.5 | 73.8 | 72.8 | 76.1 | 78.1 | 74.11 |
| T2(2) | 115 | 116 | 114 | 110 | 111 | 110 | 111 | 112 | 111 | 114 | 113 | 115 | 113.1 |
| T2(3) | 115 | 116 | 114 | 110 | 112 | 112 | 110 | 112 | 112 | 113 | 114 | 115 | 113.4 |
| T2(4) | 115 | 116 | 114 | 111 | 112 | 111 | 110 | 111 | 112 | 114 | 115 | 116 | 113.5 |
| T3(2) | 93.1 | 90.8 | 91.6 | 98.8 | 97.7 | 105 | 90.1 | 91.8 | 94.3 | 94.4 | 96.2 | 96.1 | 95.08 |
| T3(3) | 89.7 | 92.4 | 91.5 | 99.3 | 98.3 | 106 | 89.2 | 90.2 | 93.9 | 92.9 | 94.0 | 95.5 | 94.44 |
| T3(4) | 89.1 | 92.8 | 89.5 | 100 | 98.7 | 105 | 90.2 | 91.1 | 95.3 | 94.8 | 95.5 | 97.0 | 95.02 |
| T4(2) | 94.8 | 92.9 | 93.4 | 99.8 | 98.6 | 105 | 91.0 | 91.0 | 92.6 | 94.5 | 95.6 | 96.8 | 95.56 |
| T4(3) | 92.9 | 93.1 | 93.1 | 100 | 99.5 | 106 | 89.9 | 90.2 | 93.2 | 93.6 | 95.9 | 96.9 | 95.50 |
| T4(4) | 90.9 | 90.4 | 93.1 | 101 | 100 | 105 | 90.4 | 90.6 | 92.9 | 94.7 | 96.6 | 98.3 | 95.55 |
| $\sigma_D(P)(2)$ | 1.79 | 4.47 | 2.47 | 3.25 | 1.96 | 6.96 | 3.46 | 2.91 | 2.53 | 0.87 | 1.07 | 1.80 | |
| $\sigma_D(P)(3)$ | 3.03 | 4.05 | 1.84 | 3.94 | 2.60 | 7.58 | 3.93 | 3.56 | 1.89 | 1.68 | 0.64 | 1.85 | |
| $\sigma_D(P)(4)$ | 3.86 | 3.38 | 2.69 | 3.85 | 3.05 | 6.94 | 3.66 | 3.52 | 2.19 | 0.82 | 1.37 | 2.85 | |
| $\bar{D}(P)$ | 96.8 | 96.1 | 96.7 | 99.9 | 99.3 | 103.2 | 94.6 | 94.9 | 97.8 | 97.9 | 98.9 | 100.1 | |
| $\sigma_D(P)$ | 2.82 | 3.73 | 2.21 | 3.45 | 2.41 | 6.69 | 3.45 | 3.13 | 2.08 | 1.12 | 1.00 | 2.07 | |
| $\rho(P) \times 10^4$ | 9879 | 9801 | 9870 | 10190 | 10127 | 10532 | 9649 | 9685 | 9980 | 9991 | 10091 | 10212 | |
| $\sigma'_D(P)$ | 2.89 | 3.67 | 2.15 | 3.26 | 2.36 | 5.00 | 1.69 | 1.58 | 2.31 | 1.24 | 0.67 | 0.85 | |

Table 1: Mean ADC values measured in each tube on each of the three middle adjacent slices (2, 3 and 4) used; the mean and standard deviation of these values across the tubes (the rows identified by $\bar{D}(P)$ and $\sigma_D(P)$), and across the sites (the last column $\bar{D}(T)$); $\sigma'_D(P)$ refers to the standard deviation after scaling the ADC values by scale values $\rho(P)$; all phantom data were acquired in 2012; ADCs are measured in units of $10^{-5}$ $mm^2/s$.

For each column, this standard deviation $\sigma_D(P)$ is computed using ADC value from each tube together with its corresponding mean $\bar{D}(T)$, i.e.

$$\sigma_D(P) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\bar{D}_m - \bar{D}'_n(T))^2} \tag{2}$$

where $M = 15$ (for 5 tubes 3 adjacent slices each) and $m$ refers to the rows 1 to 15 in Table 1. Also while $\bar{D}_m$ includes ADC measurements from the adjacent slices (per scan), $\bar{D}'_n(T)$ refers to a single mean ADC in tube $n$ (average of the 3 mean ADCs from adjacent slices) across all scanners, i.e. $n$ is the remainder of $m/3$ (these 5 values are given below). For example, $\bar{D}'_1(T) = (\bar{D}_1(T) + \bar{D}_2(T) + \bar{D}_3(T))/3$.

Further, an overall mean $\bar{D}(PT)$ (98.02 in this case) may be computed for the mean values $\bar{D}(T)$ and used to compute a scale value $\rho(P)$ for each mean $\bar{D}(P)$.

$$\bar{D}(PT) = \frac{1}{M}\sum_{m=1}^{M}\bar{D}_m(T) \qquad ; \qquad \rho_i(P) = \frac{\bar{D}_i(P)}{\bar{D}(PT)} \tag{3}$$

where $i$ refers to the columns (2 to 13) corresponding to individual datasets (sites) A[1], A[2], ..., E[3].

Hence, $\sigma'_D(P)$ refers to the standard deviation after scaling the ADC values by scale values, i.e.

$$\sigma'_D(P) = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\rho(P)\bar{D}_m - \bar{D}'_n(T))^2} \tag{4}$$

These scale values seem to provide important information about scaling between different scanners and specific scans.

# Results

In Table 2, for each scanner we tabulate the mean S0 signal $\bar{S0}$, the mean $\chi^2$ and standard deviation of image noise $\sigma_I$ together with estimates of ADC inhomogeneity and the percentage of ice for each scan of the phantom. The quantity of ice is estimated visually as a percentage for the slice being studied and is accurate to around 5%.Table 2 and 3 are used here to generate various plots (Figures 1-10). Inhomogeneity is difficult to measure reliably, therefore computed maps of inhomogeneity were corroborated by inspecting grey levels of corrected variation along manually selected lines. We estimated errors on the inhomogeneity measurement of around 1% in good data and 2% when significant ice was present in the slice. These semi-automatic assessments could be made more automatic should we find them to be necessary for quality control when using the phantom.

Tables 3 and 4 show the mean ADC measurements and the corresponding SD values on normal liver for one ROI defined near the top of the image (L58) and another one defined near the bottom (L67). These mean and SD values have been computed from the three adjacent slices for different volunteers both for their first and second scans (e.g. V1[1] and V1[2]). we focus on these results further when discussing Figures 11-13.

| metric | A[1] | A[2] | A[3] | B[1] | B[2] | B[3] | C[1] | C[2] | C[3] | E[1] | E[2] | E[3] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\bar{S0}$ (2) | 983.8 | 780.7 | 866.4 | 620.1 | 588.2 | 596.8 | 416.8 | 357.9 | 409.1 | 462.0 | 489.2 | 485.0 |
| $\bar{S0}$ (3) | 869.9 | 776.2 | 851.7 | 619.7 | 590.9 | 598.8 | 422.8 | 371.9 | 408.9 | 459.3 | 486.3 | 484.1 |
| $\bar{S0}$ (4) | 829.5 | 998.9 | 848.3 | 621.2 | 588.3 | 602.2 | 420.2 | 364.3 | 407.2 | 461.2 | 488.4 | 407.2 |
| $\bar{S0}$ | 888.0 | 863.2 | 855.4 | 620.3 | 589.1 | 599.2 | 419.9 | 364.7 | 408.4 | 460.8 | 488.0 | 458.7 |
| $\bar{\chi^2}$ (2) | 0.818 | 0.732 | 0.574 | 0.864 | 2.376 | 1.462 | 0.560 | 0.460 | 3.940 | 0.814 | 3.176 | 3.274 |
| $\bar{\chi^2}$ (3) | 0.918 | 0.690 | 0.650 | 1.264 | 2.282 | 1.252 | 0.512 | 0.478 | 3.802 | 1.044 | 2.974 | 3.298 |
| $\bar{\chi^2}$ (4) | 0.914 | 0.810 | 0.802 | 0.868 | 2.500 | 1.962 | 0.470 | 0.524 | 3.902 | 0.904 | 3.038 | 3.172 |
| $\bar{\chi^2}$ | 0.883 | 0.744 | 0.675 | 0.998 | 2.386 | 1.558 | 0.514 | 0.487 | 3.88 | 0.920 | 3.062 | 3.248 |
| $\sigma_I$ (2) | 14.28 | 14.62 | 18.43 | 6.30 | 3.02 | 4.14 | 4.84 | 4.73 | 1.14 | 5.81 | 2.02 | 1.79 |
| $\sigma_I$ (3) | 14.61 | 15.04 | 18.20 | 5.32 | 3.07 | 4.30 | 4.72 | 4.84 | 1.16 | 5.40 | 2.13 | 1.70 |
| $\sigma_I$ (4) | 14.33 | 15.78 | 17.44 | 6.33 | 2.80 | 3.41 | 5.08 | 4.59 | 1.16 | 4.99 | 1.94 | 1.78 |
| $\bar{\sigma_I}$ | 14.40 | 15.14 | 18.02 | 5.98 | 2.96 | 3.95 | 4.88 | 4.72 | 1.15 | 5.40 | 2.03 | 1.75 |
| $\xi$ (2) | 3% | 4.5% | 3% | 4% | 6% | 5% | 2% | 6% | 3% | 4% | 5% | 3% |
| $\xi$ (3) | 2% | 5.5% | 5% | 2% | 9% | 4% | 5% | 9% | 3% | 7% | 6% | 7% |
| $\xi$ (4) | 2% | 3.5% | 3% | 3% | 7% | 4% | 5% | 6% | 4% | 2% | 3% | 2% |
| $\xi$ % | 2.33 | 4.50 | 3.66 | 3.00 | 7.33 | 4.33 | 4.00 | 7.00 | 3.33 | 4.33 | 4.67 | 4.00 |
| ice % | 95 | 95 | 90 | 35 | 10 | 15 | 95 | 95 | 3 | 60 | 5 | 10 |

Table 2: Mean S0 signal values and mean $\chi^2$ values computed in all five tubes per site (per scan) on each of the three middle adjacent slices (2, 3 and 4) used; the standard deviation of image noise and the percentage of inhomogeneity ($\xi$) on the ADC values across the phantom for these slices; an overall mean for each metric; the percentage of ice in the ice-water region of the phantom.

If the dominant source of ADC estimation is due to measurement noise in the scanner then a strong correlation is expected between the variance in ADC measurements within the tubes and the corresponding signal-to-noise ratios for each scanner. However, for these variables this correlation is very poor and cannot be accounted for by predicted measurement noise (Figure 2). While investigating this issue it was found that $\chi^2$ per degrees of freedom (DOF) values were highly variable (factor $\approx 2$) and did not conform to the unity value which was expected. The noise estimate also seemed to correlate with ice fraction. The correlation found between $\chi^2$ and signal-to-noise (Figure 3) is indicative of poor measurement of the noise (a $\chi^2$ computed for equivalently performing scanners but with errors on the estimated signal-to-noise would follow the $y = 1/x^2$ dependency shown).

A second estimate of signal-to-noise was therefore constructed based upon the average goodness of fit seen within the tubes. The spread of this new variable was seen to be smaller than the original spatial noise estimates, so

| Volunteer[scan] /site[ROI] | A[L58] | A[L67] | B[L58] | B[L67] | C[L58] | C[L67] | E[L58] | E[L67] |
|---|---|---|---|---|---|---|---|---|
| V1[1] | 104.97 | 102.97 | 93.49 | 94.48 | 102.12 | 90.77 | 105.70 | 106.73 |
| V1[2] | 119.66 | 105.59 | 97.38 | 103.43 | 107.96 | 93.02 | 116.25 | 113.15 |
| V2[1] | 60.08 | 65.52 | 84.97 | 65.40 | 133.35 | 112.22 | 108.64 | 108.53 |
| V2[2] | 78.73 | 52.49 | 83.6 | 53.49 | 135.32 | 106.42 | 109.62 | 116.60 |
| V3[1] | 95.62 | 83.45 | 97.18 | 81.10 | 105.65 | 99.25 | 100.95 | 74.41 |
| V3[2] | 94.86 | 102.55 | 90.16 | 92.39 | 109.16 | 102.55 | 95.52 | 76.86 |
| V4[1] | 105.71 | 94.87 | 108.69 | 97.70 | 122.17 | 110.56 | 119.64 | 104.91 |
| V4[2] | 103.49 | 84.09 | 88.19 | 96.70 | 110.81 | 89.95 | 105.45 | 104.51 |
| V5[1] | 94.95 | 84.34 | 46.12 | 51.72 | n/a | n/a | 102.52 | 93.56 |
| V5[2] | 121.93 | 105.21 | 64.30 | 53.32 | n/a | n/a | 99.51 | 91.35 |

Table 3: Mean ADC values on each ROI (Liver58 or Liver67) using 3 adjacent slices (for different volunteers and scan dates); ADCs are measured in units of $10^{-5}\ mm^2/s$.

| Volunteer[scan] /site[ROI] | A[L58] | A[L67] | B[L58] | B[L67] | C[L58] | C[L67] | E[L58] | E[L67] |
|---|---|---|---|---|---|---|---|---|
| V1[1] | 3.64 | 3.41 | 5.21 | 1.67 | 7.62 | 4.15 | 0.39 | 2.00 |
| V1[2] | 10.72 | 5.38 | 2.89 | 7.66 | 3.90 | 2.63 | 6.91 | 4.67 |
| V2[1] | 2.75 | 5.47 | 3.62 | 3.32 | 7.45 | 16.07 | 9.40 | 7.97 |
| V2[2] | 2.63 | 1.55 | 9.37 | 5.97 | 3.03 | 5.27 | 2.80 | 3.08 |
| V3[1] | 2.99 | 6.47 | 2.71 | 6.65 | 6.08 | 5.86 | 10.01 | 20.96 |
| V3[2] | 4.92 | 5.89 | 4.19 | 9.45 | 8.98 | 5.89 | 8.72 | 3.35 |
| V4[1] | 1.11 | 9.83 | 5.80 | 2.53 | 3.51 | 10.33 | 15.52 | 11.44 |
| V4[2] | 6.42 | 7.61 | 10.51 | 14.97 | 3.94 | 12.72 | 9.18 | 11.28 |
| V5[1] | 3.66 | 4.58 | 1.03 | 2.54 | n/a | n/a | 4.87 | 4.97 |
| V5[2] | 2.87 | 9.62 | 9.47 | 2.65 | n/a | n/a | 9.81 | 5.26 |

Table 4: Standard deviation values on ADCs measured on each ROI (Liver58 or Liver67) using 3 adjacent slices (for different volunteers and scan dates); ADCs are measured in units of $10^{-5}\ mm^2/s$.

| Volunteer[scan] /site | A | B | C | E |
|---|---|---|---|---|
| V1[1] | 2013/02/25 | 2013/03/05 | 2012/12/19 | 2012/05/03 |
| V1[2] | 2013/03/04 | 2013/03/12 | 2012/12/27 | 2012/05/15 |
| V2[1] | 2013/03/06 | 2013/03/21 | 2013/01/11 | 2012/05/21 |
| V2[2] | 2013/03/13 | 2013/03/28 | 2013/01/18 | 2012/05/28 |
| V3[1] | 2013/03/11 | 2013/03/26 | 2013/03/28 | 2012/05/24 |
| V3[2] | 2013/03/18 | 2013/04/02 | 2013/04/04 | 2012/05/28 |
| V4[1] | 2013/04/16 | 2013/03/27 | 2013/05/29 | 2012/06/11 |
| V4[2] | 2013/04/23 | 2013/04/03 | 2013/06/05 | 2012/06/14 |
| V5[1] | 2013/04/16 | 2013/05/17 | n/a | 2012/07/31 |
| V5[1] | 2013/04/23 | 2013/05/24 | n/a | 2012/08/09 |

Table 5: Scan dates for different volunteers using different scanners (at different sites).

improving the correlation with ADC variance (Figure 4). A chi-square scaled signal-to-noise ratio is a pure estimate of fitting noise (random error) with error less than 5%. However, the degree of correlation is still too poor to be explained by measurement accuracy. The integrity of the ADC variance computed using the tubes was suspected.

As the true ADC values per tube are unknown, the differences seen between measurements in tubes and the expected values (scanner averages) is only a measure of general conformity, i.e. how average is my scanner. A SD (standard deviation) can be more stably computed for tube measurements by taking out the average ADC scaling. Whist spatial variation in ADC might be expected to destabilise this process, the overall scaling of ADC seems to be scanner specific and accurate to around 1% (Figure 5). The new estimate of ADC variation is systematically

less than the original value (as expected) (Figure 6). Correlations between the new ADC SD and the goodness of fit are again improved, as is illustrated by the tighter grouping of measurements in Figure 7. It can be seen that the distributions of measurements are largely consistent with the best fit line to the data ($\chi^2$/DOF = 107/11).

ADC variance in the phantom results in a proportional error of 18% on the SD (estimated from multiple samples). and has possible contributions due to inhomogeneity. However, inhomogeneity measurement accuracy is currently too poor to see a consistent trend within scanners (Figure 8). This is due to the ice quantities which inhibit accurate measurement of inhomogeneity. No evidence is seen for reduced accuracy of ADC SD with ice fraction (Figure 9). A plot of average ADC scaling against ice fraction (Figure 10) does illustrate a correlation consistent with reducing temperature within tubes (the line shown is for a 1° drop in temperature in water around the freezing point). However, it is also consistent with percentage level systematic differences between scanners [3]. Figure 11 shows the scale values $\rho(P)$ on mean ADCs for each scan of the phantom (Table 1) against the average scale value (corresponding to three scans per site) for different sites. Here, while the spread of scale values are tighter for sites A and E, each of the sites B and C has one scale value which stands away from its corresponding group (with a difference greater than 3%). However, these differences are consistent with the estimated errors.

The average values for diffusion coefficient measured in each tube across the four different scanners ($\bar{D}'_n(T)$) was 112.3, 74.0, 113.3, 94.8 and 95.5 ($10^{-5}$ $mm^2/s$). These are systematically different from the design specification 117, 78, 117, 97 and 99 ($10^{-5}$ $mm^2/s$) [7] and produce a significant difference in computed variances and deviations. We suggest that the new averages (i.e. as measured in practice) should be those used during QA.

We have attempted to correlate the accuracy (reproducibility) of in-vivo ADC measurement with parameters measured using the IMI phantom. A mean SD for the phantom data was computed for each site (using the SD values corresponding to the three scans from that site). Figures 12 and 13 show the spread of the mean SD values for the in-vivo data from specific sites (Table 4) against the corresponding mean SD values for the phantom data. While Figures 12 and 13 correspond to the in-vivo data from separate ROI's (L58 and L67), we can put together the two SD data in Figure 14 providing up to 20 SD values per site rather than 10. Finally in Figure 15 we use the data from both ROI's to compute a single average ADC and SD value per site. Assuming that the two ROI's exhibit similar ADC values (no inhomogeneity, etc.), this way of computing the SD values is statistically more accurate as there are six ADC values used rather than three (as in Figures 12-14). While measurable differences can be seen in the quality of data between the different scanners in the phantom no correlation was found between these values and human liver measurements. Clearly, a lack of correlation between phantom and liver data implies that the ice phantom cannot currently predict biological measurement accuracy in human tissue. However, in this case the cause of the problem seems to be due to a lack of variation in human measurements rather than measurements of the phantom.

## Discussions and Conclusions

In the phantom we measured the variance between average ADC values and measurement in 5 tubes, the spatial noise, the amount of ice remaining, the total signal in all 5 tubes, the maximum percentage ADC inhomogeneity within the phantom and the average $\chi^2$ (fit error) from the exponential fits. From these values other measures were computed as descibed in equations 1-4.

It was found that both our initial estimate of signal-to-noise and ADC SD could be improved upon by using average fitting error and taking out average scalings of ADC. As absolute scalings of ADC are a potential consequence of scanner design and specifics of acquisition protocols this is a rational step. It provides a mechanism for extracting genuine measurement problems from design differences and seems to also reduce any problems with residual temperature differences within the tubes (see below). One weakness of the fit error is that (unlike tube ADC variances) it is affected by intrinsic smoothing applied during image reconstruction, which may be scanner specific. Direct comparison of these figures therefore requires an assumption of equivalent image shapness.

For the current measurement process, and assuming that the correlation seen in Figure 9 is entirely due to temperature changes, a quantity of ice corresponding to 50% of the available area at the centre of the phantom would appear to reduce any thermal effects to a level below our current ability to measure them. Variance of ADC around the average scaled values shows no evidence of a correlation with quantity of ice. With temperature effects eliminated as a significant factor (and ignoring the possibility of varying levels of smoothing as described above), the relatively poor degree of correlation seen in Figure 7 is probably best explained by spatial inhomogeneity of ADC measurement, which may be as much as 10% across the bore [3]. Inhomogeneity being the main cause of measurement error (and largely consistent between scanners) would also accord with the observed behaviour of in-vivo data. Unfortunately, until now measuring inhomogeneity has not been an important consideration and the current phantom design and protocol is unable to reliably quantify this. While the automated localisation and tube

analysis software worked as intended, inhomogeneity and spatial noise (thought be be important for understanding overall scanner performance) were very difficult to assess when large quantities of ice were present in the phantom. These data support an argument for a better design, particularly one in which ice does not play such a detrimental role during analysis.

The third scan of the phantom for site B appears as a persistent outlier in all plots. This could be due to several factors, including poor thermal equilibrium or a lack of appropriate centring of the phantom within the bore. This is the same data as used in [1] giving us a direct indication of the magnitude of the problem. Specifically, inhomogeneity could easily account for a factor of two increase in observed ADC SD in these data.

We also assessed the measurement of ADC in normal liver, by measuring the variance in mean value across three slices in 2 ROI's. The assumption of the experimental design was that the performance of a scanner is intrinsic and equivalent for all time. Measurements of the phantom were therefore often performed several months before the human subjects. However, significant (measurable) differences in performance of the scanners were seen for phantom scans even one week separate. Obtaining correlations of phantom data with in-vivo measurement was therefore always likely to be difficult with the present study design.

We need a better way of assessing the performance of the scanner in-vivo, currently all manufacturers deliver ADC measurements with the same intrinsic reproducibility, most likely driven by biology and our methodology rather than equipment. As a consequence all we can say about the use of our present results for the design of a quality control process is that all of the scanners in this study are capable of generating equivalent quality in-vivo measurement. Thus any new scanner which generates phantom results which are as good as any of these will be likely to perform as well in studies. However, there is a strong possibility that this conclusion will need to be modified once the actual measurement requirements of in-vivo studies are properly assessed. From the set of liver measurements shown in this work we can infer that any detrimental effects on performance as measured in the phantom must be at least 2-3 times less significant than the main mechanisms affecting biological data. We would therefore need to improve the accuracy of ROI average ADC estimates by a similar factor in order to hope to see correlations.

### Recommendations

Phantom analysis software should deliver average fit error (Equation 1), average ADC scaling (to agreed values) (Equation 3) and SD around the rescaled values in tubes (Equation 4). These values are the most reliable (5% error, 0.6% error and 18% error respectively) and should be output routinely from the phantom analysis software.

Inhomogeneity and spatial noise are still believed to be important parameters for understanding performance, but need to be measured better (currently 2% and 100% errors with large quantities of ice respectively). This would be best achieved by excluding ice from the region of the phantom in which measurements are made. There is no evidence in our data to suggest that by doing this the ADC measurement accuracy would be impaired, though if this is a concern other steps might be taken in order to improve thermal homogeneity (for example use of a circulation pump with its own cooled reservoir between acquisitions).

Performance of ADC measurements in-vivo need to be more accurate if we wish to see correlations with phantom measurements. This might be achieved by registration based analysis. We should also look at goodness of fit measurements for in-vivo data in order to eliminate the effects of intrinsic biological variation in ADC.

What is crucially needed is a phantom which delivers values which correlate with in-vivo performance. It is particularly worrying that the current design of phantom has more susceptibility artefacts (around tubes) than in-vivo data but also does not provide any mechanism for assessing the effectiveness of fat suppression. Both of these could be addressed with minor modifications to the phantom.

It should be emphasised that the current phantom design seems more than capable of quantifying absolute diffusion (ADC) using the present tubes and liquids. Improved accuracy of these values or more tubes would seem redundant. Any changes made need to be supported by appropriate experiments and datasets in the context of the IMI study and in accordance with this current analysis.

## References

[1] N.A. Thacker, H. Ragheb and D. Morris, Towards a Power Calculation for ADC Measurement in Clinical Trials, *Tina memo* 2012-005, 2012, *http://www.tina-vision.net/docs/memos/2012-005.pdf*.

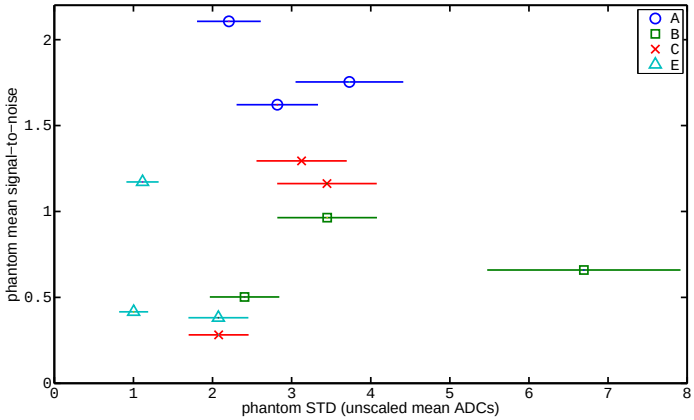


Figure 2: Phantom: mean signal-to-noise ratio against the standard deviation of the original mean ADC values for different sites A, B, C and E (three scan dates per site).

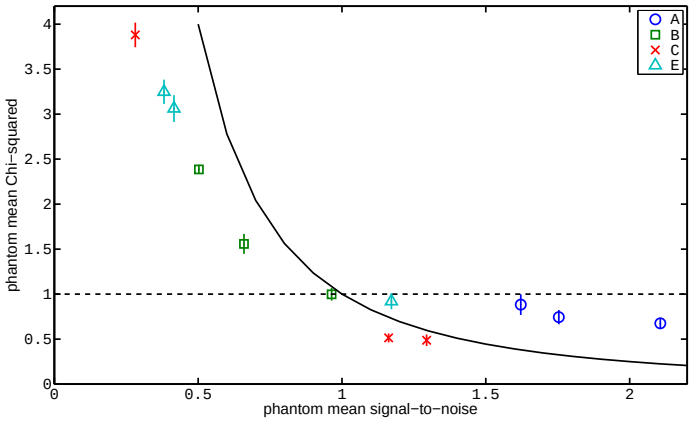


Figure 3: Phantom: mean $\chi^2$ against the mean signal-to-noise ratio for different sites A, B, C and E (three scan dates per site); the solid curve shows the $y = 1/x^2$ dependency.

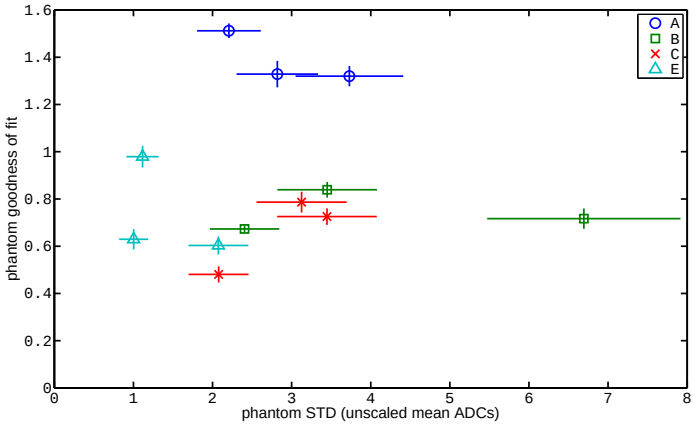


Figure 4: Phantom: mean goodness of fit against the standard deviation of the original mean ADC values for different sites A, B, C and E (three scan dates per site).

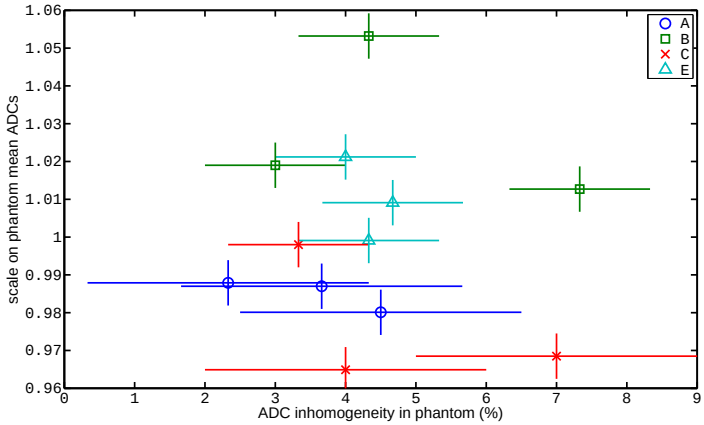


Figure 5: Phantom: the scale estimate on mean ADC values against the percentage of inhomogeneity for different sites A, B, C and E (three scan dates per site). The generally comparable level of measurement inhomogeneity across all sites is consistent with this being the dominant source of measurement error for in-vivo data (see Figures 12-15 below). However, the current method of measurement is rather imprecise and adversely affected by ice quantity.

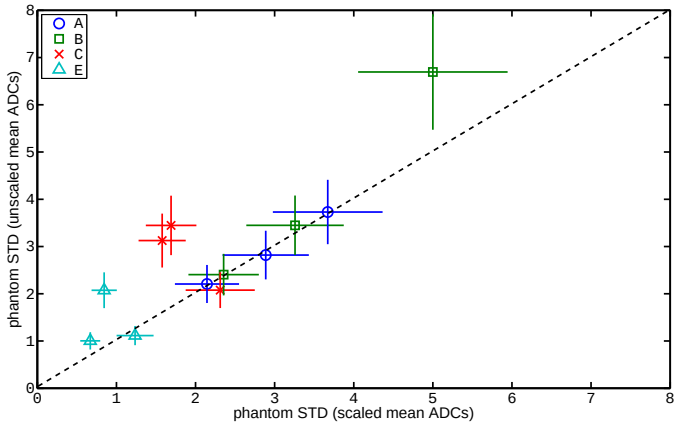


Figure 6: Phantom: standard deviation of the original mean ADC values against the standard deviation of the scaled mean ADC values for different sites A, B, C and E (three scan dates per site).

[2] T.L. Chenevert, C.J. Galban, M.K. Ivancevic, S.E. Rohrer, F.J. Londy, T.C. Kwee, C.R. Meyer, T.D. Johnson, A. Rehemtulla and B.D. Ross, Diffusion Coefficient Measurement Using a Temperature-Controlled Fluid for Quality Control in Multicenter Studies, *J. Magnetic Resonance Imaging*, 34:983-987, 2011.

[3] D. Malyarenko, C.J. Galban, F.J. Londy, C.R. Meyer, T.D. Johnson, A. Rehemtulla, B.D. Ross and T.L. Chenevert, Multi-system Repeatability and Reproducibility of Apparent Diffusion Coefficient Measurement Using an Ice-Water Phantom, *J. Magnetic Resonance Imaging*, 37:1238-1246, 2013.

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

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

[6] 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.

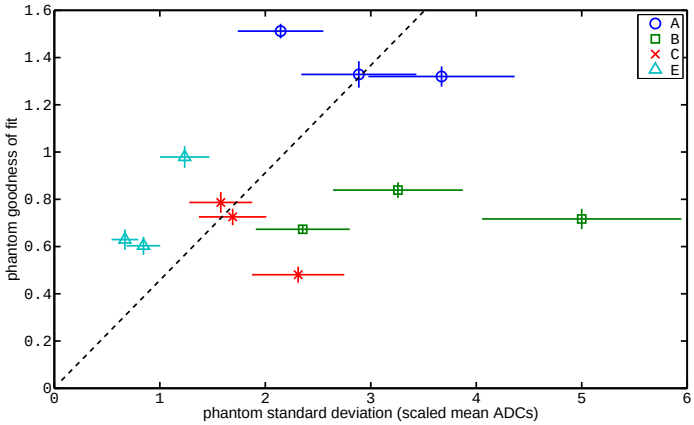


Figure 7: Phantom: mean goodness of fit against the standard deviation of the scaled mean ADC values for different sites A, B, C and E (three scan dates per site); the dashed line shows the best fit correlation between the two variables on x and y axes. The indvidual scanners show good self consistency, with sites E and C being comparable and sites B and A being significantly worse for one variable. However no correlation was found between these variables and in-vivo measurement (Figures 12-15 below).

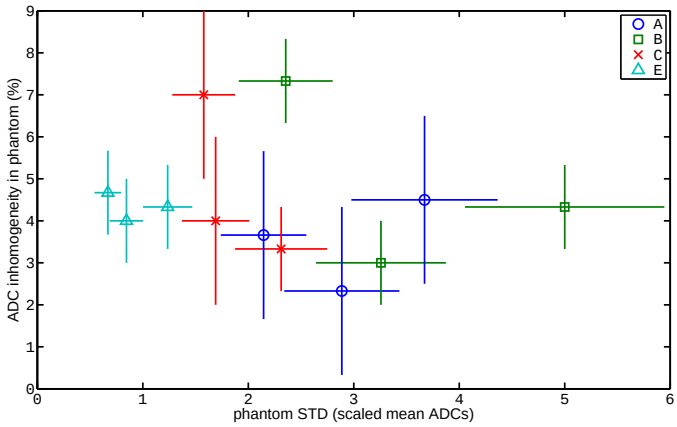


Figure 8: Phantom: the percentage of inhomogeneity against the standard deviation of the scaled mean ADC values for different sites A, B, C and E (three scan dates per site).

[7] J.M. Winfield, N.H.M. Hogg, N.M. de Souza and D.J. Collins, Phantom for Quality Assurance in Multi-Centre Trials of Diffusion-Weighted Magnetic Resonance Imaging, *CRUK and EPSRC Cancer Imaging Conference*, Imperial College, April 26th, 2012.

[8] E.A. Vokurka, N.A. Thacker and A. Jackson, A Fast Model Independent Method for Automatic Correction of Intensity Nonuniformity in MRI Data. *J. Magnetic Resonance Imaging*, 10(4):550-562, 1999.

[9] H. Ragheb, N.A. Thacker, D.M. Morris, N.H.M. Hogg and A. Jackson, Ice-Water Phantom Localisation for Diffusion Calibration, *Medical Image Understanding and Analysis (MIUA)*, Swansea, UK, July 9-11, pp. 179-184, 2012.

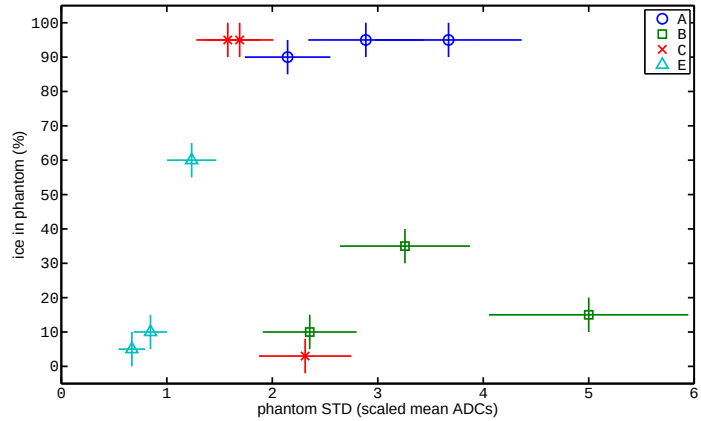


Figure 9: Phantom: the percentage of ice against the standard deviation of the scaled mean ADC values for different sites A, B, C and E (three scan dates per site).

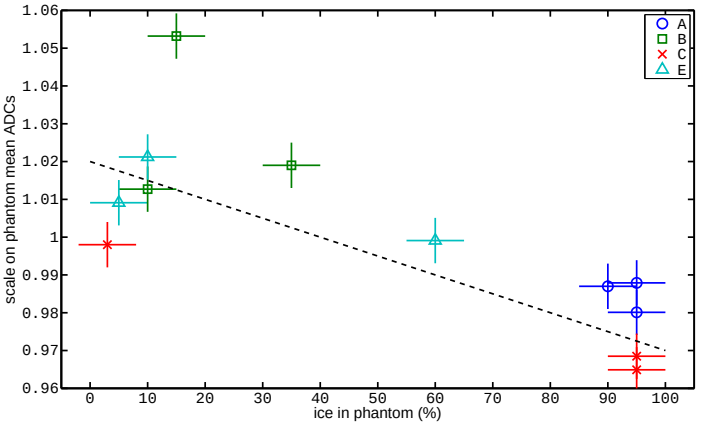


Figure 10: Phantom: the scale estimate on mean ADC values against the percentage of ice for different sites A, B, C and E (three scan dates per site); the dashed line shows the 5% variation of ADC expected for a 1°celcius change of temperature in water. Differences between datasets are therefore partly explainable as due to poor thermal control for low quantities of ice.

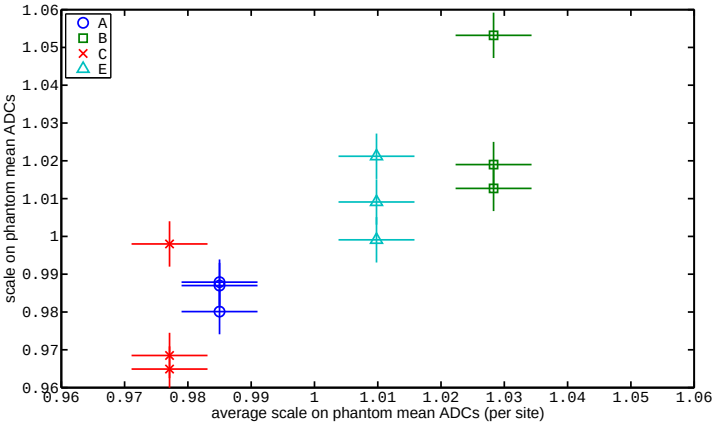


Figure 11: Phantom: the scale estimate on mean ADC values for each scan against the average scale value per site for different sites A, B, C and E (three scan dates per site).

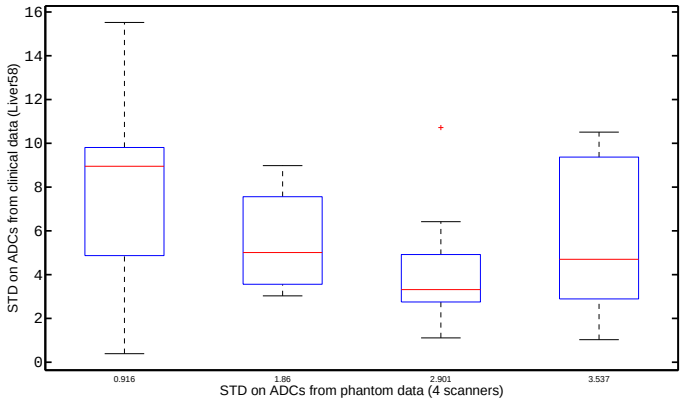


Figure 12: Clinical data against phantom data: standard deviation of the mean ADC values from healthy livers (up to five different volunteers per site each scanned on two different dates) against the standard deviation of the scaled mean ADC values from the corresponding phantom at different sites; whiskers from-left-to-right correspond to sites: E (0.91), C (1.86), A (2.9) and B (3.5); using the ROI on the top of the image of each liver (L58).

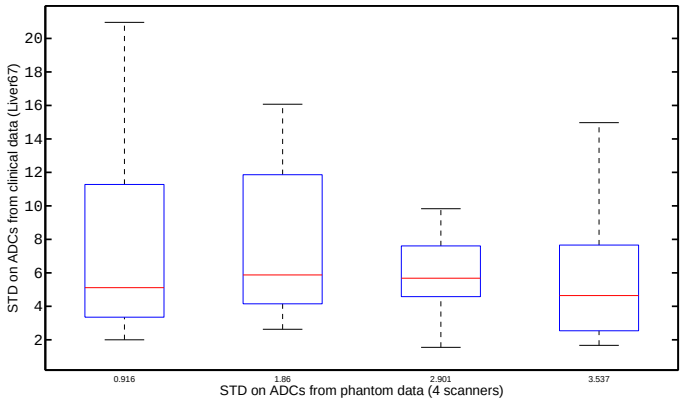


Figure 13: Clinical data against phantom data: standard deviation of the mean ADC values from healthy livers (up to five different volunteers per site each scanned on two different dates) against the standard deviation of the scaled mean ADC values from the corresponding phantom at different sites; whiskers from-left-to-right correspond to sites: E (0.91), C (1.86), A (2.9) and B (3.5); using the ROI on the bottom of the image of each liver (L67).

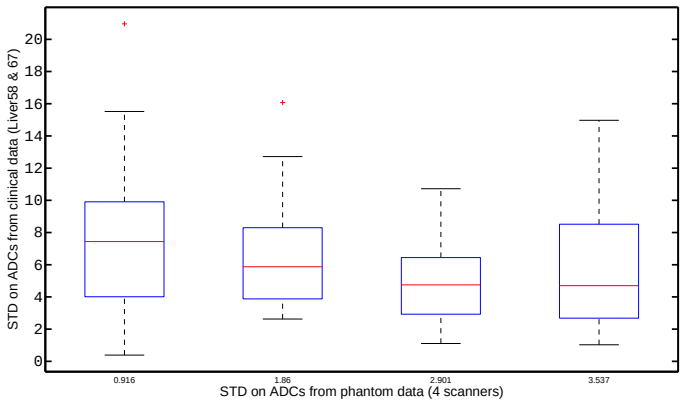


Figure 14: Clinical data against phantom data: standard deviation of the mean ADC values from healthy livers (up to five different volunteers per site each scanned on two different dates) against the standard deviation of the scaled mean ADC values from the corresponding phantom at different sites; whiskers from-left-to-right correspond to sites: E (0.91), C (1.86), A (2.9) and B (3.5); using two standard deviation values per scan each corresponding to one of the two ROI's (L58 and L67), i.e. 20 values per site.

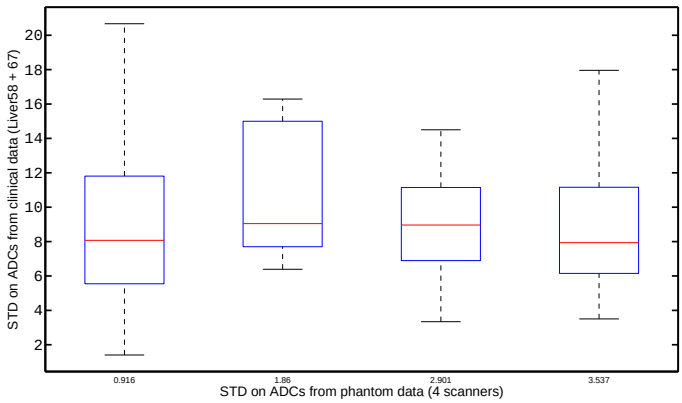


Figure 15: Clinical data against phantom data: standard deviation of the mean ADC values from healthy livers (up to five different volunteers per site each scanned on two different dates) against the standard deviation of the scaled mean ADC values from the corresponding phantom at different sites; whiskers from-left-to-right correspond to sites: E (0.91), C (1.86), A (2.9) and B (3.5); using one single standard deviation value per scan by combining the values from the two ROI's (L58 and L67), i.e. 10 values per site.