Using Artificial Intelligence to Reduce Inter-Observer Variability in Gleason Scoring/Grading of Prostate Cancer

Wei Huang, MD[1,3], Samuel Hubbard[1], Parag Jain[3] and Ramandeep Randhawa[2,3]

[1]Department of Pathology and Laboratory Medicine, University of Wisconsin – Madison, [2]University of Southern California

[3]PathomIQ Inc., California

Background

Gleason scoring is the gold standard for assessing aggressiveness of prostate cancer. It is well known that Gleason scoring suffers from high inter- and intra-observer variability with over 40% discordance between general and sub-specialty pathologists. A universal and standardized Gleason Scoring platform trained by GU pathologists is needed for better patient care and clinical research.

Materials and Methods

PathomIQ Inc's AI-based Gleason scoring/grading software was used for this study.

- Software trained using annotations from GU pathologists to identify various morphologies, including cancer of all Gleason patterns (GP3, GP4 and GP5), HGPIN, perineural invasion (PNI), vessels and lymphocytes, etc.
- The algorithm is deep learning based and comprises multiple Deep Convolutional Neural Networks that are a combination of classification and segmentation networks.
- The software automatically annotates entire whole slide images (WSI) into the various cancer and benign pattern groups, and further provides summary statistics of Gleason score, quantification of cancer area, and the percentage of each cancer pattern. (Figure 1)
- The software also allows pathologists to modify the annotations upon review.
- The software was separately validated on a set of 200 separate biopsy slides with various cancer grade groups to establish concordance with GU pathologists experts in prostate cancer and demonstrated 95% agreement (κ = 0.94); Huang et al. 2019.

Three highly-experienced GU pathologists participated in this study. They independently scored each WSI twice - first, manually, and then, taking assistance from the software before finalizing their independent score. The three of them then reviewed their results and agreed to a final score for each biopsy.

Summary

Deep learning enabled Cancer Grading/Scoring software has tremendous potential in improving inter-observer agreement, and especially in identification of high grade cancer.

References

1. Dugan, J.A., et al., *The definition and preoperative prediction of clinically insignificant prostate cancer.* JAMA, 1996. **275**(4): p. 288-94.
2. Epstein, J.I., et al., *Contemporary Gleason Grading of Prostatic Carcinoma: An Update With Discussion on Practical Issues to Implement the 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma.* Am J Surg Pathol, 2017. **41**(4): p. e1-e7.
3. de Souza, M.F., et al., *The Gleason pattern 4 in radical prostatectomy specimens in current practice - Quantification, morphology and concordance with biopsy.* Ann Diagn Pathol, 2018. **34**: p. 13-17.
4. Allsbrook, W.C., Jr., et al., *Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists.* Hum Pathol, 2001. **32**(1): p. 74-80.
5. De la Taille, A., et al., *Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays.* Hum Pathol, 2003. **34**(5): p. 444-9.
6. Al-Hussain, T.O., M.S. Nagar, and J.I. Epstein, *Gleason pattern 5 is frequently underdiagnosed on prostate needle-core biopsy.* Urology, 2012. **79**(1): p. 178-81.
7. Huang, W., et al., Prostate Cancer Diagnosis and quantification using AI-enabled Software (SW). USCAP 2019.

Results

Figure 2 compares the final scores/Grade Groups considered ground truth (i.e., consensus of the three pathologists when assisted by the software) with the manual scores/Grade Groups (without using the software). We observe that the concordance between each individual manual score with the ground truth is 69.6% (κ = 0.59) and majority of the discordance arises from under-scoring mostly for Grade Group 2 and for GP5. Additionally, we also noted that the pure AI-based Gleason scoring/grading has 96.5% concordance with the ground truth (κ = 0.96). The manual scoring/grading agreement was improved from ~60% (κ = 0.43) concordance of each independent pathologists, to ~90% (κ = 0.87 - 0.89) concordance, when scored with software assistance.

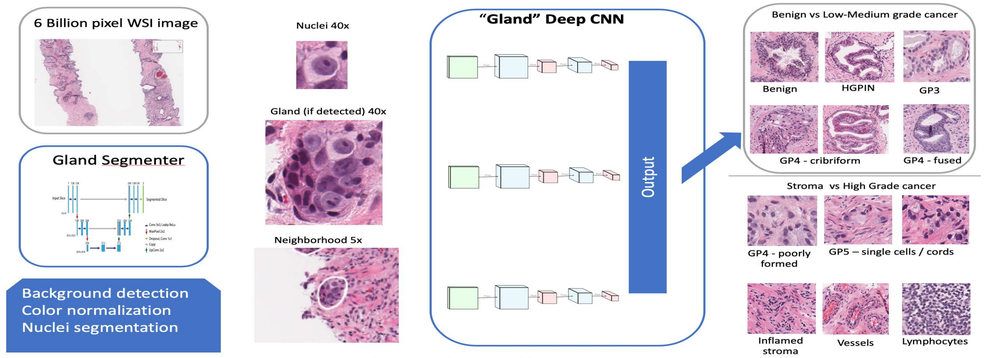

Figure 1. PathomIQ software architecture

SW Assisted Consensus Grade

Manual Consensus grade	Group 1	Group 2	Group 3	Group 4	Group 5
Group 1	35	26	1	0	0
Group 2	2	39	3	3	2
Group 3	1	1	8	0	4
Group 4	0	0	4	8	2
Group 5	0	0	0	0	23
Total	38	66	16	11	31

112 (69.8%) Consistent

8 (4.9%) Over-graded Manually

41 (25.3%) Under-graded Manually

Figure 2. Comparing software assisted final Grade Group vs each individual manual grade group to highlight the advantage of software assisted scoring

Disclosures

This research is sponsored by PathomIQ Inc.