Tina Memo No. 2017-007

Internal, preprint submitted to Scientific Reports, see also 2014-007 for original analysis.

A Data-driven Statistical Model that Estimates Measurement Uncertainty, Improves Interpretation of ADC Reproducibility in a Multi-site Study of Liver Metastases.

Ryan Pathak, Hossein Ragheb, Neil Thacker et. al.

Last updated
2 /12 / 2016

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

A data-driven statistical model that estimates measurement uncertainty, improves interpretation of ADC reproducibility in a multi-site study of liver metastases

*Ryan Pathak [1], Hossein Ragheb [1], Neil A Thacker [1], David M Morris [1], Houshang Amiri [2], Joost Kuijer [3], Nandita M deSouza [4], Arend Heerschap [2], Alan Jackson [1]

1. University of Manchester, Wolfson Molecular Imaging Centre, Manchester, Lancashire, UK.
2. Radboudumc, Radiology and Nuclear Medicine, Nijmegen, Gelderland, NL
3. VU University Medical Center, Physics & Medical Technology, PO Box 7057 Amsterdam, NL 1007MB
4. Institute of Cancer Research, MRI Unit, Downs Road, Sutton, Surrey, UK SM2 5PT

Confidence in the accuracy of Apparent Diffusion Coefficient (ADC) to reflect tumor cell density is important, as ADC is a potential quantitative imaging biomarker for early treatment changes.

We propose a strategy to improve interpretation of ADC reproducibility. A data-driven model describes sources of variation that would influence repeated measurements. Observed ADC is then standardized against this estimation of uncertainty for a given repeatability measurement.

20 patients were recruited prospectively and equitably across 4 sites, and scanned twice (test-retest) within 7 days. Repeatability measurements of defined regions (ROIs) of tumor and normal tissue were quantified as percentage change in mean ADC (test vs. re-test) and then standardized against an estimation of uncertainty. Multi-site reproducibility, (quantified as width of the 95% confidence bound between the lower confidence interval and higher confidence interval for all repeatability measurements), was compared before and after standardization to the model.

The 95% confidence interval width used to determine a statistically significant change reduced from 21.1% to 2.7% after standardization. Small tumor volumes and respiratory motion were found to be important contributors to poor reproducibility.

A look up chart has been provided for investigators who would like to estimate uncertainty from statistical error on individual ADC measurements.

Introduction

Diffusion weighted imaging (DWI) is a Magnetic Resonance Imaging (MRI) sequence acquisition that is sensitive to water mobility[1]. Regions of high cell density or reduced extra-cellular space will restrict diffusion of water relative to surrounding tissue, providing the contrast necessary to detect signal differences. Increases in extravascular-extracellular space due to cell death result in increased free water diffusion. Consequently, apparent diffusional coefficient (ADC) derived from diffusion weighted MRI has received considerable attention as a potential biomarker of early response to cytotoxic therapies[2]. The ADC is the decay constant, calculated from 2 or more DWI images, that have been acquired with increasing sensitivity to water mobility. A high ADC corresponds to increased water mobility towards free diffusion, and conversely a low ADC corresponds to restricted diffusion. In a densely cellular homogeneous tumor, such as lymphoma, treatment-related ADC changes may be as high as 50%[3], however treatment responses may be heterogeneous due to regional micro environmental factors or genetic variation[4-6]. A recent animal model study of ovarian tumors has shown on average a 7.5% increase in mean ADC after treatment, compared to control mice, however attributed spatial heterogeneity to variations in tumor ADC response[6].

In therapeutic studies using ADC early treatment induced changes are typically in the range of 10-30%. Statistically, in order to detect a 10% change in mean ADC for an individual lesion, with 95% reliability, a test-retest repeatability of 3-4% is required (assuming Gaussian distribution). If repeatability is worse than this, then the sensitivity to a 10% change in mean ADC will be lost, and with it, our ability to detect true biological change of this magnitude. A repeatability of 3% may be difficult to achieve, particularly in multi-site, multi-vendor trials, although studies in phantoms and homogenous healthy liver taken across multiple sites, have shown repeatability of 1-4% and 3- 7% respectively[9]. To our knowledge there is no published data to describe ADC reproducibility of liver metastases in a multi-site, multi-vendor setting. Factors that negatively affect repeatability relate to the tumor itself (size[10], heterogeneity[11] and site[12]), image quality (signal to noise ratio (SNR)[13], motion[14]), curve-fitting techniques[15] and errors related to the MR system[16]. Voxel-wise quantitative DWI in the liver is specifically degraded by respiratory motion artifact, with little improvement and mixed results when using on-table compensation methods

such as navigator echo and respiratory gating[17-19]. Consequently, a change in ADC due to measurement errors may be interpreted as disease progression or response where response thresholds are derived from group-wise reproducibility data.

The primary endpoint of this study was to define a statistical model of predictable sources of variability that contribute to measurement error, and fit this to observed data in order to quantify the level of uncertainty in mean ADC repeatability. Through standardization of repeatability measurements for predictable sources of statistical variability that contribute to uncertainty in the mean ADC, we sought to increase our confidence in detecting genuine post treatment changes for future studies.

Materials and methods

Patients

Ethical approval was obtained at four sites (University of Manchester, The Royal Marsden Hospital London, The VU University Medical Centre, Amsterdam, Radboud University Nijmegen Medical Centre) to recruit 5 patients per site for this study. Formal written consent was recorded for each volunteer. Inclusion criteria included; Histological diagnosis of primary colorectal carcinoma, radiological evidence of at least one liver metastasis (minimum volume $2cm^3$), new diagnosis or no ongoing treatment. Exclusion criteria included; Contraindication to MRI, ongoing treatment. Patients who met the inclusion criteria were scanned consecutively, as and when they appeared at their respective oncology clinic, prior to any new treatment commencing.

Image acquisition

Patients were imaged twice within 7 days, using 1.5T MR systems from 3 vendors (Table 1). DWI parameters were as follows; b value images for 3 orthogonal gradient directions, 4 signal averages per image, free-breathing single shot echo-planar sequence (SS-EPI), spectral attenuated inversion recovery (SPAIR) fat suppression, 5 mm axial slice thickness, 40 slices with no inter-slice gap, target FOV of 380 mm adjusted to patient size (Range 380 to 400 mm), bandwidth 1400-1800 Hz per pixel, pixel size of 3 x 3 mm, acquisition matrix 128 x 112 or 128 x 128.

Image analysis

A single lesion was chosen based on size (the largest visible single tumor or indistinguishable tumor conglomerate with a continuous circumference that was ≥ 2 cm^3) and location (right lobe, away from the heart or diaphragm where possible). A single observer manually outlined whole tumor 3-dimensional (3D) regions of interest (ROI) from b-100 images, for each test-retest measurement (Osirix v.5.8.2 32 bit viewing software) (Figure 1). The first and last MRI slices through the tumor were excluded to minimize partial volume effects, whereby voxels at the edge of tumors contain both abnormal and normal tissue, resulting in artefactual reduction of tumor signal intensity.

In order to develop and test the proposed error model (see below) we wished to maximize the range of ROI sizes available. Consequently, 2 additional single slice ROIs were defined within the delineated 3D tumor volume; 1) a slice representing the largest area through the lesion and 2) a slice that best represented mostly solid tumor. This reflects common practice in previous studies, for example where ADC metrics have been calculated from ROIs based on a single slice with the largest diameter, or occasionally a prescribed 2D area believed to be solid tumor[22]. We could expect a single observer definition of the largest diameter slice within a tumor to be fairly robust, however the largest diameter slice normally contains the most central necrotic and cystic tissue. Although more subjective, a slice with the most solid tumor may be a better representation of cell density. It is important that we emphasize at this point, the primary purpose for defining further 2D small volume ROIs was to increase the accuracy of fit to our error model, and explore the relationship between statistical error and ROI size. Studies comparing 2D axial ROIs and prescribed ROIs, to 3D volumes, have found whole tumor volumes to be more reproducible. Published consensus guidelines for diffusion imaging recommend 3D volumes[24].

In addition, ROIs of a fixed dimension were defined over normal appearing liver parenchyma away from obvious tumor.

Voxel ADC values were estimated from the mono-exponential fit of 3 b-value images (100, 500, 900 s/mm^2) corrected for high b-value SNR bias[25] (see Supplementary information appendix 1). A frequency distribution histogram of ADC values within each ROI was generated, and a mean ADC for the whole ROI calculated (Figure 1).

Statistical model of expected measurement error

The sample size chosen for this study, split equally between sites, is comparable to previous studies. We have chosen to use percentage change in ADC (ΔADC%), which provides a metric of repeatability for individual tumor measurements that can be directly compared within and between studies[28].

This is an ideal repeatability metric for monitoring post treatment changes for individual patients. For comparison between studies, the 95% confidence interval width can be used to define statistically significant change in ADC measurements. Due to the nature of sampling statistics, sample size in voxels (equivalent to tumor volume) would be expected to significantly influence the uncertainty in the measurement of mean ADC. Test-retest repeatability, expressed as individual tumor ΔADC% was therefore plotted against tumor volume expressed as the number of voxels in the tumor (log scale) to assess the relationship if any between tumor size and the sources of variation in repeated measures defined by our model (which contribute to measurement uncertainty). A single voxel volume, using the imaging protocol employed here is equivalent to 45 mm^3.

ΔADC%, (the percentage change in ADC between baselines, expressed as R_{12}, provides a direct measure of repeatability between scans 1 and 2) is calculated as

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

Where D_1 and D_2 are test and retest mean ADC values, respectively.

A proportion ($\varepsilon_{R_{12}}$) of the measured repeatability between D_1 and D_2 is due to predictable statistical measurement errors on D_1 and D_2, (σ_{D_1} and σ_{D_2} respectively).

The term $\varepsilon_{R_{12}}$ can be thought of as a measure of the uncertainty of the repeatability measurement and is estimated from error propagation of σ_{D_1} and σ_{D_2} using the equation below (see Supplementary information appendix 2 for derivation)

$$\varepsilon_{R_{12}}\left(\sigma_{D_1},\sigma_{D_2}\right)=\frac{400\sqrt{D_1^2\sigma_{D_2}^2+D_2^2\sigma_{D_1}^2}}{\left(D_1+D_2\right)^2}$$

The term $\varepsilon_{R_{12}}$ is dependent on the measurement accuracy of both the test and retest mean ADCs (σ_{D_1}, σ_{D_2}). Three parameters were defined within a model to describe $\varepsilon_{R_{12}}$ of the observed repeatability measurement. The simplest assumption is that $\varepsilon_{R_{12}}$ is due only to accumulation of systematic errors related to the MRI scanner. Systematic errors (ε_{sys}) contribute a fixed proportional error reflecting inability to accurately replicate equivalent image data on repeated attempts. Another possible source of measurement error reflects accuracy of the fitting routine used to estimate voxel ADC values; therefore a second parameter in our model assumes that these are fixed between D_1 and D_2. This is described as a fixed fitting error (σ_{fix}). The third parameter takes into consideration the ADC histogram distribution width for D_1 and D_2. This is a measure of the accuracy of the calculated mean ADC (D_1, D_2). The standard error is the ratio between the standard deviation of the mean ADC, and the square root of the number of voxels within the ROI. From a basic statistical level, the wider the distribution, the larger the standard error of the mean will be and conversely, the larger the sample size the smaller the standard error of the mean will be (hence the assumption earlier that ROI size is an important variable for repeatability). In addition to these factors, we would also expect ADC distribution width, and therefore mean ADC measurement accuracy, to be affected by SNR and tumor heterogeneity.

In summary, the 3-parameter model of statistical measurement errors include a fixed fitting error term (σ_{fix}), a term (β) proportional to ADC width and the systematic error (ε_{sys}), as described in the following equation

$$\varepsilon_{R_{12}}^2=\beta^2\varepsilon_{R_{12}}^2\left(\sigma_{D_1},\sigma_{D_2}\right)+\varepsilon_{R_{12}}^2\left(\sigma_{fix},\sigma_{fix}\right)+\varepsilon_{sys}^2$$

A maximum likelihood (MLE) routine was used to fit this general model to the defined 3D and 2D single slice ROIs in order to identify the parameter(s) most predictive of the repeatability measurements obtained (refer to Supplementary information appendix 3). The observed ΔADC% was standardized to $\varepsilon_{R_{12}}$ in order to produce an estimate of reproducibility for the entire group (95% confidence interval widths). In other words, the level of uncertainty in the repeated measures for each ROI was taken into consideration. The parameters (β, σ_{fix} and ε_{sys}) that produced the best fit of the data were used to generate a look-up chart for estimating the relationship between $\varepsilon_{R_{12}}$ and the ADC histogram width, for a range of ROI sizes. Datasets identified as having visible motion artifact were excluded from the MLE model fitting routine as we hypothesize that respiratory motion is an important additional variable affecting reproducibility, independently from the model. Once the best-fit parameters were obtained, all data including those with visible motion were included to compare the reference standard to the index test (data standardized to the level of uncertainty in each observed ΔADC%). A chi-squared goodness of fit method was applied to test the suitability of the error model as a fit for the observed data.

Data availability

The full dataset of mean and median ADC values calculated from all the ROIs defined for this study, are freely available within the following document: http://www.tina-vision.net/docs/memos/2014-007.pdf

Results

Twenty patients (5 per site) were scanned between May 2012 and October 2014 (16 males, 4 females; median age 63 years; range 44-77 years). 5 patient data sets (25%) were identified with visible motion in test, retest or both acquisitions. Table 2 is a summary of ADC values, lesion sizes and characteristics for each patient.
The average whole tumor mean ADC was 109 x 10^{-5} mm^2/s (range 76 – 198 x 10^{-5} mm^2/s).

The observed ΔADC% for each test-retest dataset is plotted against ROI size in Figure 2. There is a trend in the scatter to suggest repeatability improves with increasing ROI volume, but the overall 95% confidence limit width for all data is 21.1%.

Applying the 3-parameter model

The contribution of predictable statistical errors to each ΔADC% was estimated using the 3-parameter model described above. The parameters (scaling factors) in the error model were found to be:

$$\beta = 4.87, \ = 69.35 \ and \ = 2.65$$

In the majority of cases β had a larger contribution to the measurement error than σ_{fix}. In most cases ε_{sys} was minimal. When σ_{fix} was removed (i.e. a 2 parameter model), β = 5.48 and = 3.89.

The suitability of the error model (i.e. a null hypothesis that the model describes the data accurately) was tested using the Chi-squared (χ^2) method (see appendix 3). For 3D tumor ROIs, the= 11.33 with 15 degrees of freedom (Probability () ≤ 11.33 = 0.27). As there was no significant difference between our model and the observed data ($p>0.05$), the null hypothesis could be rejected. For the 2-parameter model, for 3D ROIs = 8.79, which was marginally worse than when σ_{fix} had been included. When only using (i.e. assuming a conventional form where measurement error is simply constant across samples) the model was rejected.

The relationship between the product of the 3-parameter model, , and ROI size was plotted for each data set (Figure 3). There is a clear inverse relationship between expected statistical error and ROI volume. The improved, despite motion, as tumor size increased. Above a threshold value of approximately 90 cm^3 (dashed line in Figure 3), the rate of improvement began to plateau.

When ΔADC% is standardized to its corresponding estimated statistical measurement error, i.e. factoring out the differences in the contribution of statistical measurement error on each ΔADC% (Figure 4), the 95% confidence interval width used to determine a statistically significant change in ΔADC% reduced from 21.1% to 2.7%.

The majority of data affected by gross motion become outliers, regardless of their size.

Using the 3-parameter approach, when the ΔADC% for each ROI used to fit the model, is standardized to its level of uncertainty, the distribution is 59.95 with 60 degrees of freedom (Probability () ≤ 59.95 = 0.48), i.e. there was no significant difference between the standardised distribution and our model, and the data was a good fit. When grouping all tumour ROIs together, distribution is 132 with 15 degrees of freedom (Probability () ≤ 132 = 2.4852e-7), therefore the model is rejected. This is to be expected, as data sets with motion artefact are included. When grouping only those tumour ROIs without visible motion, distribution is 46 with 45 degrees of freedom (Probability () ≤ 46 = 0.43), and the model is once more a good fit.

In Figure 5 a look up chart is presented that can be used to estimate for any ROI with a known ADC histogram width (SD) and size (voxels). This was developed using the parameters (β, σ_{fix} and ε_{sys}) that produced the best fit of data. For example, if an investigator measures ΔADC% of a tumor after treatment to be 25%, and the tumor volume is between 10 and 20 cm^3, the uncertainty in that estimation of ΔADC% will be approximately between 6 and 18%. This can be quantified more accurately by knowing the ADC distribution width for the ROI histogram. If the standard deviation of the ADC distribution is large e.g. 50 mm/s^2, then uncertainty is the measurement is around 18%. In comparison, if there is a narrower ADC distribution width for a tumor volume, e.g. 10 mm/s^2, then the investigator can have more confidence in the ΔADC % measurement of 25% after treatment (approximately 6% uncertainty in the measurement).

In summary, for a small tumor volume, with a wide ADC range of distribution, a higher threshold is required in the interpretation of ΔADC%, in order to overcome uncertainty in the measurement.

Discussion

Mean ADC is a potential MR imaging biomarker for use in assessment of early treatment response of colorectal liver metastasis[2]. In therapeutic trials early treatment typically induce ADC changes in the range of 10-30%. As discussed in the introduction, in order to reliably detect a 10% change in a single lesion requires an accuracy of ADC measurements sufficient to produce test-retest repeatability of 3-4%. In this prospective multi-site, multi-vendor study ΔADC% reproducibility for all tumor ROIs was 21.1% (95% confidence interval width). For completion, Coefficient of Variance (CoV) was calculated using absolute mean ADC values. A multi-site CoV of 5.3% was comparable to previous single site studies that have measured reproducibility in healthy liver, or liver tumors using 1.5T or 3T scanners with a variety of protocols and gating methods. The average whole tumor, mean ADC values from this study of 109 x 10^{-5} mm^2/s (range 76 – 198 x 10^{-5} mm^2/s) agree closely with those previously published for colorectal metastases .

Any estimate of ADC will be subject to uncertainty from a variety of sources. We applied a 3-parameter model, which includes terms for systematic MR system related errors; fitting errors in the ADC estimation and statistical errors arising from inaccuracies in estimating mean ADC. In data where there was no visible movement artifact the largest source of predictable measurement error resulted from differences in the standard error on the mean, estimated from ADC histograms. Consequently, statistical measurement error was much larger in smaller tumors. When the ROI volume is larger than approximately 90 cm^3, the benefit of reduced uncertainty with increased tumor volume begins to plateau.

The 95% confidence limit width for ΔADC% in raw data is 21.1% falling to 2.7% when the estimated ADC values were standardized to the estimated statistical measurement uncertainty. The remaining variability between test and retest values can be attributed to a combination of factors not included in the model (e.g. motion, tumor heterogeneity, SNR). Clearly our error model makes an assumption that the original datasets are accurately co-registered and does not account for movement artifact. When data is standardized against uncertainty in individual repeatability measurements a number of ROIs affected by motion become outliers (Figure 4) with the ΔADC% of the remaining ROIs lying mostly within 2% of zero bias. It is clear that accurate interpretation of the observed changes must account for or preferably correct for the level of uncertainty in a repeatability measurement for individual

tumors. Following this, the contributing effect of respiratory motion to poor reproducibility and false positive results, can be more accurately assessed. In the current dataset 25% of the data was affected by visible motion artifact.

The development of the error model has very significant implications in the interpretation of ADC data, which is likely to be equally true in other anatomical settings. Use of the model either directly, or by estimation of uncertainty from the lookup table (Figure 5) enables the investigator to understand the expected statistical errors in individual estimates of mean ADC based on the number of voxels in the sample, combined with the standard deviation of the ADC distribution within the ROI. The lookup table can be used directly to assess the likely significance of any change in ADC observed in a single tumor as a result of physiological, pathological or therapeutic response. For group studies, the model may be used to assess reproducibility and therefore significant change thresholds with greater confidence, by standardizing the observed data to the level of uncertainty in the measurement. The model may also be used to justify the selection of minimal tumor size in order to minimize measurement uncertainty.

Our study has a number of limitations. The multi-site nature of the design meant that each site had to follow a standardized protocol that may not represent the optimal results available from individual manufacturers system. We did not attempt to quantify inter observer reliability which is likely to be another source of variability in future study designs. We have not attempted to correct for the clear visible motion artifact in a subset of the patients, which we have shown to be a significant contribution to reduced accuracy in estimates of ADC. This reflects the lack of an effective motion correction technique, which must form a priority for subsequent methodological research in this area.

Conclusion

We have presented a model that describes statistical sources of variation, and illustrate how this can be used to determine the level of uncertainty in a repeatability measurement for ADC for an individual tumor based on the ROI size and the standard deviation of the ADC distribution. We have standardized observed data to their level of uncertainty, a method that can be used for group studies, to estimate with more

accuracy the confidence limits (95% confidence interval widths) that would determine a statistically significant change in ADC. For small tumor volumes with a wide ADC range of distribution, measurements are likely to have a high degree of uncertainty. A strategy of minimum tumor size could optimize statistical power from group studies. For individual tumor assessment, a higher threshold is required in the interpretation of ΔADC%, in order to overcome uncertainty in the measurement. We provide a lookup chart to allow investigators to estimate uncertainty due to statistical error, for any given tumor volume and distribution. Finally, we have also demonstrated that movement artifact is a major remaining source of error suggesting that our technique should be combined with appropriate motion correction strategies, particularly for small tumors [37].

References

1 Le Bihan, D. & Johansen-Berg, H. Diffusion MRI at 25: exploring brain tissue structure and function. *NeuroImage* **61**, 324-341, doi:10.1016/j.neuroimage.2011.11.006 (2012).

2 Sinkus, R., Van Beers, B. E., Vilgrain, V., DeSouza, N. & Waterton, J. C. Apparent diffusion coefficient from magnetic resonance imaging as a biomarker in oncology drug development. *European journal of cancer* **48**, 425-431, doi:10.1016/j.ejca.2011.11.034 (2012).

3 Huang, W. Y. *et al.* Diffusion-Weighted Imaging for Predicting and Monitoring Primary Central Nervous System Lymphoma Treatment Response. *AJNR. American journal of neuroradiology*, doi:10.3174/ajnr.A4867 (2016).

4 O'Connor, J. P. *et al.* Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clinical cancer research : an official journal of the American Association for Cancer Research* **21**, 249-257, doi:10.1158/1078-0432.CCR-14-0990 (2015).

5 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

6 Tourell, M. C. *et al.* The distribution of the apparent diffusion coefficient as an indicator of the response to chemotherapeutics in ovarian tumour xenografts. *Sci Rep* **7**, 42905, doi:10.1038/srep42905 (2017).

7 Cui, Y., Zhang, X. P., Sun, Y. S., Tang, L. & Shen, L. Apparent diffusion coefficient: potential imaging biomarker for prediction and early detection of response to chemotherapy in hepatic metastases. *Radiology* **248**, 894-900, doi:10.1148/radiol.2483071407 (2008).

8 Koh, D. M. *et al.* Predicting response of colorectal hepatic metastasis: value of pretreatment apparent diffusion coefficients. *AJR. American journal of roentgenology* **188**, 1001-1008, doi:10.2214/AJR.06.0601 (2007).

9 Winfield, J. M. *et al.* A framework for optimization of diffusion-weighted MRI protocols for large field-of-view abdominal-pelvic imaging in multicenter studies. *Medical physics* **43**, 95, doi:10.1118/1.4937789 (2016).

10 Lambregts, D. M. *et al.* Tumour ADC measurements in rectal cancer: effect of ROI methods on ADC values and interobserver variability. *European radiology* **21**, 2567-2574, doi:10.1007/s00330-011-2220-5 (2011).

11 Asselin, M. C., O'Connor, J. P., Boellaard, R., Thacker, N. A. & Jackson, A. Quantifying heterogeneity in human tumours using MRI and PET. *European journal of cancer* **48**, 447-455, doi:10.1016/j.ejca.2011.12.025 (2012).

12 Schmid-Tannwald, C. *et al.* Diffusion-weighted MR imaging of focal liver lesions in the left and right lobes: is there a difference in ADC values? *Academic radiology* **20**, 440-445, doi:10.1016/j.acra.2012.10.012 (2013).

13 Schmidt, H., Gatidis, S., Schwenzer, N. F. & Martirosian, P. Impact of measurement parameters on apparent diffusion coefficient quantification in diffusion-weighted-magnetic resonance imaging. *Investigative radiology* **50**, 46-56, doi:10.1097/RLI.0000000000000095 (2015).

14 Kwee, T. C., Takahara, T., Koh, D. M., Nievelstein, R. A. & Luijten, P. R. Comparison and reproducibility of ADC measurements in breathhold, respiratory triggered, and free-breathing diffusion-weighted MR imaging of the liver. *Journal of magnetic resonance imaging : JMRI* **28**, 1141-1148, doi:10.1002/jmri.21569 (2008).

15 Winfield, J. M. *et al.* Modelling DW-MRI data from primary and metastatic ovarian tumours. *European radiology* **25**, 2033-2040, doi:10.1007/s00330-014-3573-3 (2015).

16 Malyarenko, D. *et al.* Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. *Journal of magnetic resonance imaging : JMRI* **37**, 1238-1246, doi:10.1002/jmri.23825 (2013).

17 Kandpal, H., Sharma, R., Madhusudhan, K. S. & Kapoor, K. S. Respiratory-triggered versus breath-hold diffusion-weighted MRI of liver lesions: comparison of image quality and apparent diffusion coefficient values. *AJR. American journal of roentgenology* **192**, 915-922, doi:10.2214/AJR.08.1260 (2009).

18 Nasu, K., Kuroki, Y., Sekiguchi, R. & Nawano, S. The effect of simultaneous use of respiratory triggering in diffusion-weighted imaging of the liver. *Magnetic resonance in medical sciences : MRMS : an official journal of Japan Society of Magnetic Resonance in Medicine* **5**, 129-136 (2006).

19 Taouli, B. *et al.* Diffusion-weighted imaging of the liver: comparison of navigator triggered and breathhold acquisitions. *Journal of magnetic resonance imaging : JMRI* **30**, 561-568, doi:10.1002/jmri.21876 (2009).

20 Surov, A. *et al.* Diffusion-Weighted Imaging in Meningioma: Prediction of Tumor Grade and Association with Histopathological Parameters. *Translational oncology* **8**, 517-523, doi:10.1016/j.tranon.2015.11.012 (2015).
21 Xu, X. Q. *et al.* Diffusion Weighted Imaging for Differentiating Benign from Malignant Orbital Tumors: Diagnostic Performance of the Apparent Diffusion Coefficient Based on Region of Interest Selection Method. *Korean journal of radiology* **17**, 650-656, doi:10.3348/kjr.2016.17.5.650 (2016).
22 Kono, K. *et al.* The role of diffusion-weighted imaging in patients with brain tumors. *AJNR. American journal of neuroradiology* **22**, 1081-1088 (2001).
23 Bonekamp, D. *et al.* Interobserver agreement of semi-automated and manual measurements of functional MRI metrics of treatment response in hepatocellular carcinoma. *European journal of radiology* **83**, 487-496, doi:10.1016/j.ejrad.2013.11.016 (2014).
24 Padhani, A. R. *et al.* Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia* **11**, 102-125 (2009).
25 Gudbjartsson, H. & Patz, S. The Rician distribution of noisy MRI data. *Magnetic resonance in medicine* **34**, 910-914 (1995).
26 Deckers, F. *et al.* Apparent diffusion coefficient measurements as very early predictive markers of response to chemotherapy in hepatic metastasis: a preliminary investigation of reproducibility and diagnostic value. *Journal of magnetic resonance imaging : JMRI* **40**, 448-456, doi:10.1002/jmri.24359 (2014).
27 Heijmen, L. *et al.* Diffusion-weighted MR imaging in liver metastases of colorectal cancer: reproducibility and biological validation. *European radiology* **23**, 748-756, doi:10.1007/s00330-012-2654-4 (2013).
28 Hoang, J. K. *et al.* Diffusion-weighted imaging for head and neck squamous cell carcinoma: quantifying repeatability to understand early treatment-induced change. *AJR. American journal of roentgenology* **203**, 1104-1108, doi:10.2214/AJR.14.12838 (2014).
29 Bilgili, M. Y. Reproductibility of apparent diffusion coefficients measurements in diffusion-weighted MRI of the abdomen with different b values. *European journal of radiology* **81**, 2066-2068, doi:10.1016/j.ejrad.2011.06.045 (2012).
30 Braithwaite, A. C., Dale, B. M., Boll, D. T. & Merkle, E. M. Short- and midterm reproducibility of apparent diffusion coefficient measurements at 3.0-T diffusion-weighted imaging of the abdomen. *Radiology* **250**, 459-465, doi:10.1148/radiol.2502080849 (2009).
31 Corona-Villalobos, C. P. *et al.* Agreement and reproducibility of apparent diffusion coefficient measurements of dual-b-value and multi-b-value diffusion-weighted magnetic resonance imaging at 1.5 Tesla in phantom and in soft tissues of the abdomen. *Journal of computer assisted tomography* **37**, 46-51, doi:10.1097/RCT.0b013e3182720e07 (2013).
32 Larsen, N. E., Haack, S., Larsen, L. P. & Pedersen, E. M. Quantitative liver ADC measurements using diffusion-weighted MRI at 3 Tesla: evaluation of reproducibility and perfusion dependence using different techniques for respiratory compensation. *Magma* **26**, 431-442, doi:10.1007/s10334-013-0375-6 (2013).
33 Rosenkrantz, A. B., Oei, M., Babb, J. S., Niver, B. E. & Taouli, B. Diffusion-weighted imaging of the abdomen at 3.0 Tesla: image quality and apparent diffusion coefficient reproducibility compared with 1.5 Tesla. *Journal of magnetic resonance imaging : JMRI* **33**, 128-135, doi:10.1002/jmri.22395 (2011).
34 Koh, D. M. *et al.* Reproducibility and changes in the apparent diffusion coefficients of solid tumours treated with combretastatin A4 phosphate and bevacizumab in a two-centre phase I clinical trial. *European radiology* **19**, 2728-2738, doi:10.1007/s00330-009-1469-4 (2009).
35 Kim, S. Y. *et al.* Malignant hepatic tumors: short-term reproducibility of apparent diffusion coefficients with breath-hold and respiratory-triggered diffusion-weighted MR imaging. *Radiology* **255**, 815-823, doi:10.1148/radiol.10091706 (2010).
36 Kim, S. Y. *et al.* Reproducibility of measurement of apparent diffusion coefficients of malignant hepatic tumors: effect of DWI techniques and calculation methods. *Journal of magnetic resonance imaging : JMRI* **36**, 1131-1138, doi:10.1002/jmri.23744 (2012).
37 Ragheb, H. *et al.* The Accuracy of ADC Measurements in Liver Is Improved by a Tailored and Computationally Efficient Local-Rigid Registration Algorithm. *PloS one* **10**, e0132554, doi:10.1371/journal.pone.0132554 (2015).

Acknowledgment

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking (www.imi.europa.eu) under grant agreement number 115151, resources of which are composed of financial contribution from the European Unions Seventh Framework Programme (FP7/2007–2013) and EFPIA companies in-kind contribution. There was, however, no financial or in-kind contribution from EFPIA companies to the research specifically described in this paper. The funders of the research leading to these results had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contribution statement

We are submitting original research with data that has not yet been published in any journal. We declare that all named authors have read the manuscript and have agreed to submit in its present form. All named authors have made a sufficient contribution to the work. Dr RP has been involved with the recruitment, data acquisition, analysis and writing of the first draft. Dr HR has been involved with the data analysis and development of the statistical error model, as well as being heavily involved in the writing process. Dr NT has been heavily involved in the design and supervision of the statistical error model and data analysis.
The remaining authors have been involved in protocol development, data recruitment, second and subsequent draft edits and overall supervision and invaluable guidance

Additional Information

Competing financial interest statement

The authors declare no competing financial interest

Figure legends

Figure 1. Tumor selection and image analysis. A single lesion is chosen based on size and location from b-100 DWI images (right image) and a ROI is manually defined (green) for each test-retest data set (middle image). A parametric map of ADC values is calculated for each pixel within the ROI (left image). For 3D volumes, the voxel ADC values within each slice ROI is combined and represented as a histogram (far left).

Figure 2. Tumor reproducibility of ΔADC% as measured by the 95% confidence interval width for all multisite data. ΔADC% is plotted against ROI size (log number of voxels) for 3D and 2D tumor regions (3D circles, 2D triangles). Data affected by motion is highlighted (solid black). The fixed-sized normal parenchyma ROIs are included in the calculation of the 95% CI width of 21.1%.

Figure 3. The relationship between statistical measurement error and tumor ROI size. Measurement error improves with increasing ROI size, up to a threshold of around 2000 voxels equivalent to 90 cm^3.

Figure 4. The improvement in estimating repeatability measurements after accounting for the contribution of statistical measurement error. ΔADC% is plotted against ROI size (log scale of number of voxels) for 3D and 2D tumor regions (3D circles, 2D triangles). Data affected by motion is highlighted (solid black). When the contribution of statistical measurement error is factored out (compared to Figure 2), the 95% confidence interval width improves from 21.1% to 2.7%. The majority of data affected by motion become outliers, regardless of their size.

Figure 5. Look up chart for estimating statistical error. Using the parameters that produced the best fit of data, a look-up chart has been created, that can be utilized to estimate statistical measurement error for any ROI with a known ADC histogram width (SD) and size (voxels).

Display Items

Figures

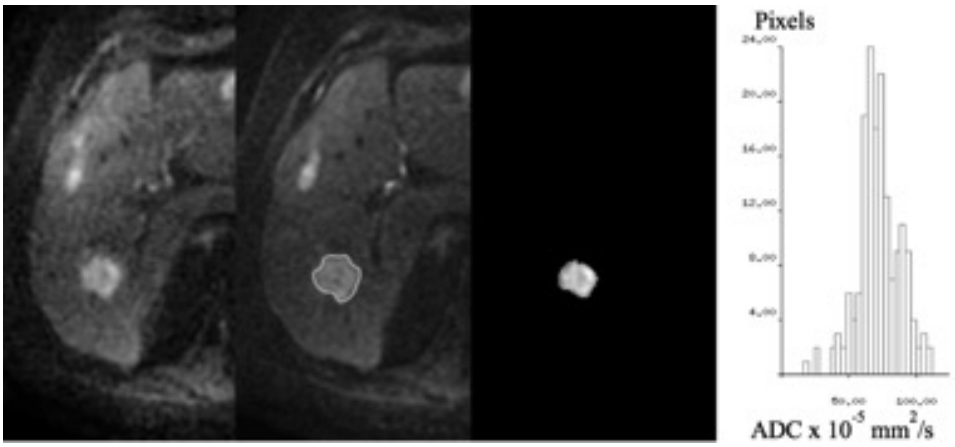

Figure 1. Tumor selection and image analysis. A single lesion is chosen based on size and location from b-100 DWI images (right image) and a ROI is manually defined for each test-retest data set (middle image). A parametric map of ADC values is calculated for each pixel within the ROI (left image). For 3D volumes, the voxel ADC values within each slice ROI is combined and represented as a histogram (far left).

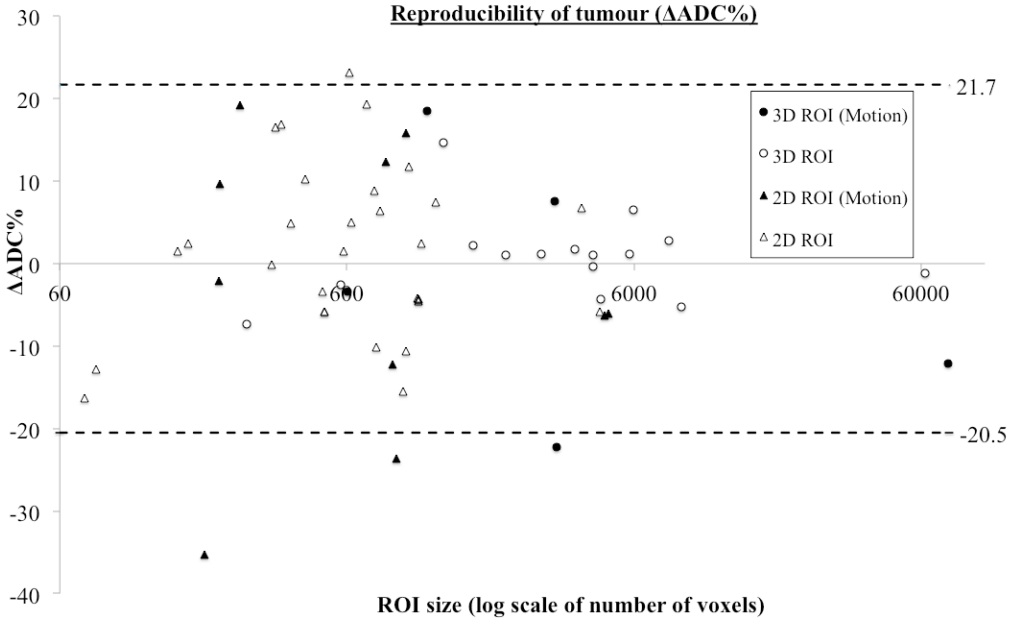

Figure 2. Tumor reproducibility of ΔADC% as measured by the 95% confidence interval width for all multisite data. ΔADC% is plotted against ROI size (log number of voxels) for 3D and 2D tumor regions (3D circles, 2D triangles). Data affected by motion is highlighted (solid black). The fixed-sized normal parenchyma ROIs are included in the calculation of the 95% CI width of 21.1%.

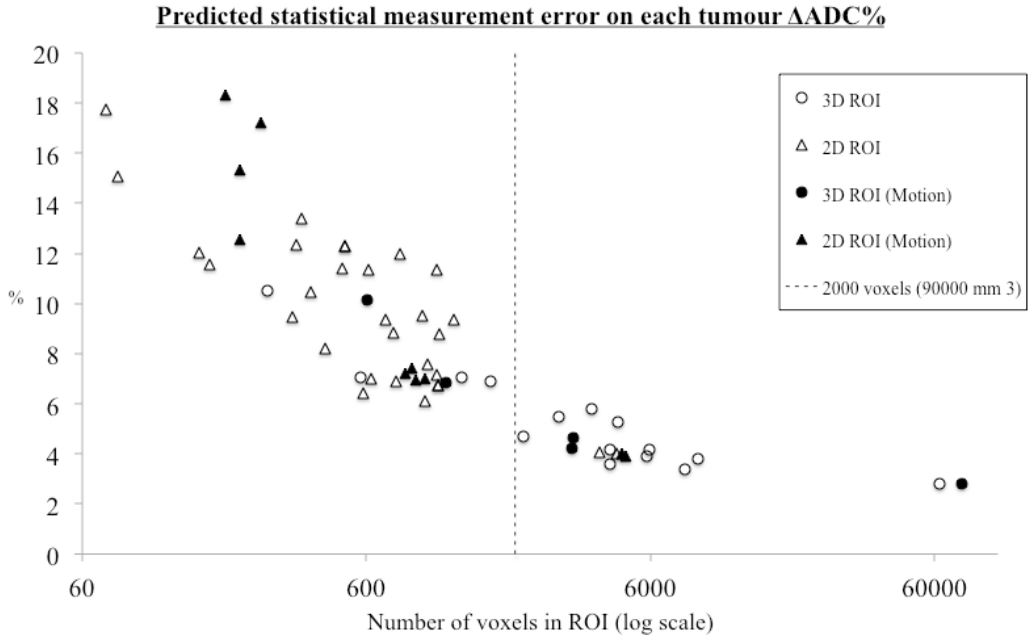

Figure 3. The relationship between statistical measurement error and tumor ROI size. Measurement error improves with increasing ROI size, up to a threshold of around 2000 voxels equivalent to 90 cm^3.

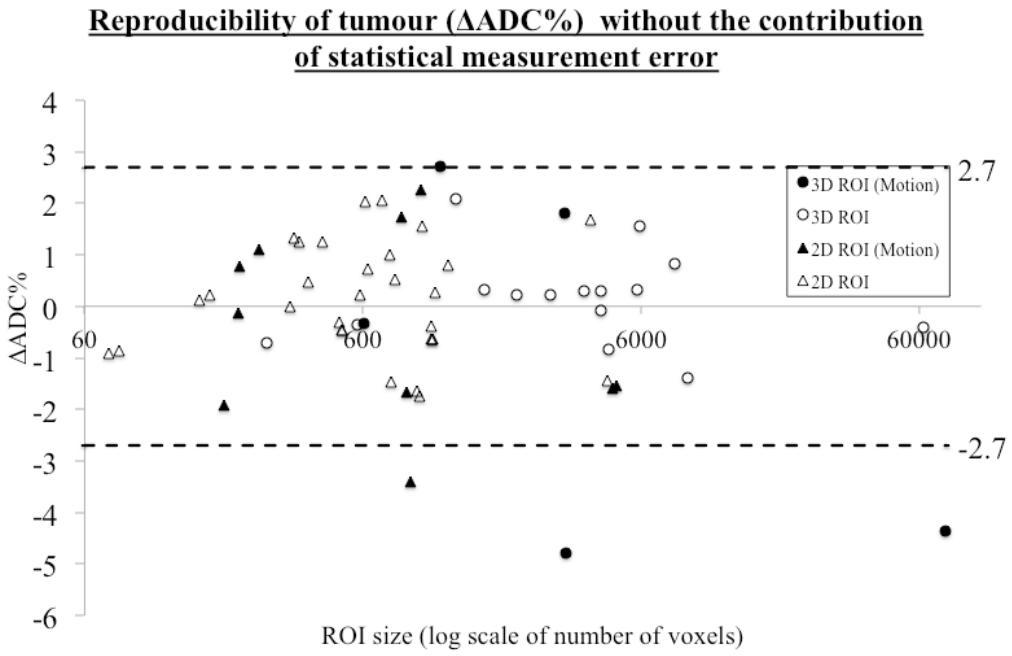

Figure 4. The improvement in estimating repeatability measurements after accounting for the contribution of statistical measurement error. ΔADC% is plotted against ROI size (log scale of number of voxels) for 3D and 2D tumor regions (3D circles, 2D triangles). Data affected by motion is highlighted (solid black). When the contribution of statistical measurement error is factored out (compared to Figure 2), the 95% confidence interval width improves from 21.1% to 2.7%. The majority of data affected by motion become outliers, regardless of their size.

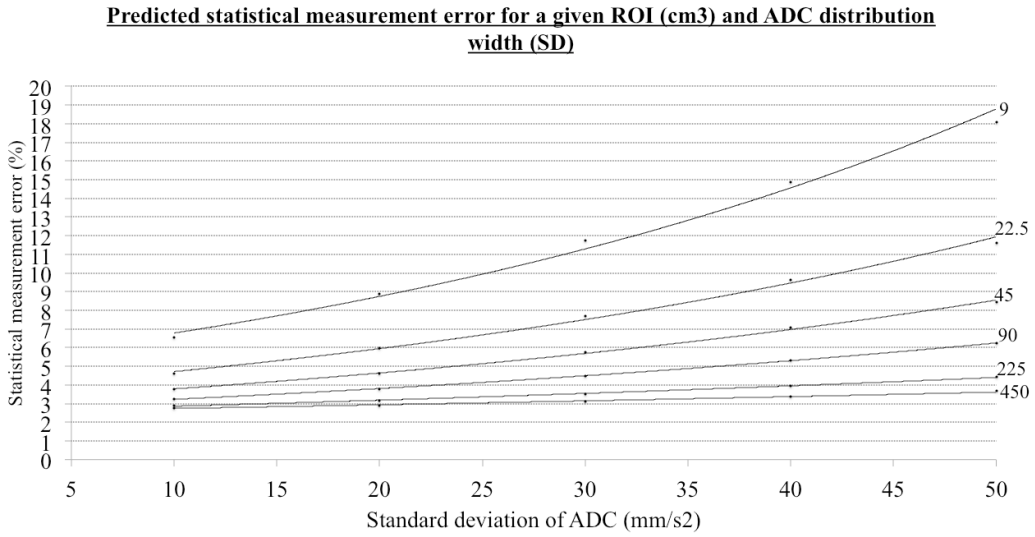

Figure 5. Look up chart for estimating statistical error. Using the parameters that produced the best fit of data, a look-up chart has been created, that can be utilized to estimate statistical measurement error for any ROI with a known ADC histogram width (SD) and size (voxels).

Tables

MRI (1.5 T)	Body coil	Parallel imaging	B-values (s/mm^2)	TR/TE (ms)
Siemens Magnetom Avanto	6 channel	GRAPPA 2	100, 500, 900	8000/76
General Electric (GE) Signa HDxt	8 channel	ASSET	100, 500, 900	8500/74
Philips Achieva	8 channel	SENSE	0, 100, 500, 900	8000/88

Table 1. List of MR systems and receiver coils used, with variable DWI acquisition parameters.

Patient	Size	ADC* (10^{-5} mm^2/s)	ΔADC%	Lesion	Image

1	1141	76	18.56		Motion
2	3214	102	-22.37	Sub-phrenic	Motion
3	2845	97	1.17	5% cystic	
4	1297	77	14.69		
5	603	98	-3.39	Sub-phrenic	Motion
6	573	87	-2.52		
7	148	123	2.25		
8	3178	102	7.60	Sub-phrenic	Motion
9	4589	95	-4.35		
10	3731	103	1.69		
11	5957	140	6.48		
12	74572	102	-12.13		Motion
13	6780	93	1.22		
14	270	93	-7.36		
15	61130	118	-1.11		
16	8788	127	-5.30	10% cystic	
17	4315	98	1.06		
18	2140	129	1.04		
19	7914	198	2.84	95% cystic	
20	4304	110	-0.31		

Table 2. The ADC values, lesion size and image characteristics for each patient. The number of voxels within 3D whole tumor ROIs is given in the 'size' column and each voxel is 45 mm^3. ADC* is the average of both test-retest mean ADC values (x 10^{-5} mm^2/s). ΔADC% is the percentage change in ADC between test-retest. The data sets visually affected by "Motion" artifact are indicated in the Image column.

Supplementary Information

Please refer to separate PDF titled

"Supplementary information_scientific reports"