326P Pathology's last exam? A curated text-based benchmark dataset for diagnostic pathology [Abstract]

Nic G. Reitsam, M. Gustav, S. Foersch, Bruno Märkl, J. N. Kather

Angaben zur Veröffentlichung / Publication details:

Reitsam, Nic G., M. Gustav, S. Foersch, Bruno Märkl, and J. N. Kather. 2025. "326P Pathology's last exam? A curated text-based benchmark dataset for diagnostic pathology [Abstract]." ESMO Real World Data and Digital Oncology 10 (Supplement): 100522. https://doi.org/10.1016/j.esmorw.2025.100522.

Legal entity responsible for the study: The authors.

Funding: Has not received any funding.

Disclosure: All authors have declared no conflicts of interest.

https://doi.org/10.1016/j.esmorw.2025.100521

326P Pathology's last exam? A curated text-based benchmark dataset for diagnostic pathology

N.G. Reitsam[1], M. Gustav[2], S. Foersch[3], B. Märkl[1], J.N. Kather[2]

[1]Pathology, Faculty of Medicine, University of Augsburg, Augsburg, Germany; [2]Else Kroener Fresenius Center for Digital Health, TUD Dresden University of Technology, Faculty of Medicine and University Hospital Carl Gustav Carus, Dresden, Germany; [3]Institute of Pathology, University Medical Center Mainz, Mainz, Germany

Background: Robust evaluation of large language models (LLMs) and agentic artificial intelligence (AI) in diagnostic pathology requires datasets that reflect the sequential, multimodal, and integrative nature of real-world practice. Existing resources rarely capture the structured interplay of clinical presentation, histology, immunohistochemistry, and molecular findings. To address this gap, we developed Pathology's Last Exam, a text-based benchmark of pathology cases designed to rigorously assess LLM-based diagnostic systems.

Methods: We curated 100 pathology cases from practice and leading journals (e.g., Am J Surg Pathol, Mod Pathol), enriched for rare and emerging entities, aberrant immunophenotypes, lesions of intermediate biological potential, and other challenging scenarios. Each case comprises a clinical summary, histopathology, special stains/immunohistochemistry, molecular findings, final diagnosis with references, and standardized metadata. All diagnostic evidence was provided to four large language models (MedGemma-27B, GPT-OSS-120B, Llama-4-Maverick-17B, GPT-5-Mini), each tasked with generating a final diagnostic interpretation. The dataset further supports stepwise information release to emulate the temporal progression of real diagnostic workflows, enabling systematic evaluation of model reasoning at both early and fully informed stages.

Results: The dataset spans several organ systems, and includes rare and complex diagnoses of neoplastic and non-neoplastic pathology cases (e.g., RUNX1-Mutant AML Mimicking B-Lymphoblastic Leukemia with aberrant B-cell immunophenotype; pilomatrix-like high-grade endometrioid carcinoma; POU2F3-positive, neuroendocrine marker low small cell carcinoma etc.). On the full-information diagnostic task, accuracy ranged from 29% (MedGemma-27B) to 75% (GPT-5-Mini).

Conclusions: Pathology's Last Exam provides a unique dataset for diagnostic reasoning in surgical pathology. Its structured, literature- and practice-derived cases support rigorous evaluation of AI models. Our findings underscore the need for expanded, pathology-specific reasoning benchmarks that combine curated literature-derived cases with new, expert-generated scenarios.

Editorial acknowledgement: During the preparation of this work the author(s) used ChatGPT 5 in order to assist with language editing. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Legal entity responsible for the study: The authors.

Funding: Has not received any funding.

Disclosure: N.G. Reitsam: Financial Interests, Personal, Invited Speaker: NanoString/Bruker. M. Gustav: Financial Interests, Personal, Invited Speaker: Techniker Krankenkasse (TK), Sartorius, AstraZeneca. S. Foersch: Financial Interests, Personal, Invited Speaker: MSD, AstraZeneca, BMS. J.N. Kather: Financial Interests, Personal, Advisory Role: Panakeia, AstraZeneca, MultiplexDx, Mindpeak, DoMore Diagnostics, Bioptimus; Financial Interests, Personal, Stocks/Shares: StratifAI, Synagen, Tremont AI, Ignition Labs; Financial Interests, Institutional, Research Grant: GSK; Financial Interests, Personal, Invited Speaker: AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, Fresenius. All other authors have declared no conflicts of interest.

https://doi.org/10.1016/j.esmorw.2025.100522

327P ClinicalTrialsMatch: An AI-assisted process for molecularly-driven patient-trial matching in oncology

M. Sánchez[1], B. Fite Abril[2], C. Viaplana[3], J. González[4], L. González[5], S. Martínez Badal[6], C. Gozalbo Barriga[7], M.A. Queralt[8], C. Vallès de Lanuza[9], C. Villaseca Sitjar[10], S. Aguilar Izquierdo[11], R. Dienstmann[12], A.P. Gómez[13]

[1]Data Engineering for Research Unit, Vall Hebron University Hospital, Vall Hebron Institute of Oncology (VHIO), Barcelona, Spain; [2]Department of Biomedical Engineering, Universitat de Barcelona, Barcelona, Spain; [3]Oncology Data Science Department (Odyssey Group), VHIO - Vall d'Hebron Institute of Oncology, Barcelona, Spain; [4]Molecular Prescreening Molecular, Vall Hebron University Hospital, Vall Hebron Institute of Oncology (VHIO), Barcelona, Spain; [5]Molecular Prescreening Program, Vall Hebron University Hospital, Vall Hebron Institute of Oncology (VHIO), Barcelona, Spain; [6]Odyssey - VHIO, VHIO - Vall d'Hebron Institute of Oncology, Barcelona, Spain; [7]Department of Data engeneering / science, VHIO - Vall d'Hebron Institute of Oncology, Barcelona, Spain; [8]Department of Data Engineering for Research Unit (DERU), VHIO - Vall d'Hebron Institute of Oncology, Barcelona, Spain; [9]Department of Data Engineering for Research Unit (DERU), Vall Hebron University Hospital, Vall Hebron Institute of Oncology (VHIO), Barcelona, Spain; [10]Department of Data Engineering for Research, Vall Hebron University Hospital, Vall Hebron Institute of Oncology (VHIO), Barcelona, Spain; [11]Molecular Prescreening Program Department, Vall d'Hebron University Hospital, Barcelona, Spain; [12]Oncology Data Science Department, VHIO - Vall d'Hebron Institute of Oncology, Barcelona, Spain; [13]Data Engineering for Research, VHIO - Vall d'Hebron Institute of Oncology, Barcelona, Spain

Background: Finding suitable clinical trials for oncology patients is challenging, as eligibility criteria are scattered across registries and written in complex, unstructured language. Manual screening is resource-intensive and may lead to patients missing opportunities for inclusion. To address this, we implemented ClinicalTrialsMatch, a process that leverages large language models (LLMs) and expert curation to enable systematic, tumor type and biomarker-based trial matching (TM).

Methods: Tumor type and molecular eligibility criteria from ∼400 active trials were retrieved from ClinicalTrials.gov and processed by an LLM, which converted unstructured text into structured, machine-readable variables. Outputs were curated and validated in REDCap by an expert team of molecular biologists and data entries. In parallel, molecular results from the institutional Prescreening Program (with next-generation sequencing, immunohistochemistry and in-situ hybridization assays of over 2,000 patients every year) were structured using a predefined biomarker vocabulary that could be matched to clinical trial inclusion/exclusion criteria. A Python-based matching agent (MA) compared trial eligibility with tumor profiles, generating candidate matches.

Results: To evaluate the MA, 2,172 patients tested in 2024 were included. The algorithm generated a list of potential matches, cross-referenced with institutional trial inclusion records. Matches corresponding to real patient enrolments were considered validated. Preliminary results showed that 72.5% of suggested matches were confirmed by real enrolments. The remaining 27.5% were not. This is mainly driven by eligibility factors extending beyond tumor type and molecular criteria, including prior lines of therapy and missed potential matches due to logistic or timing limitations.

Conclusions: ClinicalTrialsMatch demonstrates that AI-augmented TM is feasible and can be integrated into routine workflows. While molecular and tumor data are important inclusion criteria, expanding to additional clinical variables is expected to improve precision and reduce false positives. This targeted use of AI exemplifies how automation can support expert-driven decision-making in oncology.

Legal entity responsible for the study: Vall d'Hebron Institute of Oncology.

Funding: Has not received any funding.

Disclosure: All authors have declared no conflicts of interest.

https://doi.org/10.1016/j.esmorw.2025.100523

328P Explainable deep learning for FDG PET/CT tumor detection: Evaluation on a pan-cancer dataset

R. Romanò[1], L. Moiana[2], L. Provenzano[1], M. Favali[2], N. Salmistraro[3], M.B. Marino[1], B. Guirges[2], C. Silvestri[1], M. Brambilla[1], L. Mazzeo[1], M. Occhipinti[1], G. Corrao[1], C. Proto[1], F. Trovò[2], F.G.M. De Braud[1], D. Signorelli[3], G. Lo Russo[1], A. Pedrocchi[2], A. Prelaj[1], V. Miskovic[2]

[1]Medical Oncology Department 1, Fondazione IRCCS - Istituto Nazionale dei Tumori, Milan, Italy; [2]Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy; [3]Hematology, Oncology and Molecular Medicine Department, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy

Background: FDG PET/CT is central to oncologic staging and response assessment, but interpretation is time-consuming and reader-dependent. While AI methods have largely focused on CT and MRI, PET remains underexplored. Deep learning (DL) may enable reproducible, automated PET tumor detection. We adapted Häggström *et al.* pipeline (developed for binary tumor detection using 2D ResNet34 pre-trained on