Federated Learning under Combined Clinical Heterogeneity: Evaluation Protocol and Inter-Site Equity

Mohamed Achraf Talhaoui*[1] [0009-0001-9097-5164], Zakariae Alami Merrouni*[2] [0000-0002-0903-6160], Bouchra Frikh[1] [0000-0002-9711-0841], and Brahim Ouhbi[3] [0000-0002-6340-5617]

[1] LIASSE Lab, ENSA/USMBA Fez, Morocco
mohamedachraf.talhaoui@usmba.ac.ma
bouchra.frikh@usmba.ac.ma
[2] ERESS Lab, FMPDF/USMBA Fez, Morocco
zakariae.alamimerrouni@usmba.ac.ma
[3] LM2I Lab, ENSAM/UMI Meknes, Morocco
b.ouhbi@umi.ac.ma

Abstract. Federated learning makes it possible to train diagnostic models across several hospitals without moving patient data, which is especially valuable where data is sensitive and hard to share. Real hospital networks, however, are heterogeneous in many ways at once, and most federated benchmarks reduce a method to a single average score that is dominated by the largest sites and silent about the smallest. In this paper we introduce FedMedSim, a controlled and tunable benchmark that combines four sources of clinical heterogeneity, label skew, quantity imbalance, scanner appearance, and temporal drift, on a shared chest radiograph dataset. We compare five federated algorithms across three heterogeneity regimes of varying imbalance severity, and we evaluate not only aggregate performance but also how each method treats its worst-served site. Every evaluated method overfits, and the apparent ranking depends on how performance is measured. More importantly, the site that fares worst tends to be a small, under-resourced center, and among the methods we test FedBN protects it best. These differences are largely hidden by aggregate evaluation and surface only when each site is measured on its own. Our results suggest that equity metrics should complement aggregate evaluation when multi-site medical models are chosen.

Keywords: Federated Learning, Clinical Heterogeneity, Evaluation Protocol, Chest Radiography, Inter-Site Equity.

1 Introduction

Training diagnostic models for medical imaging usually needs data from many hospitals, yet that data rarely leaves the institution that holds it. Privacy rules, governance, and limited infrastructure all stand in the way, and these constraints are sharpest in African health systems, where data is both scarce and hard to move [1]. Federated

* Corresponding authors.

learning offers a way through, since it trains a shared model while each site keeps its own data [2, 3]. The difficulty is that a real network of hospitals is far from uniform. Sites differ in the diseases they see, the data they hold, and the scanners they use, all at once.

Most federated benchmarks do not reflect this. They vary a single factor, often the label distribution, and they summarize a method with one aggregate score [4]. An aggregate is convenient, but it leans toward the largest sites and stays silent about the smallest, which is a problem precisely because the smallest sites are the ones a shared model tends to serve worst. For a deployment that joins large urban hospitals with small rural clinics, the question is not how a model performs on average but how it does at the site that does worst. That question has received little attention under combined heterogeneity, and almost none with an explicit view to equity.

In this paper, we address this gap with three contributions. First, we introduce FedMedSim, a controlled benchmark for evaluation rather than a new training method. Where existing benchmarks typically vary one heterogeneity axis at a time, FedMedSim jointly controls four clinically motivated sources, label skew, quantity imbalance, scanner appearance, and temporal drift, under reproducible settings while explicitly evaluating inter-site equity. Second, we compare five federated algorithms across three regimes in which different random seeds produce varying imbalance severity rather than a hand-set level, so a result that holds across all three is structural rather than incidental. Third, we move beyond aggregate performance to an equity analysis, measuring how each algorithm treats its worst-served site and how far apart its sites end up.

Our central finding concerns that worst-served site. Across regimes it is typically a small, under-resourced center running equipment no other site uses, and among the methods we test FedBN protects it best, both lifting its weakest site and narrowing the spread between sites. This difference is largely hidden by aggregate evaluation and appears only when each site is measured on its own, which is why our results suggest that equity metrics should complement aggregate evaluation when multi-site medical models are chosen.

2 Related Work

Federated learning lets several institutions train a shared model without moving their data [3], and a large body of work studies how it behaves when that data is not identically distributed across sites. The standard way to create such conditions is a Dirichlet partition over labels [5], and the methods we benchmark were each proposed to cope with the resulting drift through a proximal term [6], control variates [7], normalized aggregation [8], or site-local normalization [9]. Most of this work, however, varies one factor at a time and reports a single aggregate score [4].

In medical imaging the dominant source of cross-site variation is the scanner. Differences in equipment measurably shift model performance on chest radiographs, though the effect is moderate rather than catastrophic [10, 11], which is the basis for the appearance shift we simulate. Closest to our setting in spirit is recent work on

federated learning for chest imaging across African sites [1], which establishes the feasibility of training one federated model rather than separate local ones but does not compare algorithms or examine how performance is distributed across sites.

Evaluating a federated model site by site, rather than only on average, has started to draw attention. A recent stress-testing study distributes brain tumor segmentation across simulated hospitals, applies graded appearance shifts, and reports worst-site Dice and inter-hospital disparity as its main metrics [12]; its heterogeneity stays within image appearance. Fairness in federated learning forms a second, longer line. q-FFL adjusts the training objective so that accuracy is spread more evenly across sites [13], and FlexFair, closer to medical imaging, optimizes a model against group-fairness criteria such as equal accuracy and equal opportunity across demographic attributes like age and sex [14]. What these share is a fairness goal built into training. We do something narrower and prior to it: we hold the five algorithms as they are, partition a single chest-radiograph dataset so that label skew, quantity imbalance, scanner appearance, and temporal drift arise together, and observe how the resulting performance across sites shifts with the method and with the evaluation protocol. Equity here is read between sites, not between demographic groups. We study this across three regimes, with an eye toward equitable deployment in African multi-site networks.

3 FedMedSim: Simulating Combined Clinical Heterogeneity

A real federation is heterogeneous along several axes at once, yet real multi-site datasets rarely let us separate these sources, let alone control them. We therefore build them deliberately. FedMedSim takes one shared chest radiograph dataset and reshapes it into a federation whose heterogeneity can be set, tuned, and reproduced for controlled comparative evaluation rather than faithful simulation of any specific network. Its four axes are summarized in Table 1: the first two decide which images each site receives and the last two how those images look.

Table 1. The four axes of heterogeneity simulated by FedMedSim.

Axis	What it varies	Mechanism	Parameter
Label skew	diseases per site	Dirichlet partition	$\alpha = 0.5$
Quantity imbalance	data per site	power law	$\beta = 1.0$
Scanner shift	image appearance	photometric profiles	$s = 0.8$
Temporal drift	drift over time	index-linked drift	$d = 0.04 \cdot i$

Label skew. Disease prevalence is regional, so no two sites see the same case mix. We reproduce this with a Dirichlet partition, the standard approach for non-IID federated studies [4, 5], drawing site proportions independently for each pathology. The concentration parameter α controls how sharp the split is. We set $\alpha = 0.5$, low enough that a pathology common at one site can be nearly absent at another, which is the regime that makes federated training difficult and the one a deployed system is likely to meet.

Quantity imbalance. Patient volume follows a long tail, a few large referral centers hold most of the data, while many small clinics hold little. We reproduce this with a power law [4], giving site i a weight proportional to $(i+1)^{-\beta}$. With $\beta = 1.0$, one site dominates and the others shrink steadily, so the federation contains both a data-rich center and the smaller sites that depend on it.

Scanner and appearance shift. Two radiographs of the same chest taken on different equipment are not identical: they differ in contrast, brightness, tonal response, sharpness, and noise. This shift is documented for radiography, where its effect is real but moderate [10, 11]. We reproduce it with three site profiles, A, B, and C, each combining a different subset of these effects together with a smooth spatial inhomogeneity, and cycle them across the five sites so that one profile is carried by a single minority site, mirroring a center that runs equipment no other site uses. A single severity parameter s scales every effect at once, moving the federation from near-identical images to clearly distinct ones with one control; we work at $s = 0.8$, strong without erasing diagnostic content. The categories of effect follow the imaging literature, while their magnitudes and their assignment to sites are set by design, which is what keeps the benchmark controllable and reproducible.

Temporal drift. Equipment ages, and its output shifts as it is used and recalibrated. We add a small drift that grows with the site index, $d = 0.04 \cdot i$, raising contrast while lowering brightness so that later sites look progressively more altered. It is a secondary effect layered on the scanner shift rather than a competing one, which keeps the appearance axes interpretable.

Combining the axes. The four axes act in two stages. The label and quantity axes are merged multiplicatively to decide which images reach each site; the scanner profile and drift are then applied to those images and cached once with a deterministic per-image seed, fixing each site's appearance as a stable property rather than a random augmentation and making the pipeline reproducible.

Evaluating twice. The same per-site transformation drives the choice that matters most here: how performance is measured. Under the common protocol, every site is scored on one neutral test set, the aggregate view most benchmarks report, though those images resemble no real site. Under the per-site protocol, each site is scored on data carrying its own profile and drift, the conditions it would face in deployment, and methods with site-local components such as FedBN [9] act through their personalized model. This distinction runs through the results that follow.

4 Experimental Setup

Our data is ChestMNIST, the chest radiograph subset of MedMNIST [15], derived from NIH ChestX-ray14 [16]: 28-by-28 grayscale images with fourteen binary pathology labels, released under the CC BY 4.0 license. We keep its official split, partitioning the 78,468 training images across five sites and evaluating on the 22,433 test images under both protocols. The low resolution is deliberate, letting us repeat the full grid of three regimes, five algorithms, and twenty rounds enough for meaningful statistics. Each site trains a DenseNet121 for one local epoch per round, over twenty rounds, using Adam at 10^{-4}, batch size 32, weight decay 10^{-5}, and gradient clipping at 1.0, within the Flower framework [17]. All runs used a single NVIDIA RTX 4060 with PyTorch 2.5.1.

We compare five federated algorithms spanning the field's main families: FedAvg [3] as the baseline, FedProx [6], FedBN [9], SCAFFOLD [7], and FedNova [8].

Three regimes from three seeds. Changing the random seed in most benchmarks just reshuffles noise. Here it does more. Because the imbalance in FedMedSim is not fixed by hand but follows from the label and quantity axes, different random seeds produce federations of varying imbalance severity, which we characterize using the imbalance ratio. Table 2 reports the three we use. The imbalance ratio rises from 4.45 to 28.19, and the largest site grows from 28 to 67 percent of all data, so by the severe regime two thirds of the data sit in one place. A finding that holds across all three is structural, not an accident of one partition.

Table 2. The three heterogeneity regimes. Both the imbalance ratio (IR) and the largest site's data share grow with the seed, giving mild, medium, and severe federations.

Seed	Regime	IR	Largest site share (S0)
123	Mild	4.45	28%
2025	Medium	6.54	56%
42	Severe	28.19	67%

Figure 1 makes one regime concrete, showing the severe partition (seed 42) where sites differ in disease mix, volume, and equipment at once.

Metrics. We report ROC-AUC averaged over the fourteen labels, under both the common and per-site protocols. Alongside it we track two equity measures used throughout the results: the worst-site AUC, and the inter-site gap between the best and worst sites. We prioritize worst-site AUC because deployment safety depends on the least-served hospital rather than the average one, and the inter-site gap complements it by showing how unevenly performance is spread. We favor these two because they target the weakest site directly, rather than summarizing the distribution as a whole. To compare the two protocols directly, we also report Δ, the difference between a method's best per-site and best common AUC, where a positive Δ means it looks better once each site is judged on its own data. All numbers are reported as the mean over the three seeds with their standard deviation.

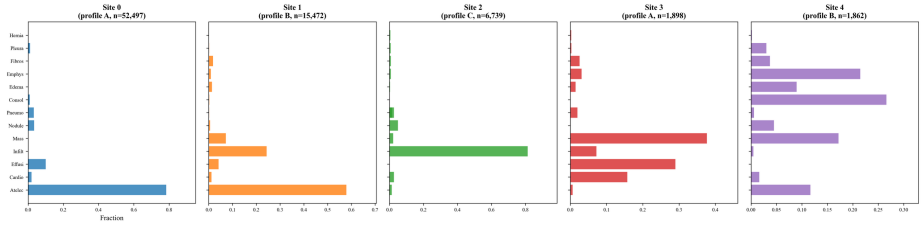

Fig. 1. The severe regime (seed 42) on the training partition: each site's pathology mix, profile, and sample count.

5 Results

Running the five algorithms across the three seeds of our setup, we observe three findings. The first is a failure mode they all share. The second shows that the evaluation protocol, not the algorithm alone, decides what we conclude. The third, and the one that matters most for deployment, is about which method protects the sites that have the least.

Federated training overfits. Every method we test improves for a few rounds, peaks between rounds five and ten, and then declines for the rest of training; across all fifteen runs, the drop from peak to final ranges from six to seventeen percent, with FedNova the worst case in every regime. One likely contributor is the growing dominance of the largest site: its local training AUC climbs toward one, approaching memorization, most sharply where that site holds much of the data, and the shared model degrades as this happens. We do not claim a single clean cause, since reducing a site's weight does not by itself remove the overfitting, but the practical lesson is clear: every evaluated method overfits (Fig. 2), so the choice of which to deploy cannot rest on a single AUC curve.

The evaluation protocol decides the ranking. For each method we measure Δ, the difference between its best per-site AUC and its best common AUC. A positive Δ means the method looks better when each site is judged on its own data and the conditions it would meet once deployed. Only FedBN comes out positive, with an average Δ of $+0.0146$ across the three seeds, while the other four methods are negative in every seed, twelve cases out of twelve (Fig. 3). The reversal is sharp. Under the common protocol FedBN sits near the bottom of the field, and it is the only method that improves under the per-site protocol. The reason is built into the method, since FedBN keeps a separate normalization for each site, so it is penalized on neutral images and rewarded on each site's own. That its design suits the per-site protocol is the whole point. A benchmark that reported only the common protocol would rank FedBN last and advise against it, while the protocol closer to real deployment turns that verdict around, a reminder that algorithm rankings are rarely robust to how performance is measured [18].

FedBN protects the worst-served site. Average AUC hides the question that matters most for deployment, which is what happens to the site that does worst. We look instead at two quantities, the worst-site AUC and the gap between the best and worst sites. FedBN leads on both. Its worst site reaches 0.673 against 0.625 for FedAvg, almost five points higher, and its inter-site gap is 0.030, roughly a third of the 0.084 under FedAvg (Fig. 4). Across the three evaluated seeds, the ordering remained consistent. FedBN ranked first on every equity criterion, and with so few seeds the strength of this result lies in that consistency.

What stands behind the numbers is the identity of the worst site. Under the other methods, the weakest site tends to be a small minority-profile center, the kind of low-volume site that runs equipment no one else does. These are the sites a shared model serves least, pulled as the federation is toward the larger, more typical centers. FedBN breaks that pattern: because the scanner-appearance axis shifts each site's intensity statistics, keeping those statistics local lets FedBN adapt to the shift instead of averaging it away, lifting the minority center substantially so that in most regimes it is no longer the one left behind. For a network that joins large urban hospitals with small rural clinics, this is the result that matters, since it is the rural clinic, not the referral center, that an equitable deployment most needs to protect.

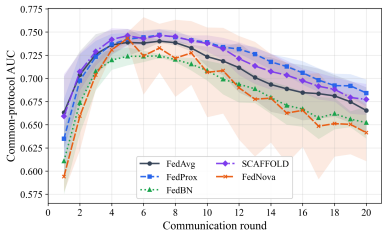

Fig. 2. Convergence curves on the test set under the common protocol, mean ± standard deviation over the three seeds.

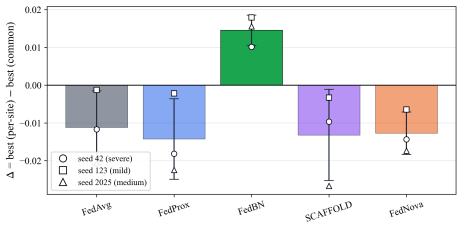

Fig. 3. Δ = best per-site AUC − best common AUC, per algorithm, on the test set. FedBN is the only method positive on all three seeds.

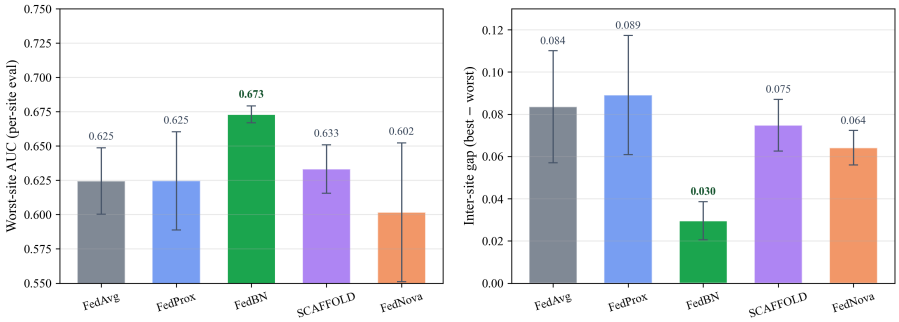

Fig. 4. Worst-site AUC (left, higher is better) and inter-site gap (right, lower is fairer) on the test set at the final round, mean ± standard deviation over the three seeds.

6 Discussion

Our results point to one practical conclusion. When hospitals differ in many ways at once, an average score is a poor way to judge a federated model, since it leans toward the largest sites and says little about the smallest. The site that does worst is not a different one each time; across our regimes, it is typically a small, minority-profile center running equipment no other site uses, easy to miss through the aggregate alone. The consequence for deployment is direct. In a network that links large urban hospitals with small rural clinics, choosing an algorithm by its average performance may pick the one that serves those rural clinics worst, because the cost stays hidden until each site is measured on its own. Of the methods we compare, FedBN raises the weakest site and pulls the others closer together, which is what an equitable deployment asks for. None of this makes FedBN a cure. It overfits like the rest, and its advantage comes from how it is built rather than from solving the deeper problem. The lasting point is not that one algorithm wins, but that the weakest sites are left behind unless equity is measured directly and weighed when a multi-site model is chosen.

7 Limitations and Future Work

The scanner shift and the temporal drift are proxies shaped for radiography: their categories come from the imaging literature, but their magnitudes and their assignment to sites are design choices rather than measurements, so the absolute numbers are not vendor-specific, and both would take a different form on modalities such as MRI or CT. Because a site's profile, drift, and size all follow from its index, the axes are entangled by construction, and we characterize their combined effect rather than any single one. We also work at low resolution, on one task, with three seeds, so our claims rest on the consistency of the direction rather than on statistical power, and we report results at fixed rounds rather than by held-out model selection, a choice a larger study would refine with a validation-based criterion. Most importantly, the benchmark is entirely simulated, with open multi-site African data still scarce, the very scarcity that motivates a controlled benchmark, a real deployment would add challenges we do not model, from heterogeneity that cannot be known in advance to differences in governance, data access, and connectivity across institutions.

These boundaries set the agenda. The most valuable step is validation on real multi-site clinical data, ideally from African networks, with uncertainty quantification across more seeds, to see whether the equity ordering survives outside simulation. Equity itself could be widened: our axes are structural rather than demographic, and extending them to attributes such as age and sex would connect inter-site equity to group fairness. Finer analyses, from per-disease performance to distributional measures beyond the worst site and the gap, would complete the picture, and disentangling the axes so that scanner, drift, and size vary independently would let the benchmark attribute an effect to a specific source. FedMedSim is available from the authors on request as a tunable starting point.

8 Conclusion

When hospitals differ in many ways at once, the way we measure a federated model decides what we are able to see. An average score rewards the largest sites and hides the smallest, yet it is the smallest, often a single under-resourced center, that a shared model serves worst. Measuring each site on its own brings this into view and shows that the choice of algorithm determines whether the weakest site is protected or left behind. We see validation on real multi-site clinical data, particularly from African networks, as the natural next step toward deployments that are equitable as well as accurate.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fabila, J., Garrucho, L., Campello, V.M., Martín-Isla, C., Lekadir, K.: Federated learning in low-resource settings: A chest imaging study in Africa – Challenges and lessons learned. ArXiv Prepr. ArXiv250514217. (2025).
2. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., others: The Future of Digital Health with Federated Learning. Npj Digit. Med. 3, 119 (2020).
3. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (2017).
4. Li, Q., Diao, Y., Chen, Q., He, B.: Federated Learning on Non-IID Data Silos: An Experimental Study. In: IEEE International Conference on Data Engineering (ICDE) (2022).
5. Hsu, T.-M.H., Qi, H., Brown, M.: Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. ArXiv Prepr. ArXiv190906335. (2019).
6. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated Optimization in Heterogeneous Networks. In: Proceedings of Machine Learning and Systems (MLSys) (2020).
7. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.: SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In: International Conference on Machine Learning (ICML) (2020).
8. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In: Advances in Neural Information Processing Systems (NeurIPS) (2020).
9. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In: International Conference on Learning Representations (ICLR) (2021).

10. Pooch, E.H.P., Ballester, P., Barros, R.C.: Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification. In: Thoracic Image Analysis (TIA), MICCAI Workshop. pp. 74–83. Springer (2020).
11. Guo, B., Lu, D., Szumel, G., Gui, R., Li, P., Wang, T., Konz, N., Mazurowski, M.A.: The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. Int. J. Comput. Assist. Radiol. Surg. (2026). https://doi.org/10.1007/s11548-026-03655-7.
12. Naseer, K., Butt, N.A.: MedFL-Stress: A Systematic Robustness Evaluation of Federated Brain Tumor Segmentation under Cross-Hospital MRI Appearance Shift. ArXiv Prepr. ArXiv260509025. (2026).
13. Li, T., Sanjabi, M., Beirami, A., Smith, V.: Fair Resource Allocation in Federated Learning. In: International Conference on Learning Representations (ICLR) (2020).
14. Xing, H., Sun, R., Ren, J., Wei, J., Feng, C.-M., Ding, X., Guo, Z., Wang, Y., Hu, Y., Wei, W., Ban, X., Xie, C., Tan, Y., Liu, X., Cui, S., Duan, X., Li, Z.: Achieving flexible fairness metrics in federated medical imaging. Nat. Commun. 16, 3342 (2025). https://doi.org/10.1038/s41467-025-58549-0.
15. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 – A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci. Data. 10, 41 (2023).
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).
17. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Parcollet, T., Lane, N.D.: Flower: A Friendly Federated Learning Research Framework. In: ICLR Workshop on Distributed and Private Machine Learning (2022).
18. 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., Full, P.M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., Kopp-Schneider, A.: Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7.