Newitt David (Orcid ID: 0000-0002-1735-6682)

Test-Retest Repeatability and Reproducibility of ADC Measures by Breast DWI: Results from the ACRIN 6698 Trial

David C. Newitt, PhD [1], Zheng Zhang, PhD [2,3,4], Jessica E. Gibbs, BA [1], Savannah C. Partridge, PhD [5], Thomas L. Chenevert, PhD [6], Mark A. Rosen, MD PhD [7], Patrick J. Bolan, PhD [8], Helga S. Marques, MS [3,4], Sheye Aliu, PhD [1], Wen Li, PhD[1], Lisa Cimino, RT [9], Bonnie N. Joe, MD PhD [1], Heidi Umphrey, MD [10], Haydee Ojeda-Fournier, MD [11], Basak Dogan, MD [12,13], Karen Oh, MD [14], Hiroyuki Abe, MD PhD [15], Jennifer Drukteinis, MD [16,17], Laura J. Esserman, MD [18], Nola M. Hylton, PhD [1]; For the ACRIN 6698 Trial Team and I-SPY 2 TRIAL Investigators

[1] Department of Radiology and Biomedical Imaging, University of California, San Francisco

[2] Department of Biostatistics, Brown University

[3] Center for Statistical Sciences, Brown University

[4] American College of Radiology Imaging Network (ACRIN)

[5] Department of Radiology, University of Washington

[6] Department of Radiology, University of Michigan

[7] Department of Radiology, Hospital of the University of Pennsylvania

[8] Department of Radiology, University of Minnesota

[9] American College of Radiology & ECOG-ACRIN Cancer Research Group

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jmri.26539

[10] Department of Radiology, University of Alabama, Birmingham

[11] Department of Radiology, University of California, San Diego

[12] Department of Radiology, University of Texas MD Anderson Cancer Center, Houston

[13] Department of Diagnostic Radiology, University of Texas Southwestern Medical Center

[14] Department of Radiology, Oregon Health & Science University

[15] Department of Radiology, University of Chicago

[16] H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida

[17] Department of Women's Imaging, St. Joseph's Women's Hospital, Tampa, Florida.

[18] Department of Surgery, University of California, San Francisco

Corresponding author:

David C. Newitt, PhD

UCSF/Mt. Zion Hospital

1600 Divisadero St., Room C254

San Francisco, CA 94115

Phone: (415) 885-7507

FAX: (415) 885-3884

Email: david.newitt@ucsf.edu

Grant Support:

This research was supported by the National Cancer Institute (NCI) through the grants U01 CA151235, R01 CA132870, U01 CA166104, R01 CA151326, P41 EB015894. *ACRIN receives funding from the NCI through the grants U01 CA079778, U01 CA080098,* U24 CA180803

Running title: Test-Retest Repeatability of Breast DWI

ABSTRACT

BACKGROUND: Quantitative diffusion-weighted MRI (DWI) is a promising technique for cancer characterization and treatment monitoring. Knowledge of the reproducibility of DWI metrics in breast tumors is necessary to apply DWI as a clinical biomarker.

PURPOSE: To evaluate the repeatability and reproducibility of breast tumor apparent diffusion coefficient (ADC) in a multi-institution clinical trial setting, using standardized DWI protocols and quality assurance (QA) procedures.

STUDY TYPE: Prospective

SUBJECTS: 89 women from nine institutions undergoing neoadjuvant chemotherapy for invasive breast cancer.

FIELD STRENGTH/SEQUENCE: DWI was acquired before and after patient repositioning using a four b-value, single-shot echo-planar sequence at 1.5T or 3.0T.

ASSESSMENT: A QA procedure by trained operators assessed artifacts, fat suppression, and signal-to-noise ratio, and determined study analyzability. Mean tumor ADC was measured via manual segmentation of the multi-slice tumor region referencing DWI and contrast-enhanced images. 20 cases were evaluated multiple times to assess intra- and inter-operator variability. Segmentation similarity was assessed via the Sørenson-Dice similarity coefficient.

STATISTICAL TESTS: Repeatability and reproducibility were evaluated using within subject coefficient of variation (wCV), intra-class correlation coefficient (ICC), agreement index (AI), and repeatability coefficient (RC). Correlations were measured by Pearson's correlation coefficients.

RESULTS: 71 cases (80%) passed QA evaluation: 44 at 1.5T, 27 at 3.0T; 60 pre-treatment, 11 after three weeks of taxane-based treatment. ADC repeatability was excellent: wCV=4.8% (95% CI 4.0, 5.7%), ICC=0.97 (95%CI 0.95, 0.98), AI=0.83 (95%CI 0.76, 0.87), and RC=0.16 * 10^{-3} mm^2/sec (95% CI 0.13, 0.19). Results were similar across field strength and timepoint subgroups. Reproducibility was excellent: inter-reader ICC=0.92 (95%CI 0.80, 0.97) and intra-reader ICC=0.91 (95%CI 0.78, 0.96).

DATA CONCLUSION: Breast tumor ADC can be measured with excellent repeatability and reproducibility in a multi-institution setting using a standardized protocol and QA procedure. Improvements to DWI image quality would reduce loss of data in clinical trials.

KEYWORDS

Breast cancer, treatment response, breast MRI, diffusion, reproducibility

INTRODUCTION

MRI has become a standard method to assess breast cancer in diagnostic, high-risk screening and treatment response scenarios. The adoption of neoadjuvant, or preoperative, chemotherapy (NAC) as standard treatment for locally advanced invasive breast cancer has provided an opportunity to use serial MRI studies for *in vivo* observation of changes in the tumor to assess treatment response. Results from the ACRIN 6657 clinical trial (1,2) demonstrated the value of MRI-derived metrics for prediction of both pathological and survival outcomes. The primary MRI technique for assessing breast cancer is dynamic contrast-enhanced (DCE) imaging, which characterizes tissue vascularity. However, growing evidence supports diffusion-weighted imaging (DWI), a functional imaging technique reflecting water diffusion properties in tissue, as a supplemental and/or alternative technique to DCE (3-12). Specifically, the apparent diffusion coefficient (ADC) measured by DWI reflects water mobility impeded by cellular constituents and interstitial tortuosity, and it has shown promise as an imaging biomarker to measure early tumor response to cytotoxic treatment (13).

In order to utilize ADC as a reliable biomarker for monitoring therapeutic effects, the underlying variability of the measurement must be understood. In addition to patient physiologic factors (e.g., menstrual phase, breast density, and menopausal status), technical sources of variability include both those more

accessible to the investigator's control such as the MRI protocol and analysis methods employed, and those more difficult to control such as variations in patient positioning, patient motion, and intra- and inter-operator variability for non-automatic analysis schemes. Previous single-site studies performed in relatively small numbers of subjects have investigated repeatability and reproducibility of breast ADC measures in normal (14-18) and cancerous (15,16,19-21) tissue. Within-subject coefficients of variation ranged from 5% to 11%, and reproducibility inter-operator intraclass correlation coefficients (ICC) varied widely from 0.47 to >0.9. These studies leave open the question of the reproducibility of breast tumor ADC measurements, and hence their overall value in the assessment of treatment response, when applied in multi-center clinical trials. The lack of appropriate reproducibility data was cited by QIBA (The Radiology Society of North America's Quantitative Imaging Biomarkers Alliance) at time of initial draft as the primary reason for excluding breast from the 2017 QIBA Profile for DWI (22), stating that several organs including the breast: "*… have been excluded for the time being due to lack of sufficient literature (test-retest data from a total of ~35 subjects, either from a single publication or in total from multiple manuscripts) support.*"

The American College of Radiology Imaging Network ACRIN 6698 Trial (3,23) is a sub-study of the multi-institution I-SPY 2 TRIAL (Investigation of serial

studies to predict your therapeutic response with imaging and molecular analysis 2) (24), a phase II treatment trial using response-adaptive randomization within breast cancer subtypes to evaluate investigational agents for women with high-risk stage II/III breast cancer. The primary aim of the ACRIN 6698 imaging trial was to evaluate whether changes in whole-tumor mean ADC measured early in the course of NAC treatment are predictive of pathologic complete response (pCR).

The purpose of this test-retest sub-study was to prospectively evaluate the repeatability and reproducibility of breast tumor ADC measures in a multi-institution, multi-MRI-platform clinical trial setting, using a standardized MR-DWI protocol and quality assurance (QA) procedure.

MATERIALS AND METHODS

Site Qualification

All actively enrolling or newly qualified I-SPY 2 imaging sites were eligible to participate in the ACRIN 6698 trial. Sites qualified for participation via a process consisting of evaluation of phantom and human exams as prescribed in the ACRIN 6698 study materials (23). Scanners could be of any manufacturer, with field strength of 1.5 or 3.0 tesla, and had to employ a dedicated bilateral breast radiofrequency coil. Qualification was specific to the particular scanner and coil combination. While qualification of multiple scanners by a site was

allowed, all MRI studies for any given patient were required to be performed using the same scanner configuration (manufacturer, field strength, model, and breast coil model). Qualification scans using a bilateral ice/water mixture phantom with a known ADC value ($1.1 * 10^{-3}$ mm^2/sec) (25) were performed using the ACRIN 6698 study DWI protocol (Table 1) and submitted to a core lab at the University of Michigan for quality control analysis. Scanner qualification required a maximum ADC bias in standardized breast imaging regions of less than 10%. In addition, two clinical DWI test scans using the ACRIN 6698 study protocol on volunteers or currently enrolled I-SPY 2 patients were required prior to study imaging. These *in vivo* scans were evaluated for protocol compliance and adequate DWI image quality for ADC measurement in breast tissue.

Patient Population

Patients were enrolled into the ACRIN 6698 trial at qualified study centers following I-SPY 2 trial procedures and inclusion/exclusion criteria (26). Both studies were HIPAA-compliant and performed under individual site IRB approval, and all patients gave informed consent prior to enrolling. Women aged 18 and over were eligible if they had biopsy-confirmed diagnosis of stage II–III disease, and clinically or radiologically measureable disease in the breast with a tumor longest diameter (LD) >2.5 cm. Patients were classified by biomarker assessments of hormone receptor (HR), human epidermal growth factor

receptor-2 (HER2), and MammaPrint (MP) status performed at pre-treatment; patients with low-risk disease (HR+/HER2-/MP-low) were excluded. For patients enrolled into ACRIN 6698, a separate IRB-approved consent was used for enrollment into the test/retest arm of the trial. Details on the population subgroup for the test/retest arm are given in the results section below.

MRI Protocol

The MRI component of the ACRIN 6698 trial consisted of four sequential studies: pre-treatment (T0), early-treatment (T1), mid-treatment (T2), and post-treatment (T3) (Figure 1). T2W, multi b-value DWI, and DCE acquisitions were taken at each study timepoint. Imaging was done in the axial plane with full bilateral coverage of the breasts. The DWI protocol was standardized to the greatest extent possible given constraints of the multiple scanner platforms, and required acquisition using a fat suppressed SS-EPI sequence with four b values (0, 100, 600, 800 sec/mm^2). Specified parameter ranges for the DWI protocol are given in Table 1. DWI was implemented with a single series multi-b DWI acquisition or with three consecutive two b-value acquisitions (0/100, 0/600 and 0/800 sec/mm^2). Test and retest DWI measurements for a given patient were performed on the same day in a single imaging session. The patient was positioned normally (prone) and scanned with initial localization, T2W, and DWI acquisitions. They were then removed from the scanner and taken off the

scanner bed, and then repositioned as before. The full ACRIN 6698 protocol was then performed, consisting of localization, T2W, DWI, and DCE acquisitions. A single test/retest study was conducted for each consented subject at either T0 or T1, with T0 specified as the preferred timepoint.

DWI Image Analysis

DICOM images from all acquisitions were deidentified and transmitted using TRIAD^{TM} (*Transfer of Images and Data*, ECOG-ACRIN, Philadelphia, Pa.) to the UCSF imaging core lab for centralized analysis. DWI images were assessed with a standardized QA protocol (27) for the three quality categories of artifacts, fat suppression, and signal-to-noise ratio (SNR), and given an overall quality rating of poor, moderate, or good. Poor quality studies were judged not analyzable and were excluded from the study. Moderate and good quality studies were then evaluated as analyzable or not analyzable based on the degree to which any negative quality issues were found to prevent confident definition of a region-of-interest (ROI) contoured to the whole primary tumor region. QA evaluation was done by trained operators, blinded to pathologic outcomes, with a minimum of five years experience evaluating breast MR images.

ADC parametric maps were created based on the classic monoexponential decay model (28):

$$S(b) = S_0 * e^{-b*ADC}$$

where S(b) is the signal intensity with diffusion weighting b. ADC was calculated using a linear least-squares fit of the log of the signal intensities at all four b-values. For studies using three consecutive two b-value acquisitions (in contrast to the standard four b-value acquisition), an automatic protocol check was used to ensure that no MR parameters other than b-value were varied and then the three acquisitions were combined into a single four b-value series for analysis. All centralized DWI image processing was done using UCSF developed software written in IDL (Exelis Visual Information Solutions, Boulder, Colorado).

Tumors were identified on post-contrast DCE subtraction images and then localized on the corresponding DWI images. Multi-slice, 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 (e.g., seroma and necrosis). For all cases, including large and multicentric/multifocal disease, all apparent disease regions were included in the ROI by using multiple distinct contours per slice and multiple slices as required to cover the entire tumor region without including intervening or adjacent tissues. All voxels from the individual contours were combined into a single composite ROI for statistical analysis. The test and retest ROIs for a given patient were defined separately and independently with no cross-referencing between the two DWI scans, and

were performed by the same operator to minimize operator variability. All ROI definitions were reviewed and adjusted, if necessary, by the senior operator (final review performed by J.E.G. with over 10 years quantitative breast MR analysis experience). Mean and median ADC values were calculated for each composite whole-tumor ROI.

Whole-tumor ROIs for the evaluation of intra- and inter-operator variability were defined independently of those for test/retest repeatability, but using the same procedures as described above. However, unlike the ROIs for the repeatability measurement, no final review of the ROIs by a senior operator was done for the reader study. For this reader reproducibility study only the 2nd "retest" acquisition on each subject was analyzed. Reader one (J.E.G., senior study operator) analyzed each case twice for intra-operator assessment (RD1, RD1B), while reader 2 (W.L.) did a single analysis (RD2) that was compared to RD1 for inter-operator assessment. RD1 and RD1B analyses for intra-operator variability were conducted 5-6 weeks apart. In addition to comparisons of the mean tumor ADC values, ROI reproducibility was directly evaluated using the Sørenson-Dice similarity coefficient (DSC) (29), given for two ROIs "A" and "B" by:

$$DSC = 2 \times V_{AB} / (V_A + V_B)$$

where V_{AB} is the volume common to both ROIs and V_A and V_B are the individual ROI volumes. Thus DSC varies from 0 for no overlap between the regions, to 1 for complete overlap (identical ROIs). For every case in the reader study DSC was evaluated between each pair of ROIs: [RD1 vs. RD1B], [RD1 vs. RD2], and [RD1B vs. RD2]; and the mean of the three values was used to estimate the ROI variability of each case:

$$DSC_{Mn} = (DSC_{[RD1,RD2]} + DSC_{[RD1B,RD2]} + DSC_{[RD1,RD1B]})/3.$$

Scatter plots and Pearson's correlation were used to estimate the dependence of the ROI variability with ADC and DCE-derived tumor characteristics.

Statistical Analysis

The reproducibility of each marker was assessed using four different indices: within subject coefficient of variation (wCV) (30), intraclass correlation coefficient (ICC) (31), agreement index (AI) (32), and repeatability coefficient (RC) (33). Both RC and wCV are based on within-subject standard deviation (wSD), where RC=2.77*wSD and wCV=100%*wSD/mean. RC has the same units as the marker, while wCV is unit-less. Both estimates are unbounded in the upper range and sensitive to extreme outliers. ICC is derived from the ANOVA model estimates, while AI is based on the data's overall ranking. Both ICC and AI are bounded (-1 to 1 for ICC and 0.5 to 1 for AI) and unit-less. The confidence intervals of AI were done with a bootstrap method.

RESULTS

Enrollment

Between August 2012 and January 2015, 406 patients were enrolled into the ACRIN 6698 trial at 10 United States institutions. 388 patients aged 23 to 77 years (median age 48 years) were found eligible under the I-SPY 2 inclusion criteria. Of these, 89 patients aged 27 to 73 years (median age 47 years) from nine institutions were consented to the test/retest sub-study. Individual sites were limited to 14 test/retest patients to better balance the accrual across different MRI scanner manufacturers and field strengths. Eleven patients from one site were imaged using three consecutive two b-value acquisitions, while the rest were done with single four b-value scans. Table 2 shows accrual numbers and QA results for the test/retest study, including a breakdown by site, visit, field strength, and MRI scanner vendor. Three patients (3.4%) were excluded for major test/retest protocol deviations: change in slice thickness, retest done after contrast injection, and changes between the successive two b-value acquisition parameters. Of the remaining 86 patients, both the test and retest scans for 71 patients (median age 46, range 27-71 years) from eight institutions were assessed as analyzable for tumor ADC through the image QA process. Fifteen subjects were rejected for image quality issues that prevented analysis of one (N=7) or both (N=8) of the DWI acquisitions, giving an overall rejection rate for image quality issues for DWI acquisitions of 13.4% (23 of 172 acquisitions).

Image quality issues in the rejected acquisitions included excessive artifacts (N=15), fat suppression failure (N= 12), and poor SNR (N=10).

ADC Repeatability

Figure 2 shows typical images and ADC maps for two example cases that illustrate poor (Subject 1, $|ADC_{difference}|/ADC_{mean}=0.117$) or good ($|ADC_{difference}|/ADC_{mean}=0.025$) repeatability. The left panels show the DCE subtraction images (nominal 2.5 minutes post-injection – pre-injection) that highlight the enhancing regions of the breast; these images were used for tumor localization. The center (test) and right (retest) panels show the corresponding slices of the ADC maps for the two DWI acquisitions. The differences in the ADC maps illustrate the effects of patient repositioning combined with the relatively thick slices in the DWI acquisitions. Repositioning makes perfect matching of ADC slices very unlikely yielding different partial-volume effects across the thick slices, thus contributing to the observed differences between the independently defined test and retest whole-tumor ROIs.

Comparisons of test and retest ADC measurements are shown in Bland-Altman plots in Figure 3. Cases measured at T0 (pre-treatment) are represented as blue diamonds and T1 (early-treatment) cases as red circles. The test and retest ADC values were similar for the 71 subject analyzable cohort (Figure 3a) with minimal bias between test and retest (Mean (SD) ADC: 1.16 (0.32) and 1.17

(0.31) x10^{-3} mm^2/sec, respectively) over tumor ADC values ranging from 0.80 to 2.62 x10^{-3} mm^2/sec. Absolute ADC values trended higher for the T1 cases, as expected due to treatment effects, but repeatability appeared similar. Figures 3b and 3c show Bland-Altman plots for subgroups defined by field strength. We observed minimal bias and no trend in ADC difference with mean ADC for these subgroups.

Repeatability of tumor ADC values was excellent by all measures for the whole cohort and within subgroups (Figure 4a,b, Table 3). For the whole cohort, RC = 0.16 x 10^{-3} mm^2/s (95% CI 0.13 to 0.19), wCV = 4.8% (95%CI 4.0, 5.7%), ICC = 0.97 (95% CI 0.95, 0.98), and AI = 0.83 (95% CI 0.76, 0.87). For the subgroups defined by the field strength or the treatment points, ICC values ranged from 0.91 to 0.99, AI from 0.81 to 0.83, and wCV from 3.6 to 5.2% — with generally tight 95% confidence intervals as shown by the error bars in Figure 4. The AI results for the 11 patients done at timepoint T1 show a wider range in the 95% CI, due to the small sample size and the use of a bootstrap technique for this calculation.

ADC and ROI Reproducibility

For the reader variability study, a random sample of 20 cases (median age 43, range 27-69 years) were selected in proportion to the vendor and field strength representations in the main analysis set (see Table 2, “N Reader Study”

column). Table 3 (bottom two rows) and Figure 5 show results for inter- and intra-operator reproducibility. All measures showed good reproducibility – e.g., ICC was 0.92 (95% CI 0.80, 0.97) for inter-operator and 0.91 (0.78, 0.96) for intra-operator variability. Bland-Altman plots (Figure 5) indicated minimal bias between the measurements and no trends in ADC difference with increasing mean ADC.

The only source of operator variability in our ADC analysis is in the definition of the whole-tumor ROIs. To quantify these variations we compared the ROIs for inter- and intra-observer tests using the Sørenson-Dice similarity coefficient (DSC) as a measure of ROI similarity. Median DSC across the 20 patient cohort was 0.69 (range 0.33-0.79) for inter-operator ROI pairs ($DSC_{[RD1,RD2]}$) and 0.70 (range 0.40-0.84) for intra-operator pairs ($DSC_{[RD1,RD1B]}$). The inter-operator and intra-operator similarity values ($DSC_{[RD1,RD2]}$ and $DSC_{[RD1,RD1B]}$) were strongly correlated (Figure 6a, Pearson's r=0.84, $p < 0.0001$, 95% CI 0.63, 0.94). To evaluate possible factors affecting ROI variability, we calculated the Pearson's correlation between the mean DSC for each case (DSC_{Mn}) and tumor characteristics – including the mean ADC, and the MRI DCE-derived functional tumor volume (34) and tumor surface area. A trend of decreasing DSC_{Mn} with increased ADC was observed (r=-0.49, p=0.03, 95% CI -0.77, -0.06; Figure 6b), consistent with the readers having increased difficulty in defining ROIs later in treatment when tumor ADC values are closer to those of

normal breast tissue. No significant correlation was found between DSC_{Mn} and ADC variability (r=-0.37, p=0.11, Figure 6c), nor between DSC_{Mn} and the DCE-based tumor size measures (r=0.34, p=0.14 for volume and r=0.32, p=0.17 for surface area).

Protocol Compliance

A post hoc detailed protocol compliance assessment was done both with respect to the ACRIN 6698 DWI acquisition specifications and for consistency between the test and retest scans on each subject. Overall compliance within the 89 patient test/retest cohort was reasonably high, with 166 (93%) of the 178 DWI acquisitions within protocol specifications (Table 1). Within the N=71 analyzable cohort, three subjects had protocol deviations that were considered major as they resulted in sizable differences in TE (78 msec test vs. 107-114 msec retest) and TR (7.1 sec test vs. 8.9 - 9.4 sec retest). Nine other subjects, while having DWI parameters within study specifications, had differences between test and retest protocols. Only one of these deviations – a swapping of frequency/phase encoding directions – was considered major. The whole cohort wCV was essentially unchanged when the four subjects with major-deviations were excluded from the analyzable cohort, dropping from 4.8% to 4.7%.

DISCUSSION

There is notable interest in using quantitative diffusion metrics as biomarkers for breast cancer treatment response, as recently summarized in the review article by Partridge et al. (35). However, given the challenges inherent in quantitative breast DWI due to motion, off-isocenter effects, and SNR considerations, as well as the analysis challenge of consistent and optimized ROI definition, the variability in breast ADC measures is a major concern. These challenges are magnified in a typical multi-site clinical trial setting due to the wide variety of scanner platforms in use and the large number of sequence variants optimized for the different platforms. In this study, we evaluated the repeatability and reproducibility of breast tumor ADC in a multi-center setting, under the conditions of a standardized protocol and centralized quality-control and analysis. Excellent agreement in repeat measures of tumor ADC was found. Repeatability did not appear to be significantly influenced by field strength or treatment time-point. Inter- and intra-operator reproducibility were also good, even though whole-tumor ROI reproducibility varied between studies. Therefore, our findings support the use of ADC values as a reproducible, functional metric for tumor response in this setting.

Multiple studies have been conducted on breast DWI reproducibility, including studies on phantoms, normal volunteers, and patients. While not directly comparable to *in vivo* studies, phantoms with known ADC values can

help establish minimum variability limits for DWI techniques. For example, using ice/water phantoms (ADC = 1.1 10^{-3} mm^2/s), Malyarenko et al. (36) demonstrated short-term test/retest repeatability wCV < 0.5% and day-to-day wCV < 2.2%, across a range of vendors and field strengths. With a similar phantom, Sorace et al. (37) reported < 0.01% average difference in ADC between repeated scans with intra-scan repositioning of the phantom for three instances of a single scanner platform. These and other phantom results generally indicate that the scanner-dependent variability in ADC measurement of uniform media is small compared to *in vivo* effects and ROI delineation within non uniform tissues.

For single-site test/retest repeatability in normal volunteers, Aliu et al. (14) found a wCV of 11% (N=9); Partridge et al. (18) showed a 4.5% wCV (N=9) for ADC derived from diffusion tensor imaging; and Sorace et al. (37) found a wCV = 3.72% (N=10). Despite the differences in methodology, these are all comparable to our findings. The higher wCV in Aliu et al. is perhaps attributable to increased time between repeat scans, which were done either within 72 hours or at the same menstrual phase in successive periods. Sorace's study also included multi-site/single-platform reproducibility on three subjects, each scanned at three different sites. The wCV for these repeat measures was 6.3%.

Patient studies of reproducibility of breast tumor ADC (15,16,19,20) have shown a range of intra- and inter-observer variability values, with wCV ranging

from 3.2% to 8.3% and ICC values >0.9 in most cases. Our results are of comparable magnitude to observer variability in these reader studies. In a retrospective single-site study of 54 patients Giannotti et al. (15) reported excellent ICC values for intra-observer variations for clinical and technical readers ranging from 0.92 to 0.99, with some dependence on ROI techniques, tumor environment, and tumor size. Inter-observer ICC values were lower however, 0.86 for a large single-slice ROI and 0.68 for a small ROI measuring minimum ADC. In a diagnostic study measuring ADC values of suspicious breast lesions (Breast Imaging Reporting and Data System [BIRADS] category 4 or 5) in 40 patients with repeat scans 1-11 days apart, Spick et al (20) found wCV of 6% and 6.6% for 2 different readers, with ICC values >0.9. We note that all these previous studies were single-platform with, except for Giannoti's work, a relatively small numbers of patients. The ACRIN 6698 study expanded this to a true clinical trial setting. Our overall repeatability values wCV=4.8%, ICC=0.97, and RC=0.16×10^{-3} mm^2/s indicate that performance similar to single-site studies is attainable in the multi-site trial setting, at least under the conditions of a standardized protocol and centralized QA and analysis. We found this true across multiple field strengths and despite the challenges of protocol standardization across multiple vendor platforms.

Previous studies in patients undergoing NAC have reported significant changes in tumor ADC with treatment response, with effect sizes generally large compared to our repeatability value of wCV=4.8%. Responders exhibited tumor ADC increases ranging from 9% to over 60% in the prior studies, depending on treatment timepoint (3-6,38).

Perhaps the most significant factor currently limiting reproducibility of breast tumor ADC measurements is operator variability in the ROI definition. For this study, we evaluated whole-tumor mean ADC as specified in the study protocol, which required multi-contour/multi-slice three dimensional ROIs. While offering a potential advantage of sampling the entire diseased region, these ROIs are also the most difficult and time consuming to define. The high values we found for intra- and inter-reader reproducibility and the similarity between these metrics, indicate that the analysis techniques are both reproducible and transferable with suitable training. However, the wide range of the Dice similarity between ROIs across the patient cohort may indicate that the excellent reproducibility is a reflection of the robustness of our whole-tumor ADC analysis protocol in the presence of ROI variations. The lack of significant correlation between ROI similarity and ADC variability in our data further indicates that tumor characteristics (e.g., homogeneity, morphology) play a more important role in

determining ADC reproducibility on an individual case than do the details of the ROI definition.

There is not currently evidence to our knowledge that the whole-tumor ADC is optimum for treatment monitoring. In particular, in a single site study of 150 patients in a diagnostic setting, Bickel et al. (39) found that small, single slice, 2D ROIs identifying a region of minimal ADC gave better performance than either larger single-slice 2D ROIs or multi-slice 3D ROIs. Further work is clearly needed in evaluating both different ROI techniques and automatic or semi-automatic segmentation techniques for optimizing any DWI treatment response metrics.

Acquisition of consistently high-quality breast DWI continues to be a challenge. The overall rejection rate of DWI acquisitions in this study due to image quality issues was 13%, representing cases where the operator felt it was not possible to confidently define an ROI for the whole tumor volume. Failed or inadequate fat suppression and excessive artifacts were the primary causes for many of these rejections. In their larger single-site diagnostic study, Bickel et al. (39) reported a loss of only 3 out of 150 studies (2%) due to technical failures of the DWI; however, they excluded cases with no MR visible lesion. Our higher rejection rate may be in part due to the compromises inherent in standardizing a protocol across multiple scanner platforms. It is possible that scanner-optimized

protocols may yield better results for treatment response monitoring in multi-site clinical trials, assuming all longitudinal scans on a given patient are restricted to a single platform. Ongoing investigations in the broader I-SPY 2 cohort may shed some light on these QA issues. While our prospective study was not powered to detect differences between scanner configurations, we did observe a somewhat lower rejection rate for 3T scanners compared with 1.5T, indicating a possible quality advantage for the higher field strength. Quality controlled studies with larger cohorts are needed to establish whether this is a valid finding.

This study was limited to exploring the variability in whole-tumor ADC due to patient repositioning, fundamental variabilities in the DWI acquisition and analysis, and reader variability. Further investigation is required to assess variability due to other factors, most particularly ROI definition protocols, and of other DWI metrics. We were also limited by the predominance of pre-treatment studies. This prohibited any substantive determination of differences in reproducibility as treatment progresses, which may be critical for treatment response determination. This may be of particular concern for measurements at time points late in the course of treatment, since experience in the ACRIN 6698 trial showed that it becomes increasingly difficult to confidently define a whole-tumor ROI after significant response to NAC. Finally, we were limited in this study to the case of centralized analysis with single-platform generated ADC maps.

On-site analysis and the use of ADC maps from multiple platforms (e.g., on-scanner parametric map generation) may further affect ADC repeatability (21,40).

In conclusion, within the prospective multi-center ACRIN 6698 trial, we found excellent reproducibility of whole-tumor ADC – with test/retest repeatability and operator variability both small compared to typical treatment induced changes and consistent across different field strengths. To facilitate use of DWI for monitoring therapy, further work is needed to improve MR-DWI image quality in the breast to reduce losses, particularly with respect to fat suppression and image artifacts.

ACKNOWLEDGEMENTS:

The authors acknowledge those individuals who have contributed substantially to the work reported in the manuscript, including the ACRIN 6698 and I-SPY 2 Trial Teams, the patients who participated in the study, and the staff members who contributed to the conduct of the study at the University of California, San Francisco; University of Minnesota; University of Pennsylvania; University of Washington; University of Alabama, Birmingham; University of California, San Diego; University of Texas MD Anderson Cancer Center; Oregon Health Sciences University; University of Chicago; and H. Lee Moffitt Cancer Center and Research Institute.

REFERENCES

1. Hylton NM, Blume JD, Bernreuter WK, et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy--results from ACRIN 6657/I-SPY TRIAL. Radiology 2012;263(3):663-672.

2. Hylton NM, Gatsonis CA, Rosen MA, et al. 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(1):44-55.

3. Partridge SC, Zheng Z, Newitt DC, et al. Diffusion-Weighted MRI Predicts Pathologic Response in Neoadjuvant Treatment of Breast Cancer: the ACRIN 6698 Multicenter Trial. Radiology 2018;(in press).

4. Galban CJ, Ma B, Malyarenko D, et al. Multi-site clinical evaluation of DW-MRI as a treatment response metric for breast cancer patients undergoing neoadjuvant chemotherapy. PloS one 2015;10(3):e0122151.

5. Li XR, Cheng LQ, Liu M, et al. DW-MRI ADC values can predict treatment response in patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. Medical oncology 2012;29(2):425-431.

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 in biomedicine 2009;22(1):104-113.

7. Manton DJ, Chaturvedi A, Hubbard A, et al. Neoadjuvant chemotherapy in breast cancer: early response prediction with quantitative MR imaging and spectroscopy. British journal of cancer 2006;94(3):427-435.

8. Nilsen L, Fangberget A, Geier O, Olsen DR, Seierstad T. Diffusion-weighted magnetic resonance imaging for pretreatment prediction and monitoring of treatment response of patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy. Acta oncologica 2010;49(3):354-360.

9. Thakur SB, Durando M, Milans S, et al. Apparent diffusion coefficient in estrogen receptor-positive and lymph node-negative invasive breast cancers at 3.0T DW-MRI: A potential predictor for an oncotype Dx test recurrence score. Journal of Magnetic Resonance Imaging 2018;47(2):401-409.

10. Bufi E, Belli P, Costantini M, et al. Role of the Apparent Diffusion Coefficient in the Prediction of Response to Neoadjuvant Chemotherapy in

Patients With Locally Advanced Breast Cancer. Clin Breast Cancer 2015;15(5):370-380.

11. Liu S, Ren R, Chen Z, et al. Diffusion-weighted imaging in assessing pathological response of tumor in breast cancer subtype to neoadjuvant chemotherapy. J Magn Reson Imaging 2015;42(3):779-787.

12. Richard R, Thomassin I, Chapellier M, et al. Diffusion-weighted MRI in pretreatment prediction of response to neoadjuvant chemotherapy in patients with breast cancer. European radiology 2013;23(9):2420-2431.

13. Chenevert TL, Stegman LD, Taylor JM, et al. Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors. Journal of the National Cancer Institute 2000;92(24):2029-2036.

14. Aliu SO, Jones EF, Azziz A, et al. Repeatability of quantitative MRI measurements in normal breast tissue. Transl Oncol 2014;7(1):130-137.

15. 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(9):1729-1736.

16. Jang M, Kim SM, Yun BL, et al. Reproducibility of Apparent Diffusion Coefficient Measurements in Malignant Breast Masses. Journal of Korean medical science 2015;30(11):1689-1697.

17. 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. European radiology 2012;22(7):1512-1518.

18. 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(3):320-328.

19. Petralia G, Bonello L, Summers P, et al. Intraobserver and interobserver variability in the calculation of apparent diffusion coefficient (ADC) from diffusion-weighted magnetic resonance imaging (DW-MRI) of breast tumours. La Radiologia medica 2011;116(3):466-476.

20. Spick C, Bickel H, Pinker K, et al. Diffusion-weighted MRI of breast lesions: a prospective clinical investigation of the quantitative imaging biomarker characteristics of reproducibility, repeatability, and diagnostic accuracy. NMR in biomedicine 2016;29(10):1445-1453.

21. 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. European radiology 2016;26(7):2291-2296.

22. 2017 QIBA Profile for DWI http://qibawiki.rsna.org/images/1/1d/QIBADWIProfilev1.45_20170427_v5_accepted.pdf. Last accessed July 18, 2018

23. 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. http://www.acrin.org/Portals/0/Protocols/6698/Protocol-ACRIN6698_v2.29.12_active_ForOnline.pdf. Last accessed February 2, 2018

24. 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(1):97-100.

25. Chenevert TL, Galban CJ, Ivancevic MK, et al. Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. J Magn Reson Imaging 2011;34(4):983-987.

26. I-SPY 2: Neoadjuvant and Personalized Adaptive Novel Agents to Treat Breast Cancer. https://clinicaltrials.gov/ct2/show/NCT01042379. Last accessed February 2, 2018

27. Aliu SO, Newitt DC, Li W, et al. Quality assessment and ranking system for quantitative breast diffusion-weighted imaging of the breast in the ACRIN 6698 trial. Proceedings of the 23rd Annual Meeting of ISMRM, Toronto, Canada 2015. (abstract 1311).

28. Stejskal EO, Tanner JE. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys 1965;42:288-292.

29. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26(3):297-302.

30. Quan H, Shih WJ. Assessing reproducibility by the within-subject coefficient of variation with random effects models. Biometrics 1996;52(4):1195-1203.

31. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. Journal of biopharmaceutical statistics 2007;17(4):529-569.

32. Zhang Z, Wang Y, Duan F. An AUC-like index for agreement assessment. Journal of biopharmaceutical statistics 2014;24(4):893-907.

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(4):231-235.

34. Hylton NM, Blume JD, Bernreuter WK, et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy--results from ACRIN 6657/I-SPY TRIAL. Radiology 2012;263(3):663-672.

35. 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(2):337-355.

36. Malyarenko D, Galban CJ, Londy FJ, et al. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. J Magn Reson Imaging 2013;37(5):1238-1246.

37. Sorace AG, Wu C, Barnes SL, et al. Repeatability, reproducibility, and accuracy of quantitative mri of the breast in the community radiology setting. J Magn Reson Imaging 2018.

38. Iacconi C, Giannelli M, Marini C, et al. The role of mean diffusivity (MD) as a predictive index of the response to chemotherapy in locally advanced breast cancer: a preliminary study. European radiology 2010;20(2):303-308.

39. Bickel H, Pinker K, Polanec S, et al. Diffusion-weighted imaging of breast lesions: Region-of-interest placement and different ADC parameters influence apparent diffusion coefficient values. European radiology 2017;27(5):1883-1892.

40. Newitt DC, Malyarenko D, Chenevert TL, et al. Multi-site Concordance of DWI Metrics: Results of the NCI Quantitative Imaging Network ADC Mapping Collaborative Project. International Society of Magnetic Resonance in Medicine; 2017.

TABLES

Table 1: ACRIN 6698 DW-MRI protocol parameters and protocol adherence for N=89 test/retest cohort

Parameter	DWI Protocol Specification	Test/retest cohort values (N)	N_{acq} out of protocol [a]
Acquisition order	DWI Pre-contrast injection	compliant (88), mixed (1) [b]	1
Sequence type	SS-EPI	all compliant	
Frequency direction	R/L [c]	R/L (60), A/P (10), mixed (1) [d]	
FOV (mm)	260 - 360	280 – 390	4
Matrix – frequency	128-192	128 – 196	2
Matrix – phase	128-192	128 – 194	2
Fat-suppression	Active fat-sat	all compliant	
TR (ms)	≥4,000	4800 – 9400	
TE (ms)	minimum TE (50-100)	59.2 – 114.0	5
Flip Angle	90 degrees	90	
B values	0, 100, 600, 800 s/mm^2	all compliant	
Slice thickness (mm)	4 – 5	4 (44), 5 (27)	
Number of slices	Complete bilateral coverage	24 - 50	
Slice Gap (mm)	≤ 1.0	0	
No. averages	≥2	2 – 5	
Parallel imaging factor	≥2	N/C	
Total acquisition time	≤ 5 minutes	N/C	

SS-EPI: Single-shot echo-planar imaging, N/C: Data not collected

[a] Number of acquisitions (test or retest) with out-of-protocol values.

[b] One study was done with the retest DWI acquisition following the contrast injection.

[c] Protocol requirement of R/L frequency encoding was relaxed after investigation indicated better image quality with A/P frequency on some scanners.

[d] One study was done with reversed phase/frequency encoding directions between test and retest scans.

Table 2. Accrual and analyzability results for test/retest studies

		N accrued	N analyzable	% Lost	N Reader study	Field strength	Vendor
	All [a]	89	71	20.2	20		
Site	Site A	5	3	40.0	0	1.5	Siemens
	Site B	11	9	18.2	3	3	GEHC
	Site C	15	14	6.7	4	3	Philips
	Site D1	7	5	28.6	2	1.5	Siemens
	Site D2	5	4	20.0	1	3	Siemens
	Site E [b]	11	7	36.4	2	1.5	GEHC
	Site F	12	11	8.3	3	1.5	Philips
	Site G	8	6	25.0	2	1.5	Philips
	Site H	14	12	14.3	3	1.5	GEHC
	Site I	1	0	100.0	0	3	GEHC
Study Visit [c]	T0	75	60	20.0	17		
	T1	14	11	21.4	3		
Field Strength	1.5T	57	44	22.8	12		
	3.0T	32	27	15.6	8		
Vendor	GEHC	37	28	24.3	8		
	Philips	35	31	11.4	9		

Siemens	17	12	29.4	3

[a] All scans were done on qualified scanners with dedicated bilateral breast coils

[b] Site E studies were done with three separate two b-value acquisitions (0/100; 0/600; 0/800 s/mm^2). All others sites used a single four b-value acquisition.

[c] T0 (baseline, pre-treatment) was specified as the preferred visit for test/retest studies

Table 3: Summary of agreement values and inter- and intra-operator variability for all cohorts for mean tumor ADC.

		RC (10^{-3} mm^2/sec)			wCV			ICC			AI		
Cohort	n	value	95% CI		value	95% CI		value	95% CI		value	95% CI	
All	71	0.16	0.13	0.19	4.8%	4.0%	5.7%	0.97	0.95	0.98	0.83	0.76	0.87
1.5T	44	0.15	0.13	0.20	4.7%	3.6%	5.8%	0.98	0.96	0.99	0.83	0.74	0.87
3.0T	27	0.16	0.12	0.21	5.1%	3.7%	6.5%	0.95	0.89	0.98	0.81	0.68	0.86
T0	60	0.16	0.13	0.19	5.2%	4.2%	6.1%	0.91	0.85	0.94	0.81	0.73	0.85
T1	11	0.16	0.12	0.28	3.6%	2.0%	5.3%	0.99	0.96	1.00	0.83	0.60	0.84
inter	20	0.18	0.14	0.26	5.4%	3.7%	7.1%	0.92	0.80	0.97	0.67	0.44	0.78
intra	20	0.17	0.13	0.25	5.6%	3.8%	7.4%	0.91	0.78	0.96	0.78	0.58	0.84

RC: repeatability coefficient; wCV: within-subject coefficient of variation; ICC: intraclass correlation coefficient;

AI: agreement index; inter: inter-operator variability; intra: intra-operator variability

FIGURE LEGENDS

Figure 1. I-SPY 2 TRIAL study schema. Patients are randomized onto the control arm (paclitaxel treatment) or onto one of the multiple experimental drug treatment arms. DCE MRI is conducted at 4 timepoints to monitor treatment response via changes in MRI functional tumor volume. For the ACRIN 6698 Trial, an advanced four b-value DWI sequence was added at each of the MRI timepoints T0-T3, and test/retest repeatability scans were done on a subset of patients at T0 or T1. (We note that the T0-T3 nomenclature is used in this paper for consistency with I-SPY 2, and differs from that defined in the ACRIN 6698 protocol.)

Figure 2. Sample images of representative slices for two subjects. The left panels show DCE subtraction images used for tumor localization. The corresponding ADC map images and ROIs are shown for the test DWI (center panels) and retest DWI (right panels). Subject 1 had relatively poor repeatability ($|ADC_{difference}|/ADC_{mean}$=0.117), while subject 2 had good repeatability ($|ADC_{difference}|/ADC_{mean}$=0.025). Each inset shows the mean whole-tumor ADC, the ROI volume segmented on the DWI, and the extent in the slice direction of the multi-slice ROI.

Figure 3. Bland-Altman plots for the difference in whole-tumor ADC measurements between test and retest for: (a) entire analyzable cohort, (b) 1.5

tesla cases, and (c) 3 tesla cases. Blue diamonds indicate the 60 measurements taken at timepoint T0 (pre-treatment) and red circles indicate the 11 measurements taken at timepoint T1 (early treatment). Mean difference and 95% CI (1.96 * SD) are shown as horizontal solid red and dashed black lines respectively. No trend was seen in ADC difference with ADC magnitude.

Figure 4. Repeatability measures intra-class correlation coefficient (ICC, blue) and agreement index (AI, red) for full cohort and sub groups defined by field strength (1.5 and 3.0 tesla) and study visit (T0: pre-treatment and T1: early-treatment). Error bars indicate 95% confidence intervals.

Figure 5. Bland-Altman plots showing inter-reader (a) and intra-reader (b) variability in whole-tumor ADC measures. Plotted is the difference in the two independent reader ADC evaluations versus the mean ADC. Mean difference and 95% CI (1.96 * SD) are shown as horizontal solid red and dashed black lines respectively. Black diamonds indicate 1.5T scans and blue squares indicate 3T scans. Plots show minimal bias and no trends in ADC difference with increasing ADC values.

Figure 6. ROI reproducibility results. a) Sørenson-Dice similarity coefficient (DSC) for the inter-reader ($DSC_{[RD1,RD2]}$, X-axis) and intra-reader ($DSC_{[RD1,RD1B]}$, Y-axis) ROI pairs for each case. Inter- and intra- values were highly correlated: Pearson's r=0.84 ($p<0.0001$, 95% CI 0.63, 0.94). b) Mean DSC

(DSC_{Mn} = ($DSC_{[RD1,RD2]}$+ $DSC_{[RD1B,RD2]}$+ $DSC_{[RD1,RD1B]}$)/3) vs. mean whole-tumor ADC for each study. A trend of decreasing DSC_{Mn} with increasing ADC was observed (Pearson's r = -0.48 [p=0.03, 95% CI -0.77, -0.06]). Correlations were somewhat driven by the high-ADC outlier. Analysis without the outlier resulted in r=0.89 ($p<10^{-6}$) for inter-DSC vs. intra-DSC and r=-0.59 (p=0.008) for DSC_{Mn} vs. mean ADC. c) DSC_{Mn} vs. ADC variability (standard deviation of the mean ADC values for three reader study observations). No significant correlation was observed (r=-0.37, p=0.14).