HISTAI: a valuable dataset with a valuable lesson

Info

PUBLISHED DATE
2026-04-15

JOURNAL
The Journal of Pathology: Clinical Research

LINK
https://doi.org/10.1002/2056-4538.70089

Authors

Katherine J Hewitt Nic G Reitsam

Sebastian Foersch Ramona Erber Qinghe Zeng

Jakob Nikolas Kather

Abstract

The application of artificial intelligence in computational pathology depends on both robust algorithms and high-quality, clinically reliable data. Progress in this field has been limited by the scarcity of large, diverse, and well-validated whole slide image (WSI) datasets. To address this gap, HISTAI introduced an open-source resource comprising over 112,000 WSIs across multiple organ systems with associated clinical metadata. Here, we present a pathologist-led evaluation of label accuracy, metadata completeness, and dataset composition across 328 selected cases from this resource. Although HISTAI reports 47,279 cases, we identified only 44,564 unique cases after accounting for missing entries and duplicate records. Basic demographic information, including age and sex, was available for only 55% of cases. Dataset composition was uneven, with dermatopathology accounting for 47.1% of cases and gastrointestinal pathology for 24.0%; however, primary specialty was explicitly reported for only 39.6% of cases, obscuring this imbalance within the provided metadata. Notably, clinical ground truth is recorded in the Conclusion column. Concordance between the dataset's Conclusion and Diagnosis fields was observed in only 20.7% of cases, while 27.1% contained conflicting diagnoses. In a focused review of 198 cases, 30.3% were found to contain unclear or ambiguous diagnostic conclusions, including eight cases in which the diagnosis was incorrect. Assessment of molecular annotation revealed that only 18.9% of analyzed lung and colorectal cancer cases included molecular information. Furthermore, among adult-type diffuse gliomas, none of the 55 cases met current World Health Organisation Classification of Tumors of the Central Nervous System 5th Edition (WHO CNS5) diagnostic criteria, with IDH mutation status reported in only 15 cases. Together, these findings highlight substantial ambiguities in ground-truth labeling, incomplete molecular annotation, and limited documentation of dataset provenance and ethical oversight. While HISTAI represents a valuable open-source resource, its effective and responsible use requires careful clinical validation and close collaboration between computational researchers and pathologists.