Title: Prediction of Clinically Significant Prostate Cancer in Patients Undergoing Fusion Biopsy Using Machine Learning: Lessons Learned from Dataset Assembly at UCSF

Authors: Kirti Magudia, MD PhD; Antonio Westphalen, MD PhD; Marc Kohli, MD; Valentina Pedoia, PhD; Peder Larson, PhD

Introduction: Prostate cancer (PCa) is a well-recognized public health problem with the highest incidence of any invasive cancer with men. Interestingly, PCa has a 100% survival if localized, but only 30% survival if metastatic. Management is dictated by the pathologically derived Gleason score, which is the dominant predictor of clinically significant PCa (CS-PCa). Prostate MR in combination with standardized reporting with PI-RADS has shown some ability to predict CS-PCa and to identify sites for subsequent TRUS-MRI fusion biopsy. However, improved identification of CS-PCa by prostate MRI is needed. We propose using the most specific pathology results correlated to prostate MRI—subsequent fusion biopsy—as the foundation of a machine learning model to predict CS-PCa. Given that dataset assembly for machine learning projects requires significant time and effort, we will share the lessons we learned to benefit of other researchers preparing for machine learning projects.

Methods: All men who had TRUS-MRI fusion biopsy of the prostate at UCSF before April 2019 were identified by querying the electronic medical record (EMR). Demographic data, pathology reports and PSA data were also obtained for the included patients. Prostate MRI exams performed for these patients in the six months preceding TRUS-MRI fusion biopsy were retrieved from PACS. Regular expression analysis was used to identify axial T2, high b-value diffusion and ADC images based on analysis of DICOM tags. These identified DICOM files were transferred to the annotation platform MD.ai using the pynetdicom python package.

Results: 1. *Identifying fusion biopsy cohort:* 1597 patients were identified with 1860 unique biopsy procedures. 2. *Identifying prostate MR exams:* 12635 accession numbers were identified for prostate MR exams performed at UCSF. After exclusion of 3D and spectroscopy exam codes, 11128 scans from 8291 patients remained. 3. *Identifying fusion and MR exam pairs:* 1442 patients had both undergone fusion biopsy and prostate MRI. Of these, 1318 represented men with unique fusion biopsy procedures with available pathology reports preceded by a prostate MR within 6 months. 4. *Processing pathology reports:* Pathology reports were processed using regular expression analysis to determine Gleason score for each MR target identified for biopsy, as well as the maximum Gleason score anywhere in the prostate. The final sample consists of 1083 men with valid pathology data for MR targets. 5. *Series selection and transferring images to MD.ai:* Transfer of selected axial T2, high b-value diffusion and ADC series are successfully underway to MD.ai, a collaborative, web-based annotation platform.

Conclusions: In pursuit of better identifying CS-PCa by prostate MRI, we identified 1083 patients that underwent prostate MRI followed by fusion biopsy at UCSF with valid pathology results for MRI-identified biopsy sites. The steps required for this dataset assembly are explicitly detailed to help other researchers.

Highlights: In pursuit of better identifying CS-PCa by prostate MRI, we identified 1083 patients that underwent prostate MRI followed by fusion biopsy at UCSF with valid pathology results for MRI-identified biopsy sites. Details will be provided regarding cohort assembly, pathology report processing, series selection and transfer of images to MD.ai.