POINT/COUNTERPOINT

Suggestions for topics suitable for these Point/Counterpoint debates should be addressed to Habib Zaidi, Geneva University Hospital, Geneva, Switzerland: habib.zaidi@hcuge.ch; Jing Cai, The Hong Kong Polytechnic University, Hong Kong: jing.cai@polyu.edu.hk; and/or Gerald White, Colorado Associates in Medical Physics: gerald.white@mindspring.com. Persons participating in Point/Counterpoint discussions are selected for their knowledge and communicative skill. Their positions for or against a proposition may or may not reflect their personal opinions or the positions of their employers.

Biomedical image analysis challenges should be considered as an academic exercise, not an instrument that will move the field forward in a real, practical way

Samuel G. Armato III, Ph.D.
Department of Radiology, The University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637, USA (Tel: 773-834-3044; E-mail: s-armato@uchicago.edu)

Keyvan Farahani, Ph.D.
Center for Biomedical Imaging and Information Technology, National Cancer Institute, Bethesda, Maryland, USA (Tel: 240-276-5921; E-mail: farahani@nih.gov)

Habib Zaidi, Ph.D., Moderator

(Received 4 February 2020; revised 5 February 2020; accepted for publication 5 February 2020; published 1 March 2020)

[https://doi.org/10.1002/mp.14081]

OVERVIEW

The biomedical imaging community has witnessed innovative algorithmic developments in quantitative imaging biomarkers taking advantage of modern multimodality imaging technologies. These developments encompass a wide methodological portfolio including but not limited to image registration, segmentation, classification, and prediction. Despite the enormous progress in technical developments, including rigorous peer-review preceding publication in the scientific literature, extensive testing, and feedback from users of the associated open-source software tools, validation of these advanced image analysis tools prior to their deployment in the clinic is still one of the main challenges faced by developers and end-users. To this end, a number of professional societies (e.g., AAPM, MICCAI, RSNA) are organizing challenges aimed at comparing competing image analysis algorithms and providing guidelines and recommendations on clinical deployment and adoption.

While there is consensus that such challenges provide a forum for developers to showcase innovative computational methodologies to solve clinically relevant problems, the exact role of this practice is the subject of debate. Some advocate that these challenges should be considered as an academic exercise, not a mechanism that will move the field forward in a real, practical way, whereas others believe that they do in fact provide a genuine approach to expand the field of biomedical image analysis. This is the topic addressed in this month's Point/Counterpoint debate.

Arguing for the proposition is Samuel G. Armato III, Ph.D. Dr. Armato earned his B.A. in Physics and Ph.D. in Medical Physics degrees from the University of Chicago. He has worked in the Department of Radiology at the University of Chicago since 1991 and is currently Associate Professor, Chair of the Committee on Medical Physics and Director of the Graduate Program in Medical Physics. He is a Fellow of the AAPM and has been very active on AAPM committees, including Chair of the Journals Business Management Committee, co-Chair of the Working Group on Grand Challenges, and a member of the *Medical Physics* and *JACMP* Editorial Boards. His research interests include computerized image analysis and computer-aided diagnostic methods for thoracic imaging, including automated lung nodule detection and analysis in CT scans, semiautomated mesothelioma tumor response assessment, radiomics for the assessment of radiotherapy-induced normal tissue complications, and the automated detection of pathologic change in temporal subtraction images. He has published over 90 papers in peer-reviewed journals. Dr. Armato has been very active in teaching at the University of Chicago and has supervised the research of numerous undergraduate, graduate, and postgraduate students.

Arguing against the proposition is Keyvan Farahani Ph.D. Dr. Farahani obtained a B.S. in Physics from Sonoma State University in California and a PhD. in

Biomedical Physics from the University of California at Los Angeles (UCLA). He is the program director in imaging informatics and the federal lead on the Imaging Data Commons (IDC) at the Center for Biomedical Informatics and Information Technology (CBIIT) at the National Cancer Institute (NCI) in Bethesda, Maryland. Dr. Farahani joined CBIIT in January 2020 after serving in the NCI Cancer Imaging Program as the program director for Image-Guided Interventions, since 2001, and the deputy director for technology development in the Quantitative Imaging Network, since 2018. Dr. Farahani has led algorithmic challenges in cancer imaging and digital pathology for the Medical Image Computing and Computer Assisted Interventions (MICCAI) meetings, the International Symposium on Biomedical Imaging (ISBI), and the AAPM since 2013. In addition, as an active member of the AAPM over the past several years, he has led AAPM Work Groups on Image-Guided Interventions, Research Funding, and Grand Challenges in a chair or a co-chair capacity. Dr. Farahani has co-authored over 60 peer-reviewed journal articles, 10 book chapters, served as a guest editor to several scientific journals and is a fellow of the American Institute for Medical and Biological Engineering (AIMBE).

FOR THE PROPOSITION: SAMUEL G. ARMATO, PH.D

Opening Statement

Biomedical image analysis challenges (often referred to as "grand challenges") have become more common as a growing number of professional societies, governmental agencies, and commercial interests have adopted such challenges as a mechanism for showcasing the key clinical problems of the day while providing a platform for interested research groups to demonstrate their computational approaches to those problems. The AAPM's Working Group on Grand Challenges (WGGC) is charged with promoting the conduct of grand challenges designed to assess or improve the use of medical imaging in diagnostic and/or therapeutic applications. In addition to conducting challenges of its own, the WGGC is now sponsoring its fourth round of challenges selected from among submitted proposals. Given the growing number of challenges and the amount of effort devoted by organizers and participants, the contributions of challenges to the advancement of medical imaging should be considered.

Challenges benefit the research community by allowing for a direct comparison of different algorithms all applied to a common dataset and evaluated with a uniform metric. Of greatest practical significance is the elimination of variability in system performance due to the composition of the database, the reference standard, and the scoring metric used to evaluate system output.[1] As a result, participants directly benefit from the efforts of the organizers to assemble a collection of images with associated metadata, including a reference standard, and from the organizers' evaluation of the results.

While each challenge calls attention to a specific real-world need and motivates research groups from around the world to focus on that single problem, the challenge itself is an academic exercise that, although undeniably important, does not directly advance the field. Most challenges are executed on a tight timeline, often with weeks (or at most a couple months) separating training set release, test set release, and final submission of results. The groups that participate, therefore, must have a mature computational methodology in place prior to the announcement of the challenge, since the timeframe of the challenge offers little (if any) opportunity for developing new algorithms (or substantively modifying existing algorithms) to address the challenge problem; instead, participants will spend their time and effort ensuring that their methods are tailored to the specific images of the challenge (in terms of format and acquisition parameters, for example) and optimizing the performance of their methods relative to the specific evaluation metric adopted by the organizers. The head-to-head comparison of computerized methods in the apples-to-apples setting of a challenge is interesting, invigorating, and even newsworthy, but the challenge environment itself neither facilitates the development of improved algorithms nor brings any of the participating methods closer to clinical reality.

The true contribution of the academic exercise of a grand challenge is the subsequent reporting in the literature of the relative merits of participating methodologies so that the research community can learn which approaches to a specific problem show promise and which methods seem more limited. Another valuable contribution of a challenge is the subsequent release of the test set reference standard as a public resource for algorithm development; some challenge organizers, however, choose to keep the test set secured to provide benchmarking for future methods, which could be a valuable alternative to release.

AGAINST THE PROPOSITION: KEYVAN FARAHANI, PH.D

Opening Statement

The wisdom of the crowd has long been utilized as a practical way for addressing innovation and research questions.[2] In recent years, the use of crowd sourcing has significantly increased, mainly due to ubiquitous access to the internet and social media for implementation and outreach. In the case of biomedical image analysis, increasing access to high-performance computing, offering significantly more efficient computation on large image datasets, the availability of public image repositories[3] for accessing additional training data, and depositing data used in challenges,[4,5] have been instrumental in making challenges practical.

Many applications (or tasks) in biomedical image analysis, ranging from technical (e.g., image segmentation, retrieval, and reconstruction) to clinical (e.g., disease detection, classification, and prediction) may lend themselves well to challenges that benchmark algorithm performance. While

suitability of the application for a challenge largely depends on the data and the ground truth available for training and testing of the solutions, the success of the challenge in conducting a fair and meaningful comparison of algorithms depends on the design, implementation, and use of robust metrics for evaluation of the results.

As with any new development, the use of biomedical imaging challenges has not been without its problems. In a comprehensive evaluation of 150 biomedical image analysis challenges conducted through 2016, Mair-Hein et al. highlighted inconsistencies in quality, evaluation, reproducibility, and ranking.[6] They identified over 50 parameters corresponding to challenge organization, design, and implementation that help improve challenges in a structured way. Implementation of these improvements, along with recent technological advances in cloud computing and requirements for submission of containerized tools to solve the challenge, will facilitate bringing models to the sequestered data[7] and help to raise quality of challenges significantly.

Biomedical image analysis challenges cannot be a substitute for the development and validation of techniques that require years of elaborate research. However, they can complement academic research by (a) providing an open-science forum for head-to-head comparison of algorithms against reference datasets, typically generated outside of any single center,[8–10] (b) promoting algorithmic excellence, and (c) driving consensus on methods and best practices.[11] Moreover, challenges often attract participation by young solvers across the globe, thus democratizing data science and allowing entry of young scientists and innovative solutions into the field of biomedical image analysis. These advances help demonstrate that challenges do in fact provide a real and practical way to advance the field of biomedical image analysis.

REBUTTAL: SAMUEL G. ARMATO, PH.D

There is no question that grand challenges contribute to the field of biomedical image analysis and are a worthwhile endeavor in terms of organizer time, participating group time, and committed financial resources. As my respected colleague (and fellow co-chair of the AAPM's Working Group on Grand Challenges) notes in his opening arguments, the growing acceptance of crowd sourcing as a problem-solving strategy has made the paradigm of challenges more familiar and valued in the culture of science. High-performance computing has greatly expanded the pool of image-based tasks around which a robust challenge may be developed, and challenges do foster the democratization of data science around the world. These arguments establish challenges as feasible and desirable but do not necessarily speak to the ability of challenges to move the field forward technologically.

The benchmarking of algorithm performance is perhaps the most valuable forward-moving aspect of grand challenges in the "real, practical way" sought by the proposition. Through a "fair and meaningful comparison of algorithms" (the "head-to-head" aspect of challenges), the community may understand which methods are most (and least) suited to address a specific task.[12–14] It should be noted that this benefit is achieved only if publications subsequent to the challenge clearly report relevant details of the top-performing (and least-performing) methods.[15] Furthermore, conclusions drawn from the relative performance of participating algorithms are only reliable to the extent that the individual algorithms were designed, implemented, and executed in a manner compatible with the challenge dataset. In other words, just because a specific algorithm performs relatively poorly on a challenge dataset does not necessarily mean the general method is incompatible with the challenge task. The algorithm, for example, may have been developed and optimized on images acquired with different technical parameters or from patients with a different distribution of disease characteristics — and the challenge environment might not allow for detailed tuning of algorithmic parameters.

Arguments both for and against this proposition have expounded on the value of grand challenges to the imaging-research community, and all members of that community should be encouraged to help organize a challenge or participate in a challenge as opportunities arise. The extent to which challenges "move the field forward in a real, practical way," however, is what remains at issue — and, I have argued, grand challenges fall short of achieving this level of impact.

REBUTTAL: KEYVAN FARAHANI, PH.D

As noted by my esteemed colleague, there are several limitations to the use of challenges, but only when viewed through a narrow perspective of what he refers to as the "challenge environment." Perhaps the most beneficial aspect of challenges is their potential to revolutionize the way we view and utilize existing data. This was particularly evident in several challenges by the ImageNet[16] Large-Scale Visual Recognition, as well as the International Symposium on Biomedical Imaging, where a series of winning convolutional neural networks, including those popularly referred to as the AlexNet,[17] the U-Net,[18] and the ResNet[19] offered breakthrough results in image classification, segmentation, detection, and localization. Such challenges have indeed helped to move the field forward, as these networks have now been deployed by many academic and industrial developers of artificial intelligence (AI) in biomedical imaging. Incidentally, it is worth noting that the general approach to a challenge, that is, training, followed by validation and testing, is very similar to that of AI, albeit on smaller scales in data, time, and production.

I would argue too that challenges can help bring participating methods closer to clinical reality. For example, the Data Science Bowl 2017 challenge, "Turning Machine Intelligence Against Lung Cancer,"[20] led by the National Cancer Institute, listed several companies[21–23] on its top 10 leaderboard.[24] These companies have continued to advance their pioneering algorithms into medical imaging AI platforms. They, and numerous other companies that had developed or benchmarked their models and algorithms through challenges, were recently exhibited at the first AI Showcase in RSNA 2019.

As Dr. Armato points out in his opening remarks, there continues to be great interest in challenges and an increasing number of professional biomedical societies, including the AAPM, have adopted them. However, promoting them as an "academic exercise" could seriously jeopardize our efforts to create opportunities to advance innovation through open data science. The medical physics community should heed the examples set by the ImageNet and the AAPM Grand Challenges, to collaborate to assemble very large datasets (i.e., thousands of cases) and ground truth data on some of the key problems in imaging and radiation therapy. By strategically selecting problems suitable for continuous, or sustainable, challenges and careful designing, and partitioning of datasets for each challenge, the AAPM can promote data science in highly targeted ways. This not only will incentivize innovation but create opportunities for collaboration and for developing advances that lead to real-world applications of AI.

CONFLICTS OF INTEREST

Dr. Armato and Dr. Farahani have no relevant conflict of interest.

REFERENCES

1. Armato SG 3rd, Hadjiiski L, Tourassi GD, et al. LUNGx Challenge for computerized lung nodule classification: reflections and lessons learned. *J Med Imaging*. 2015;2):020103.
2. Boudreau K, Lakhani K. Using the crowd as an innovation partner. *Harv Bus Rev*. 2013;91:140.
3. Prior F, Smith K, Sharma A, et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. *Sci Data*. 2017;4:170124.
4. Bakas S, Akbari H, Sotiras A, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117.
5. Elhalawani H, Mohamed A, White A, et al. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci Data*. 2017;4:170077.
6. Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun*. 2018;9:5217.
7. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. *Nature Biotechnol*. 2018;36:391–392.
8. Kalpathy-Cramer J, Zhao B, Goldgof D, et al. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. *J Digit Imaging*. 2016;29:476–487.
9. Farahani K, Kalpathy-Cramer J, Chenevert TL, et al. Computational challenges and collaborative projects in the NCI quantitative imaging network. *Tomography*. 2016;2:242–249.
10. Yang J, Veeraraghavan H, Armato SG 3rd, et al. Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. *Med Phys*. 2018;45:4568–4581.
11. Maier-Hein L, Reinke A, Kozubek M, BIAS – Transparent reporting of biomedical image analysis challenges. *ARXIV*. 2019.preprint arXiv:1910.04071v04073.
12. Athelogou M, Kim HJ, Dima A, et al. Algorithm variability in the estimation of lung nodule volume from phantom CT scans: results of the QIBA 3A public challenge. *Acad Radiol*. 2016;23:940–952.
13. Robins M, Kalpathy-Cramer J, Obuchowski NA, et al. Evaluation of simulated lesions as surrogates to clinical lesions for thoracic CT volumetry: The results of an international challenge. *Acad Radiol*. 2019;26: e161–e173.
14. Bogunovic H, Venhuizen F, Klimscha S, et al. RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. *IEEE Trans Med Imaging*. 2019;38:1858–1874.
15. Armato SG 3rd, Huisman H, Drukker K, et al. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J Med Imaging*. 2018;5:044501.
16. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211–252.
17. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84–90.
18. Ronneberger O, Fischer P, Brox T.U-Net: Convolutional networks for biomedical image segmentation. Conf. Proc. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; pp 234–241.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016, pp 770-778
20. Data Science Bowl. Turning machine intelligence against lung cancer. https://datasciencebowl.com/turning-machine-intelligence-against-lung-cancer/Accessed February 2020.
21. aidence. Artificial intelligence for radiologists. https://www.aidence.com/ Accessed February 2020.
22. md.ai. The platform for medical AI. https://www.md.ai/.Accessed February 2020.
23. Owkin. Machine learning for medical research. https://owkin.com/ Accessed February 2020.
24. Data Science Bowl 2017. Can you improve lung cancer detection? https://www.kaggle.com/c/data-science-bowl-2017/leaderboard Accessed February 2020.