Comparison of Prostate MRI Lesion Segmentation Agreement Between Multiple Radiologists and a Fully Automatic Deep Learning System

Vergleich der Kongruenz von Prostata-MRT-Läsionssegmentationen durch mehrere Radiologen und ein vollautomatisches Deep-Learning-System

Authors

Patrick Schelb[1], Anoshirwan Andrej Tavakoli[1], Teeravut Tubtawee[1], Thomas Hielscher[2], Jan-Philipp Radtke[3], Magdalena Görtz[3], Viktoria Schütz[3], Tristan Anselm Kuder[4], Lars Schimmöller[5], Albrecht Stenzinger[6], Markus Hohenfellner[3], Heinz-Peter Schlemmer[1], David Bonekamp[1]

Affiliations

1 Division of Radiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
2 Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany
3 Department of Urology, University of Heidelberg Medical Center, Heidelberg, Germany
4 Division of Medical Physics, German Cancer Research Center (DKFZ), Heidelberg, Germany
5 University Dusseldorf, Medical Faculty, Department of Diagnostic and Interventional Radiology, Dusseldorf, Germany
6 Institute of Pathology, University of Heidelberg Medical Center, Heidelberg, Germany

Key words

MRI, prostate, prostate cancer, deep learning, artificial intelligence, convolutional neural network

received 15.07.2020
accepted 29.09.2020
published online 19.11.2020

Bibliography

Fortschr Röntgenstr 2021; 193: 559–573
DOI 10.1055/a-1290-8070
ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Correspondence

Prof. Dr. med. David Bonekamp
Abteilung Radiologie (E010),
Deutsches Krebsforschungszentrum,
Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
d.bonekamp@dkfz-heidelberg.de

ZUSAMMENFASSUNG

Ziel Ein kürzlich eigens entwickeltes künstliches neuronales Netzwerk (U-Net) zeigte eine gute und mit klinischer radiologischer Befundung vergleichbare Erkennungsrate klinisch signifikanter Prostatakarzinome (sPC). In dieser Arbeit wird nun die Kongruenz der durch U-Net und mehrere Radiologen erstellten Läsionsvolumina (der Segmentationen) verglichen.

Materialien und Methoden 165 Patienten mit Verdacht auf sPC erhielten eine multiparametrische MRT (mpMRT) bei 3 Tesla, gefolgt von gezielter und systematischer MR/TRUS-Fusionsbiopsie. Fünf Segmentationen pro Untersuchung wurden erstellt: Segmentationen klinischer Läsionen, unabhängige und geblindete retrospektive PI-RADS-Befundung durch 3 Radiologen und U-Net. Die läsionsbasierte Übereinstimmung für jeden Befunder wurde durch den Dice-Koeffizienten mit überlappenden Läsionen anderer Befunder bestimmt. Die Übereinstimmung wurde durch deskriptive Statistik und lineare gemischte Modelle verglichen.

Ergebnisse Der mittlere Dice-Koeffizient war für Radiologen mit 0,48–0,52 nur moderat kongruent als Ausdruck der schwierigen visuellen Aufgabe, die Begrenzung sonst übereinstimmend detektierter Läsionen zu bestimmen. U-Net-Segmentationen waren signifikant kleiner als manuelle Segmentationen ($p < 0{,}0001$) und zeigten einen geringeren mittleren Dice-Koeffizienten von 0,22, signifikant kleiner als manuelle Segmentationen (alle $p < 0{,}0001$). Diese Unterschiede blieben nach Adjustierung für die Segmentationsgröße bestehen und wurden nicht durch das Vorliegen eines sPC oder eine zonale Lokalisation in der peripheren oder Transitionszone beeinflusst.

Schlussfolgerung Die Kenntnis der Größenordnung der Übereinstimmung manueller Segmentationen verschiedener Radiologen ist wichtig, um den Erwartungswert für Künstliche-Intelligenz (KI) -Ansätze festzulegen. Eine perfekte Übereinstimmung (Dice-Koeffizient von 1) sollte für KI nicht erwartet werden. Die geringeren Dice-Koeffizienten des U-Nets werden nur teilweise durch die geringere Segmentationsgröße des U-Nets erklärt, was durch eine Fokussierung des U-Nets auf den

Läsionskern und eine geringe Verschiebung des Läsionszentrums erklärt werden könnte. Obwohl primär die korrekte Detektion von sPC durch KI wichtig ist, kann der Dice-Koeffizient mit multiplen Befundern als sekundäres Qualitätsmaß in zukünftigen Studien herangezogen werden.

Kernaussagen:

- Intermediäre Dice-Koeffizienten der Radiologen reflektieren die Schwierigkeit der übereinstimmenden Festlegung der Berandung gemeinsam detektierter Läsionen.
- Die beobachteten geringeren Dice-Koeffizienten motivieren die Weiterentwicklung von Deep Learning Systemen mit dem Ziel der besseren Approximation menschlicher Perzeption.
- Eine vergleichbare Prädiktion des klinisch signifikanten Prostatakarzinoms erscheint unabhängig von der Übereinstimmung der Dice-Koeffizienten.
- Die Unabhängigkeit des Dice-Koeffizienten vom Vorliegen eines signifikanten Prostatakarzinoms spricht für die fehlende Unterscheidbarkeit mancher benigner von malignen Bildcharakteristika.
- Technische Verbesserungen in der Bildregistrierung zwischen DWI und T2 können in Zukunft möglicherweise die U-Net Dice-Koeffizienten erhöhen.

ABSTRACT

Purpose A recently developed deep learning model (U-Net) approximated the clinical performance of radiologists in the prediction of clinically significant prostate cancer (sPC) from prostate MRI. Here, we compare the agreement between lesion segmentations by U-Net with manual lesion segmentations performed by different radiologists.

Materials and Methods 165 patients with suspicion for sPC underwent targeted and systematic fusion biopsy following 3 Tesla multiparametric MRI (mpMRI). Five sets of segmentations were generated retrospectively: segmentations of clinical lesions, independent segmentations by three radiologists, and fully automated bi-parametric U-Net segmentations. Per-lesion agreement was calculated for each rater by averaging Dice coefficients with all overlapping lesions from other raters. Agreement was compared using descriptive statistics and linear mixed models.

Results The mean Dice coefficient for manual segmentations showed only moderate agreement at 0.48–0.52, reflecting the difficult visual task of determining the outline of otherwise jointly detected lesions. U-net segmentations were significantly smaller than manual segmentations ($p < 0.0001$) and exhibited a lower mean Dice coefficient of 0.22, which was significantly lower compared to manual segmentations (all $p < 0.0001$). These differences remained after correction for lesion size and were unaffected between sPC and non-sPC lesions and between peripheral and transition zone lesions.

Conclusion Knowledge of the order of agreement of manual segmentations of different radiologists is important to set the expectation value for artificial intelligence (AI) systems in the task of prostate MRI lesion segmentation. Perfect agreement (Dice coefficient of one) should not be expected for AI. Lower Dice coefficients of U-Net compared to manual segmentations are only partially explained by smaller segmentation sizes and may result from a focus on the lesion core and a small relative lesion center shift. Although it is primarily important that AI detects sPC correctly, the Dice coefficient for overlapping lesions from multiple raters can be used as a secondary measure for segmentation quality in future studies.

Key Points:

- Intermediate human Dice coefficients reflect the difficulty of outlining jointly detected lesions.
- Lower Dice coefficients of deep learning motivate further research to approximate human perception.
- Comparable predictive performance of deep learning appears independent of Dice agreement.
- Dice agreement independent of significant cancer presence indicates indistinguishability of some benign imaging findings.
- Improving DWI to T2 registration may improve the observed U-Net Dice coefficients.

Citation Format

- Schelb P, Tavakoli AA, Tubtawee T et al. Comparison of Prostate MRI Lesion Segmentation Agreement Between Multiple Radiologists and a Fully Automatic Deep Learning System. Fortschr Röntgenstr 2021; 193: 559–573

Introduction

Prostate MRI in combination with stereotactic MR-guided biopsies is becoming an important cornerstone of the diagnostic pathway in patients with suspicion of clinically significant prostate cancer (sPC) [1, 2]. As a result of several major clinical trials, pre-biopsy prostate MRI is increasingly utilized to better guide biopsy needles to suspicious targets and to potentially avoid biopsies in patients with unsuspicious prostate MRI [3, 4]. However, biopsy avoidance is limited by the known underestimation of multifocal lesions on prostate MRI [5, 6]. With its increased use, the demand for consistent expert-level interpretation of prostate MRI is rising. The Prostate Imaging – Reporting and Data System (PI-RADS) is the current standard in clinical prostate MRI interpretation and aims at reducing inter-reader variability and at standardizing interpretation and clinical MR protocols [7, 8]. Novel artificial intelligence approaches such as convolutional neural networks (CNN) promise to capture diagnostically decisive information directly from medical images [9, 10]. In the prostate, systems providing fully automated prostate assessment and lesion segmentation [11] or based on slice classification [12] have been developed. Other applications have utilized CNNs to evaluate predefined regions on prostate MR images [13]. It is important to ascertain that such systems have an acceptable true-negative rate in clinical

screening scenarios. This requires evaluation of clinical performance in consecutive patients screened for sPC also including patients that are ultimately not diagnosed with sPC [11], rather than evaluating pre-selected cohorts such as only patients with visible MR lesions [12]. Systems that utilize bi-parametric (T2-weighted and diffusion-weighted) MRI [11] are potentially able to extract more information than systems focusing only on a single sequence [12], as the multiparametric approach to prostate MRI has long been known to benefit from the combination of high-resolution anatomical and information-rich functional imaging [14]. It is generally expected that increasing amounts of data for the training of neural networks will improve the quality and generalizability of these models. Data annotation requires significant time and resources. High-quality 3 D lesion annotations on bi-parametric sequences [11] will not be possible for extremely large data sets. As such, in medical deep learning in general, patient-level [15] or slice-level [16] annotations are being explored as especially patient-level information is readily available in hospital information systems and any AI method trained on such data could be directly applied to the largest possible retrospectively available data sets. However, before accepting end-to-end approaches that skip lesion segmentation and proceed directly to patient-level assessment [16], it is of utmost importance to ascertain that these systems in fact base their decisions on the correct portion of the data. It is possible that prostate classification systems base their patient assessment on extra-prostatic tissues that correlate to other risk factors such as age instead of on prostate tissues. Also, it can be argued whether a segmentation CNN such as the U-Net [10] is required to exactly reproduce the segmentations of a single radiologist. From a diagnostic standpoint, it would be sufficient for the system to indicate the location and presence of a lesion without agreeing fully with the extent provided by a radiologist. Also, at present, the Dice coefficient [17], a commonly used spatial overlap index, is expected to be high for clearly defined structures such as the prostate itself, but understandably lower for prostate lesions given the known inter-observer variability in lesion assessment in the difficult task of detecting suspicious lesions on a background of hyperplastic, inflammatory, and degenerative changes [18]. We hypothesized that the inter-rater variability between different radiologists will be substantially lower than perfect (Dice coefficient of one) and comparable to that of a previously developed deep learning algorithm for automatic prostate MRI assessment.

The purpose of this study was to directly compare three-dimensional lesion segmentations based on blinded retrospective PI-RADS interpretations of three radiologists with retrospective segmentations of lesions indicated in clinical reports and with automated lesion segmentations provided by a previously established U-Net.

Materials and Methods

Patient cohort

The examined patient cohort is part of the previously published cohort which was used to train and evaluate the U-Net architecture used in this study [11]. Inclusion criteria for the study sample were MRI performed from May 2015 to February 2016; consecutive men with a clinical indication for biopsy based on prostate-specific antigen (PSA) elevation and clinical examination or participation in our active surveillance program; imaging performed with our main institutional 3.0 Tesla MRI system; MRI-transrectal US fusion biopsy performed at our institution; detection of a focal lesion by at least one of the raters and overlap of each detected lesion with at least one lesion by another rater. The exclusion criteria were history of treatment for prostate cancer (prostatectomy, radiation therapy, antihormonal therapy, focal therapy); biopsy within the past 6 months prior to MRI; and incomplete sequences or severe MRI artifacts. The institutional and governmental ethics committee approved the study and waived informed consent.

MR imaging

MR images were acquired at 3 Tesla (Magnetom Prisma, Siemens Healthcare, Erlangen, Germany) using the standard multi-channel body coil and integrated spine phased-array coil, according to the European Society of Urogenital Radiology (ESUR) guidelines [19]. As per the institutional protocol, axial, coronal and sagittal T2-weighted (T2) images, echo-planar imaging (EPI) diffusion-weighted images (DWI) and dynamic-contrast enhanced (DCE) images were acquired. Clinical interpretation by board certified radiologists included PI-RADS assessment for each detected lesion and a pictogram indicating lesion location [8, 20].

Systematic and targeted MRI/TRUS-fusion biopsies

All men underwent transperineal grid-directed biopsy performed under general anesthesia with rigid software registration using the BiopSee system (MEDCOM, Darmstadt, Germany). Targeted fusion-biopsy (FTB) of MRI-suspicious lesions (inter-quartile range (IQR) 4–8 cores, median 6 per lesion) was followed by systematic saturation biopsy (20–29 cores, median 24 cores), as previously described [1, 21]. This extended targeted and systematic biopsy approach has been validated against and confirmed to be concordant with RP specimens [21]. A median of 30 biopsies (IQR 24–37) were taken per patient with the number of biopsies adjusted to prostate volume [22]. Systematic biopsies were collected from sextants according to the Ginsburg Study group protocol [22].

Manual and U-Net MR lesion segmentations

Five sets of segmentations were recorded: prospectively called clinical lesions (CL) were retrospectively segmented by a prostate MRI researcher (blinded) with 6 months of experience using series and slice number and descriptions from clinical reports and their embedded pictograms, under supervision of a board-certified senior radiologist with 10 years of experience in prostate MRI (blinded). Here, we utilized the same segmentations also used for previous training and validation of the U-Net. Three radiologists performed previously unreported new independent retrospective interpretations according to PI-RADS version 2.1 blinded to pathology results and clinical reports but with access to PSA values. Radiologists included an expert with 12 years of prostate MRI experience (blinded, R1), a board-certified radiologist with 3 years of experience (blinded, R2) and a radiology resident with

2 years of experience and research focus in prostate MRI (blinded, R3). Segmentations were performed separately on diffusion-weighted (DWI) images (using both apparent diffusion coefficient (ADC) and b-value = 1500 s/mm2 (B1500) for lesion assessment) and T2-weighted (T2w) images using the medical imaging toolkit (MITK, www.mitk.org) [23, 24]. Three-dimensional (3 D) volumes of interest (VOI) were manually drawn using the MITK polygon tool by the investigators. Investigators were instructed to review the entire examination including T2w, DWI and DCE images first and then perform the segmentations integrating the visual appearance and the mental image of the appearance on all sequences for delineation. In addition, probability maps were provided by U-Net. Briefly, U-Net was previously trained and validated for prostate and lesion segmentation using a cohort of 250 examinations for training and cross-validation and another cohort of 62 examinations for independent testing [11] and has demonstrated comparable performance to clinical PI-RADS interpretation. For 134 of the 165 included examinations that were part of the original training set, the four U-Nets of the full ensemble of 16 U-Nets that had not been trained with each respective case were used to calculate average probability maps for each patient. For the remaining 31 examinations that were part of the original test set, all 16 U-Net members of the ensemble contributed to the probability maps. Probability maps were thresholded at 0.22, corresponding to the threshold mimicking a PI-RADS 3 assessment established during definition and validation of the U-Net [11]. In the resulting binary images, non-contiguous regions were separated, and each isolated region considered as a separate lesion segmentation. In addition, prostate contours were segmented manually on T2w images. The prostate masks were then automatically partitioned into sextants according to the Ginsburg protocol using the mid-sagittal plane and four additional angulated planes [5, 25].

Histopathology and fused ground truth

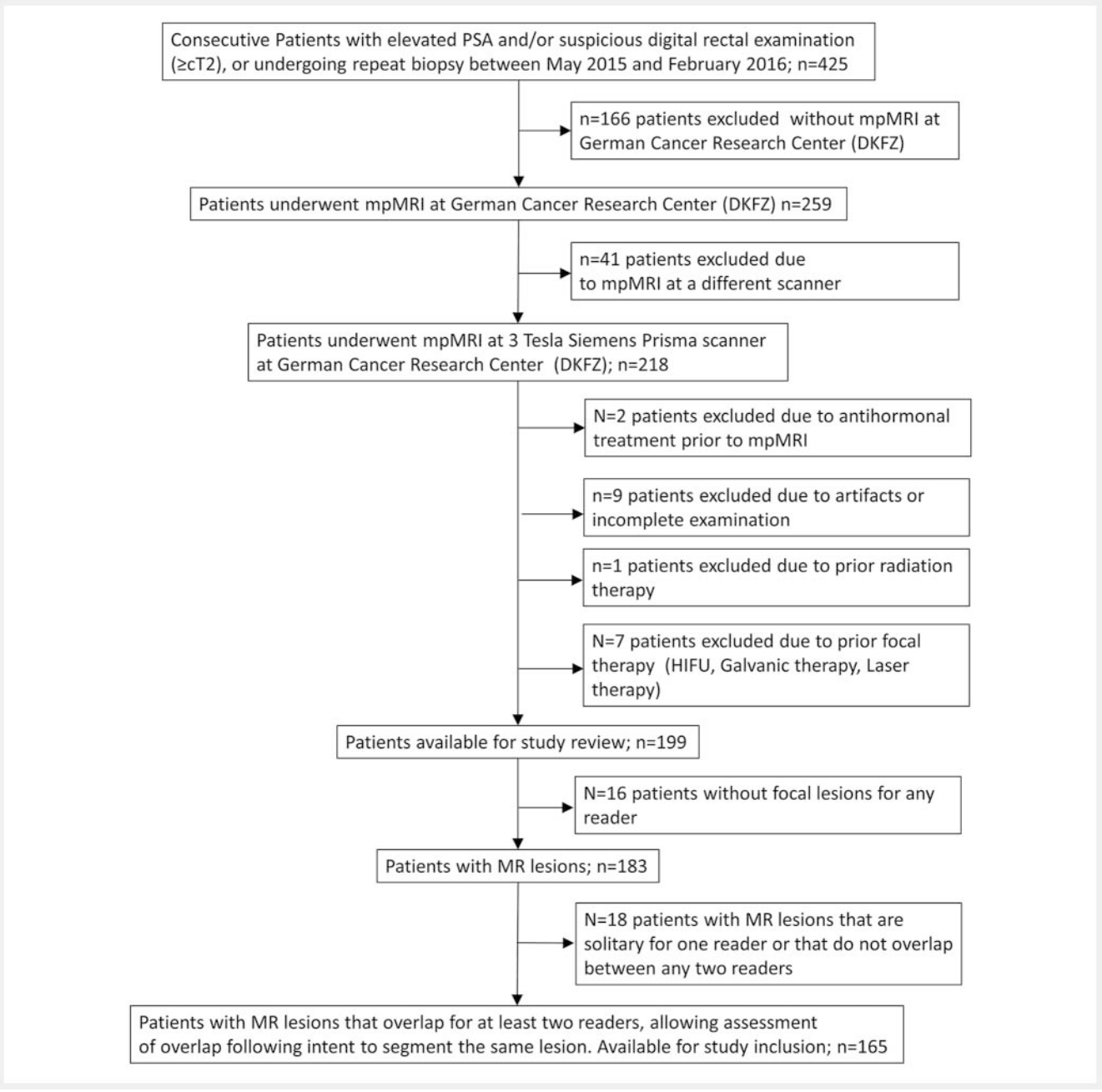
Histopathological analyses were performed under supervision of one dedicated uropathologist (blinded, 17 years of experience) according to International Society of Urological Pathology standards. sPC was defined as International Society of Urological Pathology (ISUP) grade group ≥ 2 [26]. Separate assessment of the ISUP grade group was provided for each MR targeted lesion and for each sextant of the Ginsburg biopsy scheme. As it is known that targeted biopsies can underestimate sPC in MR index lesions [21], we use a fused ground truth for optimal assessment of histopathology attributable to any MR segmentation. For this purpose, each sextant segmentation is assigned its corresponding systematic core histopathology. Then, the voxel-wise overlap between each sextant segmentation and each clinical targeted segmentation is determined and all sextants that overlap are assigned to the maximum ISUP grade group of either systematic sextant or targeted lesion histopathology. This way a high-quality ground truth is established. The utilized extended systematic and targeted biopsy approach has been previously shown to have excellent sensitivity (97 %) for the presence of sPC compared to radical prostatectomy specimens [1].

Statistical analysis

Overlap analysis was based on T2w segmentations. We used only lesions for the analysis for which at least one of the other raters provided an overlapping segmentation. Some raters provided segmentations for probably benign lesions such as BPH nodules, which were excluded from the analysis. In this way, we exclude the detection task from the assessment and focus only on the a posteriori task of outlining the lesion boundary once at least two raters have determined that a lesion should be segmented at a specific location. Descriptive statistics were used to summarize the demographic and clinical characteristics of the patient cohort and the distribution of sPC into prostate zones and ISUP grade group. A logistic mixed model with random patient effects was used to test for difference in sPC probability between MR lesions in the PZ and TZ. The mutual voxel overlap of each lesion segmentation with each other lesion segmentation per examination was determined. Dice coefficient [17] was calculated as:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

With *A* and *B* indicating the cardinality (the number of voxels) in segmentations A and B, respectively, and A ∩ B indicating the set of voxels that overlap between segmentations A and B. We determined the average Dice coefficient across lesions for each pair of readers as a measure of pairwise inter-rater agreement. To compare each reader to all other readers, each lesion of that rater was regarded as a reference lesion and the Dice coefficients of all other readers' lesions overlapping with the reference lesion were averaged. The result is a metric that represents the congruence of lesion outlines between each reference rater and all other raters that decided to segment a lesion at the same location. To study the influence of lesion size mismatch, we calculated the lesion size ratio as the ratio of the size of the larger to the smaller lesion. Linear mixed models with random patient and lesion effects were constructed for pairwise comparison of average reader Dice coefficients. All pairwise comparisons were performed using Tukey's methods for multiplicity adjustment. In addition, models corrected for log-transformed voxel ratio were calculated. Analyses were performed separately for all lesions, for lesions containing no sPC, and for those containing sPC to examine whether readers agree more in sPC-positive lesions. The results were plotted as jittered dot plots superimposed on box plots. Lesion size distribution was plotted as histograms. A correlation between voxel ratio and Dice coefficient was plotted and smoothed using local regression. Per-lesion deviation of lesion size from the mean of all raters was depicted as box plots. Statistical analyses were performed using R version 3.6 (R Foundation for Statistical Computing, Vienna, Austria, using packages emmeans and lme4) [27]. Reporting followed Standards of Reporting of Diagnostic Accuracy [28].



▶ **Fig. 1** Inclusion of patients in the study.

▶ **Abb. 1** Einschluss von Patienten in die Studie.

Results

General characteristics and descriptive statistics

From 425 men presenting at our institutions in the inclusion period, 165 men (median age: 64 years; interquartile range: 58–71 years) met the inclusion and exclusion criteria (▶ **Fig. 1**). Of the 165 patients included in the study, 82 (50 %) were biopsy naïve, 50 (30 %) had received prior negative biopsies and 33 (20 %) were enrolled in active surveillance at the time of the MR exam. Baseline epidemiological and clinical characteristics including pathological findings, and clinical assessment are given in ▶ **Table 1**. ▶ **Table 2** gives a detailed overview of all overlapping lesions, MR assessments per lesion, number of MR lesions per patient, and number of overlapping lesions with other raters for each rater individually. For all raters including CNN, complete agreement that a lesion was present at a specific location occurred most frequently (41–44 % of lesions) compared to only 3 additional raters agreeing which was the second most common occurrence for human raters (25–30 % of lesions), while for CNN it was agreement with one additional rater (30 % of lesions).

▶ **Table 1** Demographic and clinical characteristics of 165 included men.

▶ **Tab. 1** Demografische und klinische Merkmale der 165 in die Studie eingeschlossenen Patienten.

cohort	n = 165
age (years)	
▪ median (IQR)	64 (58–71)
sPC found in MRI lesion (n (%))	436/868 (50 %)
no sPC found in MRI lesion	432/868 (50 %)
lesions in the PZ	632/868 (75 %)
lesions in the TZ	220/868 (25 %)
multi-zonal lesions	16/868 (2 %)
sPC in PZ lesions	317/639 (50 %)‡
sPC in TZ lesions	109/217 (50 %)‡
sPC in multi-zonal lesions	10/12 (83 %)
per-patient maximum Gleason Score/ISUP grade group (n (%))	
▪ no PC	59 (36 %)
▪ 6 (3 + 3)/GG I	28 (17 %)
▪ 7a (3 + 4)/GG II	52 (32 %)
▪ 7b (4 + 3)/GG III	11 (7 %)
▪ 8 (4 + 4)/GG IV	4 (2 %)
▪ 9a (4 + 5)/GG V	7 (4 %)
▪ 9b (5 + 4)/GG V	4 (2 %)
▪ 10 (5 + 5)/GG V	0 (0 %)
PSA (ng/ml) median (IQR)	7.2 (5.2–10.4)
biopsy distribution per patient (n (%))	
▪ biopsy-naïve	82 (50 %)
▪ previously biopsied	50 (30 %)
▪ active surveillance	33 (20 %)

IQR = interquartile range; PSA = prostate specific antigen; MRI = magnetic resonance imaging; sPC = clinically significant prostate cancer; PZ = peripheral zone; TZ = transition zone; GG = ISUP grade group.
‡ difference in logistic mixed model not significant (p = 0.16) in sPC and non-sPC containing lesions between PZ and TZ.

CNN has lower lesion boundary agreement compared to human raters

Box plots of average per-lesion Dice coefficients between each reference reader and all other readers are given in ▶ **Fig. 2**. Averages of lesion-based Dice coefficients for all pairwise rater comparisons are given in ▶ **Table 3**, demonstrating that lesion boundary agreement according to Dice coefficients was lower for CNN compared to human raters. Linear mixed model analyses with random patient and lesion effects demonstrated significant differences in Dice coefficients between CNN and all human raters ($p < 0.0001$ for all comparisons) (▶ **Table 4**). Lesion-based mean and standard deviation of each reference rater to all other raters within the set of all lesions is given in ▶ **Table 4**. The mean Dice coefficient with all other raters was 0.22 for CNN compared to 0.48–0.52 for the other readers.

U-Net segments smaller lesions compared to human raters

▶ **Fig. 3** depicts histograms of lesion sizes in voxels for each reader, with PI-RADS categories and zonal distribution being indicated separately. CNN segments the largest number of small lesions up to 1000 voxels (N = 107 compared to 32, 28, 57 and 46 for prospective, and readers 1–3, respectively), while it segments fewer lesions of larger size compared to the human raters. ▶ **Fig. 4A** demonstrates the percent deviation of segmentation size from the per-lesion rater mean. CNN segmentations were on average 21 % smaller than the mean of all raters and significantly smaller in all pairwise comparisons (all $p <= 0.0001$). In addition, some of the large multi-zonal lesions are broken into smaller parts and thus appear both in the group of the smallest and largest lesions for CNN, while – as expected – these lesions are all segmented into the largest lesion group by all human readers.

Lesion boundary agreement depends on segmentation size

▶ **Fig. 4B** demonstrates the dependence of Dice coefficients on the lesion size ratio. Lesion size mismatch is an important factor contributing to low Dice coefficients, showing e. g. for CNN that at comparable lesion sizes Dice coefficients approximate 0.5, while they fall to about one quarter of that as the size mismatch reaches 10:1.

CNN agreement remains reduced after correction for segmentation size

Adjusting linear mixed model analyses with random patient and lesion effects for segmentation size did not remove significant differences in Dice coefficients between CNN and all human raters ($p < 0.0001$) (▶ **Table 4**). Thus, the difference in segmentation size only partially explains the overall lower Dice coefficient of CNN. This is supported by segmentations of Reader 1 being an average of 18 % larger than segmentations of all other raters and significantly larger than Reader 2 ($p < 0.0001$) and Reader 3 ($p < 0.001$), while there is an absence of a significant overlap difference for Reader 1 (▶ **Fig. 4A**, ▶ **Table 4**). Most of the size difference for Reader 1 is explained by segmentation of a smaller number of small PI-RADS 3 lesions (< 1000 voxels) and a larger number of PI-RADS 4 lesions between 1000 and 2000 voxels in size (▶ **Fig. 3**).

Agreement in sPC lesions

Additional analysis of the subgroups of sPC and non-sPC lesions is depicted in ▶ **Fig. 2** for all lesions and given in ▶ **Table 3, 4**. Dice coefficients with other readers were significantly higher ($p < 0.001$, ▶ **Table 4**) for CNN in lesions with sPC compared to non-sPC lesions, but not for human readers. However, this significance was removed ($p = 0.3$) after adjusting for voxel ratio.

▶ **Table 2** Overlapping lesions reported by different raters.

▶ **Tab. 2** Von verschiedenen Befundern detektierte überlappende Läsionen.

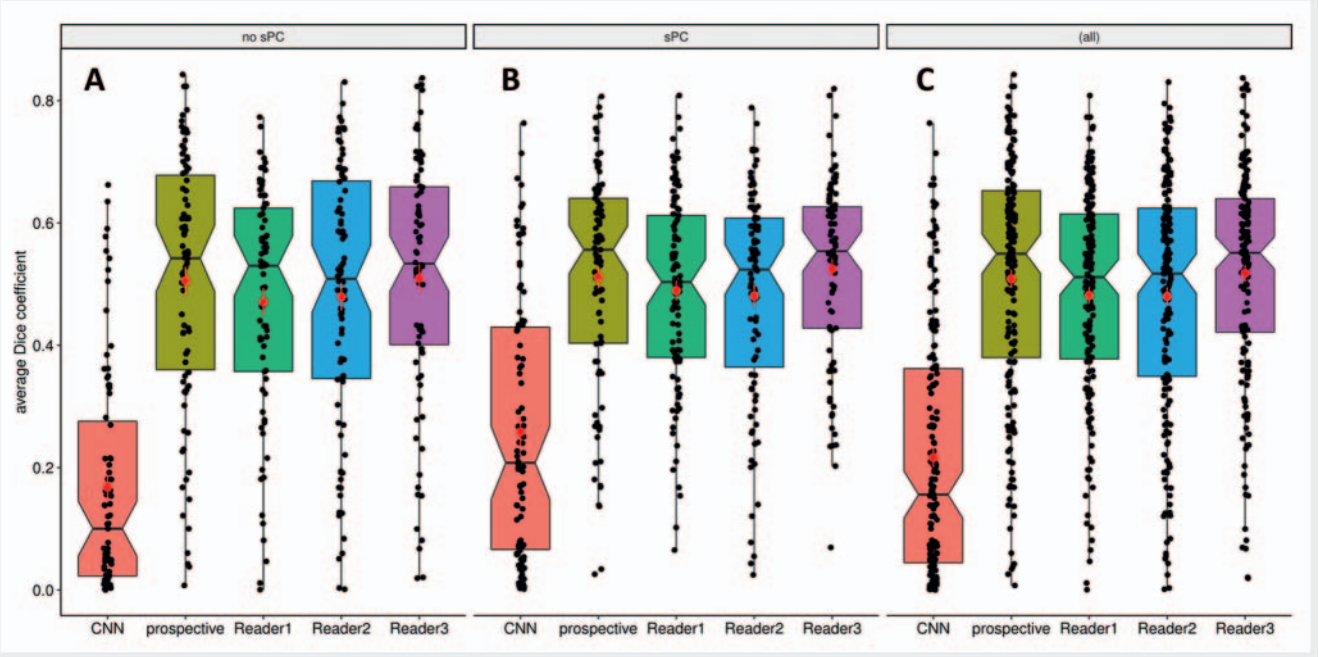
cohort	CNN	clinical	reader 1	reader 2	reader 3
number of lesions [n = 868 (100 %)] per reader (n (%))	171 (20 %)	179 (21 %)	170 (20 %)	179 (21 %)	169 (20 %)
MRI assessment (n (%))					
▪ total	171 (100 %)	179 (100 %)	170 (100 %)	179 (100 %)	169 (100 %)
▪ PI-RADS 3	18 (11 %)	47 (26 %)	31 (18 %)	56 (31 %)	40 (24 %)
▪ PI-RADS 4	77 (45 %)	88 (49 %)	97 (57 %)	83 (46 %)	76 (45 %)
▪ PI-RADS 5	76 (44 %)	44 (25 %)	42 (25 %)	40 (22 %)	53 (31 %)
patients without MRI lesions	58/165 (35 %)	26/165 (16 %)	36/165 (22 %)	26/165 (16 %)	37/165 (22 %)
patients with MRI lesions	107 (100 %)	139 (100 %)	129 (100 %)	139 (100 %)	128 (100 %)
number of MRI					
lesions per patient (n (%))					
▪ 1 lesion	70 (65 %)	102 (73 %)	95 (74 %)	104 (75 %)	95 (74 %)
▪ 2 lesions	24 (22 %)	34 (25 %)	28 (22 %)	30 (22 %)	26 (20 %)
▪ 3 lesions	6 (6 %)	3 (2 %)	5 (4 %)	5 (4 %)	6 (5 %)
▪ 4 lesions	3 (3 %)	0 (0 %)	1 (1 %)	0 (0 %)	1 (1 %)
▪ 5 or 6 lesions	4 (4 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
zonal distribution of lesions					
▪ peripheral zone (PZ)	3 (2 %)	2 (1 %)	2 (1 %)	1 (1 %)	4 (2 %)
▪ transition zone (TZ)	107 (63 %)	131 (73 %)	133 (78 %)	143 (80 %)	125 (74 %)
▪ large multi-zonal lesion	61 (36 %)	46 (26 %)	35 (21 %)	35 (20 %)	40 (24 %)
overlap of lesions with other raters (n (%))					
▪ 1 rater	51 (30 %)	30 (17 %)	31 (18 %)	37 (21 %)	22 (13 %)
▪ 2 raters	26 (15 %)	22 (12 %)	19 (11 %)	24 (13 %)	27 (16 %)
▪ 3 raters	22 (13 %)	52 (29 %)	46 (27 %)	45 (25 %)	50 (30 %)
▪ 4 raters	72 (42 %)	75 (42 %)	74 (44 %)	73 (41 %)	70 (41 %)

PI-RADS = Prostate Imaging Reporting and Data System; PZ = peripheral zone; TZ = transition zone.

▶ **Fig. 5** shows representative examples of lesion segmentations with a higher Dice coefficient of 0.71–0.77 (A) and low Dice coefficients of 0.08–0.16 (B) between CNN and other segmentations.

Influence of zonal lesion location

There was no statistical difference in the rate of sPC in reader-detected lesions between the PZ and the TZ (p = 0.17). Box plots of average per-lesion Dice coefficients between each reference reader and all other readers are depicted in ▶ **Fig. 6** stratified by the peripheral zone (PZ) and the transition zone (TZ). Additional analysis separated by PZ and TZ is given in ▶ **Table 3, 4**. When analyzing average agreement between readers stratified by zonal location, there was a slight but non-significant lower agreement in TZ lesions overall (average Dice coefficient for TZ lesions lower by 0.04, p = 0.09). Accordingly, for the readers individually – including CNN – there was no significant difference between the mean Dice coefficient of PZ and TZ lesions (all p > 0.067, ▶ **Table 4**). The reported differences in average agreement between CNN and other readers were independent of the zonal location, i. e. there was no heterogeneity in difference between readers with respect to zonal locations (interaction p = 0.9). Accordingly, pairwise comparisons of the average Dice coefficient between CNN and other readers remained significant in separate analyses of PZ and TZ lesions (all p < 0.0001), indicating significantly lower average agreement of CNN for both zonal locations at a similar magnitude compared to zone-independent analysis. Lesions were larger in the TZ (mean 6524, IQR 5999 voxels) compared to PZ (mean 3687, IQR 2695 voxels; p < 0.0001) for all raters except for CNN (TZ: mean 2272, IQR 2019 voxels; PZ: mean 2499, IQR 2464 voxels; p = 0.34). The bottom part of ▶ **Fig. 3** shows that



▶ **Fig. 2** Boxplots of per-lesion average Dice coefficient calculated between each reference reader and all remaining readers. The bold black line at the notch indicates the median, the ends of the box the 25th and 75th percentiles and the whiskers minimum and maximum. Black dots indicate individual data points. Red circles indicate the mean. Vertical red line segments indicate the standard error of the mean (SEM). **A** Boxplots depicting lesions without sPC **B** Boxplots depicting lesions with sPC. **C** Boxplots depicting data for all overlapping lesions.

▶ **Abb. 2** Box-Plots des durchschnittlichen Dice-Koeffizienten pro Läsion, berechnet zwischen jedem Referenzbefunder und allen verbleibenden Befundern. Die fette schwarze Linie an der Kerbe zeigt den Median an, die Enden der Box das 25. und 75. Perzentil und die Antennen die Minimal- und Maximalwerte. Gefüllte schwarze Kreise kennzeichnen einzelne Datenpunkte. Rote Kreise entsprechen dem Mittelwert. Vertikale rote Liniensegmente geben den Standardfehler des Mittelwerts (SEM) an. **A** Box-Plots mit Läsionen ohne sPC. **B** Box-Plots mit Läsionen mit sPC. **C** Box-Plots mit Daten für alle überlappenden Läsionen.

▶ **Table 3** Average of individual mutual Dice coefficients across lesions by reference reader, for all lesions and stratified by sPC/non-sPC and by peripheral zone (PZ) and transition zone (TZ).

▶ **Tab. 3** Über alle Läsionen gemittelte Dice-Koeffizienten, stratifiziert nach Referenzbefunder für alle Läsionen sowie separat für sPC-positive und -negative Läsionen.

reference reader	comparison	non-sPC lesions	sPC lesions	all lesions	PZ lesions	TZ lesions
		average Dice coefficient	average Dice coefficient	average Dice coefficient	average Dice coefficient	average Dice coefficient
CNN	prospective	0.23	0.29	0.26	0.29	0.21
CNN	reader 1	0.21	0.28	0.25	0.25	0.25
CNN	reader 2	0.20	0.31	0.27	0.27	0.27
CNN	reader 3	0.25	0.30	0.28	0.32	0.23
prospective	reader 1	0.58	0.54	0.56	0.57	0.52
prospective	reader 2	0.55	0.55	0.55	0.54	0.58
prospective	reader 3	0.61	0.61	0.61	0.62	0.57
reader 1	reader 2	0.53	0.51	0.51	0.54	0.44
reader 1	reader 3	0.54	0.56	0.56	0.55	0.55
reader 2	reader 3	0.56	0.56	0.56	0.56	0.54

▶ **Table 4** All pairwise comparisons of average Dice coefficient between reference readers, linear mixed model with random patient and lesion effect. Mean across lesions of mean Dice coefficients per lesion by reference reader. Test on the difference in average Dice coefficient between lesions with sPC and without sPC based on linear mixed model with random patient effect.

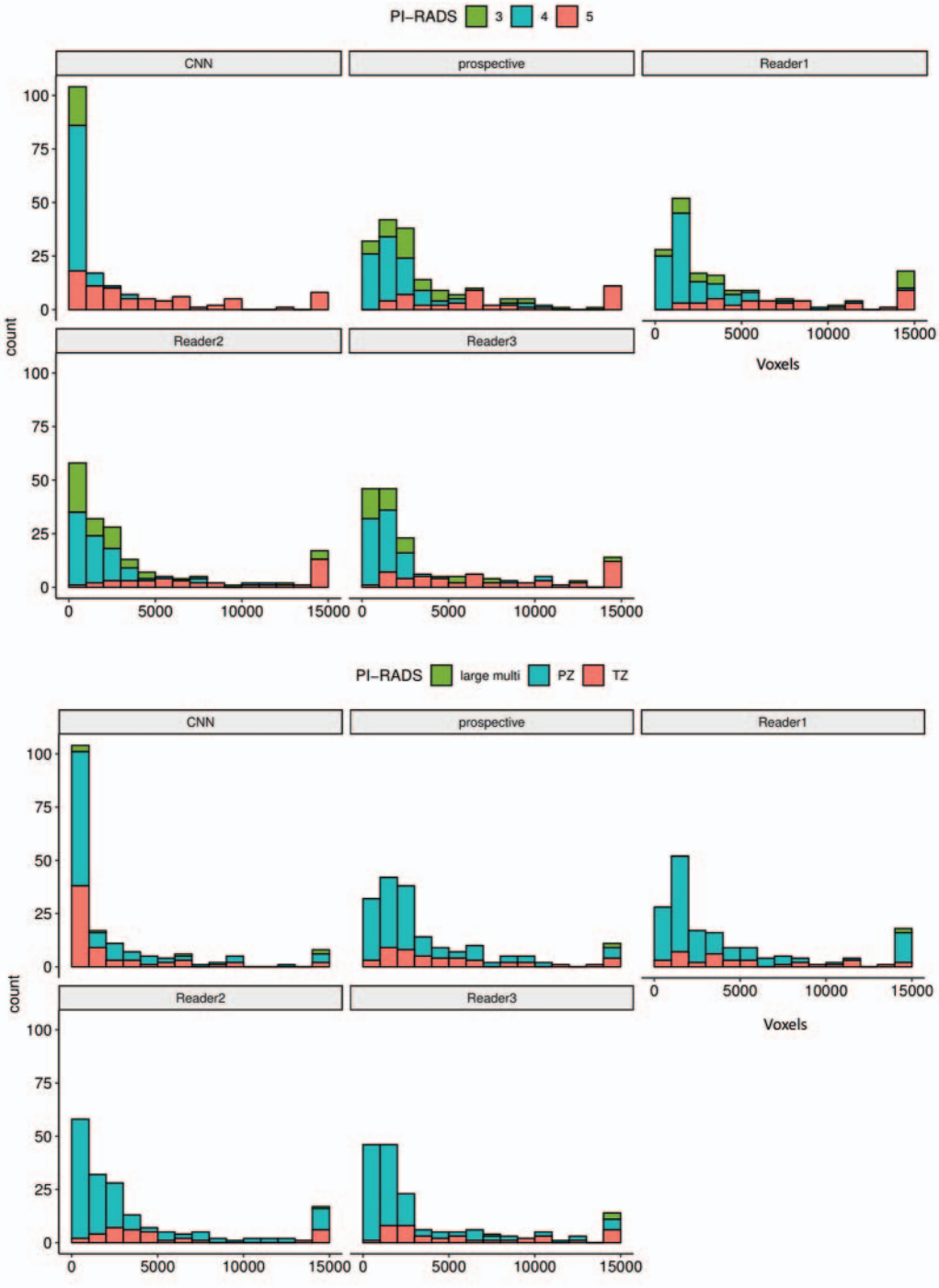
▶ **Tab. 4** Alle paarweisen Vergleiche des durchschnittlichen Dice-Koeffizienten zwischen Referenzbefundern, lineares gemischtes Modell mit Patienten und Läsionen als Random Effects. Über alle Läsionen und Vergleichsbefunder gemittelte Dice-Koeffizienten. Statistischer Test des Unterschieds des durchschnittlichen Dice-Koeffizienten zwischen Läsionen mit sPC und ohne sPC basierend auf einem linearen gemischten Random-Effects-Modell.

reference reader	comparison	non-sPC lesions		sPC lesions		all lesions	
		p-value	p-value adjusted for voxel ratio	p-value	p-value adjusted for voxel ratio	p-value	p-value adjusted for voxel ratio
CNN	prospective	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
CNN	reader 1	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
CNN	reader 2	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
CNN	reader 3	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
prospective	reader 1	0.2707	0.7120	0.1118	0.1901	0.0114	0.0779
prospective	reader 2	0.1275	0.1080	0.0624	0.1177	0.0032	0.0069
prospective	reader 3	0.9703	0.9789	0.9904	0.9949	0.9470	0.9679
reader 1	reader 2	0.9992	0.8399	0.9990	0.9992	0.9982	0.9464
reader 1	reader 3	0.6385	0.9547	0.3110	0.4053	0.0986	0.3122
reader 2	reader 3	0.4691	0.3909	0.1986	0.2805	0.0456	0.0616
reader		**mean dice coefficient non-sPC lesions**	**mean dice coefficient sPC lesions**	**mean dice coefficient all lesions**	**mean dice coefficient PZ lesions**	**mean dice coefficient TZ lesions**	**p-values (comparison PZ to TZ)**
CNN		0.17 (± 0.18)	0.26 (± 0.21)	0.22 (± 0.20)	0.23 (± 0.20)	0.18 (± 0.20)	0.179
prospective		0.51 (± 0.21)	0.51 (± 0.18)	0.51 (± 0.19)	0.52 (± 0.18)	0.46 (± 0.24)	0.067
reader 1		0.47 (± 0.19)	0.49 (± 0.16)	0.48 (± 0.17)	0.48 (± 0.17)	0.47 (± 0.20)	0.712
reader 2		0.48 (± 0.21)	0.48 (± 0.18)	0.48 (± 0.20)	0.47 (± 0.20)	0.50 (± 0.20)	0.277
reader 3		0.51 (± 0.20)	0.52 (± 0.15)	0.52 (± 0.18)	0.53 (± 0.16)	0.49 (± 0.21)	0.198
test for difference in Dice coefficient between sPC and non-sPC			**CNN**	**prospective**	**reader 1**	**reader 2**	**reader 3**
p-value			0.001	0.828	0.493	0.629	0.576
p-value adjusted for voxel ratio			0.300	0.506	0.333	0.481	0.749

the larger number of small CNN lesions does not favor the PZ or TZ specifically but affects both zones similarly.

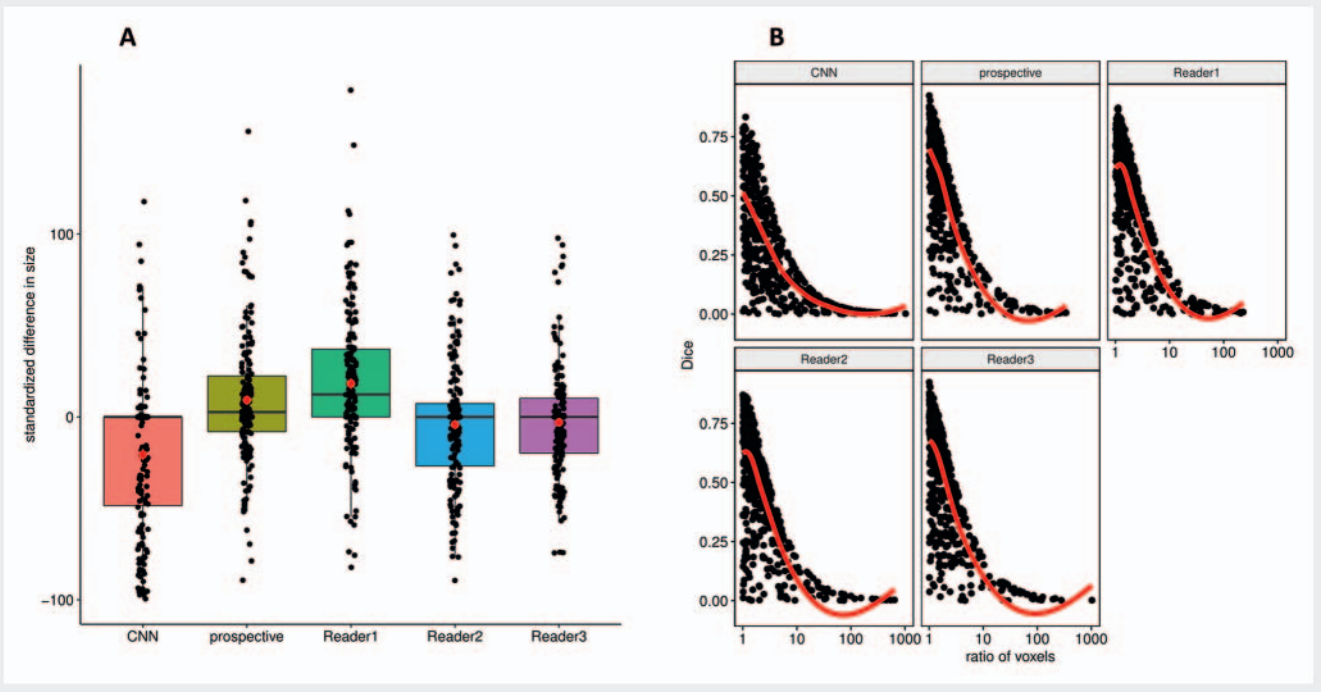
Discussion

In this study we compare the agreement in lesion segmentation on prostate MRI between 3 radiologists, retrospectively performed segmentations of clinically reported lesions and U-Net segmentations. We find that the mean Dice coefficients between manual segmentations are 0.48–0.52, establishing that radiologists do not completely agree about the boundaries of lesions that at least two readers consider reportable. This is in accordance with the known difficult task of visual prostate MRI interpretation which is also a source of inter-rater variability [18]. We find that a previously established automatic deep learning system (U-Net) which has achieved similar performance to clinical PI-RADS assessment [11] produces segmentations that are systematically smaller than segmentations of human raters. This smaller size bias of the CNN is a major contributor to lower segmentation agreement with Dice coefficients of 0.25–0.28 compared to human raters. U-Net thus appears to have a tendency to focus on the lesion core while radiologists use additional clues to also capture the lesion periphery. It is important to note that U-Net was trained using two-dimensional cross-entropy loss and not a Dice loss function. As such, training attempted to reproduce as much of the lesion area, but had to balance false-positive discoveries outside of the known tumor foci. As such, the training resulted in a delineation of the highest probability voxels within a lesion. In addition, the calibration of the U-Net which was performed on the validation set during model development to approximate the clinical performance at PI-RADS≥ 3 and ≥ 4 as closely as possible led to the selection of probability thresholds



▶ **Fig. 3** Histograms of lesion segmentation size for each reader. Bin size is 1000 voxels. The last bin (14 000–15 000 voxels) additionally contains all segmentations larger than 15 000 voxels. Top) Colors indicate PI-RADS category of lesions (see legend). Bottom) Colors indicate prostate zone. The histogram for U-Net (CNN) demonstrates a larger number of smallest lesions (< 1000 voxels) and less frequent occurrence of larger lesions compared to histograms of R1-R3 and CL (prospective). Human raters are more comparable to each other. CNN segments both lesions in the peripheral and transition zone smaller than human raters.

▶ **Abb. 3** Histogramme der Läsionssegmentierungsgröße für jeden Befunder. Die Bin-Größe beträgt 1000 Voxel. Das letzte Bin (14 000–15 000 Voxel) enthält zusätzlich alle Segmentierungen, die größer als 15 000 Voxel sind. Oben: Die Farben entsprechen der PI-RADS-Kategorie der Läsionen (siehe Legende). Unten: Die Farben entsprechen der zonalen Lokalisation der Läsionen (siehe Legende). Das Histogramm für U-Net (CNN) zeigt eine größere Anzahl kleinster Läsionen (< 1000 Voxel) und ein weniger häufiges Auftreten größerer Läsionen im Vergleich zu Histogrammen von R1-R3 und CL (prospektiv). Menschliche Bewerter sind besser miteinander vergleichbar. CNN segmentiert sowohl Läsionen in der peripheren als auch der Transitionszone kleiner als menschliche Befunder.

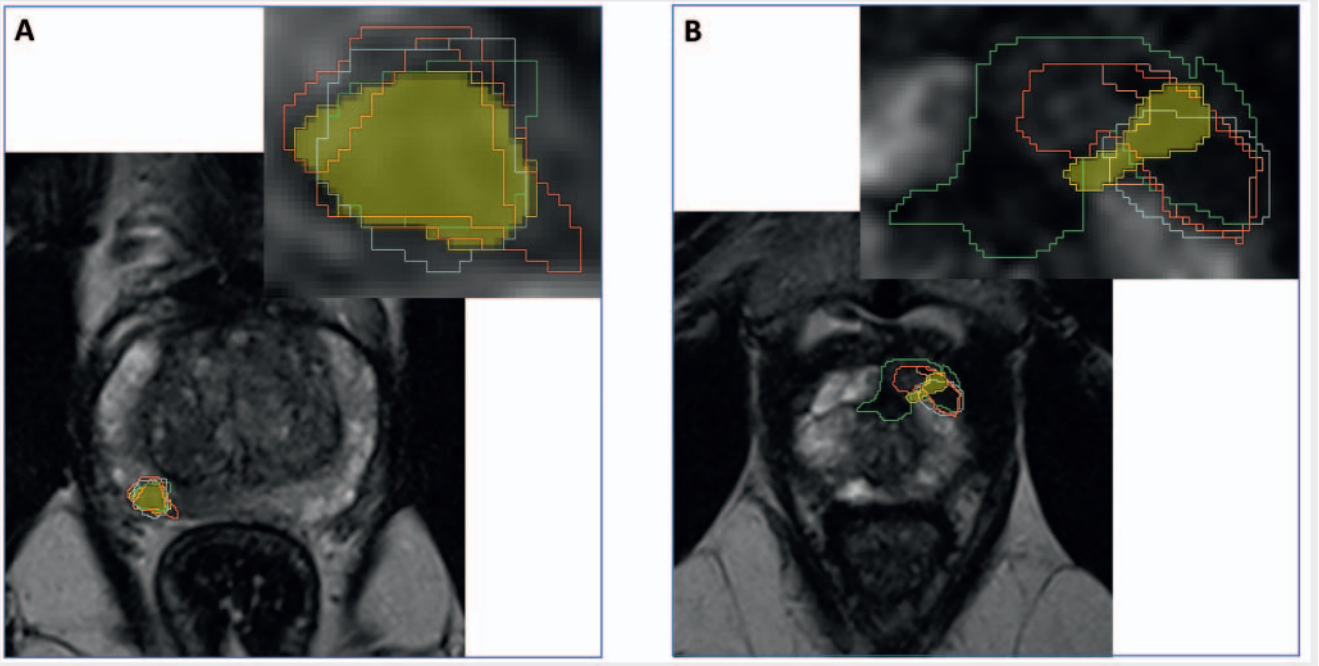


▶ **Fig. 4** **A** Normalized difference (average lesion size for each individual lesion location subtracted from all individual lesion measurements and divided by the average lesion size) shown as black dots for all lesions. Data are summarized as boxplots (see legend of Fig. 2 for further detail on boxplots). **B** Semi-logarithmic plot of the ratio of the number of voxels of the larger and smaller segmentation of each pair of overlapping lesions, given for each reference rater separately. Abscissa: logarithmic voxel ratio. Ordinate: Dice coefficient. Local regression lines in red indicate the trend of the data, showing that Dice coefficients decrease as lesion sizes are increasingly divergent. For lesions of comparable size (ratio close to one), the fit indicates Dice coefficients of nearly 0.5 for CNN and up to 0.7 for the other raters.

▶ **Abb. 4** **A** Normalisierte Differenz (durchschnittliche Läsionsgröße für jeden einzelnen Läsionsort, subtrahiert von allen einzelnen Läsionsmessungen und geteilt durch die durchschnittliche Läsionsgröße), dargestellt als schwarze Punkte für alle Läsionen. Die Daten werden als Boxp-Plots zusammengefasst (weitere Einzelheiten zu Box-Plots finden Sie in der Legende von Abb. 2). **B** Semi-logarithmisches Diagramm des Verhältnisses der Anzahl der Voxel der größeren und kleineren Segmentierungen jedes Paares überlappender Läsionen, separat angegeben für jeden Referenzbefunder. Abszisse: logarithmisches Voxelverhältnis. Ordinate: Dice-Koeffizient. Lokale Regressionslinien in Rot geben den Trend der Daten an und zeigen, dass die Dice-Koeffizienten abnehmen, wenn die Läsionsgrößen zunehmend divergieren. Für Läsionen vergleichbarer Größe (Verhältnis nahe 1) ergeben sich Dice-Koeffizienten von nahezu 0,5 für CNN und bis zu 0,7 für die anderen Befunder.

0.22 for emulation of PI-RADS≥ 3 decisions, and 0.33 for emulation of PI-RADS≥ 4 decisions. Most of the lesion area not currently included in CNN segmentations would become segmented if the threshold would be lowered below the 0.22 setting. However, this would increase the discovery of false-positive lesions elsewhere in the prostate and affect performance unless a local neighborhood region growing criterion were used to expand the lesion territory. Another likely important factor is the co-registration of DWI images to T2 used in U-Net training and predictions. U-Net reports those voxels that are suspicious on bi-parametric (T2w and DWI) co-registered data. If co-registration is not perfect, voxels of imperfect overlap will be excluded from the segmentation as these incorrectly co-register to a normal-appearing T2 signal with suspicious ADC findings or vice versa. Co-registration in prostate MRI is a known difficult challenge [29], as DWI and T2w images are often affected by significant misregistration due to DWI distortion by susceptibility interfaces to rectal air and bowel motion/varying rectal distension during the examination. The published U-Net used rigid followed by affine registration, while newer developments in our group are using optimized registration algorithms, e. g. b-spline registration and non-linear registration. It is possible that remaining misregistration accounts for some of the observed Dice coefficient reduction in this study. While lesion size is identified as an important factor resulting in reduced Dice coefficients for U-Net, it does not explain all of the differences, e. g. lesion size segmentation differences existed also within human readers. Reader 1 segmented on average larger lesions. However, there was no significantly reduced agreement between the human readers. One possible explanation is that CNN lesions are shifted relative to the common lesion core of human raters while human raters more consistently agree on the lesion center. Therefore, on average, the lesions of Reader 1 encompass the smaller lesions from the other human raters more concentrically than the CNN lesions are encompassed by all other lesions.

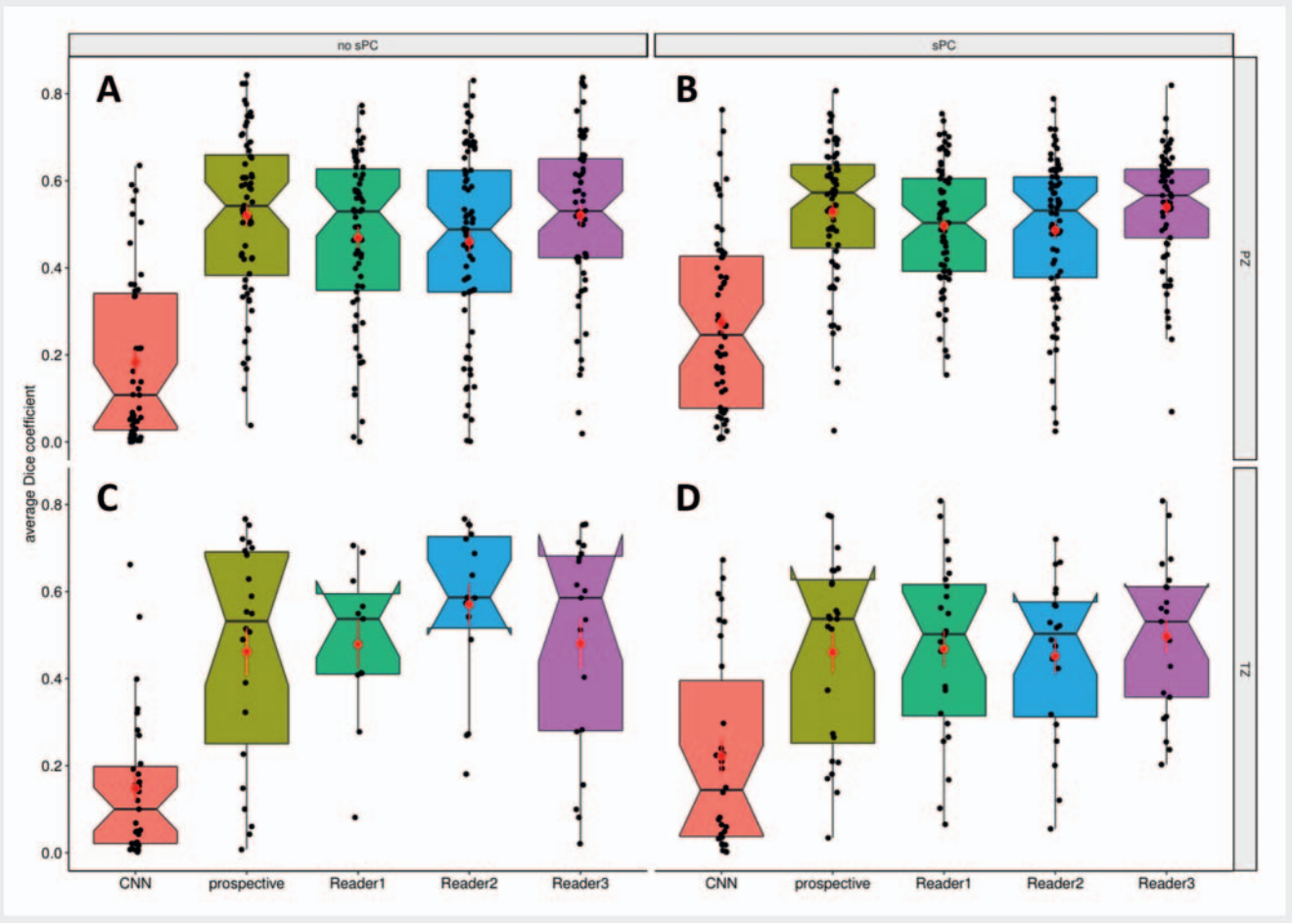
Dai et al. have previously evaluated a region proposal CNN (Mask R-CNN) to determine the Dice coefficient between radiologists and automated segmentation. They used a total of 120 patients (78 public, 42 private) with training of the lesion detection using 45 public patients and 21 private patients. These numbers are smaller compared to the 250 patients used for training of the



▶ **Fig. 5** Examples of the spectrum of overlap of multi-reader segmentations. ROIs are shown on one representative axial T2-weighted image of the prostate. CNN segmentation is shown as a yellow area, with segmentations of readers shown as outlines for clinical assessment (green), reader 1 (red), reader 2 (light blue) and reader 3 (orange) **A** 55-year-old patient with PSA of 12.0 ng/mL, no prior biopsy. Dice coefficient with CNN segmentation was 0.77 (clinical), 0.69 (reader 1), 0.69 (reader 2) and 0.71 (reader 3). **B** 65-year-old patient with PSA of 5.1 ng/mL, no prior biopsy. Dice coefficient with CNN segmentation was 0.08 (clinical), 0.11 (reader 1), 0.13 (reader 2) and 0.16 (reader 3). Targeted biopsy from both locations showed Gleason Grade Group 2 sPC. CNN segmentations were systematically smaller than segmentations performed by radiologists, with **B** representing one of the cases with the most pronounced difference.

▶ **Abb. 5** Beispiele für das Spektrum der Überlappung von Segmentierungen der verschiedenen Befunder. Die ROIs sind auf einem repräsentativen axialen T2-gewichteten Bild der Prostata dargestellt. Die CNN-Segmentierung wird als gelbe Fläche dargestellt, wobei die Segmentierungen der weiteren Befunder als Umrisse dargestellt sind: klinische Beurteilung (grün), Befunder 1 (rot), Befunder 2 (hellblau) und Befunder 3 (orange). **A** 55 Jahre alter Patient mit PSA 12,0 ng/ml, keine vorherige Biopsie. Der Dice-Koeffizient mit CNN-Segmentierung betrug 0,77 (klinisch), 0,69 (Leser 1), 0,69 (Leser 2) und 0,71 (Leser 3). **B** 65-jähriger Patient mit einem PSA von 5,1 ng/ml, keine vorherige Biopsie. Der Dice-Koeffizient mit der CNN-Segmentierung betrug 0,08 (klinisch), 0,11 (Befunder 1), 0,13 (Befunder 2) und 0,16 (Befunder 3). Gezielte Biopsien beider Läsionen ergaben Gleason Grade Group 2 sPC. CNN-Segmentierungen waren insgesamt systematisch kleiner als die Segmentierungen der Radiologen, wobei **B** ein besonders deutliches Beispiel darstellt.

currently evaluated CNN. Also, the U-Net approach used here represents a different paradigm than the region proposal approach in that it directly determines the segmentation from the data. It is not possible to determine which of the two (Mask R-CNN and U-Net) is better by comparison to the Dai et al. study. This would require comparison of both algorithms in the same cohort. Dice coefficients reported by Dai et al. in the private testing set were 0.38–0.46 compared to 0.28–0.31 in sPC lesions in this study. In our review, the public PROSTATEx-2 dataset has larger and clearer lesions than the consecutive clinical cohorts presenting in the clinical routine to our institution, thus explaining the higher Dice coefficients compared to our study. Dai et al. did not specify whether consecutive clinical patients were included, while it is important to reflect the clinical routine as closely as possible to arrive at clinically meaningful comparisons. In the present study it is an advantage that the CNN operated on the same consecutive patients that were seen by radiologists in the clinical routine. Dice coefficients have limited direct comparability between studies for several reasons. The distribution of lesions may differ markedly, e. g. patients may present with more advanced tumors to one institution and with more subtle imaging findings to another institution. Dai et al. did not report average PSA levels of the included patients. In fact, they provided almost no demographic or clinical information, which would partly have allowed better estimation of cohort differences. Also, the paradigm of lesion delineation may affect the Dice coefficient. In Dai et al. three radiologists delineated lesions using the histopathological results and clinical reports. As far as can be told, the three radiologists shared the work of delineation as no comparison of radiologist performance is given. The depth of comparison of lesion delineation in the current study is more profound, as three radiologists independently delineated lesions while performing independent re-reads of the clinical cases, thus reflecting the true clinical decision process, unbiased by knowledge of the biopsy results and previous reports. In addition, the prospective clinical performance was compared. In the current study the variability of such independent radiologist assessment can be directly compared to CNN performance, a novel insight which is



▶ **Fig. 6** Boxplots of per-lesion average Dice coefficient calculated between each reference reader and all remaining readers, stratified by peripheral and transition zone. The bold black line at the notch indicates the median, the ends of the box the 25th and 75th percentiles and the whiskers minimum and maximum. Black dots indicate individual data points. Red circles indicate the mean. Vertical red line segments indicate the standard error of the mean (SEM). **A** Boxplots depicting lesions without sPC in the peripheral zone **B** Boxplots depicting lesions with sPC in the peripheral zone. **C** Boxplots depicting lesions without sPC in the transition zone **D** Boxplots depicting lesions with sPC in the transition zone.

▶ **Abb. 6** Box-Plots des durchschnittlichen Dice-Koeffizienten pro Läsion, berechnet zwischen jedem Referenzbefunder und allen verbleibenden Befundern, stratifiziert nach Prostatazone. Die fette schwarze Linie an der Kerbe zeigt den Median an, die Enden der Box das 25. und 75. Perzentil und die Antennen die Minimal- und Maximalwerte. Gefüllte schwarze Kreise kennzeichnen einzelne Datenpunkte. Rote Kreise entsprechen dem Mittelwert. Vertikale rote Liniensegmente geben den Standardfehler des Mittelwerts (SEM) an. **A** Box-Plots mit Läsionen ohne sPC in der peripheren Zone. **B** Box-Plots mit Läsionen mit sPC in der peripheren Zone. **C** Box-Plots mit Läsionen ohne sPC in der Transitionszone. **D** Box-Plots mit Läsionen mit sPC in der Transitionszone.

of importance to correctly put CNN performance into context. Importantly, the diagnostic performance of the CNN evaluated in this study in terms of sPC detection has been previously shown [11]. The same diagnostic performance in terms of sPC detection was achieved even if the CNN segments lesions systematically smaller than the radiologists. As such, the CNN may focus more on the lesion core, while the radiologists focus more on outlining the entire lesion. It is important to note that while differences in voxel-wise agreement between radiologists and CNN exist, the resulting diagnostic performance is of primary importance such that the clinical value of a CNN cannot be determined only from examining metrics of voxel-wise agreement such as the Dice coefficient. While Dai et al. did not provide a zone-specific analysis, we also performed separate analyses in the TZ and PZ and found that the differences in Dice coefficients between CNN and the other readers are present at the same magnitude in both zones. In addition, the agreement with the other readers did not differ significantly in the PZ or TZ for any of the readers. While independence of quantitative mADC assessment [30] from zonal lesion location has been reported previously, and the diffusion component of the mpMRI examination has become more important in the most recent PI-RADS version 2.1 update [31], our findings provide further evidence that prostate zones do not govern differences in the lesion detection task neither for radiologists nor for the examined CNN. We note that sizes of lesions detected by radiologists were larger in the TZ compared to the PZ, likely reflecting the more difficult task of identifying small lesions in the often more heterogeneous TZ. This difference was absent for

CNN, indicating that CNN training did not result in the data-driven formation of a zone-specific lesion size filter representation in the network.

Lesion overlap Dice coefficient to multi-rater data as utilized in this study is surely an expensive measure as performance of segmentations by multiple radiologists is prohibitive in any large scale situation, where it is difficult enough to collect a single expert's segmentations. As such, it is of high interest to find that radiologists evaluated by the same metric as CNN do not perform much better than 0.5. This number can be regarded as a new benchmark for inter-rater agreement for the task of lesion boundary definition that can be used to evaluate if the lesion definitions agree with manual segmentations in the same way as different manual segmentations agree with each other. It is important to note that clinical usefulness is of primary concern when developing an AI system. In end-to-end AI systems [16], all that counts is that the entire processing from raw data to the clinically essential diagnostic information is included in a single trainable algorithm. Segmentation in itself is not mandatory to arrive at a prediction of the risk of a patient to harbor sPC. However, directly comparing segmentations as in the current study is crucial for AI explainability [32], since, especially for medical applications, it is important to assert that AI systems follow generally accepted patterns in medical decision making. As one finds in this study, differences may exist and require evaluation of their potential effect on the diagnostic utility of the method.

An important finding is that the lesion agreement according to Dice coefficient showed no significant changes when focusing only on sPC-positive lesions compared to sPC-negative lesions for the human raters, while U-Net demonstrated higher Dice coefficients in sPC-positive lesions before correction for lesion size mismatch, but not afterwards. For human raters this indicates that benign findings on prostate MRI can exactly mimic the appearance of cancer such that they are visually indistinguishable. For U-Net it is possible that smaller lesion segmentations in non-sPC situations reflect a stricter focus of U-Net on the lesion core which may in fact exhibit a size difference between sPC and non-sPC lesions, while human raters include more of the full lesion which may exhibit a similar size on average.

Our study has several limitations. The histopathological standard of reference was MR-TRUS fusion biopsy and not radical prostatectomy (RP). However, the sensitivity of the extended systematic and targeted biopsy performed here has previously been shown to detect 97 % of sPC compared to RP specimen [12]. In addition, only a cohort based on biopsies can encompass all men that are important to consider in a screening setting of men with suspicion for sPC. Any RP cohort will exclude many screening-relevant men, as it only focuses on men with biopsy-proven surgery-eligible sPC. The study design was retrospective. However, all eligible patients undergoing MRI and fusion biopsy in the inclusion interval were analyzed.

Conclusion

We find that human experts with various levels of training do not perfectly agree with respect to their delineation of lesions that they otherwise jointly detect, reflecting the difficulty of the visual task of lesion detection in the prostate. The examined CNN had a lower Dice coefficient compared to human raters than human raters compared to one another. Importantly, the performance of CNN in the prediction of sPC is not affected by this finding and has been validated to be comparable to clinical PI-RADS interpretation before. The remaining difference is not a primary training criterion but could be incorporated in future CNN training to assure more agreement of intermediate results such as segmentations that contribute to final diagnostic assessment. It may also be the result of U-Net focusing more on the lesion core and some contribution from remaining misregistration between T2 and DWI images, which would be large enough to affect the Dice coefficient but small enough not to impede successful lesion detection. Our study provides an example of comparative performance evaluation of CNN to human operators which may be used in addition to common comparison metrics such as precision and recall in future studies.

Conflict of Interest

Patrick Schelb, Anoshirwan Andrej Tavakoli, Teeravut Tubtawee, Thomas Hielscher have nothing to declare. Jan Philipp Radtke declares payment for consultant work from Saegeling Medizintechnik, Siemens Heathineers and for development of educational presentations from Saegeling Medizintechnik. Magdalena Görtz, Viktoria Schütz, Tristan Anselm Kuder, Lars Schimmöller have nothing to declare. Albrecht Stenzinger declares: Consulting fee and payment for lectures: Astra Zeneca, BMS, Novartis, Roche, Illumina, Thermo Fisher. Travel support: Astra Zeneca, BMS, Novartis, Illumina, Thermo Fisher. Board Member: Astra Zeneca, BMS, Novartis, Thermo Fisher. Markus Hohenfellner has nothing to declare. Heinz-Peter Schlemmer declares: Consulting fee, payment for lectures and travel support: Siemens Healthineers, Curagita, Profound, Bayer Vital. David Bonekamp declares. Consulting fee or honorarium: Bayer Vital, Payment for lectures: Bayer Vital.

References

[1] Radtke JP, Kuru TH, Boxler S et al. Comparative analysis of transperineal template saturation prostate biopsy versus magnetic resonance imaging targeted biopsy with magnetic resonance imaging-ultrasound fusion guidance. J Urol 2015; 193: 87–94. doi:10.1016/j.juro.2014.07.098

[2] Siddiqui MM, Rais-Bahrami S, Truong H et al. Magnetic resonance imaging/ultrasound-fusion biopsy significantly upgrades prostate cancer versus systematic 12-core transrectal ultrasound biopsy. Eur Urol 2013; 64: 713–719. doi:10.1016/j.eururo.2013.05.059

[3] Ahmed HU, El-Shater Bosaily A, Brown LC et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet (London, England) 2017; 389: 815–822. doi:10.1016/S0140-6736(16)32401-1

[4] Kasivisvanathan V, Rannikko AS, Borghi M et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. The New England journal of medicine 2018; 378: 1767–1777. doi:10.1056/NEJMoa1801993

[5] Bonekamp D, Schelb P, Wiesenfarth M et al. Histopathological to multiparametric MRI spatial mapping of extended systematic sextant and MR/TRUS-fusion-targeted biopsy of the prostate. European radiology 2018. doi:10.1007/s00330-018-5751-1

[6] Stabile A, Dell'Oglio P, De Cobelli F et al. Association Between Prostate Imaging Reporting and Data System (PI-RADS) Score for the Index Lesion and Multifocal, Clinically Significant Prostate Cancer. Eur Urol Oncol 2018; 1: 29–36. doi:10.1016/j.euo.2018.01.002

[7] Padhani AR, Weinreb J, Rosenkrantz AB et al. Prostate Imaging-Reporting and Data System Steering Committee: PI-RADS v2 Status Update and Future Directions. Eur Urol 2018; 75: 385–396

[8] Weinreb JC, Barentsz JO, Choyke PL et al. PI-RADS Prostate Imaging – Reporting and Data System: 2015, Version 2. Eur Urol 2016; 69: 16–40. doi:10.1016/j.eururo.2015.08.052

[9] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In, Advances in neural information processing systems 2012: 1097–1105

[10] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In, International Conference on Medical image computing and computer-assisted intervention Springer. 2015: 234–241

[11] Schelb P, Kohl S, Radtke JP et al. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. Radiology 2019; 293: 607–617. doi:10.1148/radiol.2019190938

[12] Yoo S, Gujrathi I, Haider MA et al. Prostate Cancer Detection using Deep Convolutional Neural Networks. Scientific reports 2019; 9: 19518. doi:10.1038/s41598-019-55972-4

[13] Sanford T, Harmon SA, Turkbey EB et al. Deep-Learning-Based Artificial Intelligence for PI-RADS Classification to Assist Multiparametric Prostate MRI Interpretation: A Development Study. J Magn Reson Imaging 2020. doi:10.1002/jmri.27204

[14] Bonekamp D, Jacobs MA, El-Khouli R et al. Advancements in MR imaging of the prostate: from diagnosis to interventions. Radiographics 2011; 31: 677–703. doi:10.1148/rg.313105139

[15] Campanella G, Hanna MG, Geneslaw L et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019; 25: 1301–1309. doi:10.1038/s41591-019-0508-1

[16] Wang Z, Liu C, Cheng D et al. Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network. IEEE Trans Med Imaging 2018; 37: 1127–1139. doi:10.1109/TMI.2017.2789181

[17] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945; 26: 297–302

[18] Greer MD, Brown AM, Shih JH et al. Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI: A multireader study. J Magn Reson Imaging 2017; 45: 579–585

[19] Barentsz JO, Richenberg J, Clements R et al. ESUR prostate MR guidelines 2012. European radiology 2012; 22: 746–757. doi:10.1007/s00330-011-2377-y

[20] Rothke M, Blondin D, Schlemmer HP et al. [PI-RADS classification: structured reporting for MRI of the prostate]. Rofo 2013; 185: 253–261. doi:10.1055/s-0032-1330270

[21] Radtke JP, Schwab C, Wolf MB et al. Multiparametric Magnetic Resonance Imaging (MRI) and MRI-Transrectal Ultrasound Fusion Biopsy for Index Tumor Detection: Correlation with Radical Prostatectomy Specimen. Eur Urol 2016; 70: 846–853. doi:10.1016/j.eururo.2015.12.052

[22] Kuru TH, Wadhwa K, Chang RT et al. Definitions of terms, processes and a minimum dataset for transperineal prostate biopsies: a standardization approach of the Ginsburg Study Group for Enhanced Prostate Diagnostics. BJU Int 2013; 112: 568–577. doi:10.1111/bju.12132

[23] Fritzsche KH, Neher PF, Reicht I et al. MITK diffusion imaging. Methods Inf Med 2012; 51: 441–448. doi:10.3414/ME11-02-0031

[24] Nolden M, Zelzer S, Seitel A et al. The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development. Int J Comput Assist Radiol Surg 2013; 8: 607–620. doi:10.1007/s11548-013-0840-8

[25] Kuru TH, Wadhwa K, Chang RTM et al. Definitions of terms, processes and a minimum dataset for transperineal prostate biopsies: a standardization approach of the G insburg S tudy G roup for E nhanced P rostate D iagnostics. BJU international 2013; 112: 568–577

[26] Egevad L, Delahunt B, Srigley JR et al. International Society of Urological Pathology (ISUP) grading of prostate cancer – An ISUP consensus on contemporary grading. APMIS 2016; 124: 433–435. doi:10.1111/apm.12533

[27] Team RC. R: A language and environment for statistical computing. Vienna, Austria, 2013

[28] Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Radiology 2003; 226: 24–28

[29] Litjens G, Toth R, van de Ven W et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Medical image analysis 2014; 18: 359–373

[30] Bonekamp D, Kohl S, Wiesenfarth M et al. Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values. Radiology 2018; 289: 128–137. doi:10.1148/radiol.2018173064

[31] 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

[32] Gunning D. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web 2017; 2: 2