# Beyond the Leaderboard: Insight and Deployment Challenges to Address Research Problems

**Adriënne M. Mendrik**
Netherlands eScience Center
Science Park 140 (Matrix I)
1098 XG Amsterdam, The Netherlands
a.mendrik@esciencecenter.nl

**Stephen R. Aylward**
Kitware, Inc.
101 East Weaver Street, Suite G4
Carrboro, North Carolina, 27510 USA

## 1 Introduction

In the medical image analysis field, organizing challenges with associated workshops at international conferences began in 2007 and has grown to include over 150 challenges [15]. Several of these challenges have had a major impact in the field, such as the Camelyon Challenge, which showed that deep learning algorithms can outperform a panel of 11 pathologists [1]. However, whereas well-designed challenges have the potential to unite and focus the field on creating solutions to important problems, poorly designed and documented challenges can equally impede a field and lead to pursuing incremental improvements in metric scores with no theoretic or clinical significance.

The potential detriment of challenges inspired a critical assessment of challenges at the international Medical Image Computer and Computer-Assisted Intervention (MICCAI) conference and initiated a definition of best practices for challenge design [13]. The main criticism raised is based on the observation that small changes to the underlying challenge data can drastically change the ranking order in the leaderboard. Related to this is the practice of leaderboard climbing [7, 16], which is characterized by participants focusing on incrementally improving metric results rather than advancing science or solving the challenge's driving problem. In terms of challenge design, ways to avoid these criticisms are, for example, limiting the number of resubmissions, using a hold-out test set for the final leaderboard [4], or only updating the leaderboard in case of a significant change [2].

In this abstract we look beyond the leaderboard of a challenge and instead look at the conclusions that can be drawn from a challenge with respect to the research problem that it is addressing. Research study design is well described in other research areas, such as educational research [3] and can be translated to challenge design when viewing challenges as research studies on algorithm performance that address a research problem. Based on the two main types of scientific research study design, we propose two main challenge types, which we think would benefit other research areas as well: 1) an "insight" challenge that is based on a *qualitative* study design and 2) a "deployment" challenge that is based on a *quantitative* study design.

## 2 Insight and Deployment Challenges

Although a challenge in itself is a quantitative system of comparison, the design of its elements (i.e. data, truth, and metrics) can follow a qualitative study design. Like often in a research setting, circumstances can be sub-optimal in terms of generalization. In medical image analysis, bronze standards for truth [10] are often the best that can be provided due to cost and time constraints as well as uncertainties regarding medical diagnostic categories. Therefore many challenges use small non-randomly selected datasets or in some cases incorporate indirect metrics. These uncertainties in truth, indirect metrics, and small non-randomly selected datasets are the hallmarks of a qualitative experiment. As qualitative experiments, the leaderboards of these challenges or subsequent tests cannot be used to draw quantitative conclusions about the research problem targeted by the challenge.

Attempting to compute quantitative statistics from qualitative experiments is analogous to applying a t-test when the source data was not randomly sampled or does not have a normal distribution.

Challenges that follow a qualitative design, however, can provide insights into a research problem [12, 3]. For example, a cluster analysis of a qualitative challenge's leaderboard (e.g., top 25% of the algorithms) combined with knowledge about those algorithms' methodologies can be used to identify which class of algorithms have a high potential to perform well for a class of problems. Additionally, qualitative challenges can be used to gain insight into how to achieve representative data, truth, and metrics for future challenges, including quantitative challenges. In qualitative research design this is called grounded theory [6]. It is based on systematically analyzing qualitative data to discover hypotheses and population variables, and provide evidence that is grounded in the data. When design choices for insight challenges are openly described with respect to data sampling (e.g. purposive, convenience, etc), truth protocol and used metrics, and clearly follow from a research problem, purpose statement and research question, they can be used to draw conclusions that lead to new insights for research problems. These insights can in turn inspire new challenges addressing the same research problem from a different angle, clearly advancing in the direction of solving the problem or a certain part of the research problem. This is illustrated in Figure 1.

For deployment challenges, the goal is to identify an algorithm that successfully solves a research problem. The data (randomly sampled from the statistical population and sufficiently large), truth, and metrics should be representative of the research problem. Additionally, beyond the leaderboard, a hold-out test should ultimately be used to test the hypothesis that an algorithm successfully solves the research problem, such that it could be deployed.
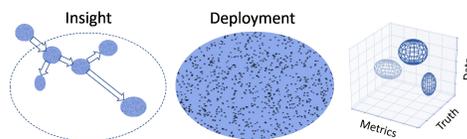


Figure 1: The dotted ellipse illustrates the problem space (statistical population) of a research problem, and how insight challenges cover a small part of this space, but can spark other insight challenges. Deployment challenges should provide data, truth, and metrics that are representative for the entire problem space to ensure generalization. The right figure illustrates how insight challenges can cover different parts of the high dimensional problem space associated with variations in data, truth, and metrics.

Another thing to consider when looking beyond the leaderboard is the way algorithms are compared with each other. One of the most common practices is to test for statistically significant differences between methods; however, statistical significance does not capture generalizability or practical significance. The widespread inappropriate reliance on statistical significance has prompted many statistical journals to discourage its use [17, 11]. Generalizability (e.g., ANOVA methods [9]) captures the potential sources of the differences between the two methods. Practical significance (e.g., scoring-guide scales [5]) indicates if findings have any decision-making utility. In medicine, practical significance is often referred to as "clinical significance" to reflect the fact that small yet statistically significant differences may have no impact on treatment or outcome. For example, a small yet statistically significant improvement in tumor boundary estimation may not enable a reduction in the radiation therapy tumor margins [8].

Some challenges focus on increasing sample size in order to increase statistical significance; however, when looking beyond the leaderboard, such efforts are often marginalized due to data saturation. Data saturation is defined as encompassing insufficiencies in the data, metrics, or truth that diminish the statistical and practical significance as well as the generalizability that can be achieved as sample size is increased. Admittedly, the concept of data saturation has been hotly debated [14]. However, it is particularly relevant to current practices of (a) data augmentation, which may increase the variance in the data in a manner that disrupts the matching of population statistics; and (b) mechanical turk, which may provide more data at the cost of decreased quality of truth which, in turn, negates any benefit from increased sample size.

## 3 Conclusion

Although challenge leaderboards provide a quick overview of algorithm performance and can boost competition, we urge researchers to look beyond the leaderboard for challenge assessment. In this abstract, we make a distinction between insight and deployment challenges to address a research problem, and describe some additional considerations with respect to challenge design.

# References

[1] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[2] A. Blum and M. Hardt. The ladder: A reliable leaderboard for machine learning competitions. *arXiv preprint arXiv:1502.04585*, 2015.

[3] J. W. Creswell. *Educational research: Planning, conducting, and evaluating quantitative*. Prentice Hall Upper Saddle River, NJ, 2002.

[4] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[5] M. Gall. Figuring out the importance of research results: Statistical significance versus practical significance. In *American Educational Research Association Annual Meeting*, 2001.

[6] B. G. Glaser, A. L. Strauss, and E. Strutzel. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364, 1968.

[7] M. Hardt. Competing in a data science contest without reading the data. `http://blog.mrtz.org/2015/03/09/competition.html`, 2015. Accessed: 2018-10-06.

[8] M. Hojat and G. Xu. A visitor's guide to effect sizes–statistical significance versus practical (clinical) importance of research findings. *Advances in health sciences education*, 9(3):241–249, 2004.

[9] K. J. Jager, C. Zoccali, A. MacLeod, and F. W. Dekker. Confounding: What it is and how to deal with it. *Kidney International*, 73(3):256–60, 2008.

[10] P. Jannin, J. M. Fitzpatrick, D. Hawkes, X. Pennec, R. Shahidi, and M. Vannier. Validation of medical image processing in image-guided therapy. *IEEE Transactions on Medical Imaging*, 21 (12):1445–9, 2002.

[11] R. Nuzzo. Scientific method: statistical errors. *Nature News*, 506(7487):150, 2014.

[12] A. J. Onwuegbuzie and N. L. Leech. A call for qualitative power analyses. *Quality & Quantity*, 41(1):105–121, 2007.

[13] A. Reinke, M. Eisenmann, S. Onogur, M. Stankovic, P. Scholz, P. M. Full, H. Bogunovic, B. A. Landman, O. Maier, B. Menze, G. C. Sharp, K. Sirinukunwattana, S. Speidel, F. v. d. Sommen, G. Zheng, H. Muller, M. Kozubek, T. Arbel, A. P. Bradley, P. Jannin, A. Kopp-Schneider, and L. Maier-Hein. How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 388–395. Springer, 2018.

[14] B. Saunders, J. Sim, T. Kingstone, S. Baker, J. Waterfield, B. Bartlam, H. Burroughs, and C. Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, 52(4):1893–1907, 2018.

[15] B. van Ginneken. Grand challenges in biomedical image analysis. `https://grand-challenge.org/challenges/`, 2018. Accessed: 2018-10-06.

[16] J. Whitehill. Climbing the kaggle leaderboard by exploiting the log-loss oracle. *arXiv preprint arXiv:1707.01825*, 2017.

[17] C. Woolston. Psychology journal bans p values. *Nature News*, 519(7541):9, 2015.