.

Large Language Models forecast Patient Health Trajectories enabling Digital Twins

* Digital twins, developed using generative AI, enable virtual patient representations to predict health trajectories, aiding treatment selection and clinical trial design, overcoming challenges like missingness and noise in real-world data.
* a digital twin for healthcare as a virtual representation of a person which allows dynamic simulation of potential treatment strategy, monitoring and prediction of health trajectory, and early
* The DT-GPT model, utilizing biomedical LLMs and EHR data, predicts clinical variables and provides preliminary explanations.
* Benchmarked on US NSCLC and ICU datasets.
* DT-GPT outperformed traditional machine learning methods in patient trajectory forecasting, achieving 3.4% MAE improvement in both long-term and short-term datasets.
* It also preserved clinical variables' cross-correlations, handled data missingness and noise, and provided forecast insights.
* Clinical forecasting involves predicting patient-specific health outcomes and clinical events over time, which is of paramount importance for patient monitoring, treatment selection and drug development.
* The application of digital twins is poised to revolutionize healthcare in areas such as precision medicine, predictive analytics, virtual testing, continuous monitoring, and enhanced decision support.
* Generative AI methods for predicting patient trajectories include recurrent neural networks, transformers and stable diffusion.5-9 These often fall short in terms of handling missing data, interpretability and performance. These challenges can be partially addressed by causal machine learning, but these algorithms face limitations related to small datasets or being confined to simulations.
* Recent advancements in generative AI include foundation models for patient forecasting, which focus on single-point predictions, and text-focused Large Language Models (LLMs), which demonstrate forecasting capabilities, including zero-shot forecasting, highlighting their generalizability.
* We propose the creation of digital twins based on LLMs that leverage data from electronic health records
* EHRs are crucial for training machine learning models in healthcare, but they face challenges like data heterogeneity, rare events, sparsity, and quality issues. Adapting models can increase complexity and introduce additional assumptions.
* We hypothesize that LLMs will empower digital twins and overcome these challenges. Here, we introduce the Digital Twin - Generative Pretrained Transformer (DT-GPT) model (Fig. 1), which enables: i) forecasting of clinical variable trajectories, ii) zero-shot predictions of clinical variables not previously trained on, and iii) preliminary interpretability utilizing chatbot functionalities.
* We analyze the performance of the model by forecasting laboratory values on both a long-term scale (up to 13 weeks) for non-small cell lung cancer (NSCLC) patients, as well as a short-term scale (next 24 hours) for Intensive Care Unit (ICU) patients.
* How it works?
* The LLM-based DT-GPT framework enables forecasting patient trajectories, identifying key variables, and zero-shot predictions. Here exemplified, a) a sparse synthetic patient timeline, which b) DT-GPT utilizes for generating longitudinal clinical variable forecasts, e.g., c) neutrophil and d) hemoglobin blood levels. DT-GPT can e) chat and respond to inquiries about important variables, as well as f) perform zero-shot forecasting on clinical variables previously not used during training
* two independent datasets, namely long-term and short-term trajectories of non-small cell lung cancer (NSCLC) and intensive care unit (ICU) patients
* he US-based NSCLC dataset, we used the nationwide Flatiron Health EHR-derived de-identified database.
* The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction.18,19 During the study period, the de-identified data originated from approximately 280 cancer clinics
* The study analyzed 16,496 NSCLC patients from 1991 to 2023, primarily from community oncology settings. The data was grouped weekly based on the last observed value, focusing on 50 common diagnoses and 80 common laboratory measurements.
* The study divided NSCLC patients into input and output segments based on therapy start dates, aiming to predict weekly values of hemoglobin, leukocytes, lymphocytes/leukocytes, neutrophils, and lactate dehydrogenase up to 13 weeks after therapy start date, reflecting key NSCLC treatment response characteristics.
* The study used the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset to analyze ICU trajectories and predict patient's future lab variables. The dataset included 300 input variables from 35,131 patients, including O2 saturation, pulse oximetry, respiratory rate, and magnesium, which were chosen due to their clinical relevance and high temporal variability.
* The study uses datasets from NSCLC and ICU to encode and interpret an LLM, BioMistral, for trajectory forecasting, evaluating output, and exploring zero-shot predictions through a chat interface.
* h datasets were randomly split at the patient level into 80% training, 10% validation, and 10% test se
* The study encoded patient trajectories using templates that converted medical histories from EHRs into a text format compatible with LLMs. The input template consisted of patient history, demographic data, forecast dates, and prompt. Output trajectories were encoded using a manually developed template and JSON-format encoding.
* The biomedical LLM BioMistral was fine-tuned using standard cross entropy loss, and 30 predictions were made for each patient sample.
* The DT-GPT model was used to create a chatbot based on patient histories for prediction explanation and zero-shot forecasting. The chatbot generated forecasting results and added task-specific prompts for prediction explanation and zero-shot forecasting, allowing for new clinical variables.
* Five multi-step, multivariate baselines were used, including a naïve model, linear regression, time series LightGBM, Temporal Fusion Transformer, and TiDE models, chosen for their ability to handle future variables.
* The study evaluated patient trajectories using mean absolute error (MAE) as a primary metric. Randomly sampled 200 patients, chatbot exploration and zero-shot forecasting were analyzed on the entire test set, evaluating the effects of RWD missingness
* he models' performance is evaluated using the Mean Absolute Error (MAE), which measures the average error between the predicted and actual values. The MAE is normalized by the standard deviation to make the comparison fairer. A lower MAE indicates better performance, meaning the model's predictions are closer to the actual values. The model named "DT-GPT" performed better than other models in most cases, as indicated by the bold highlighting in the table.
* DT-GPT outperformed LightGBM in both NSCLC and ICU datasets, with an average mean absolute error of 0·55 ± 0·04 and 0·57 ± 0·05 respectively, indicating a relative improvement of 3·4% and 1.3% respectively.
* DT-GPT predicted variables with an R2 of 0·98 and 0·99, outperforming LightGBM on NSCLC and ICU datasets, and showing consistent improvement across time.
* DT-GPT is a flexible and robust machine learning method that exhibits desired properties in various ablation studies. It is competitive with baselines, can handle increased input missingness, and is stable to misspellings in the input. Its performance degradation only shows after 25 misspellings per patient sample, unlike most established machine learning methods.
* DT-GPT's conversational capability was used for forecasting tasks, allowing user interaction and prediction reasoning. Ten predicted trajectories were generated for each patient sample, with explanatory variables extracted from 25,575 out of 27,730 responses. The most influential variables were therapy, ECOG status, and leukocyte count. Therapy significantly influenced hemoglobin dynamics, with immunotherapy and targeted therapy resulting in higher levels over time.
* The study compared zero-shot DT-GPT with a supervised LightGBM model, which was trained on over 13,000 patients. Despite being a fully supervised model, zero-shot DT-GPT outperformed LightGBM on 13 out of 69 non-target variables, indicating improved performance in target variables.