How can we do better? Pitfalls in biomedical challenge design and how to address them

Annika Reinke[1,⋆(✉)], Matthias Eisenmann[1,⋆], Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, Annette Kopp-Schneider⋆, and Lena Maier-Hein[1,⋆(✉)]

[1] Div. Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), Heidelberg, DE
{a.reinke,l.maier-hein}@dkfz.de
[2] See [2] for complete author affiliations

Abstract. Since the first MICCAI grand challenge was organized in 2007 [1], the impact of biomedical image analysis challenges on both the research field as well as on individual careers has been steadily growing. For example, the acceptance of a journal article today often depends on the performance of a new algorithm being assessed against the state-of-the-art work on publicly available challenge datasets. Furthermore, the results are also important for the individuals scientific careers as well as the potential that algorithms can be translated into clinical practice.

Yet, while the publication of papers in scientific journals and prestigious conferences, such as MICCAI, undergoes strict quality control, the design and organization of challenges do not. To investigate the effect of common practice, we have formed an international initiative dedicated to analyzing and improving a variety of aspects related to biomedical challenge design, execution and reporting [2]. In the first part of our abstract presentation at LABELS workshop, we are going to present some of the major pitfalls related to biomedical image analysis challenges today. Specifically, we will look at the following research questions:

RQ1: How robust are challenge rankings? What is the effect of
- the specific test cases used?
- the specific metric variant(s) applied?
- the rank aggregation method chosen (e.g. aggregation of metric values with the mean vs median)?

⋆ Shared first/senior authors.

- the observer who generated the reference annotation?

RQ2: Does the robustness of challenge rankings vary with different (commonly applied) metrics and ranking schemes?

Based on the findings of our study, we will further present best practice recommendations covering a range of different aspects from the size and quality of the datasets, to strategies for missing data handling and methods for computing final rankings.

References

1. van Ginneken, B., Heimann, T., Styner, M.: 3D Segmentation in the Clinic: A Grand Challenge. 3D Segmentation in the Clinic: A Grand Challenge pp. 7–15 (2007)
2. Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., et al.: Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions. arXiv preprint arXiv:1806.02051 (2018)