408P Combining rules and machine learning to improve sustainability and explicability of the extraction of pathological cancer markers from patient reports: Cross-sectional multicenter cohort study

E. Kempf[1], S. Priou[2], A. Redjdal[3], B. Dura[4], J. Calderaro[5], C. Brones[1], P. Wasjbürt[4], L. Bennani[6], X. Tannier[6]

[1]Medical Oncology, Centre Hospitalier Universitaire Henri-Mondor AP-HP, Creteil, France; [2]Industrial Engineering Laboratory, CentraleSupélec - Paris-Saclay Campus, Gif Sur Yvette, France; [3]Laboratory of Applied Biomechanics, University Gustave Eiffel, Noisy-le-Grand, France; [4]Department of Digital Services, Assistance Publique - Hopitaux De Paris, Paris, France; [5]Pathology, Henri-Mondor University Hospital, Créteil, France; [6]Limics lab, Sorbonne University, Paris, France

Background: Machine learning (ML) information extraction (IE) algorithms are associated with a concerning carbon footprint and lack explicability. We aimed to compare the performance of 2 IE models based on ML and rules, respectively, and to develop an IE method combining both approaches to optimize performance, sustainability and explicability.

Methods: We extracted 7 biomarkers from postoperative pathology reports of cholangiocarcinoma patients newly referred to the Greater Paris Teaching Hospital (APHP): cancer lymph node, vascular and perinervous invasion; tumor size, differentiation and pTNM stage; completeness of microscopic resection. We developed and validated both IE rule-based and ML models. The data set was divided in 2 by half to create development and test sets and manually annotated by oncologists. We compared the performance metrics (precision, recall) for each variable. We evaluated the rule development burden: number of analyzed reports and terminology development related to the linguistic characteristics of each variable. We developed an optimized IE method combining rules and ML.

Results: Between 2017 and 2020, 2,148 patients resulted in 289 reports. On the test set, for all the entities, the performance metrics ranged from 81% to 100% and from 79% to 100% for rules and ML, respectively. The differences in metrics never exceeded 5%, except for the precision of the 'tumor size' entity. We identified 3 pitfalls in IE rule development that ML might overcome: the gap between theoretical knowledge formalization and reporting in free texts; the slow rhythm of the dataset discovery by the rule developer; and the differentiated workload of rule development according to each entity linguistic characteristics. We suggested a combined IE approach based on both rules and ML, depending on the linguistic characteristics of each entity, and considering rules and ML as IE default and backup options, respectively.

Conclusions: Under specific text conditions, rules are a valuable IE default option to be combined with ML when necessary to optimize performance, sustainability and explicability of algorithms.

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.2025.100604

409P Benchmarking performance of reasoning large language models and agent workflows on complex clinical case domains

Z. Carrero[1], S. Jayabalan[1], J.N. Kather[2]

[1]Kather Lab, Else Kröner Fresenius Zentrum für Digitale Gesundheit, Dresden, Germany; [2]Medical Oncology Department, Technische Universität Dresden - Carl Gustav Carus Faculty of Medicine, Dresden, Germany

Background: Agentic AI systems and reasoning LLMs have attracted significant interest for clinical integration, yet their effectiveness in high-risk domains such as ethics, medical knowledge, and safety remains unclear. In this study, we evaluate their performance in complex clinical scenarios that continue to impede safe adoption in healthcare.

Methods: We evaluated performance of LLM workflows on publicly available question-answering benchmarks in ethics, medical knowledge, regulation, and safety. Each benchmark was tested under three workflows: (1) zero-shot, (2) evaluator–optimizer loop (one LLM answers, another provides feedback), and (3) multi-agent orchestration (one LLM directing tool-LLMs). We used GPT-4o (non-reasoning), GPT-5 (reasoning), and GPT-OSS-120b (reasoning, locally hosted) OpenAI models in the workflows. Performance was measured by accuracy against ground truth and model-reported confidence scores.

Results: A set of 9082 multiple-choice questions with human-derived ground truth labels were curated from 13 open-source benchmark datasets. Baseline accuracy levels were defined by performance of zero-shot GPT-4o model. Evaluator-optimizer LLM workflow and multi-agent system implemented using GPT-4o showed statistically insignificant mean increase in accuracy across all datasets of 1.52% and 2.52%, respectively, compared to baseline. Zero-shot reasoning models had statistically significant difference in mean accuracy of 17.45% (p=0.006) and 10.07% (p=0.05) for GPT-5 ad GPT-OSS-120b models, respectively. Use of GPT-OSS-120b reasoning model in evaluator-optimizer workflow yielded statistically insignificant decrease in accuracy of 0.31% from zero-shot GPT-OSS-120b.

Conclusions: Preliminary results demonstrate that LLM workflows and agent systems struggle to improve performance across benchmarks likely due to lack of decomposition into well-defined tasks, while reasoning capability provide enhanced performance for clinically complex scenarios.

Legal entity responsible for the study: The authors.

Funding: Else Kröner Fresenius Zentrum für Digitale Gesundheit (TU Dresden, Kather Lab).

Disclosure: J.N. Kather: Financial Interests, Personal, Invited Speaker, Talk on 14 November 2022: Fresenius; Financial Interests, Personal, Advisory Board, Scientific Advisory Board since 2022 till February 6, 2025: Owkin; Financial Interests, Personal, Advisory Board, Scientific Advisory Board since 2022 till February 21, 2025: DoMore Diagnostics; Financial Interests, Personal, Advisory Board, Scientific Advisory Board since 2022: Panakeia, London, UK; Financial Interests, Personal, Invited Speaker, Talk on 4 July 2023: Bayer; Financial Interests, Personal, Invited Speaker, Talk on 1 July 2023: BMS; Financial Interests, Personal, Invited Speaker, Talk on 13 November 2024: Roche; Financial Interests, Personal, Invited Speaker, Invited talks on 21 October 2023 and 31 July 2024: Pfizer; Financial Interests, Personal, Other, Expert services to select activities of AstraZeneca, e.g. Advisory Board Participation, Invited Lectures at internal events and participation in technical discussion meetings, starting in March 2023: AstraZeneca; Financial Interests, Personal, Invited Speaker, Invited lecture on 25 July 2024: Daiichi Sankyo; Financial Interests, Personal, Other, Consultancy on 5 June 2024: Bioptimus; Financial Interests, Personal, Invited Speaker, Talk on 12 January 2024: Janssen; Financial Interests, Personal, Invited Speaker, Talk on 9 May 2023: Merck Sharp and Dohme; Financial Interests, Personal, Invited Speaker, Talk on 26 September 2024: Merck; Financial Interests, Personal, Other, Consultancy on 9 October 2023 till February 28, 2025: Mindpeak; Financial Interests, Personal, Other, Consultancy since 2024: MultiplexDx; Financial Interests, Personal, Invited Speaker, Invited talk in November, 2023: Eisai; Financial Interests, Personal, Stocks/Shares, Shares and part-time activities in a company that provides artificial intelligence services for life science customers.: StratifAI GmbH; Financial Interests, Personal, Stocks/Shares, Advisory board membership and share ownership for Synagen GmbH (www.synagen.ai): Synagen GmbH; Financial Interests, Personal, Stocks/Shares, Holding shares (33%), no personal remuneration. Ignition Lab is a small VC that provides funding for life science startups.: Ignition Labs GmbH; Financial Interests, Institutional, Invited Speaker, I am PI on a research project at University Hospital Heidelberg which was funded by GSK.: GSK. All other authors have declared no conflicts of interest.

https://doi.org/10.1016/j.esmorw.2025.100605

411P From prediction to explanation: A counterfactual framework for platinum resistance in epithelial ovarian cancer

A. Traversa[1], M.N. Rosanu[2], F. Fati[1], L. De Vitis[3], G. Schivardi[2], L. Ribero[2], C. Taliento[2], D. Fumagalli[2], G. Aletti[2], N. Colombo[4], F. Multinu[2], E. De Momi[1]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy; [2]Department of Gynecology, IEO - Istituto Europeo di Oncologia IRCCS, Milan, Italy; [3]Department of Obstetrics and Gynecology, Mayo Clinic, Rochester, United States of America; [4]Gynecologic Oncology Program, IEO - Istituto Europeo di Oncologia IRCCS, Milan, Italy

Background: Platinum resistance, defined as recurrence within 6 months after platinum chemotherapy, carries poor prognosis in epithelial ovarian cancer (EOC). Early identification of high-risk patients is crucial, but predictive models often lack transparency, limiting clinical adoption. We aimed to identify actionable clinical variables that could improve patients' platinum sensitivity.

Methods: We included advanced EOC patients who underwent primary debulking surgery in a single institution (2015—2023). Relevant clinical, surgical and laboratory variables were selected. To address imbalance, Synthetic Minority Oversampling Technique was applied to the training set. Data were split into training (80%) and test (20%), with hyperparameters tuned by 5-fold cross-validation. Logistic regression, random forest and support vector machine (SVM) were compared. Performance was assessed with ROC-AUC and F1-score (medians across folds). Interpretability was explored with DiCE, generating counterfactual scenarios globally and under actionable constraints. Variables modified in >50% of cases to change prediction were considered impactful.

Results: Of 1396 patients, complete data for all 50 variables were available for 505 of them. SVM achieved the best performance (ROC-AUC 0.78 ± 0.53; F1-score 0.85 ± 0.05), significantly outperforming the other models (Table). Counterfactual analyses identified HE4, CA125 and platelets as the most influential features. When limited to clinically actionable variables, BMI, leukocyte count and surgical complexity score were frequently modified, highlighting the role of metabolic status, systemic inflammation and tumor burden in platinum resistance.

Table: 411P Results of tested models (median + IQR)

Metric	Logistic regression	SVM (p < 0.05)	Random Forest
ROC-AUC	0.67 ± 0.72	0.78 ± 0.53	0.50 ± 0.44
F1-score	0.71 ± 0.08	0.85 ± 0.05	0.85 ± 0.01

Conclusions: A counterfactual framework links robust, accurate prediction of platinum resistance with clinically meaningful explanations. The identified variables

could serve as actionable targets to improve platinum sensitivity, warranting prospective validation.

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.2025.100607

412P Structuring GDPR-compliant private networks to enable LLM-extracted oncology data on pseudonymized patient EHR data in Europe

L. Ellsworth[1], L. Groizard[2], F. Stefan[1], A. Schwarz[1], N. Viani[2], K. Harrison[3], A. Hadjigeorgiou[3], B. Adamson[3], I. Serko[2], D. Farrar[3], M. Hertstein[1], Y. Leon[4], M. Murchison[2], K. Seidl-Rathkopf[1]

[1]Flatiron Health GmbH, Berlin, Germany; [2]Flatiron Health UK, London, United Kingdom; [3]Flatiron Health, New York, United States of America; [4]Flatiron Health K.K., Minato-ku, Japan

Background: The expansion of real-world data in oncology across global markets require scalable high-quality curation of electronic health records (EHRs). Human-driven, manual extraction of data (abstraction) is resource-intensive and limits scalability, while fully automated approaches may lack the accuracy needed for regulatory and research use. Our objective was to develop a GDPR-compliant private network to enable an efficient, high-quality hybrid abstraction platform that combines large language models (LLMs) and machine learning with expert human review and supervision.

Methods: We began with EHR from partner sites of Flatiron Health in Europe. Patient-level EHR data were processed within secure, privacy-compliant environments using a "lock box" approach: data were minimized, pseudonymized, and accessed only within private architectures to ensure compliance with GDPR and local regulations. Connectivity between source data and LLMs was enabled via private network connectivity, ensuring data never reaches the public internet. LLMs were used in a static, pre-trained state, such that individual patient data were not used for model training. LLMs extract key clinical variables from unstructured documents, and expert abstractors independently review and validate outputs, all within an isolated network with strict access controls. This architecture enables usage of best in class LLMs, within a closed ecosystem, such that patient data neither informs future model development nor is accessible to the model maintainers.

Results: Across multiple countries, the private architecture keeps data secure and isolated within our cloud processing environments, such that data are never shared over the public internet. Simultaneously, we enabled industry-leading LLM tooling for efficient oncology data extraction.

Conclusions: This GDPR-compliant private network architecture enables access to the latest LLM models available, while preserving patient privacy. The integration of LLMs within our existing abstraction systems ensures compliance and privacy, laying the foundation for robust multinational research and regulatory acceptance.

Editorial acknowledgement: During the preparation of this work the authors used Dashworks AI in order to revise early drafts. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Legal entity responsible for the study: Flatiron Health, Inc.

Funding: Flatiron Health, Inc.

Disclosure: L. Ellsworth, L. Groizard, F. Stefan, N. Viani, K. Harrison, A. Hadjigeorgiou, B. Adamson, I. Serko, D. Farrar, M. Hertstein, Y. Leon, M. Murchison, K. Seidl-Rathkopf: Financial Interests, Personal, Full or part-time Employment: Flatiron Health; Financial Interests, Personal, Stocks/Shares: Roche. A. Schwarz: Financial Interests, Personal, Full or part-time Employment: Flatiron Health.

https://doi.org/10.1016/j.esmorw.2025.100608

413P AI-driven privacy-preserving synthetic data generation for mortality prediction

T. Qaiser[1], M. Rahman[2], S. Zakharkin[3]

[1]Data and Statistical Sciences Centre for RWE and EG, Daiichi Sankyo UK Ltd., Gerrards Cross, United Kingdom; [2]Department of Mathematics, University of Maryland, College Park, College Park, United States of America; [3]Data and Statistical Sciences Centre for RWE and EG, Daiichi Sankyo, Inc., Basking Ridge, United States of America

Background: Data privacy regulations restrict access to clinical trials and Real-World data, limiting their use in healthcare applications. Synthetic data replicating statistical properties of real datasets enables privacy-preserving analyses. We used the public MIMIC-IV ICU dataset to generate synthetic data with statistical and AI methods, assess the risk of patient re-identification, and evaluate utility of machine learning algorithms for mortality prediction.

Methods: The original dataset was preprocessed and missing values imputed using missForest. Synthetic datasets were generated with CART, Linear Regression, CTGAN, and TVAE. For each method, 10 datasets were created using the synthpop R package and the synthcity Python library. Fidelity was evaluated with correlation matrices and t-SNE plots. The data was split 70/30 into training/testing sets for predictive modeling, including classification (XGBoost, Random Forest, Logistic Regression) and survival analyses (XGB Survival, Survival Forest, Cox PH). Prediction performance was assessed with AUROC, F1, SHAP, and C-Index. Privacy risk was evaluated with synthcity's differential privacy metrics.

Results: All methods generated datasets similar to the original, confirmed by low nearest-neighbor distances and Kolmogorov–Smirnov tests. Linear Regression yielded the lowest reidentification risk and CART the highest. TVAE synthetic data with optimized hyperparameters achieved the highest AUROC. F1 scores were highest for XGBoost and Random Forest with CART-generated data, and for Logistic Regression with TVAE-generated data. For the survival models, CART data gave the highest Concordance-Index with XGB Survival, while TVAE performed best with Cox PH and Random Forest. SHAP and permutation importance analyses identified similar key risk drivers across synthetic and original datasets.

Conclusions: Synthetic data generation with advanced statistical and AI methods shows strong promise for privacy-preserving healthcare applications. TVAE with tuned hyperparameters demonstrated an optimal balance between predictive performance and privacy, followed by CART. These findings warrant validation using larger, more diverse datasets to ensure generalizability.

Legal entity responsible for the study: Daiichi Sankyo.

Funding: Daiichi Sankyo.

Disclosure: T. Qaiser, M. Rahman, S. Zakharkin: Full or part-time Employment: Daiichi Sankyo.

https://doi.org/10.1016/j.esmorw.2025.100609

415P How to generate open source annotated cancer clinical datasets with LLMs to support the development of smaller language models

E. Kempf[1], A.T. Vu[2], E. De La Clergerie[3], R. Flicoteaux[4]

[1]Medical Oncology, Centre Hospitalier Universitaire Henri-Mondor AP-HP, Creteil, France; [2]Medical Oncology, Assistance Publique Hôpitaux de Paris, Creteil, France; [3]Almanach, French National Institute for Research in Digital Science and Technology (INRIA), Paris, France; [4]Medical Information, Assistance Publique - Hopitaux de Paris, Paris, France

Background: The relevant clinical information in patient records is in unstructured free text. While LLMs show potential for automatic extraction, a limitation in the medical field is the lack of high-quality annotated datasets for training. Our objective was to generate open source annotated clinical datasets to support the development of smaller models for extraction of a minimal cancer dataset.

Methods: We designed complex prompts to generate realistic synthetic hospitalization reports, which included: (1) randomly sampled clinical characteristics from the French national claims database (cancer site, comorbidities, treatment modalities); (2) cancer-specific guidelines based on histology and stage; (3) randomized administrative details (patient and physician names, hospital, admission dates). Prompts were defined using LLMs and medical knowledge, and then iteratively refined by an oncologist to ensure internal consistency. The model was instructed to output both the narrative report and corresponding structured annotations. Hospitalization reports were generated with Mistral Large. An assessment study compared real and synthetic reports: 100 reports were independently reviewed by 2 physicians (oncologist and medical information specialist) and rated (1–10 scale) across 5 domains: language quality, medical consistency, completeness, conciseness, overall impression, likelihood of AI authorship.

Results: AI-generated reports achieved excellent ratings for language quality (mean 9.3 vs 9.2 for clinician-written reports) and overall impression (9.3 vs 9.2). Synthetic reports tended to include more extensive medical details, though sometimes at the expense of conciseness. Medical consistency was lower for AI-generated reports (7.9 vs 9.3), reflecting occasional clinical inconsistencies (e.g., surgical procedures in non-eligible patients). Structured annotations were of very high quality and closely matched the instructions.

Conclusions: LLMs can generate medical documents of near-human quality while simultaneously providing structured annotations linked to the narrative text. Medical inconsistencies may be addressed through prompt engineering.

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.2025.100611