HTT Test Report

24 August, 2023 - 10:29:35 AM

1 Summary of performance on the "proficiency1" test using the "camic" platform.

Table 1: Your agreement with each expert.

	Threshold	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Criterion	Pass?
LE	10	0.682	0.923	0.933	0.714	0.933	0.933	0.591	PASS
GT	10	1.000	0.933	0.933	1.000	0.882	1.000	0.647	PASS
LE	40	0.581	0.789	0.783	0.783	0.857	0.842	0.613	No
GT	40	1.000	0.889	1.000	1.000	1.000	0.909	0.444	PASS

- "Performance" in this document is understood to be "agreement with experts".
- LE: Agreement with each expert on cases less than or equal to the threshold.
- GT: Agreement with each expert on cases greater than the threshold.
- The pass criterion is the minimum Expert vs. Expert agreement observed.
- In each row, your agreement with each expert must be above the criterion to pass.

You have not passed all criteria of the proficiency1 test.

In order to participate as an expert in the HTT pivotal study, you must pass all four criteria for the **proficiency** test.

We invite you to review the rest of this document to understand the full context of the analysis and how these results are generated. In addition to the agreement analysis results, the **feedback** test report also provides the raw sTILs density data for you and the experts. This data will allow you to reproduce and understand all agreement results. This can help you understand where you might need improvement.

2 Introduction

Thank you for taking the High-Throughput Truthing (HTT) project interactive training for the assessment of stromal tumor-infiltrating lymphocytes (sTILs) in triple-negative breast cancer biopsies [1].

In this report, we refer to you as the **"reader"**.

- This is a report of your performance on the **proficiency1** test using the **camic** platform.
- The **feedback** test is considered training. You are not required to pass the feedback test criteria to participate in the HTT pivotal annotation study.
- The **proficiency** test is used to determine whether you will be considered an expert for the HTT pivotal annotation study and can participate. You must pass all the proficiency test criteria for agreement above and below all thresholds.

In this report, we compare your annotations to the annotations from a panel of six experts. The primary comparisons do not aggregate or average the expert annotations. Instead, we compare your annotations to each of six individual experts, one at a time. This means you will have six results for each agreement (performance) metric, one for each expert acting as the reference standard. We also compare each expert with the remaining experts for head-to-head comparisons of your reader-expert agreement to expert-expert agreement.

The primary agreement metrics that determine whether you "Pass" the test are based on your sTILs density scores. We apply a threshold to your score and to the score of the expert to create a three-by-three frequency table of the paired scores for the regions of interest (ROIs). The three categories are 1) "Not Evaluable", 2) Evaluable and less than or equal to the threshold (LE), and 3) Evaluable and greater than the threshold (GT). Then, we calculate your rate of agreement with each expert on cases above and below the threshold. This is detailed in the next section. The same analysis is performed to calculate the rates of agreement between experts. The lowest of these expert-expert rates of agreement is the criterion for a passing score.

To get a more holistic perspective of your scoring, we also show **scatter plots**. We first show your sTILs density against the sTILs density of each expert, followed by plots of your sTILs density against the average expert sTILs density.

In addition to the agreement analysis results, the **feedback** test report also provides the raw sTILs density data for you and the experts. This data will allow you to reproduce and understand all agreement results. This can help you understand where you might need improvement.

Finally, you can review the image of the ROI, expert results, and your data for the feedback ROIs by visiting the platform where you took the test, the reference document, and the table of raw data at the end of the feedback test report. It is worth noting that the distributions of ROIs in the feedback and proficiency tests are different. We selected the ROIs according to inter-pathologist variability as determined from the annotations of crowd-sourced pathologists that participated in the pilot study. We then selected ROIs with the highest and lowest variability in a 2:1 ratio. As a result, the ROIs in the feedback test had higher inter-pathologist variability than the ROIs in the proficiency test, making the feedback test a little harder. A more complete description of how these cases were selected can be found here [2].

3 Example for Agreement Analysis

Here we provide an **example** that describes the calculation of the primary agreement metrics with which your performance will be scored: the observed rate of agreement above or below a threshold. Below is a scatter plot and a table. The scatter plot shows the paired scores for one of the experts, who will be treated here as "The Reader", and another expert, who will be treated as "The Expert" (the reference standard). We have also added the diagonal line of equality. The axes of the scatter plot are re-scaled (by the square root operator) to better show the points at the low end. The size of each circle corresponds to the number of overlapping points at an x,y location, as shown in the legend. For example, the circle at the top right corner of the plot corresponds to four cases where Expert 4 gave a score of 90 and Expert 2 gave a score of 95. From 36 total paired observations, 25 were labeled "Evaluable" and appear in the plot, while 11 were labeled "Not Evaluable" by the reader, the expert, or both and don't appear in the plot. There are 4 paired observations clearly above the diagonal line of equality, 13 clearly below the line, and 8 very close to or on the line. The title indicates that the "Reader sTILs density scores are 4 points higher on average."

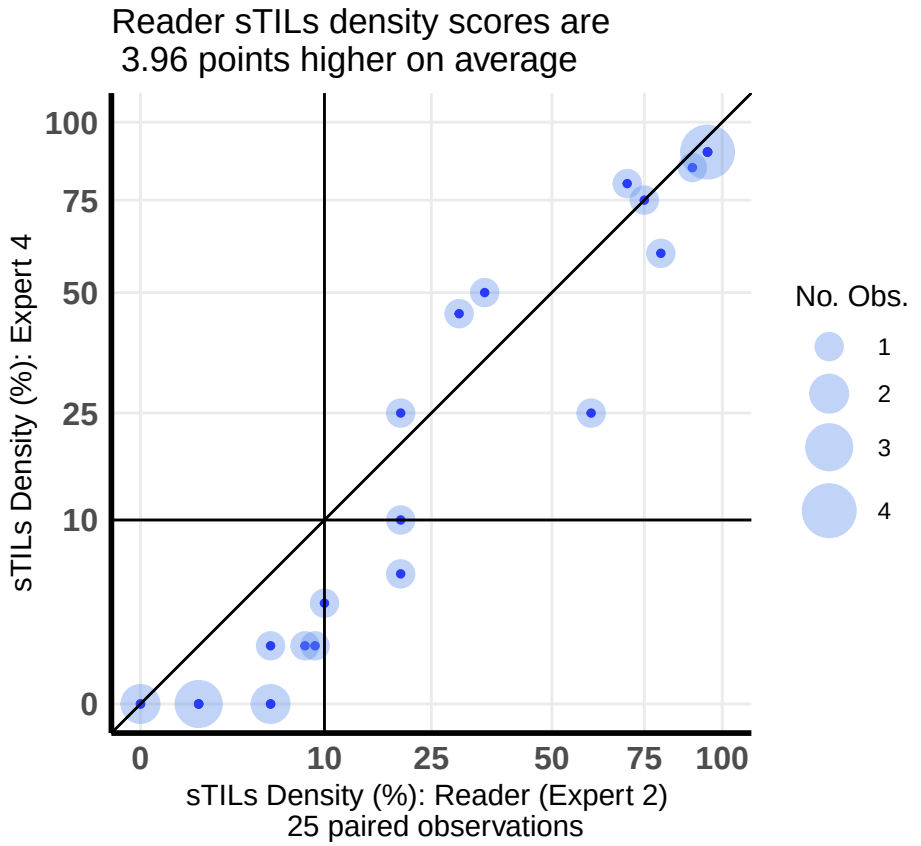


Table 2: Reader (Expert 2) and Expert 4: agreement <= 10

expert.class	reader.NotEvaluable	reader.LE	reader.GT	rateAgree
expert.GT	1	0	12	0.923
expert.LE	3	11	2	0.688
expert.NotEvaluable	7	0	0	1.000

The vertical and horizontal lines in the plot show a threshold of 10. It is possible to count the number of paired scores in each quadrant. The counts correspond to the frequencies in corresponding cells of the table. The class-specific rates of agreement are given in the "rateAgree" column. The denominator of that rate is

determined by the expert, not the reader. For example, in the first row we see that the expert scored 13 (1+0+12) cases as evaluable and greater than (GT) the threshold. The reader agreed with the expert on 12 of these, so the rate of agreement is $0.923 = 12/13$. The second row corresponds to the cases the expert scored less than or equal (LE) to the threshold (agreement $= 0.688 = 11/16$), and the last row corresponds to the cases the expert labeled as "Not Evaluable" (agreement $= 1.000 = 7/7$).

4 Agreement report for sTILs density scores <= 10

Below and in the following sections, we show your rate of agreement below and above each threshold with respect to each expert as a plot and in a table. In the plot, the labels on the x-axis indicate the expert that is acting as the reference standard. Your rates of agreement with each expert appear in the table and as black triangles in the plot (reader vs. experts). The blue circles indicate the rates of expert-expert agreement for each pair of experts. There are five of these for each expert. The horizontal line marks the lowest expert-expert rate of agreement; this is the criterion for passing this agreement metric. The criterion is given in the last column of the table, "Pass Criterion".

If your agreement falls below the criterion, you are not agreeing with the experts well on cases they score below the threshold. Please refer to the scatter plots to see this relationship.

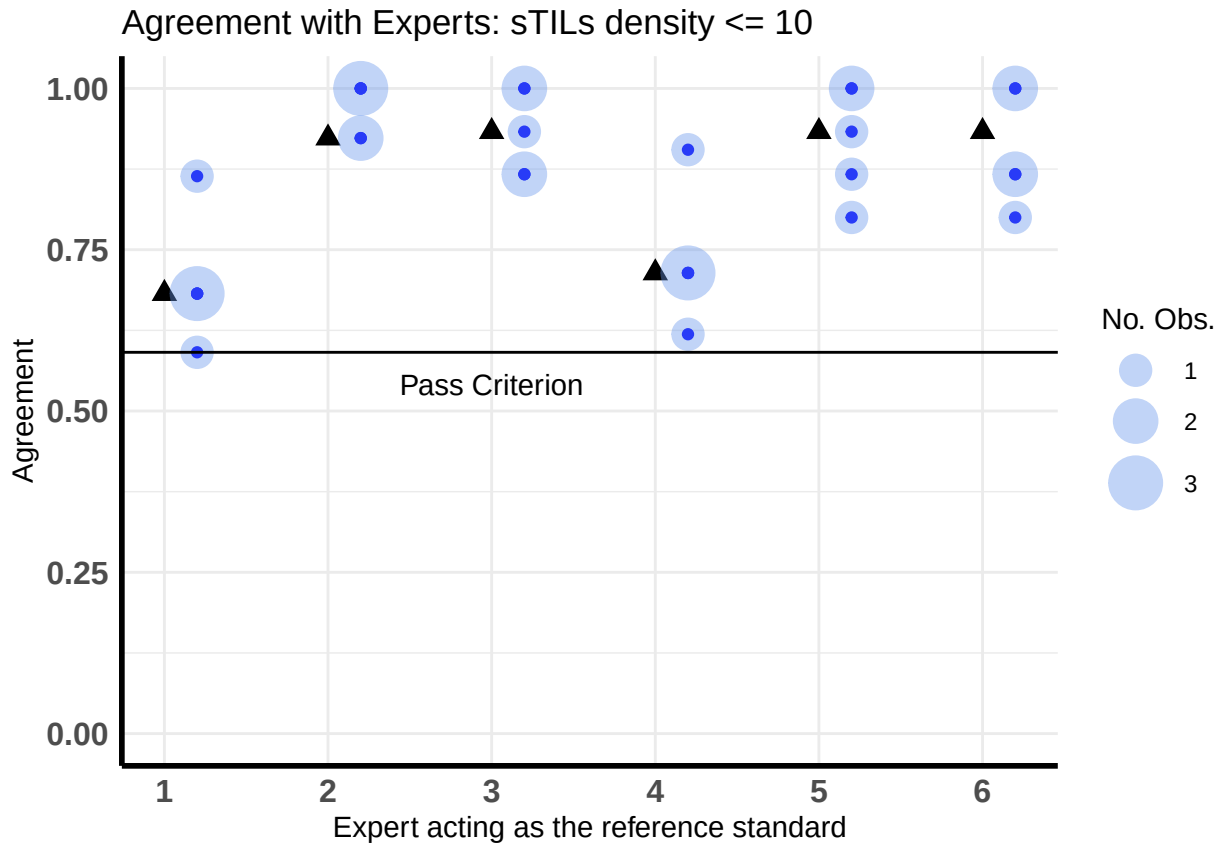


Table 3: Agreement for sTILs less than or equal (LE) to the threshold 10

expert 1	expert 2	expert 3	expert 4	expert 5	expert 6	Pass Criterion
0.682	0.923	0.933	0.714	0.933	0.933	0.591

You have passed this criterion!

5 Agreement report for sTILs density scores > 10

If your agreement falls below the criterion, you are not agreeing with the experts well on cases they score above the threshold.

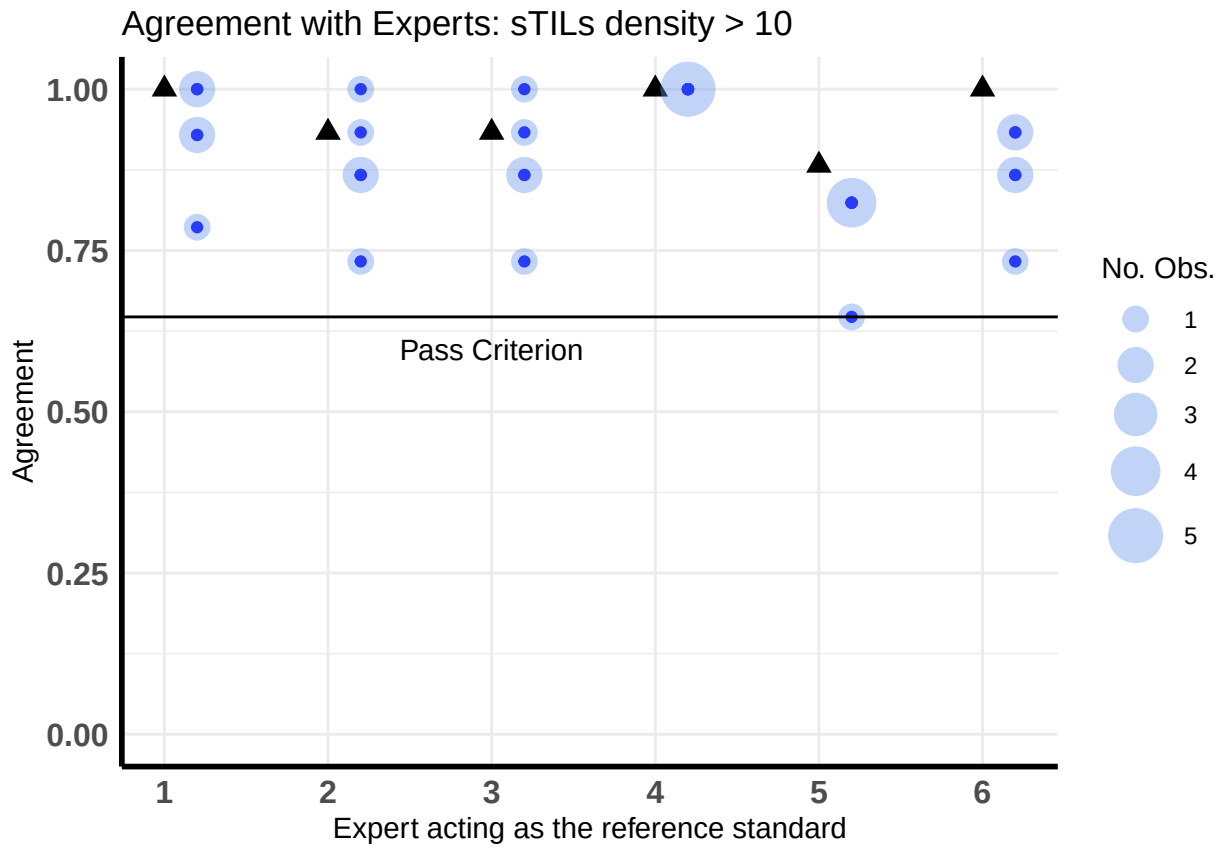


Table 4: Agreement for sTILs greater than (GT) the threshold 10

expert 1	expert 2	expert 3	expert 4	expert 5	expert 6	Pass Criterion
1.000	0.933	0.933	1.000	0.882	1.000	0.647

You have passed this criterion!

6 Agreement report for sTILs density scores <= 40

If your agreement falls below the criterion, you are not agreeing with the experts well on cases they score below the threshold.

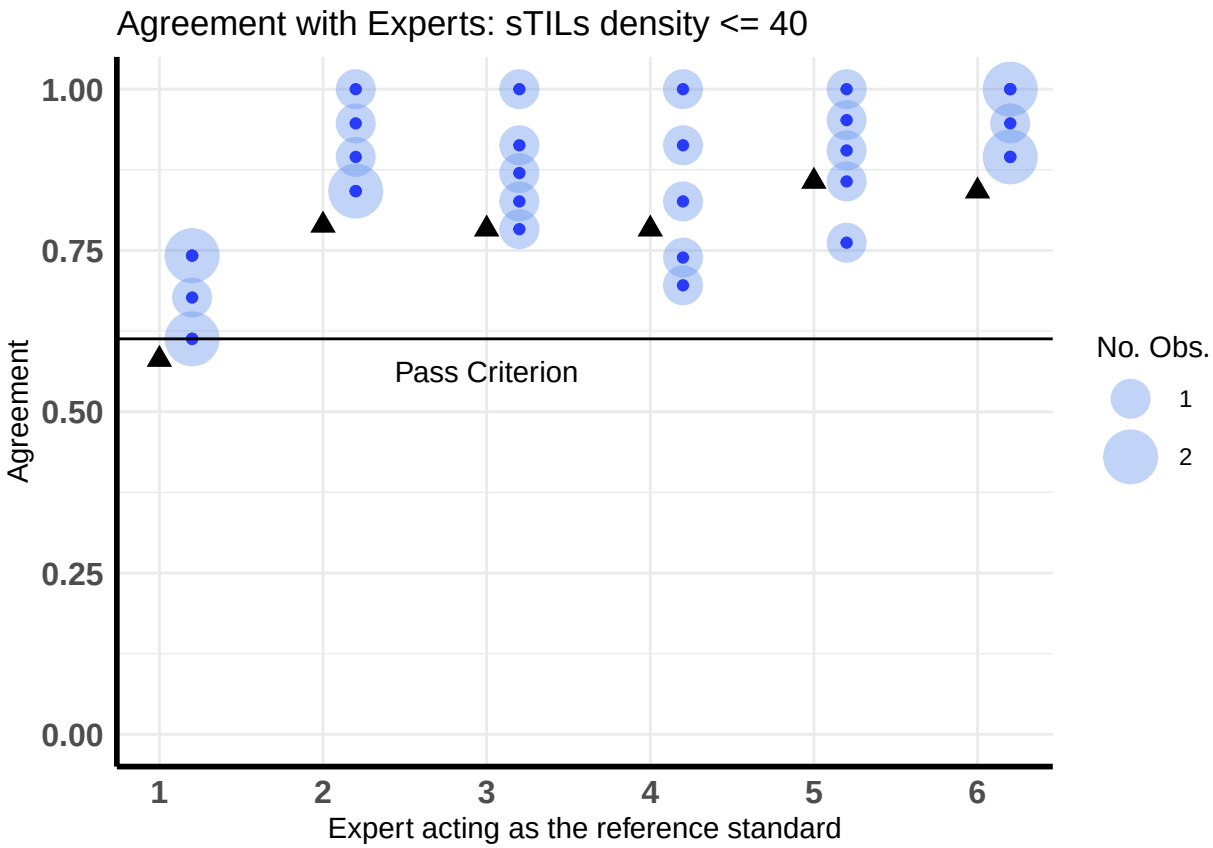


Table 5: Agreement for sTILs less than or equal (LE) to the threshold 40

expert 1	expert 2	expert 3	expert 4	expert 5	expert 6	Pass Criterion
0.581	0.789	0.783	0.783	0.857	0.842	0.613

You have not passed this criterion.

7 Agreement report for sTILs density scores > 40

If your agreement falls below the criterion, you are not agreeing with the experts well on cases they score above the threshold.

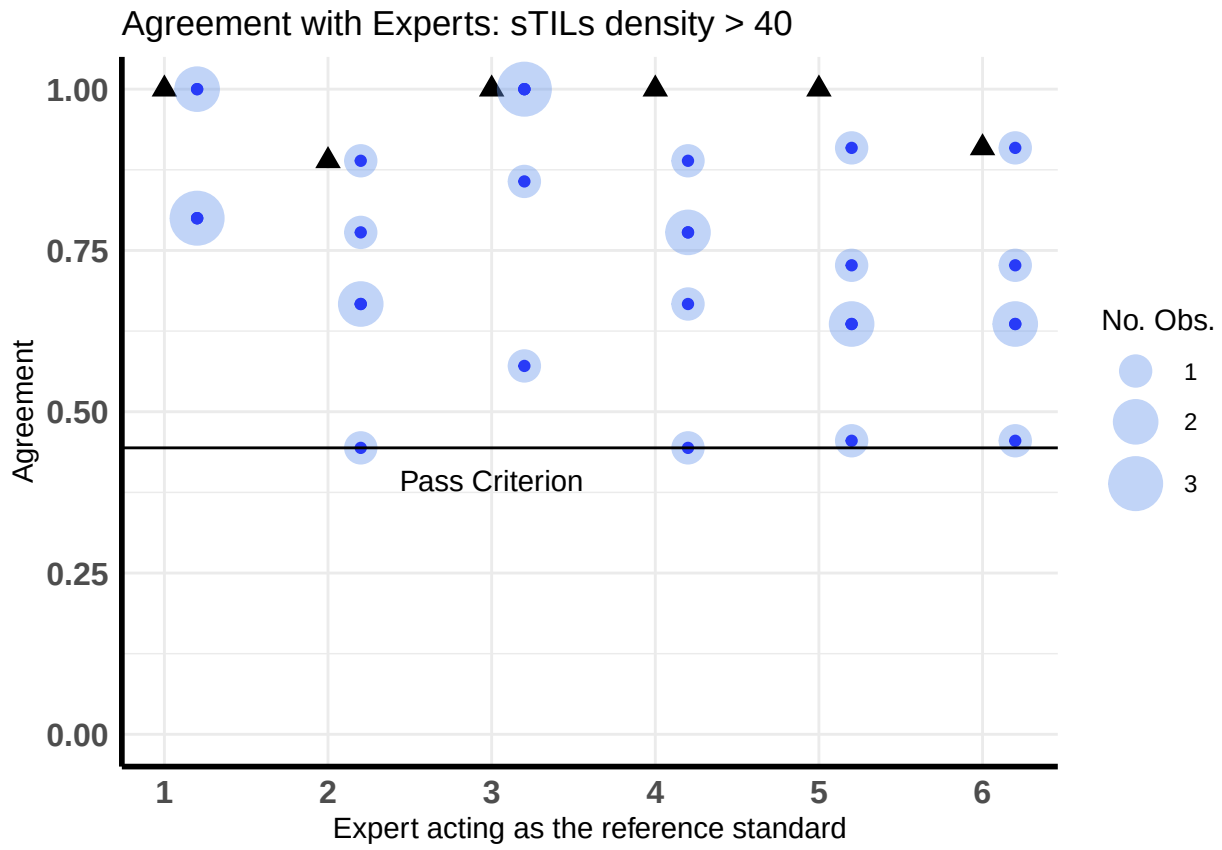


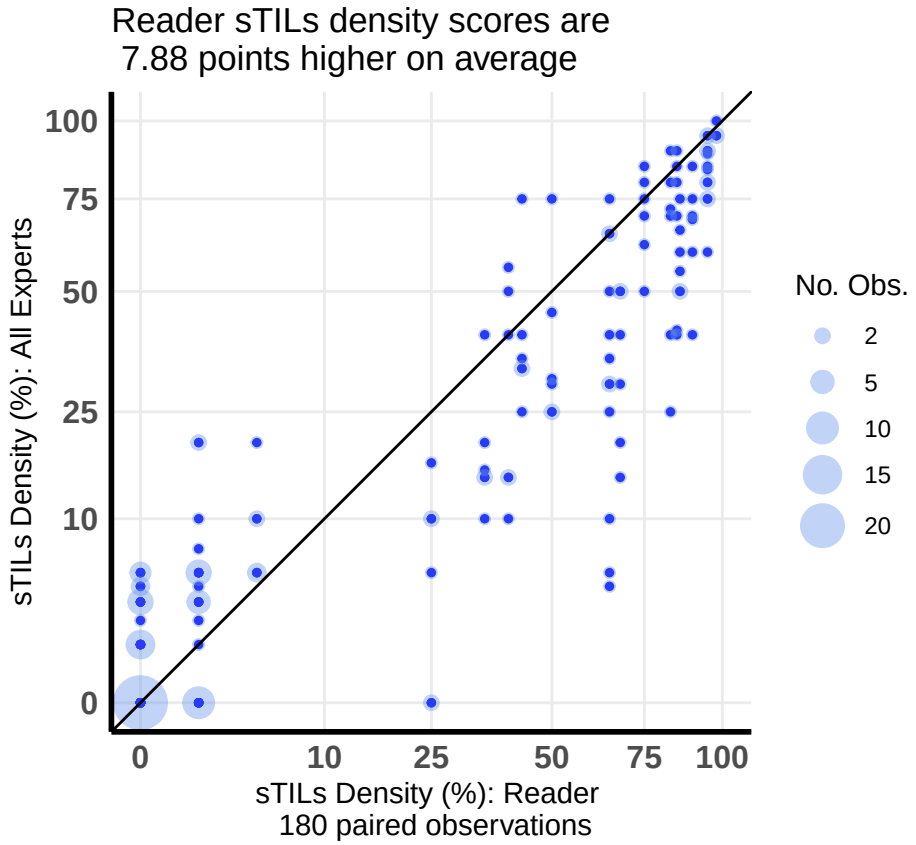
Table 6: Agreement for sTILs greater than (GT) the threshold 40

expert 1	expert 2	expert 3	expert 4	expert 5	expert 6	Pass Criterion
1.000	0.889	1.000	1.000	1.000	0.909	0.444

You have passed this criterion!

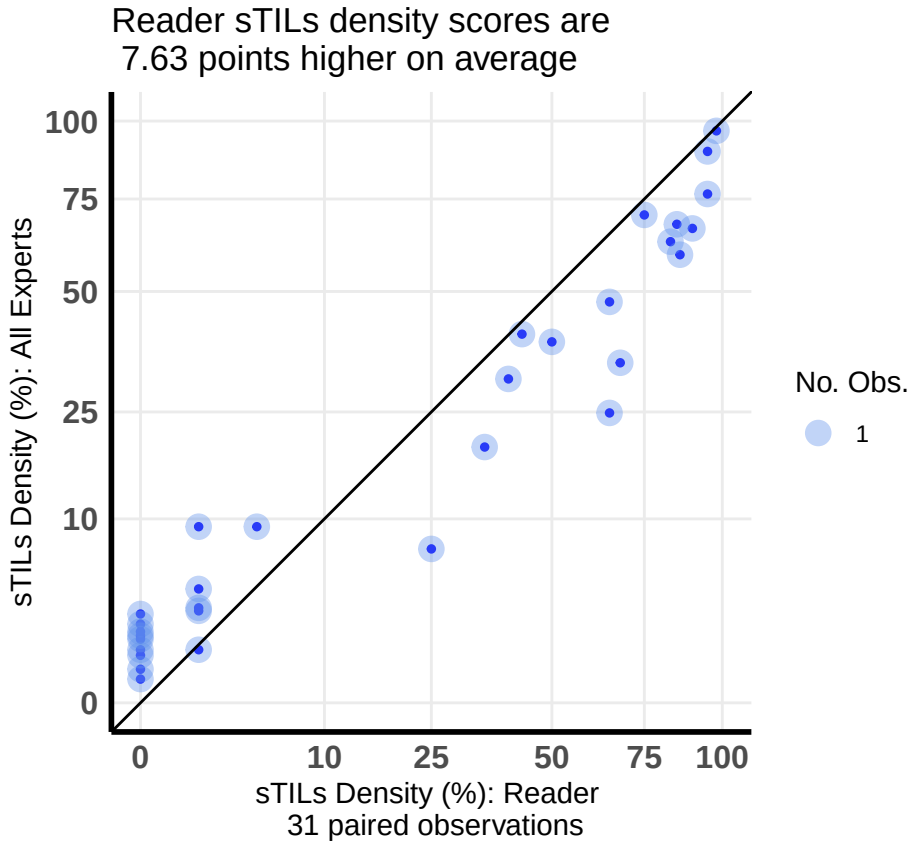
8 Scatter plot of sTILs density: reader by all experts

This plot compares your sTILs density estimates against each expert separately for each ROI. Therefore, there are six paired observations for each ROI, so that 216 (= 6 * 36) points are possible. However, ROIs labeled "Not Evaluable" by the reader or the expert do not appear in the scatter plot and reduce the number of paired observations in the plot.



9 Scatter plot of sTILs density: reader by average expert

This plot compares your sTILs density estimates against the average of the expert scores for each ROI. There are 36 points are possible. However, ROIs labeled "Not Evaluable" by the reader or by all the experts do not appear in the scatter plot and reduce the number of paired observations in the plot.



10 Agreement report for evaluable determination

The minimum expert vs. expert agreement on “Evaluable” is 0.7777778

The maximum expert vs. expert agreement on “Evaluable” is 1

The minimum reader vs. expert agreement on “Evaluable” is 0.8611111

The maximum reader vs. expert agreement on “Evaluable” is 1

We don’t summarize agreement on “Not Evaluable” because there are only a few.

Here we show the 2x2 tables for the “Evaluable” vs. “Not Evaluable” calls for the reader (you) and each expert.

Table 7: Reader vs. Expert.1

	reader.NotEvaluable	reader.Evaluable	fraction.Agree
expert.Evaluable	5	31	0.8611111
expert.NotEvaluable	0	0	NaN

Table 8: Reader vs. Expert.2

	reader.NotEvaluable	reader.Evaluable	fraction.Agree
expert.Evaluable	0	28	1.000
expert.NotEvaluable	5	3	0.625

Table 9: Reader vs. Expert.3

	reader.NotEvaluable	reader.Evaluable	fraction.Agree
expert.Evaluable	0	30	1.0000000
expert.NotEvaluable	5	1	0.8333333

Table 10: Reader vs. Expert.4

	reader.NotEvaluable	reader.Evaluable	fraction.Agree
expert.Evaluable	2	30	0.9375
expert.NotEvaluable	3	1	0.7500

Table 11: Reader vs. Expert.5

	reader.NotEvaluable	reader.Evaluable	fraction.Agree
expert.Evaluable	1	31	0.96875
expert.NotEvaluable	4	0	1.00000

Table 12: Reader vs. Expert.6

	reader.NotEvaluable	reader.Evaluable	fraction.Agree
expert.Evaluable	0	30	1.0000000
expert.NotEvaluable	5	1	0.8333333

11 References

[1] S. Dudgeon *et al.*, "A pathologist-annotated dataset for validating artificial intelligence: A project description and pilot study," *J Pathol Inform*, vol. 12, no. 1, p. 45, 2021, doi: 10.4103/jpi.jpi_83_20.

[2] V. Garcia *et al.*, "Development of Training Materials for Pathologists to Provide Machine Learning Validation Data of Tumor-Infiltrating Lymphocytes in Breast Cancer," *Cancers*, vol. 14, no. 10, p. 2467, May 2022, doi: 10.3390/cancers14102467.