Quantitative Agreement Analysis for HTT Pilot Study Data

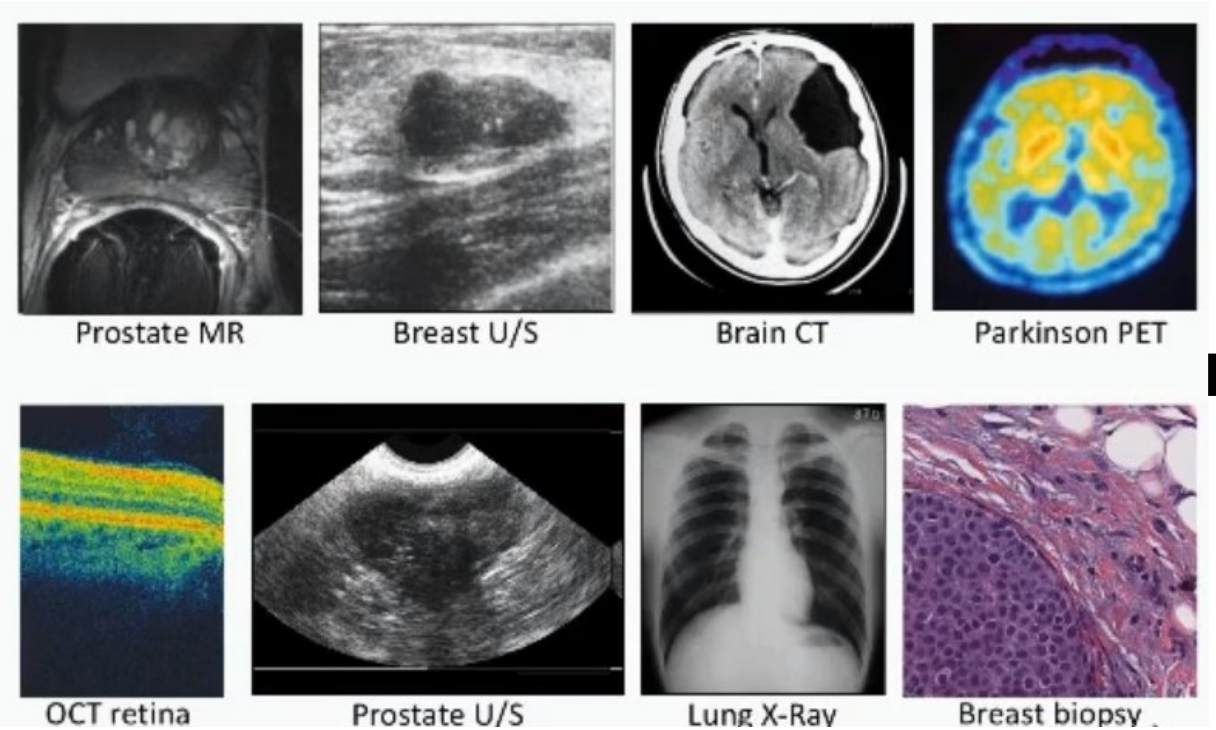
Si Wen

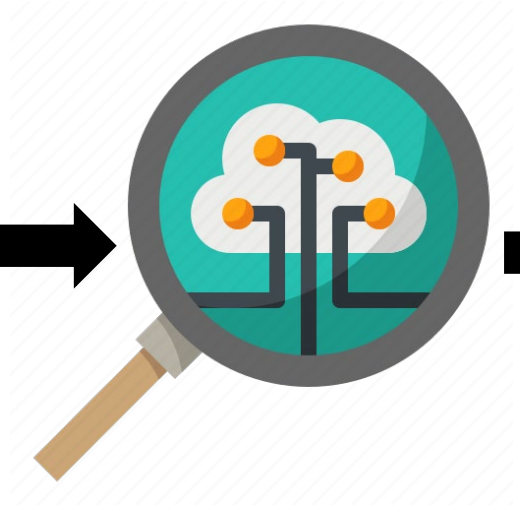
07/16

Outline

- Background
- Limits of Agreement with Ground Truth Value
- Limits of Agreement with Reference Values from Multiple Readers
  - Apply to HTT pilot study data – work in progress
- Between-Reader Agreement
  - Apply to HTT pilot study data – work in progress
- Future Work

Background





Assist Clinical Decision

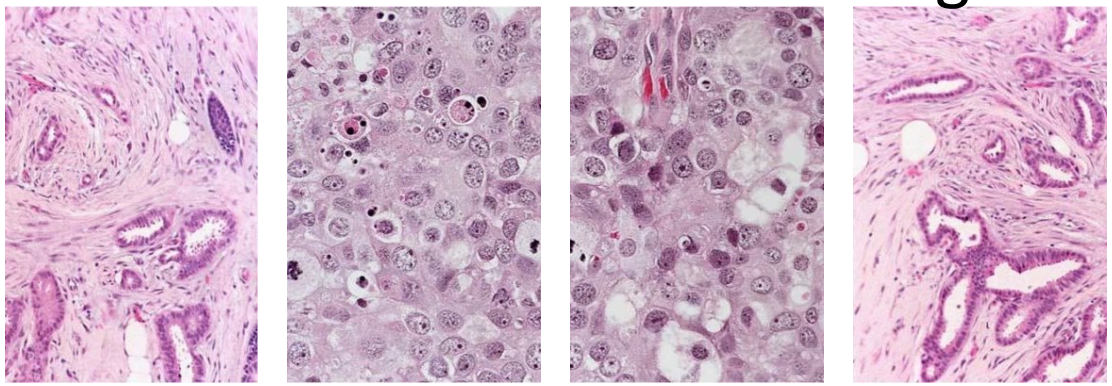
Medical Imaging

AI/ML Algorithm

VALIDATION

Background – Qualitative Assessment

Breast Cancer Grading



https://pathology.jhu.edu/breast/staging-grade/

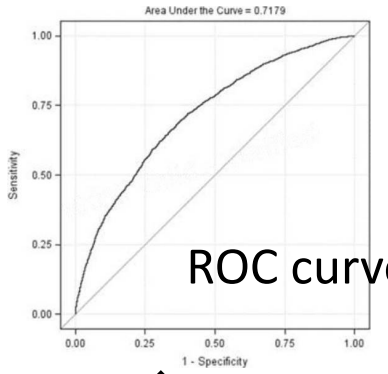
Algorithm Output

Low	High	Low	High
0.2	0.8	0.4	0.6

Ground Truth Label

Low	High	High	Low

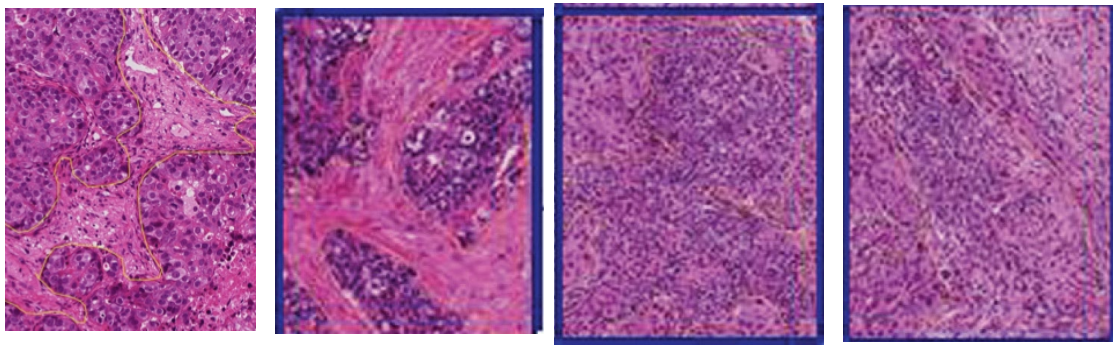
Confusion Matrix	Low	High
Low	1	1
High	1	1



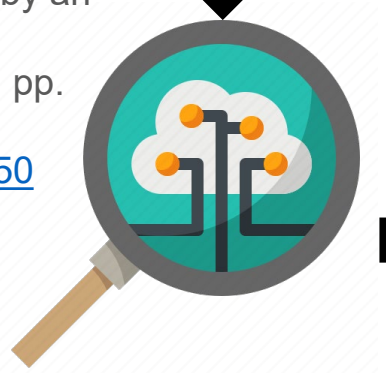
ROC curve

Background – Quantitative Measurement

Stromal Tumor Infiltrating Lymphocytes



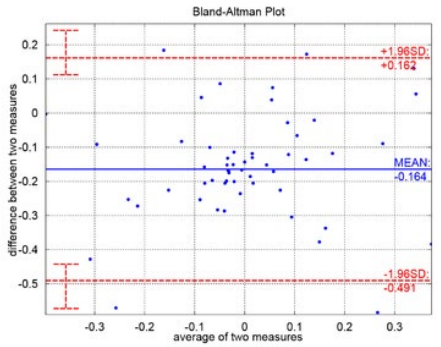
R. Salgado *et al.*, “The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014,” *Ann. Oncol.*, vol. 26, no. 2, pp. 259–271, Feb. 2015, doi: 10.1093/annonc/mdu450



Algorithm Output

Suppose we have ground truth value

10	1	90	70



Limits of Agreement Analysis

15	5	80	65

Limits of Agreement

- Suppose $\{X_k\}$ and $\{Y_k\}$ are test scores and ground truth values based on a group of subjects/cases $(k = 1, \dots, K)$

- Let d_k to denote the difference between scores on the same case

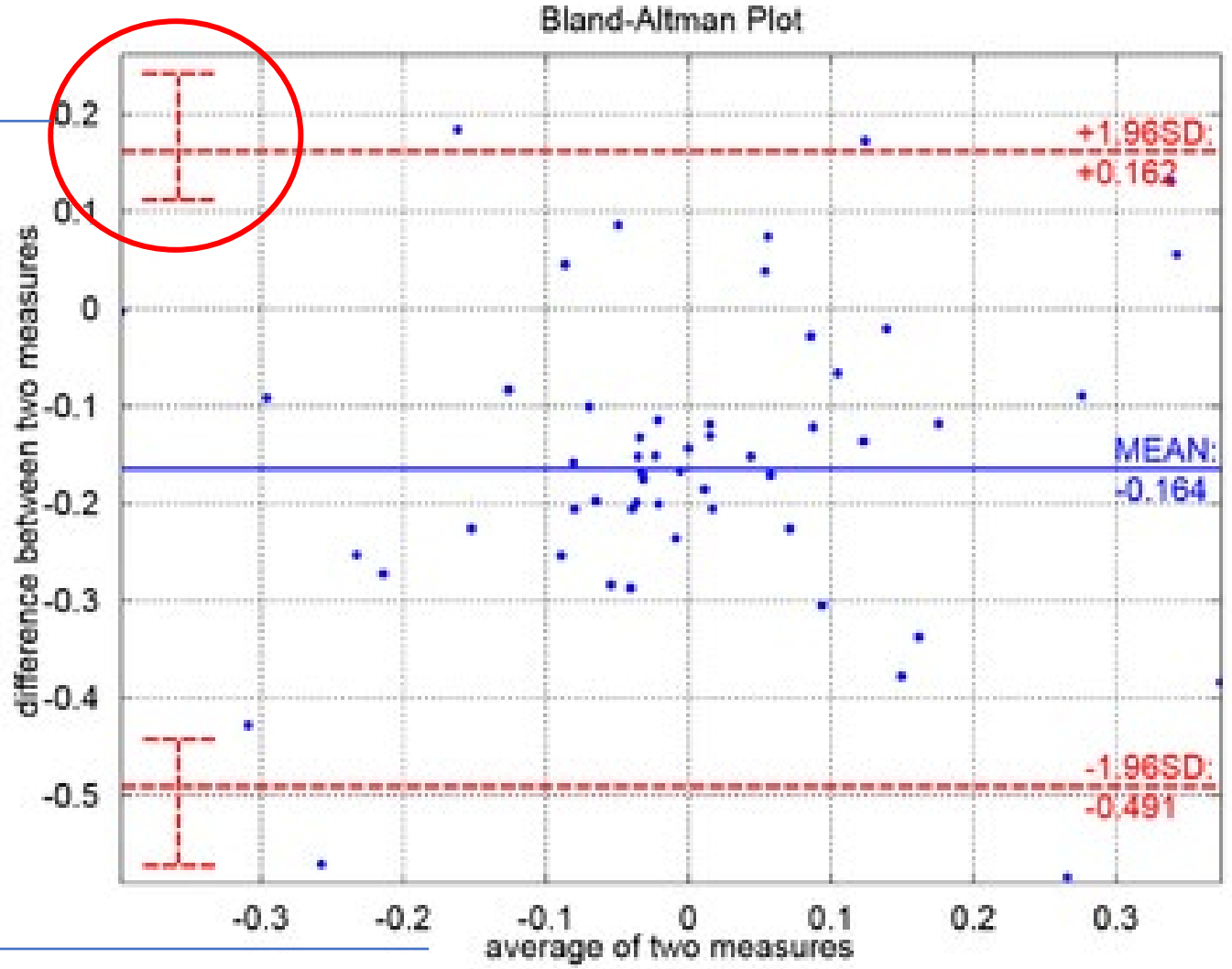
$$d_k = X_k - Y_k$$

- The mean difference : $\bar{d} = \frac{1}{K}\sum_{k=1}^{K} d_k$
- The standard deviation of the differences: $s_d = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(d_k - \bar{d})^2}$

- The 95% **limits of agreement**: $\bar{d} \pm 1.96 s_d$

Bland-Altman Plot

Confidence Interval for the limits of agreement



$\bar{d} + 1.96 s_d$

$d_k = X_k - Y_k$

$\bar{d}$

$\bar{d} - 1.96 s_d$

$(X_k + Y_k)/2$

https://en.wikipedia.org/wiki/Bland%E2%80%93Altman_plot

Limits of Agreement with Ground Truth Value

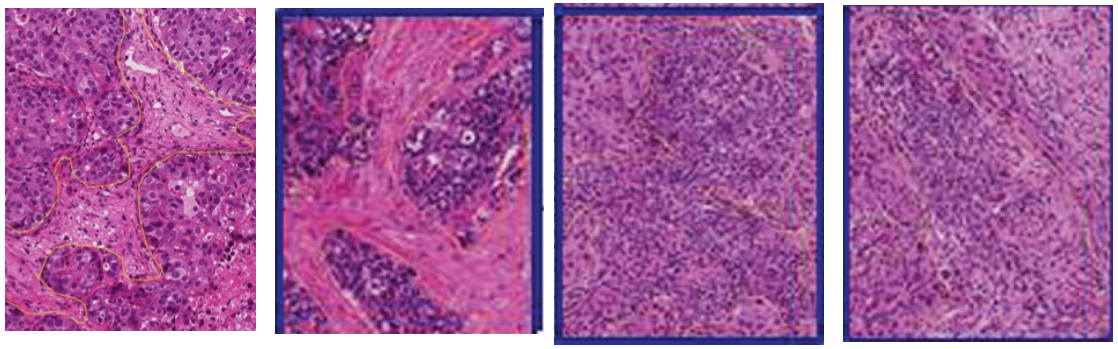
- The 95% **limits of agreement**: $\bar{d} \pm 1.96 s_d$

- Define the range within which most differences between algorithm result and ground truth value will lie

- The decision about what is acceptable agreement is a clinical one; statistics alone cannot answer the question

Quantitative Analysis- No Ground Truth Value

Stromal Tumor Infiltrating Lymphocytes



R. Salgado *et al.*, “The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014,” *Ann. Oncol.*, vol. 26, no. 2, pp. 259–271, Feb. 2015, doi: 10.1093/annonc/mdu450

Reference values from multiple readers

10	1	90	70
9	5	80	65
12	2	70	80
8	1	85	60

Algorithm Output

15	5	80	65

Quantitative Analysis- No Ground Truth Value

- Naïve Way

10	1	90	70
9	5	80	65
12	2	70	80
8	1	85	60

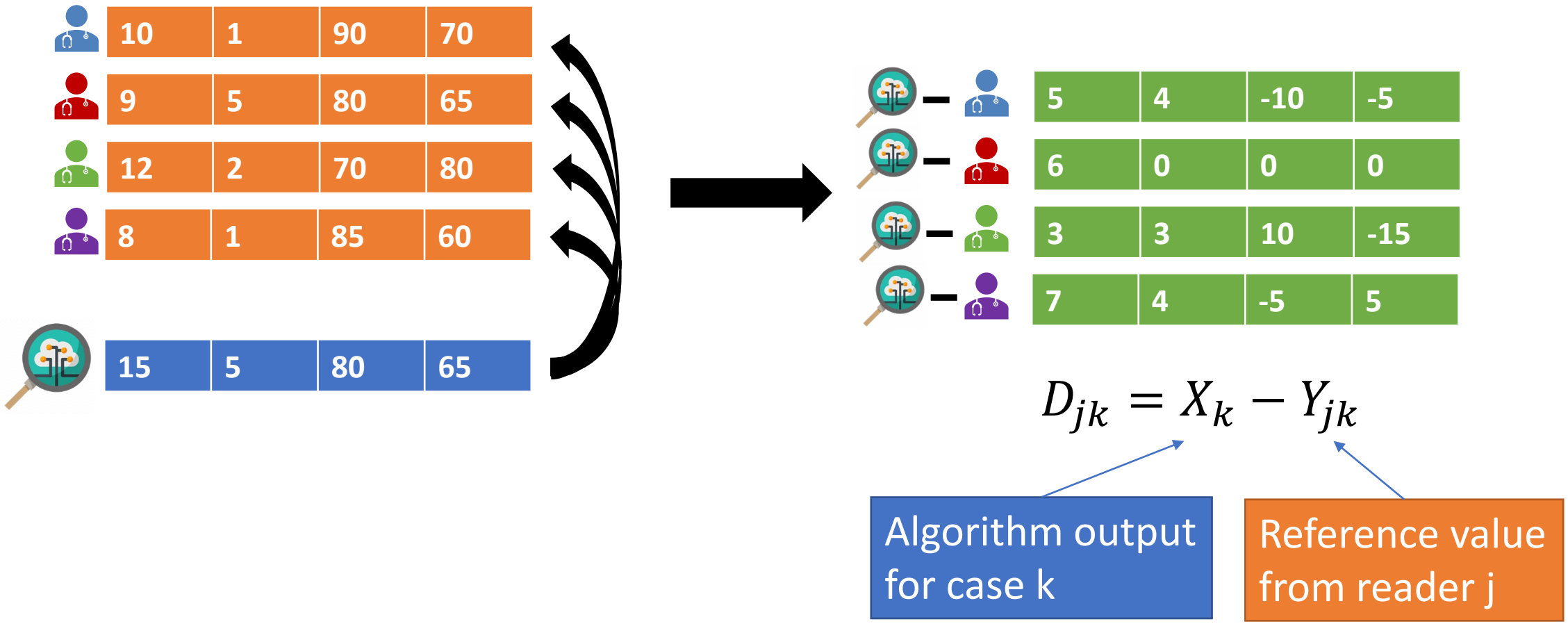
Reader averaged reference value

9.75	2.25	81.25	68.75

15	5	80	65

Ignore inter-reader variability

Quantitative Analysis- No Ground Truth Value



Quantitative Analysis- No Ground Truth Value

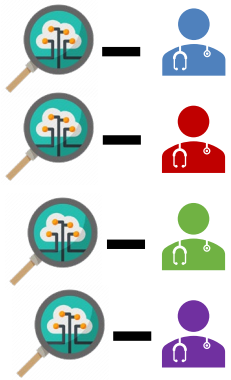
- Limits of Agreement
  - The mean difference :

$$\bar{D} = \frac{1}{JK}\sum_j \sum_k D_{jk}$$

  - The variance of the differences:

$$Var(D_{jk}) \neq \frac{1}{JK-1}\sum_j \sum_k (D_{jk} - \bar{D})^2$$

  - not independent



5	4	-10	-5
6	0	0	0
3	3	10	-15
7	4	-5	5

$$D_{jk} = X_k - Y_{jk}$$

Limits of Agreement with Reference Values from Multiple Readers

- Two-way random effect ANOVA model for the difference D_{jk}

$$D_{jk} = \mu + R_j + C_k + \varepsilon_{jk}$$

  - $R_j \sim N(0, \sigma_{dR}^2)$, $C_k \sim N(0, \sigma_{dC}^2)$, $\varepsilon_{jk} \sim N(0, \sigma_{d\varepsilon}^2)$

- The variance of D_{jk}:

$$Var(D_{jk}) = \sigma_{dR}^2 + \sigma_{dC}^2 + \sigma_{d\varepsilon}^2$$

Limits of Agreement with Reference Values from Multiple Readers

- Two-way ANOVA table

Source	DF	Sum of Square (SS)	Mean Square (MS)	E(MS)
Reader	$J-1$	$SSR = K\sum_j\left(\overline{D_{J\cdot}} - \overline{D}\right)^2$	$MSR = SSR/(J-1)$	$\sigma_{d\varepsilon}^2 + K\sigma_{dR}^2$
Case	$K-1$	$SSC = J\sum_k(\overline{D_{\cdot k}} - \overline{D})^2$	$MSC = SSC/(K-1)$	$\sigma_{d\varepsilon}^2 + J\sigma_{dC}^2$
Error	$(J-1)(K-1)$	$SSE = SST - SSR - SSC$	$MSE = SSE/(J-1)(K-1)$	$\sigma_{d\varepsilon}^2$
Total	$JK-1$	$SST = \sum_j\sum_k\left(D_{jk} - \overline{D}\right)^2$		

Limits of Agreement with Reference Values from Multiple Readers

- Variance components estimation:

$$\hat{\sigma}^2_{d\varepsilon} = MSE, \qquad \hat{\sigma}^2_{dR} = \frac{MSR - MSE}{K}, \qquad \hat{\sigma}^2_{dC} = \frac{MSC - MSE}{J}$$

- Estimated variance of difference :

$$\widehat{Var}(D_{jk}) = \hat{\sigma}^2_{dR} + \hat{\sigma}^2_{dC} + \hat{\sigma}^2_{d\varepsilon}$$
$$= \frac{1}{JK}(J * MSR + K * MSC + (JK - J - K) * MSE)$$

- The 95% **limits of agreement**: $\overline{D} \pm 1.96\sqrt{\hat{\sigma}^2_{dR} + \hat{\sigma}^2_{dC} + \hat{\sigma}^2_{d\varepsilon}}$

Limits of Agreement with Reference Values from Multiple Readers

- Naïve Way – Reader-averaged Reference Value

$$Z_k = X_k - \frac{1}{J}\sum_j Y_{jk} = \frac{1}{J}\sum_j D_{jk} = D_{\cdot k}$$

  - The mean difference :

$$\bar{Z} = \frac{1}{K}\sum_k Z_k = \frac{1}{JK}\sum_j \sum_k D_{jk} = \bar{D}$$

  - The variance of the differences:

$$\frac{1}{K-1}\sum_k (Z_k - \bar{Z})^2 = \frac{1}{K-1}\sum_k (D_{\cdot k} - \bar{D})^2 = \frac{1}{J} MSC = \hat{\sigma}_{dC}^2 + \frac{1}{J}\hat{\sigma}_{d\varepsilon}^2$$

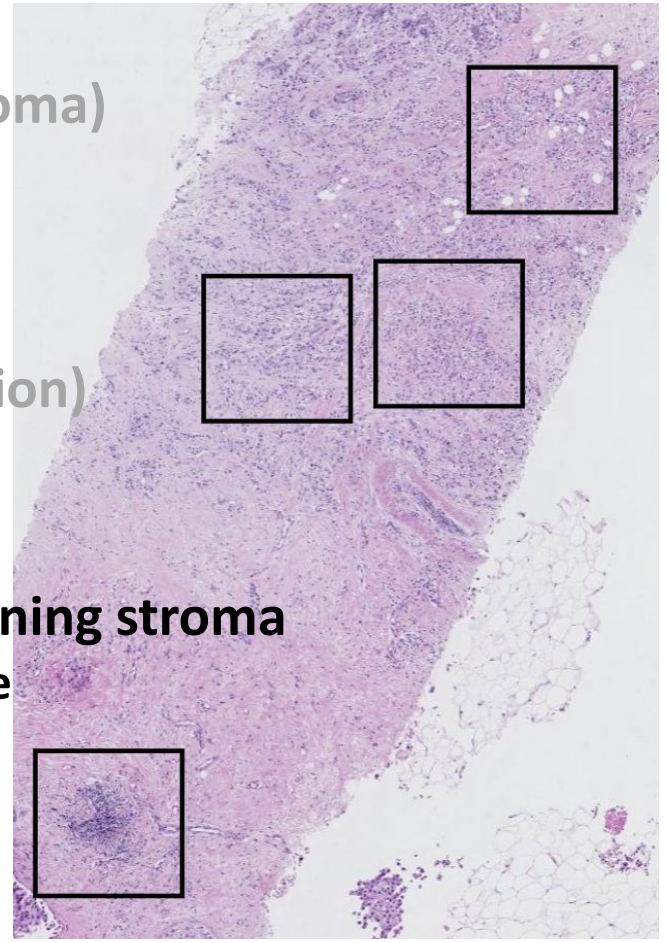
  - 95% limits of agreement for Z_k : $\bar{D} \pm 1.96\sqrt{\hat{\sigma}_{dC}^2 + \frac{1}{J}\hat{\sigma}_{d\varepsilon}^2}$

HTT Pilot Study Data

Data Collection

- Cases:
  - 64 H&E Slides
  - 10 ROIs per Slide
  - Some ROIs are not appropriate for sTIL evaluation
- Evaluation Platforms:
  - caMicroscope & PathPresenter
- Readers (finish all the ROIs):
  - 5 readers using caMicroscope
  - 2 readers using PathPresenter

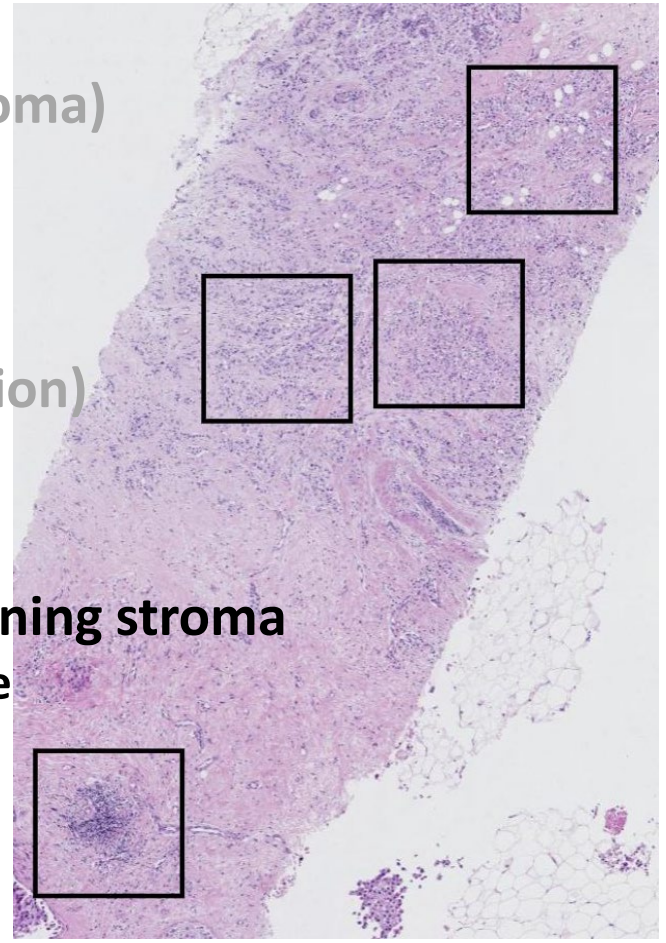
- **Intra-tumoral stroma**
  (Tumor-associated stroma)
  Select ~3 ROIs
- **Invasive margin**
  (Tumor-stroma transition)
  Select ~2 ROIs
- **Tumor with no intervening stroma**
  Select ~2 ROIs, if possible
- **Other regions**
  Select ~3-4 ROIs



HTT Pilot Study Data

Comparison between Two Readers	Intra-tumoral stroma	Invasive Margin	Other Regions	Tumor with no intervening stroma
Intra-tumoral stroma	447 (69.8%)	29	4	4
Invasive Margin	47	10	3	1
Other Regions	11	3	77	0
Tumor with no intervening stroma	1	0	0	3

- **Intra-tumoral stroma**
  (Tumor-associated stroma)
  Select ~3 ROIs
- **Invasive margin**
  (Tumor-stroma transition)
  Select ~2 ROIs
- **Tumor with no intervening stroma**
  Select ~2 ROIs, if possible
- **Other regions**
  Select ~3-4 ROIs



HTT Pilot Study Data

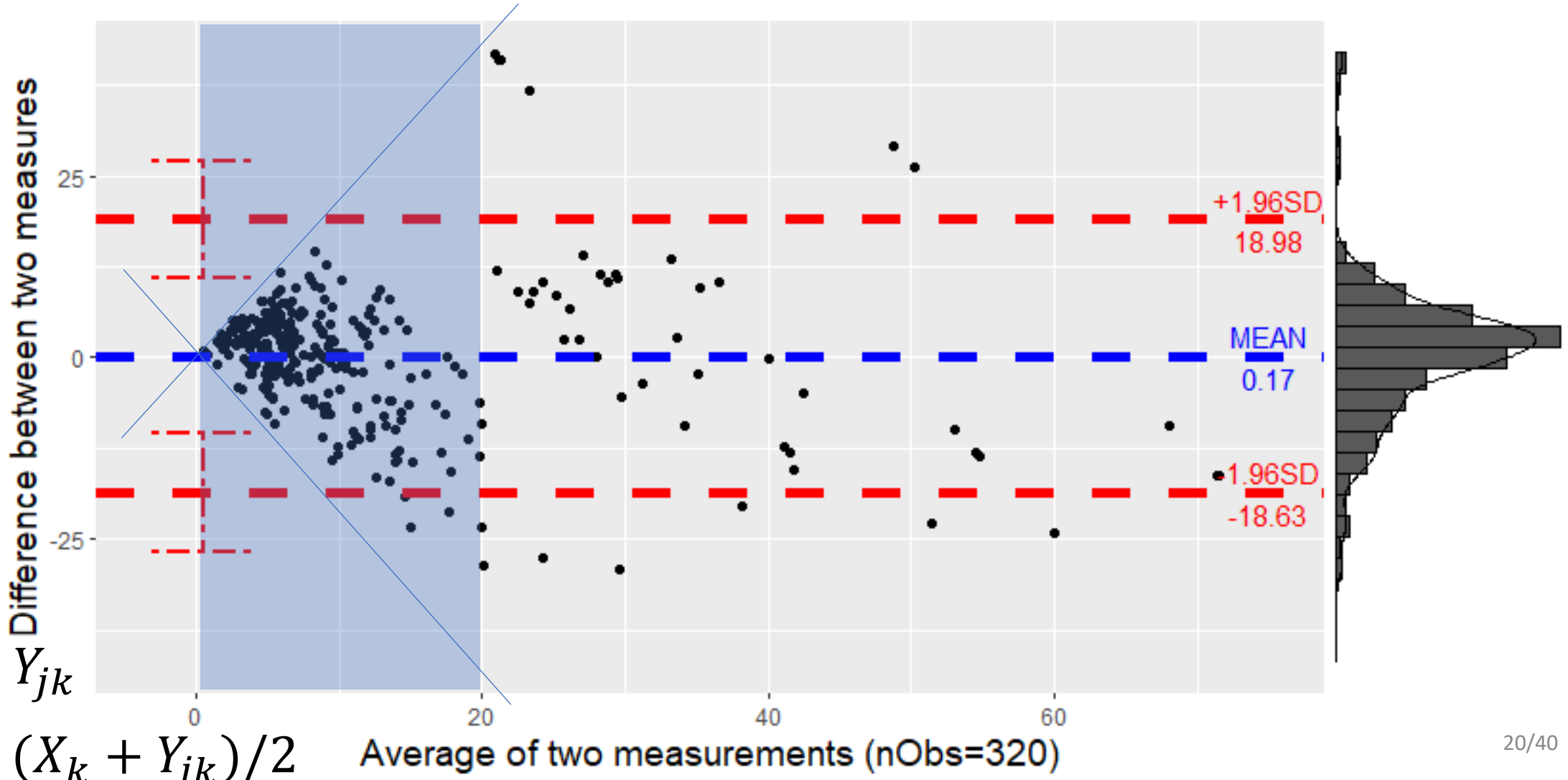
Data Collection

- Cases:
  - 64 H&E Slides
  - 10 ROIs per Slide
  - Some ROIs are not appropriate for sTIL evaluation
- Evaluation Platforms:
  - caMicroscope & PathPresenter
- Readers (finish all the ROIs):
  - 5 readers using caMicroscope
  - 2 readers using PathPresenter

Data Preparation

- Average Scores across ROIs for each Slide and Reader Combination
  - Remove the correlation among the ROIs within a slide
  - Future work: not just average over ROIs
- Algorithm Output vs Reference Values
  - Algorithm – 1 reader using PathPresenter
  - Reference Value – 5 readers using caMicroscope

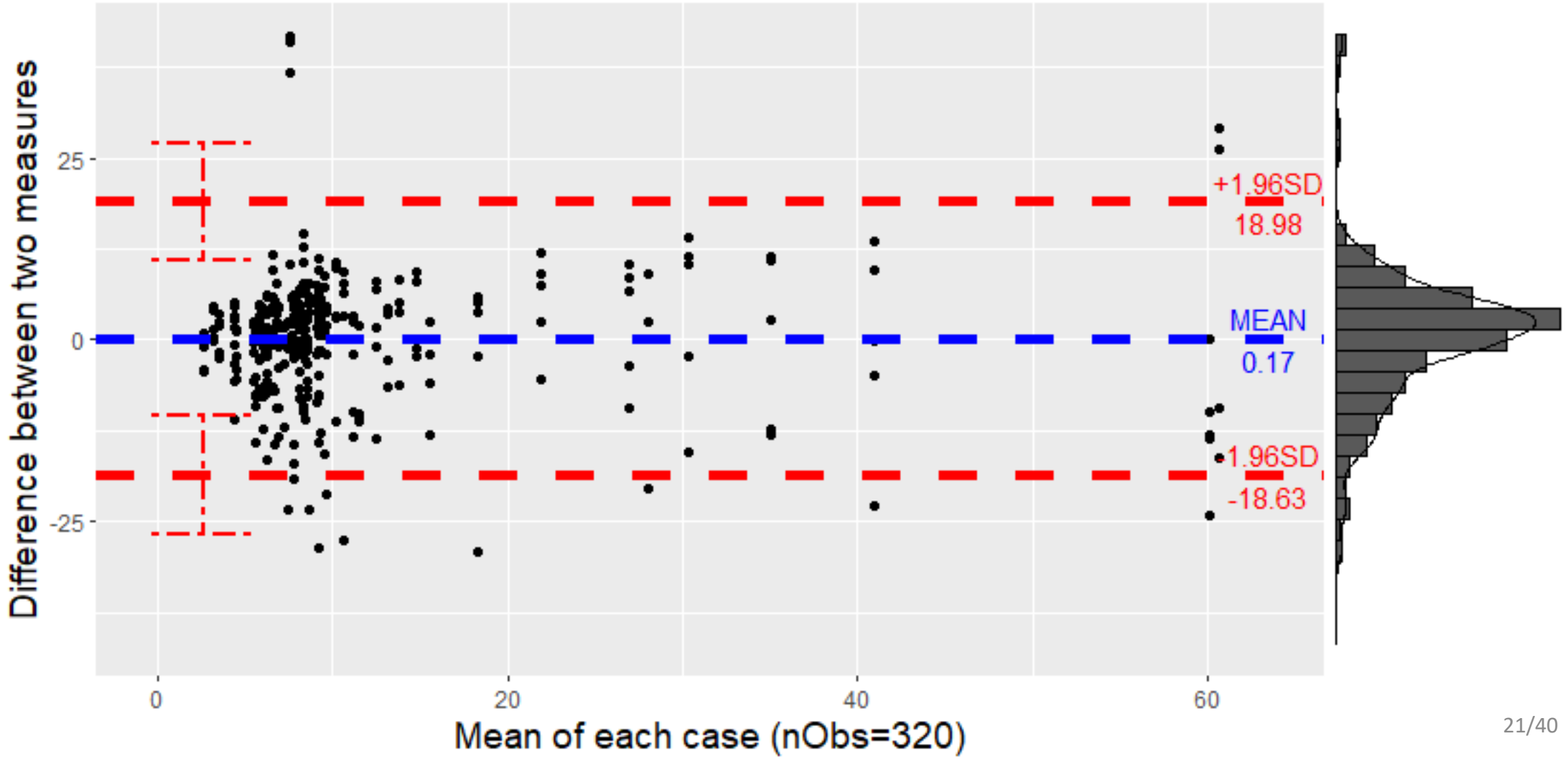
Algorithm Output Vs Reference Values



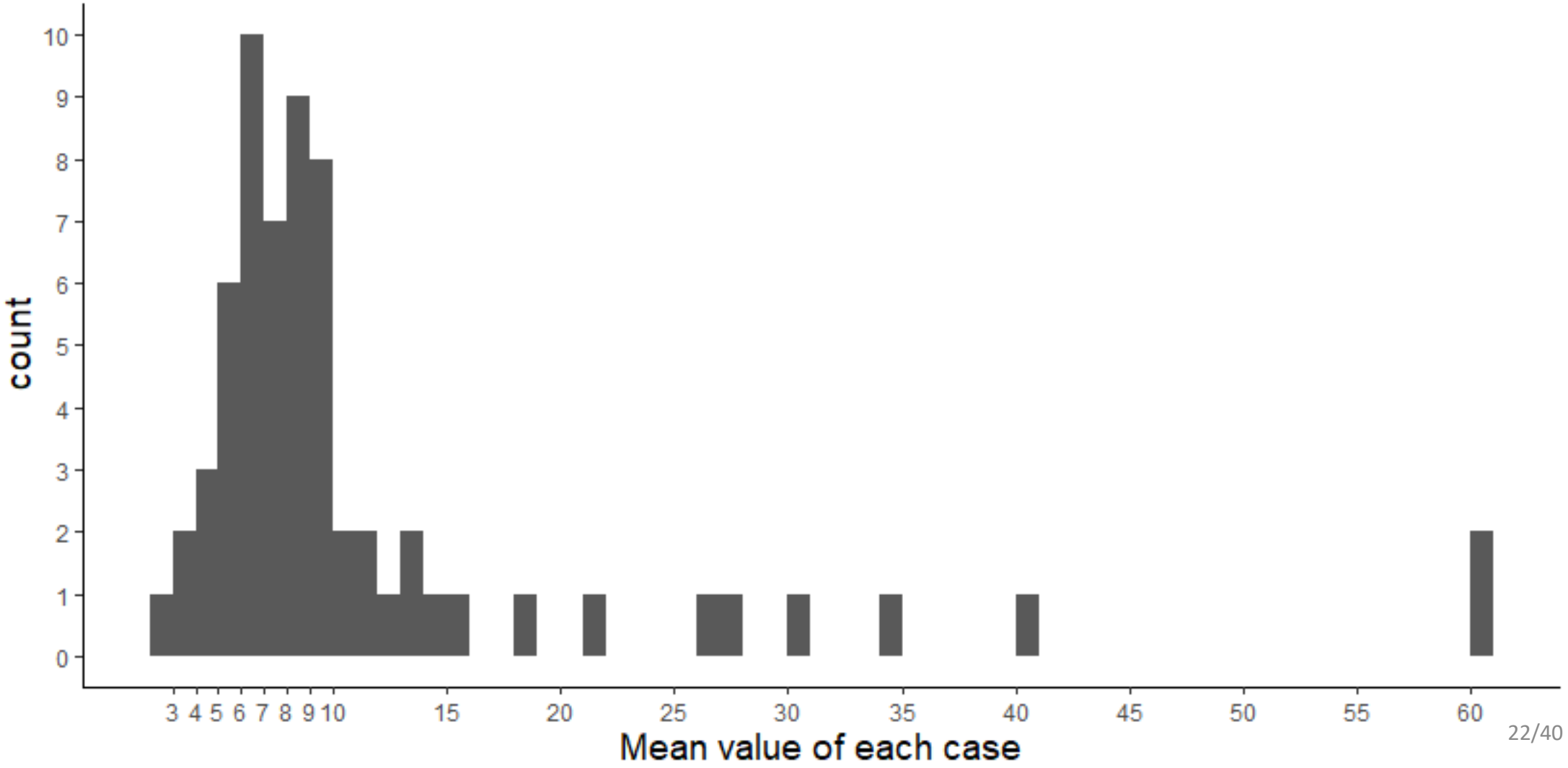
$X_k - Y_{jk}$

$(X_k + Y_{jk})/2$

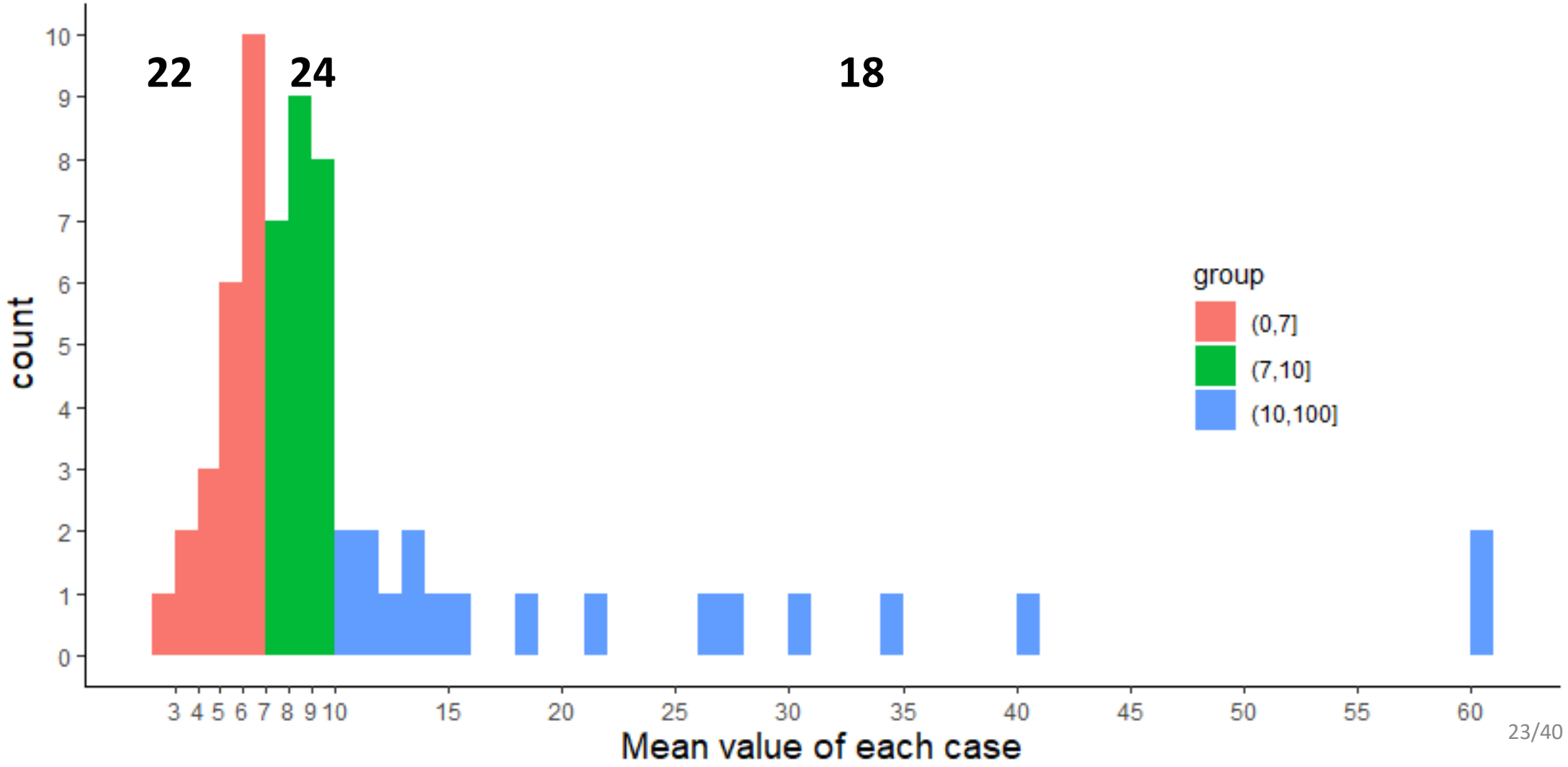
Algorithm Output Vs Reference Values



Algorithm Output Vs Reference Values

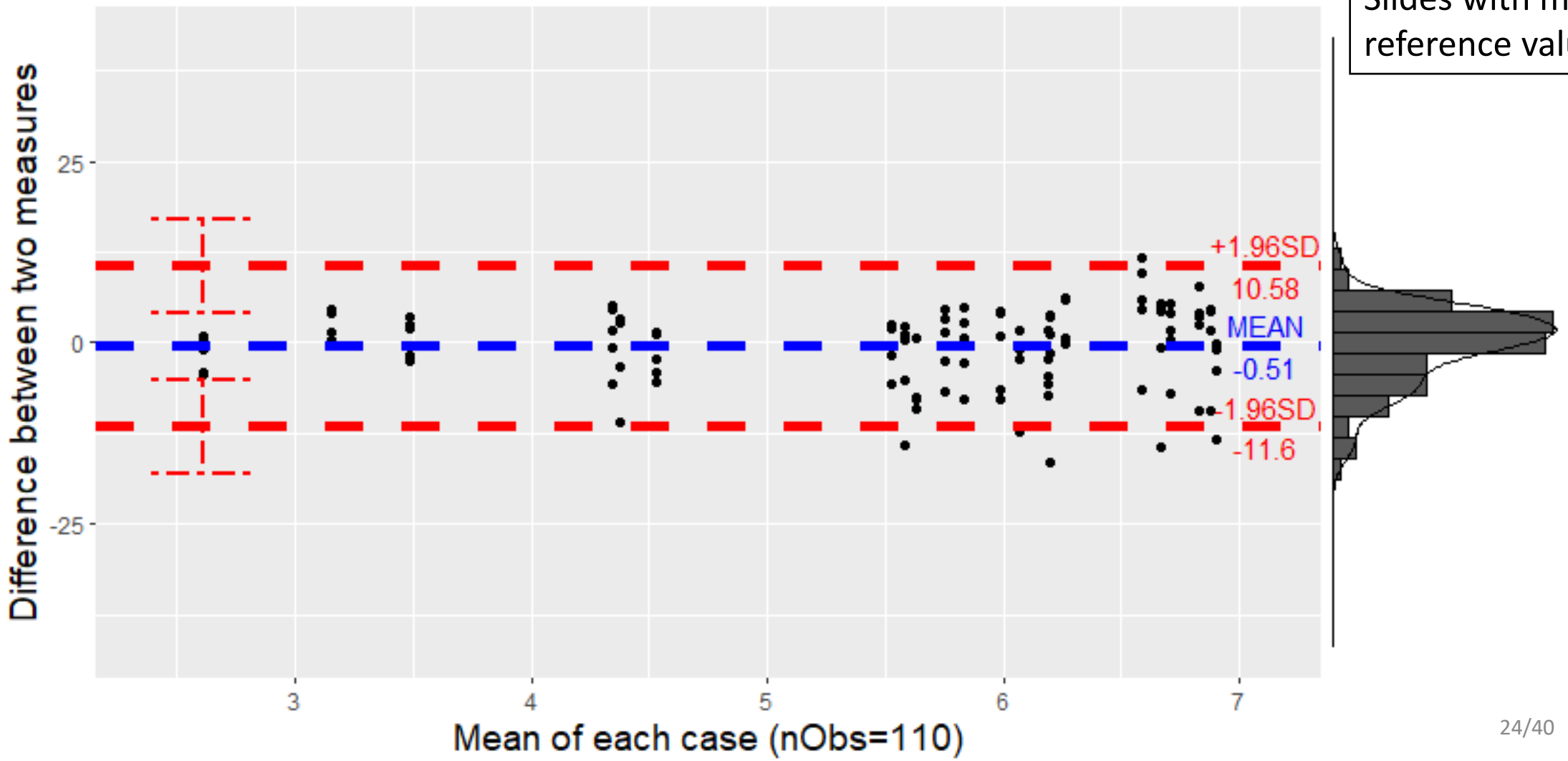


Algorithm Output Vs Reference Values

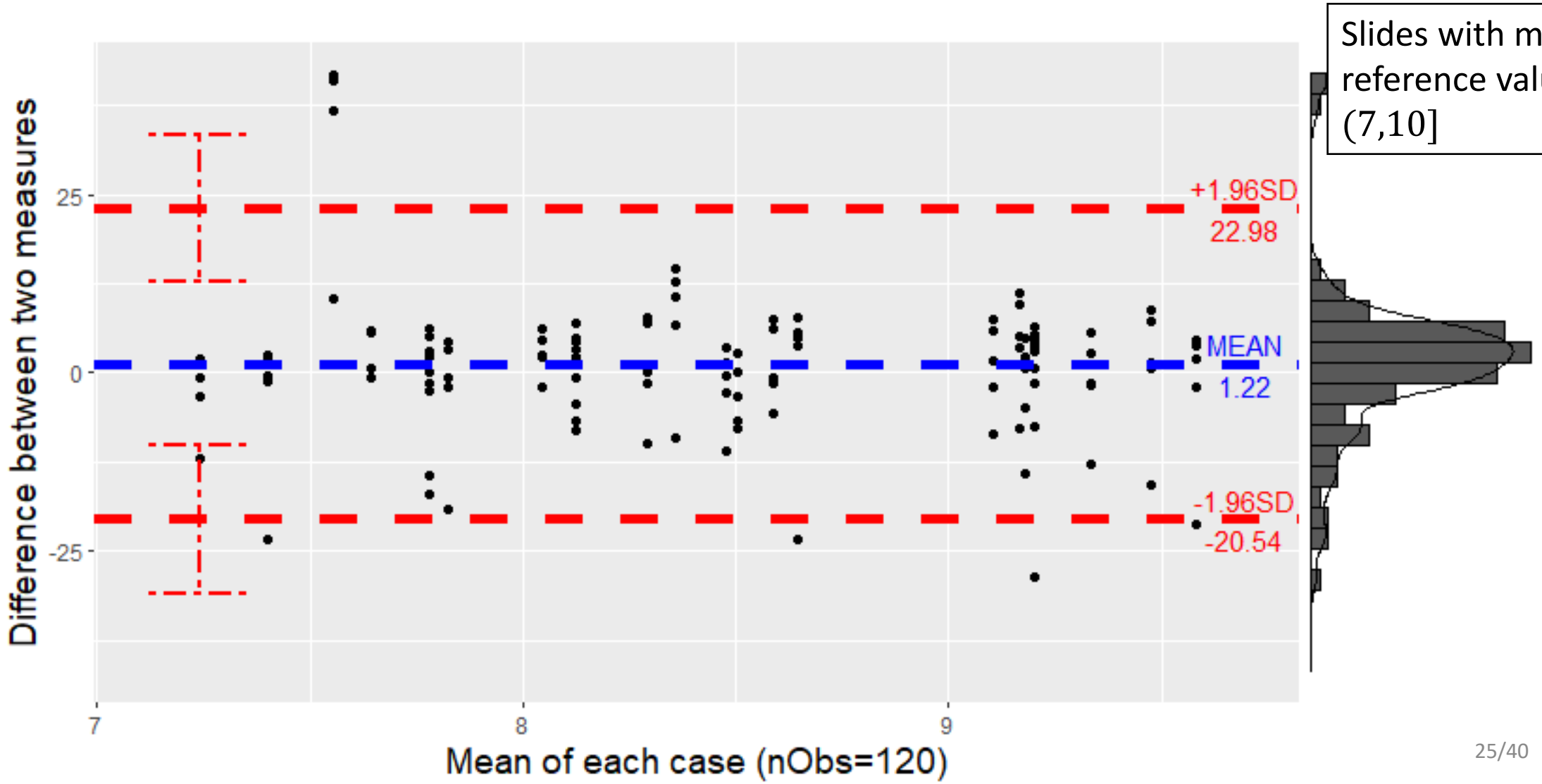


Algorithm Output Vs Reference Values

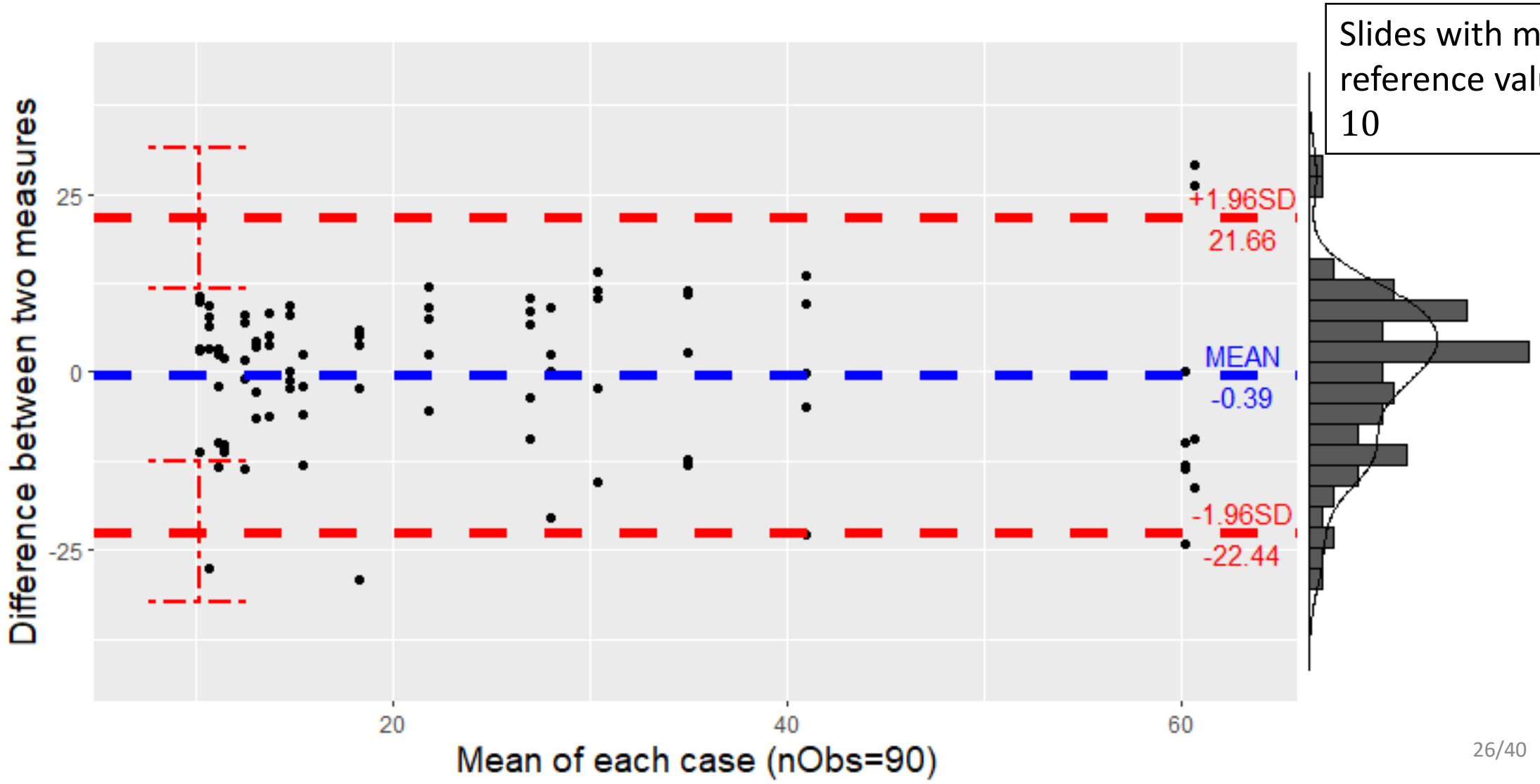
Slides with mean reference values ≤ 7



Algorithm Output Vs Reference Values



Algorithm Output Vs Reference Values



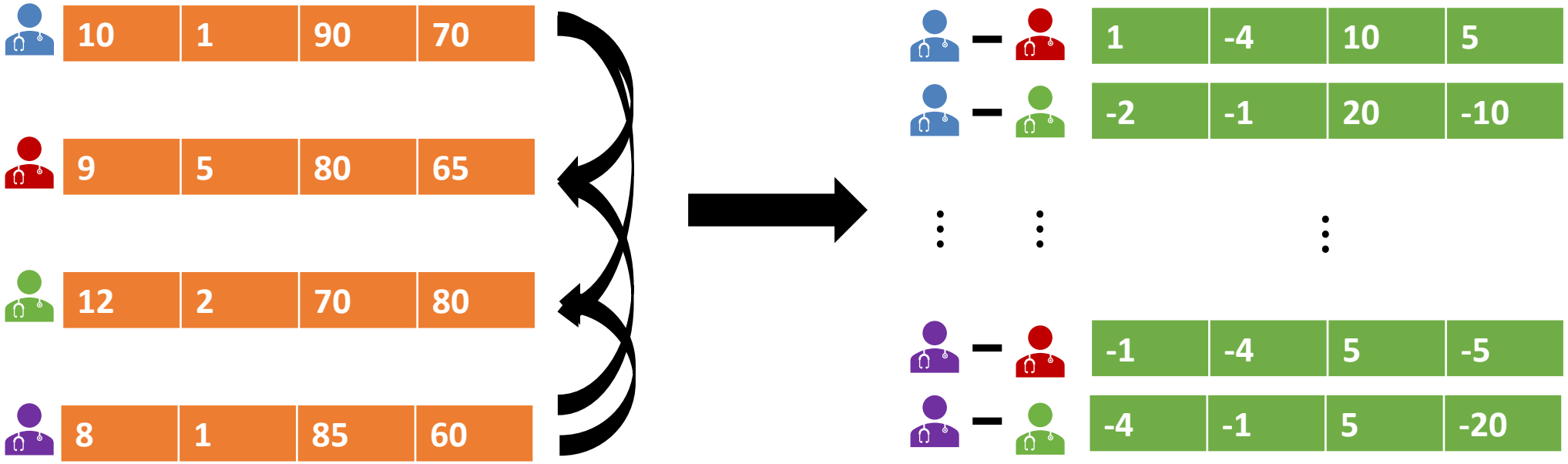
Slides with mean reference values > 10

Algorithm Output Vs Reference Values

	Mean Difference	SD of Difference	95% Limits of Agreement*		Coverage Rate
			Upper Limit	Lower Limit	
All Slides	0.17	9.59 (5.27)	-18.63 (-10.16)	18.98 (10.5)	95% (83.1%)
Slides with mean reference values ≤ 7	-0.27	5.66	-11.6	10.58	94.5%
Slides with mean reference values ∈ (7,10]	1.21	11.10	-20.54	22.98	93.3%
Slides with mean reference values >10	-0.39	11.25	-22.44	21.66	93.3%

*values in () are SD and LOA by **Naïve way**

Between-Reader Agreement



$$D_{jj',k} = Y_{jk} - Y_{j'k}$$

Between-Reader Agreement

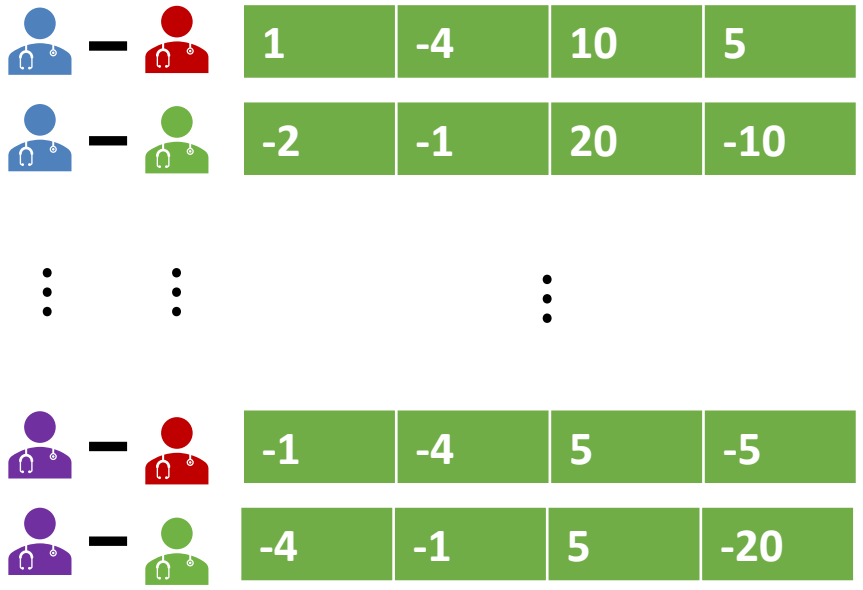
- Limits of Agreement among Readers
  - The mean difference :

$$\bar{D} = \frac{1}{J(J-1)K} \sum_j \sum_{j' \neq j} \sum_k D_{jj',k} = 0$$

  - The variance of the differences:

$$Var\left(D_{jj',k}\right) = Var\left(Y_{jk} - Y_{j'k}\right)$$

  - not independent



$$D_{jj',k} = Y_{jk} - Y_{j'k}$$

Between-Reader Agreement

- Two-way random effect ANOVA model for the reference values Y_{jk}

$$Y_{jk} = \mu + R'_j + C'_k + \varepsilon'_{jk}$$

  - $R'_j \sim N(0, \sigma^2_{yR})$, $C'_k \sim N(0, \sigma^2_{yC})$, $\varepsilon'_{jk} \sim N(0, \sigma^2_{y\varepsilon})$

- The variance of $D_{jj',k}$:

$$Var(D_{jj',k}) = Var(Y_{jk} - Y_{j'k})$$
$$= Var\left(R'_j - R'_{j'} - \varepsilon'_{jk} - \varepsilon'_{j'k}\right) = 2(\sigma^2_{yR} + \sigma^2_{y\varepsilon})$$

Between-Reader Agreement

- Two-way ANOVA table

Source	DF	Sum of Square (SS)	Mean Square (MS)	E(MS)
Reader	$J-1$	$SSR_y = K\sum_j(\overline{Y_{j\cdot}} - \bar{Y})^2$	$MSR_y = SSR_y/(J-1)$	$\sigma^2_{y\varepsilon} + K\sigma^2_{yR}$
Case	$K-1$	$SSC_y = J\sum_k(\overline{Y_{\cdot k}} - \bar{Y})^2$	$MSC_y = SSC_y/(K-1)$	$\sigma^2_{y\varepsilon} + J\sigma^2_{yC}$
Error	$(J-1)(K-1)$	$SSE_y = SST_y - SSR_y - SSC_y$	$MSE_y = SSE_y/(J-1)(K-1)$	$\sigma^2_{y\varepsilon}$
Total	$JK-1$	$SST_y = \sum_j\sum_k(Y_{jk} - \bar{Y})^2$		

Between-Reader Agreement

- Variance components estimation:

$$\hat{\sigma}^2_{y\varepsilon} = MSE_y, \qquad \hat{\sigma}^2_{yR} = \frac{MSR_y - MSE_y}{K}, \qquad \hat{\sigma}^2_{yC} = \frac{MSC_y - MSE_y}{J}$$

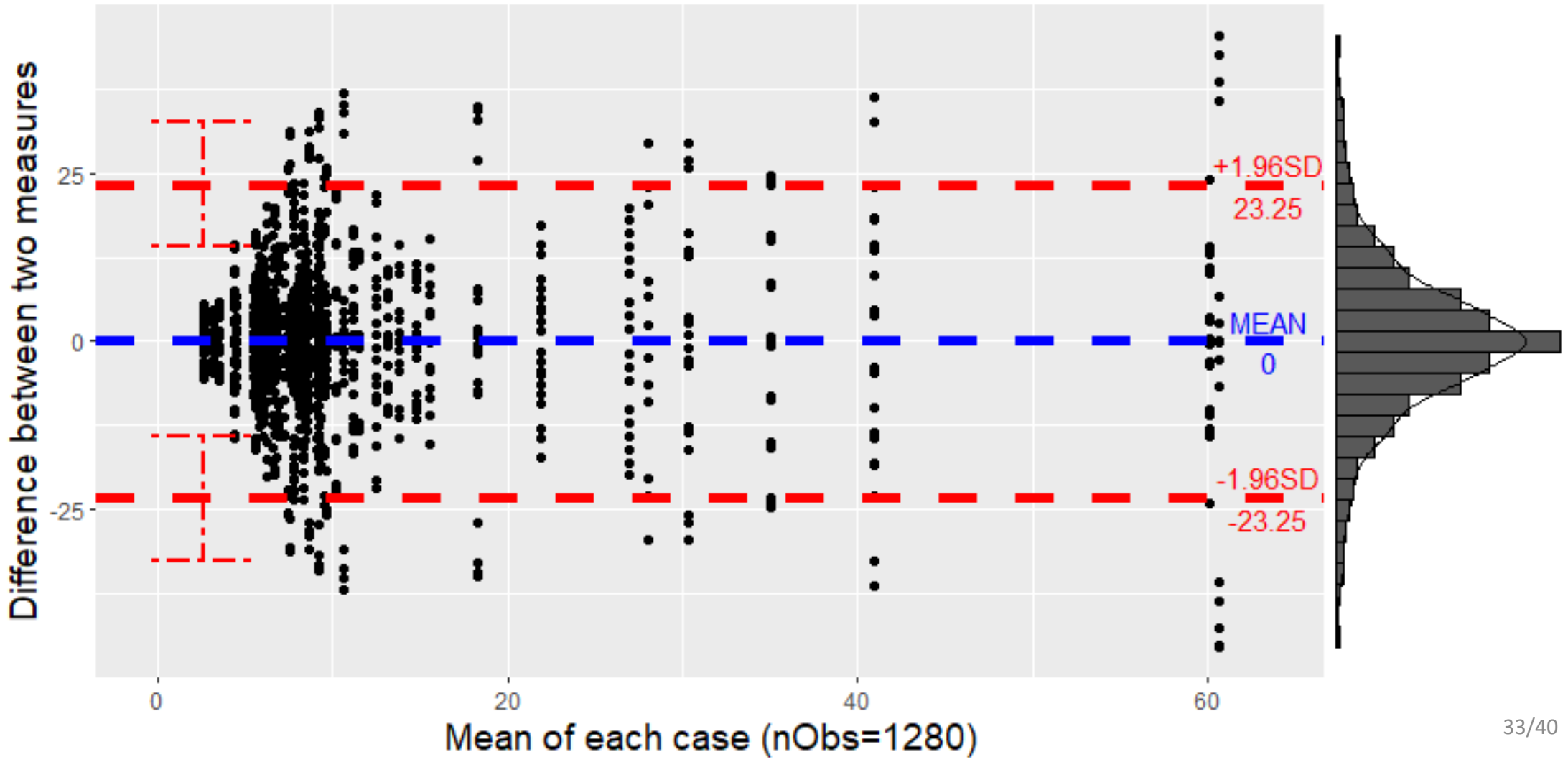
- Estimated variance of difference :

$$\widehat{Var}\left(D_{jj',k}\right) = 2\left(\hat{\sigma}^2_{yR} + \hat{\sigma}^2_{y\varepsilon}\right)$$
$$= \frac{2}{K}\left(MSR_y + (K-1) * MSE_y\right)$$

- The 95% **limits of agreement**: $0 \pm 1.96\sqrt{2\left(\hat{\sigma}^2_{yR} + \hat{\sigma}^2_{y\varepsilon}\right)}$

- The 95% **LOA for D_{jk}**: $\overline{D} \pm 1.96\sqrt{\hat{\sigma}^2_{dR} + \hat{\sigma}^2_{dC} + \hat{\sigma}^2_{d\varepsilon}}$

Between-Reader Agreement



Between-Reader Agreement

- ICC- Intraclass correlation coefficient
  - ICC(2,1) Two-way random, single measurements

$$ICC(2,1) = \frac{\sigma_{yC}^2}{\sigma_{yR}^2 + \sigma_{yC}^2 + \sigma_{y\varepsilon}^2}$$

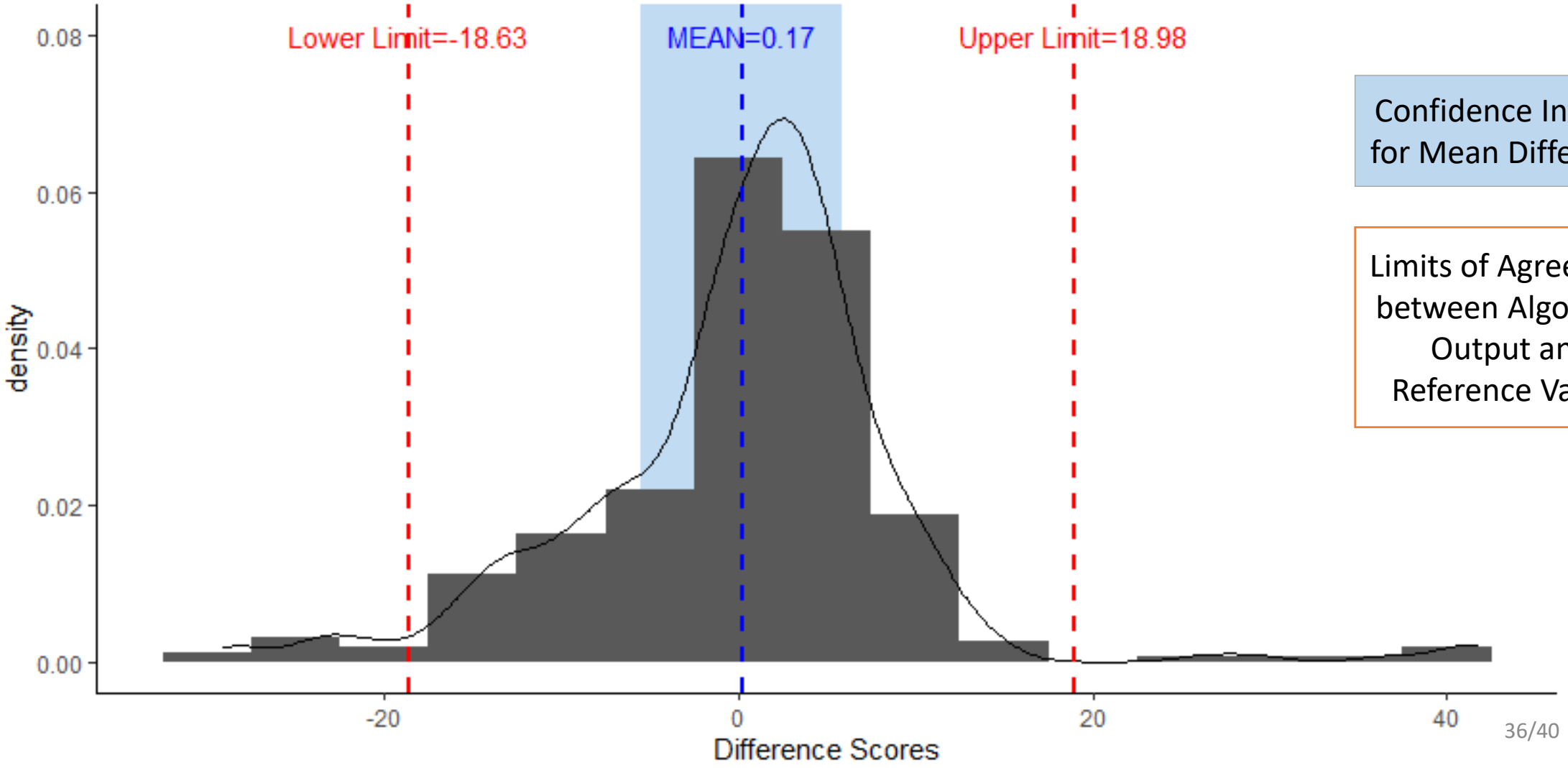
  - $Cov\left(Y_{jk}, Y_{j'k}\right) = \sigma_{yC}^2$
  - $Var\left(Y_{jk}\right) = \sigma_{yR}^2 + \sigma_{yC}^2 + \sigma_{y\varepsilon}^2$

  - When $\sigma_{yC}^2 \gg \sigma_{yR}^2$, ICC may not be able to reflect between-reader agreement

Between-Reader Agreement

	σ^2_{yR}	σ^2_{yC}	$\sigma^2_{y\varepsilon}$	LOA among Readers	ICC
All Slides	39.73	125.38	30.61	±23.25	0.64
Slides with mean reference values ≤ 10	35.84	0.28	14.23	±19.62	0.01
Slides with mean reference values > 10	54.25	242.59	67.89	±30.64	0.67

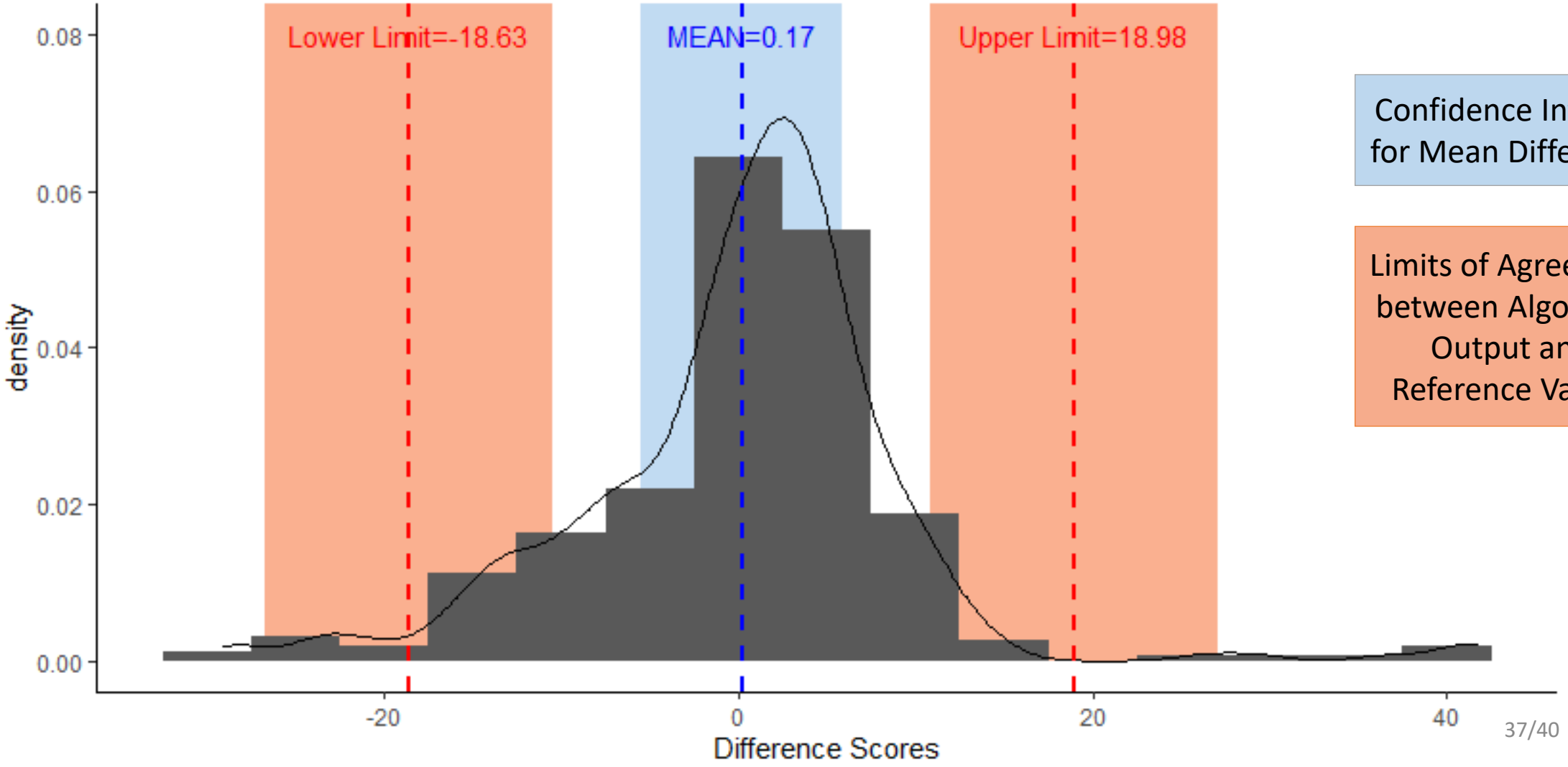
Limits of Agreement



Confidence Interval for Mean Difference

Limits of Agreement between Algorithm Output and Reference Values

Limits of Agreement

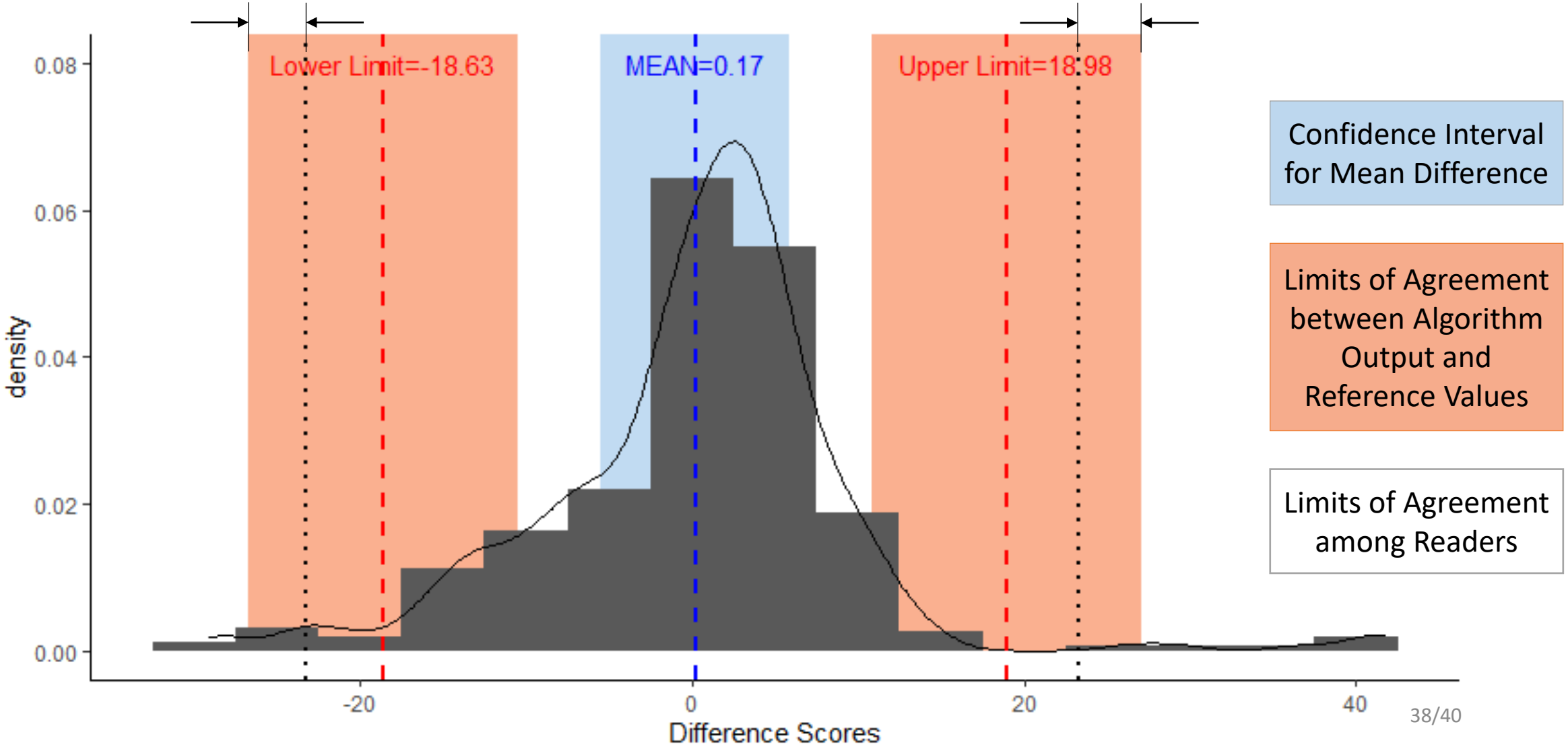


Confidence Interval for Mean Difference

Limits of Agreement between Algorithm Output and Reference Values

Limits of Agreement

Clinical tolerable non-inferiority margin?



Confidence Interval for Mean Difference

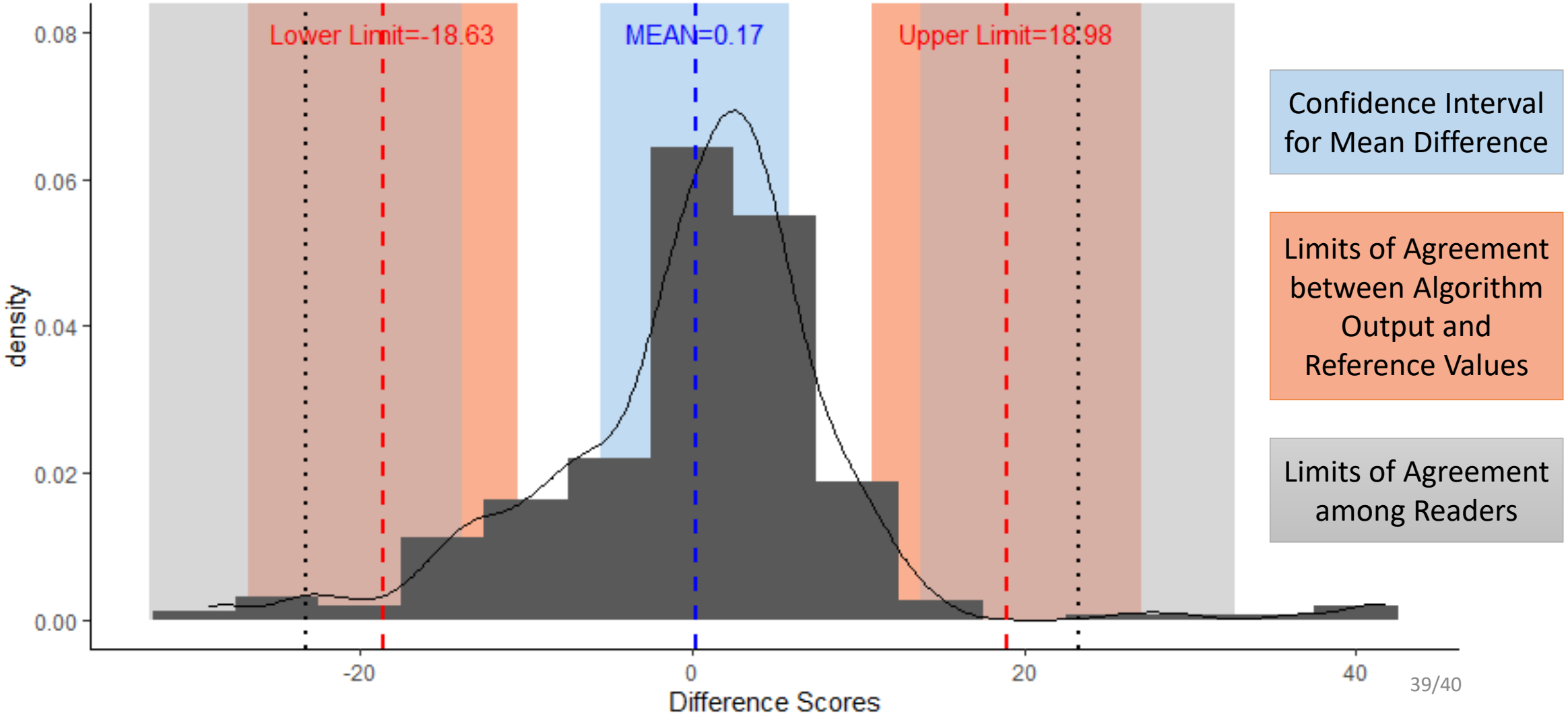
Limits of Agreement between Algorithm Output and Reference Values

Limits of Agreement among Readers

Limits of Agreement

Sample Size?

More Training?



Confidence Interval for Mean Difference

Limits of Agreement between Algorithm Output and Reference Values

Limits of Agreement among Readers

Summary

- Validate algorithm with quantitative measurement as output
  - Agreement analysis – limits of agreement
- Reference values from multiple readers
  - limits of agreement between algorithm output and reference values – ANOVA
  - Compare the algorithm result to the reader-averaged reference value – reader variability
- Between reader agreement
  - Limits of agreement among readers – ANOVA
  - Limitation of ICC

Future Work

- Analyze all the data from all the 30+ readers
- Analyze the data without averaging over the ROIs
- Non-parametric confidence interval for LOA
- Sample size calculation for the HTT pivotal study

Upcoming Plans

- GitHub Repositories
  - R package for limits of agreement analysis by using ANOVA
  - HTT pilot study data

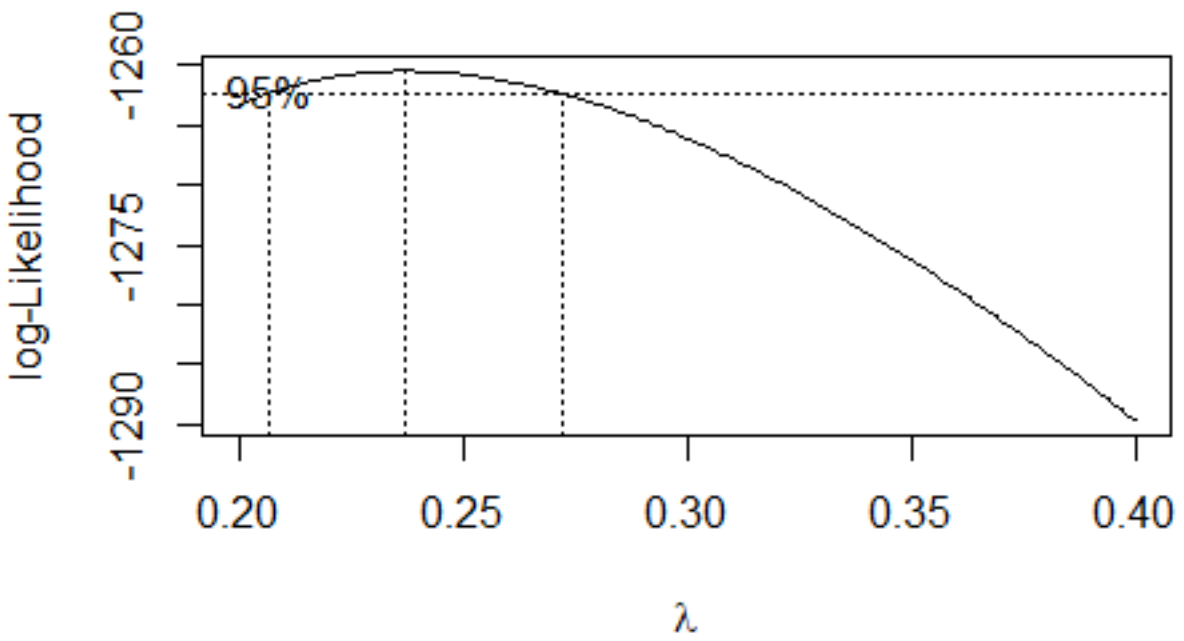
- ASA Joint Statistical Meeting
  - Presenter: Dr. Brandon Gallas
  - Title: Pathologist Agreement from Quantitative Measurements: a Pilot Study
  - Presentation Info
    - Live Speed Session: 3-4 Minute Overview August 8, 2021 (session starts at 1:30pm EDT)
    - Recorded Talk: 15 minutes
    - American Statistical Association's Joint Statistical Meeting in Seattle, August 7-12, 2021

Normality

- Box-Cox transformation

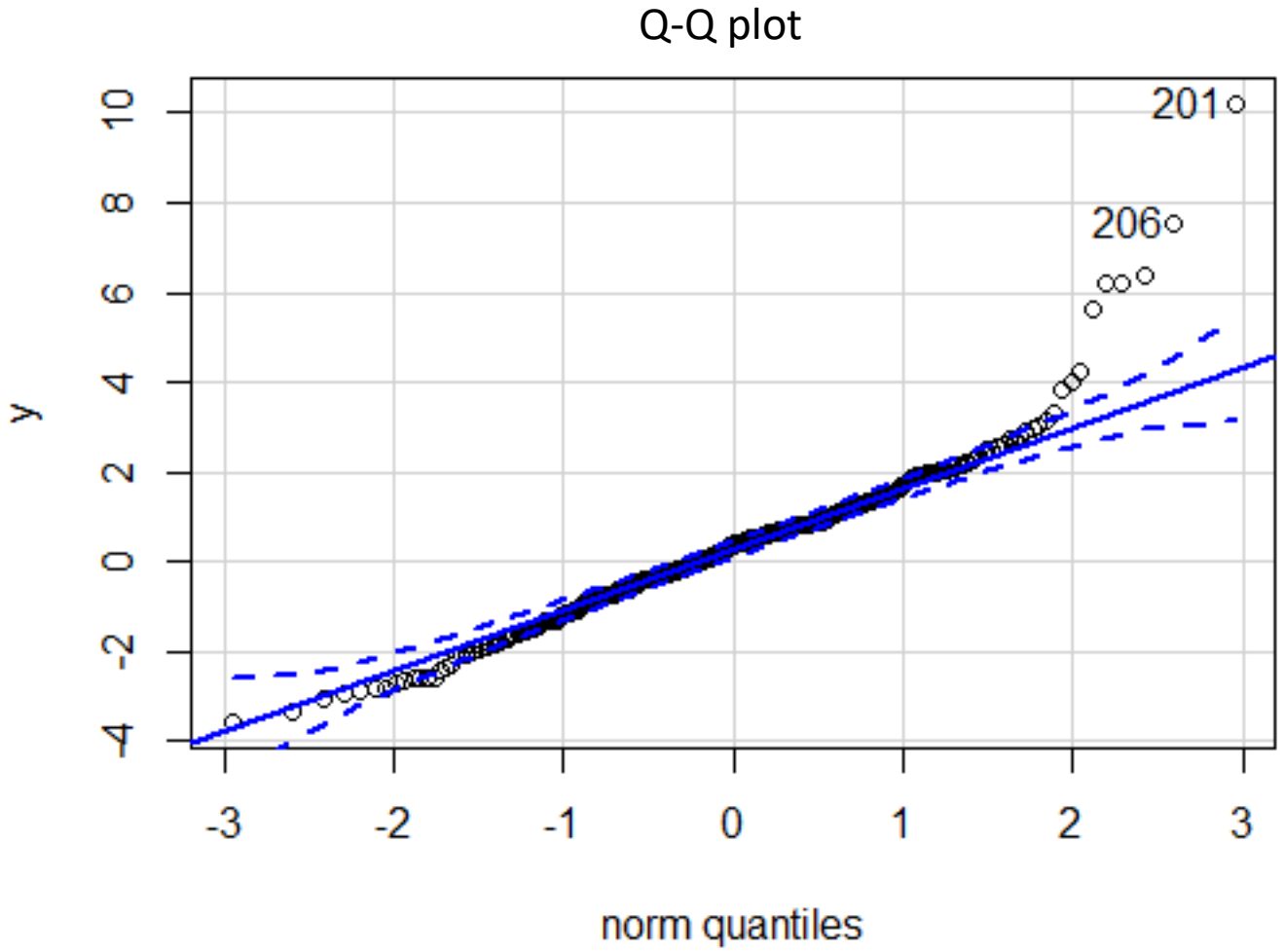
$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

- $\lambda = 0.237$



Normality

- The difference score is still not normally distributed
- Fat-tails
- Shapiro-Wilk normality test
  - W = 0.93099
  - p-value = 5.091e-11



HTT Pilot Study Data