Extracranial Soft-Tissue Tumors: Repeatability of Apparent Diffusion Coefficient Estimates from Diffusion-weighted MR Imaging[1]

Jessica M. Winfield, PhD
Nina Tunariu, MD
Mihaela Rata, PhD
Keiko Miyazaki, PhD
Neil P. Jerome, PhD
Michael Germuska, PhD[2]
Matthew D. Blackledge, PhD
David J. Collins, BA
Johann S. de Bono, MD, PhD
Timothy A. Yap, MD, PhD
Nandita M. deSouza, MD
Simon J. Doran, PhD
Dow-Mu Koh, MD
Martin O. Leach, PhD
Christina Messiou, MD
Matthew R. Orton, PhD

Purpose: To assess the repeatability of apparent diffusion coefficient (ADC) estimates in extracranial soft-tissue diffusion-weighted magnetic resonance imaging across a wide range of imaging protocols and patient populations.

Materials and Methods: Nine prospective patient studies and one prospective volunteer study, performed between 2006 and 2016 with research ethics committee approval and written informed consent from each subject, were included in this single-institution study. A total of 141 tumors and healthy organs were imaged twice (interval between repeated examinations, 45 minutes to 10 days, depending the on study) to assess the repeatability of median and mean ADC estimates. The Levene test was used to determine whether ADC repeatability differed between studies. The Pearson linear correlation coefficient was used to assess correlation between coefficient of variation (CoV) and the year the study started, study size, and volumes of tumors and healthy organs. The repeatability of ADC estimates from small, medium, and large tumors and healthy organs was assessed irrespective of study, and the Levene test was used to determine whether ADC repeatability differed between these groups.

Results: CoV aggregated across all studies was 4.1% (range for each study, 1.7%–6.5%). No correlation was observed between CoV and the year the study started or study size. CoV was weakly correlated with volume ($r = -0.5$, $P = .1$). Repeatability was significantly different between small, medium, and large tumors ($P < .05$), with the lowest CoV (2.6%) for large tumors. There was a significant difference in repeatability between studies—a difference that did not persist after the study with the largest tumors was excluded.

Conclusion: ADC is a robust imaging metric with excellent repeatability in extracranial soft tissues across a wide range of tumor sites, sizes, patient populations, and imaging protocol variations.

Online supplemental material is available for this article.

[1] From the Cancer Research UK Cancer Imaging Centre, Division of Radiotherapy and Imaging (J.M.W., N.T., M.R., K.M., N.P.J., M.G., M.D.B., D.J.C., N.M.d.S., S.J.D., D.M.K., M.O.L., C.M., M.R.O.) and Division of Clinical Studies (J.S.d.B., T.A.Y.), the Institute of Cancer Research and Royal Marsden Hospital, London, England; MRI Unit (J.M.W., N.T., M.R., K.M., N.P.J., M.G., M.D.B., D.J.C., N.M.d.S., S.J.D., D.M.K., M.O.L., C.M., M.R.O.) and Drug Development Unit (J.S.d.B., T.A.Y.), the Royal Marsden NHS Foundation Trust, Downs Road, Sutton, Surrey SM2 5PT, England. Received August 19, 2016; revision requested October 12 and received December 1; accepted December 14; final version accepted December 20. **Address correspondence to** M.O.L. (e-mail: *martin.leach@icr.ac.uk*).

Supported by Cancer Research UK and Engineering and Physical Sciences Research Council support to the Cancer Imaging Centre at the Institute of Cancer Research and Royal Marsden Hospital in association with the Medical Research Council and Department of Health C1060/A10334, C1060/A16464 and National Health Service funding to the National Institute for Health Research (NIHR) Biomedical Research Centre and the Clinical Research Facility in Imaging. M.O.L. is an Emeritus NIHR Senior Investigator.

Current address:
[2] Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, Cardiff, Wales.

Body diffusion-weighted (DW) magnetic resonance (MR) imaging is well established as a qualitative and quantitative technique in oncology (1). The most simple quantitative metric derived from DW MR imaging is the apparent diffusion coefficient (ADC), which is estimated by fitting a monoexponential curve to the measured signal at two or more diffusion weightings (*b* values). Baseline ADC estimates or posttreatment changes in ADC have been shown to be indicative of response to chemotherapy and/or chemotherapy and radiation therapy in many tumor types, including rectal adenocarcinoma (2), hepatic metastases of colorectal (3) and gastric (4) cancers, cervical cancer (5), breast cancer (6), head-and-neck squamous cell carcinoma (7), ovarian cancer (8), and non–small cell lung cancer (9).

As for all quantitative metrics, the repeatability of ADC estimates determines the ability of the technique to reveal treatment-induced changes, thereby influencing the number of patients required for clinical trials and determining the size of posttreatment changes that can be detected in individual patients. Repeatability is usefully defined as "closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement" (10), where, in imaging studies, repeatability conditions include use of the same scanner or imaging unit, imaging protocol, observers, and repetition after a short interval (typically 1 hour to 7 days). In DW MR imaging–based studies that report ADC estimates, the "measurand" is usually the mean or median of ADC estimates from voxels in a tumor. On the other hand, reproducibility may be defined as "closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement" (10)—for example, by using a different MR imaging unit. The inter–imaging unit reproducibility of ADC estimates is particularly important in multicenter studies, where it has been shown that good-quality DW images with reproducible ADC estimates across platforms can be obtained following careful optimization of imaging protocols (11).

Exploratory DW MR imaging studies in clinical trials often incorporate ADC repeatability estimates, usually by obtaining two baseline examinations with the second examination during the same visit (the so-called coffee-break repeatability study) or at a second visit 1 or more days later. The requirement for two baseline examinations increases the burden on patients, which may reduce recruitment or retention rates, and requires additional imaging unit time and resources, which may be difficult to accommodate in busy radiology departments. It would be advantageous to estimate ADC repeatability from previous studies, but this would be feasible only if repeatability was broadly the same across studies, despite variations in imaging protocol, tumor type, or patient cohort; large differences in repeatability would argue strongly for study-specific repeatability estimates. The variety of repeatability metrics reported in the literature hinders comparison between studies, and a framework for assessment of the technical performance of quantitative imaging biomarkers has been proposed by the Radiological Society of North America Quantitative Imaging Biomarkers Alliance (QIBA) (12,13). The QIBA framework recommends reporting repeatability by using the within-subject standard deviation, limits of agreement (LoAs), repeatability coefficient (RC), intraclass correlation coefficient (ICC), and within-subject coefficient of variance; the QIBA also emphasizes the importance of reporting measurement conditions. A detailed investigation of ADC repeatability across a wide range of studies using the QIBA framework is therefore desirable.

The aim of this study was to assess ADC repeatability using the framework proposed by the QIBA in extracranial soft-tissue DW MR imaging studies to investigate whether ADC repeatability differs between studies performed by using different imaging protocols and patient populations over a period of 10 years at a single institution.

Advances in Knowledge

- Repeated apparent diffusion coefficient (ADC) estimates can be obtained from extracranial soft-tissue diffusion-weighted (DW) MR imaging with coefficients of variation (CoVs) of between 2% and 7%.
- ADC repeatability does not differ markedly (CoV, 2%–7%) between DW MR imaging studies across a wide range of patient cohorts and imaging protocol variations.
- Better ADC repeatability is observed in large tumors than in smaller tumors.

Implication for Patient Care

- DW MR imaging can be used to estimate ADC with good repeatability in extracranial soft tissues, allowing a posttreatment increase of 12% or more in ADC to be distinguished.

https://doi.org/10.1148/radiol.2017161965

Content codes: GI MR

Radiology 2017; 284:88–99

Abbreviations:
ADC = apparent diffusion coefficient
CI = confidence interval
CoV = coefficient of variation
DW = diffusion weighted
ICC = intraclass correlation coefficient
LoA = limit of agreement
QIBA = Quantitative Imaging Biomarkers Alliance
RC = repeatability coefficient
ROI = region of interest
VOI = volume of interest

Author contributions:
Guarantors of integrity of entire study, M.O.L., M.R.O.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.M.W., N.T., D.J.C., M.R.O.; clinical studies, J.M.W., N.T., M.R., K.M., N.P.J., M.G., M.D.B., D.J.C., J.S.d.B., T.A.Y., N.M.d.S., D.M.K., C.M., M.R.O.; experimental studies, J.M.W., M.R., N.P.J., M.G., T.A.Y., M.R.O.; statistical analysis, J.M.W., M.R., K.M., N.P.J., D.J.C., M.R.O.; and manuscript editing, J.M.W., N.T., M.R., K.M., N.P.J., M.D.B., D.J.C., J.S.d.B., T.A.Y., N.M.d.S., S.J.D., D.M.K., M.O.L., C.M., M.R.O.

Conflicts of interest are listed at the end of this article.

Materials and Methods

Study Population

Nine patient studies and one healthy volunteer study were included in this analysis. All studies were approved by relevant National Research Ethics Committees. All patients and volunteers gave their written consent to participate in the studies. Only repeatability data from double-baseline examinations are reported here; posttreatment changes were outside the scope of this study but have been reported in the literature for some studies (14–18).

Tables 1 and 2 describe the subjects and the DW MR imaging protocols for each study (labeled A through K); further information is available in the references given. All studies were performed at 1.5 T by using MAGNETOM Avanto or Aera MR imaging units (Siemens) (Table 2). In studies where the imaging study or ADC repeatability study formed a subset of the total cohort (studies C and G), only data in patients that contributed to the ADC repeatability results are reported. In multicenter studies, only data from our center are reported (studies D, E, and K). In studies including intracranial and extracranial tumors, only extracranial data are reported (studies A and F). One result (the coefficient of variation [CoV] of ADC_{median} in study K) has been reported previously (11), but other results from study K have not been reported previously. No other results presented here have been reported previously, as publications from the original studies included data from intracranial tumors (14,15,19) or data from other centers (17), which are excluded from this analysis.

Image and Data Analysis

A total of 141 tumors and healthy organs were included in this analysis. All DW MR imaging data were fitted by using in-house software (Adept, the Institute of Cancer Research, London; or Matlab, Mathworks, Natick, Mass). ROIs were drawn as described in Table 1. Software, methods, and observers were fixed within each study; differences between studies reflect changes in technology and personnel (Table 1).

For each tumor or healthy organ, all fitted pixels in the ROIs were combined to create a VOI. Median and mean ADC (ADC_{median} and ADC_{mean}) were estimated for each VOI. Bland-Altman plots of untransformed data showed a tendency for differences between pairs of baseline measurements to scale with their ADC (see Figure E1 [online]), in which case it is recommended (13,20) that repeatability (and changes due to treatment) be quantified by using a proportional (ie, ratio-based) measure so that the same measure applies across the range of ADCs encountered. This can most easily be achieved by using the natural logarithm of the data (12,13,20–22), and this was done for all statistical analyses in this study. A paired t test was used to assess whether there was a significant difference between the first and second baseline measurements in each study. $P < .05$ was considered to indicate a significant difference.

Repeatability was assessed by using the methods recommended by QIBA (13). The within-subject standard deviation (s_W) of the log-transformed ADC estimates was estimated according to Equation (1), where d_i is the difference between two baseline estimates of $\log(ADC_{median})$ or $\log(ADC_{mean})$ for the ith VOI, and N is the number of VOIs:

$$s_W = \sqrt{\frac{1}{2N}\sum_{i=1}^{N} d_i^2}. \quad (1)$$

The within-subject CoV (23), 95% LoA, and RC, which depend only on s_W, were estimated according to Equations (2), (3), and (4), respectively.

$$\text{CoV} = 100\% \times \sqrt{\exp\left(s_W^2\right) - 1}, \quad (2)$$

$$\text{LoA} = 100\% \times \left[\exp\left(\pm 1.96\sqrt{2}s_W\right) - 1\right], \quad (3)$$

and

$$\text{RC} = 1.96\sqrt{2}s_W. \quad (4)$$

The ICC was estimated according to Equation (5), where s_B is the between-subject standard deviation:

$$\text{ICC} = \frac{s_B^2}{s_B^2 + s_W^2}. \quad (5)$$

s_B was estimated as $s_B = \sqrt{(\text{BMS} - \text{WMS})/K}$, where $\text{BMS} = K\sum_{i=1}^{N}(\bar{Y}_i - \bar{Y})^2/N$ is the between-subject mean squares, $\text{WMS} = \sum_{i=1}^{N}\sum_{k=1}^{K}(Y_{ik} - \bar{Y}_i)^2/N(K-1)$ is the within-subject mean squares, K is the number of replications ($K = 2$ for all studies in this analysis), Y_{ik} is the observed value of $\log(ADC_{median})$ or $\log(ADC_{mean})$ for the ith VOI at the kth replication, $\bar{Y}_i$ is the average over replications for the ith VOI, and $\bar{Y}$ is the grand mean of $\log(ADC_{median})$ or $\log(ADC_{mean})$ over all observations (24).

The 95% confidence intervals (CIs) for s_W were estimated as $\left(\sqrt{\text{WMS}\bullet N/\text{Inv-}\chi^2_N(0.975)}, \sqrt{\text{WMS}\bullet N/\text{Inv-}\chi^2_N(0.025)}\right)$, where $\text{Inv-}\chi^2_N(p)$ is the pth centile of the χ^2 distribution with N degrees of freedom (24).

95% CIs for ICCs were estimated as $\left(\frac{F_L - 1}{F_L + 1}, \frac{F_U - 1}{F_U + 1}\right)$, where $F_U = F_0 \bullet \text{Inv-}F_{N,N-1}(0.975)$ and $F_L = F_0/\text{Inv-}F_{N-1,N}(0.975)$, with $F_0 = \text{BMS}/\text{WMS}$, and $\text{Inv-}F_{d_1,d_2}(p)$ is the pth centile of the F distribution with d_1 and d_2 degrees of freedom (25).

In addition to each study being analyzed individually, VOIs were grouped into small, medium, and large, regardless of study (ie, smallest third, middle third, and largest third of VOIs), and repeatability was assessed for the three groups (47 VOIs per group). Finally, VOIs were aggregated from all studies, and repeatability was assessed for 141 VOIs together.

The Levene test for homoscedasticity (Matlab, LeveneAbsolute, vartestn, 2016a; Mathworks) was used to assess whether repeatability differed between studies (24). Baseline differences were calculated for each VOI for $\log(ADC_{mean})$

Table 1

Subjects and Imaging Procedures for Studies A through K

Parameter	Study A (14,15)	Study B	Study C (16)	Study D	Study E (17)	Study F (19)	Study G (18)	Study H	Study J	Studies K1, K2, and K3 (11)
Patient or volunteer cohort?	Patients (phase 1 trial population; adults)	Patients (adults)	Patients (phase 1 trial population; adults)	Patients (adults)	Patients (phase 1 trial population; adults)	Patients (pediatric)	Patients (phase 1 trial population; adults)	Patients (phase 1 trial population; adults)	Patients (phase 1 trial population; adults)	Healthy volunteers (adults, women)
Tumor site or healthy organ examined	Mixed (17 abdominal, nine pelvic lesions)	Retroperitoneal soft-tissue masses	Mixed (eight liver lesions, one splenic lesion, one renal lesion, one peritoneal lesion, one abdominal wall lesion, one pelvic lymph node)	Renal cell carcinoma	Mixed (five liver lesions, four pelvic lesions)	Mixed (extracranial solid tumors)	Mixed (four liver lesions, two pelvic lesions, two pelvic lymph nodes)	Mixed (five liver lesions, one lung lesion, one abdominal lymph node)	Liver lesions	Healthy organs (K1: kidneys; K2: liver; K3: spleen)
No. of patients or healthy volunteers	26	23	13	11	9	8	8	7	6	10
Start and end dates of study	2008–2010	2013–2016	2009–2010	2010–2016	2006–2007	2010–2014	2008–2009	2014–2016	2011–2012	2012
Interval between two examinations	7 Days	Approximately 45 minutes ("coffee-break" repeatability)	4–7 Days	24 Hours	2–10 Days	24 Hours	4–10 Days	2–3 Days	5 Days	1–7 Days

Table 1 *(continues)*

Table 1 (continued)

Subjects and Imaging Procedures for Studies A through K

Parameter	Study A (14,15)	Study B	Study C (16)	Study D	Study E (17)	Study F (19)	Study G (18)	Study H	Study J	Studies K1, K2, and K3 (11)
Method used to define VOIs*	ROIs drawn by a consultant radiologist (C.M.) with 7 years of experience, including 3 years of experience in extracranial DW MR imaging. ROIs drawn around whole area of tumor on $b = 500$ sec/mm^2 DW images on all sections on which tumor appeared (up to a maximum of 20 sections in imaging volume).	ROIs drawn by a consultant radiologist (C.M.) with 12 years of experience, including 8 years of experience in extracranial DW MR imaging. ROIs drawn around whole area of tumor on T2-weighted images on all sections on which the tumor appeared; ROIs transferred to ADC maps (excluding the most cranial and caudal sections if partial-volume effects were visible). Two stations were acquired if necessary to cover very large tumors.	ROIs drawn by a radiologist (N.T.) with 5 years of experience, including 4 years of experience in extracranial DW MR imaging. ROIs drawn around whole tumor on highest-*b*-value images on three to six sections near the center of the imaging volume.	ROIs drawn by a consultant radiologist (D.M.K.) with > 10 years of experience, including > 10 years of experience in extracranial DW MR imaging. ROIs drawn around tumor on five central sections on high-*b*-value images with reference to other imaging.	ROIs drawn by a consultant radiologist (D.M.K.) with > 10 years of experience, including > 5 years of experience in extracranial DW MR imaging. ROIs drawn around whole area of tumor on $b = 750$ sec/mm^2 DW images on all sections on which tumor appeared (excluding the most cranial and caudal sections if partial-volume effects were visible).	ROIs drawn by a consultant radiologist (D.M.K.) with > 10 years of experience, including > 10 years of experience in extracranial DW MR imaging. ROIs drawn around tumor on three sections near the center of the imaging volume; matching sections selected for second baseline examination.	ROIs drawn by a radiologist (N.T.) with 5 years of experience, including 4 years of experience in extracranial DW MR imaging. ROIs drawn around whole tumor on highest-*b*-value images on three to six sections near the center of the imaging volume.	ROIs drawn by a consultant radiologist (N.T.) with 9 years of experience, including 8 years of experience in extracranial DW MR imaging. ROIs drawn around tumor on the highest-*b*-value images on three sections at the center of the imaging volume; matching sections selected for second baseline examination.	ROIs drawn by a consultant radiologist (N.T.) with 6 years of experience, including 5 years of experience in extracranial DW MR imaging. ROIs drawn around tumor on the highest-*b*-value images on three sections at the center of the imaging volume; matching sections selected for second baseline examination.	ROIs drawn by an MR physicist (J.M.W.) with 4 years of experience, including 2 years of experience in extracranial DW MR imaging. ROIs drawn by region growing (kidneys and spleen) or freehand (liver). ROIs drawn on computed DW images ($b = 500$ sec/mm^2 for kidneys, $b = 800$ sec/mm^2 for liver, $b = 1000$ sec/mm^2 for spleen) encompassing the whole area of the organ on three contiguous sections.

Note.—References are provided for previously published studies. ROI = region of interest, VOI = volume of interest.

* ROIs were drawn by the authors named in each column; the years of experience stated for each study reflect experience at the time of the original analysis of the study.

Table 2

MR Imaging Protocol for Studies A through K

Parameter	Study A (14,15)	Study B	Study C (16)	Study D	Study E (17)	Study F (19)	Study G (18)	Study H	Study J	Studies K1, K2, and K3 (11)
MR imaging unit	MAGNETOM Avanto (Siemens, Erlangen, Germany)	MAGNETOM Aera (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)	MAGNETOM Avanto (Siemens)
Orientation of imaging plane	Axial	Axial	Coronal oblique	Coronal	Axial	Coronal	Axial	Coronal	Coronal	Axial
Section thickness (mm)	5	5	5	5	5	5	5	5	5	6
Field of view (read × phase) (mm)	380 × 380	380 × 256	380 × 380	380 × 380	340 × 298	300 × 300	380 × 308	380 × 380	380 × 380	380 × 332
Acquired matrix (read × phase)	128 × 128	160 × 108	128 × 128	128 × 128	128 × 112	128 × 128	128 × 104	128 × 128	128 × 128	128 × 112
Reconstructed matrix (read × phase)	256 × 256	320 × 216	256 × 256	256 × 256	256 × 256	256 × 256	256 × 208	256 × 256	256 × 256	256 × 224
Echo time (msec)	69	65	70	64	72	75	69	75	68	75
Repetition time (msec)	3500	9200	2500 to 7000	4000	3500	3500	3500	3500	3500	8000
Fat suppression	SPAIR	SPAIR	SPAIR	SPAIR	Chemical fat suppression	SPAIR	SPAIR	SPAIR	SPAIR	SPAIR
b Values used for ADC estimates (sec/mm^2)	0, 50, 100, 250, 500, 750	50, 600, 900	50, 100, 300, 600, 900	0, 20, 40, 60, 80, 100, 250, 500, 750, 1000*	0, 50, 100, 250, 500, 750	0, 50, 100, 300, 600, 1000	0, 50, 100, 300, 600, 900, 1050	0, 50, 100, 300, 600, 1000	150, 600, 900	100, 500, 900 (*b* = 0 acquired but not used in ADC estimation)
Diffusion-encoding scheme	Three-scan trace	Three-scan trace	Three-scan trace	Three-scan trace	Orthogonal	Three-scan trace	Three-scan trace	Three-scan trace	Three-scan trace	Three-scan trace
Receiver bandwidth (Hz/pixel)	1775	1954	1565	1628	1445	1860	1775	1954	1776	1776
Phase partial Fourier	6/8	6/8	6/8	6/8	6/8	6/8	6/8	7/8	6/8	Not used
No. of signals acquired	6	Four for *b* = 50 and 600 sec/mm^2; five for *b* = 900 sec/mm^2	4	4	5	3	6	5	10	4

Table 2 *(continues)*

Table 2 (continued)

MR Imaging Protocol for Studies A through K

Parameter	Study A (14,15)	Study B	Study C (16)	Study D	Study E (17)	Study F (19)	Study G (18)	Study H	Study J	Studies K1, K2, and K3 (11)
Breathing instructions	Free breathing	Free breathing	Respiratory triggering for liver, splenic, and renal lesions ($n = 10$ patients); free breathing for pelvic nodal, abdominal wall, and peritoneal lesions ($n = 3$ patients)	Free breathing	Free breathing	Free breathing	Free breathing	Free breathing	Free breathing	Free breathing
Acquisition time (min:sec)	6:24	6:28 (Per station)	Variable	11	4	3:30	7	4:51	5:26	5:44

Note.—References are provided for previously published studies. All studies were performed at 1.5 T. The following parameters were common to all studies: single-shot echo-planar imaging; parallel imaging using generalized autocalibrating partially parallel acquisitions, or GRAPPA, with an acceleration factor of two; a bipolar diffusion-gradient scheme; three diffusion-encoding directions; and trace-weighted images. SPAIR = spectral adiabatic nversion recovery.

* Proprietary DW MR imaging prototype packages were used.

and log(ADC_{median}), and the Levene test was used to assess whether the variance of the differences was the same for all studies; the Levene test was also used to assess whether repeatability differed between small, medium, and large VOIs.

The Pearson linear correlation coefficient (Matlab, 2016a; Mathworks) was used to assess correlation between CoV and the year the study started, the number of VOIs in the study, and the median volume of VOIs in the study.

Results

The repeatability of ADC_{mean} was similar to the repeatability of ADC_{median} in all studies (Table 3 and Table E1 [online]); for clarity, only ADC_{median} is shown in Figures 1–4. Bland-Altman plots showed no relationships between differences between pairs of baseline measurements and their means (Fig 1). None of the studies showed a significant difference between pairs of baseline measurements ($P > .05$, paired t test). The repeatability of ADC_{median} (Table 3) and ADC_{mean} (Table E1 [online]) was good, with CoVs between 1.7% and 6.3% for ADC_{median} and between 1.7% and 6.5% for ADC_{mean} for all studies (Fig 2). When we aggregated VOIs from all studies, we found that CoV was 4.1% for ADC_{median} and 3.9% for ADC_{mean}, with upper and lower 95% LoAs of 12.1% and −10.8%, respectively, for ADC_{median} and 11.5% and −10.3% for ADC_{mean}. The Levene test showed a significant difference between studies ($P = .01$ for ADC_{median} and ADC_{mean}) that did not persist after the study with the lowest CoV (study B, which included some of the largest VOIs) was excluded.

There was no correlation between the CoV and the year the studies started (Fig 3, *A*; $r = -0.4$, $P = .2$ for ADC_{median} and $r = -0.3$, $P = .3$ for ADC_{mean}) nor between the CoV and the number of VOIs in each study (Fig 3; *B*, $r = -0.3$, $P = .3$ for ADC_{median} and $r = -0.4$, $P = .2$ for ADC_{mean}). Only weak correlation was demonstrated between the CoV and the median VOI volume in each study (Fig 3, *C*; $r = -0.5$, $P = .1$ for ADC_{median} and

Table 3

Repeatability of ADC_{median}

Study	CoV (%)	95% LoA (%)		RC (log scale)	s_W (log scale)	s_B (log scale)	ICC
		Upper	Lower				
A	4.1 (3.2, 5.6)	11.9 (9.2, 16.6)	−10.6 (−14.3, −8.5)	0.112 (0.088, 0.154)	0.040 (0.032, 0.055)	0.218	0.967 (0.928, 0.985)
B	1.7 (1.4, 2.4)	4.9 (3.8, 7.0)	−4.7 (−6.5, −3.7)	0.048 (0.037, 0.068)	0.017 (0.014, 0.024)	0.279	0.996 (0.991, 0.998)
C	3.2 (2.3, 5.2)	9.4 (6.7, 15.5)	−8.6 (−13.4, −6.3)	0.090 (0.065, 0.144)	0.032 (0.023, 0.052)	0.256	0.984 (0.951, 0.995)
D	6.3 (4.5, 10.7)	19.0 (13.1, 34.4)	−16.0 (−25.6, −11.6)	0.174 (0.123, 0.296)	0.063 (0.045, 0.107)	0.251	0.941 (0.806, 0.984)
E	6.2 (4.2, 11.3)	18.6 (12.5, 36.6)	−15.7 (−26.8, −11.1)	0.171 (0.118, 0.312)	0.062 (0.042, 0.113)	0.147	0.851 (0.504, 0.964)
F	3.0 (2.1, 5.8)	8.8 (5.9, 17.5)	−8.1 (−14.9, −5.5)	0.084 (0.057, 0.162)	0.030 (0.021, 0.058)	0.217	0.981 (0.915, 0.996)
G	3.9 (2.6, 7.5)	11.4 (7.6, 23.1)	−10.3 (−18.8, −7.1)	0.108 (0.073, 0.208)	0.039 (0.026, 0.075)	0.140	0.928 (0.709, 0.985)
H	4.0 (2.7, 8.2)	11.8 (7.7, 25.6)	−10.6 (−20.4, −7.1)	0.112 (0.074, 0.228)	0.040 (0.027, 0.082)	0.150	0.932 (0.696, 0.988)
J	5.2 (3.4, 11.5)	15.5 (9.7, 37.4)	−13.4 (−27.2, −8.9)	0.144 (0.093, 0.317)	0.052 (0.034, 0.115)	0.302	0.971 (0.839, 0.996)
K1	2.6* (1.8, 4.6)	7.5 (5.2, 13.6)	−7.0 (−12.0, −4.9)	0.073 (0.051, 0.127)	0.026 (0.018, 0.046)	0.023	0.427 (−0.205, 0.816)
K2	2.9* (2.0, 5.1)	8.4 (5.8, 15.2)	−7.8 (−13.2, −5.5)	0.081 (0.056, 0.142)	0.029 (0.020, 0.051)	0.042	0.677 (0.158. 0.907)
K3	6.1* (4.3, 10.7)	18.4 (12.5, 34.5)	−15.6 (−25.7, −11.1)	0.169 (0.118, 0.297)	0.061 (0.043, 0.107)	0.023	0.126 (−0.491, 0.673)
All†	4.1 (3.7, 4.7)	12.1 (10.8, 13.8)	−10.8 (−12.2, −9.7)	0.115 (0.103, 0.130)	0.041 (0.037, 0.047)	0.309	0.982 (0.976, 0.987)

Note.—Data in parentheses are lower and upper 95% CIs. s_B = between-subject standard deviation, s_W = within-subject standard deviation. Estimates of ADC_{median} for two baseline examinations for all tumors/organs are tabulated in Table E2 (online).

* CoVs from K1, K2, and K3 reproduced from Winfield et al (11) for completeness.

† Results are shown for each study and for all VOIs ("All") analyzed together.

Figure 1

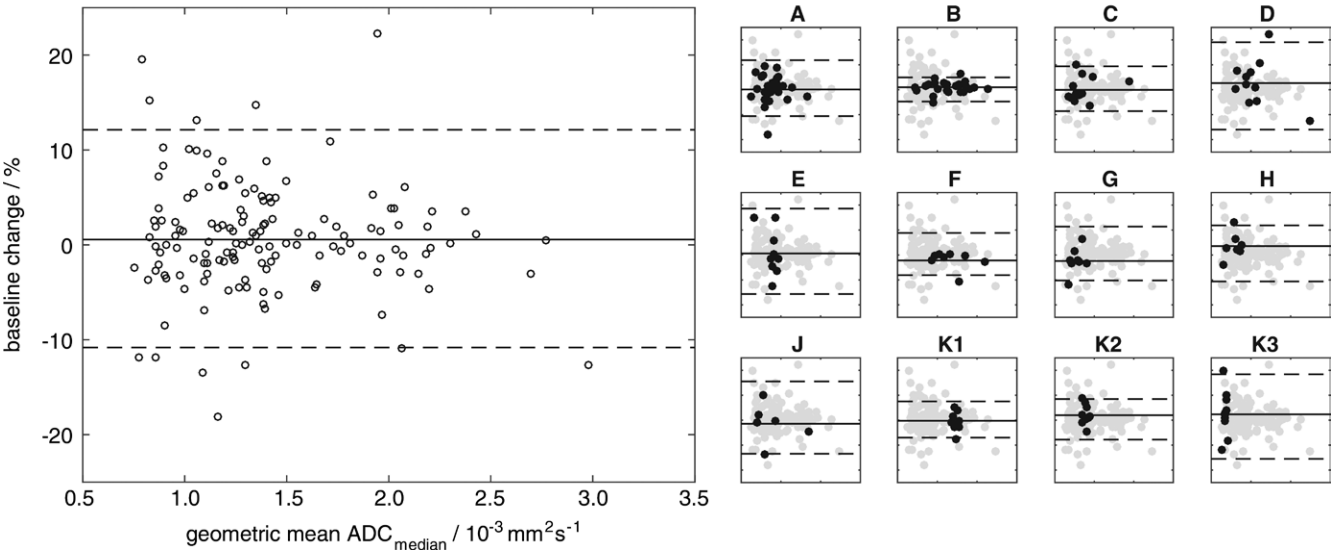

Figure 1: Bland-Altman plot shows percentage change between two baseline estimates of ADC_{median} versus their geometric mean for all VOIs in all studies. Subplots *A* through *K3* show Bland-Altman plots for each study (black dots) with VOIs from all other studies shown as gray dots; the x- and y-axis limits are the same as in the overall plot. On each plot, solid lines = the mean difference between two baseline examinations for the specified data, dashed lines = 95% LoAs.

ADC_{mean}), although the CoV was noticeably lower in one study with very large tumors (study B) compared with other studies. Grouping VOIs into small, medium, and large revealed a significant difference in ADC repeatability between sizes (Levene test; $P = .02$ for ADC_{median} and $P = .04$ for ADC_{mean}), with the lowest CoV for large VOIs (Fig 4). Although 19 VOIs in the large group were from study B, the majority (28 VOIs) were from other studies.

Discussion

The excellent repeatability of ADC_{median} and ADC_{mean} (CoV, between 1.7% and 6.5% in all studies) demonstrates that ADC is a robust metric in clinical

Figure 2

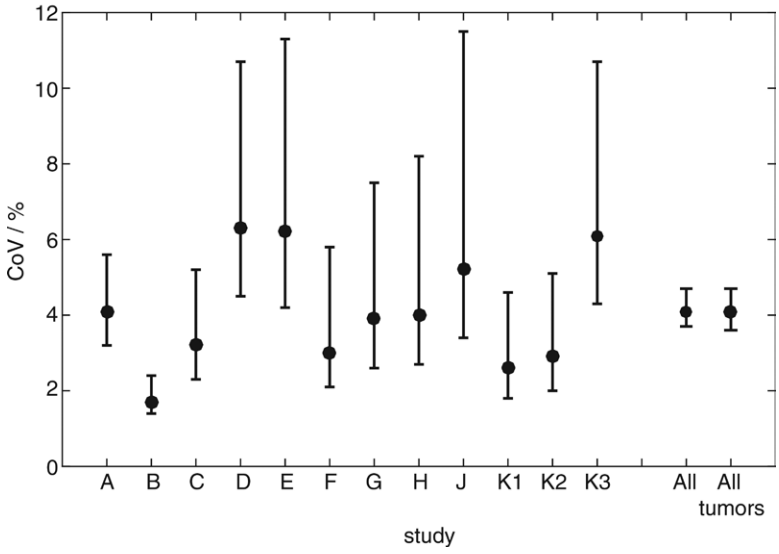

Figure 2: Graph shows CoVs of ADC_{median} for each study (A through K3); all VOIs analyzed together *(All)*; and all tumor VOIs analyzed together *(All tumors)*. Whiskers = 95% CIs for CoV estimates.

practice in oncology. The results reported in this analysis are comparable with results from similar test-retest repeatability studies, although comparison with the literature is hindered by the variety of metrics that have been reported. From the published literature, a study of malignant hepatic tumors (26) reported ICCs in the range of 0.898 to 0.933 and LoAs in the range of 18.8%–24.0% for ADC_{mean}. A study in head-and-neck squamous cell carcinoma (27) reported an RC of 15% for ADC_{mean}. A study of hepatocellular carcinoma (28) reported a CoV of 8.3% and a lower and upper LoA of −41.1% and 18.6%, respectively, for ADC_{mean}. A study in abdominal organs in healthy volunteers (29) reported RCs between 6.4% and 9.6% for ADC_{mean}. A study of normal thyroid glands in healthy volunteers that used reduced-field-of-view DW MR imaging (30) and that also followed the QIBA framework reported an s_w^2 of $0.0147 \times 10^{-3}mm^2\ sec^{-1}$, an RC of $0.3355 \times 10^{-3}mm^2\ sec^{-1}$, an ICC of 0.9273, and a CoV of 9.88%. Comparison between published studies is not straightforward because they report different repeatability metrics, but each result is similar to the present analysis for their respective metrics. Additionally, most studies do not report CIs, which further hinders comparison.

The CoV and LoA, expressed as percentages, may be more intuitive for investigators to understand, compared with s_W or RC expressed on a log scale. Although the ICC is listed in the QIBA framework for reporting repeatability, ICC may not be an appropriate metric for comparison between studies because results are scaled to the intersubject variability of the study cohort by s_B; a low ICC may therefore reflect a homogeneous cohort rather than poor repeatability (13). This is exemplified in study K, where ICCs were low (0.126–0.677 in studies K1, K2, and K3) despite the CoVs being comparable with those in other studies. Values of s_B were an order of magnitude lower than in studies A through J, reflecting the narrow range of ADC estimates in healthy organs in the tightly controlled volunteer cohort. These results strongly suggest that the ICC should not be used to compare ADC repeatability between studies.

Knowledge of ADC repeatability is essential for assessment of posttreatment changes in an individual patient (as opposed to cohort changes, which can be assessed by using a *t* test or similar); knowledge of measurement repeatability is also essential in power calculations to estimate the sample size necessary to detect a treatment effect in prospective cohort studies. Considering changes in ADC after treatment, an increase of 12% or more in ADC_{median} or ADC_{mean} would be outside the 95% LoA for all VOIs analyzed together; even considering the studies with the poorest repeatability (ie, "worst-case" studies), an increase of 20% would have been outside the 95% LoA in all studies. A tumor exhibiting such a change in ADC after treatment would therefore be assessed as exhibiting a posttreatment effect outside the expected variation of repeated measurements, with 95% confidence, when measured with the same imaging unit by using the same imaging protocol, operator, and reader (ie, in repeatability conditions). This can be compared with posttreatment changes reported elsewhere: 23% and 24% increases in ADC_{mean} in responding patients with hepatic metastases from colorectal (3) and gastric (4) cancers, respectively; and increases of 20% (ADC_{mean}) and 22% (ADC_{median}) in responding patients with ovarian cancer treated with platinum-based chemotherapy (8). In studies reporting ADC changes in individual patients, as opposed to cohort changes, posttreatment increases in ADC_{mean} of up to 100% were reported in patients with cervical cancer after chemoradiotherapy (5), and increases in ADC_{mean} of up to 50% were reported in patients with non–small cell lung cancer (9). Hence, the excellent repeatability demonstrated in the present analysis shows that ADC is sensitive to changes that are observed in clinical studies.

The significant difference between small, medium, and large VOIs shows that volume is an important factor in ADC repeatability. The weak correlation between the CoV and the median VOI volume in each study may reflect the range of tumor sizes within each study. The low CoV of 1.7% in study B may relate to the large tumors in that study. For future studies, the assumption of a CoV of 6.5% would be a

Figure 3

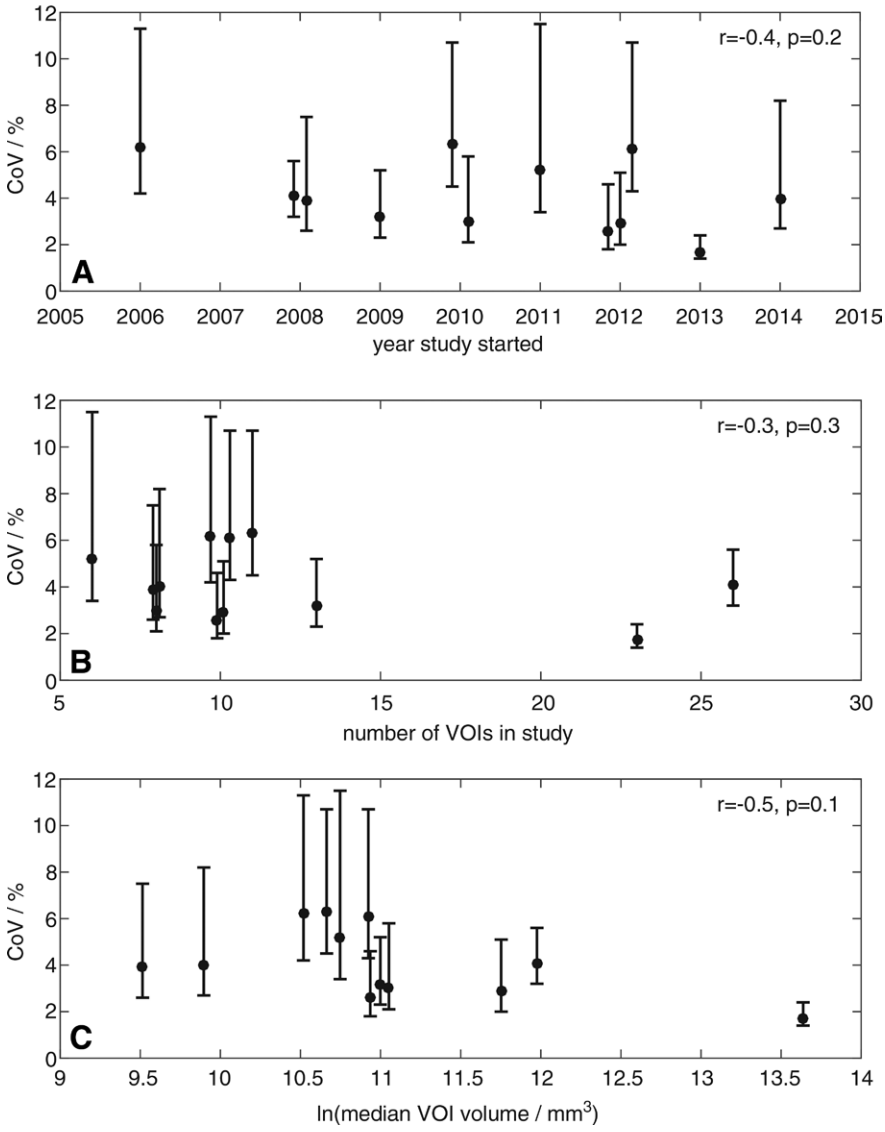

Figure 3: Plots of CoVs of ADC_{median} versus, *A,* the year the study started, *B,* the number of VOIs (subjects or lesions) in the study, and, *C,* the natural logarithm of the median volume of the VOIs in the study. Error bars = 95% CIs of CoV estimates. In *A* and *B,* studies with identical start dates or numbers of VOIs have been offset for clarity.

Figure 4

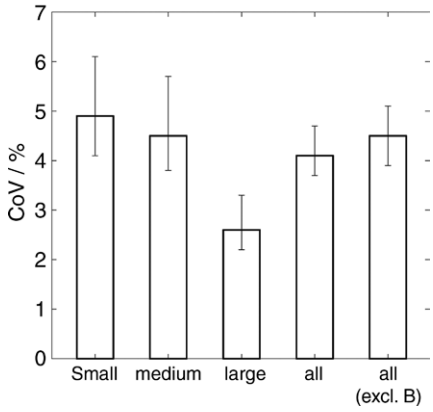

Figure 4: Bar graph shows CoVs of ADC_{median} for small, medium, and large VOIs, all VOIs together, and all VOIs excluding study B. Error bars = 95% CIs of CoV estimates.

conservative choice. It is worthwhile to note that the VOIs did not always encompass the whole tumor: ROIs were drawn around the whole area of the tumor or healthy organ on at least three sections in all studies, but studies A, B, and E included considerably more sections. Larger VOIs may provide more robust estimates of ADC_{median} and ADC_{mean} because of larger sample sizes. Furthermore, larger tumors may be less affected by motion or partial-volume effects, which may lead to better ADC repeatability. ADC repeatability in pediatric patients (study F) was not worse than that in other studies, despite the additional challenges associated with patient compliance in this group.

The apparent absence of a relationship between the CoV and the year the study commenced may suggest that ADC repeatability has not changed markedly over 10 years despite advances in MR imaging unit technology and imaging protocol methods during that time. This suggests that ADC repeatability assessments from older studies may inform future studies, although this may not apply across substantial changes in hardware and/or methods, such as a change in field strength. Although this analysis considered only ADC repeatability, imaging protocol variations may also affect overall image quality, qualitative interpretation, and absolute values of ADC estimates, but these effects are outside the scope of this analysis. Reasons for variations in imaging protocols include changes in hardware and software capabilities; advances in knowledge; requirements for imaging particular patient cohorts, such as the size of the field of view or the orientation of the imaging plane; requirements of study sponsors; and requirements to match protocols in multicenter studies.

The apparent absence of a relationship between the CoV and the number of VOIs in the study (over the range of six to 26 VOIs) may suggest that an informative estimate of repeatability may be obtained from as few as six patients, indicating that double-baseline examinations in relatively small subsets of patients may be used to efficiently

estimate repeatability for larger studies. Repeatability studies may thus be easily conducted if a center wishes to assess its DW MR imaging protocols. Inclusion of larger numbers of subjects, however, allows narrower CIs to be placed on estimated quantities and is advocated in clinical trials.

Repeatability estimates for ADC_{median} and ADC_{mean} do not apply to all summary statistics; for example, other ADC histogram centiles may exhibit poorer repeatability (31). Alternative acquisition techniques (eg, motion compensation) would also require new repeatability studies. Furthermore, it is common practice to use data from previous imaging studies to develop novel analysis methods, which require assessment of the repeatability of resulting metrics to evaluate their potential value in clinical practice. Double-baseline studies therefore provide an invaluable resource for future developments of analysis methods.

There were limitations to our analysis. First, all studies were performed at a single expert center, and senior members of staff with extensive experience of extracranial DW MR imaging were involved in the development of imaging protocols for all studies. Second, all but one of the studies were performed with the same MR imaging unit, with the remaining study performed on a unit from the same manufacturer; the generality of our conclusions for test-retest measurements across MR imaging unit from other manufacturers remains to be tested. Third, only one study in healthy volunteers was included. Fourth, many of the studies were substudies that formed part of a larger clinical trial, and there may have been selection bias because of inclusion and exclusion criteria for these trials (eg, including patients with lesions > 2 cm or excluding patients who had difficulty lying still). Generalization to routine clinical practice remains to be tested, but the repeatability of ADC estimates in less-controlled situations might be expected to be worse than the repeatability reported here.

In conclusion, ADC is a robust imaging metric that demonstrates excellent repeatability in extracranial soft-tissue DW MR imaging studies across a wide range of tumor sites, sizes, patient populations, and imaging protocol variations. Estimates of ADC repeatability obtained from similar data can inform studies where double-baseline measurements are not possible, but a double-baseline format remains critical for future studies.

Acknowledgments: We thank Thorsten Feiweier, PhD, and Berthold Kiefer, PhD, at Siemens Healthcare for providing DW imaging prototype packages. For specific studies included in this work, we acknowledge funding from Cancer Research UK Biomarkers and Imaging Discovery and Development grants C7273/A12064 and C1353/A12762; Cancer Research UK and Engineering and Physical Sciences Research Council (EPSRC) Cancer Imaging Programme at the Children's Cancer and Leukaemia Group in association with the Medical Research Council and Department of Health (England) (C7809/A10342); an Experimental Cancer Medicine Centre Network award (joint initiative, Cancer Research UK and UK Department of Health) grants C51/A7401 and C12540/A15573; EPSRC Platform Grant EP/H046526/1; Experimental Cancer Medicine Centre Network funding for support to early clinical trials; and the support of the National Institute for Health Research through the Cancer Research Network. Some of the studies in this report were supported by AstraZeneca (specifically including support for M.R.O.), Merck, Basilea, ArQule, and Genentech.

Disclosures of Conflicts of Interest: J.M.W. disclosed no relevant relationships. **N.T.** disclosed no relevant relationships. **M.R.** disclosed no relevant relationships. **K.M.** disclosed no relevant relationships. **N.P.J.** disclosed no relevant relationships. **M.G.** disclosed no relevant relationships. **M.D.B.** disclosed no relevant relationships. **D.J.C.** disclosed no relevant relationships. **J.S.d.B.** disclosed no relevant relationships. **T.A.Y.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a consultant for Clovis, Ignyta, and Pfizer; institution has grants or grants pending with AstraZeneca and Vertex; institution has received money for travel, accommodations, and/or meeting expenses from AstraZeneca, GSK, and Vertex. Other relationships: disclosed no relevant relationships. **N.M.d.S.** disclosed no relevant relationships. **S.J.D.** disclosed no relevant relationships. **D.M.K.** disclosed no relevant relationships. **M.O.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: has research agreements with Siemens Medical, Philips Medical, General Electric, and Elekta. Other relationships: disclosed no relevant relationships. **C.M.** disclosed no relevant relationships. **M.R.O.** disclosed no relevant relationships.

References

1. Taouli B, Beer AJ, Chenevert T, et al. Diffusion-weighted imaging outside the brain: consensus statement from an ISMRM-sponsored workshop. J Magn Reson Imaging 2016;44(3):521–540.

2. Dzik-Jurasz A, Domenig C, George M, et al. Diffusion MRI for prediction of response of rectal cancer to chemoradiation. Lancet 2002;360(9329):307–308.

3. Koh DM, Scurr E, Collins D, et al. Predicting response of colorectal hepatic metastasis: value of pretreatment apparent diffusion coefficients. AJR Am J Roentgenol 2007;188(4):1001–1008.

4. Cui Y, Zhang XP, Sun YS, Tang L, Shen L. Apparent diffusion coefficient: potential imaging biomarker for prediction and early detection of response to chemotherapy in hepatic metastases. Radiology 2008;248(3):894–900.

5. Harry VN, Semple SI, Gilbert FJ, Parkin DE. Diffusion-weighted magnetic resonance imaging in the early detection of response to chemoradiation in cervical cancer. Gynecol Oncol 2008;111(2):213–220.

6. Sharma U, Danishad KK, Seenu V, Jagannathan NR. Longitudinal study of the assessment by MRI and diffusion-weighted imaging of tumor response in patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. NMR Biomed 2009;22(1):104–113.

7. Kim S, Loevner L, Quon H, et al. Diffusion-weighted magnetic resonance imaging for predicting and detecting early response to chemoradiation therapy of squamous cell carcinomas of the head and neck. Clin Cancer Res 2009;15(3):986–994.

8. Kyriazi S, Collins DJ, Messiou C, et al. Metastatic ovarian and primary peritoneal cancer: assessing chemotherapy response with diffusion-weighted MR imaging—value of histogram analysis of apparent diffusion coefficients. Radiology 2011;261(1):182–192.

9. Yabuuchi H, Hatakenaka M, Takayama K, et al. Non-small cell lung cancer: detection of early response to chemotherapy by using contrast-enhanced dynamic and diffusion-weighted MR imaging. Radiology 2011;261(2):598–604.

10. National Institute of Standards and Technology. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. NIST technical note 1297. http://www.nist.gov/pml/pubs/tn1297/index.cfm. Published 1994. Accessed June 15, 2016.

11. Winfield JM, Collins DJ, Priest AN, et al. A framework for optimization of diffusion-

weighted MRI protocols for large field-of-view abdominal-pelvic imaging in multicenter studies. Med Phys 2016;43(1):95–110.

12. Raunig DL, McShane LM, Pennello G, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res 2015;24(1):27–67.
13. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology standards for quantitative imaging biomarkers. Radiology 2015; 277(3):813–825.
14. Messiou C, Orton M, Ang JE, et al. Advanced solid tumors treated with cediranib: comparison of dynamic contrast-enhanced MR imaging and CT as markers of vascular activity. Radiology 2012;265(2):426–436.
15. Orton MR, Messiou C, Collins D, et al. Diffusion-weighted MR imaging of metastatic abdominal and pelvic tumours is sensitive to early changes induced by a VEGF inhibitor using alternative diffusion attenuation models. Eur Radiol 2016;26(5):1412–1419.
16. Yap TA, Yan L, Patnaik A, et al. Interrogating two schedules of the AKT inhibitor MK-2206 in patients with advanced solid tumors incorporating novel pharmacodynamic and functional imaging biomarkers. Clin Cancer Res 2014;20(22):5672–5685.
17. Koh DM, Blackledge M, Collins DJ, et al. Reproducibility and changes in the apparent diffusion coefficients of solid tumours treated with combretastatin A4 phosphate and bevacizumab in a two-centre phase I clinical trial. Eur Radiol 2009;19(11):2728–2738.
18. Yap TA, Olmos D, Brunetto AT, et al. Phase I trial of a selective c-MET inhibitor ARQ 197 incorporating proof of mechanism pharmacodynamic studies. J Clin Oncol 2011;29(10):1271–1279.
19. Miyazaki K, Jerome NP, Collins DJ, et al. Demonstration of the reproducibility of free-breathing diffusion-weighted MRI and dynamic contrast enhanced MRI in children with solid tumours: a pilot study. Eur Radiol 2015;25(9):2641–2650.
20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307–310.
21. Keene ON. The log transformation is special. Stat Med 1995;14(8):811–819.
22. Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: keys and clues. Bioscience 2001;51(5):341–352.
23. He X, Oyadiji SO. Application of coefficient of variation in reliability-based mechanical design and manufacture. J Mater Process Technol 2001;119(13):374–378.
24. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. Transl Oncol 2009;2(4):231–235.
25. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–428.
26. Kim SY, Lee SS, Park B, et al. Reproducibility of measurement of apparent diffusion coefficients of malignant hepatic tumors: effect of DWI techniques and calculation methods. J Magn Reson Imaging 2012;36(5):1131–1138.
27. Hoang JK, Choudhury KR, Chang J, Craciunescu OI, Yoo DS, Brizel DM. Diffusion-weighted imaging for head and neck squamous cell carcinoma: quantifying repeatability to understand early treatment-induced change. AJR Am J Roentgenol 2014;203(5):1104–1108.
28. Hectors SJ, Wagner M, Besa C, et al. Intravoxel incoherent motion diffusion-weighted imaging of hepatocellular carcinoma: is there a correlation with flow and perfusion metrics obtained with dynamic contrast-enhanced MRI? J Magn Reson Imaging 2016;44(3):521–540.
29. Miquel ME, Scott AD, Macdougall ND, Boubertakh R, Bharwani N, Rockall AG. In vitro and in vivo repeatability of abdominal diffusion-weighted MRI. Br J Radiol 2012;85(1019):1507–1512.
30. Lu Y, Hatzoglou V, Banerjee S, et al. Repeatability investigation of reduced field-of-view diffusion-weighted magnetic resonance imaging on thyroid glands. J Comput Assist Tomogr 2015;39(3):334–339.
31. Jerome NP, Miyazaki K, Collins DJ, et al. Repeatability of derived parameters from histograms following non-Gaussian diffusion modelling of diffusion-weighted imaging in a paediatric oncological cohort. Eur Radiol 2017;27(1):345–353.