

User Manual

This is the user manual for the CDRH Regulatory Science Tool (RST) titled, **Dataset for the Development of Statistical Methods to Assess AI Model Performance: Images and Pathologist Annotations from the HTT Pilot Study**.

HTT refers to the High-Throughput Truthing project. Information about this project can be found here: <https://didsr.github.io/HTT.home/>.

This RST is a dataset of stromal tumor-infiltrating lymphocytes annotations. The annotation data is referred to as **pilotHTT_RST**. The data is available in the “HTT” R package Github repository:

- GitHub Repository: <https://github.com/DIDSR/HTT>
- CSV Pilot Annotation file: https://github.com/DIDSR/HTT/tree/main/inst/extdata/pilotHTT_RST.csv
- R data frame object: https://github.com/DIDSR/HTT/tree/main/data/pilotHTT_RST.rda

The HTT pilot study images are available from a server hosted by Emory University (<https://wolf.cci.emory.edu/camic/htt/>). We ask that you create an account and then request access to the images by email (<https://didsr.github.io/HTT.home/assets/pages/team>).

The R package also includes a “raw-data” version of the data (**pilotHTT**) and two R markdown files that were used to create figures for presentations and publications. These are discussed below.

This manual is meant to give users a high-level overview of the Tool, and is supplemented with further information in the **HTT R package manual**, which is available here: https://github.com/DIDSR/HTT/blob/main/inst/manual/HTT_2.0.1.pdf.

pilotHTT_RST Dataset

The **pilotHTT_RST** dataset is a simplified version of the raw data collected during the Pilot Study for the HTT Project. As this data was collected during a pilot study, some data elements were modified or added during collection. This means that there is a certain amount of missing data for some variables. The full “raw-data” version of the data collected during the Pilot Study is also available in the HTT package as **pilotHTT**. The raw data file contains additional fields, including pathologist experience, data-collection date, and viewer parameters. Explanations of the variables captured in pilotHTT_RST are available in the HTT R package manual in the “pilotHTT_RST” section (https://github.com/DIDSR/HTT/blob/main/inst/manual/HTT_2.0.1.pdf). A full description of the Pilot Study data-collection process is available in this paper:

- K. Elfer *et al.*, “Reproducible Reporting of the Collection and Evaluation of Annotations for Artificial Intelligence Models,” *Mod Pathol*, vol. 37, p. 100439, Jan. 2024, doi: [10.1016/j.modpat.2024.100439](https://doi.org/10.1016/j.modpat.2024.100439).

Supplementary Scripts and Utility Functions

There are two R markdown files (and corresponding compiled pdf files) available in the HTT R package. The R markdown files are available to recreate the figures in HTT presentations and publications. These files can also be used as a starting point to analyze similar data. Here are links to the folders containing the files and the corresponding publications containing the figures:

1. <https://github.com/DIDSR/HTT/tree/main/inst/extra/20210505-PathologyInformatics>

B. Gallas *et al.*, “High Throughput Truthing (HTT): Pathologist Agreement from a Pilot Study,” presented at the Pathology Informatics Summit, May 2021.

2. <https://github.com/DIDSR/HTT/tree/main/inst/extra/20220506-GarciaCancersPaper>

Garcia V, Elfer K, Peeters DJE, Ehinger A, Werness B, Ly A, Li X, Hanna MG, Blenman KRM, Salgado R, Gallas BD. Development of Training Materials for Pathologists to Provide Machine Learning Validation Data of Tumor-Infiltrating Lymphocytes in Breast Cancer. *Cancers (Basel)*. 2022 May 17;14(10):2467. doi: 10.3390/cancers14102467. PMID: 35626070; PMCID: PMC9139395.

Also contained in the HTT R package are four utility functions designed to aid in the statistical analysis of this or similar data with multiple readers reading multiple cases (commonly referred to as an MRMC study). These functions are further detailed in the HTT R package manual and include:

- **agreeDensityBRBM**: creates and analyzes paired-reader data
- **doStatsByCase**: summarizes scores for each case
- **getBlandAltmanWithDuplicates**: compares scores from two readers
- **lineFromTwoPoints**: finds the y value of a line defined in the input of the function.