Editorial

Value of Public Challenges for the Development of Pathology Deep Learning Algorithms

Douglas Joseph Hartman[1], Jeroen A. W. M. Van Der Laak[2,3], Metin N. Gurcan[4], Liron Pantanowitz[1]

[1]Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA, [2]Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands, [3]Centre for Medical Image Science and Visualisation, Linköping, Sweden, [4]Center for Biomedical Informatics, Wake Forest School of Medicine, Winston Salem, NC, USA

Abstract

The introduction of digital pathology is changing the practice of diagnostic anatomic pathology. Digital pathology offers numerous advantages over using a physical slide on a physical microscope, including more discriminative tools to render a more precise diagnostic report. The development of these tools is being facilitated by public challenges related to specific diagnostic tasks within anatomic pathology. To date, 24 public challenges related to pathology tasks have been published. This article discusses these public challenges and briefly reviews the underlying characteristics of public challenges and why they are helpful to the development of digital tools.

Keywords: Algorithm development, artificial intelligence, digital pathology algorithms, public challenges

INTRODUCTION

There is great excitement about the potential for artificial intelligence (AI) to favorably alter the clinical practice for diagnostic pathologists.[1] One mechanism that has facilitated the development of AI algorithms has been through public challenges for specific tasks.[2] A public challenge is a free/open to the public image task, for example, mitotic figure counting, gland segmentation, and detection of metastatic tumor foci in lymph nodes. Many of the available public challenges in medical imaging are hosted by one website – grand-challenge.org.[3] A dataset is generally provided as part of the challenge – sometimes as a single set and sometimes divided into training and testing sets. The training set is used to build the algorithm, whereas the testing set is used to evaluate the performance of the algorithm in an independent set of cases. Annotated datasets are mostly provided depending on the specific image task being examined (e.g., segmentation, drawing the boundaries of structures of interest in an image, or slide level, i.e., cancer/no cancer). As part of the data download process, users must register. The users are given a timeframe to build the algorithm using the training set and all the other available data and evaluate the developed algorithm on a test set using some evaluation criteria. For some challenges, the

results and the methods to achieve those results are presented in conjunction with a conference. The rules for the challenges are published and these typically list information about how the results will be evaluated as well as the size of the dataset. The results from different algorithms are usually posted on a public leaderboard. Some competitions are associated with awards for the top-performing algorithms. These public challenges provide a large dataset with annotations that is necessary to develop an AI algorithm.[4] The public datasets allow for a common mechanism to compare the algorithms from different developers (including both academia and industry). Public challenges are also a method to advance computational pathology by encouraging competition and enabling a direct comparison between different algorithms. They also foster the AI startup field by reducing the burden of obtaining a large dataset and enabling different groups to work on the same

Address for correspondence: Dr. Douglas Joseph Hartman,
Department of Pathology, University of Pittsburgh Medical Center,
200 Lothrop Street, A-610, Pittsburgh 15213, PA, USA.
E-mail: hartmandj@upmc.edu

Access this article online

Quick Response Code:

Website:
www.jpathinformatics.org

DOI:
10.4103/jpi.jpi_64_19

problem and learn from each other. Since these challenges provide a basis for AI development, it is critical to understand the underlying infrastructure used to build AI. Awareness of these public challenges will increase the knowledge of the field of AI development and could also be helpful in the regulatory field. Wider appreciation of how AI is developed and how it performs for given tasks can increase the acceptance of AI within the broader medical community. Although it is still evolving, the Food and Drug Administration has expressed interest in the regulation of AI as a medical device while also approving the first AI algorithm – IDx-DR,[5] a screening algorithm for the evaluation of diabetic retinopathy.

TECHNICAL BACKGROUND

These public challenges essentially offer raw data for many groups worldwide to attempt to solve some of the challenging pathology problems. The algorithms produced by the participants using the challenge data set are evaluated based on the results of their submitted outputs. In many cases, the data license prohibits the use of the dataset for reasons other than challenge participation. An exception is the Cancer Metastases in Lymph Nodes (CAMELYON) data sets, which are shared under a CC0 license allowing unlimited use of the data.[6] Although the Cancer Genome Atlas also contains whole-slide images (WSIs), the images are not annotated, limiting their usefulness for task-related challenges.[7] The challenge datasets within each challenge are variable, as are the requested tasks to be solved by AI. The challenges generally have a period where training data are provided, and then, a testing dataset is provided with the submission of the results to the host organization so that the results can be shared with the public.

PROCEDURE AND DATASETS

Using the website grand-challenge.org, the number of challenges related to anatomic pathology was collated.[3] This website is authored by the Diagnostic Image Analysis Group (led by Bram van Ginneken of the Radboud University Medical Center in Nijmegen, the Netherlands). It includes a listing of various public challenges that have been published since 2007. As of this writing, the grand challenge consisted of 191 challenges. The challenges have been hosted by various organizations or groups, but Medical Imaging Computing and Computer-Assisted Intervention, International Society for Biomedical Imaging (ISBI), and International Society for Optics and Photonics (SPIE) Medical Imaging have been some of the more prolific supporters of these challenges. For example, we will briefly review one of the more well-known challenges – Camelyon-16 Challenge.[8,9] This challenge (cosponsored by the International Society for Biomedical Imaging [ISBI]) consisted of 399 hematoxylin and eosin (H and E)-stained WSIs of sentinel lymph nodes from two hospitals in the Netherlands. This challenge was the first to provide WSIs, and those images were acquired using two different scanners – Pannoramic 250 Flash II (3DHISTECH, Budapest, Hungary) and NanoZoomer-XR (Hamamatsu

Photonics, Hamamatsu City, Japan).[8] As the ground truth, the presence of metastases was annotated under the supervision by expert pathologists. An example of an annotated WSI is presented in Figure 1. A total of 270 images were used for a training set, and 129 digital slides were available as a test set. Two tasks were requested as part of this challenge: (a) identify individual metastases in WSIs and (b) classify each WSI as containing a metastasis or not.[8] The developed solutions had to be able to detect micrometastases and macrometastases but were not required to detect isolated tumor cells. For the first task, the free-response receiver operator characteristic curve (FROC) was used to evaluate [Figure 2] the participants, whereas the second task was evaluated by area under the receiver operating characteristic curve [Table 1].[8] The organizers of the challenge also tested the pathologist results for the same tasks in two settings – unlimited and limited time. The training dataset with accompanying annotations was released for download on December 30, 2015; on March 1, 2016, the test WSI was released with a deadline for submissions on April 1, 2016. The winners were announced during the ISBI workshop on April 13, 2016. A follow-up challenge from this group – CAMELYON17 moved from slide-level evaluation to patient-level evaluation. These tasks were selected for several reasons: (a) this is a tedious, clinically relevant task that occurs in high numbers, (b) the solution would likely generalize to lymph node metastases of other cancers, and (c) the detection of metastatic clusters of tumor cells on H and E slides would require the recognition of subtle textural patterns (solutions would likely advance algorithms in histopathology in general).

RESULTS

The challenges were evaluated for their subject matter (medical discipline) ranging from radiology, pathology, cell biology, cardiology, ophthalmology, dermatology, dental, gastroenterology, and others. Out of the 191 challenges, 24 (12.5%) were related to pathology or combined radiology/pathology. Figure 3 demonstrates the relative concentration of challenges according to the medical discipline. Since
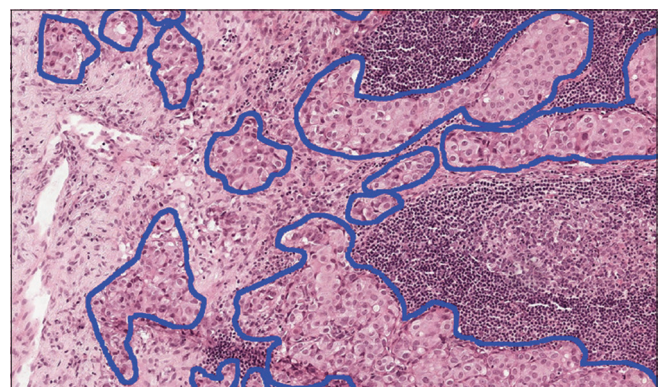

Figure 1: Example of metastatic regions in an H and E-stained sentinel lymph node tissue section, with annotations of metastases by a pathologist (blue lines)

the first challenges in 2007, the number of challenges per year has steadily been increasing [Figure 4]. The medical disciplines with challenges have diverged from initial radiology predominant studies to a much wider range of medical disciplines [Figure 4]. The first challenge involving

Table 1: Results of task 2 of Cancer Metastases in Lymph Nodes 16: Prediction of sentinel lymph node status on the slide level

Team	AUC
Harvard Medical School and MIT, Method 2 (updated)	0.9935
Harvard Medical School, Gordon Center for Medical Imaging, MGH, Method 3	0.9763
Harvard Medical School, Gordon Center for Medical Imaging, MGH, Method 1	0.9650
The Chinese University of Hong Kong (CU laboratory, Hong Kong), Method 3	0.9415
Harvard Medical School and MIT, Method 1	0.9234

AUC: Area under the receiver operating characteristic curve
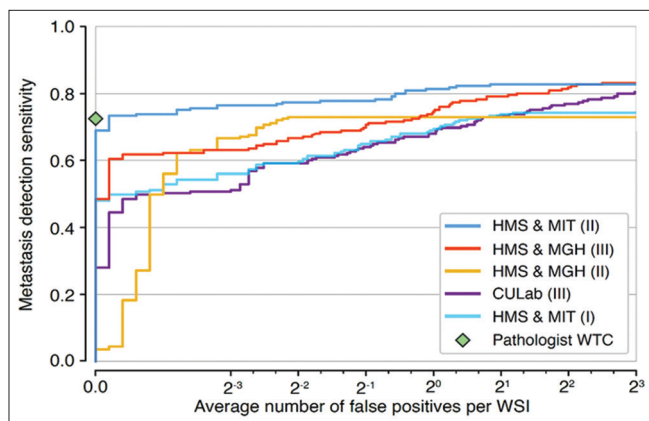

Figure 2: Results of task 1 of Cancer Metastases in Lymph Nodes 16: Detection of individual metastatic regions in SLN whole-slide image. The analysis is performed using the free-response receiver operator characteristic curve, displaying sensitivity versus the number of false positives per whole-slide image. The green diamond indicates the performance of a single pathologist who scored the slides in an experimental setting without any time constraint[8]
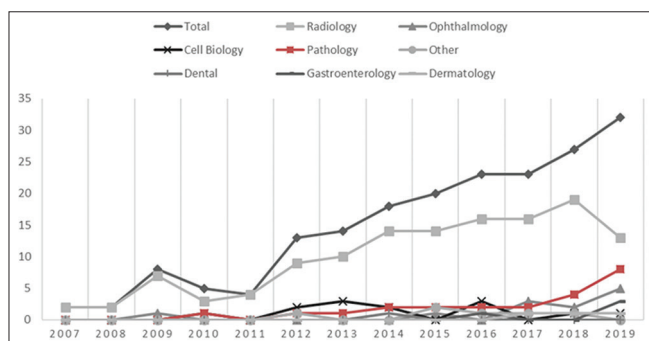

Figure 4: The number of challenges according to the medical discipline over time since the year 2007. The volume of challenges has been steadily increasing and diversifying since 2007. Radiology still account for the majority of challenges, but pathology and ophthalmology are increasing

pathology-related images was in 2010.[10] This first pathology challenge was detecting lymphocytes within H and E-stained slides and counting the number of centroblasts from the cases of follicular lymphoma and was sponsored in conjunction with the International Conference on Pattern Recognition 2010.[11] Initially, web technology and data storage were not sufficiently developed to allow the use of WSIs, and therefore, small, mostly manually selected "field of views" were used. The results in strongly limited applicability of algorithms developed within the challenging context, as the algorithms are not robust to any image content not sufficiently covered in the data sets (e.g., artifacts). A description of the pathology-related challenges is presented in Table 2. Investigation of the challenges related to pathology images demonstrate that the most frequent organ site (45.8%) used for these studies was the breast. Other organ sites included the cervix, central nervous system, thyroid, lung, and multi-organ datasets [Figure 5]. The images within these datasets consist of both "field of views" and WSIs. The number of images was variable per dataset; however, sometimes, it was not provided in the study description. The range of images available was between 15 and 1000 images. The studies inconsistently reported how many patients were represented by the number of images within the dataset. The WSIs within the datasets were sometimes from a single platform, while others provided multiple image file formats (up to 4). The listed image
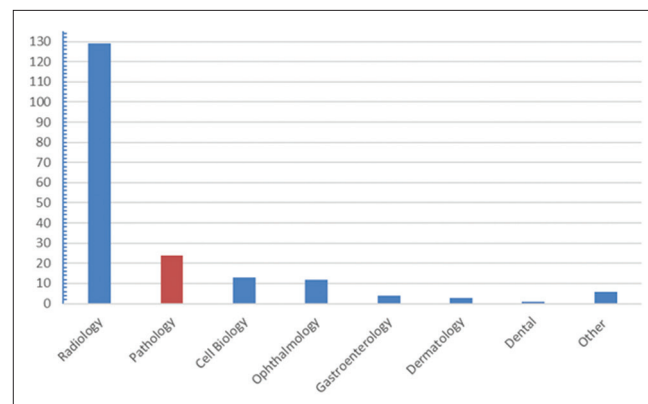
Figure 3: The breakdown of 191 challenges according to the medical discipline of the challenge. Of note, with the exception of one challenge, most of the challenges involve tasks within a single medical discipline
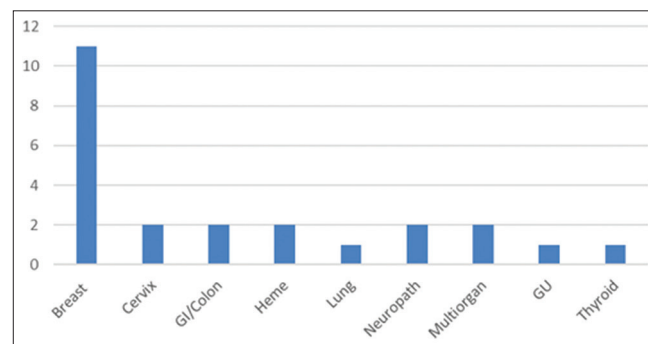
Figure 5: Breakdown of the pathology challenges according to the predominant organ site to study

Table 2: List of pathology-related public challenges since 2010

Years	Challenge name	Description	URL	Participants	Magnification	Image File format
2020	HeroHE ECDP2020	Based on H and E morphological findings, predict Her2 features in breast cancer	https://ecdp2020.grand-challenge.org/	300	NS*	mrzx
2019	Lymphocyte Assessment Hackathon (LYSTO)	Assessment of IHC-stained sections for CD3 and CD8 cells	https://lysto.grand-challenge.org/	245	NS	NS
2019	DigestPath 2019	1) Signet ring-cell detection 2) Colonoscopy tissue segmentation and classification	https://digestpath2019.grand-challenge.org/	647	×40/×20	NS
2019	Gleason 2019	Based on H and E images 1) Pixel-level Gleason grade prediction 2) Core-level Gleason score prediction	https://gleason2019.grand-challenge.org/	139	NS	NS
2019	ACDC-LungHP	Detecting and classifying lung cancer	https://acdc-lunghp.grand-challenge.org/	191	NS	TIFF
2019	ANHIR	Compares the accuracy and speed of automatic nonlinear registration methods for the same tissue stained with different biomarkers (co-registration)	https://anhir.grand-challenge.org/	169	×10-×40	svs; mrzx; ndpi; czi
2019	Patch Camelyon	Create an algorithm to identify metastatic cancer in small-image patchers taken from the larger digital pathology scans	https://www.kaggle.com/c/histopathologic-cancer-detection/rules	NA	×40	tif
2019	BreatPathQ:Cancer Cellularity	Develop an automated method for analyzing histology patches extracted from the whole-slide images and assign a score reflecting cancer cellularity in each	http://spiechallenges.cloudapp.net/competitions/14	NA	×20	NS
2019	B-ALL Classification	Automated classifier that will identify the malignant cells (leukemia) with high accuracy	https://competitions.codalab.org/competitions/20429	NA	NS	bmp
2018	MoNuSeg	This challenge will showcase the best nuclei segmentation techniques that will work on a diverse set of H and E-stained histology images	https://monuseg.grand-challenge.org/	213	×40	svs
2018	ICIAR 2018	Part A: Automatically classifying H and E-stained breast histology microscopy images into normal, benign, *in situ* carcinoma and invasive carcinoma. Part B: Performing pixel-wise labeling of whole-slide images in the same four categories as Part A	https://iciar2018-challenge.grand-challenge.org/	1142	NS	svs/Tiff
2018	Combined Radiology	Evaluate and compare the classification algorithms for lower-grade glioma cases into two subtypes - Oligodendroglioma and astrocyoma	http://miccai.cloudapp.net/competitions/82	271	NS	svs
2018	Digital Pathology Segmentation	Evaluate and compare the algorithms for the detection and segmentation of nuclei in a tissue image	http://miccai.cloudapp.net/competitions/83	NA	×20 and ×40	NS
2017	CAMELYON17	Evaluate algorithms for automated detection and classification of breast cancer metastases in whole-slide images of histologic lymph node sections	https://camelyon17.grand-challenge.org/Home/	1231	NS	TIFF (3DHistech; Hamamatsu; Philips)
2017	Tissue Microarray Analysis in Thyroid Cancer Diagnosis	Build prediction models from H and E patterns, BRAF protein expression, and patient background that produces similar results as the clinical diagnosis from size, extrathyroidal extension, lymph node metastasis, TNM stage, and BRAF mutation	http://www-o.ntust.edu.tw/~cvmi/ISBI2017/	NA	NS	NS
2016	CAMELYON16	Evaluate the algorithms for the detection of lymph node metastases on the lesion level and on the slide level	https://camelyon16.grand-challenge.org/[8]	390	NS	TIFF (3DHistech; Hamamatsu; Philips)
2016	Tumor Proliferation Assessment (TUPAC16)	Evaluate methods that predict the tumor proliferation score directly from the whole-slide images	http://tupac.tue-image.nl/	NA	×40	svs

Contd...

Table 2: Contd...

Years	Challenge name	Description	URL	Participants	Magnification	Image File format
2015	Gland Segmentation Challenge	Create an algorithm to accurately segment glands from H and E images	https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest	NA	×20	bmp (zeiss mirax)
2015	The Second Overlapping Cervical Cytology Image Segmentation Challenge	Create an algorithm that performs cell detection and cell segmentation for the automated analysis of cervical cytology specimens	https://cs.adelaide.edu.au/~zhi/isbi15_challenge/index.html	13	NS	Multilayered cytology volumes
2014	MITOS-ATYPIA-14	Give a score from nuclear pleomorphism and mitotic count	https://mitos-atypia-14.grand-challenge.org/	232	×20	svs and ndpi
2014	Overlapping Cervical Cytology Image Segmentation Challenge	Create an algorithm that performs cell detection and cell segmentation for the automated analysis of cervical cytology specimens	https://cs.adelaide.edu.au/~carneiro/isbi14_challenge/index.html	NA	NS	Extended depth field cytology images
2013	MICCAI Grand Challenge: Assessment of mitosis detection algorithms (AMIDA13)	Evaluate and compare (semi-) automatic mitotic figure detection methods that work on regions extracted from the whole-slide images		NA	×40	svs
2012	Mitotic Count (ICPR 2012)	Mitosis detection in H and E images from breast cancer	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709417/[11]	NA	×40	svs and ndpi
2010	Lymphocyte and Centroblast Count (ICPR 2010)	1) Count lymphocytes within breast cancer 2) Count centroblasts from follicular lymphoma	[10]	23	×40	svs

Please note that the "participants" information is derived from the provided participants by the sponsor of the challenge. This may be defined by the sponsor as the number of downloads of the raw data or by the number of groups who submitted solutions for the leaderboard. Some of the challenges are still open and may still be increasing the number of participants. NS: The challenges have some manipulations to the raw data which makes it difficult to ascribe a specific magnification to the raw data. Historically, this was done to minimize the size of the raw data because of computing limitations and transmission issues. *NS: Not specified, H and E: Hematoxylin and Eosin, CAMELYON: Cancer Metastases in Lymph Nodes, MICCAI: Medical Imaging Computing and Computer-Assisted Intervention, ICPR: International Conference on Pattern Recognition, NA: Not available, IHC: Immunohistochemistry, TNM: Tumor, Node, Metastasis Staging

file formats include. svs., ndpi., mrzx., tif., TIFF., bmp., czi, and extended depth field cytology images.

Evaluating the performance of the algorithms requires a benchmark or ground truth against which the output of the algorithm is compared. The majority of challenges cite "expert" or "experienced" pathologists as the source of ground truth for the dataset ($n = 16$). However, eight challenges did not describe how the ground truth was determined, which presents a major problem. Three challenges cite a single pathologist interpretation (one of which was augmented by molecular profiles). One challenge cited an "expert oncologist," another cited "two medical experts" as the ground truth, and one challenge used Her2 results without specifying the method of Her2 evaluation (immunohistochemistry or fluorescence *in situ* hybridization). Surprisingly, one study used the annotations of engineering students checked by a single pathologist.

Various statistics were used within the challenges to evaluate the outputs from the algorithms. These included F1 score, quadratic weighted kappa, instance-level recall, FROC, DICE coefficient, area under the curve, relative target registration, execution time, prediction probability, weighted precision, weighted recall, aggregated Jaccard index, overall prediction accuracy, accuracy metric, the number of correctly classified cases divided by the total number of cases, Spearman's correlation coefficient, point system for correct score, true positive, true negative, false positive, and false negative. While there is no single method to evaluate various problems, it is very important to follow some commonly accepted evaluation methodologies. These methodologies would vary depending on the nature of the problem as well as how the ground truth is generated. Each challenge needs to pay attention to (1) how the ground truth is generated, (2) what evaluation metrics would be used (e.g., DICE coefficient vs. Hausdorff distance), and (3) asking the participants to submit their evaluation results in a standard format (e.g., extended markup language) so that all the submissions could be evaluated using the same set of evaluation techniques and software. The grand challenge platform contains tools for fully automated assessment of submitted results, further increasing the reproducibility and efficiency.

CONCLUSION

As mentioned previously, AI is rapidly being developed and pathology has not been exempt from these advances. The number of public challenges, including pathology datasets, has been increasing reflecting the increased availability of digital data from pathology. One interesting observation regarding pathology challenges is that there is a disconnect between the types of organs studied and the large volume specimens typically encountered in routine clinical practice.

Dermatopathology and gastrointestinal specimens represent the large majority of specimens received in pathology laboratories; yet, there are no dermatopathology public challenges and only a few for gastrointestinal pathology. Many companies are working within these routine areas internally, but the mismatch in supply from public challenges and demand of clinical practice limit the wider adoption of AI by the pathology community.

Being aware of public challenges for AI research is important for the pathology community. As AI algorithms will likely be marketed to pathologists in the near future, it is important that the pathology community become aware of the conditions under which algorithms are being developed and the performance differences between the algorithms. Through public challenges, a common evaluation method and dataset allow for a better comparison of the performance of the algorithms. Of note, there are still some deficiencies with public challenges. For example, the image file type is sometimes only one file type or a proprietary file type,[12] which may hinder widespread deployment. Even more importantly, in many challenges, data sets consist of WSI from a single or a small number of sources. Even though data augmentation and WSI normalization may be of help,[13] generalizability of algorithms most ideally comes from a diverse data set containing images from a larger number of centers. Furthermore, diverse evaluation metrics is used in these public challenges, leading to difficulty in comparing the algorithms and methods from different challenges.[12] Public challenges offer some degree of transparency into the development process for AI as well as help expand the understanding of this relatively new (to pathology) field (i.e., opening the "black box"). An additional element of the challenges that pathologists should be aware of is the determination of the "ground truth"– this can be highly variable and although it usually involves a pathologist, it does not always. Well annotated and heavily annotated datasets are critical to the success of these challenges, but annotation can be time-consuming and costly. Several groups have explored using "crowd-sourcing" to achieve the annotations.[14,15] Whether using nonpathologists for annotation will be effective for algorithm development is yet to be determined. "Crowd-sourcing" does represent the opportunity to overcome the specific individual bias in morphologic assessment.

Aside from the common ground upon which to evaluate an algorithm, public challenges also foster the development of AI by reducing the start-up costs to commence with AI development. Historically, large curated datasets have been owned by academic medical centers and by companies who were working on developing the technology which reduces competition within the market. Public challenges also add transparency to the process by clearly describing the datasets and establishing routine practices/workflows.

We wish to commend the authors and hosts for these public challenges and encourage further such public challenges in this field. In addition to the authors and the hosts for these public challenges, numerous grants supported these challenges, and we commend those groups for their support (grants cited within the published work associated with challenges are listed in the acknowledgments). We also acknowledge the difficulties associated with generating a public challenge and administering them. The value these public challenges bring to the broader medical community needs to be emphasized.

Conflicts of interest

Douglas Hartman has received an educational honorarium from Philips. Liron Pantanowitz is on the medical advisory board for Leica and Ibex and is a consultant for Hamamatsu. Jeroen van der Laak is a member of the scientific advisory boards of Philips, The Netherlands, and ContextVision, Sweden. Jeroen van der Laak receives research funding from Sectra, Sweden and receives project remuneration from Philips, the Netherlands.

References

1. Niazi MK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019;20:e253-61.
2. Rotemberg V, Halpern A, Dusza S, Codella NC. The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. Semin Cutan Med Surg 2019;38:E38-42.
3. All Challenges. Available from: https://grand-challenge.org/challenges/. [Last accessed on 2019 Oct 24].

4. Hipp JD, Sica J, McKenna B, Monaco J, Madabhushi A, Cheng J, *et al*. The need for the pathology community to sponsor a whole slide imaging repository with technical guidance from the pathology informatics community. J Pathol Inform 2011;2:31.
5. FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems. Available from: https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm. [Last accessed on 2019 Oct 24].
6. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, *et al*. 1399 HE-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. Gigascience 2018;7:1-8.
7. The Cancer Genome Atlas. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. [Last accessed on 2019 Oct 24].
8. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199-210.
9. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast Cancer: Helping artificial intelligence be seen. JAMA 2017;318:2184-6.
10. Gurcan M, Madabhushi A, Rajpoot N. Pattern recogntion in histopathological images: An ICPR 2010 context. Int Conf Pattern Recognit 2010;226-34.
11. Roux L, Racoceanu D, Loménie N, Kulikova M, Irshad H, Klossa J, *et al*. Mitosis detection in breast cancer histological images An ICPR 2012 contest. J Pathol Inform 2013;4:8.
12. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: Challenges and opportunities. J Pathol Inform 2018;9:38.
13. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, *et al*. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal 2019;58:101544.
14. Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MAT, Abo Elnasr LS, *et al*. Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics 2019;35:3461-7.
15. Grote A, Schaadt NS, Forestier G, Wemmert C, Feuerhake F. Crowdsourcing of histological image labeling and object delineation by medical students. IEEE Trans Med Imaging 2019;38:1284-94.