Gleason grading of prostate cancer using artificial intelligence: lessons learned from the PANDA challenge

Kimmo Kartasalo[1] KIMMO.KARTASALO@KI.SE
Peter Ström[1] PETERSTROMPRIVAT@GMAIL.COM
Martin Eklund[1] MARTIN.EKLUND@KI.SE
[1] *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

Wouter Bulten[2] WOUTER.BULTEN@RADBOUDUMC.NL
Hans Pinckaers[2] HANS.PINCKAERS@RADBOUDUMC.NL
Geert Litjens[2] GEERT.LITJENS@RADBOUDUMC.NL
[2] *Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands*

Po-Hsuan Cameron Chen[3] CAMERONCHEN@GOOGLE.COM
Kunal Nagpal[3] KUNALN@GOOGLE.COM
[3] *Google Health, Palo Alto, CA, USA*

Pekka Ruusuvuori[4,5] PEKKA.RUUSUVUORI@UTU.FI
[4] *Institute of Biomedicine, Cancer Research Unit and FICAN West Cancer Centre, University of Turku and Turku University Hospital, Turku, Finland*
[5] *Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland*

PANDA Challenge Consortium

Editors: Under Review for MIDL 2022

Abstract

Assessing prostate biopsies is crucial for the clinical management of patients with suspected prostate cancer, but is associated with complications such as inter-observer variability. The PANDA challenge aimed at mitigating these issues through development and rigorous validation of image analysis algorithms for the task. In this short paper, we summarize the key insights gained from PANDA from the viewpoints of algorithm development and challenge organisation.

Keywords: Computational pathology, prostate cancer, Gleason grading, challenge.

1. Introduction

Assessing prostate biopsies according to the Gleason grading system suffers from considerable variability between different pathologists, which in turn can lead to under- and overtreatment of patients (Egevad et al., 2013). Artificial intelligence (AI) algorithms applied to digitally scanned biopsy specimens have shown promise for mitigating these issues by aiding pathologists, but have been lacking validation across diverse medical settings (Kartasalo et al., 2021). To accelerate the development of the next generation of AI algorithms for this task, we organised the Prostate cancer grade assessment challenge (PANDA), hosted on the Kaggle platform in April-July 2020, and rigorously validated the top-performing algorithms across international patient cohorts (Fig. 1) (Bulten et al., 2022).

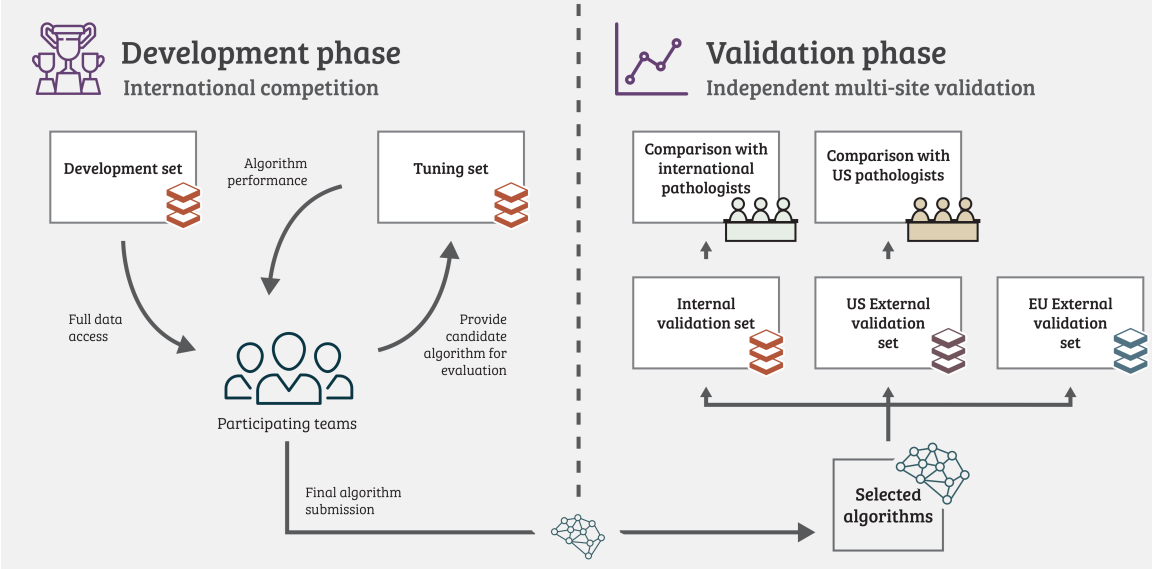

Figure 1: PANDA participants could train algorithms on a development set and evaluate them on a tuning set of biopsies. Top-performing algorithms were subjected to independent multi-site validation. Figure modified from (Bulten et al., 2022) under CC BY 4.0 (http://creativecommons.org/licenses/by/4.0/).

2. Results and Discussion

PANDA involved 12,625 whole slide images (WSI) of prostate biopsies from 6 different sites and attracted 1010 teams from 65 countries, making it the largest competition in histopathology by the time. The challenge setup proved efficient, resulting in the first team achieving pathologist-level grading performance in only 10 days. Importantly, the algorithms of the 15 teams who joined the validation phase of the study generalized successfully to fully external European (n=330) and US (n=741) cohorts, demonstrating grading performance similar to pathologists. We summarize the key takeaways from PANDA as follows:

- Algorithms should ideally be validated blinded to algorithm developers to avoid positive bias. In PANDA, Jupyter notebooks and Docker images implementing the teams' algorithms were provided to the organisers, allowing independent replication.
- Algorithms should be validated on external data representing different patient cohorts, laboratories and WSI scanners. In contrast to internal validation, there were marked differences even between the top-performing PANDA algorithms on external data.
- In spite of inter-observer and geographical grading variation, reference standards by panels of pathologists can be highly consistent even in an intercontinental setting.
- Weakly supervised training on WSI-level labels is sufficient for obtaining models with pathologist-level performance in Gleason grading. To this end, top-ranking teams

used multiple instance learning or a novel approach termed "concatenate tile pooling", which involves passing patches from a WSI through convolutional layers, followed by concatenation, pooling and feeding of the resulting features to a fully connected head.

- Various approaches for controlling label noise by excluding or relabeling samples automatically proved highly successful. Due to the subjective nature of many pathological assessments, label cleaning could be relevant in other tasks besides Gleason grading.
- The PANDA development set of 10,616 WSIs remains one of the largest openly available digital pathology datasets (https://panda.grand-challenge.org/) and is a potentially useful resource for diverse computational pathology applications.

Acknowledgments

PANDA Challenge Consortium: Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Rober Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Américo Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo G. O. Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, Ming Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manuel Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia-Lun Hsieh, igor Zubarev, Habib S. T. Bukhar, Wenyuan Li, Jiayun Li, William Speier, Corey Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong-Seok Lee and Jeonghyuk Park.

References

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, pages 1–10, 2022.

Lars Egevad, Amar S Ahmad, Ferran Algaba, Daniel M Berney, Liliane Boccon-Gibod, Eva Compérat, Andrew J Evans, David Griffiths, Rainer Grobholz, Glen Kristiansen, et al. Standardization of gleason grading among 337 european pathologists. *Histopathology*, 62 (2):247–256, 2013.

Kimmo Kartasalo, Wouter Bulten, Brett Delahunt, Po-Hsuan Cameron Chen, Hans Pinckaers, Henrik Olsson, Xiaoyi Ji, Nita Mulliqi, Hemamali Samaratunga, Toyonori Tsuzuki, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer in biopsies—current status and next steps. *European Urology Focus*, 7(4):687–691, 2021.