Title

Repeatability and Reproducibility of ADC Histogram Metrics from the ACRIN 6698 Breast Cancer Therapy Response Trial

Journal

Tomography, 6(2)

ISSN

2379-1381

Authors

Newitt, David C

Amouzandeh, Ghoncheh

Partridge, Savannah C

et al.

Publication Date

2020-06-01

DOI

10.18383/j.tom.2020.00008

Repeatability and Reproducibility of ADC Histogram Metrics from the ACRIN 6698 Breast Cancer Therapy Response Trial

David C. Newitt[1], Ghoncheh Amouzandeh[2], Savannah C. Partridge[3], Helga S. Marques[4], Benjamin A. Herman[4], Brian D. Ross[2], Nola M. Hylton[1], Thomas L. Chenevert[2], and Dariya I. Malyarenko[2]

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, CA; [2]Department of Radiology, University of Michigan, Ann Arbor, MI; [3]Radiology, University of Washington, Seattle, WA; and [4]Brown University—Center for Statistical Sciences, ECOG-ACRIN Biostatistics Center, Providence, RI

Corresponding Author:
David C. Newitt, PhD
Department of Radiology and Biomedical Imaging, UCSF/Mt. Zion Hospital, 1600 Divisadero St., Room C254, San Francisco, CA 94115;
E-mail: david.newitt@ucsf.edu

Key Words: Clinical imaging trials, breast cancer therapy response, apparent diffusion coefficient, ADC repeatability, ADC histogram analysis

Abbreviations: American College of Radiology Imaging Network (ACRIN), apparent diffusion coefficient (ADC), confidence interval (CI), Digital Imaging Communication in Medicine (DICOM), diffusion-weighted imaging (DWI), quality control (QC), Quantitative Imaging Biomarker Alliance (QIBA), Quantitative Imaging Network (QIN), repeatability coefficient; (RC), Response Evaluation Criteria In Solid Tumors (RECIST), region of interest (ROI), single-shot echo-planar imaging (SS-EPI), standard deviation (SD), within-subject coefficient of variance (wCV)

ABSTRACT

Mean tumor apparent diffusion coefficient (ADC) of breast cancer showed excellent repeatability but only moderate predictive power for breast cancer therapy response in the ACRIN 6698 multicenter imaging trial. Previous single-center studies have shown improved predictive performance for alternative ADC histogram metrics related to low ADC dense tumor volume. Using test/retest (TT/RT) 4 b-value diffusion-weighted imaging acquisitions from pretreatment or early-treatment time-points on 71 ACRIN 6698 patients, we evaluated repeatability for ADC histogram metrics to establish confidence intervals and inform predictive models for future therapy response analysis. Histograms were generated using regions of interest (ROIs) defined separately for TT and RT diffusion-weighted imaging. TT/RT repeatability and intra- and inter-reader reproducibility (on a 20-patient subset) were evaluated using wCV and Bland–Altman limits of agreement for histogram percentiles, low-ADC dense tumor volumes, and fractional volumes (normalized to total histogram volume). Pearson correlation was used to reveal connections between metrics and ROI variability across the sample cohort. Low percentiles (15th and 25th) were highly repeatable and reproducible, wCV $< 8.1\%$, comparable to mean ADC values previously reported. Volumetric metrics had higher wCV values in all cases, with fractional volumes somewhat better but at least 3 times higher than percentile wCVs. These metrics appear most sensitive to ADC changes around a threshold of 1.2 $\mu m^2/ms$. Volumetric results were moderately to strongly correlated with ROI size. In conclusion, Lower histogram percentiles have comparable repeatability to mean ADC, while ADC-thresholded volumetric measures currently have poor repeatability but may benefit from improvements in ROI techniques.

INTRODUCTION

Serial magnetic resonance imaging (MRI) studies during neoadjuvant chemotherapy (NAC) for breast cancer allow for in vivo observation of changes in the tumor to assess treatment response. Multiparametric breast MRI studies typically include a primary dynamic contrast-enhanced (DCE-MRI) acquisition for lesion visualization. These images can be used for morphologic characterization, quantitative and qualitative enhancement characterization of both lesion and background parenchyma, and quantification of lesion size. DCE-MRI-derived metrics have shown value for prediction of both pathological and survival outcomes for patients with breast cancer (1–3). Functional diffusion-weighted imaging (DWI), which reflects water mobility impeded by cellular constituents and interstitial tortuosity (4), can help in evaluating therapeutic efficacy by reflecting changes in tumor cellularity (5). The apparent diffusion coefficient (ADC) measured by DWI has been shown to improve specificity and positive predictive value of breast magnetic resonance (MR) examinations and to identify early tumor response to cytotoxic effects of breast cancer therapy (6–9).

ISSN 2379-1381 https://doi.org/10.18383/j.tom.2020.00008

Unlike conventional qualitative diagnostic imaging relying on differences in signal intensities, interpretation of changes in quantitative metrics such as ADC requires the measurement of confidence intervals (CIs). Metric changes exceeding the CI will, with 95% confidence, correspond to true parameter changes (beyond measurement error) (10). These intervals are determined by precision (repeatability) and accuracy (bias) of the applied DWI protocol and the physical model for a derived quantitative imaging biomarker (11). The baseline precision can be determined from test/retest (TT/RT) examination performed with identical imaging protocol for study subjects. To reliably detect changes in breast tumor diffusion characteristics, the measured changes in any lesion ADC metric must be compared to corresponding CIs.

Previous single-site studies performed in relatively small subject cohorts have investigated repeatability and reproducibility of breast ADC measures in normal (12–16) and cancerous (13, 14, 17–19) tissue. Within-subject coefficients of variance ranged from 5% to 11%. Recent findings from the multicenter ACRIN 6698 Trial investigating DWI biomarkers for predicting treatment response in breast cancer NAC (20) indicated excellent repeatability of mean and median tumor ADC metrics (21). These results were achieved with a standardized imaging protocol, centralized processing and extensive quality assurance and control procedures. However, mean tumor ADC measures provided only moderate power for predicting treatment outcome (22). Other single-center research suggests that improved tumor characterization may be achieved using alternative histogram metrics (23–26), as well as showing potential relations of volume-based metrics to the clinical standard Response Evaluation Criteria In Solid Tumors (RECIST) criteria (27). The evaluation of precision for such alternative breast tumor ADC histogram metrics is still sparse and based on single-center studies (18, 28).

An objective measure of tumor burden is essential for clinical management and evaluation of cancer therapeutics. Radiographic assessments of solid lesions, including longest diameter, estimators of cross-sectional tumor area, and total tumor volume, have been used as indicators of tumor size and have formed the basis for objective criteria of response by mass shrinkage, as well as disease progression (27, 29, 30). Water mobility, on the other hand, is sensitive to tissue microenvironment, such that lower mobility (reflected by low ADC) implies higher cellular density. There is therefore potential to derive novel useful biomarkers that combine features of both tumor volume and density by means of ADC histogram analysis. Conceptually, the cumulative volume of voxels both within the tumor region of interest (ROI) and having an ADC value below a specified threshold (thus excluding presumably less-dense or already necrotic tissues reflected by higher ADC) provides an estimate of dense tumor volume (23).

In this retrospective study of data from the TT/RT arm of the multisite ACRIN 6698 trial, we analyzed repeatability and reproducibility of ADC histogram and volumetric characteristics to establish confidence intervals for corresponding biomarkers for use in treatment response assessment during breast cancer NAC.

METHODOLOGY

Patient Population

The DW-MRI data for this study was acquired as part of the ACRIN 6698 Trial "Diffusion Weighted MR Imaging Biomarkers for Assessment of Breast Cancer Response to Neoadjuvant Treatment" (20), a sub-study of the multicenter I-SPY 2 TRIAL evaluating novel treatments for breast cancer. ACRIN 6698 was performed at a subset of I-SPY 2 sites that met additional prequalification requirements for performing DW-MRI. Both studies were HIPAA-compliant and performed under IRB approval, and all patients gave informed consent before enrolling. Women of age $\geq$18 were eligible if they had biopsy-confirmed diagnosis of stage II–III disease, and clinically or radiologically measurable disease in the breast with a tumor longest diameter (LD) of $>$2.5 cm. Patients were classified by hormone receptor (HR), human epidermal growth factor receptor-2 (HER2), and MammaPrint (MP) status, and patients with low-risk disease (HR+/HER2−/MP-low) were excluded. A subset of patients participated in the repeatability arm of the trial. For this subset, "coffee break" style TT/RT DWI scans were acquired (as described below) for evaluation of whole-tumor ADC repeatability (21), and were retrospectively analyzed for this current study.

DWI Acquisitions

Details of the multivisit I-SPY 2 MRI protocol, the standardized ACRIN 6698 DWI protocol, and the TT/RT DWI protocol have been previously reported (20, 21, 31). In brief, for repeatability evaluation, T2W and multi b-value DWI images were acquired; the patient was removed from the scanner and repositioned; the scans were repeated. A DCE acquisition was subsequently performed. All imaging was done in the axial plane with full bilateral coverage of the breasts, in the prone position. The standardized DWI protocol required acquisition using a fat-suppressed SS-EPI sequence using b values of 0, 100, 600, and 800 s/mm^2. TT/RT DWI measurements for a given patient were performed on the same day in a single imaging session. A single TT/RT study was conducted for each consented subject at either pretreatment (T0 time-point) or early treatment (T1 time-point, after 3 weeks of treatment), with T0 specified as the preferred time-point. DWI images were assessed with a standardized QA protocol (32), and subjects with either TT or RT scans judged not analyzable owing to protocol deviations or poor image quality were excluded from further analyses.

Whole-Tumor ADC Histogram Analysis

ADC histogram analysis was conducted using the ADC maps and tumor regions of interest (ROIs) defined for the primary study analysis (21). The TT and RT ADC maps were calculated using all b values and a monoexponential decay model. Multislice, whole-tumor ROIs were manually defined by selecting regions with hyperintensity on high b-value DWI (b = 600 or 800 s/mm^2) and relatively low ADC, while avoiding adjacent adipose and fibroglandular tissue, biopsy clip artifacts, and regions of high T2 signal (eg, seroma and necrosis). All apparent disease regions were included in the ROI by using multiple distinct contours per slice and multiple slices as necessary. All voxels from the individual contours were combined into a single composite ROI for histogram analysis. The TT and RT ROIs for a given patient were defined separately and independently with no cross-referencing between the 2 DWI scans, and were defined by the same operator to minimize operator variability. All ROI definitions were reviewed and adjusted, if necessary, by the senior operator

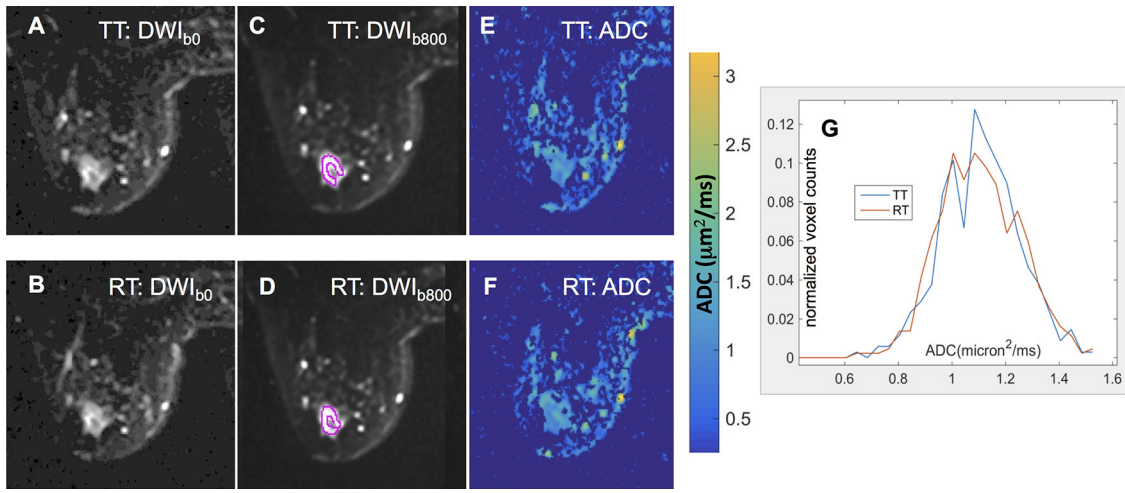

Figure 1. Sample images and apparent diffusion coefficient (ADC) histograms (bin size, 0.04 μm^2/ms) from a typical ACRIN 6698 patient with invasive breast cancer. Grayscale diffusion-weighted imaging (DWI) images for b = 0 s/mm^2 (A, B) and b = 800 s/mm^2 (C, D) illustrate solid tumor region of interest (ROI) segmentation on 1 slice for (A, C) test (TT) and (B, D) retest (RT) scans. The color images (E, F) show the corresponding ADC maps using the quantitative scale provided in the color bar. Normalized ADC histograms (G) are plotted for the full multislice tumor ROI (red: RT, blue: TT).

(reader 1; >10 years of quantitative breast MR analysis experience). The composite ROIs were applied to the derived ADC maps and used to define subject-specific TT and RT histograms. Standard histogram statistics, including mean, standard deviation, skew, kurtosis, median, ranges, and percentiles (5th, 15th, 25th, 50th, 75th, and 95th) were calculated for each histogram. Dense tumor volumes (V_{ADC}), defined as the volume of tissue within the ROI with ADC values below a specified threshold ADC, were calculated by summing the appropriate histogram bins and multiplying by image voxel volume found in the DICOM header. Fractional dense tumor volumes (fV_{ADC}) was calculated as the volume at the ADC threshold (V_{ADC}) divided by the volume at an ADC threshold of 3.0 μm^2/ms ($V_{3.0}$). $V_{3.0}$ corresponds to approximately the full ROI volume, discounting isolated voxels with ADC > 3.0 μm^2/ms resulting from noise. ADC thresholds used were 0.5, 0.6,..., 2.0, 2.5, and 3.0 μm^2/ms.

Repeatability Analysis

The measurement repeatability of each metric across subjects was quantified using Bland–Altman (BA) 95% limits of agreement (LOA) = Mean(TT − RT) ± 1.96 × SD(TT − RT), where Mean(TT − RT) and SD(TT − RT) are the mean and standard deviation of the difference between TT and RT values. The repeatability coefficient, RC = 1.96 × SD(TT − RT) was used for comparisons between metrics of the same units (33). Within-subject coefficient of variance (11),

$$\mathrm{wCV} = \sqrt{\mathrm{mean}\left[\frac{\mathrm{variance}(\mathrm{TT_i}, \mathrm{RT_i})}{\mathrm{mean}^2(\mathrm{TT_i}, \mathrm{RT_i})}\right]}, \qquad [1]$$

was calculated, with 95% upper/lower confidence intervals estimated as (34):

$$CI(95\%) = \sqrt{\left(N \times wCV^2 / \chi^2_{(N,\alpha)}\right)} \qquad [2]$$

where $\chi^2_{(N,\alpha)}$ is the αth percentile of the chi-square distribution with N degrees of freedom; α = 0.975 and 0.025 for the upper and lower bounds, respectively. BA plots were used to compare LOAs among percentile metrics and ADC-threshold volumes. The sources of variability for select metrics were analyzed from populationwise distributions and with intersubject Pearson correlation, R, between metrics and ROI area.

Reproducibility Study

As part of the primary analysis of the TT/RT arm of ACRIN 6698 a reader study for determining intra- and interoperator reproducibility was conducted using the RT scans from a subset of 20 patients. Reader 1 defined whole-tumor ROIs on the DWI twice ("RD1" and "RD1b") while Reader 2 (4 y experience at quantitative breast MRI analysis) defined a single set of ROIs on the studies ("RD2"). The readers operated independently. The ROIs were defined independently from those used in the repeatability analysis, but using the same ROI protocol. The second set of ROIs for intra-operator measures (RD1b) were defined 5–6 weeks after the first set. Reproducibility results for ROI characteristics and for mean tumor ADC were previously reported (21). For the current study we applied the tumor segmentations from the reader study to calculate the intra- and interoperator reproducibility of the histogram percentile, V_{ADC} and fV_{ADC} metrics. Reproducibility was determined using wCV and BA LOA analysis as described above.

All image and statistical analyses were performed using in-house IDL software (Exelis Visual Information Solutions, Boulder, CO), Matlab R2015b toolboxes (MathWorks, Natick, MA) and SAS™ software version 9.4 (SAS, Cory, NC).

Table 1. Test/Retest Repeatability of Histogram Metrics

	Units	Mean[a]	wCV (%)	wCV 95% CI (%)		Delta[b]	BA RC[c]
Mean ADC	μm^2/ms	1.17	5.36	4.63	6.46	−0.010	0.156
15th Pctl	μm^2/ms	0.93	8.07	6.97	9.73	−0.010	0.174
25th Pctl	μm^2/ms	1.00	6.58	5.69	7.95	−0.006	0.160
50th Pctl	μm^2/ms	1.15	5.44	4.70	6.56	−0.009	0.158
$V_{0.9}$	cm^3	1.53	44.0	38.1	53.1	−0.035	1.451
$V_{1.1}$	cm^3	3.32	36.5	31.6	44.1	−0.111	2.184
$V_{1.3}$	cm^3	4.96	29.1	25.1	35.1	−0.259	3.281
$fV_{0.9}$		0.24	42.4	36.7	51.2	0.018	0.230
$fV_{1.1}$		0.51	30.9	26.7	37.3	0.010	0.211
$fV_{1.3}$		0.72	15.5	13.4	18.7	0.002	0.165

[a] Mean = mean[(TT + RT)/2].
[b] delta = mean(TT − RT);.
[c] BA RC = 1.96 × SD(TT − RT): repeatability coefficient of the BA LOA.

RESULTS

The ACRIN 6698 Trial consented 89 patients (median age, 47 years; range, 27–73 years) from 9 institutions to the TT/RT substudy. Of those, 18 patients were excluded from analysis owing to either MRI protocol inconsistencies between TT and RT acquisitions (N = 3) or unacceptable image quality on TT and/or RT scans (N = 15). Scans from the remaining 71 patients (median age, 46 years; range, 27–71 years), including 60 pretreatment

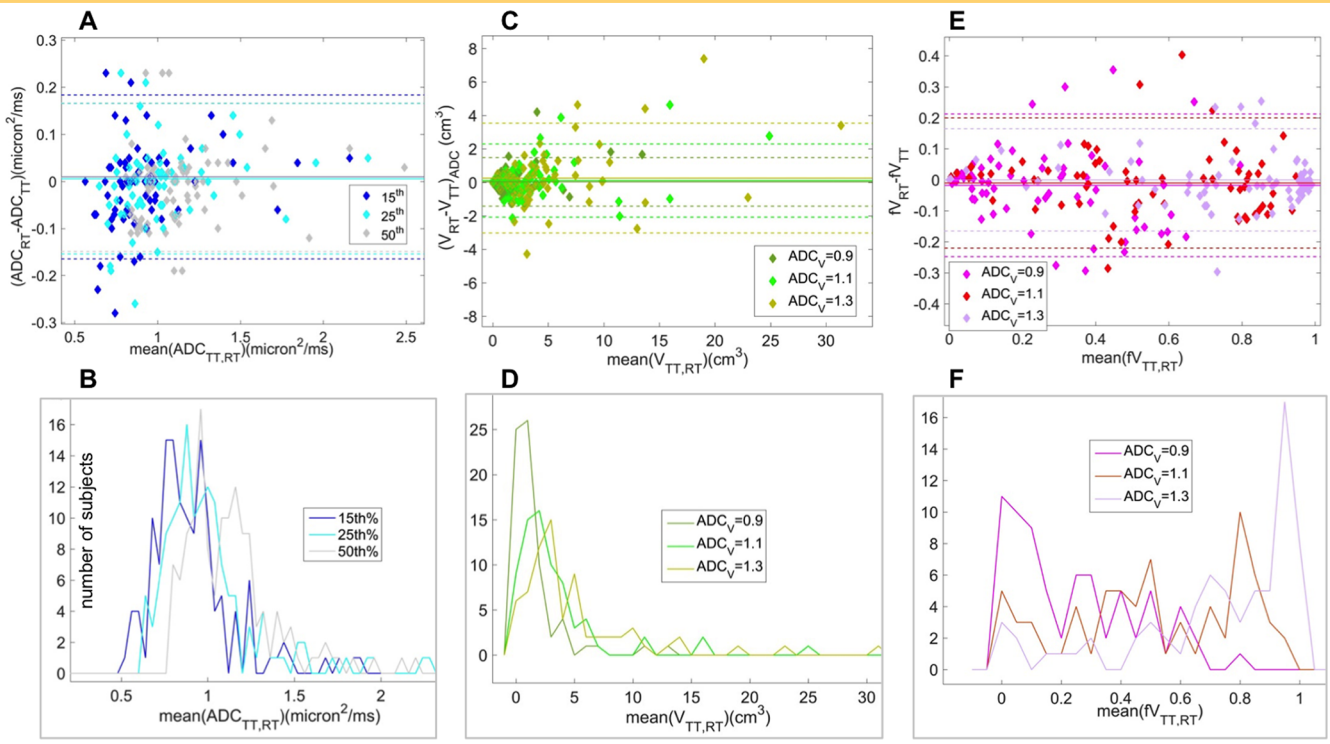

Figure 2. Bland–Altman plots (top) and corresponding test/retest (TT/RT) mean distributions (bottom) are shown (left-to-right) for: (A, B) 15th, 25th, and 50th ADC percentiles; (C, D) low ADC volumes, V_{ADC}, thresholded at ADC < 0.9, 1.1, and 1.3 μm^2/ms; and (E, F) fractional volumes, fV_{ADC} (V_{ADC} normalized by total histogram volume $V_{3.0}$). The 95% limits of agreement for mean metrics are shown by dashed lines. The symbol and line assignments are color-coded in the legends. Mean value histograms (B, D, F) were calculated using bin sizes of 0.04 μm^2/ms for percentiles, 1 cm^3 for V_{ADC} and 0.05 for fV_{ADC}. For fV_{ADC} mean values (F), we see that for thresholds around 1.1 μm^2/ms, the mean values are reasonably evenly distributed across most of the range of the metric (0.0 to 1.0).

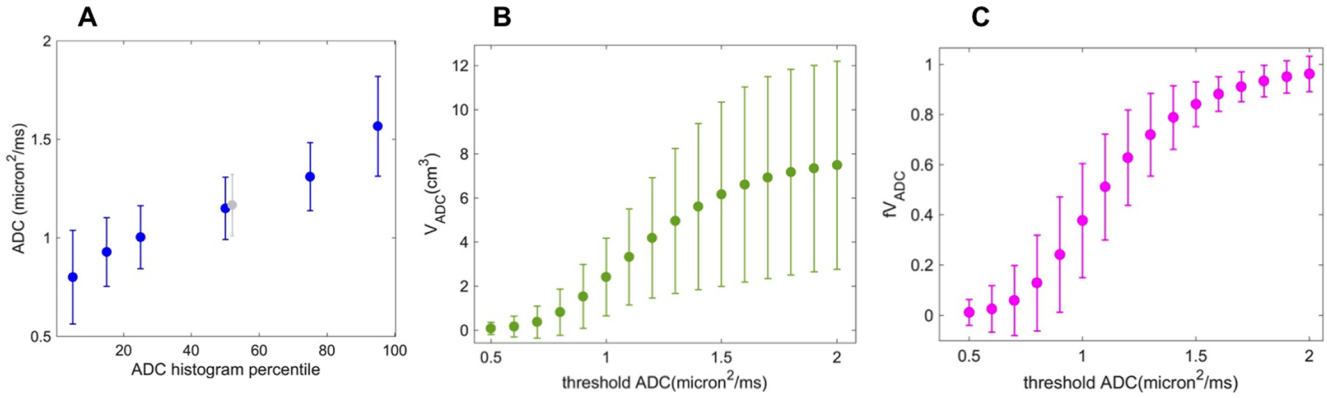

Figure 3. Sample mean values of the histogram metrics are plotted for: (A) ADC histogram percentiles; (B) low ADC volumes V_{ADC}; and (C) fractional volumes fV_{ADC} (V_{ADC} normalized to total histogram volume $V_{3.0}$). The central gray data point in (A) corresponds to mean tumor ADC. The error bars illustrate the 95%CI [mean ± RC, RC = repeatability coefficient = 1.96 × SD(TT − RT)] on the individual measurements. CI are tight for histogram percentiles indicating good precision, except at the extremes of 5th and 95th percentiles. For V_{ADC} repeatability is poor. At lower thresholds we see large repeatability coefficient (RC) values relative to the means, while at high thresholds, CI are very wide owing to ROI variability. At high thresholds for fV_{ADC}, the CI are tight, but this is likely just reflecting the very small number of high ADC voxels included in the manually selected ROIs.

(T0) and 11 early-treatment (T1) visits, were analyzed for this study. This cohort was identical to that analyzed for the original study mean ADC repeatability analysis (21). Figure 1 shows T2-weighted images (b = 0 s/mm^2), high-b-value DWI images, the corresponding ADC maps, and the segmented tumor ADC histograms for TT and RT acquisitions from 1 subject. These illustrate typical differences in TT and RT tumor ROI segmentation and ADC map noise, leading to variations of the respective histogram characteristics.

Repeatability results for DWI histogram metrics are given in Table 1 and presented graphically in Figure 2, with values for the mean ADC included for comparison. Highest precision (wCV = 5.44%) was observed for the 50th percentile (median) metric, whose sample distribution overlapped with the distribution for the mean ADC metric. This overlap is consistent with the Gaussian measurement noise being the main source of observed TT-RT variations for ADC histogram percentile metrics. Precision was also good for moderately lower percentile metrics, with wCV = 8.1% and 6.6% for 15th and 25th percentiles, respectively, but was degraded to 13.9% at the fifth percentile. The BA plot for selected histogram percentile values (Figure 2A) illustrates consistent repeatability patterns for 15th, 25th, and 50th percentiles. The LOAs were very similar for these metrics: RC values = 0.174, 0.160, and 0.158 μm^2/ms (LOA shown as horizontal dashed lines in the BA plot). The histograms of the binned mean values for these percentile metrics across our cohort are shown in Figure 2B. For the 15th percentile metric, 85% (60/71) of all cases and 92% of pretreatment cases (50/60) had ADC values < 1.1 μm^2/ms, indicating the presence of appreciable dense tumor tissue in these cases.

Precision was lower for ADC-thresholded volume metrics (V_{ADC}), and it had considerable variation across the tested ADC thresholds (Table 1; Figure 2, C and D). wCV values were >50% for ADC thresholds <0.9 μm^2/ms, indicating very poor repeatability for these measures. At higher ADC thresholds (≥1.5 μm^2/ms), V_{ADC} is dominated by the volume of the whole-tumor ROI for the majority of cases, as tissue with ADC above these thresholds would be included in the ROI only by error. This resulted in wCV ≅ 27% for these thresholds, representing the repeatability of the ROI size. $V_{1.5}$ was >80% of the total ROI volume for 56 (79%) of all cases and 50 (83%) of the pretreatment cases. We therefore focused analysis on moderate threshold values of 0.9, 1.1, and 1.3 μm^2/ms, finding wCV = 44.0%, 36.5%, and 29.1%. Figure 2 C and D shows the BA plots and mean value histograms for the $V_{0.9}$, $V_{1.1}$, and $V_{1.3}$ volume measures. The sample means (RC) for these 3 thresholds were 1.5 (1.5), 3.3 (2.2), and 5.0 (3.3) cm^3 (LOAs shown in Figure 2C). RC values exceeded the mean metric values for all lower threshold volumes, consistent with low repeatability of these metrics. Results for fractional volumes ($fV_{ADC} = V_{ADC}/V_{3.0}$; ADC = 0.9, 1.1, 1.3 μm^2/ms) are shown in Table 1 and Figure 2, E and F. wCV values were lower than respective V_{ADC} values but still a factor of 3–5 times greater than the percentile measure wCV values. For ADC thresholds 0.9 and 1.1 μm^2/ms, fractional volumes were distributed fairly uniformly across the range from 0 to 1 (Figure 2F, dark and light green).

Figure 3 shows the dependence on metric parameter of the sample means and 95%CI [mean ± RC, RC = repeatability coefficient = 1.96 × SD(TT − RT)] for ADC percentiles (A), V_{ADC} volumes (B), and fV_{ADC} fractional volumes (C) across the full range of the parameters examined. The tightest CI were observed for 15th to 75th ADC percentiles, with some drop off in precision at the extremes of 5th and 95th percentiles, indicating good precision measurements across a very wide range. For V_{ADC} measures at thresholds at or above 1.4 μm^2/ms, wide CI and relatively small changes in sample means with ADC threshold changes indicate the limiting effects of ROI size dependence and ROI

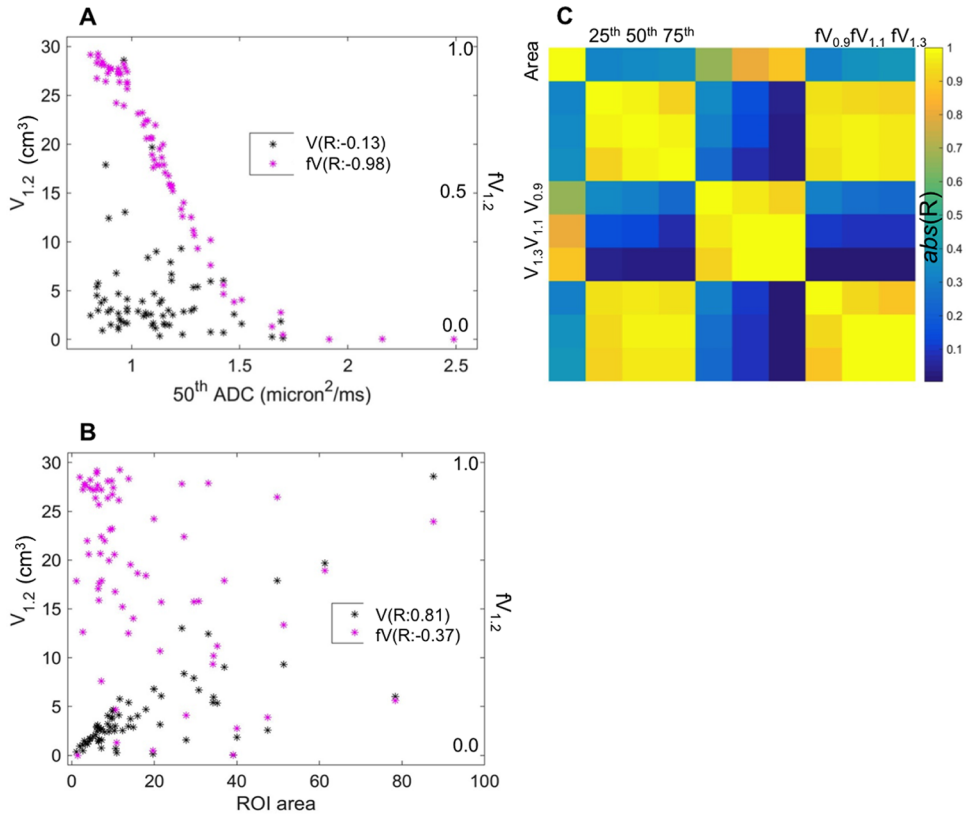

Figure 4. Scatter plots (A, B) for low ADC volume (V_{ADC}, left Y-axis) and fractional volume (fV_{ADC}, right Y-axis), both at threshold ADC = 1.2 μm^2/ms, illustrate relative correlations to the 50th ADC percentile metric (A) and to ROI area (B). $fV_{1.2}$ (magenta) showed much greater correlation to histogram percentile, and somewhat smaller correlation to ROI area, as compared to the absolute volume $V_{1.2}$. The corresponding Pearson correlation coefficients (R) are listed in the legends. Correlation map (C) graphically illustrates intermetric correlations for (top left to bottom right): ROI area, ADC percentiles (25th, 50th, and 75th), low ADC volumes ($V_{0.9}$, $V_{1.1}$, and $V_{1.3}$), and corresponding fractional volumes ($fV_{0.9}$, $fV_{1.1}$, and $fV_{1.3}$). High correlation ($|R| > 0.8$) is observed within each metric type (3×3 boxes along diagonal). Fractional volumes fV_{ADC} have greater correlation to ADC percentiles and lesser correlation to ROI measures relative to the corresponding V_{ADC}. *P*-values were consistent with significant correlations ($P < 10^{-4}$) within each metric type, between fractional volumes and histogram percentiles, and between absolute volumes and ROI area.

variability on these measurements. Specifically, in this threshold range, V_{ADC} is reflecting a manually determined total tumor volume based primarily on the high-b-value image intensity. This volume has high variability owing to operator subjectivity. We therefore expect generally poor sensitivity for detecting volume changes with treatment in this parameter range. Figure 3B also indicates that confident measurement of typical V_{ADC} changes, in particular the negative tumor volume changes most commonly associated with therapy response, may be limited at thresholds <0.9 μm^2/ms. In this range the RC is near to or greater than the mean, putting the lower CI below 0 cm^3. With current ROI techniques the most promising threshold range for low ADC volume measurements would appear to be between 0.9 and 1.3 μm^2/ms. The fV_{ADC} measures (Figure 3C) showed a similar effect of excessive variability at low ADC thresholds, and also lost sensitivity at higher thresholds due to compression of values against the upper limit of fV_{ADC} = 1.0. In the moderate threshold range, the 95%CI for fV_{ADC} metric appeared somewhat tighter than that for the absolute V_{ADC} volumes.

Correlations between different histogram metrics and with ROI total area are shown in Figure 4. For the correlation analysis, the cohort was limited to the 60 patients with TT/RT acquisitions at the pretreatment (T0) study time point, to avoid complications from the upward shift in the overall population tumor ADC distribution with NAC treatment. The scatter plots illustrate the correlations between absolute and fractional volumes $V_{1.2}$ and $fV_{1.2}$ with the median ADC (50th percentile, Figure 4A) and the ROI area (Figure 4B). $fV_{1.2}$ indicates a strong correlation with the median (R = −0.98), which was not seen for the absolute $V_{1.2}$ (R = −0.13). Figure 4B shows the strong correlation (R = 0.81) between volume $V_{1.2}$ and ROI area, pointing toward ROI variability as the most significant contributor to the poor repeatability for the volume metrics. This correlation is reduced but still moderate (R = −0.37) for the corresponding fractional volume. The color chart in Figure 4C shows the correlation results (Pearson R values) for pairwise comparisons between ROI area (left column and top row) and the 9 ADC histogram metrics. High correlation ($|R| > 0.8$) was observed within each metric type (3×3 arrays

Table 2. Intraoperator Reproducibility of Histogram Metrics

	Units	Mean[a]	wCV (%)	wCV 95% CI (%)		Delta[b]	BA RC[c]
Mean ADC	$\mu m^2/ms$	1.14	4.97	3.80	7.18	0.028	0.173
15th Pctl	$\mu m^2/ms$	0.92	4.68	3.58	6.75	0.036	0.142
25th Pctl	$\mu m^2/ms$	0.99	3.82	2.92	5.52	0.026	0.113
50th Pctl	$\mu m^2/ms$	1.12	4.83	3.70	6.98	0.025	0.155
$V_{9.0}$	cm^3	1.20	41.2	31.5	59.5	−0.073	0.363
$V_{1.1}$	cm^3	3.01	37.5	28.7	54.2	−0.075	0.707
$V_{1.3}$	cm^3	4.44	37.1	28.4	53.6	0.159	1.832
$fV_{0.9}$		0.19	41.2	31.5	59.5	−0.031	0.083
$fV_{1.1}$		0.51	36.9	28.2	53.3	−0.067	0.206
$fV_{1.3}$		0.77	33.7	25.8	48.7	−0.027	0.208

[a] Mean = mean[(RD1 + RD1b)/2].
[b] delta = mean(RD1 − RD1b).
[c] BA RC = 1.96 × SD(RD1 − RD1b): repeatability coefficient of the BA LOA.

along the diagonal), between percentile values and fV_{ADC}, and between V_{ADC} and ROI area. Normalization of the volumes to create fractional volumes reduced the correlation to ROI area, but it was still moderate. *P*-values were consistent with significant correlations ($P < 10^{-4}$) for all the comparisons indicating high correlation.

Tables 2 and 3 give intra- and interoperator reproducibility, respectively, for histogram metrics evaluated using the RT data (second acquisition on each patient) on a 20-patient subset. Results for reproducibility followed similar patterns to the repeatability results: the histogram percentiles between 15th and 50th showed good reproducibility, with wCV(95%CI) ranging from 3.8% (2.9, 5.5) to 4.8% (3.7, 7.0) for intraoperator variability and 3.8% (2.9, 5.5) to 5.3% (4.1, 7.7) for interoperator variability. Volume-based measures were considerably less reproducible. In our primary range of interest for ADC thresholds from 0.9 to 1.3 $\mu m^2/ms$ there was no discernable dependence on wCV of V_{ADC} with threshold. For fV_{ADC} reproducibility values showed the expected trend of lower wCV (higher precision) with higher thresholds, with wCV(95%CI) values for threshold ADC = 1.3 $\mu m^2/ms$ the lowest at 34% (26%, 49%) and 13% (10%, 20%) for intra- and interoperator reproducibility, respectively. The poor wCV values and large CI for volume measures were consistent with greater dependence on ROI characteristics as seen in the repeatability measures presented above. However, substantial variability in wCV estimates may be also due to the small number of patients in the reproducibility cohort, and for fV_{ADC} measures, the wCV model constraints may not be well satisfied, as errors in these measures may not be proportional to the mean values.

Table 3. Interoperator Reproducibility of Histogram Metrics

	Units	Mean[a]	wCV (%)	wCV 95% CI (%)		Delta[b]	BA RC[c]
Mean ADC	$\mu m^2/ms$	1.15	5.57	4.26	8.04	0.001	0.177
15th Pctl	$\mu m^2/ms$	0.93	4.35	3.33	6.29	0.014	0.108
25th Pctl	$\mu m^2/ms$	0.99	3.83	2.93	5.53	0.014	0.105
50th Pctl	$\mu m^2/ms$	1.13	5.32	4.07	7.68	0.006	0.168
$V_{9.0}$	cm^3	1.10	39.64	30.15	57.90	0.121	0.631
$V_{1.1}$	cm^3	2.84	29.55	22.47	43.16	0.270	1.117
$V_{1.3}$	cm^3	4.32	27.28	20.75	39.85	0.408	2.016
$fV_{0.9}$		0.18	31.35	23.84	45.80	−0.014	0.135
$fV_{1.1}$		0.49	20.23	15.38	29.54	−0.027	0.261
$fV_{1.3}$		0.75	13.35	10.16	19.50	0.001	0.260

[a] Mean = mean[(RD1 + RD2)/2].
[b] delta = mean(RD1 − RD2).
[c] BA RC = 1.96 × SD(RD1 − RD2): repeatability coefficient of the BA LOA.

DISCUSSION

This study provides baseline precision and reproducibility for ADC histogram-based metrics along both ADC (ADC percentile) and volume dimensions. In our study cohort of patients undergoing NAC for invasive breast cancer, repeatability is better for ADC percentiles versus low ADC volumes, the latter appearing more sensitive to ROI segmentation variations. Fractional-volumes, that is low ADC volumes normalized to the total tumor ROI volume, show reduced sensitivity to segmentation variability. However, compared with all volumetric measures, the low ADC percentiles (15th and 25th), which are of interest for quantifying changes in dense tumor tissue with treatment, showed at least 3-fold better repeatability and lower sensitivity to segmentation variability.

For precise measurement of response during NAC, it is critical to quantify changes in malignant tumor burden. This can be done with a variety of techniques including linear dimension measurements by clinical examination or from imaging studies (eg, RECIST), or volumetric measures such as functional tumor volume from DCE-MRI examinations (35). DW-MRI has the ability to more specifically quantify solid or viable tumor volumes, based on their low ADC values of <1 $\mu m^2/ms$ (23). However, our analysis indicates relatively low precision of such volume measurements when coupled with manual segmentation. Improved segmentation consistency, either through better prescribed procedures or preferably through more highly automated techniques, is likely needed for useful measurement of treatment-induced changes in ADC-based solid tumor volumes in the breast cancer NAC realm. The use of fractional dense tumor volumes, normalized to the full ROI volume, alleviated some of the dependence on segmentation reproducibility. The wCV values were still relatively poor, but this may be reflective in part of breakdown of the wCV model when the errors are not proportional to the mean. The fractional volume metrics did show strong correlations to the histogram percentile metrics, indicating a possible functional equivalence between them. The changes in these metrics over treatment will be explored as potential biomarkers for therapy response prediction in a future study.

The most significant limitation to this study was the restriction to manually defined whole-tumor ROIs. Given the great heterogeneity among breast cancer lesions, the wide variety of imaging platforms in the multicenter study, and the complex ROI definition procedure, there was a lot of variability introduced in the analysis. This limits the determination of a true repeatability value for the tested metrics. We were also limited by relatively small sample sizes, having only 71 analyzable studies for repeatability, with a very unequal split between the T0 and T1 time-points, and a further limitation to 20 subjects for the reader study.

In conclusion, development and validation of quantitative imaging tools for supporting cancer patient trials and ultimately routine clinical adoption have been the focus of the National Institutes of Health Quantitative Imaging Network over the past decade (34, 36–38). In this present study, we found that ADC histogram percentiles down to 15% have high repeatability and reproducibility, comparable to mean ADC, while low-ADC volumetric measures were substantially less repeatable. Tumor segmentation variability appeared to be the main source of TT/RT error for volume-based ADC histogram metrics. High correlation of ADC percentiles to fractional tumor volumes indicated functional equivalence, and both low percentile distribution and fractional volume analysis suggested best sensitivity to volume changes for ADC between 0.9 and 1.3 $\mu m^2/ms$. The diagnostic and predictive performance of these biomarkers will be evaluated in future work.

ACKNOWLEDGMENTS

National Institutes of Health Grants: U01CA225427, R01CA132870, U01CA166104, R01CA190299, P01CA085878, P30CA008748, and R35CA197701.

This study was conducted in cooperation with the ECOG-ACRIN Cancer Research Group (Peter J. O'Dwyer, MD and Mitchell D. Schnall, MD, PhD, Group Co-Chairs) and supported by the National Cancer Institute of the National Institutes of Health under the following award numbers: CA180794 and CA180820. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

We are thankful to QIBA RSNA statistician, Dr. Nancy Obuchowski, for advice on repeatability analysis.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the United States and other countries. ® indicates United States registration.

Disclosure: Dr. Newitt reports research support from Kheiron Medical Technology to institution, outside the submitted work. Dr. Hylton reports research support from Kheiron Medical Technology to institution and research support from GE Healthcare to institution, outside the submitted work. Dr. Chenevert and Dr. Ross are coinventors on intellectual property assigned to and managed by the University of Michigan licensed by Imbio for histogram and fDM analysis. All other authors declare no conflict of interest.

REFERENCES

1. Ah-See M-LW, Makris A, Taylor NJ, Harrison M, Richman PI, Burcombe RJ, Stirling JJ, d'Arcy JA, Collins DJ, Pittam MR, Ravichandran D, Padhani AR. Early changes in functional dynamic magnetic resonance imaging predict for pathologic response to neoadjuvant chemotherapy in primary breast cancer. Clin Cancer Res. 2008;14:6580–6589.
2. Hylton NM, Gatsonis CA, Rosen MA, Lehman CD, Newitt DC, Partridge SC, Bernreuter WK, Pisano ED, Morris EA, Weatherall PT, Polin SM, Newstead GM, Marques HS, Esserman LJ, Schnall MD. Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival-results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. Radiology. 2016;279:44–55.
3. Li K-L, Partridge SC, Joe BN, Gibbs JE, Lu Y, Esserman LJ, Hylton NM. Invasive breast cancer: predicting disease recurrence by using high-spatial-resolution signal enhancement ratio imaging. Radiology. 2008;248:79–87. .
4. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology. 1988;168:497–505.
5. Chenevert TL, Stegman LD, Taylor JM, Robertson PL, Greenberg HS, Rehemtulla A, Ross BD. Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors. J Natl Cancer Inst. 2000;92:2029–2036.
6. Galbán CJ, Ma B, Malyarenko D, Pickles MD, Heist K, Henry NL, Schott AF, Neal CH, Hylton NM, Rehemtulla A, Johnson TD, Meyer CR, Chenevert TL, Turnbull LW,

Ross BD. Multi-site clinical evaluation of DW-MRI as a treatment response metric for breast cancer patients undergoing neoadjuvant chemotherapy. PloS One. 2015;10: e0122151.
7. Li X-R, Cheng L-Q, Liu M, Zhang Y-J, Wang J-D, Zhang A-L, Song X, Li J, Zheng Y-Q, Liu L. DW-MRI ADC values can predict treatment response in patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. Med Oncol. 2012;29:425–431.
8. Partridge SC, Nissan N, Rahbar H, Kitsch AE, Sigmund EE. Diffusion-weighted breast MRI: clinical applications and emerging techniques. J Magn Reson Imaging. 2017;45:337–355.
9. 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;22:104–113.
10. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, Wahl RL, Kurland BF, Schwarz AJ, Gönen M, Zahlmann G, Kondratovich MV, O'Donnell K, Petrick N, Cole PE, Garra B, Sullivan DC. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res. 2015;24:27–67.
11. Shukla-Dave A, Obuchowski NA, Chenevert TL, Jambawalikar S, Schwartz LH, Malyarenko D, et al. Quantitative Imaging Biomarkers Alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. J Magn Reson Imaging. 2019;49:e101–e21.
12. Aliu SO, Jones EF, Azziz A, Kornak J, Wilmes LJ, Newitt DC, Suzuki SA, Klifa C, Gibbs J, Proctor EC, Joe BN, Hylton NM. Repeatability of quantitative MRI measurements in normal breast tissue. Transl Oncol. 2014;7:130–137.
13. Giannotti E, Waugh S, Priba L, Davis Z, Crowe E, Vinnicombe S. Assessment and quantification of sources of variability in breast apparent diffusion coefficient (ADC) measurements at diffusion weighted imaging. European journal of radiology. 2015;84:1729–1736.
14. Jang M, Kim SM, Yun BL, Ahn HS, Kim SY, Kang E, Kim S-W. Reproducibility of apparent diffusion coefficient measurements in malignant breast masses. J Korean Med Sci. 2015;30:1689–1697.
15. O'Flynn EA, Morgan VA, Giles SL, deSouza NM. Diffusion weighted imaging of the normal breast: reproducibility of apparent diffusion coefficient measurements and variation with menstrual cycle and menopausal status. Eur Radiol. 2012;22:1512–1518.
16. Partridge SC, Murthy RS, Ziadloo A, White SW, Allison KH, Lehman CD. Diffusion tensor magnetic resonance imaging of the normal breast. Magn Reson Imaging. 2010;28:320–328.
17. Petralia G, Bonello L, Summers P, Preda L, Malasevschi A, Raimondi S, Di Filippi R, Locatelli M, Curigliano G, Renne G, Bellomi M. Intraobserver and interobserver variability in the calculation of apparent diffusion coefficient (ADC) from diffusion-weighted magnetic resonance imaging (DW-MRI) of breast tumours. Radiol Med. 2011;116:466–476.
18. Spick C, Bickel H, Pinker K, Bernathova M, Kapetas P, Woitek R, Clauser P, Polanec SH, Rudas M, Bartsch R, Helbich TH, Baltzer PA. Diffusion-weighted MRI of breast lesions: a prospective clinical investigation of the quantitative imaging biomarker characteristics of reproducibility, repeatability, and diagnostic accuracy. NMR Biomed. 2016;29:1445–1453.
19. Clauser P, Marcon M, Maieron M, Zuiani C, Bazzocchi M, Baltzer PA. Is there a systematic bias of apparent diffusion coefficient (ADC) measurements of the breast if measured on different workstations? An inter- and intra-reader agreement study. Eur Radiol. 2016;26:2291–2296.
20. ACRIN-6698: Diffusion Weighted MR Imaging Biomarkers for Assessment of Breast Cancer Response to Neoadjuvant Treatment: A sub-study of the I-SPY 2 TRIAL. 2012 [cited 2018 February 2]. Available from: http://www.acrin.org/Portals/0/Protocols/6698/Protocol-ACRIN6698_v2.29.12_active_ForOnline.pdf.
21. Newitt DC, Zhang Z, Gibbs JE, Partridge SC, Chenevert TL, Rosen MA, Bolan PJ, Marques HS, Aliu S, Li W, Cimino L, Joe BN, Umphrey H, Ojeda-Fournier H, Dogan B, Oh K, Abe H, Drukteinis J, Esserman LJ, Hylton NM. Test-retest repeatability and reproducibility of ADC measures by breast DWI: results from the ACRIN 6698 trial. J Magn Reson Imaging. 2019;49:1617–1628.
22. Partridge SC, Zheng Z, Newitt DC, Gibbs JE, Chenevert TL, Rosen MA, Bolan PJ, Marques HS, Romanoff J, Cimino L, Joe BN, Umphrey HR, Ojeda-Fournier H, Dogan B, Oh K, Abe H, Drukteinis JS, Esserman LJ, Hylton NM; ACRIN 6698 Trial Team and I-SPY 2 Trial Investigators. Diffusion-Weighted MRI Predicts Pathologic Response in Neoadjuvant Treatment of Breast Cancer: the ACRIN. Multicenter Trial Radiology. 2018;289:618–627.
23. Chenevert TL, Malyarenko DI, Galbán CJ, Gomez-Hassan DM, Sundgren PC, Tsien CI, Ross BD. Comparison of voxel-wise and histogram analyses of glioma ADC maps for prediction of early therapeutic change. Tomography. 2019;5:7–14.
24. Pope WB, Kim HJ, Huo J, Alger J, Brown MS, Gjertson D, Sai V, Young JR, Tekchandani L, Cloughesy T, Mischel PS, Lai A, Nghiemphu P, Rahmanuddin S, Goldin J. Recurrent glioblastoma multiforme: ADC histogram analysis predicts response to bevacizumab treatment. Radiology. 2009;252:182–189.
25. Pope WB, Lai A, Mehta R, Kim HJ, Qiao J, Young JR, Xue X, Goldin J, Brown MS, Nghiemphu PL, Tran A, Cloughesy TF. Apparent diffusion coefficient histogram analysis stratifies progression-free survival in newly diagnosed bevacizumab-treated glioblastoma. AJNR Am J Neuroradiol. 2011;32:882–889.
26. Wen Q, Jalilian L, Lupo JM, Molinaro AM, Chang SM, Clarke J, Prados M, Nelson SJ. Comparison of ADC metrics and their association with outcome for patients with newly diagnosed glioblastoma being treated with radiation therapy, temozolomide, erlotinib and bevacizumab. J Neurooncol. 2015;121:331–339.
27. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–247.
28. Jensen LR, Garzon B, Heldahl MG, Bathen TF, Lundgren S, Gribbestad IS. Diffusion-weighted and dynamic contrast-enhanced MRI in evaluation of early treatment effects during neoadjuvant chemotherapy in breast cancer patients. J Magn Reson Imaging. 2011;34:1099–1109.
29. Hersberger KE, Mendiratta-Lala M, Fischer R, Kaza RK, Francis IR, Olszewski MS, Harju JF, Shi W, Manion FJ, Al-Hawary MM, Sahai V. Quantitative imaging assessment for clinical trials in oncology. J Natl Compr Canc Netw. 2019;17:1505–1511.
30. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, Lin NU, Litière S, Dancey J, Chen A, Hodi FS, Therasse P, Hoekstra OS, Shankar LK, Wolchok JD, Ballinger M, Caramella C, de Vries EGE, RECIST working group. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol. 2017;18:e143–e52.
31. I-SPY2. I-SPY 2: Neoadjuvant and Personalized Adaptive Novel Agents to Treat Breast Cancer. 2010 [updated January 25, 2018; cited 2018 February 2]. Available from: https://clinicaltrials.gov/ct2/show/NCT01042379.
32. Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharmacol Ther. 2009;86:97–100.
33. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. Transl Oncol. 2009;2:231–235.
34. Paudyal R, Konar AS, Obuchowski NA, Hatzoglou V, Chenevert TL, Malyarenko DI, Swanson SD, LoCastro E, Jambawalikar S, Liu MZ, Schwartz LH, Tuttle RM, Lee N, Shukla-Dave A. Repeatability of quantitative diffusion-weighted imaging metrics in phantoms, head-and-neck and thyroid cancers: preliminary findings. Tomography. 2019;5:15–25.
35. Newitt DC, Aliu SO, Witcomb N, Sela G, Kornak J, Esserman L, Hylton NM. Real-time measurement of functional tumor volume by MRI to assess treatment response in breast cancer neoadjuvant clinical trials: validation of the Aegis SER software platform. Transl Oncol. 2014;7:94–100.
36. Farahani K, Tata D, Nordstrom RJ. QIN benchmarks for clinical translation of quantitative imaging tools. Tomography. 2019;5:1–6.
37. Malyarenko DI, Swanson SD, Konar AS, LoCastro E, Paudyal R, Liu MZ, Jambawalikar SR, Schwartz LH, Shukla-Dave A, Chenevert TL. Multicenter repeatability study of a novel quantitative diffusion kurtosis imaging phantom. Tomography. 2019;5:36–43.
38. Sorace AG, Wu C, Barnes SL, Jarrett AM, Avery S, Patt D, Goodgame B, Luci JJ, Kang H, Abramson RG, Yankeelov TE, Virostko J. Repeatability, reproducibility, and accuracy of quantitative mri of the breast in the community radiology setting. J Magn Reson Imaging. 2018;48. [Epub ahead of print]