How to Exploit Weaknesses in Biomedical Challenge Design and Organization

Annika Reinke[1,⋆,✉], Matthias Eisenmann[1,⋆], Sinan Onogur[1], Marko Stankovic[1], Patrick Scholz[1], Peter M. Full[1], Hrvoje Bogunovic[2], Bennett A. Landman[3], Oskar Maier[4], Bjoern Menze[5], Gregory C. Sharp[6], Korsuk Sirinukunwattana[7], Stefanie Speidel[8], Fons van der Sommen[9], Guoyan Zheng[10], Henning Müller[11], Michal Kozubek[12], Tal Arbel[13], Andrew P. Bradley[14], Pierre Jannin[15], Annette Kopp-Schneider[16,⋆], 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] Christian Doppler Laboratory for Ophthalmic Image Analysis, Dept. of Ophthalmology, Medical University Vienna, Vienna, AT
[3] Electrical Engineering, Vanderbilt University, Nashville, Tennessee, US
[4] Inst. Medical Informatics, University of Lübeck, Lübeck, DE
[5] Inst. Advanced Studies, Dept. of Informatics, Technical University of Munich, Munich, DE
[6] Dept. Radiation Oncology, Massachusetts General Hospital, Boston, Massachusetts, US
[7] Inst. Biomedical Engineering, University of Oxford, Oxford, GB
[8] Div. Translational Surgical Oncology (TCO), National Center for Tumor Diseases Dresden, Dresden, DE
[9] Dept. Electrical Engineering, Eindhoven University of Technology, Eindhoven, NL
[10] Inst. Surgical Technology and Biomechanics, University of Bern, Bern, CH
[11] Information System Inst., HES-SO, Sierre, CH
[12] Centre for Biomedical Image Analysis, Masaryk University, Brno, CZ
[13] Dept. of Electrical & Computer Engineering, McGill University, Montreal, QC, CA
[14] Science and Engineering Faculty, Queensland University of Technology, Brisbane, Queensland, AU
[15] Laboratoire du Traitement du Signal et de l'Image, INSERM, University of Rennes 1, Rennes, FR
[16] Div. Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, DE

Abstract. Since the first MICCAI grand challenge organized in 2007 in Brisbane, challenges have become an integral part of MICCAI conferences. In the meantime, challenge datasets are widely recognized as international benchmarking datasets and thus have a great influence on the research community and individual careers. In this paper, we show several ways how to exploit weaknesses related to current challenge design and organization. Our experimental analysis based on MICCAI segmentation challenges organized in 2015 demonstrates that both challenge organizers and participants may undertake measures to substantially tune

[⋆] Shared first/senior authors.

rankings. To overcome these problems we present best practice recommendations for improving challenge design and organization.

1 Introduction

In many research fields, organizing challenges for international benchmarking has become increasingly common. Since the first MICCAI grand challenge was organized in 2007 [4], the impact of challenges on both the research field as well as individual careers has been steadily growing. For example, the acceptance of a journal article today often depends on the performance of a new algorithm compared to state-of-the-art work as assessed on publicly available challenge datasets. Yet, while the publication of papers in scientific journals and prestigious conferences, such as MICCAI, undergoes a strict quality control, the design and organization of challenges does not. Given the discrepancy between challenge impact and quality control, the contribution of this paper can be summarized as follows:

1. Based on an analysis of past MICCAI challenges we show that current practice is heavily based on trust in challenge organizers and participants.
2. We experimentally show how "security holes" related to current challenge design and organization can be used to potentially manipulate rankings.
3. To overcome these problems we propose best practice recommendations against cheating.

2 Methods

Analysis of common practice: To review common practice in MICCAI challenge design, we systematically captured the publicly available information from publications and websites. Based on the data acquired, we generated descriptive statistics on the ranking scheme and several further aspects related to the challenge organization, with a particular focus on segmentation challenges.

Experiments on rank manipulation: While our analysis demonstrates the great impact of challenges on the field of biomedical image analysis it also revealed several weaknesses related to challenge design and organization that can potentially be exploited by challenge organizers and participants to manipulate rankings (see Tab. 2). To experimentally investigate the potential effect of these weaknesses, we designed experiments based on the most common challenge design choices. As detailed in sec. 3, our comprehensive analysis revealed *segmentation* as the most common algorithm category, *single-metric ranking* with *mean* and *metric-based aggregation* as the most frequently used ranking scheme and the *Dice similarity coefficient (DSC)* as the most commonly used segmentation metric. We thus consider single-metric ranking based on the DSC (aggregate with mean, then rank) as the default ranking scheme for segmentation challenges in this paper. For our analysis, the organizers of the MICCAI 2015 segmentation challenges provided the following datasets for all tasks ($n_{tasks} = 50$ in total) of their challenges[1] that met our inclusion criteria[2]: For each participating algorithm ($n_{algo} = 445$ in total) and each test case, the metric values for those metrics $\in$ {DSC, HD, HD95} (HD: *Hausdorff distance (HD)*; HD95: 95% variant) that had been part of the original challenge ranking were provided. Note in this context that the DSC and the HD/HD95 were the most frequently used segmentation metrics in 2015. Based on this data, the following three scenarios were analyzed:

Scenario 1: Increasing one's rank by selective test case submission
According to our analysis, only 33% of all MICCAI tasks provide information on missing data handling *and* punish missing values in some way when determining a challenge ranking (see sec. 3). However, out of the 445 algorithms who participated in the 2015 segmentations tasks we investigated, 17% did not submit results for all test cases. For these algorithms, the mean/maximum amount of missing values was 16%/73%. In theory, challenge participants could exploit the practice of missing data handling by only submitting the results on the easiest cases. To investigate this problem in more depth, we used the MICCAI 2015 segmentation challenges with default ranking scheme to perform the following analysis: For each algorithm and each task of each challenge that met our inclusion criteria[2], we artificially removed those test set results (i.e. set the result to N/A) whose DSC was below a threshold of $t_{DSC} = 0.5$. We assume that these cases could have been relatively easily identified by visual inspection even without having access to the reference annotations. We then compared the new ranking position of the algorithm with the position in the original (default) ranking.

Scenario 2a: Decreasing a competitor's rank by changing the ranking scheme
According to our analysis of common practice, the ranking scheme is not published in 20% of all challenges. Consulting challenge organizers further revealed that roughly 40% of the organizers did not publish the (complete) ranking scheme before the challenge takes place. While there may be good reasons to do so (e.g. organizers want to prevent algorithms from overfitting to a certain assessment method), this practice may – in theory – be exploited by challenge organizers to their own benefit. In this scenario, we assume that the challenge organizers do not want the winning team according to the default ranking to become the challenge winner (e.g. because the winning team is their main competitor). Based on the MICCAI 2015 segmentation challenges, we performed the following experiment for all tasks that met our inclusion criteria[2] and had used both the DSC and the HD/HD95[3]: We simulated 12 different rankings based on the

[1] A challenge may comprise several different tasks for which dedicated rankings/ leaderboards are provided (if any).

[2] Number of participating algorithms > 2 and number of test cases > 1.

[3] Leading to $n = 45$ tasks and $n_{algo} = 424$ for Scenario 2a and 2b.

most commonly applied metrics (DSC, HD, HD95), rank aggregation methods (rank then aggregate vs aggregate then rank) and aggregation operators (mean vs median). We then used Kendall's tau correlation coefficient [6] to compare the 11 simulated rankings with the original (default) ranking. Furthermore, we computed the maximal change in the ranking over all rank variations for the winners of the default ranking and the non-winning algorithms.

Scenario 2b: Decreasing a competitor's rank by changing the aggregation method As a variant of Scenario 2a, we assume that the organizers publish the metric(s) they want to use before the challenge, but not the way they want to aggregate metric values. For the three metrics DSC, HD and HD95, we thus varied only the rank aggregation method and the aggregation operator while keeping the metric fixed. The analysis was then performed in analogy to that of scenario 2a.

3 Results

Between 2007 and 2016, a total of 75 grand challenges with a total of 275 tasks have been hosted by MICCAI. 60% of these challenges published their results in journals or conference proceedings. The median number of citations (in May 2018) was 46 (max: 626). Most challenges (48; 64%) and tasks (222; 81%) dealt with segmentation as algorithm category. The computation of the ranking in segmentation competitions was highly heterogeneous. Overall, 34 different metrics were proposed for segmentation challenges (see Tab. 1), 38% of which were only applied by a single task. The DSC (75%) was the most commonly used metric, and metric values were typically aggregated with the mean (59%) rather than with the median (3%) (39%: N/A). When a final ranking was provided (49%), it was based on one of the following schemes:

Metric-based aggregation (76%): Initially, a *rank for each metric* and algorithm is computed by aggregating metric values over all test cases. If multiple metrics are used (56% of all tasks), the final rank is then determined by aggregating metric ranks.

Case-based aggregation (2%): Initially, a *rank for each test case* and algorithm is computed for one or multiple metrics. The final rank is determined by aggregating test case ranks.

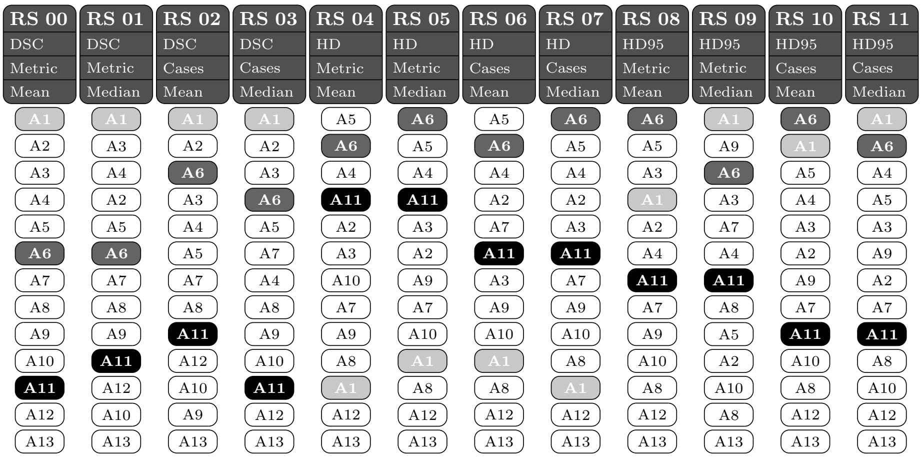
Other (2%): Highly individualized ranking scheme (e.g. [2])

No information provided (20%)

As detailed in Tab. 2, our analysis further revealed several weaknesses of current challenge design and organization that could potentially be exploited for rank manipulation. Consequences of this practice have been investigated in our experiments on rank manipulation:

Scenario 1: Our re-evaluation of all MICCAI 2015 segmentation challenges revealed that 25% of all 396 non-winning algorithms would have been ranked first if they had systematically not submitted the worst results. In 8% of the 50 tasks investigated, every single participating algorithm (including the one ranked last) could have been ranked first if they had selectively submitted results. Note that a threshold of $t_{DSC} = 0.5$ corresponds to a median of 25% test cases set to N/A. Even when leaving out only the 5% worst results, still 11% of all non-winning algorithms would have been ranked first.

Scenario 2a: As illustrated in Fig. 1, the ranking depends crucially on the metric(s), the rank aggregation method and the aggregation operator. In 93% of the tasks, it was possible to change the winner by changing one or multiple of these parameters. On average, the winner according to the default ranking was only ranked first in 28% of the ranking variations. In two cases, the first place dropped to rank 11. 16% of all (originally non-winning) 379 algorithms became the winner in at least one ranking scheme.



RS 00	RS 01	RS 02	RS 03	RS 04	RS 05	RS 06	RS 07	RS 08	RS 09	RS 10	RS 11
DSC	DSC	DSC	DSC	HD	HD	HD	HD	HD95	HD95	HD95	HD95
Metric	Metric	Cases	Cases	Metric	Metric	Cases	Cases	Metric	Metric	Cases	Cases
Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
A1	A1	A1	A1	A5	A6	A5	A6	A6	A1	A6	A1
A2	A3	A2	A2	A6	A5	A6	A5	A5	A9	A1	A6
A3	A4	A6	A3	A4	A4	A4	A4	A3	A6	A5	A4
A4	A2	A3	A6	A11	A11	A2	A2	A1	A3	A4	A5
A5	A5	A4	A5	A2	A3	A7	A3	A2	A7	A3	A3
A6	A6	A5	A7	A3	A2	A11	A11	A4	A4	A2	A9
A7	A7	A7	A4	A10	A9	A3	A7	A11	A11	A9	A2
A8	A8	A8	A8	A7	A7	A9	A9	A7	A8	A7	A7
A9	A9	A11	A9	A9	A10	A10	A10	A9	A5	A11	A11
A10	A11	A12	A10	A8	A1	A1	A8	A10	A2	A10	A8
A11	A12	A10	A11	A1	A8	A8	A1	A8	A10	A8	A10
A12	A10	A9	A12	A12	A12	A12	A12	A12	A8	A12	A12
A13	A13	A13	A13	A13	A13	A13	A13	A13	A13	A13	A13

Fig. 1. Effect of different ranking schemes (RS) applied to one example MICCAI 2015 segmentation task. Design choices are indicated in the grey header: *RS 0x* defines the different ranking schemes. The following three lines indicate the used *metric* $\in$ {DSC, HD, HD95}, the *aggregation method* based on {Metric, Cases} and the *aggregation operator* $\in$ {Mean, Median}. *RS 00* (single-metric ranking with DSC; aggregate with mean, then rank) is considered as the default ranking scheme. For each RS, the resulting ranking is shown for algorithms $A1$ to $A13$. To illustrate the effect of different RS on single algorithms, $A1, A6$ and $A11$ are highlighted.

Scenario 2b: When assuming a fixed metric (DSC/HD/HD95) and only changing the rank aggregation method and/or the aggregation operator (three ranking variations), the winner remains stable in 67% (DSC), 24% (HD) and 31% (HD95) of the experiments. In these cases 7% (DSC), 13% (HD) and 7% (HD95) of all (originally non-winning) 379 algorithms became the winner in at least one ranking scheme. To overcome the problems related to potential cheating, we compiled several best practice recommendations, as detailed in Tab. 2.

Table 1. Metrics used by MICCAI segmentation tasks between 2007 and 2016.

Metric	Count	%
Dice similarity coefficient (DSC)	206	75
Average surface distance	121	44
Hausdorff distance (HD)	94	34
Adjusted rand index	82	30
Interclass correlation	80	29
Average symmetric surface distance	52	19
Recall	29	11
Precision	23	8
95% Hausdorff distance (HD95)	18	7
Kappa	15	5
Specificity	15	5
Euclidean distance	14	5
Volume	12	4
F1-Score	11	4
Accuracy	11	4
Jaccard index	10	4
Absolute surface distance	6	2
Time	6	2
Area under curve	6	2
Metrics used in < 2% of tasks	61	22

4 Discussion

To our knowledge, we are the first to investigate common practice and weaknesses related to MICCAI challenge design and organization. According to our experiments, a number of different ranking design choices (metrics, aggregation method, missing data handling) have a substantial influence on the ranking. Further, the instability of the rankings combined with common practice of reporting/challenge organization can – in theory – be exploited by both challenge participants and organizers to manipulate rankings. Our analysis also revealed that challenge design and organization of MICCAI challenges are highly heterogeneous and lot of relevant information is commonly not reported. While initial valuable steps towards more quality control related to MICCAI challenges have meanwhile been taken, these initiatives have so far been focusing on the selection of challenge proposals, while no quality control process has been put in place to monitor the implementation of the proposed design. A weakness of our experimental analysis could be seen in the fact that we simulated the removed test case results by applying a threshold to the DSC values based on the known reference annotations rather than performing a visual inspection. Yet, we strongly believe that the poorly performing cases with a DSC below 0.5 would have also been identified visually. Our approach, in turn, ensured an objective, scalable and reproducible process. Note that an investigation with the HD/HD95 as metric in an analogous manner would not have been reasonable as a threshold would strongly depend on the task/images. Secondly, it is worth mentioning that instead of applying the different variations of ranking schemes as used in the challenges we focused on the most commonly used ranking scheme in order to perform a statistical analysis that enables a valid comparison across challenges. Given that all rankings of the challenges investigated are based on the DSC as metric, we consider this procedure as valid. Finally, it could be argued that our work is of limited practical value as challenge organizers and participants are fair in general. While this may hold true for the majority, we expect every ”se-

Table 2. Weaknesses of current challenge design and organization that can potentially be exploited by challenge organizers and participants along with best practice recommendations to address existing issues.

Source of problem **→ Consequence**	**Best practice recommendation**
Ranking schemes are often not published before the challenge → Challenge organizers may tune rankings (cf. sec. 3)	Challenge organizers should ... ... consider not generating a final ranking at all ... publish the whole challenge design before the challenge ... make changes in the ranking scheme transparent ... publish their evaluation software
Challenge participants often have access to test data → They may do manual corrections of the algorithm output and/or use the knowledge of the test data to tune their algorithms	Challenge organizers should... ... consider releasing more test cases than are used for validation (and keeping the real ones for which annotations are available confidential). ... consider not releasing test data at all and requiring submission of algorithms [1] or ... arrange on-site competitions and ... ask participants to release their source code
Challenge organizers have access to test data annotations → They may manipulate their results	Challenge organizers and members of the organizers' institute(s) ... ... should not be eligible for awards ... should not participate in their own challenge or otherwise ... should make their participation transparent in the leaderboard Provision of (non-competing) baseline algorithms by the organizers, on the other hand, is encouraged.
Missing data may be ignored when aggregating metric values → Challenge participants may selectively submit test cases to get a better rank (cf. sec. 3)	Missing cases should not be allowed or be punished, e.g. by ... assigning the last rank to those cases in case-based aggregation (see e.g. [7]) ... setting the result to the worst metric value (e.g. 0 for the DSC) in metric-based aggregation, if possible (see e.g. [8])
Sometimes arbitrary number of resubmissions possible → Participants can tune their algorithms based on the performance on the test set	Feedback after a submission should not reveal information on individual cases Only the final submission should be based on the full test set [3].

curity hole" to be exploited sooner or later [5]. Furthermore, our study not only investigates the effect of challenge weaknesses in the context of cheating but also demonstrates the instabilities of rankings for the first time.

In conclusion, we believe that the insights of this study along with the best practice recommendations provided should be carefully considered in future MICCAI challenges. A key message from this paper is to make the challenge design, organization and results as transparent as possible.

Acknowledgments. We thank all of the organizers of the 2015 segmentation challenges who are not co-authoring this paper. We further thank A. Laha, D. Mindroc-Filimon, B. Pekdemir and J. Yoganathan (DKFZ, Germany) for helping with the comprehensive challenge capturing. Finally, we acknowledge support from the European Union through the ERC starting grant COMBIOSCOPY under the New Horizon Framework Programme under grant agreement ERC-2015-StG-37960.

References

1. Boettiger, C.: An Introduction to Docker for Reproducible Research. ACM SIGOPS Operating Systems Review **49**(1), 71–79 (2015). doi: 10.1145/2723872.2723882
2. Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., et al.: Longitudinal Multiple Sclerosis Lesion Segmentation: Resource and Challenge. NeuroImage **148**, 77–102 (2017). doi: 10.1016/j.neuroimage.2016.12.064
3. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A.: The Reusable Holdout: Preserving Validity in Adaptive Data Analysis. Science **349**(6248), 636–638 (2015). doi: 10.1126/science.aaa9375
4. 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)
5. Ioannidis, J.P.: Why Most Published Research Findings are False. PLoS medicine **2**(8), e124 (2005). doi: 10.1371/journal.pmed.0020124
6. Kendall, M.G.: A New Measure of Rank Correlation. Biometrika **30**(1/2), 81–93 (1938)
7. Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al.: ISLES 2015 – A Public Evaluation Benchmark for Ischemic Stroke Lesion Segmentation from Multispectral MRI. Medical Image Analysis **35**, 250–269 (2017). doi: 10.1016/j.media.2016.07.009
8. Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., Urbiola, A., España, T., Venkatesan, S., Balak, D.M., et al.: A Benchmark for Comparison of Cell Tracking Algorithms. Bioinformatics **30**(11), 1609–1617 (2014). doi: 10.1093/bioinformatics/btu080