Bland-Altman Analysis for Tabata2019_Diagn-Pathol_v14p65

Brandon D. Gallas

April 2, 2019

This document contains analysis and text for the following manuscript:

- Tabata, K.; Uraoka, N.; Benhamida, J.; Hanna, M. G.; Sirintrapun, S. J.; Gallas, B. D.; Gong, Q.; Aly, R. G.; Emoto, K.; Matsuda, K. M.; Hameed, M. R.; Klimstra, D. S. & Yagi, Y. (2019), 'Validation of mitotic cell quantification via microscopy and multiple whole-slide scanners.', *Diagn Pathol* **14**, 65.

Methods: Bland-Altman analysis

Intra-observer agreement between the scanner and microscope data was also analyzed with Bland-Altman plots and related summary statistics. For each modality we plot the differences in log counts between the paired scanner and microscope data for each pathologist against the average of each pair (citation: Bland1999_Stat-Methods-Med-Res_v8p135). The log transform stabilizes the variance in the count differences as a function of the mean (citation: Veta2016_PloS-One_v11pe0161286). The summary statistics include the mean differences in log counts and the standard deviation of the log-count differences (uncertainty). Twice the standard deviation of the log-count differences above and below the mean give the limits of agreement (LA). LA are similar to but different from confidence intervals, which typically quantify uncertainty in a mean. For this analysis, we counted all the cells marked as MFs for each reader in a WSI. This aligns with what is done in clinical practice (citation: clinical ref?). Therefore, we have four counts for each reader and modality. The uncertainties estimated in this Bland-Altman analysis account for the variability from the pathologists and the correlations that are arise when the pathologists evaluate the same cases, a so-called multi-reader multi-case analysis (citation: Gallas2007_J-Opt-Soc-Am-A_v24pB70).

- Bland1999_Stat-Methods-Med-Res_v8p135: Bland, J. M. & Altman, D. G. (1999), 'Measuring Agreement in Method Comparison Studies', *Stat Methods Med Res* **8**(2), 135-160.
- Veta2016_PloS-One_v11pe0161286: Veta, M.; van Diest, P. J.; Jiwa, M.; Al-Janabi, S. & Pluim, J. P. W. (2016), 'Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method', *PloS One* **11**(8), e0161286.
- Gallas2007_J-Opt-Soc-Am-A_v24pB70: Gallas, B. D.; Pennello, G. A. & Myers, K. J. (2007), 'Multi-reader Multicase Variance Analysis for Binary Data', *J Opt Soc Am A, Special Issue on Image Quality* **24**(12), B70-B80.

Methods: Accuracy

Accuracy was analyzed using the average of sensitivity and specificity, giving the 2 x 2 tables of true and false MFs vs. positive and negative determinations of all candidate MFs. Sensitivity is defined as the number of MFs detected by an observer divided by the number of true MFs. Specificity is defined as one minus the false-positive fraction, where the false-positive fraction is the number of false MFs that were positively marked, divided by the total number of false MFs. This average is equivalent to the area under the receiver operating characteristic curve for binary scores and is proportional to Youden's index (28, 29); it is also correlated with Cohen's Kappa (30). We reported the accuracy for each reader and modality and then the average over readers for each modality. We also performed a multiple-reader multiple-case (MRMC) analysis of reader-averaged accuracy using the Obuchowski-Rockette (OR) method (cite: Obuchowski1995_Commun-Stat-Simulat_v24p285, Hillis2014_Stat-Med_v33p330). This method takes as input the covariances between the

AUCs from all the reader by modality combinations (five readers times five modalities). These covariances account for within-slide correlation between measurements obtained on ROIs within the same slide (cite: Obuchowski1997_Biometrics_v53p567, Obuchowski1997_clusteredROC_software).

Hillis, S. L. (2014), 'A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data.', Stat Med 33(2), 330–360.

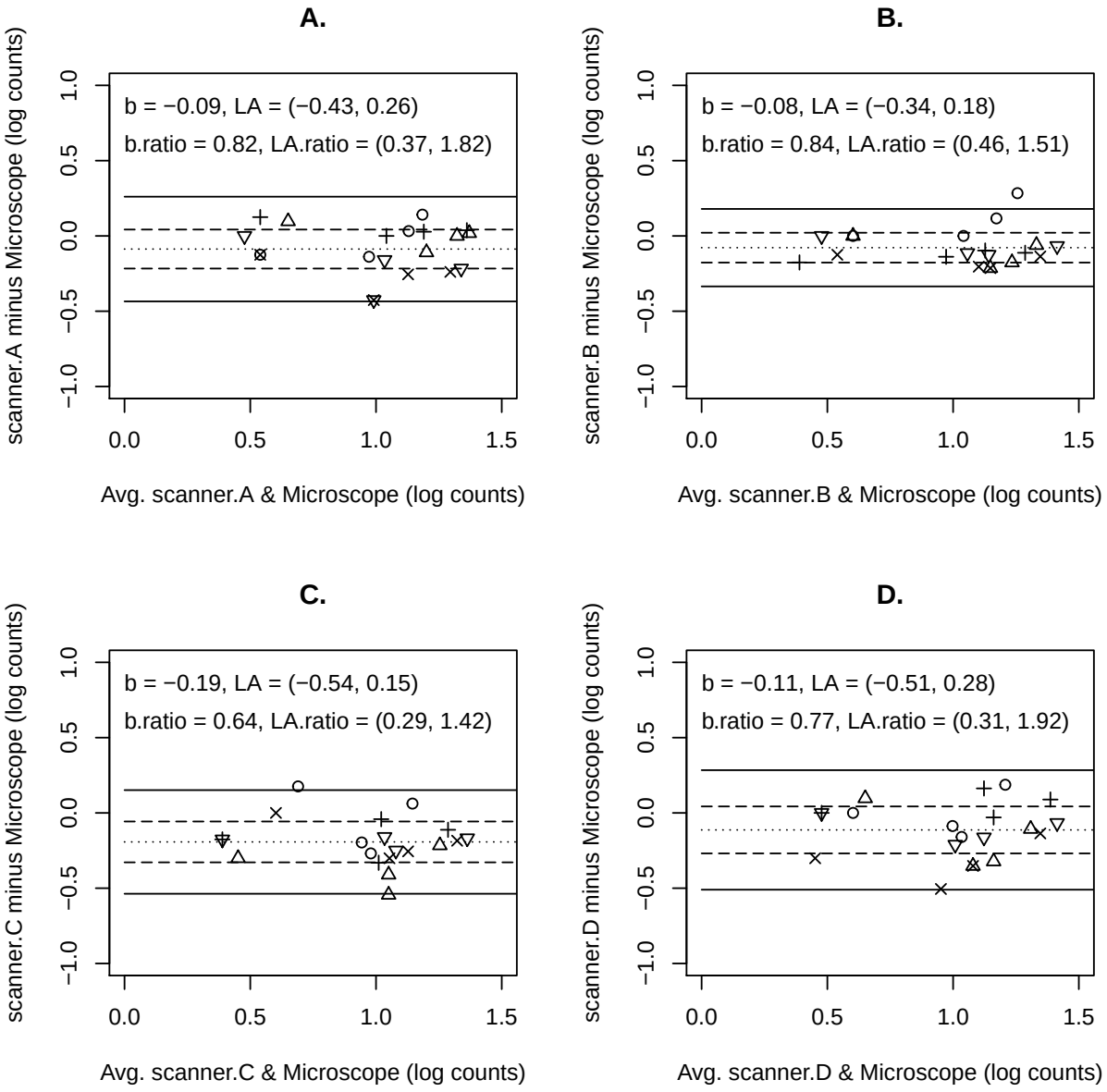
Obuchowski, N. A. & Rockette, H. E. (1995), 'Hypothesis Testing of Diagnostic Accuracy for Multiple Readers and Multiple Tests: An ANOVA Approach with Dependent Observations', Commun Stat B-Simul 24(2), 285-308.

Obuchowski, N. A. (1997), 'Nonparametric Analysis of Clustered ROC Curve Data', Biometrics 53(2), 567-578.

Obuchowski, N. (1997), 'funcs_clusteredROC.R: Nonparametric Analysis of Clustered ROC Curve Data', Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Silver Spring, MD. URL: https://www.lerner.ccf.org/qhs/software/roc_analysis.php, accessed 5/22/2019.

Results

Figure 1: "Within-Reader Log-Count Differences""




```
## pdf
##   2
```

Figure 1 Caption: Bland-Altman plots of within-reader differences in log (base 10) counts between each scanner (A, B, C, D) and the microscope. Each symbol corresponds to a different reader. The dotted line in each plot is b, the mean difference in the log counts. The dashed lines show the 95% MRMC confidence interval for b. The solid lines show the MRMC limits of agreement (LA). We map b and LA to ratios of counts with the inverse log: 10^b, 10^LA.

In Fig. 1 we show the within-reader Bland Altman plots comparing log-count differences from the scanners to those with the microscope. The biases observed in the log counts show that the pathologists marked fewer MFs with the scanners compared to the microscope. They marked between 16% to 36% fewer on average and 70% fewer in some cases.

```
##            ScannerA ScannerB ScannerC ScannerD Microscope
## Observer 1    0.713    0.743    0.685    0.706      0.764
## Observer 2    0.700    0.715    0.631    0.648      0.778
## Observer 3    0.704    0.738    0.717    0.802      0.806
## Observer 4    0.691    0.726    0.699    0.717      0.842
## Observer 5    0.698    0.754    0.738    0.785      0.802
## Average       0.701    0.735    0.694    0.732      0.798
## SE            0.021    0.023    0.028    0.035      0.021
## botCI         0.659    0.689    0.636    0.653      0.754
## topCI         0.743    0.780    0.752    0.810      0.842
## p-value       0.001    0.009    0.001    0.062         NA
```

Table 4 footnote: Accuracy refers to the average of sensitivity and specificity. SE, standard error; CI, confidence interval. The p-value corresponds to a two-sided hypothesis test comparing reader-averaged accuracy with each scanner viewing mode to the accuracy of the microscope. The p-values of the four hypotheses are compared following the sequentially rejective Bonferroni test with alpha = 0.05 (33). Statistical significance is indicated with an asterisk *. All analyses account for the correlations and variability from the readers reading the same ROIs, and the correlations arising from MFs contained within the same slides.

To compare all detected mitotic cell candidates with ground truth, we analyzed accuracy which is the average of sensitivity and specificity. Accuracy was between 0.631 and 0.842 across all readers and modes (Table 4, Figure 3). After averaging over readers for each detection method, we found (33) that mitosis detection accuracy of each of the three scanners, A, B, and C, was significantly less than that of the microscope.

Discussion

One interesting finding worth investigating with a larger study is that the pathologists found fewer mitotic figures on the scanners than on the microscope.