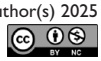
Training With Local Data Remains Important for Deep Learning MRI Prostate Cancer Detection

Canadian Association of Radiologists Journal
2026, Vol. 77(2) 338–347



DOI: 10.1177/08465371251367620
journals.sagepub.com/home/caj

Shawn G. Carere[1,2], John Jewell[2], Paola V. Nasute Fauerbach[1], David B. Emerson[2], Antonio Finelli[3], Sangeet Ghai[4], and Masoom A. Haider[1,4]

Abstract

Objectives: Domain shift has been shown to have a major detrimental effect on AI model performance however prior studies on domain shift for MRI prostate cancer segmentation have been limited to small, or heterogenous cohorts. Our objective was to assess whether prostate cancer segmentation models trained on local MRI data continue to outperform those trained on external data with cohorts exceeding 1000.

Methods: We simulated a multi-institutional consortium using the public PICAI dataset (PICAI-TRAIN: *1241 exams*, PICAI-TEST: *259*) and a local dataset (LOCAL-TRAIN: *1400 exams*, LOCAL-TEST: *308*). IRB approval was obtained and consent waived. We compared nnUNet-v2 models trained on the combined data (CENTRAL-TRAIN) and separately on PICAI-TRAIN and LOCAL-TRAIN. Accuracy was evaluated using the open source PICAI Score on LOCAL-TEST. Significance was tested using bootstrapping.

Results: Just 22% (309/1400) of LOCAL-TRAIN exams would be sufficient to match the performance of a model trained on PICAI-TRAIN. The CENTRAL-TRAIN performance was similar to LOCAL-TRAIN performance, with PICAI Scores [95% CI] of 65 [58-71] and 66 [60-72], respectively. Both of these models exceeded the model trained on PICAI-TRAIN alone which had a score of 58 [51-64] ($P < .002$). Reducing training set size did not alter these relative trends.

Conclusion: Domain shift limits MRI prostate cancer segmentation performance even when training with over 1000 exams from 3 external institutions. Use of local data is paramount at these scales.

Résumé

Objectifs : Il a été démontré que la divergence entre domaines nuit considérablement à la performance des modèles d'intelligence artificielle. Toutefois, les études antérieures portant sur cette problématique dans le cadre de la segmentation du cancer de la prostate à l'IRM se limitaient à des cohortes de petite taille ou hétérogènes. L'objectif de cette étude était d'évaluer si des modèles de segmentation du cancer de la prostate entraînés à partir de données d'IRM internes conservent une performance supérieure à ceux entraînés sur des données externes, même lorsque les cohortes dépassent 1 000 examens.

Méthodes : Nous avons simulé un consortium multi-institutionnel en utilisant la base de données publique PICAI (PICAI-TRAIN : *1 241 examens*, PICAI-TEST : *259*) et une base de données locale (LOCAL-TRAIN : *1 400 examens*, LOCAL-TEST : *308*). L'approbation du comité d'éthique a été obtenue et le consentement a été levé. Nous avons comparé des modèles nnUNet-v2 entraînés sur l'ensemble combiné (CENTRAL-TRAIN) ainsi que séparément sur les ensembles PICAI-TRAIN et LOCAL-TRAIN. La performance des modèles a été évaluée à l'aide du score PICAI (code source libre) sur l'ensemble LOCAL-TEST. La signification statistique a été déterminée au moyen de la méthode bootstrap.

Résultats : Il suffirait de 22 % (309/1 400) des examens de l'ensemble LOCAL-TRAIN pour égaler la performance d'un modèle entraîné exclusivement sur l'ensemble PICAI-TRAIN. Les performances des modèles CENTRAL-TRAIN et LOCAL-TRAIN étaient comparables : leurs scores PICAI [IC à 95 %] étaient de 65 [58–71] et 66 [60–72], respectivement. Ces deux modèles surpassaient celui entraîné uniquement sur PICAI-TRAIN, dont le score était de 58 [51–64] ($P < 0{,}002$). La réduction de la taille de l'ensemble d'entraînement n'a pas modifié ces tendances relatives.

Conclusion : La divergence entre domaines limite la performance des modèles de segmentation du cancer de la prostate à l'IRM, même lorsque l'entraînement repose sur plus de 1 000 examens provenant de trois établissements externes. L'intégration de données locales demeure essentielle à cette échelle.

Keywords

MRI, prostate cancer, artificial intelligence, deep learning, data sharing, segmentation

Introduction

Magnetic Resonance Imaging (MRI) is the recommended imaging test for clinically significant prostate cancer (csPCa) detection.[1,2] However, MRI requires interpretation by highly trained radiologists, with significant variation in performance.[3] Artificial intelligence (AI) may help in improving radiologist performance and consistency.[4] The ideal would be a single AI model that can identify and localize suspicious lesions with expert-level performance despite inherent variations such as patient populations, scanner types, and MRI acquisition protocols.

A long-standing belief is that training a single model on as much multisite data as possible will, given enough samples, eventually achieve expert-level performance.[5] While this assumption generally holds true, how much data is sufficient in the prostate MRI domain remains uncertain.[5] Moreover, emerging research in bi-parametric MRI (bpMRI) csPCa segmentation suggests that performance gains plateau when extending data beyond a single institution using data sharing, and that such models struggle to generalize to institutions that did not contribute data.[6-9] This decline is largely attributed to domain shift—where subtle differences in imaging protocols, scanner types, and patient demographics degrade model performance on unseen data.[6-10] AI model development assumes that the training set is representative of all technical and population variations for sufficient generalizability. Covering all variations is a major challenge and tuning with local data may be more efficient than trying to create such a diverse data set.

Given the availability of models based on multi-institutional data, this raises the question of whether training with local data is still necessary. Previous studies exploring this question for csPCa segmentation have been limited by small datasets, custom models, inconsistent metrics, and mixtures of endorectal and non-endorectal coil MRI protocols the latter being an obvious cause of domain shift.[6-9] Given the availability of an established open-source model (nnUNet-v2)[11] a sizable open-source multi-institutional dataset and metric library shared by the PICAI group,[12] we can study the effect of local data on performance with contemporary prostate MRI at a larger scale. The purpose of this study was to assess, using standardized metrics and methodologies, whether csPCa segmentation models trained on local MRI data continue to outperform those trained on external data at dataset sizes exceeding 1000 exams.

Materials and Methods

Study Design

This retrospective study employed 2 bpMRI datasets to compare local versus external data for csPCa segmentation: PICAI and LOCAL. Exclusion criteria are detailed in Figure 2. The study was approved by the institutional review board and informed consent was waived.

The LOCAL dataset represented a single local institution and the PICAI dataset represented an external multi-institutional cohort. Each dataset was split into training and testing cohorts resulting in 4 subset datasets (LOCAL-TRAIN, LOCAL-TEST, PICAI-TRAIN, PICAI-TEST). We defined a 5th dataset (CENTRAL-TRAIN) as the union of LOCAL-TRAIN and PICAI-TRAIN (Figure 1). Using identical model architectures and training procedures, we trained 3 models: one on local data (LOCAL-TRAIN), one on the external PICAI data (PICAI-TRAIN), and one on the combined local and external data (CENTRAL-TRAIN). This process was repeated with reduced dataset sizes (40%, 20%, and 10% of the original PICAI and LOCAL training sets), resulting in 12 total models (3 per subset size). To further contextualize results, we compared performance to the nnUNet-v2 model provided from the PICAI competition (PICAI-PROVIDED)—a semi-supervised model that ranked 11th of 26 in the open phase and 8th of 12 in the closed phase. We reported results using both the official pretrained weights[13] on the full publicly available PICAI dataset (1500 exams) and our own version trained on the PICAI-TRAIN cohort (subset of 1241 exams) using the authors' publicly released code.[14] All models were evaluated on the LOCAL-TEST dataset. Secondary testing on the PICAI-TEST set is provided in Supplemental Material.

Participants

The PICAI dataset comprised 1500 publicly available prostate MRI exams from 1476 men suspected of having csPCa based on elevated prostate-specific antigen (PSA) levels or abnormal digital rectal examination. These exams were acquired between 2012 and 2021 across 3 centers in the Netherlands.[12] For context PICAI only publicly released a portion of their full dataset which included an additional 8707 exams from 4 institutions. The LOCAL dataset included 2771 consecutive exams acquired between 2012 and 2020 from treatment-naive patients referred for prostate MRI at our institution due to elevated risk. After applying the exclusion

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada
[2]Vector Institute, Toronto, ON, Canada
[3]Division of Urology, Department of Surgical Oncology, Princess Margaret Hospital, University of Toronto, ON, Canada
[4]Joint Department of Medical Imaging, University of Toronto, Princess Margaret Hospital and Sinai Health System, ON, Canada

Corresponding Author:
Masoom A. Haider, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, 600 University Avenue, Toronto, ON M5G 1X5, Canada.
Email: m.haider@utoronto.ca

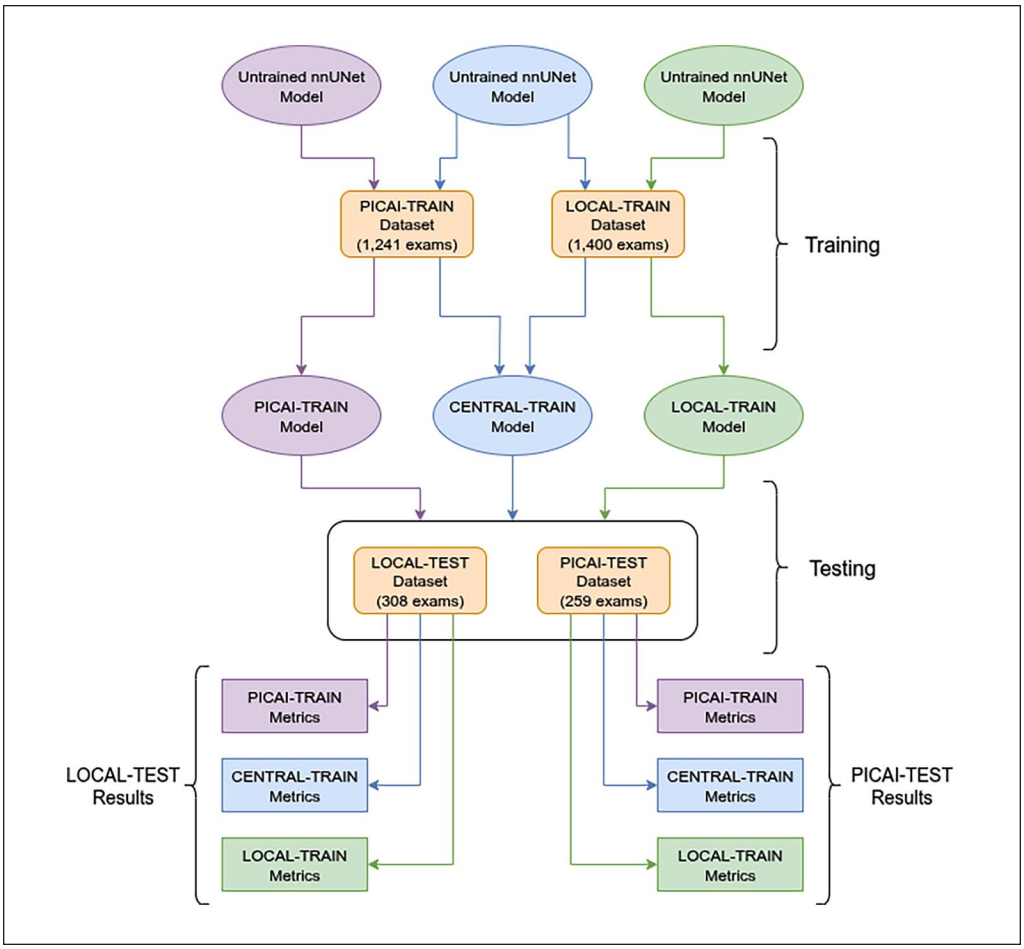


Figure 1. Dataset combinations used for training and naming of the corresponding models. Datasets are indicated in rounded corner rectangles, models in ellipses, and metrics in rectangles. The primary test dataset for all models was the LOCAL-TEST dataset. Additional results on the PICAI-TEST dataset are in Supplemental Material. The training cohorts consisted of the LOCAL training cohort, the PICAI training cohort of the combination of both termed the CENTRAL training cohort. This same experimental design was repeated using 40%, 20%, and 10% of the PICAI-TRAIN and LOCAL-TRAIN datasets. 100% of the LOCAL and PICAI TEST datasets were always used during evaluation.

criteria (Figure 2) to the LOCAL institution data, 1708 MRI examinations from 1514 patients were included.

Image Acquisition

All exams across both datasets consisted of an axial T2-weighted (T2W) image, a high b-value image (BHIGH) and an apparent diffusion coefficient (ADC) map derived from the diffusion weighted images. All images in the LOCAL cohort were acquired on Siemens 3T systems (Avanto, Vario, Skyra, Siemens Healthineers, Erlangen, Germany). Protocols varied over time however the following common protocol parameters were used: a surface phased array coil without an endorectal coil, 3 mm slice thickness, field of view from 16 to 20 cm, and b values ranging from 0-100 s/mm^2 for the lowest b-value to 900-1600 s/mm^2 for the highest b-value. For BHIGH, extrapolated b-value maps were calculated and used at $b = 1600$ s/mm^2.

Labels

For both the LOCAL and PICAI datasets, available labels were the segmentation masks, PIRADS v2.1 scores assigned by radiologists[15] and biopsy-confirmed pathology results. Patients with multiple exams were treated independently. In the LOCAL dataset, segmentation masks were drawn using ITK-SNAP v4.0.2[16] by an abdominal radiologist under the supervision of a uroradiologist with more than 25 years of experience reporting more than 1000 exams with pathology correlation. For the PICAI dataset training we used the official annotations provided by Saha et al.[12] At the voxel-level these included 220 annotations completed by a trained investigator or expert radiologist, while the annotations for the remaining 205 csPCa-positive exams were AI-generated using a model by the PICAI group.[17]

Both the PICAI and LOCAL datasets were randomly split into training (≈80%) and testing (≈20%) cohorts resulting in

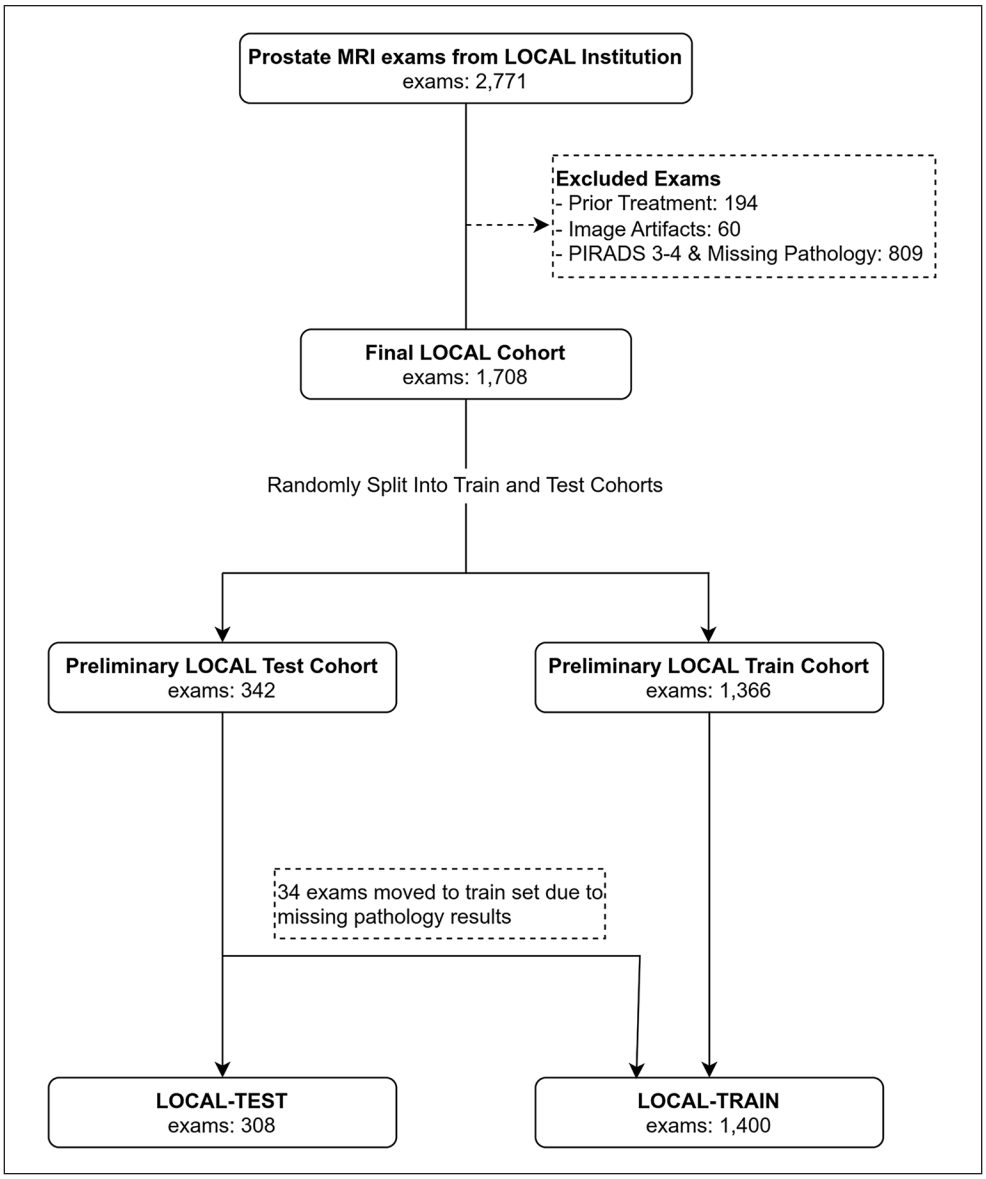


Figure 2. Flow diagrams showing exam selection from LOCAL Institution for LOCAL Dataset.
Note. PIRADS = Prostate Imaging Reporting and Data System.

Table 1. Exam-Level Annotations for LOCAL and PICAI Datasets.

Cohort	csPCa-Positive	csPCa-Negative
LOCAL-TRAIN	PIRADS 3-5 & GG >1 OR PIRADS 5 & any GG OR PIRADS 5 & no pathology	PIRADS 1-4 & GG <2 OR PIRADS 1-2 & no pathology OR no PIRADS & GG <2
LOCAL-TEST	PIRADS 3-5 & GG >1	PIRADS 1-5 & GG <2
PICAI-TRAIN	GG >1 & Lesion segmented by radiologist or AI	GG <2 OR PIRADS 1-2 & no pathology
PICAI-TEST	GG >1 & Lesion segmented by radiologist	GG <2

Note. GG = International Society of Urological Pathology Grade Group; PIRADS = Prostate Imaging Reporting and Data System; csPCa = clinically significant prostate cancer.

Table 2. Dataset Characteristics.

	LOCAL			PICAI		
Dataset	Total	Train	Test	Total	Train	Test
MRI Exams (N, % [$\frac{N}{Total}$])	1708 (100)	1400 (82)	308 (18)	1500 (100)	1241 (83)	259 (17)
MRI Exams with csPCa (N, %)	555 (32)	443 (32)	112 (36)	425 (28)	362 (29)	63 (24)
Lesions (N)	758	606	152	437	369	68
Patients (N)	1514	1226	288	1476	1221	255
Age (y, 95% CI)	65 (49-79)	64 (49-79)	64 (49-79)	66 (52-80)	66 (52-80)	65 (52-78)
PSA level (ng/mL, 95% CI)	9.0 (0-23)	9.2 (0-24)	8.1 (0-19)	11.9 (0-41)	11.8 (0-42)	12.5 (0-37)
Maximum exam GG (N, %)	N ($\frac{N}{1708}$)	N ($\frac{N}{1400}$)	N ($\frac{N}{308}$)	N ($\frac{N}{1500}$)	N ($\frac{N}{1241}$)	N ($\frac{N}{259}$)
N/A	682 (40)	682 (49)	0 (0)	499 (33)	499 (40)	0 (0)
0	323 (19)	239 (17)	84 (27)	348 (23)	225 (18)	123 (47)
1	361 (21)	249 (18)	112 (36)	228 (15)	155 (12)	73 (28)
2	268 (16)	177 (13)	91 (30)	234 (16)	200 (16)	34 (13)
3	47 (3)	34 (2)	13 (4)	99 (7)	82 (7)	17 (7)
4	13 (1)	9 (1)	4 (1)	40 (3)	33 (3)	7 (3)
5	14 (1)	10 (1)	4 (1)	52 (3)	47 (4)	5 (2)
Maximum exam PIRADS (N, %)	N ($\frac{N}{1708}$)	N ($\frac{N}{1400}$)	N ($\frac{N}{308}$)			
N/A	122 (7)	122 (9)	0 (0)	–	–	–
1	14 (1)	13 (1)	1 (0)	–	–	–
2	665 (39)	615 (44)	50 (16)	–	–	–
3	164 (10)	106 (8)	58 (19)	–	–	–
4	401 (23)	260 (19)	141 (46)	–	–	–
5	342 (20)	284 (20)	58 (19)	–	–	–

Note. Continuous parameters are reported as means with 95% confidence intervals (CI) in parentheses. csPCa = clinically significant prostate cancer; PSA = Prostate Specific Antigen; GG = International Society of Urological Pathology Grade Group; N/A = Not Applicable (ie, The exam was missing this information); PIRADS = Prostate Imaging Reporting and Data System.

4 dataset subsets (PICAI-TRAIN, PICAI-TEST, LOCAL-TRAIN, LOCAL-TEST). There were 171 and 23 patients with multiple exams in LOCAL and PICAI respectively. To ensure independence between subsets, patients with multiple MRI exams were assigned entirely to either training or testing. This was consistent with the approach taken by the PICAI group.[12] It is worth noting that individual lesion PIRADS scores were not publicly released for the PICAI dataset. As a result slightly different label criteria were used in training and testing. A summary of the annotation criteria used for training and testing is provided in Table 1. Pathology labels were available for all exams in the PICAI dataset, some LOCAL training data and all LOCAL test data. A detailed overview of the datasets and labeling criteria are in Table 2. Exams with PIRADS scores 1, 2, or 5 that lacked pathology were included only in training with 1 and 2 considered negative and 5 considered positive. PIRADS 3 and 4 exams without pathology labels were not used. For LOCAL-TEST, only pathologically proven cases were included with Grade Group (GG) greater than 1 being considered positive for csPCa. For model training, standard 5-fold cross-validation was used.

Model

We used nnUNet-v2 to provide a well understood baseline model for comparison.[11] This choice promotes reproducibility and aligns with tools commonly used in clinical research settings. The "3d_fullres" configuration was used, with 3 volumes (T2W, BHIGH, ADC) and outputting a continuous csPCa heat-map. For each exam, the BHIGH image and ADC map were resampled to match the resolution of the T2W image as per PICAI preprocessing methods.[12] We used the nnUNet-v2 plans file generated from LOCAL-TRAIN, and hence the same model architecture and training settings, for all experiments, updating only the stored dataset statistics used by nnUNet-v2 for normalization prior to training. Training followed nnUNet-v2 defaults, except we replaced the loss function with an equal-weighted sum of Focal Loss and Binary Cross Entropy Loss, as this was used in training the PICAI-PROVIDED model.[12] Additional details on data preprocessing, model architecture, and model training are included in the Supplemental Material.

Table 3. PICAI Scores for Each Full Dataset Trained Model on LOCAL-TEST Dataset.

Model	Train set	Exam-level AUROC (%)	PICAI Score (%)	Δ PICAI Score	*P*-value
PICAI-PROVIDED nnUNet-v2 (Publicly Available Weights)	Full PICAI Dataset	75.6 (70-81)	59.1 (52-65)	−7.0	<.001
PICAI-PROVIDED nnUNet-v2 (Publicly Available Code)	PICAI-TRAIN	73.3 (67-79)	56.2 (50-63)	−9.8	<.001
PICAI-TRAIN	PICAI-TRAIN	74.2 (68-80)	58.0 (51-64)	−8.1	<.001
LOCAL-TRAIN	LOCAL-TRAIN	79.9 (74-85)	66.1 (60-72)	—	—
CENTRAL-TRAIN	PICAI-TRAIN & LOCAL-TRAIN	78.3 (73-84)	64.7 (58-71)	−1.4	.090

Note. The PICAI Score with 95% confidence intervals in parentheses for each model on the LOCAL holdout test dataset (LOCAL-TEST) are shown. The Δ PICAI Score indicates the difference compared to the LOCAL model with the final column providing the *P*-value for whether or not the PICAI Score difference is significant. The area under receiver operating characteristic curve (AUROC) is also shown (along with 95% confidence intervals) which is a patient level performance metric.

Evaluation Metrics and Statistics

Previous studies have used various voxel and exam level metrics to evaluate performance,[5-10] but inconsistent post-processing methods limit comparability.[18] To address this, we adopted the standardized evaluation pipeline from the PICAI competition,[12] which emphasizes lesion-level assessment over voxel-level accuracy (see Supplemental Material). As noted in the competition guidelines, the PICAI Score "mandates coupling the tasks of lesion detection and patient diagnosis to promote interpretability and disincentivize AI solutions that produce inconsistent outputs."[19] The PICAI Score was the primary metric and is the mean of 2 submetrics, an area under the receiver operating characteristic (AUROC) based on the model's overall confidence that an exam contains csPCa and an average precision (AP) based on the model's csPCa confidence for individual predicted lesions (for which there may be multiple in a single exam). We report both of these submetrics in our results. For evaluation, we selected the best-performing model checkpoint for each fold based on validation performance and used these checkpoints to generate predictions on the test sets. For each test sample, an ensembled prediction was generated by averaging the outputs from all folds. These ensembled predictions were used to compute the PICAI Score, exam-level AUROC, and lesion-level AP. Reported values for these metrics represent the mean performance over 1000 bootstrap iterations. We used bootstrapping to estimate 95% confidence intervals and compute *P*-values for model comparisons using the PICAI Score. A *P*-value of $<.05$ was considered significant. Additional details are provided in the Supplemental Material.

Results

The 1500 publicly available exams from the PICAI dataset were included in this study. For the LOCAL dataset, 1063 exams were not used in this study based on the exclusion criteria with 1708 exams from 1514 patients remaining. Both the PICAI and LOCAL datasets were partitioned into independent training and testing sets (PICAI-TRAIN, PICAI-TEST, LOCAL-TRAIN, LOCAL-TEST). Dataset characteristics for all partitions are presented in Table 2.

Model Performance on Full Datasets

We evaluated the LOCAL-TRAIN, PICAI-TRAIN, CENTRAL-TRAIN, and the PICAI-PROVIDED models on the LOCAL-TEST cohort (Table 3). Both LOCAL-TRAIN (66, $P<.001$) and CENTRAL-TRAIN (64, $P<.002$) models significantly outperformed PICAI-TRAIN (58). The performance of the CENTRAL-TRAIN model closely matched the LOCAL-TRAIN model. The best results were obtained from models that included LOCAL data in training with the CENTRAL-TRAIN and LOCAL-TRAIN models improving the PICAI Score by 6.7 and 8.1 points, respectively, over PICAI-TRAIN model which had no local data. For context, the dynamic range of PICAI Scores across the PICAI competition leaderboard is 16 points in the open phase and 9.6 in the final ranking (excluding 2 outlier scores). These results underscore the substantial impact of domain shift on model generalization. For completeness, we included the performance of the models on a hold out PICAI-TEST dataset in the Supplemental Material. When assessing performance on the PICAI-TEST data the trends were similar however the performance of the LOCAL-TRAIN model on the PICAI-TEST data was only 1.6% worse (Supplemental Material) than the PICAI-TRAIN model while the spread for the LOCAL-TEST data was 8.1% (Table 3).

Model Performance on Partial Datasets

To estimate how many local exams are needed to match the performance of a model trained on the full PICAI training set and explore the effect of dataset size on model performance, we retrained the LOCAL-TRAIN model on progressively smaller subsets of its data. The PICAI-TRAIN and CENTRAL-TRAIN models were similarly retrained on subsets of their data to assess how performance scales with training volume. PICAI Scores for all 12 models (3 models $\times$ 4 dataset sizes) on the LOCAL-TEST dataset are shown in

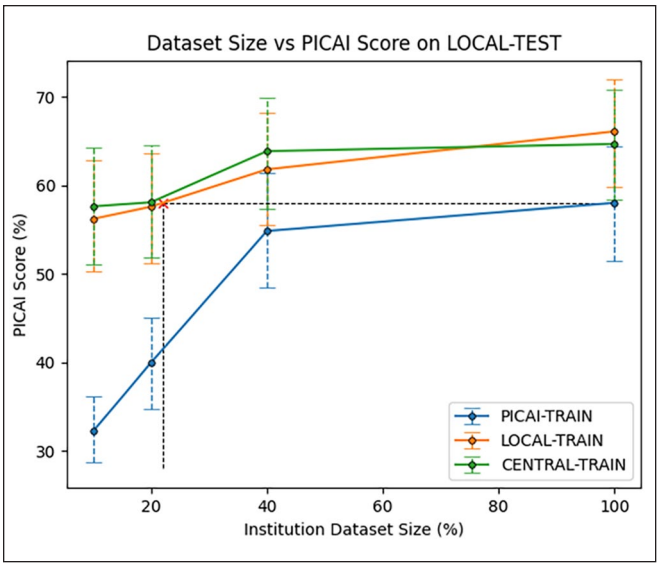


Figure 3. Institutional Dataset Size versus PICAI Score on LOCAL-TEST Dataset. The *x*-axis represents the percentage of training exams used for model training, with the percentage applied separately to the total number of exams in each of PICAI-TRAIN and LOCAL-TRAIN. The *y*-axis shows the PICAI Score on LOCAL-TEST computed using bootstrapping with error bars indicating 95% confidence intervals. The red *x* ($x = 22.1\%$, $y = 58.0\%$) represents the point along the curve where LOCAL-TEST matches the performance of the PICAI-TRAIN model trained on 100% of the PICAI training set.

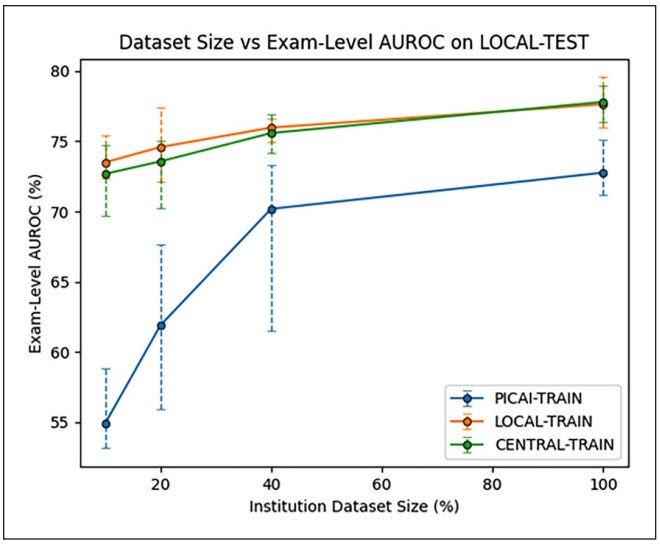


Figure 4. Institutional Dataset Size versus Exam-Level AUROC on LOCAL-TEST Dataset. The *x*-axis represents the percentage of training exams used for model training, with the percentage applied separately to the total number of exams in each of PICAI-TRAIN and LOCAL-TRAIN. The *y*-axis shows the exam-level area under the receiver operating characteristic (AUROC) on LOCAL-TEST computed using bootstrapping with error bars indicating 95% confidence intervals.

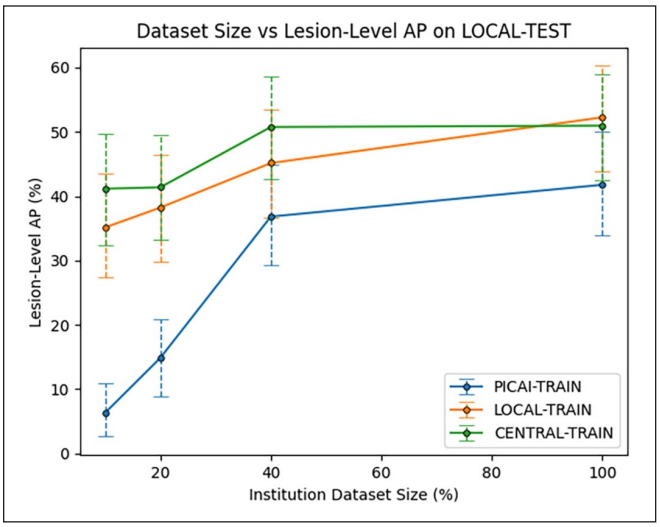


Figure 5. Institutional Dataset Size versus Lesion-Level AP on LOCAL-TEST Dataset. The *x*-axis represents the percentage of training exams used for model training, with the percentage applied separately to the total number of exams in each of PICAI-TRAIN and LOCAL-TRAIN. The *y*-axis shows the lesion-level average precision (AP) on LOCAL-TEST computed using bootstrapping with error bars indicating 95% confidence intervals.

Figure 3. Exam-Level AUROC and Lesion-Level AP are shown in Figures 4 and 5, respectively, with a summary of all metrics across all subsets in Table 4. Results for the PICAI-TEST set are provided in the Supplemental Material.

CENTRAL-TRAIN and LOCAL-TRAIN performance were consistently better than PICAI-TRAIN performance on the LOCAL-TEST dataset across all dataset sizes ($P < .006$). To estimate the number of local training exams required to match the full PICAI-TRAIN performance (without a reduction in training data), we linearly interpolated the LOCAL-TRAIN performance across dataset sizes and identified the intersection with the PICAI-TRAIN score at 100% data usage. As shown in Figure 3, this occurred at 22% of the original LOCAL-TRAIN training set, corresponding to 309 exams. Finally, we note that the CENTRAL-TRAIN model achieved the highest performance across all dataset sizes; however, its improvement over LOCAL-TRAIN was small and not statistically significant despite being trained on nearly twice the number of exams.

Discussion

Using a nnUNet-v2 architecture trained on varying amounts and combinations of local and external data, we investigated the relative value of local versus external exams for training AI models to segment and detect csPCa in bpMRI. Models trained with local data consistently outperformed those tested on an external cohort, even when using substantially less data. We found that just 309 local exams would be sufficient to match the local performance of training with a much larger

Table 4. Results on the LOCAL-TEST Dataset With Partial Dataset Sizes.

		Institution dataset size (%)			
Metric	Model	10%	20%	40%	100%
PICAI Score	PICAI	32.2 (29-36)	39.9 (35-45)	54.8 (48-61)	58.0 (51-64)
	LOCAL	56.2 (50-63)	57.6 (51-64)	61.8 (56-68)	**66.1** (60-72)
	CENTRAL	**57.6** (51-64)	**58.1** (52-65)	**63.8** (57-70)	64.7 (58-71)
Exam-Level AUROC	PICAI	58.0 (54-62)	65.0 (59-71)	72.8 (66-79)	74.2 (68-80)
	LOCAL	**77.2** (72-82)	**76.9** (71-82)	**78.4** (73-84)	**79.9** (74-85)
	CENTRAL	74.0 (68-80)	74.7 (69-80)	76.9 (71-82)	78.3 (73-84)
Lesion-Level AP	PICAI	6.3 (3-11)	14.9 (9-21)	36.8 (29-45)	41.8 (34-50)
	LOCAL	35.1 (27-44)	38.2 (30-46)	45.2 (37-53)	**52.3** (44-60)
	CENTRAL	**41.2** (32-50)	**41.4** (33-50)	**50.8** (43-59)	51.0 (42-59)

Note. Metrics with 95% confidence intervals in parentheses. For each combination of a metric and institution dataset size the best value between the 3 models is in bold. The dataset size indicates the percentage of exams used for model training. AUROC = area under receiver operating characteristic curve; AP = average precision.

multi-institutional dataset of 1241 external exams. The model provided by PICAI also underperformed on our local test set, suggesting that in-house implementation biases were not a factor. Even combining all available data into a centralized model yielded only minimal, statistically insignificant gains over local training alone.

While previous studies have explored the impact of dataset composition on model generalization,[6-9] our work uniquely assesses these impacts on a much larger scale using established public-domain model architectures and metrics. Importantly, we conduct our analysis using standardized, publicly available models and metrics tailored for csPCa by the PICAI group,[12] facilitating reproducibility and cross-study comparison. In contrast, many prior studies employ a wide range of models and evaluation metrics. The latter are often inconsistently applied making it difficult to determine which metric should serve as the primary measure of performance.[7,10,20] Existing metrics frequently lack either clinical relevance or the ability to comprehensively assess segmentation performance.[21,22] Moreover, prior studies are typically based on relatively small cohorts. For example, Netzer et al, Rodrigues et al, and Provenzano et al used only 640, 733, and 175 exams, respectively.[6,7,9] To our knowledge, only 3 related studies have exceeded 1000 exams. Rajagopal et al used 1959 exams, however nearly half were acquired with an endorectal coil, making the domain shift overt and undermining its relevance.[8] Hosseinzadeh et al used 2734 exams[5] but their only domain shift-related analysis involved testing their local model on 296 external exams, which showed no significant drop in performance. Furthermore, their data lacked biopsy confirmation, instead using PIRADS scores as the reference standard and PIRADS 3 lesions were treated as csPCa-negative.

In contrast to prior work, our study explicitly compares models trained on local versus external data across 3208 prostate MRI exams, more than any previous study. This allows the assumption that training with a sufficiently large and diverse external dataset can match or exceed the performance of local training, to be tested at an unprecedented scale and level of data diversity. Despite the growing availability of large, multi-institutional datasets,[12,23] our findings emphasize the continued importance of local data. Even when trained on over 1200 external exams from 3 institutions, the PICAI-TRAIN model still underperformed on data from a new independent institution. Training with less than one quarter the number of cases using local data matched the external-data trained model performance. Even when combining all the available data to train on a total of 2641 exams, the additional 1241 external exams included in training (representing a 47% increase in data volume) did not result in any significant performance improvements. This suggests that training or fine-tuning with local data remains essential in real-world deployment, even when leveraging external multi-institutional datasets on the order of 1000 exams.

It is noteworthy that when studying the PICAI-TEST dataset the LOCAL-TRAIN model was only 1.6% worse than the PICAI-TRAIN model while the spread for the LOCAL-TEST dataset was 8.1% (Table 3) suggesting our locally trained model was more generalizable. This requires further investigation. We did not insist on pathology confirmation for every case in our local training (just testing) while all PICAI cases were pathology confirmed. We can speculate that perhaps this improves training diversity and generalizability by including more cases that did not go to biopsy.

While some hypothesize that domain shift can be mitigated with sufficiently large and diverse datasets[6-9] our findings suggest that 1200 exams across 3 sites are insufficient. The true threshold may be much higher, and maintaining such a model over time would be costly given evolving local practices. Our study shows that even small amounts of local data significantly improve local performance. This suggests frameworks of deployment should support ongoing data sharing or fine-tuning to allow clients to update the central model. How to do this is an important direction for future research in domains such as federated learning, quality assurance, and models that can correct for domain shift with minimal local data.

Limitations of our study include the absence of pathological confirmation for all LOCAL-TRAIN cases, whereas PICAI-TRAIN was completely biopsy-confirmed. However, this would bias against local training, the opposite of what we observed. Additionally, all local test cases were biopsy confirmed. Another limitation is that the publicly available PICAI dataset is only a subset of the full PICAI data, which remains private. Notably, we have not studied the performance of much larger external data sets trained over larger numbers of institutions and encompassing all major MRI manufacturers and field strengths as, to our knowledge, such a dataset is not publicly-available. It remains possible that an order of magnitude larger centralized data may obviate the need for local training.

In conclusion, domain shift limits MRI prostate cancer segmentation performance even when training with over 1000 exams from 3 external institutions. Use of local data is paramount at these scales.

Abbreviations

ADC apparent diffusion coefficient
AP average precision
AUROC area under the receiver operating characteristic
BHIGH high b-value
bpMRI biparametric MRI
csPCa clinically significant prostate cancer
GG International Society of Urological Pathology grade group
PIRADS Prostate Imaging Reporting and Data System
PSA prostate specific antigen
T2W T2 weighted image

ORCID iDs

Sangeet Ghai https://orcid.org/0000-0001-9451-0594
Masoom A. Haider https://orcid.org/0000-0002-7165-8315

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially funded by the Sinai Health Foundation, Toronto, Canada.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

1. Haider MA, Brown J, Chin JLK, Perlis N, Schieda N, Loblaw A. Evidence-based guideline recommendations on multiparametric magnetic resonance imaging in the diagnosis of clinically significant prostate cancer: a Cancer Care Ontario updated clinical practice guideline. *Can Urol Assoc J*. 2022;16:16-23. doi:10.5489/cuaj.7425
2. Cornford P, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer-2024 update. Part I: screening, diagnosis, and local treatment with curative intent. *Eur Urol*. 2024;86:148-163. doi:10.1016/j.eururo.2024.03.027
3. Westphalen AC, McCulloch CE, Anaokar JM, et al. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: experience of the society of abdominal radiology prostate cancer disease-focused panel. *Radiology*. 2020;296:190646. doi:10.1148/radiol.2020190646
4. Forookhi A, Laschena L, Pecoraro M, et al. Bridging the experience gap in prostate multiparametric magnetic resonance imaging using artificial intelligence: a prospective multi-reader comparison study on inter-reader agreement in PI-RADS v2.1, image quality and reporting time between novice and expert readers. *Eur J Radiol*. 2023;161:110749. doi:10.1016/j.ejrad.2023.110749
5. Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *Eur Radiol*. 2022;32:2224-2234. doi:10.1007/s00330-021-08320-y
6. Netzer N, Eith C, Bethge O, et al. Application of a validated prostate MRI deep learning system to independent same-vendor multi-institutional data: demonstration of transferability. *Eur Radiol*. 2023;33:7463-7476. doi:10.1007/s00330-023-09882-9
7. Rodrigues NM, de Almeida JG, Verde ASC, et al. Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data. *Comput Biol Med*. 2024;171:108216. doi:10.1016/j.compbiomed.2024.108216
8. Rajagopal A, Redekop E, Kemisetti A, et al. Federated learning with research prototypes: application to multi-center MRI-based detection of prostate cancer with diverse histopathology. *Acad Radiol*. 2023;30:644-657. doi:10.1016/j.acra.2023.02.012
9. Provenzano D, Melnyk O, Imtiaz D, et al. Machine learning algorithm accuracy using single- versus multi-institutional image data in the classification of prostate MRI lesions. *Appl Sci*. 2023;13:1088. doi:10.3390/app13021088
10. Andrade-Miranda G, Vega PS, Taguelmimt K, et al. Exploring transformer reliability in clinically significant prostate cancer segmentation: a comprehensive in-depth investigation. *Comput Med Imaging Graph*. 2024;118:102459. doi:10.1016/j.compmedimag.2024.102459
11. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203-211. doi:10.1038/s41592-020-01008-z
12. Saha A, Bosma JS, Twilt JJ, et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. *Lancet Oncol*. 2024;25:879-887. doi:10.1016/S1470-2045(24)00220-1
13. Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands. GitHub. DIAGNijmegen/picai_nnunet_semi_supervised_gc_algorithm: semi-supervised nnUNet model for 3D csPCa detection/diagnosis in bpMRI, deployable on grand-challenge.org. 2023. Accessed April 9, 2025. https://github.com/DIAGNijmegen/picai_nnunet_semi_supervised_gc_algorithm
14. Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands. GitHub. DIAGNijmegen/

picai_baseline. 2023. Accessed, April 9, 2025. https://github.com/DIAGNijmegen/picai_baseline
15. 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;76:340-351. doi:10.1016/j.eururo.2019.02.033
16. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31:1116-1128. doi:10.1016/j.neuroimage.2006.01.015
17. Bosma JS, Saha A, Hosseinzadeh M, Slootweg I, de Rooij M, Huisman H. Semisupervised learning with report-guided pseudo labels for deep learning-based prostate cancer detection using biparametric MRI. *Radiol Artif Intell*. 2023;5:e230031. doi:10.1148/ryai.230031
18. Isensee F, Wald T, Ulrich C, et al. nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: Linguraru MG, Dou Q, Feragen A, et al, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer Nature Switzerland; 2024:488-498.
19. Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands. GitHub. DIAGNijmegen/picai_eval. 2025. Accessed, April 9, 2025. https://github.com/DIAGNijmegen/picai_eval
20. Liu Q, Dou Q, Heng P-A. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Martel AL, Abolmaesumi P, Stoyanov D, et al, eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Springer International Publishing; 2020:475-485.
21. Yan W, Yang Q, Syer T, et al. The impact of using voxel-level segmentation metrics on evaluating multifocal prostate cancer localisation. In: Wu S, Shabestari B, Xing L, eds. *Applications of Medical Artificial Intelligence*. Springer Nature Switzerland; 2022:128-138.
22. Nai Y-H, Teo BW, Tan NL, et al. Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset. *Comput Biol Med*. 2021;134:104497. doi:10.1016/j.compbiomed.2021.104497
23. Tibrewala R, Dutt T, Tong A, et al. FastMRI Prostate: a public, biparametric MRI dataset to advance machine learning for prostate cancer imaging. *Sci Data*. 2024;11:404. doi:10.1038/s41597-024-03252-w