Tina Memo No. 2012-007
Internal, IMI preliminary work

Monte-Carlo ADC Measurement Sensitivity Tables for Null Hypothesis Tests in Organs with Heterogeneity.

Neil.A.Thacker and Hossein Ragheb.

Last updated
12 / 1 / 2013

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Monte-Carlo ADC Measurement Sensitivity Tables for Null Hypothesis Tests in Organs with Heterogeneity.

Abstract

The purpose of this work is to develop and document simple statistical methods for detection of changes observed on ADC (Apparent Diffusion Coefficient) measurements in cancerous organs (as in the IMI Quic Concept project). We expect similar methods will be used while monitoring the early response of tumours to therapy and during clinical management. It is important to understand the operation and statistical limits of various approaches, so that we can estimate the expected statistical power of any study, and to understand how best to make use of the information available in phantom data.

The main reason for partial response of cancerous cells to chemotherapy is heterogeneity. This means the ADC histogram computed using DW-MR images (corresponding to the pixels in the region of interest (ROI)) will only change in a small portion of an organ volume some time after administering drugs. However in practice, large changes may be seen because of systematic errors due to changes in the vendor machine, the effective b values, ROI, etc.

After describing the typical change detection scenario encountered in experiments, a simple Monte-Carlo model is introduced which is thought to be characteristic of the main mechanisms of variation seen in real data. This allows us to explore the level of statistically significant changes using parameters which would be obtained from patient and phantom data. We conclude with general statements regarding sensitivity of a variety of simple approaches, and with calibrated tables and graphs suitable for computing statistical tests relating to the null hypothesis (no change seen).

Introduction

In conventional experiments [18], [11] the effects of a drug treatment are observed in groups of subjects. Mean (or median) ADC values are taken from a specified region (or organ) [8]. The pre and post treatment means are then averaged over all subjects and assessed for change at the group level [14], [12]. Variances across the group can be computed in order to construct a T-test, and thereby obtain a P value for the observation of significant change. The specific mean value and distribution width will be organ specific. For example, typical (ball-park) values of ADC in an organ might be $132 \times 10^{-5}\ mm^2/s$, with a standard deviation on the distribution of value of $30 \times 10^{-5}\ mm^2/s$, i.e. the distribution extends almost all the way from a value of $300 \times 10^{-5}\ mm^2/s$ down to zero. In the same units, water at blood temperature has a value of approximately[1] $290 \times 10^{-5}\ mm^2/s$, while dense tissue on the other-hand may have a value of typically $80 \times 10^{-5}\ mm^2/s$ [16]. Typical results from pre-clinical trials in regions of interest within a mouse tumour are shown from two of our collaborators in Figure 1. While the initial distribution may be well approximated by a Gaussian, or skewed on one side (typically towards high values) when there is necrotic tissue in the region of interest.

In previous studies in our group[2] effective treatments have been found to increase the mean ADC in later stages of treatment by approximately $\Delta ADC = 12 \times 10^{-5}\ mm^2/s$ i.e. 8% of the measured value or 1/3rd of a standard deviations of the approximating distribution. We must now convert this observation into an approximate model of change. Assuming the regions affected by treatment will have their ADC values elevated to on average that close to water (typically $210 \times 10^{-5}\ mm^2/s$) a proportion ($\alpha = 0.2$) of the organ must be affected by treatment in order to generate an increase in the mean of $12 \times 10^{-5}\ mm^2/s$. We will take this as an upper limit on the amount of change we might reasonably expect to see when attempting early detection of treatment effect.

In conventional studies the mean estimates are highly variable across subjects, and the within group standard deviation on the mean is of the order of 10 %. Under these circumstances it is impossible to identify a significant treatment effect ($P < 0.01$ for the null hypothesis) in a single individual, although significant effects do begin to emerge when taking an average over 15 or more individuals. An example is provided below in the Conclusions, using the tables derived from this current work. Unfortunately, the inability to observe significant effects on a case

[1] When attempting to obtain accurate measurements of ADC, even for such simple systems, there is a wide variety of results in the literature (e.g. $200.0 - 280.0 \times 10^{-5}\ mm^2/s$) [10, 7]. Our preferred result is taken from [4]. Average values for organs and metastases vary even more, some even beyond what might be regarded as physically sensible [16]. Here we have assumed expected means and variances will be provided by those using the statistical methods, so that what we believe these values to be currently is largely unimportant.

[2] For example the DREAM dataset provided by the Wolfson Molecular Imaging Centre of the University of Manchester.

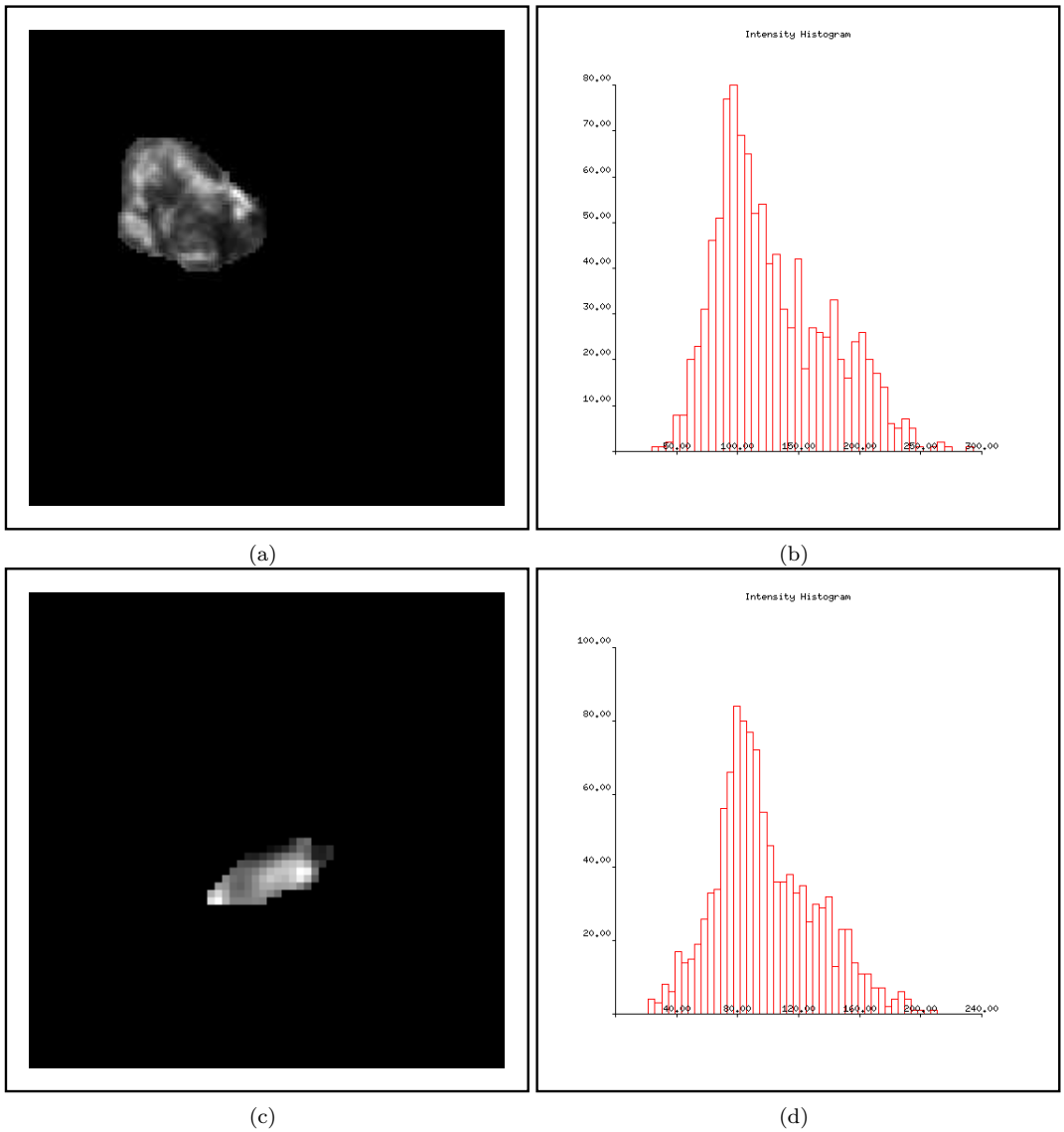

Figure 1: Regions of interest in ADC image for mouse tumour data (a) and (c), with respective ADC histograms in units of $mm^2/s \times 10^{-5}$ (b)(this slice) and (d) (entire ROI volume). These data illustrate a bimodal behaviour, due to heterogenous regions of tumour and necrotic tissues. These histograms are typical of such data, though the proportions of the two distributions can vary and in some there is almost no necrotic component.

by case basis (see for example [6] means this is not a suitable approach for patient management. Indeed, even for a group analysis we would benefit from a more sensitive mechanism of change detection.

There is a general lack of consistency between published ADC values in the literature. This can be due to several causes, including varying imaging protocols and variations in analysis methods. Effects such as these can however be reduced via standardisation. Our recent studies with ice-water phantoms lead us to believe that a high variability in measured ADC values can be accounted for by the lack of an absolute calibration, combined with spatially varying changes in the 'effective b' values. The precise details of this effect will vary depending on intervals between scans and the location of the region of interest, but the range of these variations (even on a 1.5T machine), can already be seen to be large enough to account for an overall systematic 10% error in reproducibility (i.e. it does not reduce by taking an average over a large region).

While the mean of a distribution may be statistically insensitive to the effects of drug treatment under these circumstances, there are other forms of statistically summary variable (for example the 95 percentile and differences of this to the median), which are expected to be both robust to significant variation in the initial distribution and also sensitive to the changes which accompany treatment response. While we may not have precise details of the distributions involved, it is still possible to investigate these methods by setting up simulation experiments designed to cover a range of possible circumstances. The purpose of this document is to quantify the various levels

of sensitivity for a variety of summary variables, and to investigate how the information obtained from phantom data can be used to develop better measurements and change detection methods. As only statistical sampling effects and systematic measurement problems are modelled this is a simplified case. Without additional effects, such as biological variability (for which we currently have no detailed model), our results must be considered a best case. In furture work the intention is to ensure that our image aquisition and analysis methodologies reduces these additional effects to negligible levels, so that our best case assessment of sensitivity to measurement of ADC change is realisable.

Methods

Several methods have been suggested as alternatives to the mean for the detection in change in ADC histograms. These include; kurtosis [3], maximum value [5], median and interquartile range [9]. Our criteria here are those of both sensitivity and robustness, particularly robustness to large fluctuations in a small number of data due to fitting errors. In addition the method needs to be simple and based upon familiar statistical principles if it is to be widely adopted.

The median and 95 percentile meet two of the above criteria by design. They are simple and their robustness stems from the complete insensitivity to some forms of change in distribution shape. Unlike a mean, maximum or kurtosis (which are all known to be highly sensitive to outlier data), data can be moved around within the distribution and provided these movements do not move data across these locations the measurements remain entirely unaffected. Of the three requirements it therefore only remains to assess their sensitivity.

Distribution Modelling

The accuracy of variables such as a mean, median or 95 percentile, is a function of two key variables, the quantity of data N and the width of the parent distribution σ_{ADC}. When generating tables for measured level of statistical significance, the effects of distribution width can be eliminated by re-normalising. As accuracy is always proportional to distribution width, the variable $X = ADC/\sigma_{ADC}$ has measurement precision which is only a function of data quantity. ADC data which is distributed around a mean of 132 $\times 10^{-5}$ mm^2/s with a width of 30 $\times 10^{-5}$ mm^2/s has a distribution as seen in Figure 2 (left). These means and variances can be obtained easily from real data. For skew or bi-modal distributions we can model the initial distribution with $\alpha > 0.0$, for example a value of 0.2 seems to be a good approximation to the data in Figure 1.

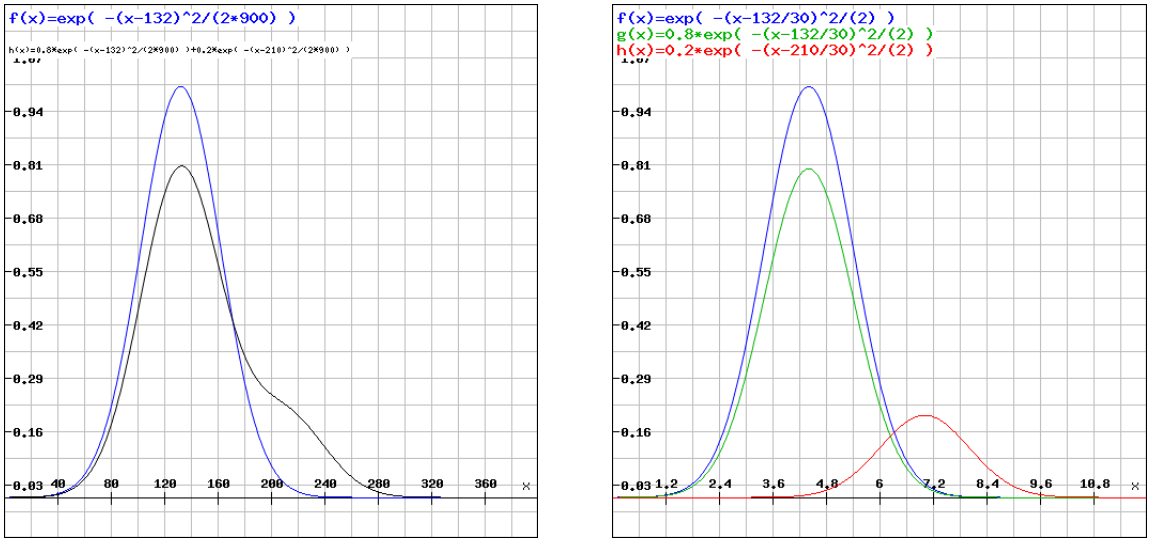

Figure 2: Modelled distribution of ADC values in units of $mm^2/s \times 10^{-5}$ (left), and normalised to distribution width (right). The blue curve is for pre-treatment distribution and the black curve is post-treatment. The green and red curves are the equivalent non-responding and responding distributions post-treatment with $\alpha = 0.2$ and $X = 2.6$ in our Monte-Carlo for Bimodal data (see main text).

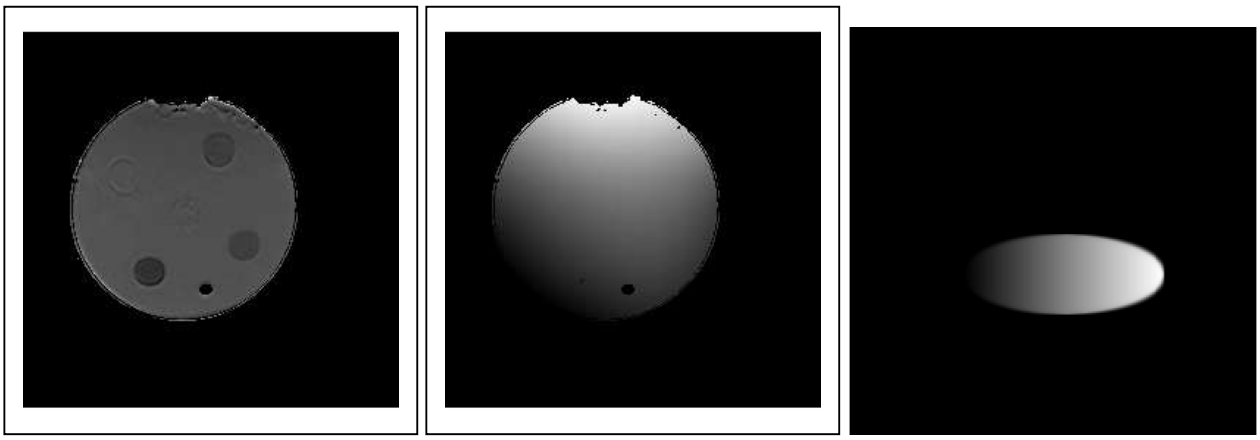

Figure 3: ADC map of the phantom (left), and the magnified ADC correction map within the phantom (middle); approximate systematic variation in ADC within a synthetic region of interest (right).

Monte-Carlo Modelling of Systematic Errors

The sensitivity to change of any summary statistic when using this X variable will be a function of the statistical and systematic errors on ADC. In our previous work we explained how statistical errors arise from sampling, while systematic effects (due to secondary physics processes in the scanner) can be modelled as a spatial multiplicative variation in the measurements of ADC. Although work has been published regarding the origin of statistical error [2], [1], [15], we have stated previously, and now show here, that it is the systematic errors which generate the limits on reproducibility of derived statistical summaries, such as the median. This is because the statistical accuracy of a summary variable generally improves by a factor proportional to the square root of the data quantity ($\sqrt{N}$), while systematic errors constitute a source of fixed irreducible bias. For data with large N the random statistical errors rapidly become negligible. In accordance with our previous work [17], we now model the systematic errors as two multiplicative factors applied to the Monte-Carlo data (Figure 3). The first factor is applied to the data which is perturbing around the mean, and is intended to model the mean ADC error within a randomly selected region of interest. So it not only widens the data histogram, but also shifts the mean position. The second factor is intended to model the local relative rescaling of ADC across a region of interest, and only applied in a way which widens the histogram without shifting its mean position.

Assuming the spatial variation is randomly sampled from the observed inhomogeneity of ADC measurement via the selection of a region of interest, the likely distribution for these parameters can be inferred from phantom data by measuring the standard deviation of spatial variation of multiplicative ADC inhomogeneity ($\sigma_s \approx 6\%$), and the standard deviation of its spatial derivative scaled to the mean region size (σ_r). For this work we have assumed a fixed upper limit corresponding to a region of interest with a maximum linear dimension of 1/6 th of the measurable region. This is equivalent to approximately 4000 pixels ($\sigma_r \approx 1\%$). We observe that σ_r has only a minor impact on results. As a simplification we can therefore assume these values have a fixed relationship ($\sigma_r = \sigma_s/6$) which corresponds to a largest allowable region of interest size. These values can then be used as the basis for statistical effect size in a Monte Carlo simulation to quantify measurement repeatability and statistical sensitivity in the presence of these effects.

Selection of Summary Statistics

In this study we have investigated five separate approaches to quantifying a change between ADC distributions within a fixed region of interest following treatment. These are:

- Changes in the mean $\bar{X}' - \bar{X}$.
- Changes in the median $X'_m - X_m$.
- Changes in the 95 percentile $X'_{95\%} - X_{95\%}$.
- Changes in the difference between the median X_m and 95 percentile, i.e. $(X'_{95\%} - X'_m) - (X_{95\%} - X_m)$.
- Changes in the difference in 95 percentile scaled to the median, i.e. $X_m X'_{95\%}/X'_m - X_{95\%}$.

In order to evaluate and calibrate a null hypothesis statistic we are interested in knowing two key characteristics of the method. Firstly, the accuracy of the difference measurement when there is no change expected beyond the variations in statistical and systematic error (for the initial distributions; Gaussian (A) and Bimodal (B)), and secondly the scale of variations due to effects of likely change (C). The first of these can be used to compute P values for the null hypothesis in longitudinal studies, while the second is used to allow us to predict the likely degree of effect which would be visible in a successful experiment, i.e. a power calculation.

MonteCarlo Experiments and Data

We evaluate these processes as follows;

- (A) generate repeat Monte-Carlo data (for various quantities pixel data) for a uni-modal Gaussian (tens of thousand histograms); we measure the sensitivity for repeat measurement for the variables of interest.
- (B) generate repeat Monte-Carlo data for a bi-modal distribution ($\alpha = 0.2$), and again measure the sensitivity for repeat measurement.
- (C) generate repeat Monte-Carlo data (for various quantities of pixel data) for a bi-modal Gaussian (tens of thousand histograms) described as a proportion of responding voxels α and normalised scale of response $X = ADC/\sigma_{ADC}$; we measure the mean change in all variables.

In accordance with our earlier discussion of expected ADC values, the normalised ADC change (ΔX) is varied between 0.5 and 3.0 S.D. in order to encompass the expected degree of change in a range of organs. Here, the proportion of responding voxels is limited to 30 % and below in order to concentrate specifically on early response. These studies are conducted for a range of systematic effects ranging from $\sigma_s = 0$ up to 12%. with a value of $\sigma_s = 6\%$ expected to be appropriate for our 1.5T scanner.

The results for (A) as a function of scale of systematic error as a function of data quantity (500-4000 pixels) are summarised in tables 1-4. It can be seen that for moderate systematic effects ($\sigma_s > 2\%$) the reproducibility is dominated by systematic effects, also simpler measures based upon absolute ADC loose all sensitivity.

The results for (B) are similarly summarised in tables 5-8. These results show a slight reduction in the accuracy of repeat measurements for some variables.

The results for (C) are shown in Figures 4-8. Now using tables 1-4 (dependant upon data quantity) it is possible to predict the sensitivity of measured change for each variable and estimate the level of statistical significance in comparison to the changes induced by various values of α and X. Transferring typical values for 2000 pixels, and $\sigma_s = 6\%$ and $\sigma_r = 1\%$ to Figures 4-8 we can see that only relative measures are capable of detecting longitudinal change in an individual for the range of treatment responses modelled. The final method ($X_m X'_{95\%}/X'_m - X_{95\%}$) is up to 12 times more sensitive than using a mean for the detection of treatment response for 4000 points, $\alpha = 10\%$, $\sigma_s = 6\%$.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.031	0.064	0.066	0.075
0.02 + 0.0033	0.099	0.155	0.082	0.075
0.04 + 0.0066	0.191	0.289	0.115	0.077
0.06 + 0.0100	0.283	0.426	0.156	0.079
0.09 + 0.0150	0.424	0.636	0.221	0.085
0.12 + 0.0200	0.565	0.845	0.288	0.093

Table 1: Null hypothesis sensitivities for 500 points; sensitivities listed for the mean point (σ_{mean}) can equally be used for the median point.

Conclusions

The current work has sought to investigate the effects of statistical and systematic errors on various summary parameters measured from ADC histograms. In order to deal with the large possible variations in specific distributions we have done this by modelling a range of possibilities covering a range of values seen in typical datasets. The results are summarised as a series of tables, which can be used if the specific distribution characteristics are known.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.022	0.045	0.047	0.053
0.02 + 0.0033	0.096	0.148	0.066	0.053
0.04 + 0.0066	0.190	0.286	0.105	0.055
0.06 + 0.0100	0.283	0.424	0.148	0.059
0.09 + 0.0150	0.425	0.636	0.217	0.066
0.12 + 0.0200	0.566	0.846	0.285	0.075

Table 2: Null hypothesis sensitivities for 1000 points; sensitivities listed for the mean point (σ_{mean}) can equally be used for the median point.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.015	0.032	0.033	0.037
0.02 + 0.0033	0.095	0.144	0.057	0.038
0.04 + 0.0066	0.188	0.282	0.099	0.041
0.06 + 0.0100	0.283	0.424	0.145	0.045
0.09 + 0.0150	0.424	0.635	0.214	0.054
0.12 + 0.0200	0.564	0.844	0.283	0.065

Table 3: Null hypothesis sensitivities for 2000 points; sensitivities listed for the mean point (σ_{mean}) can equally be used for the median point.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.011	0.023	0.023	0.026
0.02 + 0.0033	0.094	0.142	0.052	0.027
0.04 + 0.0066	0.189	0.283	0.097	0.031
0.06 + 0.0100	0.282	0.422	0.142	0.036
0.09 + 0.0150	0.424	0.634	0.212	0.046
0.12 + 0.0200	0.565	0.845	0.282	0.058

Table 4: Null hypothesis sensitivities for 4000 points; sensitivities listed for the mean point (σ_{mean}) can equally be used for the median point.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{50\%}$	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.031	0.041	0.078	0.084	0.082
0.02 + 0.0033	0.107	0.109	0.178	0.103	0.082
0.04 + 0.0066	0.207	0.206	0.327	0.143	0.083
0.06 + 0.0100	0.309	0.307	0.483	0.192	0.083
0.09 + 0.0150	0.459	0.456	0.711	0.266	0.085
0.12 + 0.0200	0.609	0.606	0.936	0.340	0.090

Table 5: Null hypothesis sensitivities for a Bimodal distribution ($\alpha = 0.2$) with 500 points.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{50\%}$	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.022	0.029	0.055	0.059	0.057
0.02 + 0.0033	0.105	0.105	0.168	0.083	0.058
0.04 + 0.0066	0.206	0.204	0.322	0.130	0.058
0.06 + 0.0100	0.306	0.304	0.477	0.182	0.060
0.09 + 0.0150	0.459	0.455	0.709	0.261	0.063
0.12 + 0.0200	0.610	0.606	0.937	0.336	0.068

Table 6: Null hypothesis sensitivities for a Bimodal distribution ($\alpha = 0.2$) 1000 points.

We have currently assumed effects due to region of interest selection, and secondary imaging problems (such as distortion, fat suppression and wrap around), can all be reduced to acceptable levels via appropriate methodology

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{50\%}$	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.015	0.020	0.039	0.042	0.041
0.02 + 0.0066	0.103	0.103	0.163	0.071	0.041
0.04 + 0.0100	0.205	0.203	0.319	0.122	0.042
0.06 + 0.0100	0.307	0.304	0.477	0.177	0.043
0.09 + 0.0150	0.458	0.454	0.707	0.257	0.047
0.12 + 0.0200	0.608	0.604	0.933	0.332	0.054

Table 7: Null hypothesis sensitivities for a Bimodal distribution ($\alpha = 0.2$) with 2000 points.

$\sigma_s + \sigma_r$	σ_{mean}	$\sigma_{50\%}$	$\sigma_{95\%}$	$\sigma_{95\%-50\%}$	$\sigma_{\Delta 95\%}$
0.00 + 0.0000	0.011	0.014	0.028	0.030	0.029
0.02 + 0.0033	0.103	0.102	0.161	0.065	0.029
0.04 + 0.0066	0.205	0.202	0.318	0.119	0.030
0.06 + 0.0100	0.306	0.303	0.476	0.174	0.032
0.09 + 0.0150	0.461	0.457	0.711	0.256	0.036
0.12 + 0.0200	0.611	0.607	0.938	0.333	0.045

Table 8: Null hypothesis sensitivities for a Bimodal distribution ($\alpha = 0.2$) with 4000 points.

and quality assurance. This is an area for future work.

Our data demonstrate that for the entire range of distributions modelled, absolute measurements of ADC are detrimentally affected by the spatial variation in "effective b" values as measured in our phantom work. However, significant improvements can be obtained by using more appropriate summary statistics and also measurements sensitive to relative differences in ADC distribution rather than absolute measurement. We would also expect spatial correction of ADC using the phantom might reduce the systematic effects to 1/3 of their initial value (i.e. $6\% \rightarrow 2\%$), allowing absolute measurements to once again be competitive. However, transferring such spatial calibration to a clinical environment may be problematic. Alternatively, improvements may yet be seen in the base level performance of modern scanners, particularly the high field machines (3 Tesla and above).

These results have been generated for a specific scenario of ADC change measurement, for a mean ADC to σ_{ADC} ratio of the target organ of 132/30 (i.e. 4.4). Significant changes in this value would lead to differences in the significance tables. In particular, the sensitivities for all methods but for the median to 95 percentile measurement can be corrected to better than 10% by adjusting the assumed σ_s by $ADC/(4.4 * \sigma_{ADC})$ for $ADC/\sigma_{ADC} < 10$. Consequently, these tables should be correct within a factor of 30 % for values of $3.0 < ADC/\sigma_{ADC} < 6.0$ (see Appendix B). Effects measured in units of standard deviation can be treated as a z score and converted to P values via the error function (erf())), in the usual way.

The response of ADC to therapy is complicated, with both upward and downward trends seen at different stages [13]. Power calculations for the design of drug trials therefore require that the mean value of ADC and expected percentage of responding tissue (at some specified interval following treatment) to be predicted by pre-clinical work. These values can then be used to estimate the level of significance obtainable for a scanner of known σ_s and σ_r. These values will be measured from phantom data.

For example, assume;

- an organ has data consitstent with a Gaussian, with mean ADC of $150 \times 10^{-5}\ mm^2/s$,
- a width of $50 \times 10^{-5}\ mm^2/s$
- a predicted treatment response generating a mean ADC of $210 \times 10^{-5}\ mm^2/s$ in 10% of tissue.
- a region interest size of 2000 pixels
- a scanner with $\sigma_s = 12\%$, $\sigma_r = 2\%$.

Staring with the median difference, our four statistical methods show a gradual improvement in sensitivity (Figures 5-8), with the normalised 95 percentile being the best. The robustness to outliers of the median difference in comparison to the mean, is bought at the cost of a slight loss in measurable effect. For the 95 percentile the sensitivity to changes in ADC/σ_{ADC} of 0.844 (table 3). This has a normalised variable of $\Delta X = 1.2$ which

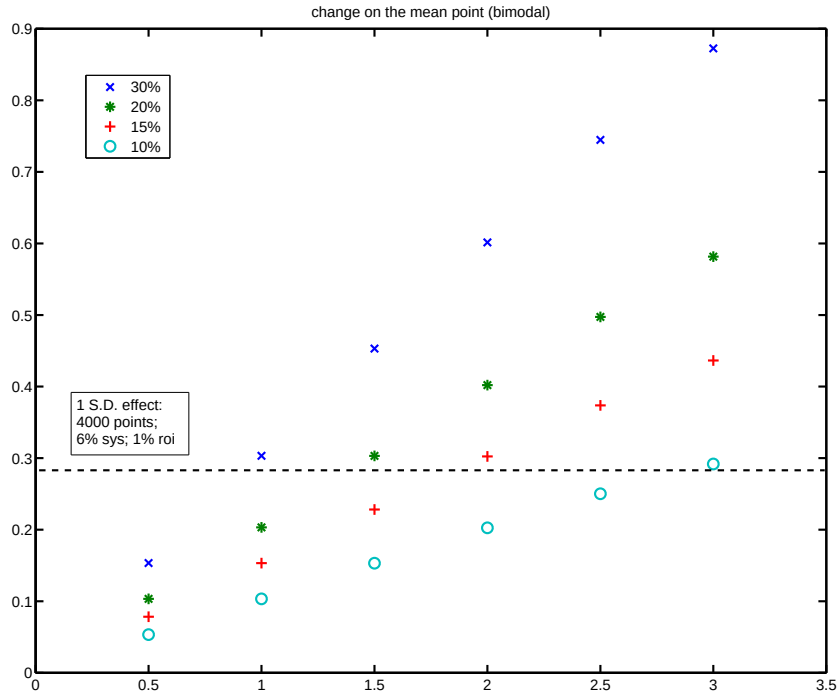

Figure 4: Change of the mean point against the number of standard deviations used as distance between the mean point of each pair of Gaussian distributions; the four curves correspond to percentages of data belonging to the second Gaussian in the bimodal distribution (with the larger mean point).

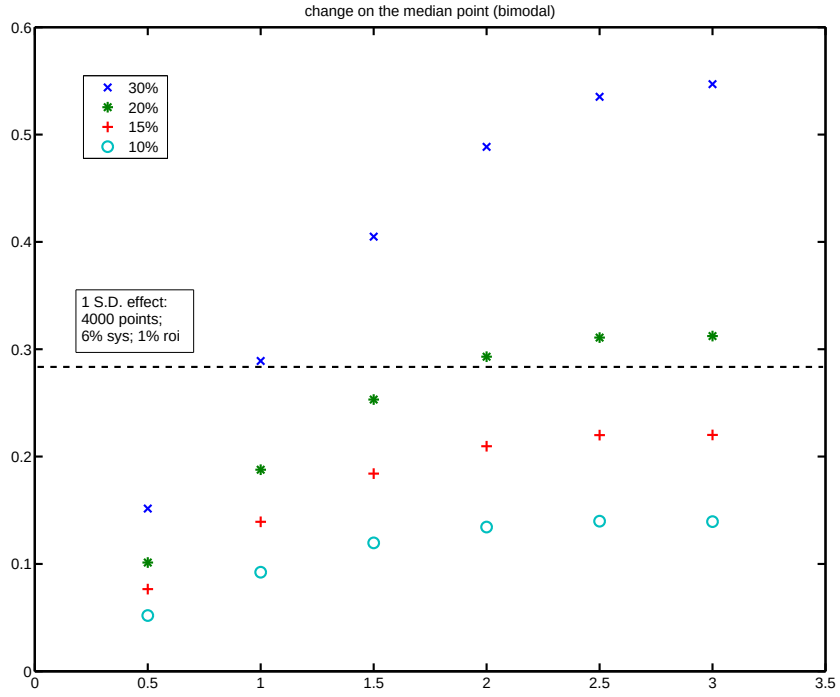

Figure 5: Same plot as Fig. 3, but for the change on the median point.

corresponds to an expected change of 0.22 (interpolated from Figure 6). As this value is much less than 0.844 we

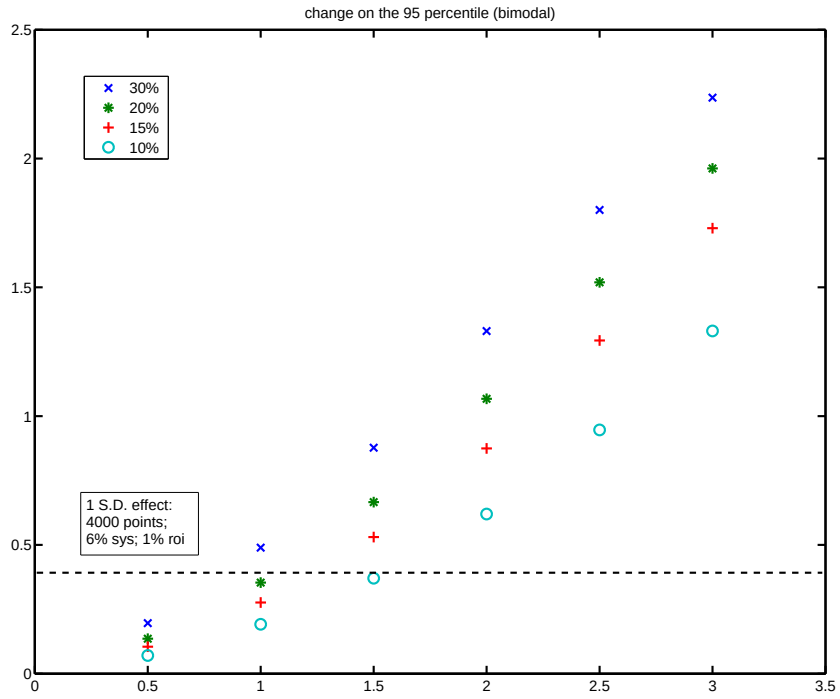

Figure 6: Same plot as Fig. 3, but for the change on the 95 percentile.

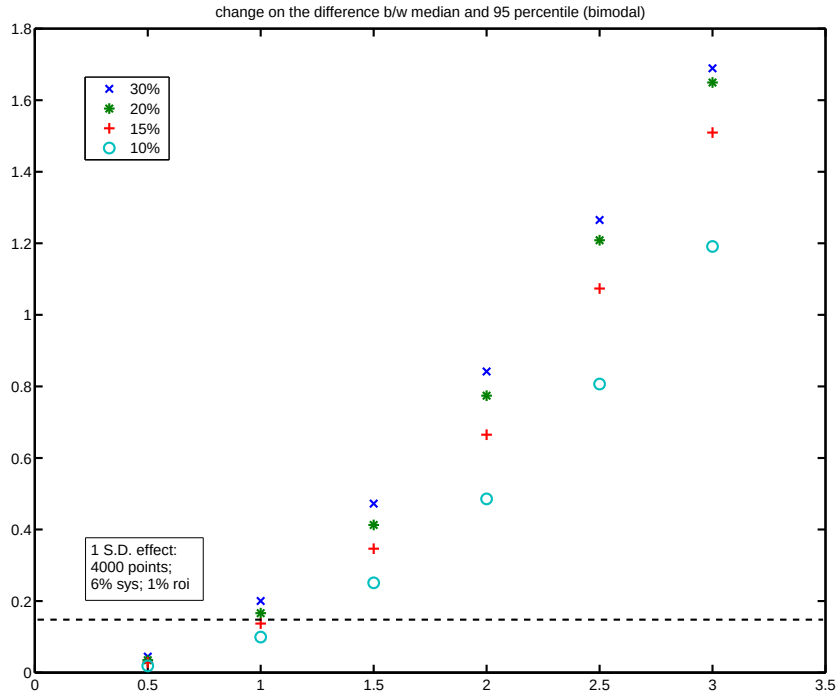

Figure 7: Same plot as Fig. 3, but for the change on the difference between the 95 percentile and the median points.

need to combine multiple measurements in order to reach statistical significance (see Appendix A). To reach a 2 S.D. effect in this study we would need $(2 * 0.844/0.22)^2 = 59$ subjects. Correcting for spatially varying ADC using a phantom calibration would reduce systematic effects by a factor of 3, and thereby increase our sensitivity

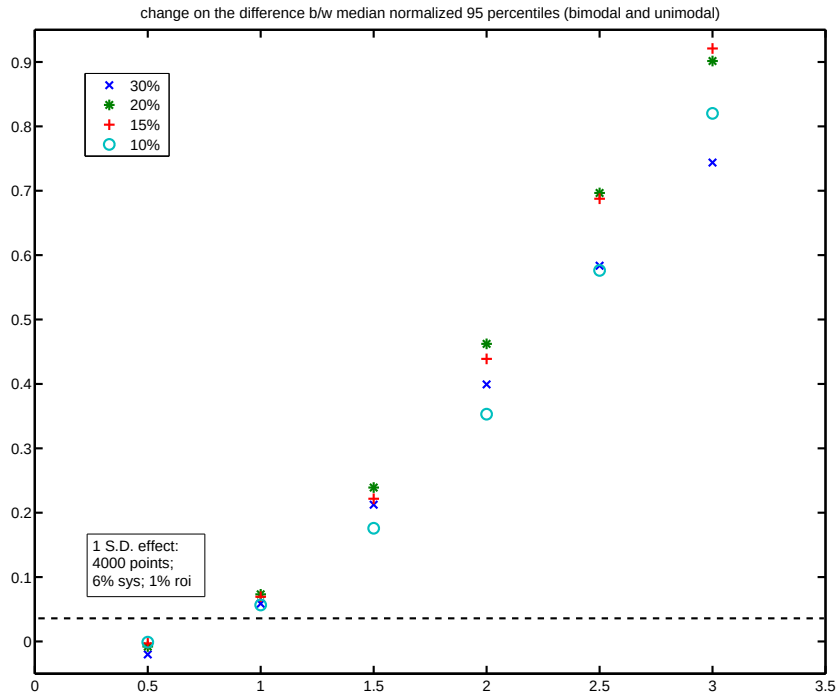

Figure 8: Same plot as Fig. 3, but for change in the 95 percentile from the unimodal distribution and the normalised 95 percentile from the bimodal distribution; normalisation is based on the ratio of the median points.

to 0.28 (table 3), and reducing the numbers of subjects required to see a 2 S.D. effect to $(0.56/0.22)^2 = 6.6$.

Alternatively, when using the median normalised 95 percentile difference we are sensitive to changes in ADC/σ_{ADC} of 0.065 (Table 3). The expected effect for this variable is 0.1 (interpolated from Figure 8), at $0.1/0.065 = 1.8$ S.D this is potentially sufficient for a decision in patient management. The capabilities for detection of change across the full range of methods is shown in Table 9.

In our approach, the uncertainties associated with imprecise estimation of a S_Y based upon the sample are replaced by imprecision of the Monte-Carlo model used to predict measurement sensitivity (Appendix B). While the first of these we can do nothing about, the second is left to our own discretion. The Monte-Carlo approach also allows us to obtain results from sample sizes as small as $n = 1$, which is entirely precluded in conventional t-tests.

The main purpose here was only to demonstrate that slight changes in the way we summarise distributions can have large benefits with regard to statistical efficiency. As seen in the introduction, real data sets may not be well approximated by a Gaussian, with skewing to one side. In this case we can model this with an initial non-zero value for α. Tables 5-8 show a marginal loss in sensitivity, but at the same time the expected change in measured variable is also modified. On this basis, some of the techniques above appear to loose sensitivity as $\alpha \to 0.5$. The value needed to approximate data will depend upon the quantity of necrosis within the region of interest at the time of the pre-treatment scan. This will be heavily influenced by or methods for region of interest selection. For pre-clinical work we have seen that this may be as high as $\alpha = 0.2$. In addition any region of interest needs to be selected so that the ADC histogram is constructed in a highly repeatable way and such that a sufficiently large proportion of its contents are expected to respond to treatment.

Future Work

Now that we can see the benefits of more sensitive statistical estimates it would be tempting to suggest that we simply use the most sensitive for all experiments. However, we do not know that spatial calibration of ADC (Method 2 above), will be possible using the phantom, it remains to be tested. More work is needed to develop methods for the extraction of suitable estimates of σ_r and σ_s from phantom data. This requires more data sets than are currently avaliable. For preclinical work we will not have the benefit of the phantom, however, the tables

method	S.D. of effect (M1)	subjects (M1)	S.D. of effect (M2)	subjects (M2)
$\bar{X}' - \bar{X}$	0.20	100	0.58	12
$X'_m - X_m$	0.18	123	0.52	15
$X'_{95\%} - X_{95\%}$	0.26	59	.78	6.6
$(X'_{95\%} - X'_m) - (X_{95\%} - X_m)$	0.57	12	1.6	1.5
$X_m X'_{95\%}/X'_m - X_{95\%}$	1.8	1.2	2.4	0.70

Table 9: Results for the example (2000 points); level of measured effect and number of subjects required to reach 2 S.D; M1 refers to the results without and M2 refers to those with spatial ADC correction.

provided here can be used to infer effective values of these parameters from observed reproducibility. In this respect the mean difference may prove to be the most useful, as this measure is the one most affected by effective b changes. The difference in median normalised 95 percentile was specifically chosen to be maximally robust to this effect.

What we have described are Null Hypothesis tests, which will detect not only the changes we seek, but potentially also any failure in methodology or measurement which introduces measurement bias. These represent the best case scenario for the detection of change in a clinical setting using these summary statistics. In order to use the more sensitive methods, we will need to develop techniques for robust identification of regions of interest and associated quality assurance. so that we can guarantee the uncertainties modelled in these statistical methods dominate in the process of repeat measurement. Without this step our hypothesis tests will simply be identifying poor region of interest repeatability or biological variation. These techniques thus need to be tested in reproducibility studies. The intention now is to apply these methods to preclinical and human data, for circumstances where inhomogeneity is known to be present in the spatial distribution of response.

Appendix A: Power Scaling of T-tests with Sample Size

We assume that a paired t-test is computed from the means of two groups of summary statistic $\bar{Y}_1$ and $\bar{Y}_2$ from n paired samples, and the expected standard deviation S_Y of these differences, to give a student t variable

$$t_n \;=\; \frac{|\bar{Y}_1 - \bar{Y}_2|}{S_Y\sqrt{1/n}} \qquad (1)$$

So that the significance of any given difference scales with $\sqrt{n}$. Therefore the number of subjects needed in order to reach a specified level of target significance T is given by

$$n \;=\; (T^2/ < t_1 >)^2$$

where in this case $< t_1 >$ is of course the average expected size of the statistical effect for one sample, or the expected mean change divided by its reproducibility.

Appendix B: Small Sample Stability of T-Tests

We wish to make some remarks here regarding the accuracy of statistical significance tables by drawing a comparison between alternative methods for estimation of S_Y. In more conventional studies the value of S_Y is not provided from Monte-Carlo but estimated from the available samples. For large samples a probability can be computed on the assumption that t'_n is drawn from a Gaussian distribution. The number of degrees of freedom available for this estimate are then $n-1$, so that the definition of t_n needs to be adjusted to

$$t'_n \;=\; \frac{|\bar{Y}_1 - \bar{Y}_2|}{S_Y\sqrt{1/(n-1)}} \qquad (2)$$

However, this distribution is affected by small samples, so that our noisy estimate t'_n no longer has unit variance. Using error propagation we can make an approximation of the magnitude of this effect as

$$var(t'_n) \;\approx\; 1 \;+\; (\partial t_n/\partial S_Y)^2 var(S_Y)$$

where $var(S_Y)$ is the square of the standard error on S_Y and is given by $S_Y/2(n-1)$. Then

$$var(t'_n) \;\approx\; 1 + t_n^2/2(n-1)$$

So that S_Y is not a valid basis for the estimate of distribution width, indeed assuming a Gaussian distribution for samples sizes less than $n < t_n^2$ will significantly underestimate the probability of the null hypothesis. A first order correction can be made using the following formula[3].

$$t''^2_n = (n-1)(\sqrt{1 + 2t'^2_n/n} - 1) \qquad (3)$$

so allowing us to use the terminology of " an equivalent number of S.D." of the measured effect, with reasonable accuracy for $n >= 8$ (Table 6). This can therefore be used as a good approximation to using the correct T-distribution, which is expected to remove any remaining bias.

n	eq(1)	eq(2)	eq(3)
5	6.6	4.8	1.15
6	5.1	3.8	0.92
8	3.4	2.6	0.72
10	2.59	2.04	0.62
15	1.68	1.40	0.53
20	1.32	1.13	0.51
50	0.79	0.74	0.50
100	0.63	0.61	0.50

Table 10: Percentage of null hypothesis failure rates at $p = 0.005$ for a two sided t-test of various approximate forms, estimated from a Monte-Carlo with 10,000,000 sets of n samples. The required value is by definition 0.5.

Having confirmed the approximate behaviour of t_n, we can say that in comparison to knowing the S_Y in advance, t_n estimated using the sample S_Y will incur an approximate additional estimation error of $t_n\sqrt{1/(2n-2))}$ so that a $t_9 = 3.0$ effect is approximately 3.0 ± 0.75. Using the T-distribution for construction of the hypothesis probability takes appropriate account of this additional source of error but does not remove it. For $n < 20$, 10 % of S_Y's (and hence t_n) are no better than 30% accurate. Thus having significance tables generated from Monte-Carlo which are accurate to around the same level can be seen as consistent with, and generally better than, current practice with $n < 20$.

References

[1] O. Brihuega-Moreno, F.P. Heese and L.D. Hall, Optimization of Diffusion Measurements Using Cramer-Rao Lower Bound Theory and its Application to Articular Cartilage, *Magnetic Resonance in Medicine*, 50(5):1069-1076, 2003.

[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(4):983-987, 2011.

[3] J. Hansmann, A. Lemke, J. Wambsganss, M. Meyer, M. Reichert, S.O. Schoenberg, and U.I. Attenberger, Impact of Kurtosis Diffusion Weighted Imaging on the Detection of Liver and Kidney Abnormalities at 1.5 and 3 Tesla, *ISMRM*, Melbourne, Australia, May 2012.

[4] M. Holz, S.R. Heil and A. Sacco, Temperature Dependant Self-Diffusion Coefficients of Water and Six Selected Molecular Liquids for Calibration of Accurate H1 NMR PFG Measurements, Phys., Chem., (DOI 10.1039/b005319h), 2, 4740-4742, 2000.

[5] M. Jolapara, S.N. Patro, C. Kesavadas, J. Saini, B. Thomas, A.K. Gupta, N. Bodhey and V.V. Radhakrishnan, Can Diffusion Tensor Metrics Help in Preoperative Grading of Diffusely Infiltrating Astrocytomas? A Retrospective Study of 36 Cases, *Neuroradiology*, 53(1):63-8, 2011.

[6] M. Kawamura, H. Satake, S. Ishigaki, A. Nishio, M. Sawaki and S. Naganawa, Early Prediction of Response to Neoadjuvant Chemotherapy for Locally Advanced Breast Cancer Using MRI, *Nagoya J. Medical Science*, 73(3-4):147-156, 2011.

[3]Setting the observed t_n value to $t_n\sqrt{1 + t_n^2/(2n-2)}$ and solving the resulting quadratic generates this function but with $(n-1)$ in the square-root term. Empirically this modified formula corrects the resulting t_n in all cases 30% better.

[7] S.M. Kealey, T. Aho, D. Delong, D.P. Barboriak, J.M. Provenzale and J.D. Eastwood, Assessment of Apparent Diffusion Coefficient in Normal and Degenerated Intervertebral Lumbar Disks: Initial Experience, *Radiology*, 235(2):569-74, 2005.

[8] Dow-Mu Koh and Harriet C. Thoeny (editors), Diffusion-Weighted MR Imaging: Applications in the Body, Springer, New York, 2009.

[9] F.H. Miller, Y. Wang, R.J. McCarthy, V. Yaghmai, L. Merrick, A. Larson, S. Berggruen, D.D. Casalino and P. Nikolaidis, Utility of Diffusion-Weighted MRI in Characterization of Adrenal Lesions, *American Journal of Roentgenology*, 194(2):179-185, 2010.

[10] R. Mills, Self-Diffusion in Normal and Heavy Water in the Range $1-45°$, *The Journal of Physical Chemistry*, 77(5):685-688, 1973.

[11] 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.

[12] W.B. Pope, X.J. Qiao, H.J. Kim, A. Lai, P. Nghiemphu, X. Xue, B.M. Ellingson, D. Schiff, D. Aregawi, S. Cha, V.K. Puduvalli, J. Wu, W.K. Yung, G.S. Young, J. Vredenburgh, D. Barboriak, L.E. Abrey, T. Mikkelsen, R. Jain, N.A. Paleologos, P.L. Rn, M. Prados, J. Goldin, P.Y. Wen and T. Cloughesy, Apparent Diffusion Coefficient Histogram Analysis Stratifies Progression-free and Overall Survival in Patients with Recurrent GBM Treated with Bevacizumab: A Multi-center Study, *J Neurooncol.* 108(3):491-8, 2012.

[13] C. Schraml, N.F. Schwenzer, P. Martirosian, M. Bitzer, U. Lauer, C.D. Claussen and M. Horger, Diffusion Weighted MRI of Advanced Hepatocellular Carcinoma during Sorafenib Treatment: Initial Results, *AJR*, 193:301-307, 2009.

[14] A. Srinivasan, T.L. Chenevert, B.A. Dwamena, A. Eisbruch, K. Watcharotone, J.D. Myles and S.K. Mukherji, Utility of Pretreatment Mean Apparent Diffusion Coefficient and Apparent Diffusion Coefficient Histograms in Prediction of Outcome to Chemoradiation in Head and Neck Squamous Cell Carcinoma, *J Comput Assist Tomogr*, 36(1):131-7, 2012.

[15] G. Tang, Y. Liu, W. Li, J. Yao, B. Li and P. Li, Optimization of b Value in Diffusion-weighted MRI for the Differential Diagnosis of Benign and Malignant Vertebral Fractures, *Skeletal Radiol.*, 36(11):1035-41, 2007.

[16] B. Taouli and D.M. Koh, Diffusion-weighted MR Imaging of the Liver, *Radiology*, 254(1): 47-66, 2010.

[17] 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.*

[18] B. Turkbey, O. Aras, N. Karabulut, A.T. Turgut, E. Akpnar, S. Alibek, Y. Pang, S.M. Erturk, R.H. El Khouli, D.A. Bluemke and P.L. Choyke, Diffusion-weighted MRI for Detecting and Monitoring Cancer: A Review of Current Applications in Body Imaging, *Diagnostic and Interventional Radiology*, 18(1):046-059, 2012.