Multi-Site Concordance of Diffusion-Weighted Imaging Quantification for Assessing Prostate Cancer Aggressiveness

Sean D. McGarry, PhD,[1] Michael Brehler, PhD,[2] John D. Bukowy, PhD,[3] Allison K. Lowman, BS,[2] Samuel A. Bobholz, BS,[1] Savannah R. Duenweg, BS,[1] Anjishnu Banerjee, PhD,[4] Sarah L. Hurrell, BS,[2] Dariya Malyarenko, PhD,[5] Thomas L. Chenevert, PhD,[5] Yue Cao, PhD,[5,6] Yuan Li, PhD,[6] Daekeun You, PhD,[6] Andrey Fedorov, PhD,[7] Laura C. Bell, PhD,[8] C. Chad Quarles, PhD,[8] Melissa A. Prah, BS,[1] Kathleen M. Schmainda, PhD,[1] Bachir Taouli, MD,[9] Eve LoCastro, MS,[10] Yousef Mazaheri, PhD,[10,11] Amita Shukla-Dave, PhD,[10,11] Thomas E. Yankeelov, PhD,[12] David A. Hormuth II PhD,[12] Ananth J. Madhuranthakam, PhD,[13] Keith Hulsey, PhD,[13] Kurt Li,[14] Wei Huang, PhD,[15] Wei Huang, MD,[16] Mark Muzi, PhD,[17] Michael A. Jacobs, PhD,[18] Meiyappan Solaiyappan, MS,[18] Stefanie Hectors, PhD,[19] Tatjana Antic, MD,[20] Gladell P. Paner, MD,[20] Watchareepohn Palangmonthip, MD,[21,22] Kenneth Jacobsohn, MD,[23] Mark Hohenwalter, MD,[2] Petar Duvnjak, MD,[2] Michael Griffin, MD,[2] William See, MD,[23] Marja T. Nevalainen, MD, PhD,[21] Kenneth A. Iczkowski, MD,[21] and Peter S. LaViolette, PhD[2,24]*

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.27983

Received Jul 27, 2021, Accepted for publication Oct 22, 2021.

*Address reprint requests to: P.S.L., 8701 Watertown Plank Rd., Milwaukee, WI 53226, USA. E-mail: plaviole@mcw.edu

[Correction added on 08 December 2021, after first online publication: Affiliation 2 has been changed to Department of Radiology Medical College of Wisconsin, Milwaukee, Wisconsin, USA.]

From the [1]Department of Biophysics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA; [2]Department of Radiology, Medical College of Wisconsin, Milwaukee, WI, USA; [3]Department of Electrical Engineering and Computer Science, Milwaukee School of Engineering, Milwaukee, WI, USA; [4]Division of Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA; [5]Department of Radiology, University of Michigan, Ann Arbor, Michigan, USA; [6]Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan, USA; [7]Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, USA; [8]Division of Neuroimaging Research, Barrow Neurological Institute, Phoenix, Arizona, USA; [9]Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [10]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, New York, USA; [11]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, New York, USA; [12]Department of Biomedical Engineering, Diagnostic Medicine, Oncology, Oden Institute for Computational Engineering and Sciences, Livestrong Cancer Institutes, The University of Texas, Austin, Texas, USA; [13]Department of Radiology, The University of Texas Southwestern Medical Center, Dallas, Texas, USA; [14]International School of Beaverton, Aloha, Oregon, USA; [15]Advanced Imaging Research Center, Oregon Health Sciences University, Portland, Oregon, USA; [16]Department of Pathology, Oregon Health and Science University, Madison, Wisconsin, USA; [17]Department of Radiology, Neurology, and Radiation Oncology, University of Washington, Seattle, Washington, USA; [18]The Russell H. Morgan Department of Radiology and Radiological Science and Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; [19]Department of biomedical engineering and imaging institute, Weill Cornell Medical College, New York City, New York, USA; [20]Department of Pathology, University of Chicago, Chicago, Illinois, USA; [21]Department of Pathology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA; [22]Department of Pathology, Chiang Mai University, Chiang Mai, Thailand; [23]Department of Urology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA; and [24]Department of Biomedical Engineering, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

Additional supporting information may be found in the online version of this article

Background: Diffusion-weighted imaging (DWI) is commonly used to detect prostate cancer, and a major clinical challenge is differentiating aggressive from indolent disease.
Purpose: To compare 14 site-specific parametric fitting implementations applied to the same dataset of whole-mount pathologically validated DWI to test the hypothesis that cancer differentiation varies with different fitting algorithms.
Study Type: Prospective.
Population: Thirty-three patients prospectively imaged prior to prostatectomy.
Field Strength/Sequence: 3 T, field-of-view optimized and constrained undistorted single-shot DWI sequence.
Assessment: Datasets, including a noise-free digital reference object (DRO), were distributed to the 14 teams, where locally implemented DWI parameter maps were calculated, including mono-exponential apparent diffusion coefficient (MEADC), kurtosis (K), diffusion kurtosis (DK), bi-exponential diffusion (BID), pseudo-diffusion (BID*), and perfusion fraction (F). The resulting parametric maps were centrally analyzed, where differentiation of benign from cancerous tissue was compared between DWI parameters and the fitting algorithms with a receiver operating characteristic area under the curve (ROC AUC).
Statistical Test: Levene's test, $P < 0.05$ corrected for multiple comparisons was considered statistically significant.
Results: The DRO results indicated minimal discordance between sites. Comparison across sites indicated that K, DK, and MEADC had significantly higher prostate cancer detection capability (AUC range = 0.72–0.76, 0.76–0.81, and 0.76–0.80 respectively) as compared to bi-exponential parameters (BID, BID*, F) which had lower AUC and greater between site variation (AUC range = 0.53–0.80, 0.51–0.81, and 0.52–0.80 respectively). Post-processing parameters also affected the resulting AUC, moving from, for example, 0.75 to 0.87 for MEADC varying cluster size.
Data Conclusion: We found that conventional diffusion models had consistent performance at differentiating prostate cancer from benign tissue. Our results also indicated that post-processing decisions on DWI data can affect sensitivity and specificity when applied to radiological–pathological studies in prostate cancer.
Level of Evidence: 1
Technical Efficacy: Stage 3

J. MAGN. RESON. IMAGING 2022;55:1745–1758.

Prostate cancer accounts for one in five new cancer diagnoses in men, with an estimated 193,000 new cases in 2020,[1] although not all cases are high risk. Ongoing imaging evaluations are aimed at better differentiating aggressive from indolent disease to avoid over-treatment of non-aggressive prostate cancer and to accurately detect tumors that have high metastatic potential.[2] Advancements in multi-parametric magnetic resonance imaging (MP-MRI) such as T_2-weighted and diffusion-weighted imaging (DWI) have yielded substantial improvement for prostate cancer detection and MP-MRI is increasingly used for justifying and guiding biopsy.[3]

DWI is commonly used for diagnosing prostate cancer and is weighted heavily as a deciding factor in the Prostate Imaging Reporting and Data System (PIRADSv2.1) grading scale for radiographic diagnosis.[2,4,5] Tissue micro-structure strongly influences diffusion properties and abnormalities such as dense cellularity or atrophic glands can result in distinct imaging signatures.[6] However, the calculation of quantitative diffusion values varies by fitting algorithms and recent collaborative studies have looked to quantify differences between sites.[7]

There are three common fitting schemes for deriving quantitative maps from DWI. The apparent diffusion coefficient (ADC) is calculated from a mono-exponential fit of the different *b*-values from the DWI data and is the most common metric used for evaluation of prostate cancer.[2,4,5] More complex diffusion models have been developed to separate tissue diffusivity from capillary microperfusion.[8,9] By assuming a bi-exponential relationship between both diffusion and perfusion effects, the intra-voxel incoherent motion (IVIM) computes both pseudo-diffusion (BID*) and perfusion fraction (F).[8,9] Kurtosis (K) and diffusion kurtosis (DK) models measure deviations of diffusion from a Gaussian distribution[10] due to cellular restriction.

The aim of this study, by a collaborative group[11–14] organized by the National Cancer Institute (NCI), was to undertake a multi-institutional study to quantify whether prostate cancer detection varies due to differences in DWI fitting algorithms. In addition, we also measure changes in perceived cancer differentiation due to varying post-processing parameters were investigated as were changed due to varying the pathologist performing the ground truth annotations.

Methods

This study was proposed and organized through an NCI working group. Investigators from the central organizing institution and 14 other institutions participated. Data were collected at the central site and distributed to each satellite institution for processing. Fourteen implementations were included in this project from investigators at MCW (Team 2), University of Washington, Johns Hopkins University, University of Michigan, University of Texas at Austin, University of Texas Southwestern Medical Center, Oregon Health and Science University, Memorial Sloan Kettering Cancer Center, Mount Sinai, Brigham and Women's Hospital, and Barrows Neurological Institute, in no particular order. Resulting maps were then sent back to the central site for analysis. A diagram showing the design in this study can be seen in Fig. 1.

Patient Population and Data Acquisition

This study was IRB-approved at the central site. All patients provided written informed consent. Inclusion criteria required

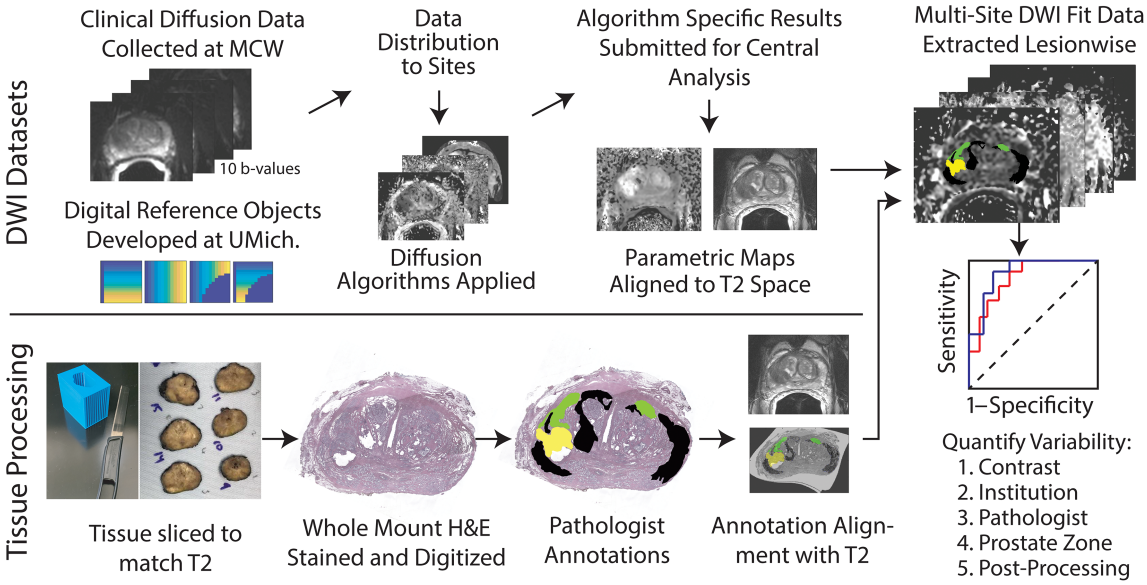

FIGURE 1: Schematic representation of the experimental design. Top: Raw diffusion data distributed to partner institutions in DICOM format, partner institutions return fits to MCW where they were manually aligned to the T2-weighted image. Bottom: Post-surgery, tissue was sliced to match the T2-weighted image using patient-specific slicing jigs. Whole-mount samples were stained and annotated by a pathologist. Annotations were then aligned to the T2-weighted image.[15–18] Right: Pathologist annotations and fits from multiple institutions were combined for analysis to determine variability in prostate cancer sensitivity and specificity.

patients to undergo MRI prior to prostatectomy and have high-quality images. Thirty-nine consecutive patients met the first inclusion criterion, scanned between December 2014 and August 2016. Six patients were subsequently excluded due to excessive motion on their MRI. Imaging from the remaining 33 patients (demographics and cancer stage indicated in Table 1) was acquired on a 3-T MRI scanner (General Electric, Waukesha, WI) using an endorectal coil and phased-array torso coil. The MP-MRI sequences comprised of field-of-view optimized and constrained undistorted single shot (FOCUS) DWI[19] with 10 *b*-values (b = 0, 10, 25, 50, 80, 100, 200, 500, 1000, and 2000 seconds/mm^2), NEX: 1, 2, 1, 1, 1, 2, 2, 4, 8, 12 respectively, repetition time/echo time (TR/TE) = 4/69–99 msec; interpolated resolution = 0.625 mm × 0.625 mm × 4 mm voxels, acquisition matrix 80 × 80, FOV 160 mm × 160 mm, echo train length 1 (80 echos). Additionally, an anatomical T_2-weighted multi-slice dataset was acquired (acquisition matrix 384 × 256, TR = 5000 msec, TE = 0.125 s, FA = 111 echo train length 24, interpolated 0.234 mm × 0.234 mm × 3 mm, FOV 120 mm × 120 mm). Robotic prostatectomy was performed approximately 2 weeks later and the extracted prostate was sectioned using patient-specific custom three-dimensional-printed slicing jigs to match orientation and 3 mm slice thickness of the T_2-weighted image.[6,15,20]

Histo-Pathological Analysis

Prostate samples were cut at 4 μm thickness, and whole-mount sections were hematoxylin and eosin (H&E) stained, digitized, and annotated by a urological fellowship-trained pathologist (K.I., 23 years of experience) (Fig. 1). A total of 169 slides were included. Each slice was manually aligned to the T_2-weighted image using control points and a non-linear transform. Regions with tears and histology artifacts were excluded with manually placed ROIs applied after the spatial transform. Annotations of different Gleason patterns were brought into MRI space using the same non-linear transform.[6,19] Pathologist-annotated (K.I.) regions that consisted of at least 200 contiguous voxels axially (11 mm^2 in plane, 33 mm^3) were included, which resulted in 231 cancer (CA) regions of interest (ROIs), and 564 ROIs not associated with cancer (benign atrophy, BA). These ROIs were used to extract the quantitative parametric diffusion values. A subset of slides was annotated by five pathologists from four universities with 23 (K.I.), 15 (W.H.), 13 (G.P.), 11 (T.A.), and 1 (W.P.) year of experience. This subset included 33 slides from 28 patients.[15]

Diffusion Signal Fitting

DICOM datasets obtained with FOCUS DWI were de-identified to meet HIPPA compliance and distributed to the collaborating sites for analysis. Each site was asked to calculate diffusion parameter maps using publicly available or locally developed software, implemented to fit DWI signals. The individual methods used for each site implementation are detailed in the supplement, and in Table 2.[7–9,16–18,21–31] These methods included a mono-exponential fit (parameter: MEADC), diffusion kurtosis (parameters: kurtosis [K] and diffusion [DK]),[10] and a bi-exponential fit (parameters: diffusion (BID), BID*, and F).[8] Each site submitted the calculated maps back to the central site for comparative analysis.

TABLE 1. Patient Demographics and Clinical Data ($N = 33$, Age 59.7 ± 5.7)

Patient No.	Age (Years)	PSA (ng/mL)	Gleason Score	Gleason Grade				T Stage	EPE	Number of PIRADS Lesions	PIRADS Score
				G3	G4-FG	G4-Cr	G5				
1	61	13.1	3 + 4 (=7)	1	1	0	0	T3a	1	1	PR4
2	68	4.5	3 + 4 (=7)	1	1	0	0	T2c	0	2	PR5, PR5
3	59	6.6	3 + 4 (=7)	1	0	0	0	T2c	0	2	PR3, PR4
4	56	4.4	5 + 4 (=9)	1	1	1	1	T3a	1	1	PR5
5	64	6.3	4 + 3 (=7)	1	1	1	0	T3a	1	1	PR5
6	55	4.9	3 + 4 (=7)	1	1	0	0	T3b	0	1	PR4
7	58	21.9	3 + 4 (=7)	1	1	1	0	T2c	0	1	PR5
8	60	3.0	3 + 4 (=7)	1	1	1	0	T2c	0	2	PR4, PR2
9	71	6.6	3 + 4 (=7)	1	1	1	0	T2c	0	2	PR5, PR3
10	59	5.5	3 + 4 (=7)	1	1	1	0	T3a	1	1	PR5
11	57	5.0	3 + 4 (=7)	1	1	1	0	T3a	1	3	PR4, PR4, PR2
12	49	4.9	3 + 3 (=6)	1	0	0	0	T2c	0	2	PR4, PR4
13	58	6.5	3 + 3 (=6)	1	0	0	0	T2c	0	3	PR4, PR4, PR4
14	60	4.5	3 + 3 (=6)	1	0	1	0	T2a	0	1	PR3
15	66	11.0	3 + 4 (=7)	1	1	1	1	T3a	1	1	PR4
16	52	4.9	3 + 4 (=7)	1	1	0	0	T2c	0	1	PR4
17	63	5.2	3 + 4 (=7)	1	1	1	0	T3a	1	2	PR4, PR4
18	62	6.9	3 + 4 (=7)	1	1	1	1	T2c	0	0	0
19	56	6.4	3 + 3 (=6)	1	0	0	0	T2a	0	1	PR2
20	55	3.4	3 + 3 (=6)	1	0	0	0	T2c	0	1	PR3
21	61	10.3	4 + 5 (=9)	1	1	1	0	T3b	0	1	PR4
22	45	7.2	3 + 3 (=6)	1	0	0	0	T2a	0	1	PR4
23	53	18.5	3 + 4 (=7)	1	1	0	0	T2c	0	1	PR3
24	59	7.3	4 + 3 (=7)	1	1	1	1	T2c	0	1	PR5

TABLE 1. Continued

Patient No.	Age (Years)	PSA (ng/mL)	Gleason Score	Gleason Grade				T Stage	EPE	Number of PIRADS Lesions	PIRADS Score
				G3	G4-FG	G4-Cr	G5				
25	61	5.0	3 + 4 (=7)	1	0	0	0	T2a	0	3	PR4, PR4, PR2
26	54	17.2	3 + 4 (=7)	1	1	1	0	T2c	0	3	PR4, PR5, PR4
27	68	18.7	3 + 4 (=7)	1	1	1	0	T3b	0	2	PR5, PR4
28	63	4.9	3 + 4 (=7)	1	1	1	0	T2c	0	3	PR4, PR4, PR4
29	59	4.0	3 + 4 (=7)	1	1	1	0	T2c	0	1	PR4
30	59	2.8	3 + 3 (=6)	1	1	0	0	T2c	0	2	PR5, PR4
31	66	5.9	3 + 4 (=7)	1	1	1	1	T3a	1	1	PR4
32	66	5.2	3 + 4 (=7)	1	1	1	0	T2c	0	0	0
33	67	8.2	4 + 5 (=9)	1	1	0	0	T2c	0	1	PR4

EPE = extraprostatic extension from pathology report; PSA = prostate specific antigen.

Sites were not required to fit each model in order to maximize participation in this collaborative research project. The site-specific parametric maps were aligned and resampled to the T2-weighted image at the coordinating site to ensure the same resampling code was used.

Digital Reference Object Design

Two separate digital reference objects (DROs) were created for the IVIM and Kurtosis models.[32] Methods for the DRO analysis are detailed in the supplement.

Correlation Analysis

To determine concordance of the quantitative parametric maps submitted, median values were calculated from pathologist-defined region, and a percent difference calculation and a Pearson correlation coefficient were calculated. This was done in both cancerous regions (G3+) and benign atrophy.

In Vivo Data Extraction and Cancer Differentiation

For each parametric map submitted by the sites, median values were calculated from each pathologist-defined region. An empirical receiver operating characteristic (ROC) curve was calculated for each fit to determine the ability of each contrast to differentiate regions of cancer. Two classification tasks were considered: cancer (G3+) vs. benign atrophy, and low-grade (LG = G3) vs. high-grade (HG = G4+). The area under the curve (AUC) served as the metric of interest to assess concordance between site implementations.

Clustergram Analysis

To visually measure group concordance and similarity, clustergrams were created comparing the value within each lesion across all sites who submitted a given fit. Median values were extracted from all lesions greater than 200 voxels in-plane. For each lesion, a SD was calculated quantifying variability across implementations for a given fit. SDs were then sorted and displayed using Matlab (Mathworks Inc, Natick, MA).

Zonal Anatomy

ROIs defining prostate peripheral zone (PZ) and transition zone (TZ) were manually drawn on the T_2-weighted image and verified by a radiologist. Zone masks were used to determine the location of each pathologist annotation. In cases where a lesion crossed the zone boundary, the mode was used to determine the predominant zone. The ROC analysis was repeated within each contrast, plotting cancer vs. benign atrophy stratified by zone.

Index Lesion

The ROC analysis was repeated including only the index lesion to mirror the experimental setup of biopsy-confirmed

TABLE 2. Site Implementation Methods and Submitted Image Format

	Site-Specific Processing										Central Analysis Processing	
	MEADC	IVIM	K	MEADC *b*-Values	IVIM *b*-Values	K *b*-Values	Constraints	Submission Image Format	Code/Tool	Link	Citation	Reorient	Scaling
SI1	X	X	X	<2000	≤200	0, 100, 200, 500, 1000, 2000		NIFTI	In House Matlab		8, 9, 23	X	X
SI2	X	X	X	All	All	All		NRRD	DWModeling SlicerProstate extension	DWIModeling (https://github.com/SlicerProstate/SlicerProstate/tree/master/DWModeling)	26		X
SI3	X	X		All	All			DICOM	IB Diffusion, Osirix Plug-in		8	X	X
SI4	X		X	All		<2000		DICOM	In House Matlab		23		X
SI5	X	X	X	All	≥200	All	X	NIFTI	In House Matlab/ QUAMPER		8, 9, 23	X	X
SI6		X			All			NIFTI	In House Matlab		9, 21	X	X
SI7	X	X	X	≥200	All	≥200	X	MHD	In House Matlab		21, 23, 27	X	X
SI8		X	X		All	All	X	NIFTI	In House Matlab		9, 23	X	X
SI9	X	X	X	All	≥200 (BID), ≤50 (F BID*)	≥200	X	DICOM	In House Matlab		8, 9, 23		X
SI10	X	X		All	≥200		X	NIFTI	ADCmap Osirix Plugin	ADCmap (https://github.com/mribri999/ADCmap)	8	X	X
SI11	X		X	All		All		DICOM	In House imFIAT		10	X	X
SI12	X	X		All	All			DICOM	ImageJ and Custom C ++ Code		8	X	X
SI13	X	X	X	All	All	All		NIFTI	Osirix UMM Diffusion Plugin		8, 10, 32		X
SI14		X			All		X	NIFTI	ADCmap Osirix Plugin	ADCmap (https://github.com/mribri999/ADCmap)	8	X	X

MEADC = mono-exponential apparent diffusion coefficient; IVIM = intra-voxel incoherent motion; K = kurtosis.

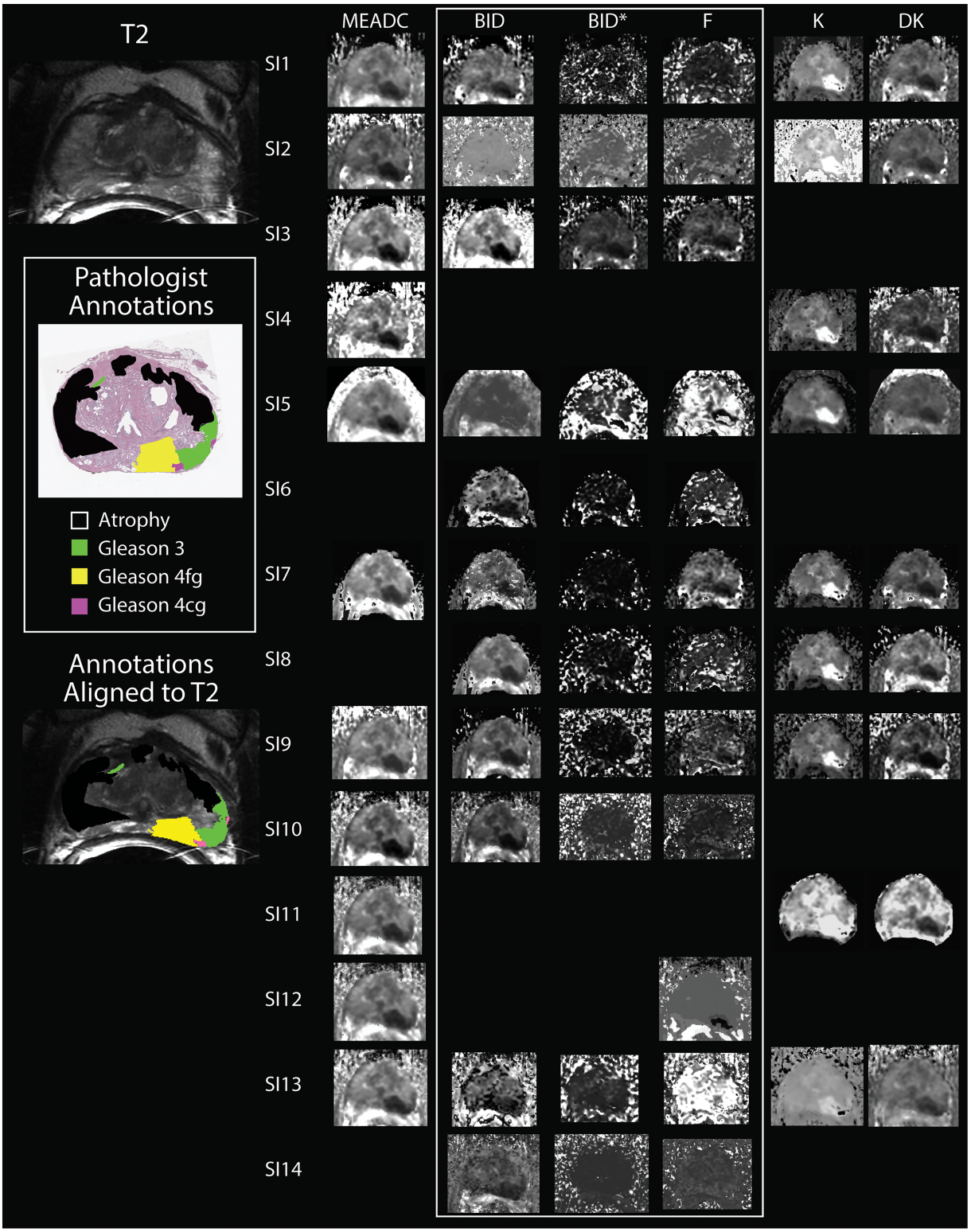

FIGURE 2: A summary of submitted site diffusion-weighted imaging (DWI) fit parameter maps aligned to a pathologist-annotated whole-mount histology slide. Left: The corresponding T2-weighted slice, pathologist annotations in histology space containing a dominant G4 fused gland (G4FG) tumor (yellow) with a secondary G3 region (green) and two small G4 cribriform gland (G4CG) tumors (Pink).[6] Left bottom shows the pathologist annotations aligned in MRI space and overlaid on the T2. Right: Site implementations included mono-exponential apparent diffusion coefficient (MEADC), bi-exponential diffusion (BID), pseudo-diffusion (BID* [$\times 10^{-3}$ mm^2/second]), and perfusion fraction (BID [$\times 10^{-3}$ mm^2/second], BID* [$\times 10^{-3}$ mm^2/second], and F), and kurtosis and kurtosis diffusion (K and DK). Some sites submitted multiple sets of fits, each implementation is separated and treated separately. Relative contrast differences between sites are notable in the MEADC images, but independent of implementation the tumor has decreased diffusion compared to surrounding tissue. Bi-exponential fits showed notable contrast differences between site implementation while kurtosis fits were notably similar.

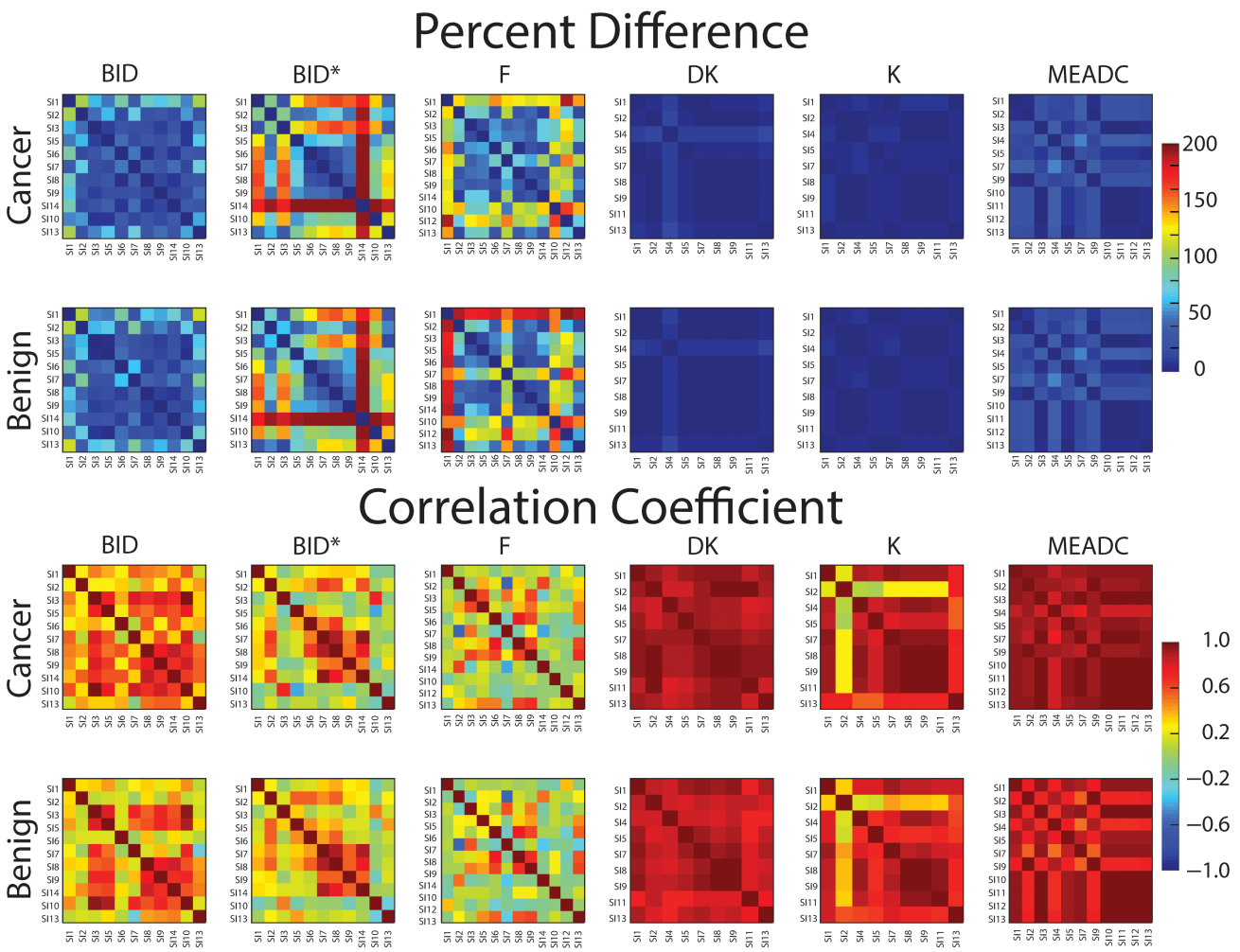

FIGURE 3: Percent difference matrix comparing DWI parameters between site implementation (SI) and between classes of cancer and normal (Top). Pearson correlation coefficient matrices comparing DWI parameters between SI and classes of cancer and non-cancer (Bottom). MEADC, K, and DK show the least percent difference across sites and the highest correlation. Data are shown in Tables S2 and S3 in the Supplemental Material. DWI = diffusion-weighted imaging; MEADC = mono-exponential apparent diffusion coefficient; K = kurtosis; DK = diffusion kurtosis.

radiology studies. The index lesion was defined as the largest in-plane pathologically confirmed cancerous region. A matching number of benign atrophy regions were included in the analysis.

Annotation Extraction Metric

The metric for extracting values from the region of interest was varied and the receiver operating characteristic analysis was repeated. Mean, median, and 10th percentile values were tested (90th percentile for kurtosis fits). A cluster limit of 200 was used for this analysis.

Cluster Limit

The ROC analysis was repeated varying the minimum lesion size required to be included in the analysis. Cluster limits of 100, 200, 300, 400, and 500 voxels were tested. With T_2-weighted voxels being 0.234 mm × 0.234 mm × 3 mm, this corresponded to within slice areas of 5.5, 11.00, 16.5, 22, and 27.5 mm^2 (16.5, 33, 49.5, 66, and 82.5 mm^3). In DWI image space, this was approximately 10, 20, 30, 40, and 50 voxels. Both conditions, cancer vs. benign, and low-grade vs. high-grade were evaluated.

Multi-Pathologist Analysis

The ROC analysis was repeated varying the pathologist annotating the ground truth. This analysis was performed on a subset of 33 slides annotated by five pathologists. A cluster limit of 200 was used and median values were taken from the ROIs. Cancer vs. regions left unlabeled by all five pathologists (unlabeled consensus)[15] was tested in addition to HG vs. LG.

Statistical Comparisons

Basic descriptive statistics of mean and SD values of ROC AUC analysis for each contrast, site implementation, and condition were calculated. To measure differences between implementations, we used a Levene's test applied to the

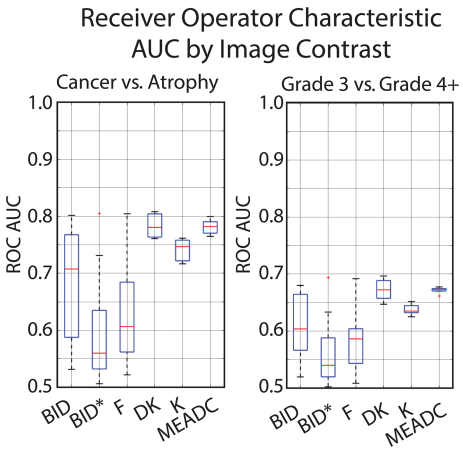

FIGURE 4: Boxplot showing area under the curve receiver operating characteristic (ROC AUC) variability by site implemented fits. Left: Cancer (G3+) vs. benign atrophy. Right: Gleason 3 vs. Gleason 4+. ROC AUC was calculated lesion-wise using the median value in each pathologist annotated region larger than 200 voxels. A tighter boxplot indicates less cancer differentiation variability between site implementations.

SD. To quantify differences in the ROC AUC between contrasts and implantations, we used a linear model with contrast as a covariate, with MEADC as the baseline category, with the sandwich standard error estimates being used to account for lack of homoscedasticity between groups.[33] Pairwise comparisons were performed (consistency and contrast comparisons), with the Tukey's honestly significant difference procedure used to correct *P*-values for multiple comparisons. $P < 0.05$ was considered significant (R-software, v3.6.3; www.r-project.org).

Results

Sample images from each site and software implementation applied to the same slide can be seen in Fig. 2. Sites were not required to fit each model to maximize participation. Submitted maps varied in noise levels and visual interpretability, which was most evident in BID* and F. Universally, regions of cancer showed a decrease in ADC, BID, and DK compared to benign atrophy, and an increase in K.

Correlation Analysis

The correlation analysis revealed similar patterns in percent difference and correlation coefficient in both normal and cancerous ROIs. Mono-exponential ADC, K, and DK were more similar between sites than the IVIM fits (Fig. 3). Value ranges are shown in Tables S1 and S2 in the Supplemental Material. Larger variability of bi-exponential model parameters was also consistent with observations for noise-free DRO (Fig. S1 in the Supplemental Material), although with smaller absolute percent-deviations.

TABLE 3. Statistical Results Comparing Site Implementation ROC AUC Values Between Contrasts and Conditions Cancer vs. Benign Atrophy (CAvBA), and Low-Grade vs. High-Grade Cancer (LGvHG)

CAvBA	BID	BIDS	BIPF	DK	K	MEADC
BID		0.988	0.982	0.054	0.040*	0.011*
BIDS			1.000	0.203	0.162	0.060
BIPF				0.204	0.162	0.058
DK					1.000	0.998
K						1.000
MEADC						
LGvHG	**BID**	**BIDS**	**BIPF**	**DK**	**K**	**MEADC**
BID		0.999	0.898	0.094	0.017	0.005*
BIDS			0.983	0.196	0.044*	0.014*
BIPF				0.512	0.169	0.069
DK					0.988	0.942
K						1.000
MEADC						

ROC AUC = receiver operating characteristic area under the curve; BID = bi-exponential diffusion; DK = diffusion kurtosis; K = kurtosis; MEADC = mono-exponential apparent diffusion coefficient.
*$P < 0.05$.

Variation in Cancer Differentiation

The ROC analysis calculated using a cluster limit of 200 and a median value from each ROI is shown in Fig. 4. Comparing cancer to benign atrophy, MEADC had a median AUC of 0.78, range 0.76–0.80, while BID, BID*, and F had median values of 0.71, 0.56, 0.61 respectively, and ranges of 0.53–0.80, 0.51–0.81, and 0.52–0.80 respectively. Kurtosis models resulted in median AUC of 0.78 and 0.75 for DK and K respectively with ranges of 0.76–0.81 and 0.72–0.76 respectively. Comparing G3 to G4+, MEADC had a median AUC of 0.67, range 0.66–0.68, while BID, BID*, and F had median values of 0.60, 0.54, 0.59 respectively, and ranges of 0.52–0.68, 0.50–0.69, and 0.51–0.69 respectively. Kurtosis models resulted in median AUC of 0.67 and 0.64 for DK and K respectively with ranges of 0.65–0.70 and 0.63–0.65 respectively. Values are summarized in Table S3 in the Supplemental Material. Across all contrasts cancer vs. benign atrophy resulted in a higher AUC than low-grade vs. high-grade.

Statistical Comparisons

Comparing the ROC AUCs between contrasts and conditions (BA vs. CA, and HG vs. LG), MEADC significantly outperformed all other contrasts with the exception of DK and K. DK outperformed all bi-exponential parameters across conditions. Statistical results are detailed in Table 3. Comparing contrast-specific ROC AUC variance between conditions, we found that MEADC and K had significantly less variance between site-specific ROC AUC compared to BID, for both conditions, and trended towards significance for the other bi-exponential parameters, consistent with what can visually be seen in Fig. 4 (Levene's test $P < 0.05$ corrected for multiple comparisons). Statistical results from each comparison are detailed in Table 4.

Zonal Anatomy

The results from the zone analysis are shown in Fig. S3, with data shown in Table S4 in the Supplemental Material. The median AUCs for PZ were 0.81, 0.77, 0.77, 0.73, 0.58, and 0.63 for MEADC, DK, K, BID, BID*, and F respectively. For TZ, median AUCs were 0.84, 0.74, 0.86, 0.72, 0.60, and 0.62 respectively. Summary values with ranges are shown in Table S4 in the Supplemental Material, where in general, the IVIM parameters showed greater variability in range and overall lower performance compared to the kurtosis and mono-exponential parameter maps. Across site implementations, kurtosis performed better in the TZ than the PZ; however, all other parameter maps were roughly equivalent independent of zone.

Index Lesion

The results of the index lesion analysis are shown in Table S5 and Fig. S4 in the Supplemental Material. The median AUCs

TABLE 4. Statistical Results Comparing the Contrast-Specific Variances Between ROC AUC Across Conditions CvBA and LGvHG

CvBA	BID	BIDS	BIPF	DK	K	MEADC
BID		0.177	0.538	0.029	0.505	0.032*
BIDS			0.955	<0.001*	<0.001*	<0.001*
BIPF				<0.001*	<0.001*	<0.001*
DK					<0.001*	0.998
K						<0.001*
MEADC						
LGvHG	**BID**	**BIDS**	**BIPF**	**DK**	**K**	**MEADC**
BID		0.260	0.696	0.015*	0.710	0.012*
BIDS			0.939	<0.001*	<0.001*	<0.001*
BIPF				<0.001*	0.006*	<0.001*
DK					<0.001*	1.000
K						<0.001*
MEADC						

ROC AUC = receiver operating characteristic area under the curve; BID = bi-exponential diffusion; DK = diffusion kurtosis; K = kurtosis; MEADC = mono-exponential apparent diffusion coefficient.

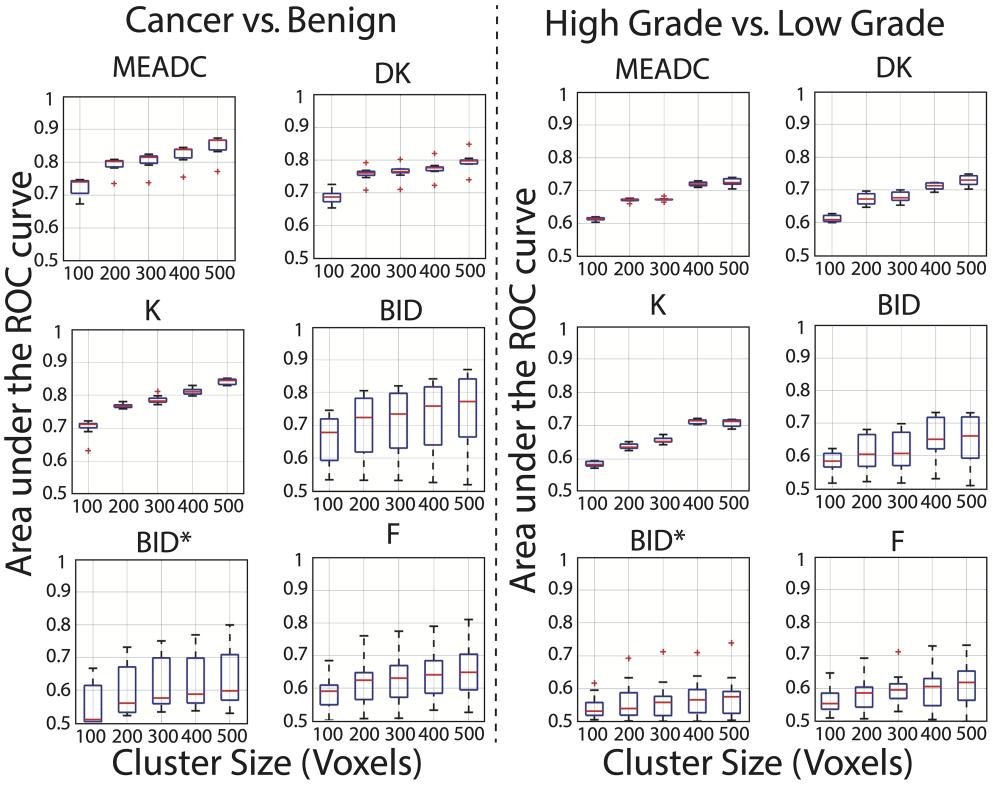

FIGURE 5: Receiver operator characteristic area under the curve (ROC AUC) for all institutions grouped by fit and repeated varying the minimum lesion size included in the analysis. Lesion size limit was varied from 100 voxels to 500 voxels stratifying G3+ vs. benign atrophy (Left) and stratifying G3 from high-grade tumors (Right). There is a trend towards increasing AUC as the cluster limit for inclusion becomes more selective in both cancer vs. benign and low-grade vs. high-grade. Fits that are highly variable between sites remain highly variable independent of cluster limit.

were 0.82, 0.75, 0.58, 0.62, 0.79, and 0.77 for MEADC, DK, K, BID, BID*, and F respectively. While MEADC remains similar to the other experiments, the BID and BID* parameter maps become less variable under this condition, while the DK and K maps become more variable between site implementations.

Clustergram Analysis

The results of the clustergram analysis are shown in Fig. S2 in the Supplemental Material for each of the contrasts. The heat maps shown indicate SD from the mean for each value. More consistency and grouping were seen in the MEADC, K, DK, and BID, with less consistency seen in BID* and F. For MEADC, K, and DK, results indicated that four site implementations were virtually identical in values.

Cluster Limit

The results of the cluster limit analysis are detailed in Fig. 5, with values shown in Table S4 in the Supplemental Material. Across both conditions (high-grade vs. low-grade and cancer vs. benign atrophy) parameter maps AUC increased as minimum cluster to be included was increased from 100 to 500 T_2-resolution voxels in 100 increments. Increases in median AUC went from 0.74 to 0.87, 0.69 to 0.80, 0.71 to 0.85, 0.68 to 0.77, 0.51 to 0.60, and 0.59 to 0.65 for MEADC, DK, K, BID, BID*, and F respectively. Independent of cluster limit, mono-exponential ADC and the kurtosis fit parameters showed smaller ranges of variability between sites. Additionally, the variability between sites in the IVIM parameter maps tended to increase as cluster limit was increased (Fig. 5). Figure S5 in the Supplemental Material shows the number of lesions included in the analysis at each step, indicating that the number of lesions across all conditions decreased by more than half as the cluster limit increased from 100 to 500.

Extraction Metric

The results of varying the extraction metric (median, mean, or 10th percentile) are shown in Table S7 and Fig. S6 in the Supplemental Material. While the median value across all sites is relatively consistent independent of which metric is chosen, the variability between sites is highly dependent on the metric used to extract a value from an ROI.

Multi Pathologist

The results of the multi pathologist experiment are shown in Fig. 6. Varying the ground truth had a substantial effect on both the median AUC as well as the extent of the inter-site-

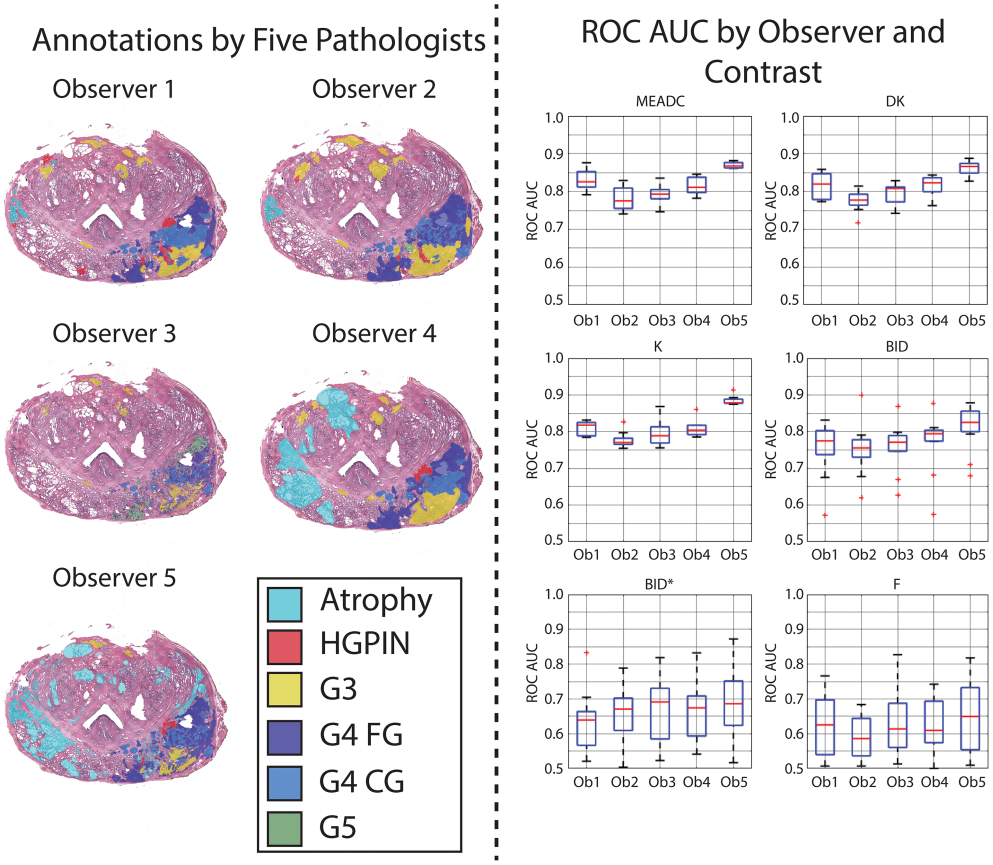

FIGURE 6: Area under the curve receiver operator characteristic (ROC AUC) for cancer vs. regions left unlabeled by all pathologists (unlabeled consensus) annotations varying the pathologist annotating the slides. Left: Sample annotations from all five observers on a representative whole-mount prostate slide. Right: Boxplots showing AUCs varying the image contrast and observer annotating the slides. Median values were extracted from regions of interest (ROIs) greater than 200 voxels in plane.

variability. MEADC, DK, and K values calculated from observer 5's annotations had the greatest AUC and tightest range of AUC between sites at differentiating cancer from regions left unlabeled by all pathologists (unlabeled consensus). Numeric results are shown in Table S8 in the Supplemental Material. The AUCs from BID varied by observer showed consistency between site implementation with the exception of a few outliers, while BID* and F both showed large ranges of AUC regardless of the observer defining ground truth.

Discussion

This study tested inter-site concordance of diffusion-derived parametric maps on the same pathologically validated prostate cancer dataset under a variety of post-processing conditions. In addition to measuring the consistency of values between sites, inter-site variability in performing a diagnostic task was measured. We found that mono-exponential and kurtosis diffusion models were reliably calculated independent of implementation (high correlation between site implementations) and performed well at differentiating prostate cancer (consistently high ROC AUC between implementations). Values calculated from IVIM algorithms varied more between sites (low correlation between site implementations, large range of ROC AUC between sites), although those that applied physical constraints performed better at differentiating prostate cancer (high ROC AUC). In addition, we found that post-processing decisions made at the central analysis site such as ROI sizes and varying the observer defining ground truth, affected the diagnostic potential of all DWI parametric maps, as measured by ROC AUC.

The correlation analysis demonstrated the stability of each fit across sites. The mono-exponential and kurtosis fits had a low percent difference and high correlation coefficient independent of which pair of sites was analyzed. Of the diffusion fits included in this study, six MEADC fit implementations resulted in almost identical maps and values. Kurtosis was likewise consistent across institutions and provided as good or better contrast than ADC with respect to identifying high-grade tumors. The IVIM contrasts were much less similar between implementations, both in normal and cancerous regions.

A number of post-processing parameters were tested. Varying the minimum lesion size included in the analysis caused approximately 0.1 increase in AUC independent of contrast and site implementation. With the exception of this

analysis explicitly testing size, an ROI cluster limit of 200 voxels (11 mm^2) was selected to capture all clinically significant tumors as outlined in PI-RADSV2. For DWI acquisition, typical cluster sizes are only 2–15 acquired voxels, highly susceptible to partial volume at lesion boundaries.[6] However, anatomical boundaries are more clearly seen in T_2 imaging, and thus aligning the annotations with the T_2 images results in a more accurate alignment. This limitation partially explains cluster-size sensitivity of the corresponding lesion AUC analysis for DWI-derived parameters.

Prior work measuring inter-pathologist variability annotating Gleason patterns has been done on tissue microarrays,[34,35] and in whole-mount prostate samples.[15] Interestingly, varying the pathologist performing the gold-standard annotations changed the resulting ROC AUC in the contrasts that varied minimally between site implementations (MEADC, K, and DK). While in most cases observers marked similar areas overall, the size and boundaries of the lesions varied between observers as expected. Partial volume and lesion size limitations resulted in different numbers of ROIs included from each pathologist, which may partially explain the differences in ROC AUC.

The *b*-values used to calculate the IVIM fits varied between implementations. Additionally, some sites chose to apply post-calculation filters such as upper and lower bounds, non-negativity constraints, or other error reduction techniques on their parameter maps to ensure physical values. Those that included physical constraints and other post-fitting filters had the highest ROC AUC (sites 2, 3, 5, 7, and 10). The *b*-values used in the DWI fitting also varied in the implementations of K, DK, and MEADC at different sites. This variability in implementation may explain why some sites MEADC values were consistently higher than others, though the ability to differentiate cancer was not adversely affected with MEADC.

The top performing site implementations for MEADC varied only slightly, so no general recommendations can be made by our conclusions. For the IVIM submissions, in general, the sites that chose to implement constraints on the values calculated performed better due to having less outlier values. The choice of *b*-values included in the fitting did not appear to affect the top performing implementations, as there was a mix of submissions that used all provided *b*-values, and those that limited the *b*-values included in fitting. Regarding kurtosis, the top performing implementations used all *b*-values provided, but generally all performed similarly so no consensus recommendations can be offered beyond constraining values.

Limitations

One major limitation to this study is the relatively small cohort of 33 patients. We felt there was a balance between including a larger cohort and increasing the analysis burden on the external sites. Future studies should increase the *N* and reduce the scope to less fitting models. Regarding the patient cohort, there were wide ranges in the PIRADS scores, Gleason scores, and PSA levels, and there may be potential bias as all the subjects included had a prostatectomy. While this was essential for the pathological validation, our conclusions may not generalize to patients that, for example, undergo radiation treatment rather than surgery. Future studies should determine whether DWI performance between sites varies dependent on PIRADS score, Gleason score at diagnosis, and National Comprehensive Cancer Network risk stratification, as these analyses were beyond the scope of this study. Unfortunately, with our small cohort, we were statistically under powered to split it into smaller subgroups. Additional future studies should determine whether cancer detection varies between repeated pre-surgical quantitative DWI, both in the same scanner, and between vendors.

Anatomical landmarks are more readily apparent on the higher resolution T_2-weighted images and thus using T_2 space for an analysis using aligned pathology is the best practice for creating a reliable ground truth. However, efforts to convert, align, scale, and resample the diffusion maps to the T_2 resolution for comparison to the ground truth pathologist annotations may have introduced minor alignment differences between submissions. These potential sources of error should be mitigated in the future with a consensus on data format and orientation standards for large multicenter research studies.

Conclusion

This study tested inter-site concordance of diffusion-derived parametric maps on the same pathologically validated prostate cancer dataset under a variety of post-processing conditions. We found that conventional diffusion models (mono-exponential and kurtosis fits) had less variability between algorithms in differentiating prostate cancer and performed significantly better overall. More complex IVIM models, in some implementations, also performed well at differentiating prostate cancer, although were more inconsistent between algorithms due to varying constraints and resulted in non-diagnostic AUCs of less than 0.70. We also found that post-processing decisions made at the central analysis site affected the diagnostic potential of all DWI parametric maps, as measured by ROC AUC. These results indicate that a careful selection, explanation of methods, understanding of their effects on the ROC AUC, and code sharing will ease the adoption of advanced quantitative imaging into the clinical setting.

Acknowledgments

This work was supported by Advancing a Healthier Wisconsin, the State of Wisconsin Tax Check off Program for Prostate Cancer Research, National Center for Advancing Translational Sciences: UL1TR001436, TL1TR001437, National Cancer Institute: UG3CA247606, 5P30CA006973 (Imaging Response Assessment Team-IRAT), P01CA085878, P30CA008748, R01CA158079, R01CA160902, R01CA190299,

R01CA218144, R01CA221938, R01CA248192, R01CA249882, R21CA231892, R50CA211270, U01CA140204, U01CA142565, U01CA151261, U01CA154602, U01CA166104, U01CA172320, U01CA176110, U01CA183848, U01CA207091, U01CA211205, U24CA180918, and T.E.Y. is a CPRIT Scholar of Cancer Research.

Conflict of Interest

Author KMS has ownership interest in IQ-AI and financial interest in Imaging Biometrics LLC.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70(1):7-30.
2. 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. Eur Urol 2019;75(3):385-396.
3. Kasivisvanathan V, Rannikko AS, Borghi M, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. N Engl J Med 2018;378(19): 1767-1777.
4. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. Eur Urol 2019;76(3):340-351.
5. Vargas HA, Hotker AM, Goldman DA, et al. Updated prostate imaging reporting and data system (PIRADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: Critical evaluation using whole-mount pathology as standard of reference. Eur Radiol 2016;26(6):1606-1612.
6. Hurrell SL, McGarry SD, Kaczmarowski A, et al. Optimized b-value selection for the discrimination of prostate cancer grades, including the cribriform pattern, using diffusion weighted imaging. J Med Imaging (Bellingham) 2018;5(1):011004.
7. Newitt DC, Malyarenko D, Chenevert TL, et al. Multisite concordance of apparent diffusion coefficient measurements across the NCI Quantitative Imaging Network. J Med Imaging (Bellingham) 2018;5(1):011003.
8. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology 1988;168(2):497-505.
9. 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.
10. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magn Reson Med 2005;53(6):1432-1440.
11. Clarke LP, Nordstrom RJ, Zhang H, et al. The quantitative imaging network: NCI's historical perspective and planned goals. Transl Oncol 2014;7(1):1-4.
12. Farahani K, Kalpathy-Cramer J, Chenevert TL, et al. Computational challenges and collaborative projects in the NCI quantitative imaging network. Tomography 2016;2(4):242-249.
13. Yankeelov TE, Mankoff DA, Schwartz LH, et al. Quantitative imaging in cancer clinical trials. Clin Cancer Res 2016;22(2):284-290.
14. Hadjiiski LM, Nordstrom RJ. Quantitative imaging network: 12 years of accomplishments. Tomography 2020;6(2):55.
15. McGarry SD, Bukowy JD, Iczkowski KA, et al. Radio-pathomic mapping model generated using annotations from five pathologists reliably distinguishes high-grade prostate cancer. J Med Imaging (Bellingham) 2020;7(5):054501.
16. Branch MA, Coleman TF, Li Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM J Sci Comput 1999;21(1):1-23.
17. Constantinides CD, Atalar E, McVeigh ER. Signal-to-noise measurements in magnitude images from NMR phased arrays. Magn Reson Med 1997;38(5):852-857.
18. Du J, Li K, Zhang W, et al. Intravoxel incoherent motion MR imaging: Comparison of diffusion and perfusion characteristics for differential diagnosis of soft tissue tumors. Medicine (Baltimore) 2015;94(25):e1028.
19. McGarry SD, Hurrell SL, Iczkowski KA, et al. Radio-pathomic maps of epithelium and lumen density predict the location of high-grade prostate cancer. Int J Radiat Oncol Biol Phys 2018;101(5):1179-1187.
20. McGarry SD, Bukowy JD, Iczkowski KA, et al. Gleason probability maps: A radiomics tool for mapping prostate cancer likelihood in MRI space. Tomography 2019;5(1):127-134.
21. Dyvorne HA, Galea N, Nevers T, et al. Diffusion-weighted imaging of the liver with multiple b values: Effect of diffusion gradient polarity and breathing acquisition on image quality and intravoxel incoherent motion parameters—A pilot study. Radiology 2013;266(3):920-929.
22. Hectors SJ, Semaan S, Song C, et al. Advanced diffusion-weighted imaging modeling for prostate cancer characterization: Correlation with quantitative histopathologic tumor tissue composition—A hypothesis-generating study. Radiology 2018;286(3):918-928.
23. Jensen JH, Helpern JA. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR Biomed 2010;23(7):698-710.
24. Koh DM, Collins DJ, Orton MR. Intravoxel incoherent motion in body diffusion-weighted MRI: Reality and challenges. AJR Am J Roentgenol 2011;196(6):1351-1361.
25. Kristoffersen A. Optimal estimation of the diffusion coefficient from non-averaged and averaged noisy magnitude data. J Magn Reson 2007;187(2):293-305.
26. Langkilde F, Kobus T, Fedorov A, et al. Evaluation of fitting models for prostate tissue characterization using extended-range b-factor diffusion-weighted imaging. Magn Reson Med 2018;79(4):2346-2358.
27. Lewin M, Fartoux L, Vignaud A, Arrive L, Menu Y, Rosmorduc O. The diffusion-weighted imaging perfusion fraction f is a potential marker of sorafenib treatment in advanced hepatocellular carcinoma: A pilot study. Eur Radiol 2011;21(2):281-290.
28. Lu Y, Jansen JF, Mazaheri Y, Stambuk HE, Koutcher JA, Shukla-Dave A. Extension of the intravoxel incoherent motion model to non-gaussian diffusion in head and neck cancer. J Magn Reson Imaging 2012;36(5):1088-1096.
29. Pang Y, Turkbey B, Bernardo M, et al. Intravoxel incoherent motion MR imaging for prostate cancer: An evaluation of perfusion fraction and diffusion coefficient derived from different b-value combinations. Magn Reson Med 2013;69(2):553-562.
30. Paudyal R, Konar AS, Obuchowski NA, et al. Repeatability of quantitative diffusion-weighted imaging metrics in phantoms, head-and-neck and thyroid cancers: Preliminary findings. Tomography 2019;5(1):15-25.
31. Stejskal EO, Tanner JE. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. J Chem Phys 1965; 42(1):288-292.
32. Malyarenko D, Pang Y, Amouzandeh G, Chenevert T. *Numerical DWI phantoms to optimize accuracy and precision of quantitative parametric maps for non-Gaussian diffusion*. Vol 11313: SPIE; 2020. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11313/113130W/Numerical-DWI-phantoms-to-optimize-accuracy-and-precision-of-quantitative/10.1117/12.2549412.short?SSO=1
33. Huber PJ. The behavior of maximum likelihood estimates under non-standard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; 1967.
34. Nir G, Karimi D, Goldenberg SL, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. JAMA Netw Open 2019;2(3):e190442.
35. Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Med Image Anal 2018;50:167-180.