Boss Michael A (Orcid ID: 0000-0002-9492-767X)
Choi Haesun (Orcid ID: 0000-0002-3565-5475)

Repeatability and Reproducibility Assessment of the Apparent Diffusion Coefficient in the Prostate: A Trial of the ECOG-ACRIN Research Group (ACRIN 6701)

Michael A. Boss, PhD[1†], Bradley S. Snyder, MS[2], Eunhee Kim, PhD[2*], Dena Flamini, RT[1], Sarah Englander, PhD[3], Karthik M. Sundaram, MD[3], Naveen Gumpeni, MD[4], Suzanne L. Palmer, MD[5], Haesun Choi, MD[6], Adam T. Froemming, MD[7], Thorsten Persigehl, MD[8], Matthew S. Davenport, MD[9], Dariya Malyarenko, PhD[9], Thomas L. Chenevert, PhD[9], Mark A. Rosen MD, PhD[3]

[1]Center for Research and Innovation, American College of Radiology Philadelphia, Pennsylvania, USA; [2]Department of Biostatistics, Brown University, Providence, Rhode Island, USA; [3]Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania; [4]Department of Radiology, Weill Cornell Medical Center, New York, New York, USA; [5]Department of Radiology, University of Southern California, Los Angeles, California, USA; [6]Department of Abdominal Imaging, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA; [7]Department of Radiology, May Clinic, Rochester, Minnesota, USA; [8]Department of Radiology, University Hospital Cologne, Cologne, Germany; [9]Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA

[†]Corresponding author, email: mboss@acr.org

[*]Presently at Merck Research Laboratories, Kenilworth, New Jersey, USA

Acknowledgements

This project has been funded in part with Federal funds from the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN268201000050C, and funding from the National Cancer Institute, under Award No. U24CA180803.

This study was coordinated by 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: U10CA180820, U10CA180794. 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 U.S. government.

Professor Edward F. Jackson, PhD played an instrumental role in ACRIN 6701 as study co-chair, and a leader in QIBA, including time as Chair. We are grateful for his many contributions, sage advice, and unending enthusiasm and patience in advancing standards for quantitative imaging. We endeavor to continue his mission and realize his vision.

Running Title: ACRIN 6701 Prostate ADC Reproducibility

Repeatability and Reproducibility Assessment of the Apparent Diffusion Coefficient in the Prostate: A Trial of the ECOG-ACRIN Research Group (ACRIN 6701)

Abstract

BACKGROUND:

Uncertainty regarding the reproducibility of the apparent diffusion coefficient (ADC) hampers the use of quantitative diffusion-weighted imaging (DWI) in evaluation of the prostate with MRI. The Quantitative Imaging Biomarker Alliance (QIBA) Profile for quantitative DWI claims a within-subject coefficient of variation (wCV) for prostate lesion ADC of 0.17. Improved understanding of ADC reproducibility would aid the use of quantitative diffusion in prostate MRI evaluation.

PURPOSE:

Evaluation of the repeatability (same-day) and reproducibility (multi-day) of whole-prostate and focal-lesion ADC assessment in a multi-site setting.

STUDY TYPE:

Prospective multi-institutional.

SUBJECTS:

29 males, ages 53–80 (median 63), following diagnosis of prostate cancer, 10 with focal lesions.

FIELD STRENGTH/SEQEUNCE:

3T, single-shot spin-echo diffusion-weighted echo-planar sequence with four *b*-values.

ASSESSMENT:

Sites qualified for the study using an ice-water phantom with known ADC. Readers performed DWI analyses at visit 1 (“V1”) and visit 2 (“V2”, 2–14 days after V1), where V2 comprised scans before (“V2pre”) and after (“V2post”) a “coffee-break” interval with subject removal and repositioning. A single reader segmented the whole prostate. Two readers separately placed region-of-interests for focal lesions.

STATISTICAL TESTS:

Reproducibility and repeatability coefficients for whole prostate and focal lesions derived from median pixel ADC. We estimated the wCV and 95% confidence interval using a variance stabilizing transformation and assessed interreader reliability of focal lesion ADC using the intraclass correlation coefficient (ICC).

RESULTS:

The ADC biases from b_0–b_{600} and b_0–b_{800} phantom scans averaged 1.32% and 1.44%, respectively; mean *b*-value dependence was 0.188%. Repeatability and reproducibility of whole prostate median pixel ADC both yielded wCVs of 0.033 (N=29). In ten subjects with an evaluable focal lesion the individual reader wCVs were 0.148 and 0.074 (repeatability) and 0.137 and 0.078 (reproducibility). All time points demonstrated good to excellent interreader reliability for focal lesion ADC (ICC_{V1}=0.89; ICC_{V2pre}=0.76; ICC_{V2post}=0.94).

DATA CONCLUSION:

This study met the QIBA claim for prostate ADC. Test-retest repeatability and multi-day reproducibility were largely equivalent. Interreader reliability for focal lesion ADC was high across time points.

Introduction

Diffusion-weighted magnetic resonance imaging (DWI) is a technique that incorporates the use of position-encoding magnetic field gradients to weight the magnetic resonance signal intensity by the self-diffusion due to Brownian motion of hydrogen species, primarily from water molecules.[1] The resulting signal attenuation generates diffusion-weighted contrast and can provide insight into tissue microstructure. This process is well-described mathematically and allows for calculation of a quantitative imaging biomarker (QIB), the apparent diffusion coefficient (ADC), on a voxel-by-voxel basis, generating quantitative parametric maps of diffusion.[2] These ADC maps facilitate comparison of exams across subjects, time, vendors, and sites.[1-3]

Diffusion-weighted imaging is an especially important part of the diagnostic evaluation for prostate cancer by MRI and is part of the multiparametric MRI exam of the prostate, also including anatomic T_2-weighted imaging and dynamic–contrast enhanced imaging. The application of these MRI techniques for prostate evaluation are ensconced in the PI-RADS (Prostate Imaging Reporting and Data System) guidelines, first proposed in 2012, and revised as PI-RADS v2.0 in 2017 and PI-RADS v2.1 in 2019.[4,5] In these guidelines, imagers use DWI qualitatively to assess for the presence of significant cancer in the prostate, with “restriction” defined in visual terms as the presence of hyper- and hypointense areas on the high *b*-value DW image and ADC maps, respectively.

The current PI-RADs v2.1 document does not endorse the use of quantitative analysis of ADC maps, recommending qualitative visual assessment, noting that “ADC calculations are influenced by choice of *b*-values and have been inconsistent across vendors”.[6] As such, more work is required to better understand sources of variability in ADC measures; otherwise there will remain limits on the use of quantitative criteria for defining the probability of cancer based on DWI ADC maps.

Elucidating the sources of variability and establishing a baseline for ADC reproducibility can establish the best use of quantitative DWI in evaluation of prostate cancer. The Quantitative Imaging Biomarkers Alliance (QIBA) seeks to harmonize quantitative imaging approaches by creating Profile documents that allow for technical assessment of quantitative imaging methods, e.g., the QIBA DWI Profile.[7] QIBA aims to achieve harmonization through a technical claim statement for a QIB, which provides bounds on its reproducibility. These claims are met by following the imaging protocols and assessment methodology which the Profile prescribes in detail. QIBA claims derive from test-retest literature for QIBs; however, there is a dearth of appropriate test-retest literature for ADC in the prostate.

The main goal of ACRIN 6701 was to assess the reproducibility and repeatability of quantitative MRI metrics in the prostate by means of test-retest imaging. Reproducibility in ACRIN 6701 refers to the stability of these metrics across time (2–14 days) on the same subjects imaged on the same scanners; repeatability refers to an immediate re-positioning and re-scanning of the same subjects on the same scanners by the same operators.[8,9] The primary aims were assessment of K^{trans} and ADC in the whole prostate; the secondary aims were assessment of the same QIBs in the focal lesion, and a comparison of test-retest of T_1-dependent and T_1-independent DCE-MRI models. Exploratory aims were to correlate DCE and DWI metrics in whole prostate and tumor, and to compare the repeatability of ADC from a test-retest experiment conducted on the same day with the reproducibility of ADC across different days of scanning the same subject. This report details results of DWI analysis. DCE results will be reported in a future publication.

Materials and Methods

ACRIN 6701, activated in August 2012 (subjects enrolled 10Jan2014–12Jan2016), was HIPAA-compliant and approved by the individual site Institutional Review Boards. All subjects gave written informed consent prior to enrolling. All centers invited to participate had an active prostate MR program (> 50 cases per year). Participating centers designated one or more 3T MRI scanners for the study from one of three major MRI vendors at the time of site initiation (Siemens, Philips, or GE), provided a site body MR radiologist (to oversee study implementation), and performed phantom and subject DWI scans.

Site Qualification

ACRIN 6701 employed 8 scanners across 7 sites. Scanners at 3 Tesla from GE (3 Discovery MR750 and 1 Signa HDxt), Philips (1 Achieva and 1 Ingenia) and Siemens (1 Skyra and 1 Trio) imaged the subjects.

Prior to site activation, each scanner scanned and analyzed a DWI ice-phantom with known diffusivity (1.1 x 10^{-3} mm^2/s) at 0 °C.[10,11] Phantom scanning derived from the ACRIN 6698 Breast DWI study procedures, using a single, equilibrated ice-water phantom vertically in the scanner, with an appropriate loading coil and a torso-array coil.[12] We performed single-volume imaging in the FOV center, employing a single-shot spin-echo diffusion-weighted echo planar imaging sequence (320 mm x 260 mm, 160 x 128, bandwidth = 1500–2500 Hz, repetition time (TR) ≥ 8 s, echo time (TE) = 75–100 ms), utilizing *b*-values of 0, 100, 600, and 800 s/mm^2.

The ACR Imaging Core Laboratory, in collaboration with partnering MRI physics teams (at the University of Wisconsin-Madison and the University of Michigan), performed centralized quality assurance (QA) and quality control (QC) evaluation of each DWI exam as previously described.[10] Sites passed DWI qualification if they met the qualitative and quantitative requirements, including adequate SNR (> 75 at b_{800}), ADC bias < 5%, and ADC *b*-value dependence < 2%. Centralized QA/QC also confirmed compliance with the trial specific DWI acquisition parameters using information in the DICOM headers.[10,13] The core lab and physics teams documented quantitative performance evaluation results in a QC report and acceptance letter; sites obtained feedback from the core lab via email. If the qualifying exams did not obtain approval, feedback included required corrections for rescanning. ACRIN 6701 required requalification with a repeat phantom scan 1) for any new scanner introduced to the study, 2) after any major changes to scanner hardware or software (minor software updates excluded), or 3) failure to complete all subject scanning within one year of initial scanner qualification.

Subject Population

Eligible subjects for the study were males over the age of 18 with prostate cancer diagnosed by transrectal ultrasound-guided biopsy 28–90 days before enrollment and able to tolerate the MR imaging protocol. The minimum tumor burden for inclusion was at least one of the following criteria: 1) a single core with ≥ 50% cancer burden and ≥ 5 mm tumor length; 2) two or more cores in the same prostate region, each with ≥ 30% cancer burden; 3) three or more cores positive for prostate cancer (of any magnitude of cancer burden) in the same prostate region; 4) Gleason score of 7 or higher cancer burden; or 5) prostate-specific antigen level (PSA) ≥ 10 ng/mL.

The study excluded subjects unable to tolerate MRI due to claustrophobia, incompatible metallic implants, excessive weight for the MR table, or who could not receive gadoliniumbased contrast agent due to poor renal function (globular filtration rate, GFR < 30 mL/min/1.73 m^2 based on a serum creatinine level obtained within 48 hours prior to enrollment). We also did not allow enrollment by subjects with prior anti-androgen therapy within the past 30 days; previous external beam radiation, proton beam, or brachytherapy to the prostate; or prior major pelvic surgery, including hip replacement.

Each qualified site was limited to an initial enrollment of up to five subjects. This capping of enrollment ensured that there was adequate representation from at least two different imaging centers for each of the represented vendors. If any enrolled subject was disqualified from analysis, the site was allowed one additional enrollment slot.

MRI Scanning Protocol

The ACRIN 6701 study required MRI scanning to begin within 28 days of enrollment and comprised two separate MRI visits, separated by 2 to 14 calendar days. The imaging requirements for each visit are shown in Figure 1.

The study used torso array coils for the DWI elements of the study. Subjects lay on the scanner bed in a supine position with the torso-array receive coil applied to the pelvis.

On visit 1 (V1, 0 to 28 calendar days after enrollment), sites acquired sub-mm axial and coronal T_2-weighted (T2w) fast-spin echo (FSE) images, optionally oblique to the angle of the prostate. The DWI protocol consisted of a 2D spin-echo echo-planar imaging (SE-EPI) sequence with fat suppression, axial orientation, TE < 90 ms, TR > 4 s, *b*-values = 0, 100, 600, 800 s/mm^2, field-of-view 320–400 x 220–360 mm, acquisition matrix 128–192 x 128–192, 5 mm slices, with 3–6 averages per *b*-value. Sites were able to perform additional DWI to meet clinical requirements.

During visit 2 (V2, 2 to 14 calendar days after V1), a localizing axial single-shot FSE guided further anatomic prescription. Sites then performed two study-specific DWI series, identical to those from V1, but separated by a “coffee break” (~5 minutes between V2pre and V2post), during which site personnel removed and disconnected the torso array coil from the scanner. The subject exited the scanner briefly and then repositioned, with consistent landmarking. The site then relocalized the subject in the axial plane and performed a post-coffee-break DWI scan.

At each time point, the ACRIN 6701 imaging protocol included T2w and SE-EPI diffusion-weighted imaging. All imaging occurred prior to surgery, radiation, or hormonal therapy.

MRI Scan QA

Subject images underwent local site and central quality assurance. Each study was assessed locally for proper subject positioning, coil placement, and anatomic coverage, and adherence to the protocol-specific parameters. Sites could conduct repeat visits for any study requirements that were deemed unsuitable.

The ACRIN Core Lab performed a central analysis. A dedicated technologist (DF, 15 years of experience in MRI analysis) assessed each exam for missing images/sequences, appropriate image anonymization, complete anatomical coverage of the prostate, and adherence of all sequences to imaging protocol. The lead body radiologist (MAR, 13 years of experience in prostate MRI analysis) then performed a dedicated review. This analysis confirmed any technical deficiencies/variations noted by the core lab technologist and assessed for any artifacts that might interfere with quantitative DWI analysis. Failure to adhere to prescribed *b*-values and deviation from the prescribed FOV, matrix, or slice thickness by more than 30% were considered major variances. These major variances and/or severe artifacts, such as significant distortion due to rectal gas, led to disqualification of that DWI series from further analysis. All decisions regarding imaging suitability were performed prior to image segmentation and tumor analysis.

When QA disqualified a subject for analysis for major technical violations or severe image artifact, the core lab notified the site and invited it to offer a repeat visit to the enrolled subject. If the enrolled subject could not return for additional imaging within the study window, sites were invited to enroll up to one additional subject.

Tumor Identification, segmentation, and analysis

We imported the DWI data into in-house software for viewing, segmentation, and analysis (IDL©, L3 Harris Geospatial, Broomfield, CO) similar to the ACRIN 6698 analysis protocol and calculated ADC maps with a mono-exponential decay model with linear least-squares fits of the log of the signal intensities at three *b*-values (100, 600, and 800 s/mm^2), [2,14]

$$S(b) = S_0 e^{-b \cdot ADC} \tag{1}$$

where $S(b)$ is the signal intensity with diffusion-weighting of b and S_0 is the signal intensity with zero diffusion-weighting.[15] We created a calculated b_{1400} diffusion-weighted image from the ADC curve fit on a voxel-by-voxel basis, using the line fit from the 3 *b*-values to extrapolate the signal intensity at b=1400 s/mm^2.

The lead body MR radiologist for the study reviewed the multi-parametric imaging (including T2w-FSE, DWI, and vendor-provided ADC maps) on a dedicated prostate MRI analysis platform (DynaCAD©, iCAD Inc., Nashua, NH) to determine the presence of restricting focal lesion in the peripheral zone with minimum bi-dimensional axes of 6 mm x 6 mm and with the area of diffusion restriction spanning at least two adjacent slices. If there was no such lesion, the lead reader instead chose a distinct benign prostatic hyperplasia (BPH) lesion demonstrating restriction, if present. Remaining subjects were designated as "no focal lesion".

For subject DWI analysis, the lead reader manually segmented the whole prostate on the b_{600} images in all cases; this reader similarly segmented focal lesions, when present, on the calculated b_{1400} images. ADC maps were available as a reference. However, the reader avoided direct segmentation on the ADC maps to minimize potential bias.

To evaluate the robustness of the determination of lesion ADC repeatability and reproducibility estimates a second central reader (KMS, 2 years of experience in prostate MRI) repeated the lesion segmentation in the same manner as the initial central reader. This reader was provided the identities of the cases where a focal lesion was identified by the lead reader, and the center of this lesion was indicated in a single axial T_2-weighted image. The second reader had access to all images but did not have access to the segmentation boundaries from the lead reader.

ADC values for both whole prostate and tumor were reported as the voxel medians. ROI volumes (prostate and focal lesion) were calculated as voxel sizes multiplied by number of voxels in the ROI.

Statistical Methods

We assessed the test-retest performance of DWI ADC under reproducibility conditions using the V1 and V2pre scans, where the two scans were separated between 2 and 14 calendar days.[16] We produced Bland-Altman plots and estimated both the reproducibility coefficient (RDC) and the within-subject coefficient of variation (wCV) for the whole prostate. To estimate the RDC, we used a one-way ANOVA model with random subject effect to determine the within-subject standard deviation (wSD), with normality of error terms assessed using a quantile-quantile (QQ) plot of the model residuals. RDC was then calculated as [2.77*wSD].[17] The wCV is a unitless ratio defined as [wSD/mean], and was estimated according to Quan and Shih, with 95% confidence interval calculated using the variance stabilizing transformation suggested by Shoukri et al.[18,19] We conducted identical analyses for focal lesions (when available).

We assessed the test-retest performance of DWI ADC under repeatability conditions using the same-day V2pre and V2post scans. We produced Bland-Altman plots and estimated both the repeatability coefficient (RC) and the wCV for the whole prostate. We again estimated the wCV as described above and conducted identical analyses for focal lesions.

For the subset of subjects with a focal lesion, ADC data was independently provided by two readers. We assessed the interreader reliability of the focal lesion ADC by calculating the RDC, RC and wCV separately by reader. In addition, we report the intraclass correlation coefficient (ICC), based on a single measurement, absolute agreement, two-way random effect model.[20] We calculated the ICC separately for each time point, with the degree of reliability determined using the guidelines suggested by Koo and Li.[21]

As the squared difference between the two measurements within a subject is proportional to the within subject variance, we conducted a two-sided paired t-test at a significance level of 0.05 on the log transformation of the squared differences between repeatability and reproducibility assessments to determine whether the within-subject variances differed significantly between the same-day and different-day approaches, which would imply a significant difference between the RDC and RC.[17,22] Finally, we explored the potential relationship between focal lesion size and repeatability by plotting the subject-specific coefficient of variation (CV) against ROI volume, and calculating the Spearman correlation coefficient. This was done for each reader under both reproducibility (V1/V2pre) and repeatability (V2pre/V2post) conditions, where ROI volume was obtained by averaging over the two time points in question.[23]

A description of the sample size calculations for ACRIN 6701 can be found in the supplement.

Results

All sites passed site qualification requirements using the ice-water phantom. The ADC biases for b_0–b_{600} and b_0–b_{800} calculations averaged (standard deviation in parentheses) 1.32% (1.85%) and 1.44% (1.70%), respectively. The mean *b*-value dependence was 0.188% (0.169%), and the SNR was 149 (50). We validated the IDL ADC calculation using a DWI digital reference object (DRO), provided by the University of Michigan team and available from the RSNA Quantitative Imaging Data Warehouse (qidw.rsna.org*).[24] Using DRO data over an appropriate range of *b*-values, with an SNR equivalent of 60 (at b_0), we found that our ADC calculation method resulted in bias of <5% over the typical prostate tissue ADC of 0.5–1.5 x 10^{-3} mm^2/s.

Cohort Selection

Thirty-five subjects from seven centers were recruited for this study (Table 1). This included a total of 11 subjects from Siemens scanners, 14 subjects from GE scanners, and 10 subjects from Philips scanners. Software version was consistent on each scanner for all subjects across time points, except for one subject imaged using a different software version than other subjects on the same scanner.

We excluded 2 subjects due to non-protocol imaging at both visits (N=1) or failure to undergo visit 2 imaging (N=1). Of the remaining 33 subjects, we excluded 4 additional subjects from analysis due to technical deviations or severe artifacts on one or more of the DWI visits. This left 29 subjects for the “whole prostate” cohort with analyzable images on all three DWI series (Cohort 1). Of these 29 patents, 9 demonstrated a focal restricting

* https://qidw.rsna.org/#folder/5bb7cfa7b3467a6a9210bfe2

lesion in the peripheral zone that met minimum size requirements; one subject demonstrated a dominant restricting BPH nodule. These 10 subjects formed the "focal lesion" cohort (Cohort 2). Figure 2 provides a schematic of the subject populations in this study; Table 2 provides demographic information, as well as baseline Gleason scores and PSA levels for both cohorts.

We also analyzed subjects with incomplete imaging for whole prostate and/or focal lesion DWI. These results are not included in Cohorts 1 or 2 but are in the supplemental material.

ADC repeatability and reproducibility in the prostate

Figure 3 demonstrates a representative example of the segmented prostate and focal lesion in a subject across the V1, V2pre, and V2post series. Table 3 summarizes the whole prostate and focal lesion results for each cohort across all three imaging exams (V1, V2pre, V2post), where reader 1 supplied data for both Cohort 1 (whole prostate) and Cohort 2 (focal lesion), and reader 2 supplied data for Cohort 2.

We also determined the wCVs, reproducibility coefficients (RDC, between V1 and V2pre), and repeatability coefficients (RC, between V2pre and V2post) for both cohorts (Table 4). There was no significant difference between the RDC and RC for either Cohort 1 (reader 1, *p*=0.52) or Cohort 2 (reader 1, *p*=0.14; reader 2, *p*=0.57). These results indicate that in the context of our study, there was no significant difference between for the same-day (repeatability) or different-day (reproducibility) wCVs estimates for either the whole prostate or focal lesions. Figure 4 demonstrates a box-and-whiskers plot of the two distinct pairwise comparisons for both cohorts. Corresponding Bland-Altman plots can be found in the supplemental material.

Interreader reliability of ADC measurements.

We estimated the wCV, RDC and RC for Cohort 2 separately for reader 2 (Table 4). Agreement plots for the focal lesion ADC between reader 1 and reader 2 can be found in the supplemental material. Based on the ICC, all three time points demonstrated good to excellent reliability between readers for focal lesion ADC (Table 5).

Effect of tumor ROI size on reliability of ADC measurements.

Figure 5 demonstrates a scatter plot showing the relationship between mean ROI size and subject-specific CV of the ADC measurement for each reader. For reader 1, smaller ROIs tended to have a larger subject-specific CV, hence demonstrating a negative association, although this association did not consistently achieve statistical significance across all time points (Spearman correlation coefficients: -0.54 (*p*=0.11) for V1/V2pre, -0.67 (*p*=0.03) for V2pre/V2post). However, this trend was not as apparent for reader 2 (Spearman correlation coefficients: 0.01 (*p*=0.99) for V1/V2pre, -0.01 (*p*=0.99) for V2pre/V2post).

Discussion

In this study, we sought to determine the reproducibility (different-day evaluation) and repeatability (same-day evaluation) of whole-prostate and focal-lesion ADC values in a test-retest setting. We defined a cohort of subjects across multiple imaging centers using different MRI vendors and software platforms. Furthermore, we sought to undertake a direct comparison of the estimates of ADC repeatability and reproducibility in the same subject cohort and to assess the interreader reliability of focal lesion ADC. To our knowledge, this type of test-retest analyses of prostate DWI in a multi-institutional setting has not previously been undertaken.

Our wCV of the median ADC value for whole prostate indicates that this quantitative ADC analysis for the whole prostate is highly reproducible, similar to that obtained in other studies of ADC test-retest performance in brain[25-27], breast[12,14,28], and prostate[29-31]. Our results further demonstrate that reproducibility and repeatability are similar. Medved et al. noted significant differences in whole prostate ADC pre- and post-ejaculation, suggesting that such ADC values might be expected to vary significantly over time.[32] Our whole prostate results, demonstrating no difference in wCV estimates between same-day and different-day exams is therefore somewhat unexpected. However, our cohort was significantly older than that of Medved et al. It is possible that in our population, the frequency and/or physiologic effects of ejaculation between V1 and V2 visits is markedly reduced from that which may be expected in a younger cohort.

The wCV of the median ADC for focal lesions is greater than that of the whole prostate, with values of 7.4–14.8%. While this suggests that there will be more variability in quantitative assessment of ADC of focal prostate lesions, our wCV results align with the

17% wCV claim cited in the DWI ADC Profile published by the Quantitative Imaging Biomarker Association (QIBA) of the RSNA.[7]

Although ACRIN 6701 was not powered to compare RDC and RC, we found no meaningful difference in ADC reliability metrics in either whole prostate or focal lesion between the V1/V2pre (RDC) and V2pre/V2post (RC). These two methods for estimating reliability of ADC values have not previously been studied in a head-to-head comparison with the same subject population. As most test-retest studies of DWI in humans are performed in a single visit (i.e., the coffee-break approach), our study suggests that same day test-retest evaluation may be a dependable measure of ADC reliability over a period of time, assuming consistent scan protocol. Many quantitative ADC evaluations in cancer imaging seek to determine the threshold for significant change after initiation of anti-tumor therapy.[14,27,33-35] Our results suggest the use of coffee-break CV standards for determining minimal threshold change in ADC that can be attributed to the effects of anti-tumor therapies or biologic evolution of lesions rather than random variation.

Our values for reproducibility are similar to those reported by others. Fedorov et al. reported a single-institution cohort of 15 subjects who underwent multiparametric MRI with endorectal coil at two time points separated by less than two weeks.[30] They reported mean ADC value RCs of 0.295 for the whole prostate (N=15) and 0.418 for the focal lesions (N=11). Barrett et al. also studied a single-institution 10 subject cohort in a "coffee-break" test-retest setting using phased-array coil, obtaining a RC of 0.271 for tumor nodules in these subjects (whole gland values were not reported).[29] These investigators also evaluated ROI "erosion" to evaluate the effect of edge ROIs on median ADC results, showing little change. Other investigators in single-institution studies have found similar results.[36,37] Our RDC/RC values of 0.132/0.131 (whole gland), 0.350/0.385 (reader 1, focal lesion), and 0.205/0.196 (reader 2, focal lesion) are very similar to these single-institutional results, suggesting the robustness of prostate ADC quantification on most modern MRI scanners.

Our DWI acquisition protocol varied slightly from the standards of the PI-RADS recommendations, and those discussed in the QIBA DWI Profile. Most notably, we used a slice thickness of 5 mm, whereas PI-RADS recommends a slice thickness of 3 mm or less. As the primary endpoint of our protocol was geared toward the evaluation of the whole prostate as a surrogate for a generic nodule in the abdominal/pelvis region, this difference was reasonable to ensure a manageable number of slices for whole prostate evaluation. The subsequent development of the PI-RADS guidelines sought to maximize the imaging evaluation of small tumor nodules. As such, reduction in slice thickness (while maintaining adequate image SNR) may improve the reliability of ADC estimates for smaller focal prostatic lesions.

While we sought to directly compare reproducibility vs. repeatability of whole prostate and tumor ADC values in the same subject cohort, our reproducibility study differed from the repeatability study only in the use of a different day for V2 (2–14 days after V1). We strictly maintained other variables in our study, including the use of the same scanner and software, and the same image acquisition parameters for each subject. Variation in scanner technology could introduce other sources of variability not reflected in the reproducibility estimates in our study.

Tamada et al. evaluated the effect of reader on the RC of the mean ADC of focal lesions. They found that reader effects contributed approximately 10% (intrareader) and 20%

(interreader) to ADC variation.[38] More studies will elucidate the variability introduced by different MRI readers, and the effects of ROI placement errors on mean and median prostate ADC values.

Limitations

Notwithstanding a cohort of 35 enrolled subjects across 7 imaging centers, we were only able to undertake an analysis of reproducibility/repeatability of prostate lesion ADC across 10 subjects. Despite the use of eligibility criteria to limit enrollment to subjects with PSA ≥ 10 ng/mL, Gleason score ≥ 7, or minimum core biopsy nodule "density", many of our subjects did not demonstrate either a focal restricting lesion or a restricting BPH nodule of adequate size. Focal-lesion-exhibiting subjects (Cohort 2) were not distinguished from the broader cohort by PSA value or Gleason score, with a slight difference in core biopsy tumor density between Cohorts 1 and 2. Future studies should seek alternate means of creating more restrictive entry criteria in hopes of obtaining an adequate number of tumor-bearing subjects for analysis.

We sought to evaluate the robustness in the wCV estimate by utilizing two separate readers who were unaware of each other's segmentation results. While our study was not powered to determine the exact contribution of reader to wCV variability, our results confirm that the DWI profile claim was met for each reader and that interreader reliability was consistently high. Additional research is required to better evaluate the effects of individual readers on focal lesion ADC estimates in prostate MRI. While we did not test reader effects on the estimated wCV for whole prostate ADC, it is expected than any such reader effects will be small.

Conclusion

We have demonstrated in a test-retest analysis that DWI ADC measurements are robust in prostate MRI performed in a multi-institutional setting, and that values in line with the estimates provided in the QIBA profile can be achieved even in a setting of multiple MRI hardware and imaging environments. Furthermore, we demonstrated good interreader reliability for focal lesion ADC and our results suggest minimal differences in the repeatability (same day) and reproducibility (different day) estimates of median ADC for both whole prostate and focal lesions. Although our study was not able to establish formal equivalence between same day and different day results, it lends some support to the continued use of same day (coffee-break) test-retest analysis as a facile means of assessing the reliability of tumor or tissue ADC estimates.

References

1. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. Radiology 1986;161(2):401-407.
2. Stejskal EO, Tanner JE. Spin diffusion measurements: spin echoes in the presence of a time‐dependent field gradient. The journal of chemical physics 1965;42(1):288-292.
3. Le Bihan D, Turner R, Douek P, Patronas N. Diffusion MR imaging: clinical applications. AJR Am J Roentgenol 1992;159(3):591-599.
4. Barentsz JO, Weinreb JC, Verma S, et al. Synopsis of the PI-RADS v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use. European urology 2016;69(1):41.
5. Padhani AR, Weinreb J, Rosenkrantz AB, Villeirs G, Turkbey B, Barentsz J. Prostate imaging-reporting and data system steering committee: PI-RADS v2 status update and future directions. European urology 2019;75(3):385-396.
6. Committee P-R. PI-RADS: Prostate Imaging–Reporting and Data System, Version 2.1. Accessed June 23, 2021. Volume 2021. https://www.acr.org/-/media/ACR/Files/RADS/Pi-RADS/PIRADS-V2-1.pdf?la=en; 2019.
7. QIBA DWI Biomarker Committee. QIBA Profile: Diffusion-Weighted Magnetic Resonance Imaging (DWI), Consensus Version. Accessed 06Jan2021. http://qibawiki.rsna.org/images/6/63/QIBA_DWIProfile_Consensus_Dec2019_Final.pdf; 2020.
8. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res 2015;24(1):9-26.
9. BIPM I, IFCC, ILAC, IUPAC, IUPAP, ISO, OIML The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn. http://www.bipm.org/vim; 2012.
10. 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.
11. Holz M, Heil SR, Sacco A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements. Physical Chemistry Chemical Physics 2000;2(20):4740-4742.
12. Newitt DC, Zhang Z, Gibbs JE, et al. Test-retest repeatability and reproducibility of ADC measures by breast DWI: Results from the ACRIN 6698 trial. J Magn Reson Imaging 2018.
13. 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.
14. Partridge SC, Zhang Z, Newitt DC, et al. Diffusion-weighted MRI Findings Predict Pathologic Response in Neoadjuvant Treatment of Breast Cancer: The ACRIN 6698 Multicenter Trial. Radiology 2018;289(3):618-627.
15. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. Radiology 1986;161(2):401-407.
16. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology Standards for Quantitative Imaging Biomarkers. Radiology 2015;277(3):813-825.

17. 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.

18. Shoukri MM, Elkum N, Walter SD. Interval estimation and optimal design for the within-subject coefficient of variation for continuous and binary variables. BMC Med Res Methodol 2006;6:24.

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

20. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420-428.

21. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15(2):155-163.

22. Bland JM. Comparing within-subject variances in a study to compare two methods of measurement. Accessed 02Sep2020. https://www-users.york.ac.uk/~mb55/meas/compsd.htm; 2010.

23. Spearman Rank Correlation Coefficient. The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 502-505.

24. Malyarenko DI, Pang Y, Amouzandeh G, Chenevert TL. Numerical DWI phantoms to optimize accuracy and precision of quantitative parametric maps for non-Gaussian diffusion. Medical Imaging 2020: Image Processing. Volume 11313: International Society for Optics and Photonics; 2020. p. 113130W.

25. Bonekamp D, Nagae LM, Degaonkar M, et al. Diffusion tensor imaging in children and adolescents: reproducibility, hemispheric, and age-related differences. Neuroimage 2007;34(2):733-742.

26. Pfefferbaum A, Adalsteinsson E, Sullivan EV. Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. J Magn Reson Imaging 2003;18(4):427-433.

27. Paldino MJ, Barboriak D, Desjardins A, Friedman HS, Vredenburgh JJ. Repeatability of quantitative parameters derived from diffusion tensor imaging in patients with glioblastoma multiforme. J Magn Reson Imaging 2009;29(5):1199-1205.

28. 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.

29. Barrett T, Lawrence EM, Priest AN, et al. Repeatability of diffusion-weighted MRI of the prostate using whole lesion ADC values, skew and histogram analysis. European journal of radiology 2019;110:22-29.

30. Fedorov A, Vangel MG, Tempany C, Fennessy F. Multiparametric MRI of the prostate: repeatability of volume and apparent diffusion coefficient quantification. Invest Radiol 2017;52:538-546.

31. Gibbs P, Pickles, M.D., L.W. Turnbull. Repeatability of echo-planar-based diffusion measurements of the human prostate at 3T. Magn Reson Imaging 2007;25(10):1423-1429.

32. Medved M, Sammet S, Yousuf A, Oto A. MR imaging of the prostate and adjacent anatomic structures before, during, and after ejaculation: qualitative and quantitative evaluation. Radiology 2014;271(2):452-460.

33. Vollenbrock SE, Voncken FEM, Bartels LW, Beets-Tan RGH, Bartels-Rutten A. Diffusion-weighted MRI with ADC mapping for response prediction and assessment of oesophageal cancer: A systematic review. Radiother Oncol 2020;142:17-26.

34. Klompenhouwer EG, Dresen RC, Verslype C, et al. Transarterial Radioembolization Following Chemoembolization for Unresectable Hepatocellular Carcinoma: Response Based on Apparent Diffusion Coefficient Change is an Independent Predictor for Survival. Cardiovasc Intervent Radiol 2018;41(11):1716-1726.

35. Saleh MM, Abdelrahman TM, Madney Y, Mohamed G, Shokry AM, Moustafa AF. Multiparametric MRI with diffusion-weighted imaging in predicting response to chemotherapy in cases of osteosarcoma and Ewing's sarcoma. Br J Radiol 2020;93(1115):20200257.
36. Sadinski M, Medved M, Karademir I, et al. Short-term reproducibility of apparent diffusion coefficient estimated from diffusion-weighted MRI of the prostate. Abdom Imaging 2015;40(7):2523-2528.
37. Michoux NF, Ceranka JW, Vandemeulebroucke J, et al. Repeatability and reproducibility of ADC measurements: a prospective multicenter whole-body-MRI study. Eur Radiol 2021:1-14.
38. Tamada T, Huang C, Ream JM, Taffel M, Taneja SS, Rosenkrantz AB. Apparent diffusion coefficient values of prostate cancer: comparison of 2D and 3D ROIs. American Journal of Roentgenology 2018;210(1):113-117.

Tables

Table 1: Site Enrollment and Scanner Information

Site	**Scanner**	**Software Version**	**Enrolled subjects**	**Cohort 1:** Whole prostate	**Cohort 2:** Focal lesion
Site A	GE Discovery MR750	DV24.0	2	0	0
Site B	GE Discovery MR750	DV23.1	6	6	1
Site C	Siemens Skyra	D13/E11[†]	5	5	1
Site D	Philips Achieva	3.2.2	4	4	0
Site E	Philips Ingenia	4.1.3	6	5	4
Site F	Siemens Trio	B17	6	4	1
Site G	GE Signa HDXT	HD16.0	6	5	3
Total			**35**	**29**	**10**

All subjects in Cohorts 1 and 2 had analyzable V1, V2pre, and V2post images.

[†] One subject (subject 33) scanned under deviation from qualification following software upgrade to E11

Table 2: Baseline demographics and clinical characteristics

	All enrolled subjects N=35		**Cohort 1:** Whole prostate [V1/V2pre/V2post all analyzable] **N=29**		**Cohort 2:** Focal lesion [V1/V2pre/V2post all analyzable] **N=10**	
	N	**%**	**N**	**%**	**N**	**%**
Age						
Mean (std)	64.5 (7.3)		64.4 (7.1)		62.0 (6.7)	
Median (Range)	63 (52,80)		63 (53, 80)		60.5 (53, 78)	
Race						
Black/African American	3	9%	3	10%	0	0%
White	29	83%	23	79%	9	90%
Not reported/Unknown	3	9%	3	10%	1	10%
Ethnicity						
Hispanic/Latino	3	9%	3	10%	0	0%
Not Hispanic/Latino	31	89%	25	86%	10	100%
Not reported/Unknown	1	3%	1	3%	0	0%
Insurance Status						
Private insurance	22	63%	19	66%	9	90%
Medicare/Other government insurance	7	20%	6	21%	0	0%
Medicaid/Uninsured	6	17%	4	14%	1	10%
PSA (ng/mL)						
Mean (Std)	14.7 (25.8)		10.6 (9.7)		12.2 (15.4)	
Median (Range)	8.4 (3.7, 153.5)		8.4 (3.7, 55.2)		7.3 (4.1, 55.2)	
Gleason Score						
Unknown	1	3%	0	0%	0	0%
6	3	9%	3	10%	1	10%
7	26	74%	23	79%	7	70%
8	2	6%	2	7%	2	20%
9	3	9%	1	3%	0	0%
Core Biopsy Tumor Density [1]						
Low	13	37%	9	31%	2	20%
High	22	63%	20	69%	8	80%

[1] High density: one core ≥ 50% tumor burden and ≥ 5mm in tumor length, two or more cores ≥ 30% tumor burden, or three or more positives cores in the same sextant.

Table 3: Distributional statistics for ADC (x 10^{-3} mm^2/s) by whole prostate and focal lesion regions

Region	Time point	Mean	Std Dev	Median	Min	Max
Reader 1 Whole prostate (Cohort 1, N=29)	V1	1.44	0.146	1.46	1.02	1.73
	V2pre	1.44	0.132	1.45	1.03	1.68
	V2post	1.44	0.122	1.45	1.06	1.66
Reader 1 Focal lesion (Cohort 2, N=10)	V1	0.893	0.266	0.808	0.561	1.28
	V2pre	0.951	0.258	0.910	0.624	1.30
	V2post	0.932	0.247	0.858	0.600	1.33
Reader 2 Focal lesion (Cohort 2, N=10)	V1	0.927	0.198	0.855	0.697	1.23
	V2pre	0.958	0.224	0.915	0.704	1.39
	V2post	0.963	0.207	0.918	0.681	1.29

Table 4: Agreement statistics for DWI ADC by whole prostate and focal lesion regions

Region	Reproducibility (V1, V2pre)		Repeatability (V2pre, V2post)	
	RDC (95% CI) x 10^{-3} mm^2/s	wCV (95% CI)	RC (95% CI) x 10^{-3} mm^2/s	wCV (95% CI)
Reader 1 Whole prostate (Cohort 1, N=29)	0.132 (0.105, 0.177)	0.0330 (0.0255, 0.0429)	0.131 (0.104, 0.176)	0.0328 (0.0253, 0.0426)
Reader 1 Focal lesion (Cohort 2, N=10)	0.350 (0.244, 0.614)	0.137 (0.0868, 0.222)	0.385 (0.269, 0.675)	0.148 (0.0937, 0.238)
Reader 2 Focal lesion (Cohort 2, N=10)	0.205 (0.143, 0.360)	0.0785 (0.0500, 0.125)	0.196 (0.137, 0.344)	0.0736 (0.0469, 0.117)

Table 5: Intraclass correlation coefficient (ICC) by time point for assessing interreader reliability of the focal lesion DWI ADC.

Timepoint	**ICC (95% CI)** [1,2]
V1	0.89 (0.65, 0.97)
V2pre	0.76 (0.28, 0.94)
V2post	0.94 (0.79, 0.98)

[1] ICC corresponds to ICC(2,1) in the nomenclature of Shrout and Fleiss, based on a single measurement, absolute agreement, two-way random effects model.

[2] Koo and Li give the following guidelines for interpreting the ICC:

Below 0.50: poor reliability
Between 0.50 and 0.75: moderate reliability
Between 0.75 and 0.90: good reliability
Above 0.90: excellent reliability

Repeatability and Reproducibility Assessment of the Apparent Diffusion Coefficient in the Prostate: A Trial of the ECOG-ACRIN Research Group (ACRIN 6701)

Michael A. Boss, PhD[1†], Bradley S. Snyder, MS[2], Eunhee Kim, PhD[2*], Dena Flamini, RT[1], Sarah Englander, PhD[3], Karthik M. Sundaram, MD[3], Naveen Gumpeni, MD[4], Suzanne L. Palmer, MD[5], Haesun Choi, MD[6], Adam T. Froemming, MD[7], Thorsten Persigehl, MD[8], Matthew S. Davenport, MD[9], Dariya Malyarenko, PhD[9], Thomas L. Chenevert, PhD[9], Mark A. Rosen MD, PhD[3]

[1]Center for Research and Innovation, American College of Radiology Philadelphia, Pennsylvania, USA; [2]Department of Biostatistics, Brown University, Providence, Rhode Island, USA; [3]Department of Radiology, University of Pennsylvania, Philadelphia, Pennsylvania; [4]Department of Radiology, Weill Cornell Medical Center, New York, New York, USA; [5]Department of Radiology, University of Southern California, Los Angeles, California, USA; [6]Department of Abdominal Imaging, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA; [7]Department of Radiology, May Clinic, Rochester, Minnesota, USA; [8]Department of Radiology, University Hospital Cologne, Cologne, Germany; [9]Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA

[†]Corresponding author, email: mboss@acr.org

[*]Presently at Merck Research Laboratories, Kenilworth, New Jersey, USA

Acknowledgements

This project has been funded in part with Federal funds from the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN268201000050C, and funding from the National Cancer Institute, under Award No. U24CA180803.

This study was coordinated by 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: U10CA180820, U10CA180794. 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 U.S. government.

Professor Edward F. Jackson, PhD played an instrumental role in ACRIN 6701 as study co-chair, and a leader in QIBA, including time as Chair. We are grateful for his many contributions, sage advice, and unending enthusiasm and patience in advancing standards for quantitative imaging. We endeavor to continue his mission and realize his vision.

Running Title: ACRIN 6701 Prostate ADC Reproducibility

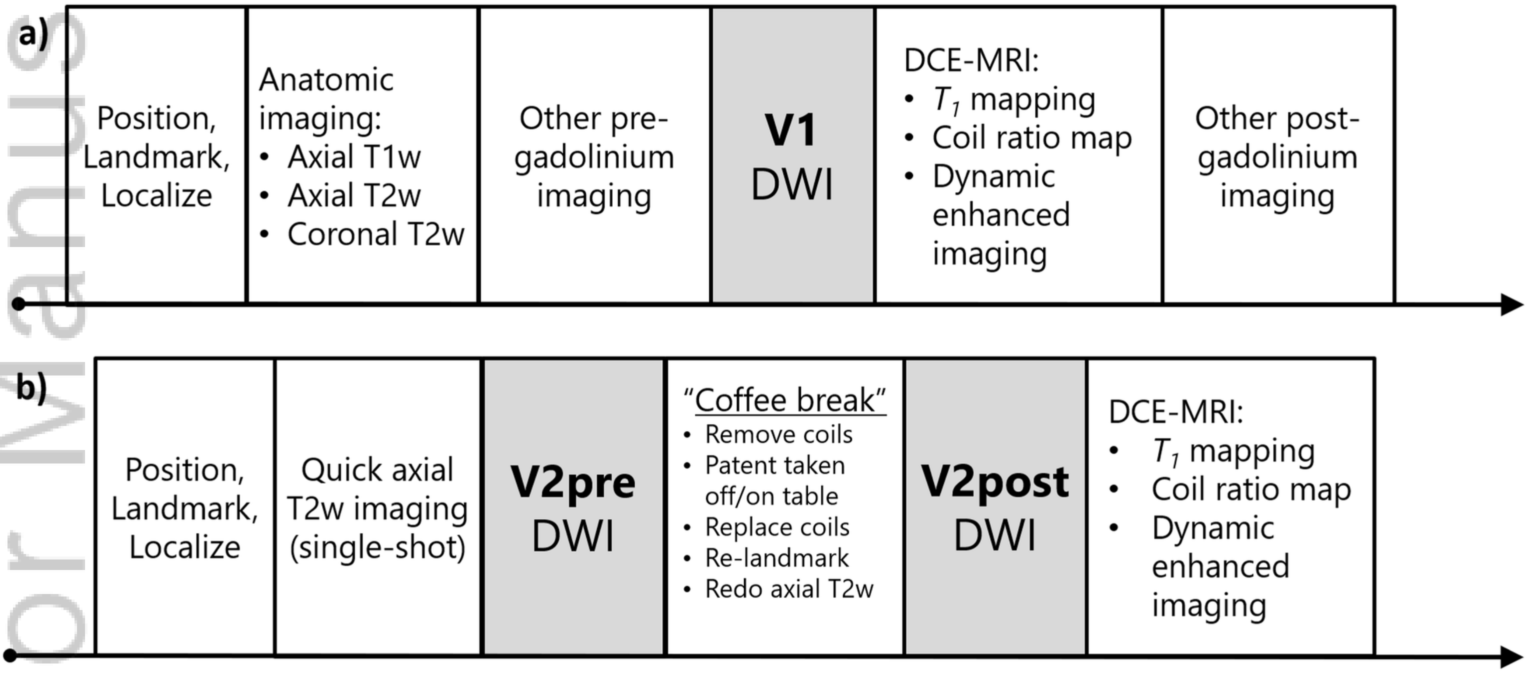

Figure 1_flat.tif

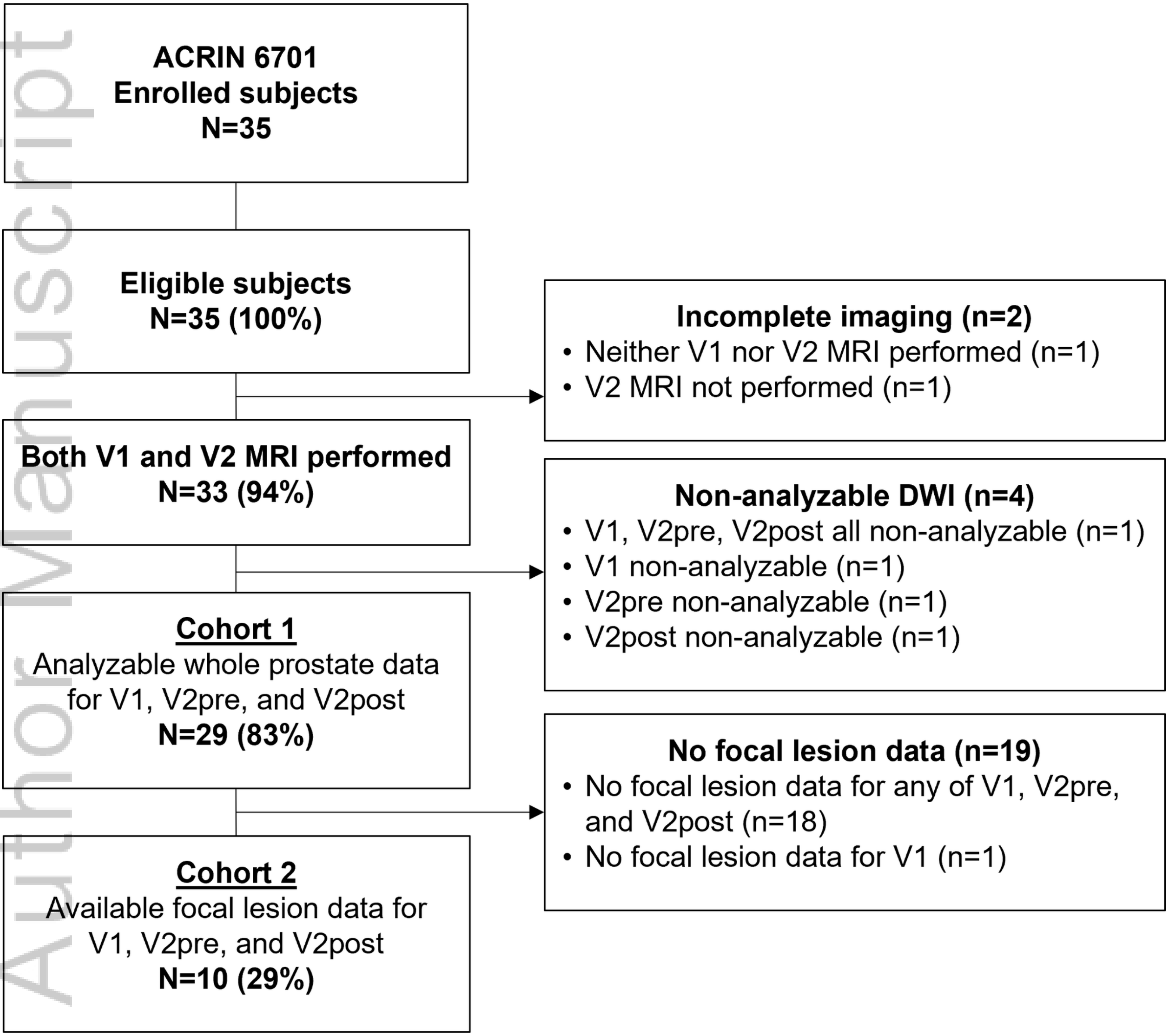
ACRIN 6701
Enrolled subjects
N=35
Eligible subjects
N=35 (100%)
Incomplete imaging (n=2)
• Neither V1 nor V2 MRI performed (n=1)
• V2 MRI not performed (n=1)
Both V1 and V2 MRI performed
N=33 (94%)
Non-analyzable DWI (n=4)
• V1, V2pre, V2post all non-analyzable (n=1)
• V1 non-analyzable (n=1)
• V2pre non-analyzable (n=1)
• V2post non-analyzable (n=1)
Cohort 1
Analyzable whole prostate data for V1, V2pre, and V2post
N=29 (83%)
No focal lesion data (n=19)
• No focal lesion data for any of V1, V2pre, and V2post (n=18)
• No focal lesion data for V1 (n=1)
Cohort 2
Available focal lesion data for V1, V2pre, and V2post
N=10 (29%)

Figure 2_flat.tiff

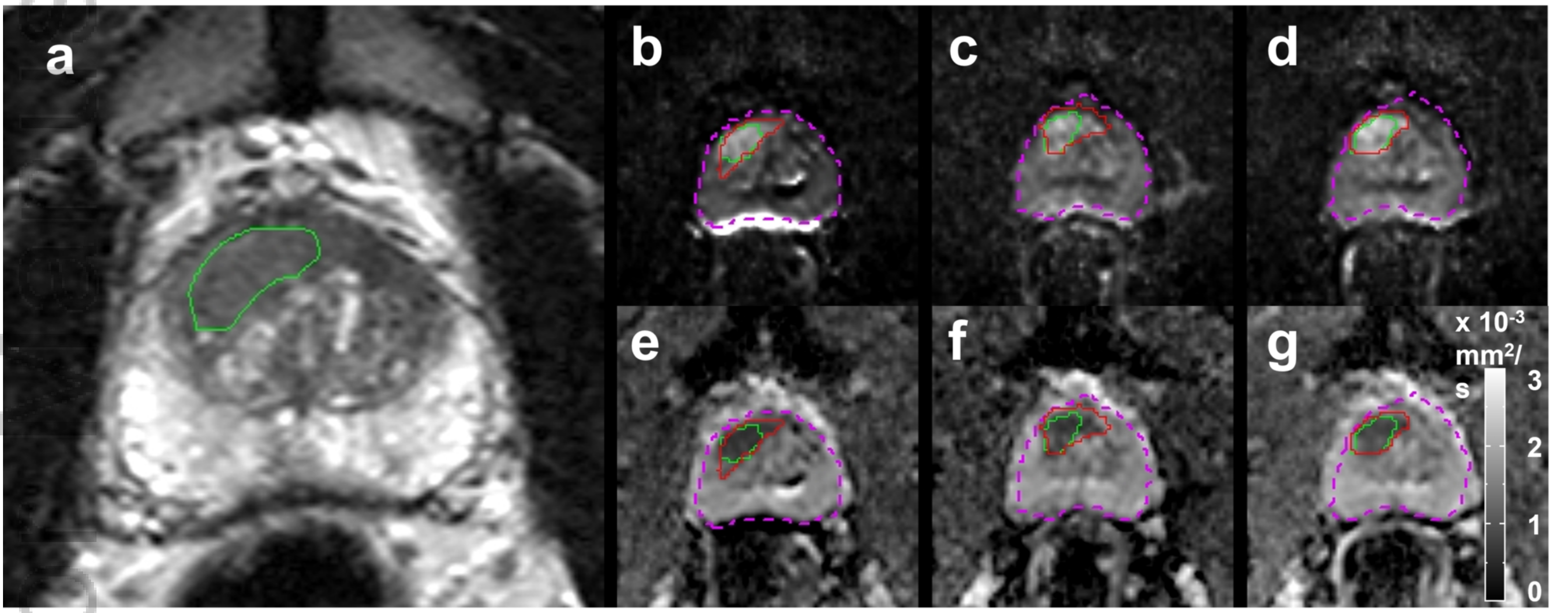

Figure 3.tif

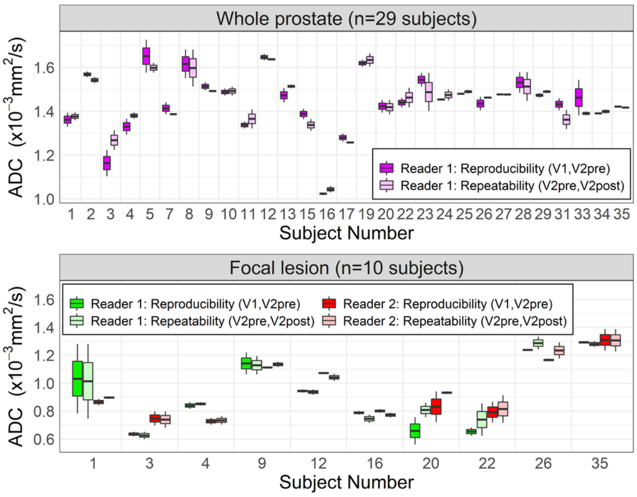
Whole prostate (n=29 subjects)
ADC (x10^{-3}mm^2/s)
1.0
1.2
1.4
1.6
Reader 1: Reproducibility (V1,V2pre)
Reader 1: Repeatability (V2pre,V2post)
1 2 3 4 5 7 8 9 10 11 12 13 15 16 17 19 20 22 23 24 25 26 27 28 29 31 33 34 35
Subject Number
Focal lesion (n=10 subjects)
ADC (x10^{-3}mm^2/s)
0.6
0.8
1.0
1.2
1.4
1.6
Reader 1: Reproducibility (V1,V2pre)
Reader 1: Repeatability (V2pre,V2post)
Reader 2: Reproducibility (V1,V2pre)
Reader 2: Repeatability (V2pre,V2post)
1 3 4 9 12 16 20 22 26 35
Subject Number

Figure 4.tif

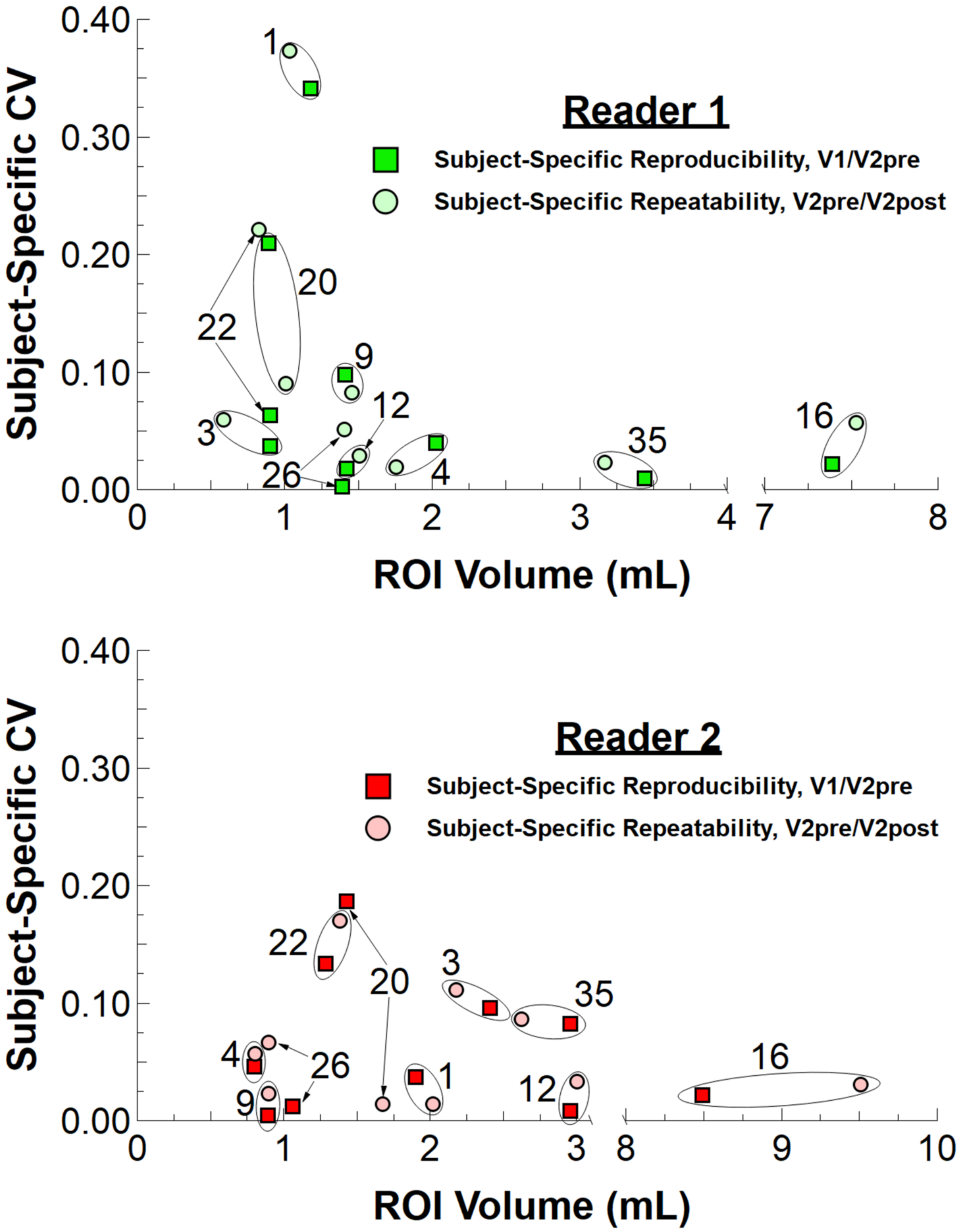

Figure 5_combined.tif