Title

High Throughput Truthing (HTT): pathologist agreement from a pilot study

Authors

Brandon D. Gallas [a], Katherine Elfer [a], Mohamed Amgad [b], Weijie Chen [a], Sarah Dudgeon [c], Rajarsi Gupta [d], Matthew Hanna [e], Steven Hart [f], Richard Huang [g], Evangelos Hytopoulos [h], Denis Larsimont [i], Xiaoxian Li [j], Anant Madabhushi [k], Hetal Marble [g], Roberto Salgado [l], Joel Saltz [d], Manasi Sheth [m], Rajendra Singh [n], Evan Szu [o], Darick Tong [o], Si Wen [a], Bruce Werness [o]

a. United States Food and Drug Administration, Center for Devices and Radiologic Health, Office of Science and Engineering Laboratories, Division of Imaging Diagnostics & Software Reliability, White Oak, MD
b. Department of Pathology, Northwestern University, Chicago, IL
c. CORE Center for Computational Health Yale-New Haven Hospital, New Haven, CT
d. Stony Brook Medicine Dept of Biomedical Informatics, Stony Brook, NY
e. Memorial Sloan Kettering Cancer Center, New York, NY
f. Department of Health Sciences Research, Mayo Clinic, Rochester, MN
g. Massachusetts General Hospital/Harvard Medical School, Boston, MA
h. iRhythm Technologies Inc, San Francisco, CA
i. Department of Pathology, Institut Jules Bordet, Brussels, Belgium
j. Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA
k. Case Western Reserve University, Cleveland, OH
l. Division of Research, Peter Mac Callum Cancer Centre, Melbourne, Australia; Department of Pathology, GZA-ZNA Hospitals, Antwerp, Belgium.
m. United States Food and Drug Administration, Center for Devices and Radiologic Health, Office of Product Quality and Evaluation, Office of Clinical Evidence and Analysis, Division of Biostatistics, White Oak, MD
n. Northwell health and Zucker School of Medicine, New York, NY
o. Arrive Bio, San Francisco, CA

Corresponding Author: Brandon.Gallas@fda.hhs.gov

Abstract

Background

Artificial intelligence algorithms in digital pathology have enormous potential to increase diagnostic speed and accuracy. However, the performance of these algorithms must be validated against a reference standard before deployment in clinical practice. In this work, pathologists are considered as the reference standard. We studied interobserver variability in pathologists who evaluate stromal tumor-infiltrating lymphocytes (sTILs) in hematoxylin and eosin stained breast cancer biopsy specimens. Our ultimate goal is to create a validation dataset that is fit for a regulatory purpose.

Methods

Following an IRB exempt determination protocol, we obtained informed consent of volunteer pathologist annotators prior to completing data collection tasks via two modalities: an optical light microscope and two digital platforms (slides were scanned at 40X). Pathologists were trained on the clinical task of sTIL density estimation before annotating pre-specified regions of interest (ROIs) across multiple platforms. The ROI selection protocol sampled ROIs in the tumor, tumor boundary, and elsewhere. Inter-pathologist agreement was characterized with the root mean-squared difference, which is analogous to the root mean-squared error but doesn’t require ground truth.

Results

The pilot study accumulated 6,257 sTIL density estimates from 34 pathologists evaluating 64 cases, with 10 ROIs per case. The variability of sTIL density estimates in an ROI increases with the mean; the reader-averaged root mean-squared differences were 8.3%, 17.7%, and 40.4% as the sTIL density reference score increased from 0-10%, 10-40%, and 40-100%, respectively. We also found that the root mean-squared differences for some pathologists were considerably larger than others (as much as 120% larger than the next largest root mean-squared difference).

Conclusions

Slides, images, and annotations were successfully provided by volunteer collaborators and participants, which created an innovative and thorough method for data collection and truthing. This pilot study will inform the development of statistical methods, simulation models, and sizing analyses for pivotal studies. The development and results of this validation dataset and analysis tools will be made publicly available to serve as an instructive tool for algorithm developers and researchers. Furthermore, the methods used to analyze pathologist agreement between density estimates are applicable to other quantitative biomarkers.