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

Emmanuelle Kempf, A.T. Vu, Eric Villemonte de La Clergerie, Rémi Flicoteaux

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