High Throughput Truthing (HTT) of pathologist annotations as a reference standard for validating artificial intelligence in digital pathology

Elfer, Katherine N[1]; Dudgeon, Sarah N[2]; Gallas, Brandon D[1]

1. FDA/CDRH/OSEL/Division of Imaging, Diagnostics and Software Reliability
2. CORE Center for Computational Health Yale-New Haven Hospital

Abstract

Background: Recent advancements in whole slide imaging (WSI) technology have exponentially increased the development of tools in digital pathology analysis, including artificial intelligence (AI) algorithms, allowing pathologists to be more efficient and potentially improve their diagnoses. Before coming to market, AI algorithm performance must first be validated against a reference standard. There are few examples of creating a reference standard using pathologist-annotated ground truth. This project will fill this gap.

Purpose: This work harnesses the flexibility of the Medical Device Development Tool (MDDT) program and promotes innovative methods in tool design. The work will support good Machine Learning Practices for digital pathology and serve as a demonstration of quality data collection methods and accompanied statistical handling of algorithm performance with truth by clinician observation with no given truth.

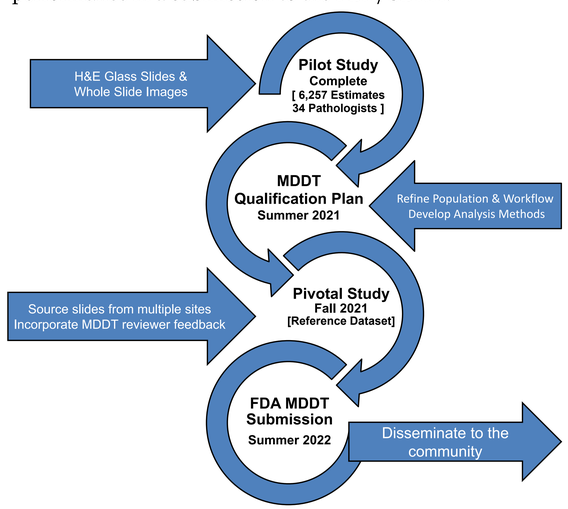
Methodology: The clinical use case of this work is the evaluation of the density of tumor-infiltrating lymphocytes (TILs) in breast cancer tissue stained with hematoxylin and eosin. Pathologist annotations are collected through two modalities to compare pathologist agreement between the clinical standard technology, an optical microscope, and two digital platforms. The microscope was used to provide the reference standard, eliminating bias from slide digitization. Image data is sourced from multiple clinical sites, ensuring generalizable results. Multiple pathologists will annotate each image across platforms, allowing for intra- and inter- pathologist agreement analysis. Algorithm performance will compare algorithm-pathologist differences to pathologist-pathologist differences.

Results: Pilot data collected through the microscope system and through the digital platforms have led to critical evaluation of image and annotator parameters needed to design a generalizable dataset. The final dataset will include pathology glass slides, corresponding digital slide images, and location-specific annotations.

Conclusions: The dataset created through this work will be submitted for qualification as an MDDT. This effort creates a resource typically only available to large companies, promoting advancement of AI as a medical device for innovators of all sizes. Partnerships with external collaborators are a critical component of this work to reach community consensus on a standard reference dataset in digital pathology aligned with leading standards in clinical practice.

High-Throughput Truthing Project

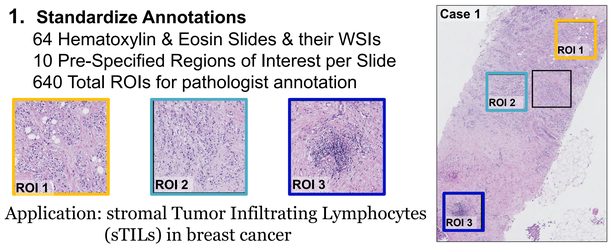
OBJECTIVE: We are crowdsourcing pathologists to collect data (images + pathologist annotations) that can be qualified by the FDA/CDRH medical device development tool program (MDDT). If successful, the MDDT qualified data along with a statistical software package for data analysis would be available to any algorithm developer to be used to validate their algorithm performance in a submission to the FDA/CDRH.



Materials and Methods

1. Standardize Annotations

64 Hematoxylin & Eosin Slides & their WSIs
10 Pre-Specified Regions of Interest per Slide
640 Total ROIs for pathologist annotation

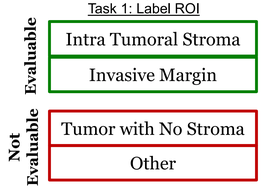


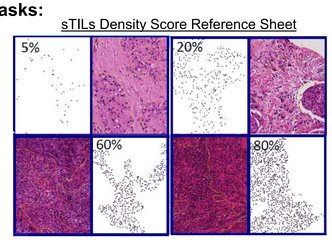
Application: stromal Tumor Infiltrating Lymphocytes (sTILs) in breast cancer

2. Pathologists complete three tasks:

Task 1: Label the ROI
Task 2: Record percent Stroma
Task 3: Record percent TILs

Task 1: Label ROI

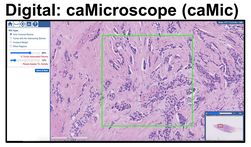




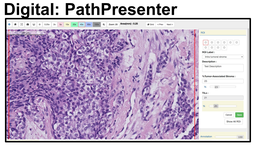
If the ROI is determined "not evaluable" the pathologist will not annotate the % density estimates

3. Annotations collected on three platforms: 2 Digital, 1 Microscope

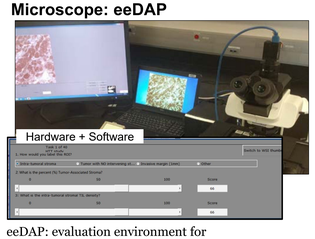
Digital: caMicroscope (caMic)



Digital: PathPresenter



Microscope: eeDAP



eeDAP: evaluation environment for digital and analogue pathology

Reader Progress caMicroscope

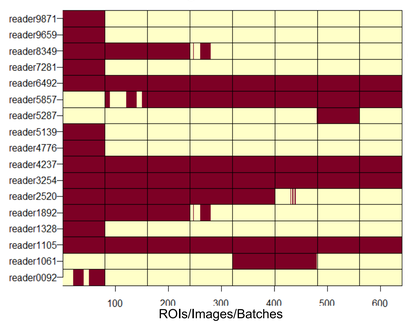


Figure 1. Current reader Progress across the caMicroscope platform. Red segments correspond to data collected by ROI (10 ROIs per image, 8 images per batch). Vertical lines separate batches

Discussion: **The pilot study ran during Spring 2020-April 2021 and recruited the most observers on caMicroscope (n=17). Future work will focus on recruiting observers on all three platforms: caMicroscope, Path Presenter, and eeDAP.**

Results and Discussion

Fig 2: Coefficient of Variation vs Mean sTILs for each ROI (n=571, caMic)

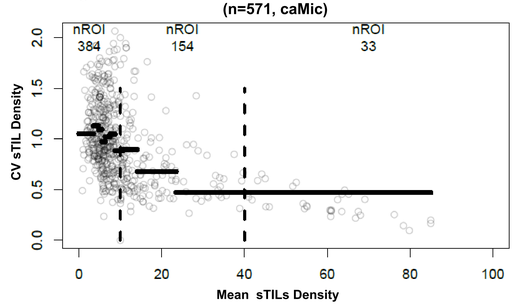


Figure 2. Coefficient of Variation averaged over all readers on caMicroscope. Each point represents one ROI. The horizontal lines show the average CV in 10% bins of the data (57 ROIs). Vertical dashed lines split the data into low (≤ 10%), medium (>10% & ≤ 40%), and high (>40%) sTIL density.

Discussion: **Relative variability decreases with density estimates and ~67% of the ROIs were scored with a "low" density estimate of ≤ 10%.**

Only observations on caMicroscope were included as it was the platform with the most data collected across all ROI/Image/Batches for the pilot study.

Fig 3.a: Histogram of Paired Observations (n=28126)

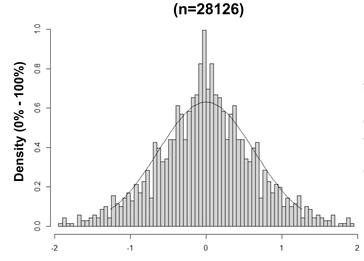


Fig 3.b: Normal Q-Q Plot (n=28126)

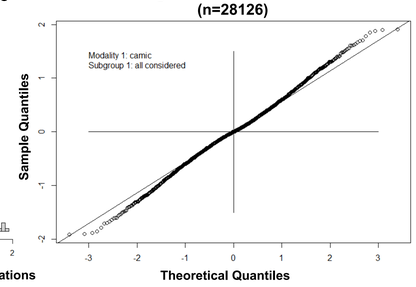


Figure 3.a. Histogram of the differences in the log of reader scores across all observations with Gaussian overlay.

Figure 3.b. Normal Q-Q plot of the sample vs. theoretical quantiles of all log-transformed observations.

Discussion: **The log-transformed data appears approximately normal, even with the noise: wide variances in scores, reader expertise, completion of training, and data collection across platforms.**

Fig 4: Between Reader Symmetrized Scatter Plot (n= 28,126)

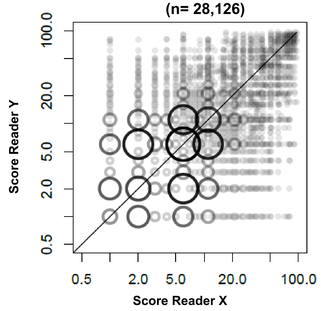


Figure 4. Between-Reader Symmetrized Scatter Plot: N=28,126 paired observations on a log-log scaled plot. Symmetrized means we plot (x,y) and (y,x) since we are pooling over readers and none is the reference. Size of symbol and transparency are scaled with number of paired observations; largest symbol = 922 paired observations.

Discussion: **We can see very large differences. The first column of points corresponds to one reader giving a score of 1. For some of these, another reader gives scores above 50.**

Who Contributes?

Connect with us: https://ncihub.org/groups/eedapstudies

Conclusion

Complete:

- ✓ Pilot study data-collection target achieved (still open): information gained informs future workflows, sizing for the pivotal study, and analysis methods
- ✓ Feedback from MDDT program informs pivotal study population selection

Future Work and Collaborators Needed:

- **NEED:** Recruitment of data collectors and physical slides for pivotal study
- **NEED:** Identification of host-sites for in-person data collection events with eeDAP
- **To-Do:** Identify subset of readers to serve as reference
- **To-Do:** Calculate limits of agreement of an "algorithm" and check non-inferiority

References:

- Dudgeon SN, *et. al.* "A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study." *J Pathol Inform*. In press.
- Wen, S. and Gallas B. "Three-way Mixed Effect ANOVA to Estimate MRMC Limits of Agreement." Submitted 2020, *Statistics in Biopharma Research*. Correspondence: si.wen@fda.hhs.gov
- Mable HD, *et. al.* "A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients." *J Pathol Inform* 11:22. 2020.