Table 5: qDESS scan acquisition parameters for the SKM-TEA dataset. RO – readout, PE – phase encode, TE – echo time, TR – repetition time.

Matrix (RO × PE)	416×512
Resolution (mm^2)	0.38×0.31
TE - Echo 1 (ms)	5.7
TE - Echo 2 (ms)	30.1
Number of Echoes	2
TR (ms)	17.9
Flip Angle (°)	20
Parallel Imaging	2×1

A Dataset: Additional Details

A.1 Acquisition Parameters

All shared scan parameters are shown in Table 5. qDESS scans were acquired with 2x1 parallel imaging with multiple receiver coils. Scans were acquired with 15 or 16 coils. Number of slices were varied based on the knee size, ranging from 80 to 88 slices.

A.2 ZIP2 Zero-Padding

K-space data for each scan was zero-padded along the readout dimension to $k_y \times k_z$ matrix size of 512×512. The data was also zero-padded along the slice dimension so as to double the matrix size along this dimension, following the ZIP2 convention. The resulting $k_x \times k_y \times k_z$ matrix size is $512 \times 512 \times (2s)$, where s is the number of slices acquired.

The raw data distributed publicly is zero-padded. When undersampling this data, only the true acquisition region (416×512) should be undersampled to the extent corresponding to the prescribed acceleration. The undersampling masks are generated such that the undersampling occurs only among the data acquisition region – i.e. they do not include zero-padded region.

A.3 Gradient-Warping Correction for Segmentations

Scanner-generated DICOM images undergo vendor- and scanner-specific gradient warping to correct for gradient imperfections (between the nominal and actual magnetic fields) that results in a non-linear spatial deformation of the reconstructed image. As a result, segmentations annotated on the gradient-warped DICOM images did not overlap with appropriate regions in the SENSE-reconstructed images. To correct for this, a regional b-spline registration algorithm (available in DOSMA [12]) was used to register the DICOM images and the corresponding segmentation masks to SENSE reconstructions for each scan. We refer to these as the *gradient-warp-corrected segmentations*. All analysis or end-to-end inter-operation between SENSE reconstructions and tissue segmentations should use the gradient-warp-corrected segmentation. From manual inspection, no such process was required for the coarser bounding box pathology labels.

A.4 Annotator Details

Detection bounding boxes and tissue segmentations were created by two researchers with 3-4 years experience with knee MR image interpretation, supervised by two board-certified musculoskeletal radiologists with 26 and 24 years of experience. All four individuals had semi-structured clinical radiology reports which were used to instruct bounding box labels. Each scan was annotated by a single annotator, such that no one scan had labels from multiple annotators.

For ground-truth cartilage and meniscus segmentations, annotators used both qDESS echoes that provide separate image contrasts to distinguish between the neighboring cartilage and meniscus pixels, as well as additional tissues such as bone, muscle, and joint fluid. Segmentations were performed slice-by-slice in the sagittal plane and volumetric consistency was enforced by correcting segmentations in the axial and coronal planes in the ITK-SNAP software. Every image volume segmentation was quality controlled by the two researchers with 4 and 3 years experience with knee MR image interpretation.

Table 6: Available data, whether it is a model input or output, the tracks to be used with, and corresponding tasks. N/A indicates the data should not be used for benchmarking any task.

Data	Data Type	Raw Data Track	DICOM Track	Task
Raw data (k-space)	Input	✓	✗	Recon
Undersampling masks (per-scan)	Input	✓	✗	Recon
Sensitivity maps	Input	✓	✗	Recon
SENSE reconstruction	Output	✓	✗	Recon, Seg, Detection
DICOM images	Input	✗	✓	Seg, Detection
DICOM T_2 parametric maps	Input	✗	✗	N/A
DICOM segmentations	Output	✗	✓	Seg
Gradient-warp-corrected segmentations	Output	✓	✗	Seg
Pathology bounding boxes	Output	✓	✓	Detection

A.5 Distribution, Hosting, and Maintenance

All public data is distributed under the Stanford University School of Medicine (http://www.stanford.edu/site/terms/) license and the terms listed for the Lower Extremity Radiographs dataset (https://aimi.stanford.edu/lera-lower-extremity-radiographs-2). Data is hosted and maintained by the authors and Microsoft Azure as part of a partnership with the Stanford Center for Artificial Intelligence in Medicine and Imaging. All data and corresponding artifacts (annotations, etc.) will be semantically versioned and available for future use.

Instructions for downloading and using the dataset, versioned data splits and annotations, starter code, and baselines can be found on the dataset GitHub page: https://github.com/StanfordMIMI/skm-tea.

A.6 Usage

Table 6 summarizes the available data and the tracks with which they are compatible.

Raw Data Track: All raw data and artifacts originating from this data (sensitivity maps, SENSE reconstructions, etc.) should be used solely in the Raw Data Track. For segmentation-related analysis in this track, gradient-warp-corrected segmentations should be used in place of DICOM segmentations. For reconstruction, all evaluation results should be reported on data undersampled using the precomputed undersampling masks (at the appropriate acceleration) that are distributed with the raw data. Complex-valued SENSE reconstructed images should be used as ground-truth images for reconstruction evaluation.

DICOM Track: This track pertains to all tasks enabled by DICOM images and their artifacts (e.g. DICOM segmentations). For segmentation-related analysis, DICOM segmentations should be used. DICOM images should not be used for any part of the reconstruction task.

A.7 Author Statement

We, the authors, confirm that we bear all responsibility in case of violation of rights, etc. Public data are distributed under the Stanford University School of Medicine (http://www.stanford.edu/site/terms/) license and the terms listed for the Lower Extremity Radiographs dataset (https://aimi.stanford.edu/lera-lower-extremity-radiographs-2).

B Tissue Subregions

In this appendix, we detail the relevance of subregional tissue analysis in qMRI and the method by which different tissue subregions are extracted.

B.1 Relevant Tissue Subregions

Acute knee injuries and knee degeneration are predominantly localized processes, where specific subregions of the knee undergo more change than others [17]. To quantify local qMRI parameter profiles, specific subregions of relevant tissues must be precisely segmented. Recent work has shown

that subregions in articular cartilage and the meniscus are sensitive to early-onset of degenerative diseases, such as osteoarthritis [18]. Thus, we include subregional analysis of the segmented tissues in our proposed qMRI evaluation framework, as was previously described [33].

B.2 Subregion Extraction

Segmentations for each of the four nominal tissues (patellar cartilage, femoral cartilage, tibial cartilage, and meniscus) served as the base ROIs. These segmentations were then divided into subregions using shape-based priors and center-of-mass (COM) estimates, which are detailed below. Subregions were abbreviated based on their anatomical location. For example, a subregion for the deep cartilage compartment *d*, anterior part of the knee *A*, and the lateral condyle *L* would have the abbreviation *dA-L*. All sub-regions were extracted automatically using DOSMA (v0.1.0).

Patellar cartilage: Patellar cartilage was divided into four subregions: deep-lateral (d-L), deep-medial (d-M), superficial-lateral (s-L), and superficial-medial (s-M). The medial/lateral boundary was determined by the COM of the patellar cartilage segmentation along the sagittal plane. The deep/superficial boundary was computed by finding the midpoint of each column in the patellar cartilage segmentation along the coronal plane.

Femoral cartilage: Femoral cartilage segmentations were divided into a total of 12 regions along the three primary axes: deep/superficial (d/s), anterior/central/posterior(A/C/P), and medial/lateral (M/L). Resulting subregions were named following the nomenclature of these axes; for example, *dA-M* corresponds to the deep anterior cartilage in the medial compartment. A/C/P and M/L compartments were delineated based on COM measurements of the base segmentation. The deep-superficial boundary was computed using the "unrolling technique" [33], where the boundary is determined by the midpoint between radii of concentric cylindrical fits to the femoral cartilage shape.

Tibial cartilage: Tibial cartilage was also divided into 12 subregions: inferior/superior (i/sup), A/C/P, and M/L. A/C/P and M/L compartments were divided based on the COM of the base segmentation. The inferior/superior (i/sup) boundary was determined by finding the COM for each tibial cartilage column in the axial direction [9].

Meniscus: The meniscus was divided into medial/lateral compartments using COM between the two compartments.

C qMRI T_2 Pipeline

In this section, we discuss the recommended pipeline for computing ground truth and predicted T_2 estimates for benchmarks in the Raw Data Track and DICOM Track. For fairness of comparison, this pipeline should be used when comparing results from future benchmarks and methods for the SKM-TEA dataset to results detailed in this work.

Raw Data Track – Reconstruction: Image reconstruction methods were used to generate reconstructions for echo 1 (E1) and echo 2 (E2). Two T_2 parameter maps were computed for each scan, one using the network reconstruction and the other using the ground-truth SENSE reconstruction. Ground-truth gradient-warp-corrected segmentations were used to identify relevant tissues in both parameter maps. Differences in T_2 estimates were computed between regional T_2 estimates from the two parameter maps.

DICOM Track – Segmentation: For each scan, a single T_2 parameter map was computed from the DICOM images. Image segmentatation methods were used to generate predicted masks for relevant tissues. Differences in T_2 estimates were computed between regional T_2 estimates extracted using the ground truth mask and the predicted mask.

D Training Details

In this section, we cover details pertaining to the model architecture, training setup, and compute resources used for the Raw Data Track reconstruction and DICOM Track segmentation benchmarks. Model training and evaluation was conducted in PyTorch. Code to reproduce all results with detailed instructions and evolving benchmarks are available at `https://github.com/StanfordMIMI/skm-tea`.

D.1 Raw Data Track – Reconstruction

Training setup: In this problem, models are trained to reconstruct complex-valued, 2D undersampled axial ($k_y \times k_z$) slices for both qDESS echoes. Scans are undersampled using 2D Poisson Disc undersampling at acceleration factors of R=6x, 8x. Models are trained separately at each acceleration and evaluated on simulated undersampled scans at the respective acceleration. All models were trained with the complex-L1 loss with a fixed random seed.

Data normalization: Input data is normalized by dividing by the 95$^{\text{th}}$ percentile of the magnitude image. Outputs are re-normalized by undoing the scaling operation prior to computation of evaluation metrics. Outputs are not re-normalized during training, prior to computing the training loss.

Undersampling masks: During training, 100,000 undersampling masks are precomputed and cached to ensure all training runs use the same set of undersampling masks. For evaluation, each scan in the test dataset is prescribed a fixed undersampling mask (for the specific acceleration) that is distributed as part of the dataset. All masks are generated such that only the acquisition region is undersampled - i.e. all zero-padded regions in the kspace are not included in the generated undersampling mask.

U-Net baseline: We consider a 2D U-Net model, a popular model for fully convolutional and image-to-image tasks, following the implementation in [24] as one baseline architecture. This U-Net implementation has four max pooling layers with compounding number of channels (32, 64, 128, 256, 512), instance normalization, and leaky-relu activation with slope α=-0.2. All U-Net models were trained for 20 epochs using the Adam optimizer with the following hyperparameters: batch size 24, learning rate η=1e-3, weight decay 1e-4.

Unrolled baseline: We consider the 2D proximal-gradient unrolled network, which has achieved state-of-the-art performance on MRI reconstruction tasks, as another baseline architecture. We follow the unrolled network in [44] with minimal hyperparameter changes. Each unrolled block consists of a shallow residual network with two, 128-channel residual blocks with relu activation. The network consists of a total of eight sequential unrolled blocks with weighted data consistency between each block. All unrolled models were trained for 20 epochs using the Adam optimizer. Due to hardware memory constraints, the same batch size as the U-Net could not be used. Instead a smaller batch size of 4 with 6 gradient accumulation steps was used so that the effective batch size is the same as that of U-Net baselines. A learning rate of η=8e-4 and weight decay of 1e-4 were used.

Hardware: All models were trained on Titan RTX (24GB) or GCP-supported Titan V100 (16GB) GPUs. Models trained on Titan RTX GPUs were constrained so that the total available memory was identical to the Titan V100 GPU (16GB).

D.2 DICOM Track – Segmentation

Training setup: In this problem, models are trained segment patellar cartilage, femoral cartilage, tibial cartilage, and the meniscus from 2D sagittal slices of the DICOM images. All models were trained with a soft Dice loss with a fixed random seed.

Input normalization: All inputs are zero-mean, unit standard deviation normalized by mean and standard deviation values computed over the full volume of the echo. For multi-channel inputs (i.e. E1$\oplus$E2), each channel is normalized independently. Root-sum-of-squares (RSS) inputs are normalized by mean and standard deviation values computed on the RSS volume.

V-Net baseline: Another baseline used a 2D V-Net architecture as implemented in MONAI [32]. This network has 4 pooling layers with doubling number of channels after each pooling step (16, 32, 64, 128, 256). Neither dropout nor early stopping was not used.

U-Net baseline: One baseline used a 2D U-Net architecture as implemented in [14]. This network has 5 pooling layers with doubling number of channels after each pooling step (32, 64, 128, 256, 512, 1024). Convolutional blocks at each encoder and decoder level are composed of two convolutional layers each with relu activations followed by a batch normalization layer.

Default training hyperparameters: Models were trained using the Adam optimizer with initial learning rate η_0=1e-3, minimum learning rate η_{min}=1e-8, and step decay by 0.9x every 2 epochs. A maximum training length of 100 epochs was prescribed with early stopping (δ=1e-5, τ=12 epochs). Training batch size was set to 16 without gradient accumulation. All benchmarks used these hyperparameters unless otherwise mentioned.

Table 7: Performance of U-Net segmentation models measured by standard ML pixel and surface segmentation metrics with absolute T_2 error. Models were trained with echo 1 only (E1), echo 2 only (E2), multi-channel echo1 and echo2 (E1⊕E2), and the root-sum-of-squares (RSS) of both echoes.

Metric	Tissue	E1	E2	E1⊕E2	RSS
DSC	Patellar Cartilage	0.87 (0.097)	0.85 (0.11)	0.87 (0.088)	0.87 (0.10)
	Femoral Cartilage	0.88 (0.035)	0.86 (0.032)	0.88 (0.033)	0.88 (0.032)
	Tibial Cartilage	0.86 (0.036)	0.82 (0.049)	0.86 (0.041)	0.86 (0.037)
	Meniscus	0.84 (0.062)	0.82 (0.047)	0.84 (0.067)	0.84 (0.065)
ASSD (mm)	Patellar Cartilage	0.86 (1.4)	0.58 (0.82)	0.84 (1.6)	1.52 (2.0)
	Femoral Cartilage	0.36 (0.25)	0.36 (0.20)	0.40 (0.60)	0.33 (0.19)
	Tibial Cartilage	0.46 (0.29)	0.45 (0.18)	0.66 (1.2)	0.66 (0.70)
	Meniscus	0.63 (0.42)	0.64 (0.29)	0.91 (1.5)	1.24 (1.4)
Abs T2 Error (ms)	Patellar Cartilage	0.70 (0.59)	0.92 (0.95)	0.71 (0.63)	0.75 (0.59)
	Femoral Cartilage	0.50 (0.36)	0.92 (0.50)	0.53 (0.37)	0.51 (0.36)
	Tibial Cartilage	0.49 (0.47)	0.98 (0.66)	0.51 (0.53)	0.50 (0.52)
	Meniscus	0.60 (0.78)	1.07 (1.0)	0.96 (1.0)	0.65 (0.74)

Hardware: All models were trained on Quadro RTX 8000 (48GB) GPUs, but were constrained to only use 24GB memory.

E Additional Results

E.1 Additional Segmentation Baselines

In addition to the V-Net models, we trained baseline segmentation models with the U-Net architecture. Dice, ASSD, and absolute T_2 error are reported in Table 7. Like V-Net, U-Net models trained on only the second echo (E2) performed considerably worse across all metrics. V-Net models achieved slightly higher performance among standard ML segmentation metrics for patellar cartilage, but had similar performance among T_2 error metrics.

Following the convention of previous segmentation challenges [13, 22, 27], we also compute volumetric overlap error (VOE) and coefficient of variation (CV). A summary of segmentation model performance on these metrics is shown in Table 8. Top performing models – U-Net (E1), U-Net (E1 ⊕ E2), U-Net (RSS) – achieved similar performance and outperformed U-Net (E2) across both metrics.

E.2 T_2 Error

In addition to absolute T_2 error, we measure the standard T_2 error, which can help characterize the bias and variance of the errors in T_2 estimates.

Raw Data Track - Reconstruction: Table 9 summarizes T_2 error for all benchmarked reconstruction models. All models except U-Net (E1+E2) underestimate T_2 across all tissues. While unrolled networks have lower variance in T_2 estimates than U-Net models, the bias of unrolled networks is often larger that that of U-Net models. Thus, unrolled models may be more precise in estimating T_2, but may still need to be optimized to reduce bias in these estimates.

DICOM Track - Segmentation: T_2 error profiles for different segmentation models are summarized in Table 8. Top performing models have low bias for femoral cartilage and tibial cartilage compared to the patellar cartilage and meniscus. Segmentations from all models overestimate T_2 for articular cartilage but underestimate T_2 for the meniscus. Variance in T_2 estimates is also the highest for patellar cartilage and the meniscus. Thus, T_2 estimates in both patellar cartilage and meniscus may be more sensitive to changes in segmentation quality than estimates in femoral cartilage or tibial cartilage.

Table 8: Performance [mean (standard deviation)] of segmentation models on the DICOM Track as measured by volumetric overlap error (VOE), coefficient-of-variation (CV), and T_2 error (in milliseconds). T_2 error values are not bolded as it is unclear if better performance is characterized by smaller bias or lower variance.

Metric	Tissue	V-Net (E1)	V-Net (E2)	V-Net (E1⊕E2)	V-Net (RSS)
VOE	Patellar Cartilage	0.205 (0.108)	0.242 (0.127)	0.193 (0.102)	0.201 (0.114)
	Femoral Cartilage	0.214 (0.0551)	0.237 (0.0496)	0.205 (0.0457)	0.210 (0.0511)
	Tibial Cartilage	0.241 (0.0547)	0.288 (0.0675)	0.238 (0.0516)	0.238 (0.0515)
	Meniscus	0.259 (0.0818)	0.289 (0.071)	0.257 (0.0769)	0.256 (0.0829)
CV	Patellar Cartilage	0.078 (0.0854)	0.077 (0.0558)	0.066 (0.0913)	0.078 (0.0964)
	Femoral Cartilage	0.085 (0.0613)	0.077 (0.059)	0.076 (0.0539)	0.080 (0.0573)
	Tibial Cartilage	0.095 (0.0691)	0.092 (0.084)	0.094 (0.0726)	0.092 (0.0645)
	Meniscus	0.084 (0.0707)	0.081 (0.0661)	0.074 (0.0698)	0.074 (0.0662)
T2 Error (ms)	Patellar Cartilage	0.486 (0.637)	0.873 (0.928)	0.531 (0.934)	0.474 (0.673)
	Femoral Cartilage	0.172 (0.632)	0.772 (0.657)	0.287 (0.532)	0.282 (0.624)
	Tibial Cartilage	0.215 (0.681)	0.805 (0.831)	0.208 (0.669)	0.243 (0.698)
	Meniscus	-0.121 (0.813)	-0.911 (1.01)	-0.666 (0.857)	-0.325 (0.821)
Metric	Tissue	U-Net (E1)	U-Net (E1⊕E2)	U-Net (E2)	U-Net (RSS)
VOE	Patellar Cartilage	0.219 (0.121)	**0.216 (0.110)**	0.245 (0.131)	0.222 (0.118)
	Femoral Cartilage	0.220 (0.054)	0.217 (0.052)	0.246 (0.048)	**0.216 (0.050)**
	Tibial Cartilage	**0.245 (0.053)**	0.246 (0.060)	0.297 (0.067)	0.247 (0.055)
	Meniscus	**0.271 (0.084)**	0.274 (0.089)	0.300 (0.066)	0.275 (0.088)
CV	Patellar Cartilage	0.0641 (0.085)	0.0835 (0.101)	**0.0640 (0.081)**	0.0724 (0.083)
	Femoral Cartilage	0.0754 (0.060)	0.0772 (0.057)	**0.0746 (0.058)**	0.0757 (0.057)
	Tibial Cartilage	0.0811 (0.058)	0.0882 (0.076)	**0.0799 (0.076)**	0.0848 (0.064)
	Meniscus	**0.0716 (0.067)**	0.0803 (0.082)	0.0816 (0.063)	0.0824 (0.076)
T2 Error (ms)	Patellar Cartilage	0.468 (0.790)	0.510 (0.805)	0.484 (1.23)	0.441 (0.854)
	Femoral Cartilage	0.152 (0.596)	0.388 (0.522)	0.837 (0.639)	0.237 (0.585)
	Tibial Cartilage	0.147 (0.666)	0.179 (0.714)	0.887 (0.783)	0.180 (0.698)
	Meniscus	-0.360 (0.923)	-0.954 (1.048)	-1.06 (1.05)	-0.499 (0.847)

Table 9: Performance [mean (standard deviation)] of qDESS reconstruction models with respect to T_2 estimates (in milliseconds) for articular cartilage and the meniscus localized with ground truth segmentations. Typical cartilage T_2 values are 30-40ms, while meniscus T_2 values are 10-15ms.

Acc	Tissue Model	Patellar Cartilage	Femoral Cartilage	Tibial Cartilage	Meniscus
6x	U-Net (E1/E2)	-1.93 (1.98)	-0.228 (1.42)	-1.17 (1.48)	-2.70 (1.35)
	U-Net (E1+E2)	-0.231 (3.46)	1.83 (2.51)	0.201 (1.73)	-1.88 (1.59)
	U-Net (E1⊕E2)	-1.25 (1.96)	-0.838 (1.08)	-1.44 (1.16)	-1.78 (1.03)
	Unrolled (E1/E2)	-0.516 (0.327)	-0.765 (0.283)	-1.03 (0.419)	-2.48 (0.786)
	Unrolled (E1+E2)	-0.555 (0.269)	-0.836 (0.319)	-1.12 (0.444)	-2.52 (0.780)
	Unrolled (E1⊕E2)	-0.639 (2.09)	-2.01 (0.917)	-1.30 (0.650)	-1.25 (0.910)
8x	U-Net (E1/E2)	-3.48 (1.74)	-2.71 (1.38)	-3.21 (1.24)	-3.76 (1.10)
	U-Net (E1+E2)	0.335 (3.38)	2.75 (2.42)	0.153 (1.89)	-2.24 (1.55)
	U-Net (E1⊕E2)	-0.247 (1.68)	-0.889 (1.29)	-1.93 (1.39)	-1.88 (2.45)
	Unrolled (E1/E2)	-0.702 (0.340)	-0.866 (0.415)	-1.20 (0.618)	-2.78 (0.868)
	Unrolled (E1+E2)	-0.971 (0.419)	-0.977 (0.421)	-1.26 (0.590)	-2.86 (0.882)
	Unrolled (E1⊕E2)	-0.482 (0.449)	-0.817 (0.717)	-1.15 (0.931)	-2.69 (0.998)

Table 10: Concordance between standard reconstruction (pSNR, SSIM) and segmentation (DSC, ASSD) metrics versus absolute error in T_2 estimates across different tissue sub-regions as measured by absolute value of the Pearson's correlation coefficient ($|\rho|$). Tissue subregions are defined in Appendix B.2. Values are averaged over all baseline reconstruction and segmentation models.

MainTissue	Base Subregion	ASSD (mm)	DSC	SSIM	pSNR (dB)
Patellar Cartilage	d-L	0.31	0.67	0.16	0.36
	d-M	0.21	0.59	0.23	0.53
	s-L	0.29	0.64	0.25	0.46
	s-M	0.11	0.29	0.30	0.49
Femoral Cartilage	dA-L	0.36	0.68	0.32	0.37
	dA-M	0.15	0.29	0.27	0.36
	dC-L	0.08	0.20	0.08	0.22
	dC-M	0.10	0.21	0.10	0.30
	dP-L	0.03	0.06	0.21	0.37
	dP-M	0.02	0.05	0.25	0.31
	sA-L	0.20	0.44	0.32	0.37
	sA-M	0.12	0.29	0.34	0.42
	sC-L	0.03	0.04	0.27	0.44
	sC-M	0.03	0.08	0.42	0.54
	sP-L	0.02	0.02	0.40	0.46
	sP-M	0.10	0.09	0.35	0.31
Tibial Cartilage	iA-L	0.14	0.31	0.21	0.35
	iA-M	0.01	0.10	0.19	0.34
	iC-L	0.10	0.22	0.23	0.32
	iC-M	0.17	0.41	0.13	0.28
	iP-L	0.00	0.19	0.25	0.38
	iP-M	0.03	0.14	0.24	0.36
	supA-L	0.11	0.35	0.31	0.40
	supA-M	0.00	0.01	0.40	0.48
	supC-L	0.12	0.36	0.41	0.47
	supC-M	0.05	0.16	0.39	0.49
	supP-L	0.03	0.18	0.35	0.41
	supP-M	0.03	0.05	0.36	0.45
Meniscus	L	0.08	0.17	0.20	0.18
	M	0.05	0.25	0.17	0.13

E.3 ML-T_2 Metric Concordance in Tissue Sub-Regions

As mentioned in §B.1, subregional qMRI analysis is a pivotal tool for understanding localized changes in tissue structure. To understand sensitivity of ML metrics to these biomarker-driven metrics, we quantify the concordance between standard ML metrics for image reconstruction and segmentation and subregional absolute T_2 error using Pearson's correlation coefficient. For reconstruction, the global pSNR and SSIM, which are standard metrics computed for image quality, were compared to subregional T_2 estimate errors. For segmentation, subregional T_2 estimate errors were compared to segmentation metrics computed on the corresponding parent tissue structure. For example, T_2 error in the deep-superficial-lateral compartment of femoral cartilage was compared to DSC and ASSD of femoral cartilage. Tissue subregions are computed using methods detailed in §B.2. Table 10 summarizes the results.

SSIM and pSNR have very weak correlation with subregional T_2 error ($|\rho| \leq 0.42, 0.54$ respectively). Because these metrics measure the global image quality, they are likely not sensitive to local regional changes in image quality, and thus, may be even less sensitive to subregional T_2 error. ASSD is also very weakly correlated with T_2 error across all subregions ($|\rho| \leq 0.36$). Because ASSD is a surface metric, it may not capture the changes in volumetric subregions of the image over which these estimates are computed. DSC is a volumetric metric, which may explain why the average correlation is stronger with DSC than with ASSD. However, despite its volumetric nature, DSC is

weakly correlated with T_2 error among almost all subregions. In particular, it is weakly correlated along the subregions in the medial compartment, which are the most likely to undergo degeneration in chronic degenerative diseases such as osteoarthritis [31, 35].

This may further warrant the need for direct biomarker-based evaluation metrics that the qMRI evaluation framework enables.