64

Synthetic data generation using generative AI to support biomedical innovation: A health policy perspective

Rachele Hendricks-Sturrup and Maryam Nafie
Duke University, Duke-Margolis Institute of Health policy

OBJECTIVES/GOALS: Synthetic data holds potential for inclusion in medical product development pipelines. Therefore, we explored the current regulatory and practice landscape to identify best practices used to ensure, and communicate to key stakeholders, synthetic data quality, relevance, and reliability in regulatory settings. METHODS/STUDY POPULATION: We identified areas in which synthetic data created using generative AI holds the most value for health researchers. We reviewed regulatory documents, published literature, and expert insights to examine how regulators currently use, define, apply, and govern synthetic data created using generative AI. Next, we identified data management tools, best practices, ethical considerations, and regulatory developments that are necessary for generating fit-for-use synthetic datasets. Using this information, we developed a risk-based credibility assessment framework that aligns with current governmental standards that can be useful for users of synthetic data derived from generative AI applications in regulatory settings. RESULTS/ANTICIPATED RESULTS: Synthetic data created using generative AI holds value for health researchers across four key areas: Acting as a 1) privacy-enhancing technology, 2) data science 'sandbox' for training and exploration, 3) mechanism to navigate legalities around data sharing and/or use, and 4) method to augment underrepresented subgroups in datasets. However, synthetic data raise ethical and legal concerns, particularly regarding privacy, consent, stakeholder engagement, and ownership. Regulators and health technology assessment bodies, including FDA, EMA, MHRA, and Canada's Drug Agency, are exploring synthetic data to supplement medical datasets, validate external control arms, enhance model performance, and inform regulatory decision-making. DISCUSSION/SIGNIFICANCE OF IMPACT: Our work underscores the need to continue cultivating transparent, ethical, and fit-for-use approaches to synthetic data generation using generative AI. Moving forward, effective synthetic data use and development requires a culture of learning and transparency among regulators, end users, and those involved in data generation and exchange.

65

Integrating pathology reports into multimodal machine learning models to predict thyroid cancer recurrence

Denise Lee[1], Steven Yen[2], Grace Deng[3], Yehia Elkersh[2], David S Matteson[3], Andrew Thomas[4], Michael L Marin[2] and Girish N Nadkarni[5]
[1]Icahn School of Medicine at Mount Sinai; [2]Department of Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [3]Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA; [4]Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA and [5]Windreich Department of Artificial Intelligence, Icahn School of Medicine at Mount Sinai, New York, NY

OBJECTIVES/GOALS: The objective of this study was to evaluate the performance of multimodal machine learning (ML) models trained to predict differentiated thyroid cancer (DTC) recurrence using clinical data combined with novel natural language processing (NLP) derived features extracted from patient cytopathology and surgical pathology reports. METHODS/STUDY POPULATION: This was a retrospective study of adult thyroid cancer patients treated at an academic medical center. Patients were classified as having cancer recurrence or no recurrence. NLP features were extracted from cytopathology and surgical pathology reports using Term Frequency–Inverse Document Frequency (TF-IDF), latent Dirichlet allocation (LDA), and a zero-shot large language model (LLM) classification. 5 multimodal ML models were trained to predict cancer recurrence utilizing a combination of NLP and LLM features and clinical variables. Model performance was evaluated using area under the receiver operating characteristic curve (ROC-AUC) and precision recall area under the curve (PR-AUC). The top performing model was optimized with a 5-fold cross-validation. Feature importance was calculated. RESULTS/ANTICIPATED RESULTS: 480 patients with differentiated thyroid cancer diagnosed on surgical pathology were included in this study. The baseline model (clinical variables only) had a F1-score of 0.52 and an AUC of 0.53. The optimized gradient boosting model utilizing all features (EMR, LDA, TF-IDF, and LLM) had a F1-score of 0.87 and an AUC of 0.86. Topic words and themes from the patient cytopathology and surgical pathology reports were generated using LDA. Topic themes in cytopathology reports include malignancy, lymph node evaluation, and molecular testing. Topic themes in surgical pathology reports include histologic subtype, orientation of nodule, and intraoperative biopsy. The LDA themes of malignancy and histologic subtype ranked the highest in terms of feature importance. DISCUSSION/SIGNIFICANCE OF IMPACT: Multimodal models utilizing novel NLP features derived from unstructured pathology reports may enable improved prediction of recurrence in patients with DTC. Our optimized model demonstrated that 4 of the top 6 highest features were LDA topics. Topic modeling may be a valuable tool to extract relevant information from unstructured clinical notes.

66

Piloting an AI-driven conversational agent for social needs screening and referral in pediatric primary care

Emre Sezgin[1], Daniel Jackson[2], Amad Hussain[2], Faith Logan[2], Elizabeth Clarkson[2] and Joshua Stokes[2]
[1]Ohio State University and [2]Nationwide Children's Hospital

OBJECTIVES/GOALS: To evaluate the feasibility, acceptability, and usability of DAPHNE, an AI-powered chatbot for health-related social needs screening and referral in pediatric primary care, and to assess feasibility of study procedures to inform a future large-scale trial.

https://doi.org/10.1017/cts.2026.10291 Published online by Cambridge University Press