Author's Reply

Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, and Amit Sethi

***Abstract*— We had released MoNuSAC2020 as one of the largest publicly available, manually annotated, curated, multi-class, and multi-instance medical image segmentation datasets. Based on this dataset, we had organized a challenge at the International Symposium on Biomedical Imaging (ISBI) 2020. Along with the challenge participants, we had published an article summarizing the results and findings of the challenge (Verma *et al.*, 2021). Foucart *et al.* (2022) in their "Analysis of the MoNuSAC 2020 challenge evaluation and results: metric implementation errors" have pointed ways in which the computation of the segmentation performance metric for the challenge can be corrected or improved. After a careful examination of their analysis, we have found a small bug in our code and an erroneous column-header swap in one of our result tables. Here, we present our response to their analysis, and issue an errata. After fixing the bug the challenge rankings remain largely unaffected. On the other hand, two of Foucart *et al.*'s other suggestions are good for future consideration, but it is not clear that those should be immediately implemented. We thank Foucart *et al.* for their detailed analysis to help us fix the two errors.**

***Index Terms*— Nucleus segmentation, MoNuSAC, computational pathology, challenges.**

I. Introduction

The main objective of creating the multi-organ nucleus segmentation and classification (MoNuSAC) dataset [3] was to encourage the computer vision and computational pathology research community to develop approaches for instance segmentation of nuclei in hematoxylin and eosin (H&E) stained tissue images for further studies of tumor pathobiology. The dataset is one of the largest and most diverse of fully manually annotated and curated medical imaging datasets for any imaging modality, as it contains 46,000 hand-annotated nuclei covering 71 patients, 31 hospitals, four cancer types (kidney, lung, prostate, and breast), and four cell types (epithelial, lymphocytes, neutrophils, and macrophages). The design of the associated challenge (MoNuSAC2020) went through rigorous evaluation for acceptance to the International Symposium on Biomedical Imaging (ISBI) 2020. The objective of the associated article submitted to IEEE Transactions on Medical Imaging [1] was to provide descriptions of the algorithms that were submitted as part of the challenge and inform the reader about directional utility of different techniques. We hope that the resources provided with the article – the dataset and the challenge participants' algorithm details – will support future research in quantitative characterization of the tumor microenvironment.

We used panoptic quality (PQ) [4] as a metric to evaluate the submissions to the challenge and provided a detailed description of PQ computation in our article [1]. We also released the code for further transparency into how the metric was computed [5]. However, the main issues with any metric that we know of are two fold. Firstly, there is always a subjective trade-off between relatively overemphasizing one type or error or the other, or even double-counting some errors. For instance, it is not clear if each error be weighted equally at the object instance-level or the pixel-level. Secondly, when the number of classes, instances, images, and patients increase, there is no perfect aggregation strategy with no disadvantages. For instance, it is not clear if a patient contributes multiple images to a dataset, should a metric be computed at an image-level (especially, if image sizes can be very different) or at the patient-level before aggregation using methods such as simple averaging, or should the size of the images or the number of objects be used as weights. In particular, for multi-class and multi-image segmentation, the best way to aggregate the results is an open problem. We did realize even at the time of proposing the average panoptic quality (a-PQ) metric for the challenge that it may not be a perfect metric, and different readers will find different pros and cons with it. We took a decision to take the arithmetic mean of per-class and per-image PQ results, as stated clearly in Section III.C (including Equation 1) of the post-challenge article [1].

In their analysis [2], Foucart *et al.* seem to have spent considerable effort analyzing our computation of the challenge results. They have claimed that there is an error in our implementation of the evaluation metric and another error in one of the supplementary tables. Additionally, they have made further interpretations based on some of the files on our GitHub repository [5]. Additionally, they have hosted a detailed analysis on their own GitHub repository [6].

We have carefully analyzed their assertions, and have found merit in one code bug and one column header swap pointed out by Foucart *et al.* [2]. The rest of their suggestions are also worth considering for the future, but it is not clear if their suggested solution needs to be immediately incorporated. We recap the assertions by Foucart *et al.* [2], present our assessment of these, and issue the errata based on those that require immediate fixing. Overall, the conclusions of the MoNuSAC2020 challenge [1], [3] remain largely unaffected.

II. Assessment of the Claims by Foucart *et al.*

In this section we distill the assertions by Foucart *et al.* in their Analysis [2] for completeness and conciseness, and present our assessment of each of these, starting with the ones that require immediate fixing.

Manuscript received March 3, 2022; accepted March 3, 2022. Date of current version April 1, 2022. *(Corresponding author: Ruchika Verma.)*

Ruchika Verma is with the Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106 USA, and also with the Alberta Machine Intelligence Institute, Edmonton, AB T5J 3B1, Canada (e-mail: verma.ece.ruchika@gmail.com).

Neeraj Kumar is with the Department of Computing Science, University of Alberta, Edmonton, AB T6G 2R3, Canada, and also with the Alberta Machine Intelligence Institute, Edmonton, AB T5J 3B1, Canada (e-mail: neeraj.kumar.iitg@gmail.com).

Abhijeet Patil, Nikhil Cherian Kurian, and Amit Sethi are with the Department of Electrical Engineering, IIT Bombay, Mumbai 400076, India (e-mail: abhijeetptl5@gmail.com; nikhilcherian30@gmail.com).

Swapnil Rane is with Tata Memorial Centre—ACTREC, HBNI, Mumbai 410210, India (e-mail: raneswapnil82@gmail.com).

Digital Object Identifier 10.1109/TMI.2022.3157048

A. Computation of the Challenge Metric

In the Section II.A of their analysis [2], Foucart *et al.* point that there is a bug in the code released on our GitHub repository [5] in computing the PQ metric, where in certain cases *"a False Positive will be incorrectly added to the tally."* In particular, they point out that *"the problem happens when removing the index of the predicted object. The published code removes the elements where pred indx list == [indx][0], when it should be pred indx list == matched instances[indx][0]. In the challenge version, [indx][0] will resolve to indx, which is the ground truth object index. If this particular ground truth index is not present in the predicted index list, no object will be removed."*

Indeed, this is true. This bug was an oversight on our part when we translated the algorithm for PQ metric described in our article [1] (Section III.C, Equation 1) into code in this single line. We have thoroughly rechecked the code and it seems free of other bugs. Fortunately, as we show in Section III, the impact of fixing this bug is small, and we are releasing the fixed code, corrected PQ scores, and the corrected rankings.

B. Column Header Swap in a Supplementary Table

As pointed in the Table I and Section II.B of Foucart *et al.*'s Analysis [2], there indeed was an inadvertent swap in the column headers of one of tables in the Supplementary Material. The participating entries in the challenge were compared for their overall a-PQ scores, while to add nuance to the results, we also reported results by cell-type. In Supplementary Table S2 we accidentally swapped the column headers for macrophages and neutrophils. However, this has no material effect on the ranking of the challenge entries. We have now thoroughly checked that the rest of the results are free of errors. However, in light of the changes identified in Section II-B, all these numbers also have to updated after the column header swap (see Section III).

C. Claim About Undetected False Positives

In the Section III of their Analysis [2], Foucart *et al.* claim that our implementation of the PQ metric misses a certain type of false positive. According to them, *"the problem is that there is nothing in the provided code that checks for additional files in a team predictions directory without corresponding files in the ground truth directory."*

Here, we would like to point that the false positive for one class will be false negative for another class in an mutually exclusive and exhaustive multi-class multi-instance detection problem. We did not want to double count an error, and therefore our loops for error counting run over the ground truth objects. The interpretation of positives and negatives in multi-class problems is a matter interpretation until settled, and this leads to multiple ways of computing the PQ metric for multi-class problems. We do find merit in Foucart *et al.*'s interpretation also, but because this is not a binary segmentation problem with only one interpretation of a positive match, we think that a further evolution of a multi-class segmentation metric is needed.

D. Aggregation of PQ Metric

In Section IV of their analysis, Foucart *et al.* suggest that if an average of the per-class PQ metric is computed across all the test images in a per-image sense, then it will unfairly weigh the performance of those images that have very few nuclei of that class at par with the other images that have more nuclei of the same class [2]. They suggest that the average should be computed per-patient, and not per-image.

This is a feedback worth considering for future challenges. An antidote to assigning equal weight to all images could be to compute an aggregated PQ index over patients (as suggested by Foucart *et al.*), or even the entire dataset, wherein all the sums in the numerator and denominator of a-PQ computation (Equation 1 in the MoNuSAC article [1]) run over all nuclei of all images. Such an approach would naturally weigh each image by the number of entities or relevant pixels present. Indeed, we had previously taken this latter approach in proposing the aggregated Jaccard index for the single-class classification for the MoNuSeg challenge in 2016 [7]. However, this approach does not allow for a robust computation of confidence intervals, as one computes the metric only once per class for all the images. Therefore, we took the approach described in our article [1] and our code [3], [5].

However, Foucart *et al.*'s claim [2] that *"the post-challenge publication makes it clear that the PQ metric should be computed per-class c and per-patient p as PQ_c^p, and not per sub-image i as PQ_c^i"*, seems to be based on some confusion. Nowhere in our post-challenge publication [1] (see Section III.C and Equation 1 therein) or our GitHub page or code [3], [5] is it mentioned that PQ is first computed for each patient. We repeatedly mention computing the score for each image (called sub-image in [2]) indexed by i. We think that the source of this confusion is a typing error in Section III.C of the post-challenge publication [1], where the number of test images should have been 101 and not 25.

TABLE I

MoNuSAC2020 Challenge Results Update (Errata): The Previous and Updated Ranks of the Teams That Submitted Their Results Till the Challenge Deadline Appear With an "L" Prefix, While Those of the Teams That Submitted Their Results by the Post-Challenge Deadline Appear With a "PL" Prefix – Same as the Convention Followed in Our Original Article [1]

Team Name	Previous Results		Updated Results	
	PQ	Rank	PQ	Rank
Challenge Leaderboard Teams				
TIA-Lab	0.6119	L1	0.6583	L1
SJTU_426	0.5793	L2	0.6183	L2
IVG	0.5084	L3	0.5531	L3
LSL000UD	0.4969	L4	0.5492	L4
Sharif HooshPardaz	0.4808	L5	0.5172	L6
xperience.ai	0.4490	L6	0.4681	L7
Team Tiger	0.4264	L7	0.4530	L8
Amirreza Mahbod	0.3890	L8	0.5478	L5
DeepBlueAI	0.3365	L9	0.3522	L9
Debut_Kele	0.2630	L10	0.2837	L11
the_great_backpropagator	0.1838	L11	0.1912	L12
StevenSmiley	0.1659	L12	0.1728	L13
NUKMLMA	0.1494	L13	0.3154	L10
Post-Challenge Teams				
Amirreza Mahbod	0.5607	PL1	0.6211	PL1
IIAI	0.5493	PL2	0.6049	PL2
Sharif_HooshPardaz	0.5363	PL3	0.5816	PL3
Trnal	0.4540	PL4	0.4896	PL4
Cp&ig	0.3989	PL5	0.4213	PL5
Onward	0.3889	PL6	0.4179	PL6

E. Use of Color-Coded Scratch Results From GitHub

In their Analysis [2], Foucart *et al.* have made several conclusions based on color-coded images that were found on our GitHub repository [5]. We reproduce a list of those statements below:

1) *"Predictions of the "top teams" were not released in this format. Instead, "color-coded" predictions were released, with a single RGB image per sub-image (each class being associated with a color), and borders being added to the objects to show the separation of close or overlapping nuclei."*
2) *"We also test the impact of the bug using a single image from the "SJTU 426" team's prediction. We demonstrate that offsetting the label indices without changing anything else*

TABLE II

AMENDED SUPPLEMENTARY TABLE S2 SHOWING PER-ORGAN AND PER-NUCLEUS-CLASS AVERAGE PQ SCORES FOR THE PARTICIPATING TEAMS (WITH 95% CONFIDENCE INTERVAL) BASED ON UPDATED RANK

Organ	Rank	Epithelial	Lymphocytes	Macrophages	Neutrophils
Breast	L1	0.712 (0.665 - 0.759)	0.621 (0.492 - 0.750)	0.684 (0.634 - 0.735)	0.755 (0.713 - 0.797)
	L2	0.688 (0.602 - 0.774)	0.606 (0.444 - 0.768)	0.614 (0.478 - 0.750)	0.710 (0.622 - 0.798)
	L3	0.489 (0.370 - 0.607)	0.530 (0.438 - 0.623)	0.474 (0.280 - 0.668)	0.532 (0.445 - 0.620)
	L4	0.568 (0.417 - 0.720)	0.538 (0.340 - 0.735)	0.630 (0.606 - 0.655)	0.661 (0.558 - 0.765)
	L5	0.579 (0.429 - 0.729)	0.535 (0.328 - 0.741)	0.509 (0.351 - 0.668)	0.580 (0.356 - 0.805)
	L6	0.630 (0.538 - 0.722)	0.593 (0.439 - 0.746)	0.536 (0.414 - 0.658)	0.659 (0.565 - 0.752)
	L7	0.139 (0.075 - 0.202)	0.154 (0.071 - 0.238)	0.149 (0.012 - 0.286)	0.259 (0.103 - 0.416)
	L8	0.424 (0.322 - 0.527)	0.401 (0.260 - 0.542)	0.385 (0.263 - 0.507)	0.401 (0.315 - 0.486)
	L9	0.298 (0.130 - 0.467)	0.218 (0.048 - 0.388)	0.170 (-0.016 - 0.356)	0.301 (0.180 - 0.422)
	L10	0.486 (0.375 - 0.597)	0.416 (0.230 - 0.601)	0.473 (0.358 - 0.588)	0.508 (0.455 - 0.560)
	L11	0.480 (0.411 - 0.548)	0.390 (0.245 - 0.536)	0.501 (0.391 - 0.611)	0.417 (0.352 - 0.483)
	L12	0.136 (0.089 - 0.183)	0.143 (0.091 - 0.195)	0.104 (0.080 - 0.127)	0.187 (0.141 - 0.233)
	L13	0.112 (0.060 - 0.163)	0.078 (0.020 - 0.135)	0.031 (0.003 - 0.060)	0.113 (0.047 - 0.179)
	PL1	0.644 (0.535 - 0.753)	0.599 (0.439 - 0.759)	0.593 (0.513 - 0.673)	0.738 (0.707 - 0.770)
	PL2	0.715 (0.678 - 0.751)	0.538 (0.342 - 0.733)	0.668 (0.619 - 0.718)	0.644 (0.481 - 0.806)
	PL3	0.595 (0.517 - 0.673)	0.581 (0.482 - 0.681)	0.527 (0.453 - 0.601)	0.627 (0.562 - 0.691)
	PL4	0.599 (0.530 - 0.668)	0.551 (0.417 - 0.686)	0.552 (0.493 - 0.611)	0.639 (0.610 - 0.668)
	PL5	0.381 (0.314 - 0.449)	0.409 (0.285 - 0.532)	0.410 (0.318 - 0.501)	0.454 (0.382 - 0.527)
	PL6	0.539 (0.451 - 0.628)	0.508 (0.338 - 0.679)	0.541 (0.475 - 0.608)	0.589 (0.513 - 0.665)
Kidney	L1	0.582 (0.438 - 0.726)	0.617 (0.521 - 0.714)	0.567 (0.437 - 0.698)	0.602 (0.461 - 0.742)
	L2	0.670 (0.590 - 0.750)	0.586 (0.506 - 0.667)	0.554 (0.456 - 0.653)	0.581 (0.443 - 0.720)
	L3	0.642 (0.572 - 0.712)	0.392 (0.285 - 0.499)	0.569 (0.437 - 0.701)	0.622 (0.546 - 0.699)
	L4	0.544 (0.373 - 0.715)	0.403 (0.232 - 0.574)	0.511 (0.368 - 0.653)	0.545 (0.375 - 0.715)
	L5	0.595 (0.529 - 0.662)	0.440 (0.298 - 0.583)	0.432 (0.296 - 0.568)	0.542 (0.407 - 0.678)
	L6	0.598 (0.508 - 0.688)	0.531 (0.411 - 0.651)	0.505 (0.380 - 0.629)	0.634 (0.534 - 0.734)
	L7	0.489 (0.339 - 0.640)	0.421 (0.259 - 0.582)	0.439 (0.309 - 0.568)	0.456 (0.302 - 0.610)
	L8	0.497 (0.395 - 0.599)	0.393 (0.288 - 0.499)	0.503 (0.383 - 0.624)	0.543 (0.404 - 0.681)
	L9	0.435 (0.335 - 0.535)	0.269 (0.129 - 0.410)	0.325 (0.183 - 0.467)	0.388 (0.226 - 0.550)
	L10	0.431 (0.344 - 0.519)	0.272 (0.146 - 0.399)	0.168 (0.090 - 0.245)	0.336 (0.216 - 0.457)
	L11	0.444 (0.369 - 0.519)	0.255 (0.160 - 0.350)	0.365 (0.253 - 0.477)	0.356 (0.221 - 0.491)
	L12	0.302 (0.213 - 0.390)	0.236 (0.144 - 0.328)	0.280 (0.150 - 0.409)	0.249 (0.170 - 0.329)
	L13	0.224 (0.118 - 0.329)	0.070 (0.027 - 0.112)	0.058 (-0.002 - 0.117)	0.080 (0.015 - 0.144)
	PL1	0.690 (0.630 - 0.749)	0.536 (0.414 - 0.657)	0.562 (0.456 - 0.667)	0.608 (0.497 - 0.719)
	PL2	0.600 (0.474 - 0.725)	0.519 (0.391 - 0.647)	0.539 (0.415 - 0.663)	0.603 (0.459 - 0.747)
	PL3	0.574 (0.490 - 0.658)	0.496 (0.383 - 0.610)	0.500 (0.372 - 0.628)	0.596 (0.471 - 0.721)
	PL4	0.643 (0.581 - 0.705)	0.414 (0.274 - 0.554)	0.535 (0.414 - 0.656)	0.576 (0.430 - 0.722)
	PL5	0.541 (0.476 - 0.605)	0.346 (0.249 - 0.443)	0.484 (0.377 - 0.590)	0.493 (0.344 - 0.642)
	PL6	0.534 (0.419 - 0.649)	0.387 (0.234 - 0.539)	0.365 (0.226 - 0.504)	0.489 (0.330 - 0.648)
Lung	L1	0.298 (-0.183 - 0.780)	0.566 (0.356 - 0.776)	0.502 (0.265 - 0.739)	0.541 (0.410 - 0.672)
	L2	0.380 (-0.003 - 0.763)	0.598 (0.347 - 0.849)	0.692 (0.490 - 0.893)	0.529 (0.238 - 0.819)
	L3	0.323 (0.077 - 0.570)	0.355 (0.144 - 0.567)	0.228 (0.003 - 0.452)	0.441 (0.189 - 0.694)
	L4	0.153 (-0.152 - 0.458)	0.246 (-0.031 - 0.523)	0.042 (-0.075 - 0.159)	0.344 (0.020 - 0.668)
	L5	0.385 (-0.105 - 0.874)	0.600 (0.355 - 0.846)	0.477 (0.121 - 0.833)	0.547 (0.254 - 0.839)
	L6	0.289 (-0.155 - 0.733)	0.356 (0.129 - 0.584)	0.474 (0.208 - 0.741)	0.374 (0.083 - 0.664)
	L7	0.362 (0.061 - 0.663)	0.431 (0.228 - 0.635)	0.561 (0.244 - 0.878)	0.439 (0.171 - 0.708)
	L8	0.179 (-0.116 - 0.474)	0.386 (0.245 - 0.527)	0.059 (-0.047 - 0.165)	0.251 (0.077 - 0.425)
	L9	0.167 (0.048 - 0.286)	0.264 (0.043 - 0.486)	0.229 (0.077 - 0.381)	0.239 (-0.012 - 0.491)
	L10	0.236 (-0.211 - 0.683)	0.311 (0.112 - 0.510)	0.211 (0.148 - 0.273)	0.395 (0.228 - 0.563)
	L11	0.094 (-0.206 - 0.395)	0.134 (0.013 - 0.254)	0.025 (-0.044 - 0.093)	0.079 (-0.043 - 0.202)
	L12	0.009 (-0.020 - 0.039)	0.000 (0.000 - 0.000)	0.031 (-0.056 - 0.118)	0.006 (-0.004 - 0.015)
	L13	0.081 (-0.178 - 0.340)	0.177 (0.025 - 0.329)	0.136 (0.010 - 0.262)	0.214 (0.053 - 0.376)
	PL1	0.316 (-0.220 - 0.853)	0.491 (0.280 - 0.702)	0.567 (0.237 - 0.897)	0.540 (0.255 - 0.825)
	PL2	0.225 (-0.210 - 0.659)	0.546 (0.343 - 0.749)	0.473 (0.218 - 0.728)	0.414 (0.211 - 0.618)
	PL3	0.305 (-0.020 - 0.630)	0.467 (0.197 - 0.737)	0.475 (0.360 - 0.590)	0.432 (0.159 - 0.706)
	PL4	0.215 (-0.079 - 0.510)	0.266 (0.051 - 0.481)	0.200 (-0.014 - 0.415)	0.401 (0.154 - 0.648)
	PL5	0.028 (-0.061 - 0.116)	0.041 (-0.022 - 0.103)	0.210 (-0.064 - 0.483)	0.117 (0.027 - 0.207)
	PL6	0.141 (-0.240 - 0.521)	0.141 (-0.051 - 0.332)	0.264 (0.011 - 0.516)	0.094 (-0.010 - 0.197)
Prostate	L1	0.681 (0.522 - 0.841)	0.699 (0.555 - 0.844)	0.814 (0.754 - 0.874)	0.751 (0.652 - 0.849)
	L2	0.503 (0.187 - 0.819)	0.484 (0.298 - 0.671)	0.697 (0.562 - 0.833)	0.687 (0.504 - 0.870)
	L3	0.745 (0.588 - 0.901)	0.534 (0.355 - 0.714)	0.587 (0.337 - 0.837)	0.726 (0.619 - 0.832)
	L4	0.664 (0.350 - 0.978)	0.586 (0.429 - 0.744)	0.597 (0.368 - 0.825)	0.757 (0.690 - 0.824)
	L5	0.422 (-0.074 - 0.917)	0.480 (0.288 - 0.672)	0.400 (0.100 - 0.701)	0.499 (0.270 - 0.728)
	L6	0.349 (-0.037 - 0.735)	0.363 (0.213 - 0.513)	0.267 (-0.033 - 0.566)	0.473 (0.304 - 0.642)
	L7	0.499 (0.164 - 0.834)	0.653 (0.479 - 0.826)	0.441 (0.203 - 0.678)	0.593 (0.397 - 0.789)
	L8	0.541 (0.200 - 0.883)	0.472 (0.295 - 0.650)	0.326 (0.116 - 0.535)	0.525 (0.375 - 0.676)
	L9	0.237 (-0.073 - 0.547)	0.386 (0.225 - 0.546)	0.321 (0.134 - 0.508)	0.423 (0.211 - 0.634)
	L10	0.407 (-0.074 - 0.887)	0.061 (-0.016 - 0.139)	0.192 (-0.026 - 0.410)	0.333 (0.147 - 0.519)
	L11	0.098 (-0.214 - 0.411)	0.013 (-0.015 - 0.042)	0.035 (-0.048 - 0.119)	0.219 (0.004 - 0.434)
	L12	0.062 (-0.054 - 0.178)	0.138 (0.034 - 0.241)	0.190 (-0.001 - 0.381)	0.221 (0.085 - 0.356)
	L13	0.271 (-0.118 - 0.659)	0.296 (0.165 - 0.427)	0.228 (0.063 - 0.393)	0.299 (0.090 - 0.508)
	PL1	0.513 (0.393 - 0.633)	0.632 (0.470 - 0.793)	0.494 (0.216 - 0.772)	0.703 (0.629 - 0.778)
	PL2	0.672 (0.407 - 0.937)	0.609 (0.441 - 0.776)	0.789 (0.723 - 0.856)	0.719 (0.619 - 0.819)
	PL3	0.598 (0.197 - 1.000)	0.614 (0.487 - 0.742)	0.501 (0.270 - 0.732)	0.680 (0.577 - 0.784)
	PL4	0.448 (-0.112 - 1.007)	0.213 (0.048 - 0.378)	0.452 (0.176 - 0.727)	0.667 (0.524 - 0.810)
	PL5	0.230 (-0.113 - 0.573)	0.513 (0.364 - 0.661)	0.386 (0.133 - 0.639)	0.475 (0.277 - 0.673)
	PL6	0.378 (-0.085 - 0.840)	0.302 (0.155 - 0.449)	0.319 (0.096 - 0.542)	0.239 (0.026 - 0.451)
All	L1	0.603 (0.520 - 0.685)	0.635 (0.572 - 0.699)	0.631 (0.557 - 0.704)	0.665 (0.603 - 0.727)
	L2	0.622 (0.558 - 0.687)	0.560 (0.485 - 0.635)	0.612 (0.551 - 0.672)	0.630 (0.555 - 0.704)
	L3	0.567 (0.502 - 0.632)	0.458 (0.385 - 0.531)	0.512 (0.423 - 0.602)	0.600 (0.544 - 0.655)
	L4	0.520 (0.420 - 0.620)	0.461 (0.369 - 0.552)	0.490 (0.393 - 0.586)	0.595 (0.510 - 0.679)
	L5	0.545 (0.471 - 0.619)	0.498 (0.412 - 0.584)	0.445 (0.356 - 0.535)	0.542 (0.455 - 0.629)
	L6	0.543 (0.468 - 0.617)	0.459 (0.385 - 0.534)	0.458 (0.370 - 0.545)	0.564 (0.495 - 0.634)
	L7	0.362 (0.271 - 0.453)	0.445 (0.354 - 0.537)	0.403 (0.314 - 0.492)	0.439 (0.352 - 0.527)
	L8	0.441 (0.372 - 0.511)	0.419 (0.351 - 0.487)	0.389 (0.305 - 0.472)	0.463 (0.392 - 0.534)
	L9	0.336 (0.261 - 0.411)	0.296 (0.219 - 0.373)	0.284 (0.203 - 0.365)	0.354 (0.272 - 0.437)
	L10	0.423 (0.353 - 0.493)	0.239 (0.168 - 0.310)	0.233 (0.168 - 0.298)	0.385 (0.320 - 0.450)
	L11	0.374 (0.302 - 0.446)	0.183 (0.125 - 0.241)	0.278 (0.196 - 0.360)	0.298 (0.222 - 0.375)
	L12	0.186 (0.130 - 0.241)	0.147 (0.099 - 0.194)	0.198 (0.123 - 0.272)	0.193 (0.145 - 0.241)
	L13	0.176 (0.116 - 0.236)	0.162 (0.107 - 0.217)	0.098 (0.051 - 0.145)	0.159 (0.099 - 0.219)
	PL1	0.610 (0.542 - 0.679)	0.571 (0.498 - 0.643)	0.554 (0.480 - 0.629)	0.652 (0.596 - 0.708)
	PL2	0.601 (0.520 - 0.682)	0.556 (0.480 - 0.631)	0.605 (0.533 - 0.677)	0.613 (0.540 - 0.687)
	PL3	0.552 (0.490 - 0.614)	0.545 (0.479 - 0.612)	0.502 (0.430 - 0.574)	0.600 (0.537 - 0.663)
	PL4	0.556 (0.484 - 0.627)	0.351 (0.269 - 0.433)	0.478 (0.394 - 0.562)	0.587 (0.517 - 0.658)
	PL5	0.392 (0.320 - 0.464)	0.359 (0.286 - 0.432)	0.415 (0.339 - 0.491)	0.426 (0.346 - 0.505)
	PL6	0.471 (0.390 - 0.552)	0.341 (0.261 - 0.421)	0.374 (0.292 - 0.457)	0.397 (0.304 - 0.489)

about the prediction mask leads to computed PQs ranging from 0.385 (completely unaligned indices) to 0.501 (completely aligned indices) for the Lymphocyte class."

3) *"The provided color-coded images, however, are sufficient to demonstrate the problems in the computation of the metric."*
4) *"To check if this was the case in the challenge results, we recomputed the full PQ metric of the SJTU 426 team based on the color-coded masks."*
5) *"Looking at the history of the PQ metric.ipynb file2, we can see a "result dump" for several of the participating teams,*

showing the per-image, per-class PQ computed on the entire test set. According to this result dump, the score computed for the image and class mentioned above of the SJTU 426 team was 0.381, which is very close to our "worst case scenario" score of 0.385."

We would like to emphasize that the color-coded images hosted on GitHub [5] were never meant to be a proxy for official challenge results, and we did not include any explanatory text to give an impression otherwise. Some of these images were generated using preliminary results and preliminary codes. These images should not be used for making any conclusions about the challenge results.

Additionally, we had committed to the challenge participants that their code and raw results will not be released, in order to encourage participation, for example, from commercial entities that are sensitive about their intellectual property falling into public domain.

Although, we appreciate the efforts made by Foucart *et al.* in analyzing these color-coded images, we caution them and the other readers against reading too much into the color-coded results of some of the test images in our GitHub dump [5], as these might be old or even incorrect, and these are not the official results.

F. Foucart et al.'s Implementation of Metric Computation

In their Analysis [2], Foucart *et al.* have asserted that they have shown where to fix our code, reproduced erroneous results using the previous version of the code, and hosted their own code and results on their GitHub repository [6]. We reproduce a list of those statements below:

1) *"A full technical description of the errors as well as all the code necessary to reproduce our results are available on GitHub: https://github.com/adfoucart/monusac-resultscode-analysis"*
2) *"In our code on GitHub, we use synthetic data to check that the published code behaves as we described above, and that replacing the problematic line in the code with our fix makes the problem disappear."*
3) *"Based on our analysis (see supplementary materials on GitHub), 439 nuclei detected by the SJTU 426 team are with no corresponding ground truth object but are not counted as False Positives."*

We have examined Foucart *et al.*'s code and found it to be in line with their analysis. In particular, the suggested change to fix the bug in the indexing of nuclei for calculating the challenge metric is correct, and we have incorporated it in our errata. On the other hand, for the part of their code that claims to fix the missed false positives, we leave the choice of whether or not to double-count an error (false positive for one class and false negative for another class in a multi-class problem) to the reader, until a more unambiguous metric comes along.

III. Errata and Impact

We now show the impact of the bug discovered by Foucart *et al.* in our code for computation of the PQ metric, and the inadvertent column header swap in a supplementary results table.

A. Fixing the PQ Metric Computation

Based on fixing the bug in our code for computing the PQ metric discovered by Foucart *et al.* (Section II.A of the Analysis [2] and Section II-A of this Errata), there were changes in the a-PQ metric of the challenge entries. Fortunately, the changes are small in magnitude and the affect on the resulting ranking is also small. The changes in the metric and the rankings are shown in Table I. We have made these changes on the challenge webpage [3] as well, and clarified the change and the bug-fix on both the challenge webpage and our GitHub repository [3].

B. Typing Error for the Number of Images

In our post-challenge publication, the number of test images in Section III.C of our post-challenge publication [1] should be 101, and not 25. The number of test patients was 25, but the PQ metric was averaged at the per-image level, as clearly stated previously [1].

C. Changes in Per-Organ Results in Supplementary Material

Based on fixing the bug in our code discovered by Foucart *et al.* (Section II.A of [2] and Section II-A of this Errata) and the correction of the inadvertent column-header swap in the supplementary results, the Supplementary Table S2 in the Supplementary Material of the post-challenge publication [1] and the Challenge webpage [3] was recomputed, and is shown here as Table II. We have also reformatted the table for more clarity.

IV. Conclusion and Future Directions

We thank Foucart *et al.* for their thorough and careful analysis [2] of the MoNuSAC challenge results in the manuscript [1] and its webpage [3], as well as carefully combing through our metric computation and aggregation code [5]. Due to Foucart *et al.*'s analysis, we were able confirm and fix a bug in our code.

Through the impact analysis and the errata here we confirm that the impact of the bug is small, and the overall directional findings of the MoNuSAC Challenge remain largely unaffected [1], [3].

Our publication of the error metric and its code along with Foucart *et al.*'s analysis highlights importance of openness and transparency in organizing medical image analysis challenges. Various such challenges are, fortunately, going in this direction of increased openness and robustness in the evaluation of the results. For instance, the Challenges Track for ISBI 2022, of which of the co-authors of this Errata is a Co-Chair, is not only recommending to the challenge organizers to declare their metric publicly beforehand, but also recommending that the organizers allow a certain period within which the challenge participants can suggest amendments or alternatives with proper reasoning before fixing the final metric.

Another valuable outcome of Foucart *et al.*'s analysis is to highlight the need for better metrics for multi-class, multi-instance, and multi-image segmentation problems, and their aggregation methods. This search for metrics is even more relevant to medical image datasets, where there are hierarchies of possible aggregation levels from patients to cohorts to hospitals or labs to populations, for instance. As is clear from this debate, there are pros and cons of various aggregation methods, and we hope that more widely debated and accepted multi-class and multi-instance segmentation metric and their aggregation methods will evolve with such analyses.

References

[1] R. Verma *et al.*, "MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3413–3423, Dec. 2021.
[2] A. Foucart, O. Debeir, and C. Decaestecker, "Analysis of the MoNuSAC 2020 challenge evaluation and results: Metric implementation errors," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 997–999, Apr. 2022.
[3] *MoNuSAC 2020*. Accessed: Dec. 1, 2021. [Online]. Available: https://monusac-2020.grand-challenge.org/
[4] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9404–9413.
[5] *MoNuSAC Grand-Challenge Organized at ISBI 2020*. Accessed: Dec. 1, 2021. [Online]. Available: https://github.com/ruchikaverma-iitg/MoNuSAC
[6] *Analysis of MoNuSAC*. Accessed: Dec. 1, 2021. [Online]. Available: https://github.com/adfoucart/monusac-results-code-analysis
[7] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1550–1560, Jul. 2017.