Clinically applicable Gleason grading (GD) system for prostate cancer based on deep learning

Yun Niu1, Can-Cheng Liu2, Bing-Lin Zhang1, Zhi-Gang Song3, Huang Chen1, Ping-Ping Liu1, Jing-Si Chen2, Shu-Hao Wang2,3, Huai-Yin Shi4, Ding-Rong Zhong1

1Department of Pathology, China-Japan Friendship Hospital, Beijing 100029, China; 2R&D, Thorough Images, Beijing 100102, China; 3Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China; 4Department of Pathology, Chinese People’s Liberation Army General Hospital, Beijing 100085, China.

To the Editor: Prostate cancer is one of the most common malignant tumors of the male genital system, with approximately 1.1 million new cases in 2012.[1] The accurate diagnosis of prostate cancer leads to a better chance of successful treatment when it is still confined to the prostate gland. The Gleason grading (GD) system was first established by Donald Gleason during 1966 to 1974.[2,3] The Gleason pattern ranges from 1 to 5. A higher score corresponds to poorer differentiation, which indicates a worse prognosis and higher metastasis possibility. The total score is calculated with the first half of the dominant Gleason pattern and the second half based on the non-dominant one.

Although the Gleason score (GS) remains one of the most powerful predictors of oncological outcomes for men, there are differences in interpretation among pathologists, and the assessment of the proportion of each grade in the specimen is subjective, leading to the poor repeatability of diagnosis among pathologists and even to misdiagnosis for small lesions.

To reveal the objectiveness of artificial intelligence, we proposed a GD system using deep learning to assist the prostate cancer histopathological diagnosis. All prostate biopsy slides used in this study were collected from the China-Japan Friendship Hospital. This study was approved by the Institutional Review Board (No. 2018-106-K75). A total of 123 hematoxylin-eosin (HE)-stained slides of the prostate cancer histopathological diagnosis, respectively. The labels included Gleason patterns 3–5, high-grade prostate intraepithelial neoplasia (HPIN), inflammation, and normal tissue. The slides were first assigned to the first pathologist and then reviewed by the senior pathologist. During the labeling process, the pathologists used the corresponding immunohistochemistry (IHC) slides as references.

Before model training, we divided the tissue area into 320 × 320-pixel patches with 200× field-of-views (0.5 μm/pixel). We obtained 152,139 training patches, including Gleason patterns 3 (25,316), 4 (31,176), and 5 (25,344), HPIN (3252), inflammation (2744), and normal tissue (64,307).

As shown in Figure 1, we used the DeepLab v3 image segmentation model with ResNet-50 to establish the GD system.[4] During the model training process, the parameters of the gastric cancer detection model[5] were used as the initial values, and the model parameters were fine-tuned using the prostate training data by transfer learning. The model training was performed with TensorFlow on 8 NVIDIA GTX1080Ti GPUs. The optimizer was ADAM, used to assist the labeling process, including p63, 34βE12, and p504S. We have also collected 137 HE-stained slides for model testing. All slides were digitalized using KFPRO-005 (KFBIO, Ningbo, Zhejiang, China) scanner with 400× magnification. The detailed data distribution is provided in Supplementary Table 1, http://links.lww.com/CM9/A382.

All whole-slide imagings (WSIs) were reviewed by two licensed pathologists via our in-house labeling system. The pathologists had 11 and 30 years of experience in prostate pathological diagnosis, respectively. The labels included Gleason patterns 3–5, high-grade prostate intraepithelial neoplasia (HPIN), inflammation, and normal tissue. The slides were first assigned to the first pathologist and then reviewed by the senior pathologist. During the labeling process, the pathologists used the corresponding IHC slides as references.
with the learning rate, batch size, training iteration fixed at 0.0001, 256, 28,000, respectively. We also applied listopathological-oriented data augmentation. The slide-level prediction was defined as the average of the top 100 probabilities of the pixel-level predictions.

The multiple classification could be evaluated in a binary manner. We defined “malignant” as Gleason patterns 3–5 and “benign” as HPIN, inflammation, and normal tissue. The deep learning model achieved a sensitivity, specificity, and accuracy of 100.00%, 87.04%, and 94.89%, respectively [Supplementary Table 2, http://links.lww.com/CM9/A382].

We illustrated several model predictions in Supplementary Figure 1, http://links.lww.com/CM9/A382. The variably sized individual glands that were well-formed with discrete units were marked as Gleason pattern 3 by the model. All cribriform patterns were classified as Gleason pattern 4, including poorly formed or fused glands. The sheets of the tumor, individual cells, cords, line arrays, and solid nests of cells were classified as Gleason pattern 5. In addition, cribriform glands with comedonecrosis were detected as Gleason pattern 5. The inflammatory regions and HPINs were also accurately identified.

The confusion matrix of the model prediction based solely on the HE-stained slides against the senior pathologist’s diagnosis is shown in Supplementary Figure 2A, http://links.lww.com/CM9/A382. Most prediction results were consistent with those of the senior pathologist (100/137). In 22 cases, the results predicted by the model were very close to the senior pathologist’s diagnosis, with a difference of only one score. We also invited an attending pathologist to diagnose all the cases in the test dataset with the reference to both the HE- and IHC-stained slides. The confusion matrix is shown in Supplementary Figure 2B, http://links.lww.com/CM9/A382, 107 diagnoses were consistent with those of the senior pathologist. There was no obvious advantage over the model (P = 0.325).

The model performed better than the attending pathologist in several cases [Supplementary Figure 3, http://links.lww.com/CM9/A382]. The model could identify cancer with small foci and local Gleason pattern 4 lesions with a Gleason pattern 3 background. Moreover, 20 samples with GS ≥ 8 were correctly predicted by the model, while only 13 were correctly predicted by the attending pathologist. For cases with HPIN, the model sensitivity surpassed that of the attending pathologist, with P = 0.006. HPIN with multiple foci in punctured tissue was a high-risk factor for the subsequent detection of prostate cancer. The system was of great significance in detecting HPIN and prompting patient follow-ups.

Furthermore, we performed a trial running on the history prostate samples collected from May 2013 to July 2015 at the China-Japan Friendship Hospital. The model achieved a sensitivity of 100.0% and a specificity of 91.4% for malignant tumor detection. We also collected 166 slides from the Chinese PLA General Hospital. The sensitivity and specificity reached 97.0% and 77.4%, respectively. The confusion matrices were provided in Supplementary Figure 4, http://links.lww.com/CM9/A382.

In summary, the deep learning-based GD system could intuitively identify lesions and provide a reference for...
pathologists. Moreover, it could output the GS objectively, saving much time for pathologists. Despite all this, there were still some defects in the model, including false-positive cases and inaccurate GD [Supplementary Figure 5, http://links.lww.com/CM9/A382]. More training samples were required to optimize the model and improve its specificity continuously.

Funding
This work was supported by grants from the National Natural Science Foundation of China (No. 61532001) and the Tsinghua Initiative Research Program (No. 20151080475).

Conflicts of interest
Shu-Hao Wang is the co-founder and chief technology officer (CTO) of Thorough Images. Can-Cheng Liu and Jing-Si Chen are algorithm researchers of Thorough Images. All remaining authors have declared no conflicts of interest.

References
1. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO classification of tumours of the urinary system and male genital organs-part A: renal, penile, and testicular tumours. Eur Urol 2016;70:53–103. doi: 10.1016/j.eururo.2016.02.029.
2. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. Modern Pathol 2004;17:292–306. doi: 10.1038/modpathol.3800054.
3. Gleason DF, Mellinger GT. Veterans Administration Cooperative Urological Research Group. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol 2017;197:S134–S139. doi: 10.1016/j.juro.2016.10.099.
4. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder With Atrous Separable Convolution for Semantic Image Segmentation Computer Vision- ECCV 2018. Cham: Springer; 2018. 833–851.
5. Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020;11:4294. doi: 10.1038/s41467-020-18147-8.

How to cite this article: Niu Y, Liu CC, Zhang BL, Song ZG, Chen H, Liu PP, Chen JS, Wang SH, Shi HY, Zhong DR. Clinically applicable Gleason grading system for prostate cancer based on deep learning. Chin Med J 2021;134:859–861. doi: 10.1097/CME9.000000000001220