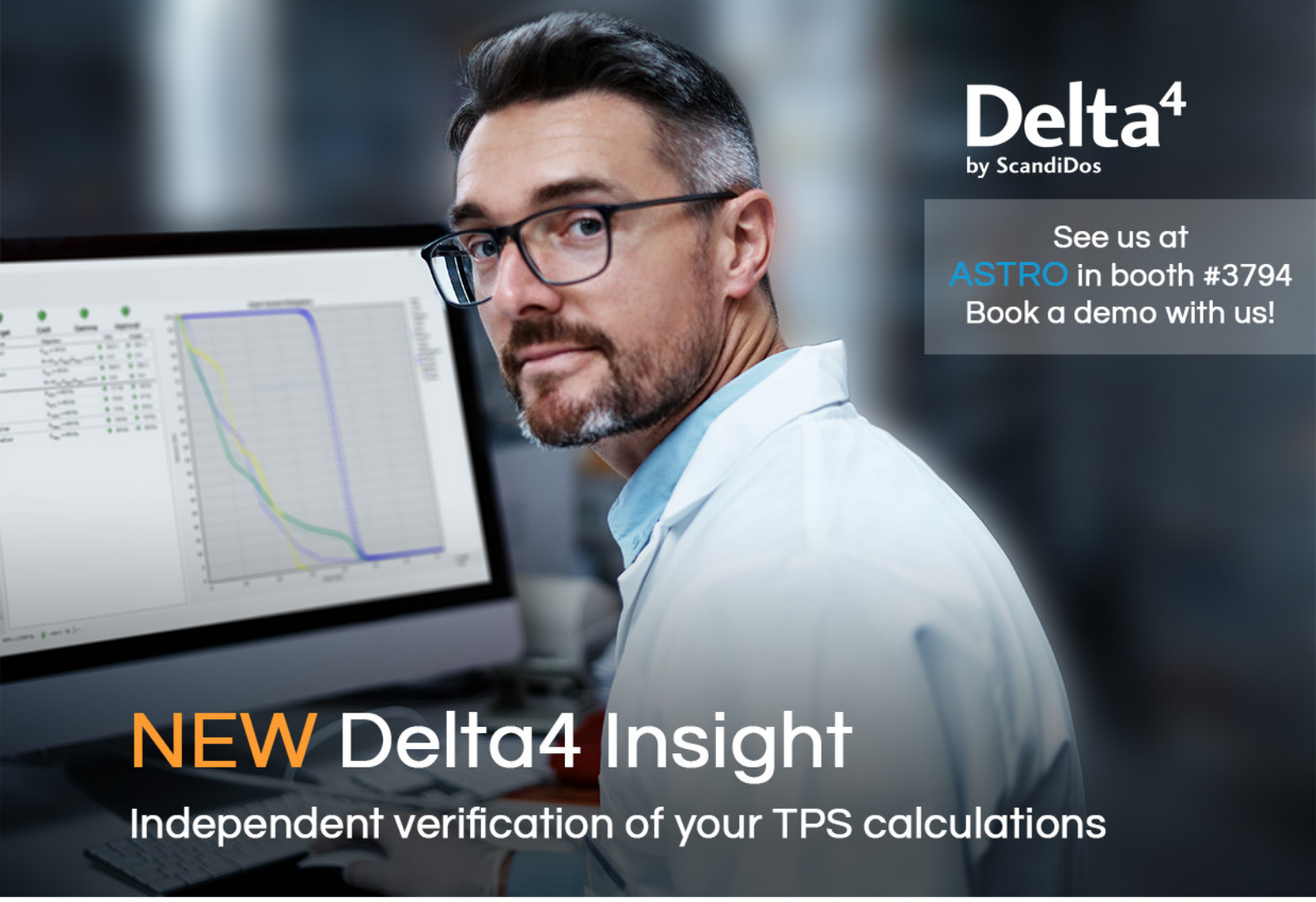


We provide medical physicists with patient QA from prescription to final fraction!

Delta4 Insight* is an independent secondary 3D dose calculation software that utilizes a Monte Carlo engine to verify the calculations of a clinic's Treatment Planning System (TPS).

"Our ambition when developing Delta4 Insight was to extend our Delta4 family of products into a full end-to-end verification solution where the customer can be confident that Delta4 provides the true values and not just the deviations."
Görgen Nilsson, CTO, and founder

Delta4 Insight completes the ScandiDos patient QA product portfolio which now covers - patient QA from prescription to final fraction.

*Currently available together with Delta4 Phantom+.
Not available for all markets, contact ScandiDos for more information.

Learn more ▶

delta4family.com

Multicenter validation of prostate tumor localization using multiparametric MRI and prior knowledge

Cuong Viet Dinh, Peter Steenbergen, Ghazaleh Ghobadi, Henk van der Poel, Stijn W.T.P.J. Heijmink, and Jeroen de Jong
The Netherlands Cancer Institute, Amsterdam, The Netherlands

Sofie Isebaert, Karin Haustermans, Evelyne Lerut, and Raymond Oyen
University of Leuven, University Hospitals Leuven, Leuven, Belgium

Yangming Ou
Massachusetts General Hospital, Boston, MA, USA

Davatzikos Christos
University of Pennsylvania, Philadelphia, PA, USA

Uulke A. van der Heide[a)]
The Netherlands Cancer Institute, Amsterdam, The Netherlands

(Received 25 May 2016; revised 12 December 2016; accepted for publication 26 December 2016; published 16 March 2017)

Purpose: Tumor localization provides crucial information for radiotherapy dose differentiation treatments, such as focal dose escalation and dose painting by numbers, which aim at achieving tumor control with minimal side effects.

Multiparametric (mp-)MRI is increasingly used for tumor detection and localization in prostate because of its ability to visualize tissue structure and to reveal tumor characteristics. However, it can be challenging to distinguish cancer, particularly in the transition zone. In this study, we enhance the performance of a mp-MRI-based tumor localization model by incorporating prior knowledge from two sources: a population-based tumor probability atlas and patient-specific biopsy examination results. This information typically would be considered by a physician when carrying out a manual tumor delineation.

Materials and methods: Our study involves 40 patients from two centers: 23 patients from the University Hospital Leuven (Leuven), Leuven, Belgium and 17 patients from the Netherlands Cancer Institute (NKI), Amsterdam, the Netherlands. All patients received a mp-MRI exam consisting of a T2-weighted, diffusion-weighted, and dynamic contrast-enhanced MRI before prostatectomy. Thirty-one features were extracted for each voxel in the prostate. Among these, 29 were from the multiparametric-MRI, one was from the population-based tumor probability atlas and one from the biopsy map. T2-weighted images of each patient were registered to whole-mount section pathology slices to obtain the ground truth. The study was validated in two settings: single-center (training and test sets were from the same cohort); and cross-center (training and test sets were from different cohorts). In addition, automatic delineations created by our model were compared with manual tumor delineations done by six different teams on a subset of Leuven cohort including 15 patients.

Results: In the single-center setting, mp-MRI-based features yielded area under the ROC curves (AUC) of 0.690 on a pooled set of patients from both cohorts. Including prevalence into mp-MRI-based features increased the AUC to 0.751 and including all features achieved the best performance with AUC of 0.775. Using all features always showed better results when varying the size of the training set. In addition, its performance is comparable with the average performance of six teams delineating the tumors manually. The error rate using all features was 0.22. The two prior knowledge features ranked among the top four most important features out of the 31 features.

In the cross-center setting, combining all features also yielded the best performance in terms of the mean AUC of 0.777 on the pooled set of patients from both cohorts. In addition, the difference in performance between the single-center setting and cross-center setting was not significant.

Conclusions: The results showed significant improvements when including prior knowledge features in addition to mp-MRI-based features in both single- and cross-center settings. *© 2016 The Netherlands Cancer Institute-Antoni van Leeuwenhoek. Medical Physics published by Wiley periodicals, Inc. on behalf of American Association of Physicists in Medicine.* [https://doi.org/10.1002/mp.12086]

Key words: multiparametric MRI, prior knowledge, prostate tumor localization, registration between MRI and H&E staining image

1. INTRODUCTION

The current standard in radiotherapy of prostate cancer is to treat the entire gland to a homogeneous dose. Recent improvements in dose delivery techniques and imaging methods facilitate the exploration of focal dose differentiation, such as focal dose escalation and dose painting by numbers.[1] These treatment strategies aim at controlling the tumor with minimal side effects by giving a high dose to the tumor area and a lower dose to the rest of the gland. Accurate delineation of prostate tumors is therefore crucial for these treatment options.

Structural and functional MRI modalities, such as T2-weighted (T2w-), diffusion-weighted imaging (DWI-), and dynamic contrast-enhanced (DCE-) MRI are increasingly used for tumor detection and localization.[2] Accurate tumor delineation is, however, still challenging as visually interpreting these images is both labor-intensive and prone to interobserver variability. In the study by Steenbergen et al.,[3] six teams consisting of a radiation oncologist and a radiologist delineated tumors on mp-MRI of 20 prostate patients. The kappa indices for the agreement between the delineations of the teams were quite low: 0.61 ± 0.19 (mean $\pm$ standard deviation). Several automatic tumor localization models have therefore been proposed to make this interpretation easier and more robust.[4–6]

A common approach to build tumor localization models is to extract first the relevant features for each voxel from one or a few modalities and then classify these voxels into normal and tumor tissue using basic classifiers such as support vector machines or logistic regression. Viswanath et al.[5] used texture features such as Gabor wavelet and Haar wavelet transformation extracted from T2w-MRI scans to represent each voxel. Groenendaal et al.[4] represented each voxel by several local statistics, e.g. minimum, maximum, and median of intensities obtained on apparent diffusion coefficient (ADC) maps and volume transfer constant K^{trans} maps, which are derived from the DWI-MRI and DCE-MRI modalities respectively. Another approach for tumor localization is to detect regions of interest in prostate first using, for example, ADC map-based blob detection[6] or clustering,[7] and then classifying the detected regions into normal and tumor regions. Despite its success, mp-MRI-based tumor localization has some limitations. It is hard to distinguish prostate cancer from confounders such as benign prostatic hyperplasia (BPH), postbiopsy hemorrhage, and atrophy.[8,9]

Tumors are not distributed equally within the prostate.[10,11] Ou et al.[10] constructed a tumor atlas, which is in fact a statistical map of the spatial probability distribution of prostate cancer based on 158 prostatectomy specimen. The value of each voxel in the tumor probability atlas represents the number of specimen having tumor at that voxel. Figure 1 shows an example of a slice of the tumor probability atlas. The higher values on the left and right sides of the peripheral zone indicated higher tumor probabilities. Rusu et al.[11] extended the work of Ou et al.[10] further by constructing a population-based atlas which integrated both histology and imaging information. A sophisticated registration method between T2w and the tumor atlas was proposed, which took into account MRI intensity as well as anatomical information, such as transition zone and peripheral zone area, in the registration process.

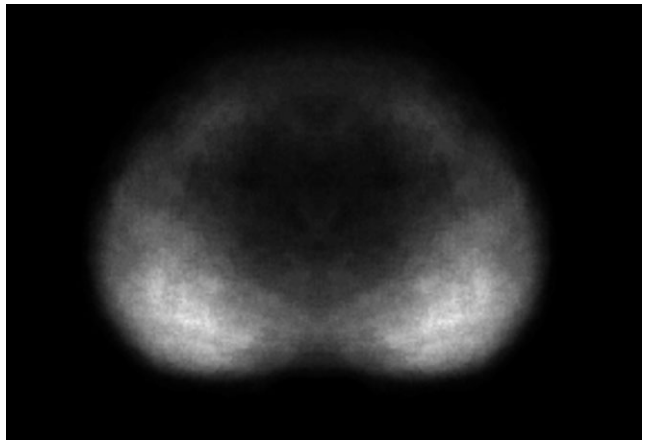


FIG. 1. A slice of the tumor prevalence map in the transversal view.

In this study, we propose a method to improve the performance of a mp-MRI-based tumor localization for individual patients by incorporating two sources of prior knowledge, which are typically available to a physician delineating a tumor manually. The first source is the tumor atlas constructed in[10] and the second is a patient-specific map indicating tumor-positive regions based on the results of a TRUS-based biopsy procedure.

2. MATERIALS AND METHODS

2.A. Patient characteristics

The study involves 23 patients from the University Hospital Leuven, Belgium (Leuven) and 17 patients from Netherlands Cancer Institute (NKI), Amsterdam, the Netherlands. All patients received an MRI exam prior to prostatectomy. Table I shows patient characteristics of the two cohorts.

For Leuven cohort, there are 25 and 7 tumors located in the peripheral zone and transition zone, respectively. The volumes in terms of mean and standard deviation were 2.91 ± 2.52 cm^3 for the peripheral zone and 4.29 ± 2.92 cm^3 for the transition zone. For NKI cohort, there are 18 and 5 tumors located in the peripheral zone and transition zone, respectively. The volumes in terms of mean and standard deviation were 1.86 ± 0.1 cm^3 for the peripheral zone and 2.86 ± 2.47 cm^3 for transition zone. Tumors with volume smaller than 0.5 cm^3 were not counted. In general, tumors in NKI cohort are smaller than those in Leuven cohort. Tumors in the transition zone are bigger than those in the peripheral zone.

Twenty patients from the Leuven cohort were previously involved in another study done by Steenbergen et al.[3] of which the main focus was on interobserver variability of prostate tumor delineation. In this study, six teams consisting

TABLE I. Characteristics of patients from Leuven and NKI cohorts. Higher Gleason score and pathology stage shows higher tumor aggressiveness.

Patient characteristics	Leuven	NKI
No. of patients	23	17
Gleason score		
3 + 3	0	2
3 + 4	2	11
4 + 3	11	3
4 + 4	7	1
4 + 5	1	0
5 + 4	2	0
T-stage		
T1	0	0
T2b	0	1
T2c	9	10
T3a	8	4
T3b	6	2

of a radiation oncologist and a radiologist delineated tumors on mp-MRI. From the 20 patients, biopsy reports were not available for five of these. Thus, only 15 patients from the Leuven cohort were included in our analysis described in Section 4 to compare between manual and automatic tumor delineations.

2.B. MRI Protocols

Patients received an mp-MRI exam consisting of a T2w, DWI-MRI, and DCE-MRI scans. Although both protocols were consistent with recently published guidelines,[2] there were significant differences in imaging protocol and acquisition between the two institutes: at Leuven,[12] the exam was performed on a 1.5T MRI (Siemens SONATAVision) scanner with a combination of a six-channel phased array body coil and a spine coil. At NKI, the exam was performed on a 3T MRI (Philips Achieva) scanner with a six-channel phased array coil in combination with an endorectal coil. Details of the MRI protocols for the two centers are given in Table II. Particularly for the DCE scans, the protocols at the two centers differ from each other in terms of interval time and scan duration.

ADC maps were derived from DWI scans using a mono-exponential model.[13] Similarly, K^{trans} maps were derived from DCE scans following the generalized kinetic model.[14] For Leuven, a constant precontrast T1 value of 1434 ms was used[15] as T1 mapping was not included in the MRI exams.

2.B.1. Image normalization

As there were significant differences in imaging protocol and acquisition between the two centers, image normalization is used to reduce these differences. In this study, we normalized the T2w and K^{trans} maps to the median value in the peripheral zone following.[4] A T2w image is known to provide no quantitative measure. K^{trans} is in theory quantitative but the values tend to vary among patients and institutes. For ADC maps, normalization was not necessary.

2.C. Feature representation

To combine information from different parametric maps, the T2w, ADC, and K^{trans} images from both centers were resampled to the grid with voxel size $0.49 \times 3.3 \times 0.49$ mm^3, corresponding to the T2w images from Leuven. The first and third dimensions of the grid correspond to the transverse plane and the second dimension corresponds to the slice direction of the scan.

Each voxel was represented by features derived from the mp-MR scan (29 features) and prior knowledge data (two features). Table III summarizes all extracted features.

2.C.1. MRI-based features

Intensity ($f_1..f_3$): Tumors typically appear dark on T2w and ADC and bright on K^{trans}.[2] For each voxel, intensity values of the normalized T2w, ADC, and normalized K^{trans} maps were included.

Textural features ($f_4..f_{29}$): Tumors may exhibit different textural characteristics than normal tissues. For example, on T2w images tumors often exhibit a so-called "erased charcoal sign", a smudge-like dark texture, and appear to have a blob shape on ADC and K^{trans} images.[6,8] In this study, we adopted the textural features ($f_4..f_{27}$) proposed by Litjes et al.[8] and Vos et al.[6] We extracted from the T2w image the Gaussian derivatives up to second order at four exponentially increasing scales (σ = 1.5; 2.4; 3.8, and 6.0 mm). Normalized multiscale blobness features[16] over the same four scales, f_{28} and f_{29}, were calculated on the ADC and K^{trans} images, respectively.

2.C.2. Prior knowledge-based feature representation

Prevalence map (f_{30}): We used the tumor atlas introduced by Ou et al.[10] The value of each voxel in this tumor atlas corresponds to the number of specimen having tumor at that voxel.

To transfer this population-based tumor probability atlas to each specific patient, we registered it to the patient's T2w image using an implementation of the b-spline deformation algorithm described previously.[17] First, the volumes of the tumor probability atlas and the prostate were manually delineated and converted into binary masks. These masks were then registered using the normalized cross-correlation (NCC) similarity measure with a regularization term, which minimizes the bending energy in the deformation, to minimize unrealistic deformations inside the binary masks. The registration used a gradient

TABLE II. Scan protocols from Leuven and NKI.

Center	Sequence	TR/TE [ms]	No of slices	Voxel dimension [mm^3]	Remark
Leuven	T2w	7120–13550/124–136	21–56	0.49 × 3.3 × 0.49 or 0.58 × 3.0 × 0.58	
	DWI-MRI	4000–9900/67–83	24–42	2.97 × 5.0 × 2.97 or 2.73 × 4.0 × 2.73	*b*-values: 0, 50, 100, 500, 750, 1000 (s/mm^2). A B0 map, which allows correcting for geometrical distortions, is not available.
	DCE-MRI	4.65–7.36/1.56–3.6	14	1.37 × 4.0 × 1.37	Interval time: 9 s Scan duration: 144 s (16 frames in total). No T1 mapping was available
NKI	T2w	3140–3626/120	25	0.27 × 3.0 × 0.27	
	DWI-MRI	3453–3492/67.8–69	20	1.03 × 2.7 × 1.03	*b*-values: 0, 188, 375, 563, 750 (s/mm^2) A B0 map is available.
	DCE-MRI	4/1.9	20	1.02 × 3.0 × 1.02	Interval time: 2.5 s Scan duration: 300 s (120 frames in total) T1 mapping using the variable flip angle method;[15] Flip angles were corrected for B1 inhomogeneities.

TABLE III. Feature representation based on mp-MRI and prior knowledge.

Feature ID	Feature name	Description
f_1	Normalized T2w	Median normalization is used
f_2	ADC intensity	No normalization is used
f_3	Normalized K^{trans}	Median normalization is used
f_4–f_{27}	T2w texture features	Gaussian derivatives of T2w up to 2nd order with four scales σ = 1.5, 2.4, 3.8, 6.0 mm. Number of features per scale is six
f_{28}	ADC blobness	Multiscale blobness of ADC map calculated over the same above four scales
f_{29}	K^{trans} blobness	Multiscale blobness of K^{trans} map calculated over the same above four scales
f_{30}	Prevalence map	Obtained by registering the population-based tumor probability atlas to T2w image
f_{31}	Biopsy map	Obtained from the biopsy report

descent-based multiresolution approach with a final control point field space, i.e., the space between control points in the deformable registration, of 5 mm. Finally, the resulting transformation was applied to the tumor probability atlas. The transformed atlas, hereafter referred to as prevalence map, was used as an additional feature for voxel representation.

Ultrasound biopsy map (f_{31}): Ultrasound-guided biopsies are used by urologists to confirm prostate cancer and to derive an estimate of the tumor location for staging purpose. There is substantial variation in the way biopsy results are reported. In some institutes, the location of each biopsy is reported systematically. However, for most institutes only the numbers of positive biopsies and the total number of biopsies at each side (left/right) of the prostate are reported. This is the case for the patients from NKI. At Leuven, on each side, three biopsies were taken from the peripheral zone (toward the apex, middle, and base areas) and two biopsies from the transition zone (toward the base and apex) and the tumor state in each core was recorded [Fig. 2(a)].[18] Figure 2(b) shows an example of biopsy sections indicated by different grayscale levels in three directions (left: axial view, top right: coronal view, and bottom right: sagittal view). For this patient, the tumor area indicated by the white contour appeared on the left, middle peripheral zone section according to the biopsy scheme.

We reconstructed a biopsy map on the T2w image by assigning voxel values at each side of the prostate as the ratio of the number of positive biopsies to the total number of taken biopsies on that side. The reconstructed biopsy map was smoothed by convolving it with a Gaussian kernel (σ = 4.5) to consider the fact that the biopsy status was less certain for voxels close to the midline of the prostate. The smoothed biopsy map was used as a feature for voxel representation.

2.D. Registration between H&E-stained slice and T2w image

Hematoxylin and eosin (H&E)-stained slices of the prostatectomy specimen were used as ground truth. Therefore, they were registered to the T2w images. This was done via slice matching and point matching steps by three independent observers, who then combined their results to achieve consensus.

2.D.1. Slice matching

A T2w slice was manually assigned to each delineated H&E slice taking into account: (a) the relative order of the slices, (b) the location of apex and base of the prostate, and (c) the relative size and shape of the subsequent H&E and T2w slices.

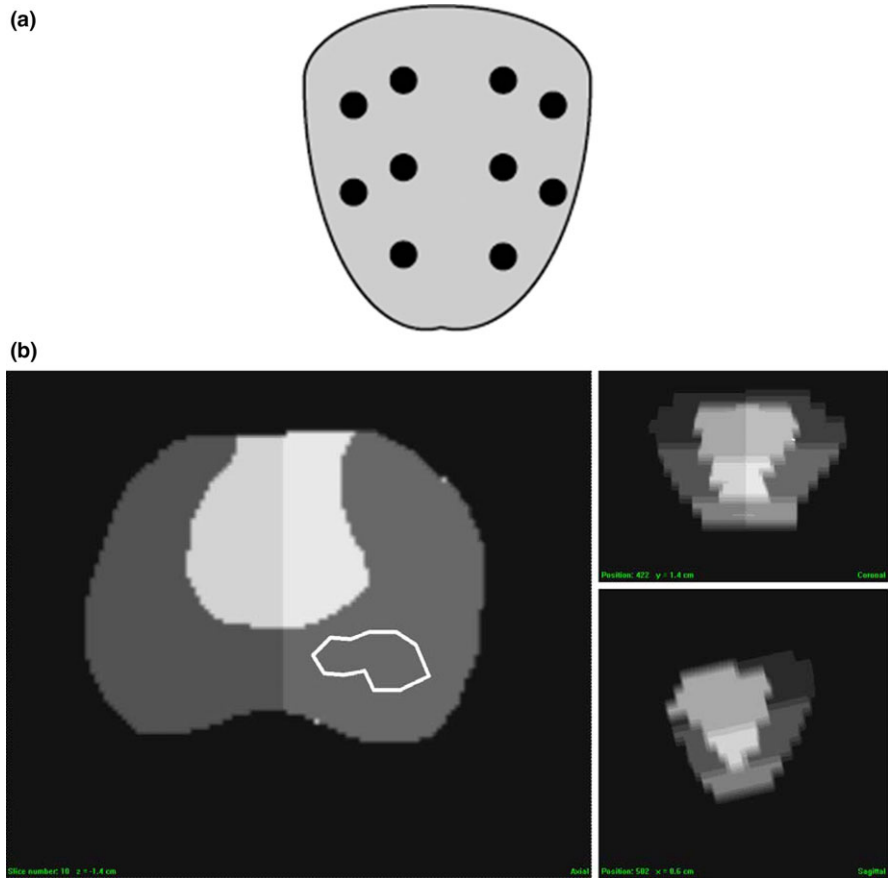


FIG. 2. (a) Biopsy locations of a ten-cores biopsy scheme in the coronal view from Leuven center. (b) Biopsy sections indicated by different grayscale levels in three different directions (left: axial view, top right: coronal view, and bottom right: sagittal view). For this patient, tumor marked by a white contour appeared on the left, middle peripheral zone section according to the biopsy scheme. [Colour figure can be viewed at wileyonlinelibrary.com]

2.D.2. Point matching

Each H&E slice was then registered to its matched T2w slice using a deformable method based on landmark points (Coherent Point Drift).[19] We selected landmark points that were visible on both images from the prostate boundary and features such as the transitions between prostate and seminal vesicles. After registration, the tumor delineations on the H&E slice were transferred to the MRI scan. We estimated the registration error by selecting one landmark, which was mostly the urethra, per pathology slide and measuring the distance between this point in the T2w MRI and registered pathology slide. Sample images of selected landmarks are shown in Fig. 3.

The average errors we found were 2.1 mm for the Leuven datasets and 2.6 mm for the NKI datasets. The largest error in both databases was of 5 mm. The average error was slightly larger for the NKI datasets, which may be attributed to the use of an endorectal coil at NKI, which induces tissue deformations in the MRI scan in the prostate area close to the rectum.

To reduce the influence of registration errors, we ignored voxels within $\pm$ 1.25 mm margin (margin's width: 2.5 mm) of the pathological tumor contours when constructing the model in both centers.

2.E. Model creation

After features were extracted for each voxel, a model was first created by fitting a logistic regression model to the data of a training cohort of patients in which each voxel was assigned a normal/tumor tissue label obtained from the delineations of a pathologist on the H&E-stained slices of the prostatectomy specimen. The trained model was then applied to the data of a validation cohort of patients to estimate an individual tumor probability map for each of the patients. The tumor probability map can be converted into a tumor segmentation by applying a threshold t, i.e., assigning voxels with tumor probability larger than t to the tumor class.

2.F. Model general setting

2.F.1. Single-center and cross-center validation

The performance of the feature options was validated in both single-center setting and cross-center setting. In the single-center setting, a leave-one-dataset-out cross-validation was used to separate training and test sets. In the cross-center setting, all patients from a center were used to fit the model.

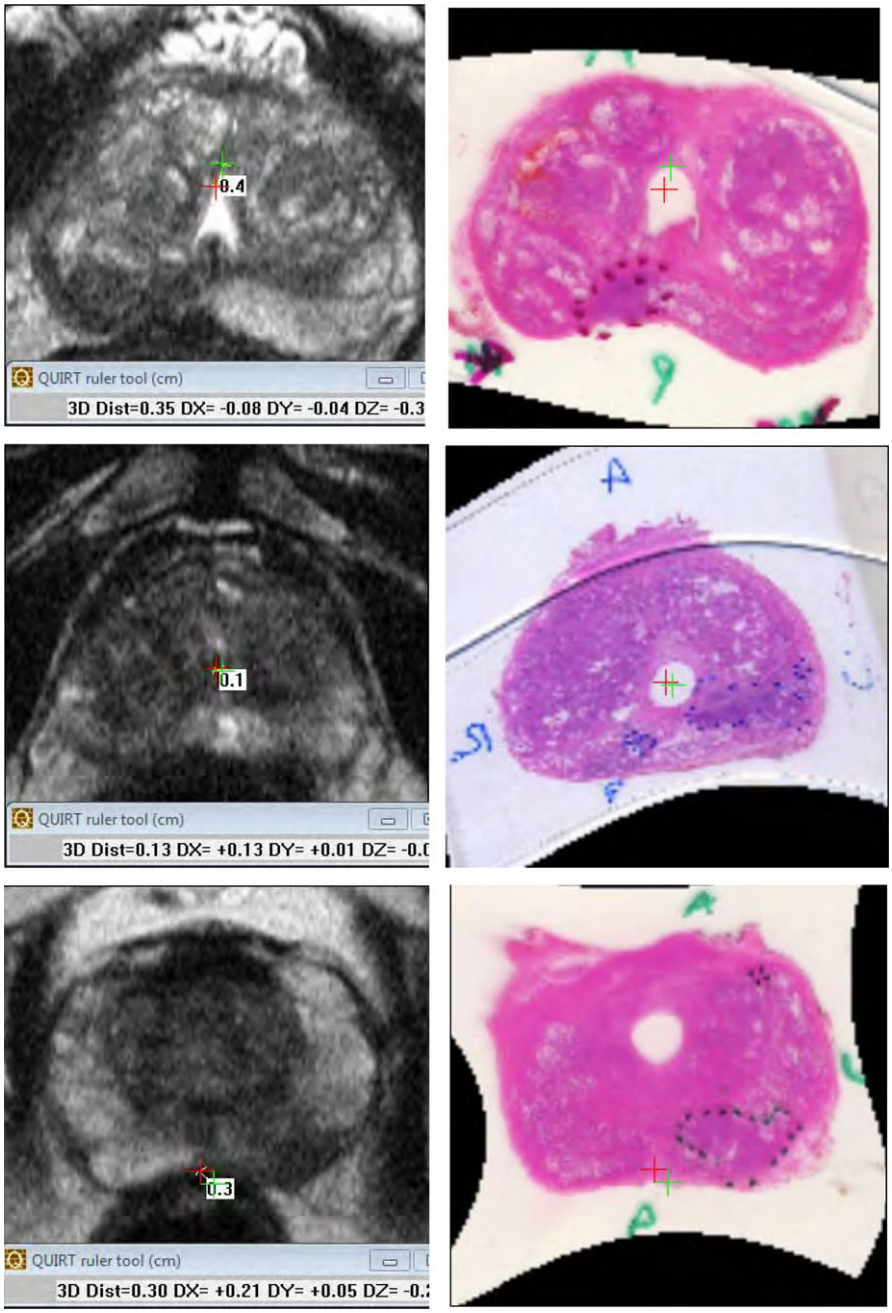


FIG. 3. Examples of selected landmarks for measuring registration error. Landmarks based on T2w and pathology images are marked by crosses (in red and green, respectively, in online version). Numbers indicates the distance between them. [Colour figure can be viewed at wileyonlinelibrary.com]

2.F.2. Metrics for evaluation

We evaluated the performance of tumor localization models on the datasets from the two cohorts. The area under the curve (AUC) was used to assess the model performance for each patient. Average performance over the pooled set of patients from both cohorts was reported. This is justified as we are interested in the effect of adding information from prevalence and biopsy and less interested in the difference between institutes. The AUC was used as it is known to handle unbalanced class situations well. This is the case in our prostate tumor localization problem as the relative tumor area is quite small in some patients.

The threshold to convert from the tumor probability map to a tumor delineation map was chosen as the optimal point on the ROC curve. The optimal point was defined as the closest point to the top left corner of the ROC curve, based on the training data. Performances of automatic and manual delineations were measured in terms of error rate, which is defined as the average classification error from both normal and tumor tissue classes. AUC was not used for comparison because it is not possible to calculate AUC for manual

delineations which only provide a binary decision (not probabilistic decision) at a voxel level.

The Wilcoxon signed-rank test with a Bonferroni multiple comparison correction was used to evaluate whether the difference in performance between two models at patient level was statistically significant. There were nine tests in total in our study. Thus, the new significant level (*P*-value) after Bonferroni correction is 5.5×10^{-3}. The tests used in our study and their corresponding *P*-values are shown in Table V of the Appendix S1.

2.F.3. Feature combination options

To study the effect of including prior knowledge features, three feature combination options were compared: (a) the MRI features only (MRI); (b) combining MRI features with feature from the prevalence map (MRI_prevalence); and (c) all features, i.e., combining MRI features with features from the prevalence map and biopsy map (MRI_prevalence_biopsy). In addition, we applied a commonly used forward feature selection method[20] to reveal the features that contribute the most to the model performance when all features were included (MRI_prevalence_biopsy option). The feature selection procedure starts from an empty set. It then sequentially adds the feature that maximizes a predefined criterion when combined with the feature set that has already been selected. In this study, the Mahalanobis distance between two classes (normal vs. tumor)[21] was used as the criterion to select features in the feature selection procedure.

2.F.4. Learning curve

We investigated the robustness of the three feature combinations with respect to various number of training data sizes by generating a learning curve. This was done in the single-center setting.

The database at each center was randomly split into several partitions, each containing *k* datasets. At each round, one partition was selected for training and the remaining partitions were used for evaluation. The mean classification performance in terms of AUC over all partitions was then calculated. This process was repeated for *n* times. The mean and standard deviation of the mean classification error for the two centers are reported.

TABLE IV. Classification performances in terms of AUC of three feature combination options on the whole prostate in the single-center setting. Patients were pooled from both cohorts. The differences in performances between MRI_prevalence_biopsy and MRI_prevalence and between MRI_prevalence and MRI were significant. Higher AUC shows better performance.

Feature combination options	AUC
MRI	0.690 ± 0.15
MRI_prevalence	0.751 ± 0.11
MRI_prevalence_biopsy	0.775 ± 0.13

In this study, the numbers of included training sets *k* varied between 3 and 23 for Leuven and between 3 and 17 for NKI. The number of repetitions *n* was set to 100. For *k* equal to 23 (Leuven) and 17 (NKI), this was identical to the leave-one-dataset-out cross-validation situation. Thus, the standard deviation of the mean classification error is equal to zero in this case.

3. EXPERIMENTAL RESULTS

3.A. Single-center validation

3.A.1. Performance of the three feature options

Whole prostate tumor classification: Table IV shows the average and standard deviation of AUC values for the three feature options, over the pooled set of patients from both cohorts, when performed on the whole prostate level. For Leuven, the MRI_prevalence_biopsy yielded the highest mean AUC of 0.775, which was 0.024 higher than the MRI_prevalence and the difference was significant (*P*-value is 2×10^{-5}). The MRI_prevalence also performed significantly (*P*-value is 4×10^{-3}) better than the MRI feature option with an AUC improvement of 0.061. Model performances per cohort are shown in Table I of the Appendix S1. In general, the results for the NKI cohort were worse than for the Leuven cohort.

We also evaluated the performance of the three feature options using other two classifiers: random forest and linear support vector machine (SVM). For these classifiers, subsampling the training data, i.e., only 1% of the data was used for training, was performed as these classifiers are computational expensive for large datasets. In our case, each prostate can contain tens of thousands voxels. Results using three different classifiers, i.e., logistic regression, random forest, and linear SVM, on the same subsampled database are shown in Tables III and IV of the Appendix S1. Performances of the models using linear SVM and random forest were worse than those using the logistic regression. However, we do see a consistent pattern in the performance using different feature options, i.e., MRI_prevalence_biopsy always performs the best among the three feature options.

Comparison between automatic and manual delineations: Table V shows the average error rates obtained by automatic delineations (three feature options) and manual delineations (six teams) on 15 patients from the Leuven cohort. Automatic delineations were done by binarizing the corresponding tumor probability maps using the threshold determined from ROC analysis. MRI_prevalence_biopsy feature option again worked the best among the three feature option with the lowest error rate of 0.22. Mean error rate varied among six teams in the range between 0.19 and 0.27. On average over six teams, the error rate returned by a manual delineation was 0.224,

TABLE V. Comparison in terms of error rate between automatic delineations (three feature combination options) and manual delineations (six teams).

Delineation options	Average error rate
MRI	0.29 ± 0.15
MRI_prevalence	0.26 ± 0.11
MRI_prevalence_biopsy	0.22 ± 0.13
Team 1	0.24 ± 0.10
Team 2	0.27 ± 0.11
Team 3	0.24 ± 0.13
Team 4	0.19 ± 0.13
Team 5	0.21 ± 0.10
Team 6	0.20 ± 0.14
Average over six teams	0.224 ± 0.03

which was similar to that returned by MRI_prevalence_biopsy feature option.

Peripheral zone and transition zone tumor classification: Table VI shows the performance of the three feature options in the peripheral zone and transition zone of the prostate. Here also the MRI_prevalence_biopsy performed better than MRI_prevalence; MRI_prevalence performed better than MRI feature option for both PZ and TZ. The true positive rates of the MRI, MRI_prevalence, and MRI_prevalence_biopsy within the transition zone are 0.60, 0.574, and 0.632, respectively. Inclusion of the prevalence map significantly reduced the false-positive rates in the transition zone. The false-positive rates of the MRI, MRI_prevalence, and MRI_prevalence_biopsy within the transition zone are 0.378, 0.288, and 0.295, respectively. As a result, the overall performances in terms of AUC of the MRI_prevalence and the MRI_prevalence_biopsy feature options were better than the MRI feature option in the transition zone.

Figure 4 shows an example of the classification results. In this example, the tumor is located on the left side of the peripheral and transition zone of the prostate. This is consistent with the information provided by the biopsy map [Fig 2(e)]. Figures 3(f)-3(h) demonstrate the tumor probability maps established using the MRI, MRI_prevalence, and MRI_prevalence_biopsy features, respectively. Red contours indicate the estimated tumor areas. Using only MRI features resulted in a false-positive area on the right side of the patient prostate. Incorporating prior knowledge with low prevalence map values in the transition zone and with positive biopsies only on the left side of the prostate markedly reduced tumor probability of voxels on the right side and thus, did not produce the false-positive area [Fig. 4(h)].

TABLE VI. Classification performances in terms of AUC of three feature combination options on the pooled set of patients from both cohorts in the peripheral zone and transition zone.

Feature combination options	Peripheral zone	Transition zone
MRI	0.70 ± 0.16	0.658 ± 0.19
MRI_prevalence	0.745 ± 0.13	0.715 ± 0.16
MRI_prevalence_biopsy	0.777 ± 0.17	0.736 ± 0.17

Standard biopsy scheme versus detailed biopsy scheme: For the patients from the Leuven cohort, we used the additional information in the biopsy reports to construct detailed biopsy maps in which each voxel is assigned a binary value corresponding to the status of the biopsy in that region. The same smoothing algorithm used in the standard biopsy scheme was applied. The AUC of the MRI_prevalence_biopsy in the single-center validation (Table IV) when standard biopsy map was replaced by the detailed biopsy map improved from 0.811 to 0.825. The difference was however not significant (P-value = 0.059). Comparison between the biopsy report and the pathological image of Leuven patients showed that there were three cases in which the biopsy report did not match with the information from the pathology image. For example, a tumor on the right side of the patient on the pathology slice was indicated in the left side in the biopsy report. Excluding these three patients, the mean AUC increased from 0.821 to 0.866 with the detailed biopsy map. This increase is significant with P-value = 0.001.

3.A.2. Learning curves of the three feature combination options

Figures 5 and 6 show the learning curves for Leuven and NKI data using three feature combination options. The figures show clear distinctions in performance among the three options for all different numbers of patients in the training set. In addition, higher performance was obtained when more data were included for training each model. The largest improvement in model performance was observed by increasing small training set sizes, in the range from three to ten.

3.A.3. Feature ranking

Figures 7(a)–8(a) show the Mahalanobis distances calculated during the forward feature selection method with different number of features on the Leuven and NKI cohorts, respectively. The Mahalanobis distance curves on both Leuven and NKI data showed that the first eight features contributed the most to the separation between the normal and tumor classes. Figures 7(b)–8(b) listed the 10th highest ranked features on Leuven and NKI centers, respectively. The lists show that biopsy and prevalence features rank among the top four features in both cohorts.

3.B. Cross-center validation

Table VII showed the mean and standard deviation of AUC values of the three feature options in the cross-center validation. Similar to the single-center setting, the

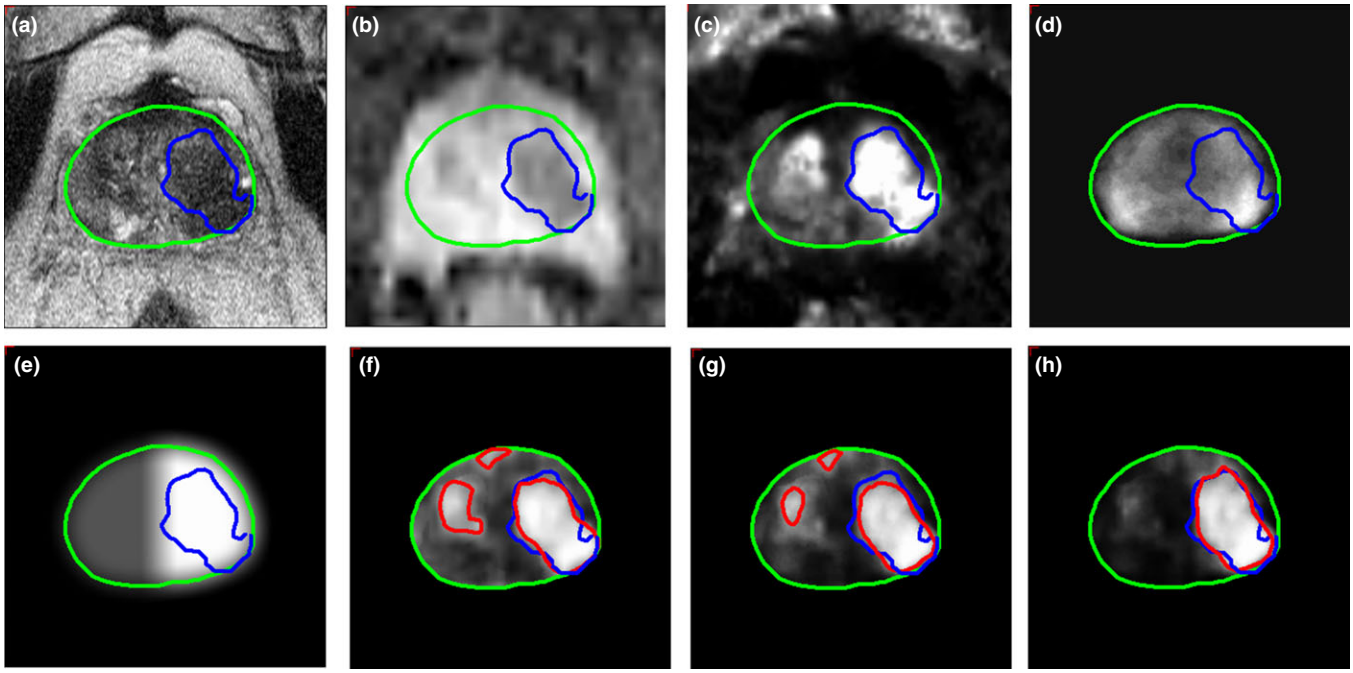


FIG. 4. Example of transverse input images and classification results. (a)–(e) show the input images T2w, ADC, K^{trans}, prevalence, and the smoothed clinical biopsy maps. Contours on these images indicate the delineated prostate and tumor areas based on the H&E-stained slices. (f)–(h) display the tumor probability maps and the estimated tumor regions established using MRI, MRI_prevalence, and MRI_prevalence_biopsy features, respectively. The tumor is on the left side of the patient. [Colour figure can be viewed at wileyonlinelibrary.com]

MRI_prevalence_biopsy yielded the best. However, its performance was only moderately different from the MRI_prevalence feature option (AUC improvement of 0.015, *P*-value = 12×10^{-3}). The MRI_prevalence performed significantly better than MRI feature option with an AUC improvement of 0.067 (*P*-value = 2×10^{-4}). Model performances per cohort in the cross-center setting are shown in Table II of the Appendix S1. In all cases, the difference between single-center setting (Table IV) and cross-center setting (Table VI) using the same feature option, e.g., MRI, was not significant (*P*-value > 0.05).

4. DISCUSSIONS

The experimental results showed that the prevalence feature enhanced the performance of the mp-MRI-based model substantially both for the entire prostate and for the peripheral and transition zones separately. This can be explained by the

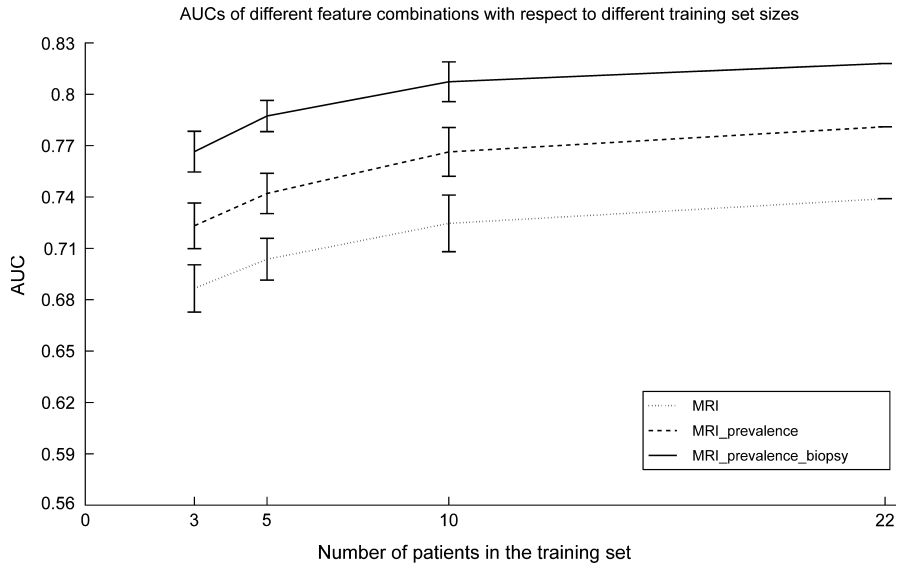


FIG. 5. Learning curve of the three feature combination options for Leuven data center.

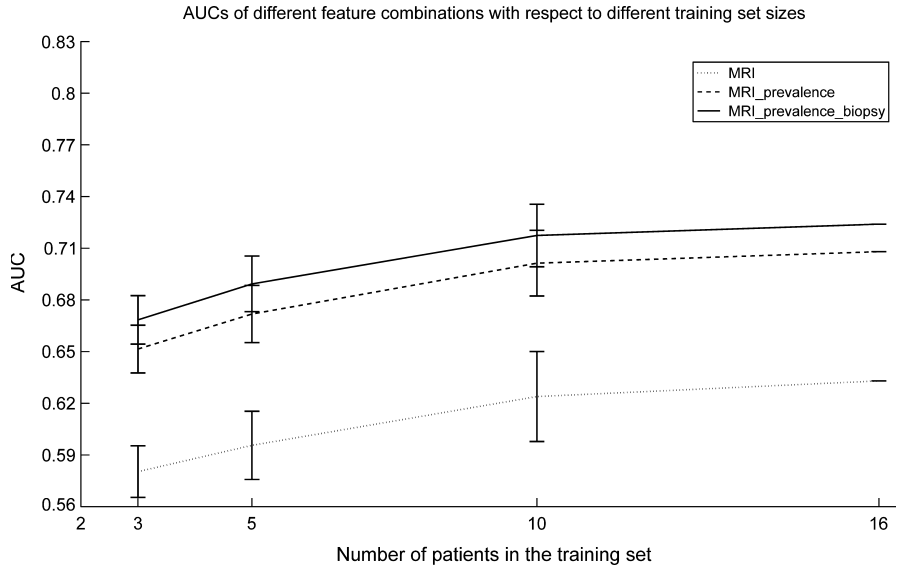
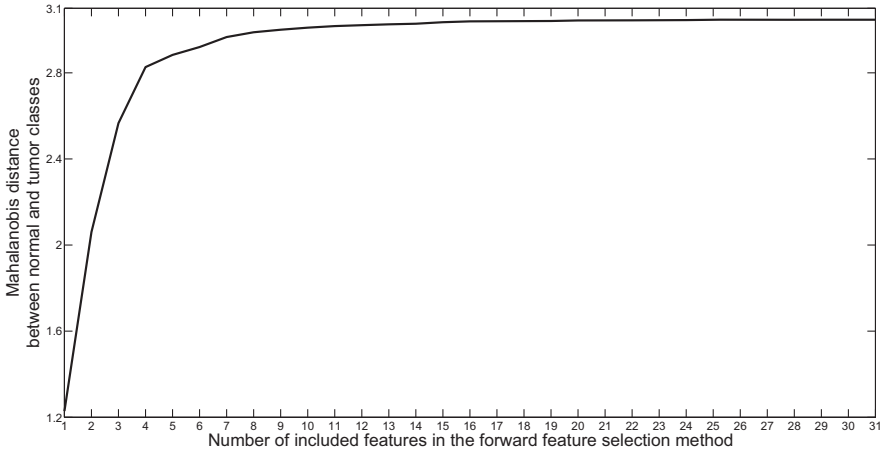


FIG. 6. Learning curve of the three feature combination options for NKI data center.

(a)

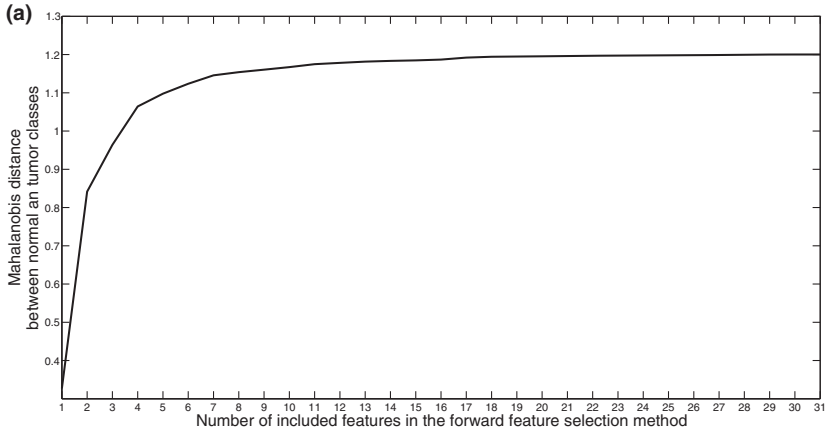


(b)

Rank	Feature Name
1	Biopsy
2	ADC
3	Prevalence
4	K^{trans} blob
5	T2w texture (σ = 6, order = yy)
6	T2w texture (σ = 6, order = x)
7	T2w texture (σ = 3.8, order = xx)
8	T2w texture (σ = 3.8, order = xx)
9	K^{trans}
10	T2w texture (σ = 6, order = y)

FIG. 7. Feature ranking at Leuven. (a) Mahalanobis distances calculated during the forward feature selection method; (b) List of 10 highest ranking features. For T2w texture features, order indicates the order used in calculating the Gaussian derivatives.

(a)



(b)

Rank	Feature Name
1	ADC
2	Prevalence
3	Biopsy
4	K^{trans}
5	T2w texture (σ = 6, order = xx)
6	T2w texture (σ = 1.5, order = x)
7	T2w texture (σ = 1.5, order = xx)
8	T2w texture (σ = 6, order = xy)
9	K^{trans} blob
10	T2w texture (σ = 6, order = x)

FIG. 8. Feature ranking at NKI. (a) Mahalanobis distances calculated during the forward feature selection method; (b) List of 10 highest ranking features.

TABLE VII. Classification performances in terms of AUC of three feature combination options on the whole prostate in the cross-center setting, i.e., model is trained in one cohort and then evaluated in the other cohort. Patients were pooled from both cohorts. The difference in performances between MRI_prevalence and MRI was significant. However, there was only a moderate difference between MRI_prevalence_biopsy and MRI_prevalence. Higher AUC shows better performance.

Feature combination options	AUC
MRI	0.695 ± 0.16
MRI_prevalence	0.762 ± 0.10
MRI_prevalence_biopsy	0.777 ± 0.12

observation that MRI may introduce false positives due to benign confounders especially in the transition zone. Population statistics indicates that tumors occur in the transition zone with low probability (around 25%)[5] and hence, the corresponding prevalence values are small. In our experiments, this indeed reduced the false-positive rate in the entire prostate. While this might decrease the true-positive rate in the transition zone, it still markedly improved the true-positive rate in the entire prostate.

In contrast to the prevalence map, the biopsy map provides patient-specific, albeit rudimentary information about tumor locations. The fact that biopsy and prevalence maps make different types of mistakes from MRI can be appreciated from Figs. 1 and 2(b). In the example in Fig. 2(b), the biopsy scheme indicated tumor presence on the left, middle peripheral zone section of the biopsy scheme. As a result, all voxels in that section have the same likelihood to be tumor. A delineation based only on the clinical biopsy therefore would involve the entire section and would make a wrong prediction for all healthy voxels in that section (voxels outside the white contour). The prevalence feature provides symmetric tumor probability values between the left and right side of the prostate (Fig. 1). Thus, a model using only the prevalence map cannot be used to delineate the actual tumor for a specific patient. In the case of the example in Fig. 2(b), the prevalence map would be reasonable on the left, but wrong on the right peripheral zone.

The MRI-based model provides information on the outline of the tumor, but is more prone to making wrong predictions particularly in areas such as the transition zone, due to benign confounders. As the prior knowledge features make different mistakes than the MRI features alone, including them substantially improved the performance of a mp-MRI-based tumor localization model.

Although a biopsy report is always available for prostate cancer patients before treatment, e.g., with radiotherapy, this information has not been used for computer-aided tumor localization before. This is because the report is mainly used for diagnostic purposes, i.e., to know whether a patient has prostate cancer and not particularly for precise tumor localization. We are the first to combine information from the biopsy report with imaging data, our results showing that the inclusion of the biopsy map further improved the combined MRI and prevalence feature option in both single- and cross-center settings.

The result was further improved by replacing the standard biopsy map by the detailed map in the Leuven cohort. If biopsy schemes are more elaborate and detailed, the information about tumor localization derived from the schemes becomes also more valuable. For tumor localization, this would reduce the added value of MRI. However, as the prostate still would be divided into relatively large sections per biopsy, this would still be insufficient for tumor delineation. This is clearly shown in Fig. 2(b) in which the actual tumor area is much smaller than the biopsy section (left, middle peripheral zone) following the biopsy scheme. As shown in this study, a more detailed representation of the cancer in the prostate is achieved with the combination of both biopsies and imaging.

In comparison with manual tumor delineations, automatic delineation created by the model using all features provided a similar performance when compared to the average performance of six teams consisting of a radiation oncologist and a radiologist. In addition, the high variation in delineations between teams indicates an uncertainty about the edge of the tumor. One could argue that voxels should be treated differently, e.g., higher doses should be given voxels more likely to be tumors, following the idea of dose painting by numbers. Our model fits to that idea as it provides a probabilistic value for tumor presence rather than a binary decision at a voxel level as in the case of manual delineation.

The feature ranking results reconfirmed the importance of the prior knowledge. The biopsy and prevalence features both ranked among the top four features. In addition, the top four features in both centers represent four different modalities: diffusion-weighted MRI (ADC), DCE-MRI (K^{trans} for Leuven and K^{trans} blob for NKI), biopsy, and prevalence. They provided complementary information and thus helped to distinguish between the normal and tumor classes.

The main focus of our paper was to predict the change whether a voxel is a normal or tumor tissues, which is shown in Section 4.A.1 at both prostate and zone levels. Nevertheless, the ability of models to distinguish cancer from benign confounders was indirectly reflected in these experiments by their ability to classify voxels into their correct class. Thorough analysis on how specifically prior knowledge features improve MRI-based model's performance on those factors, which can be found out from pathological data, should be the subject of further investigation in the future.

As noted in the Section 2.D, voxels within $\pm$ 1.25 mm margin of the pathological tumor contours were ignored to reduce the influence of registration errors. Nevertheless, we found the same improvement in model performance by adding prior information when no margin was used. However, classification results were consistently lower for all feature options.

Overall, the performance is better for Leuven cohort than NKI cohort. This may be explained by the higher disease stages at Leuven (Table I) compared to NKI. The conspicuity of high-stage tumors on mp-MRI tends to be higher compared to lower stage disease.[22]

This study used a simple deformable registration between patient's T2w scan and the population-based tumor atlas, which relies heavily on the prostate contour to create the patient-specific prevalence map. A more sophisticated registration method between T2w and the tumor atlas, which takes MRI intensity as well as anatomical information into account in the registration process,[11] could be used as well. This might further boost the performance of our combined prior and MRI model for prostate tumor localization in individual patients.

We evaluated the robustness of our model in a multicenter setting. There was a significant difference in imaging protocols between the two centers, e.g., they used different scanners with different field strength. In all cases, the difference in performance between single-center setting (Table IV) and cross-center setting (Table VII) using the same feature option was not significant. Applying normalization techniques on the raw T2w and K^{trans} maps and then extracting invariant features, such as Gaussian derivatives and blobness, helped to reduce but does not account for all these differences and acquisition-related issues. However, for the purpose of tissue classification, variation in imaging features is acceptable as long as the discriminant information between classes derived from these features is maintained. This is corroborated by our experimental result that knowledge learned from one center can be transferred to the other center even though the imaging protocols are vastly different from each other. This result is consistent with those shown in transfer learning techniques, concerning with the issue of how to learn when training and testing data follow different distributions. Transfer learning has been recently applied to medical imaging problems and showed promising results on, for example, brain segmentation tasks where images were obtained with different scanners and imaging protocols.[23]

Nevertheless, we note that the experiments on cross-center setting were performed on just a small set of patients. Experiments on a larger population with larger variety in imaging protocols are needed to verify our results.

Variations in delineation of the tumor have an impact on the treatment plans. If we consider integrated boosting, such as done in the FLAME trial,[24] the dose outside the delineated tumor volume drops from the boost dose to the standard dose. In the FLAME trial,[24] this was a drop of 18 Gy, from 95 to 77 Gy. The steepness of the dose gradient depends on many details of the specific treatment plan, but if we assume a penumbra width (20–80% of dose) of 6 mm, a contour variation of 1 mm would correspond to a dose difference of 10% of 18 Gy. A treatment planning study, also taking individual factors such as the proximity of the delineated tumor to organs at risk into account, falls outside the scope of this study.

5. CONCLUSIONS

We presented a method for tumor localization in prostate based on mp-MRI in combination with prior knowledge. Experimental results validated on patients from two cohorts showed that including prior knowledge features substantially improved the performance of a mp-MRI-based tumor localization model in both single-center and cross-center settings.

Other patient-specific information, such as tumor extension status from radiology reports and the Gleason score of biopsies from biopsy reports, may also be investigated in the future to improve the performance of tumor localization and to add mapping of tumor aggressiveness estimations, respectively.

ACKNOWLEDGMENT

The authors thank Simon van Kranen for his support in registering between T2w and population-based tumor probability maps. This research was supported by the Dutch Cancer Society (NKI 2013-5937) and the Dr. Therapat project (Digital Radiation Therapy Patient).

CONFLICTS OF INTEREST

There is no conflict of interest.

[a)]Author to whom correspondence should be addressed. Electronic mail: u.vd.heide@nki.nl; Telephone: +31 (0)20 5122350.

REFERENCES

1. Dinh CV, Steenbergen P, Ghobadi G, et al. Magnetic resonance imaging for prostate cancer radiotherapy. *Physica Med*. 2016;32:446–451.
2. Barentsz JO, Richenberg J, Clements R, et al. ESUR prostate MR guidelines 2012. *Eur Radiol*. 2012;22:746–757.
3. Steenbergen P, Haustermans K, Lerut E, et al. Prostate tumor delineation using multiparametric magnetic resonance imaging: inter-observer variability and pathology validation. *J Radiother Oncol*. 2015;115:186–190.
4. Groenendaal G, Borren A, Moman MR, et al. Pathologic validation of a model based on diffusion-weighted imaging and dynamic contrast-enhanced magnetic resonance imaging for tumor delineation in the prostate peripheral zone. *Int J Radiat Oncol Biol Phys*. 2012;82:537–544.
5. Viswanath SE, Bloch NB, Chappelow JC, et al. Central gland and peripheral zone prostate tumors have significantly different quantitative imaging signatures on 3 tesla endorectal, in vivo T2-weighted MR imagery. *J Magn Reson Imaging*. 2012;36:213–224.
6. Vos PC, Barentsz JO, Karssemeijer N, Huisman HJ. Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis. *Phys Med Biol*. 2012;57:1527.
7. Shah V, Turkbey B, Mani H, et al. Decision support system for localizing prostate cancer based on multiparametric magnetic resonance imaging. *Med Phys*. 2012;39:4093–4103.
8. Litjens GJS, Elliott R, Shih N, et al. Distinguishing prostate cancer from benign confounders via a cascaded classifier on multi-parametric MRI. In: *SPIE Medical Imaging*, 903512–903512; 2014.
9. Hoeks CM, Barentsz JO, Hambrock T, et al. Prostate cancer: multiparametric MR imaging for detection, localization, and staging. *Radiology*. 2011;261:46–66.
10. Ou Y, Shen D, Zeng J, Sun L, Moul J, Davatzikos C. Sampling the spatial patterns of cancer: optimized biopsy procedures for estimating prostate cancer volume and Gleason Score. *Med Image Anal*. 2009;13:609–620.
11. Rusu M, Bloch BN, Jaffe CC, et al. Prostatome: a combined anatomical and disease based MRI atlas of the prostate. *Med Phys*. 2014;41:072301.
12. Isebaert S, Van den Bergh L, Haustermans K, et al. Multiparametric MRI for prostate cancer localization in correlation to whole-mount histopathology. *J Magn Reson Imaging*. 2013;37:1392–1401.
13. Vilgrain V, Daire JL, Sinkus R, Van Beers BE. Diffusion-weighted MR imaging of the liver. *J Radiol*. 2010;91:381–390.
14. Tofts PS, Brix G, Buckley DL, et al. Estimating kinetic parameters from dynamic contrast-enhanced T 1-weighted MRI of a diffusable tracer: standardized quantities and symbols. *J Magn Reson Imaging*. 1999;10:223–232.
15. Fennessy FM, Fedorov A, Gupta SN, Schmidt EJ, Tempany CM, Mulkern RV. Practical considerations in T1 mapping of prostate for dynamic contrast enhancement pharmacokinetic analyses. *Magn Reson Imaging*. 2012;30:1224–1233.
16. Lindeberg T. Feature detection with automatic scale selection. *Int J Comput Vision*. 1998;30:79–116.
17. Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Non-rigid registration using free-form deformations: application to breast MR images. *IEEE Trans Med Imaging*. 1999;18:712–721.
18. Presti JC Jr. Prostate biopsy: current status and limitations. *Reviews in Urology*. 2007;9:93.
19. Myronenko A, Song X. Point set registration: coherent point drift. *IEEE Trans Pattern Anal Mach Intell*. 2010;32:2262–2275.
20. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell*. 1997;19:153–158.
21. De Maesschalck R, Jouan-Rimbaud D, Massart DL. The mahalanobis distance. *Chemometr Intell Lab Syst*. 2000;50:1–18.
22. Turkbey B, Mani H, Shah V, et al. Multiparametric 3T prostate magnetic resonance imaging to detect cancer: histopathological correlation using prostatectomy specimens processed in customized magnetic resonance imaging based molds. *J Urol*. 2011;186:1818–1824.
23. van Opbroek A, Ikram MA, Vernooij MW, De Bruijne M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans Med Imaging*. 2015;34:1018–1030.
24. Lips IM, van der Heide UA, Haustermans K, et al. Single blind randomized phase III trial to investigate the benefit of a focal lesion ablative microboost in prostate cancer (FLAME-trial): study protocol for a randomized controlled trial. *Trials*. 2011;12:255.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Appendix S1: Performance of three feature options evaluated per institute and using different classifiers.