**Task**

**Please prepare a concise document summarizing your general understanding of ideas and approach from the given sample papers and how Digital Twins are leveraged in AI-driven Drug Discovery and Clinical Trials.**

**In your submission, we expect you to:**

**1. Explain your approach and clearly outline your understanding/ideas.**

**2. Include concepts, frameworks, and tools you will use to enhance Drug Trial Outcome Prediction based on your understanding from the papers.**

**3. Explore knowledge graphs and multi-agent AI workflows to enhance the process.**

**4. Visualize your approach using diagrams or flowcharts, where possible.**

**Large Language Models forecast Patient Health Trajectories enabling Digital Twins**

**Abstract:**

This paper presents the DT-GPT model, a digital twin powered by generative AI, to predict patient health outcomes using electronic health records The model excels in forecasting clinical variables, even for data it hasn't been trained on, and provides explanations for its predictions through a chatbot interface. Tested on cancer and ICU patient data, DT-GPT outperforms traditional machine learning models, showing significant improvements in prediction accuracy. This approach could revolutionize healthcare by enhancing patient monitoring, treatment decisions, and clinical trial design.

**Introduction:**

**A diagram of a human body

Description automatically generated**

Fig 1. Digital twin in healthcare

In recent years, the concept of digital twins has gained traction in various fields, including healthcare. Digital twins is the virtual representation of a person which allows dynamic simulation of potential treatment strategy, monitoring and prediction of health trajectory, and early. These were developed using generative AI, offer a powerful tool for predicting patient health trajectories, aiding in treatment selection, and improving the design of clinical trials.

However, there were various challenges such as missing or noisy data in real-world healthcare settings can complicate these tasks.This study introduces the Digital Twin - Generative Pretrained Transformer model, which utilizes large biomedical language models (LLMs) and data from electronic health records to forecast clinical variables and provide preliminary explanations. This model not only improves the accuracy of patient trajectory predictions but also addresses the challenges of data missingness and noise while preserving the relationships between clinical variables.The DT-GPT model marks a significant advancement in the application of AI in healthcare, with the potential to revolutionize precision medicine, predictive analytics, and patient monitoring through its ability to generate accurate, interpretable predictions based on complex and diverse clinical data.

**Problem Definition:**

The main problem this study addresses is the challenge of accurately predicting patient health outcomes using data from electronic health records. Traditional methods struggle with issues like missing or noisy data, and they often can't handle the complexity of patient information effectively. This study aims to solve these issues by developing a new model, DT-GPT, that uses advanced AI techniques to make better predictions about a patient's future health, even when data is incomplete or difficult to interpret.

**Methodology:**

Fig 2.Flowchart of methodology

**Implementation:**

1. **Dataset Collection**

The dataset collection for the DT-GPT model involves two main sources:

1. **NSCLC Dataset (Non-Small Cell Lung Cancer):**

The data is derived from the Flatiron Health EHR database, which is a longitudinal database containing de-identified patient-level data. This data is collected from approximately 280 cancer clinics across the United States.The dataset includes structured and unstructured data from 16,496 NSCLC patients spanning from 1991 to 2023. It contains detailed information on 50 common diagnoses and 80 laboratory measurements, such as hemoglobin levels, leukocyte counts, and neutrophil levels.The data is grouped weekly based on the last observed value, and it is used to predict key variables for up to 13 weeks after the start of therapy, which is crucial for assessing treatment response.

1. **ICU Dataset:**

The data is taken from the Medical Information Mart for Intensive Care database, which is a publicly available dataset containing de-identified health data from patients in intensive care units .This dataset includes data from 35,131 ICU patients, with around 300 input variables such as oxygen saturation, respiratory rate, pulse oximetry, and magnesium levels. These variables are selected for their clinical relevance and high temporal variability. The data is used to forecast short-term clinical variables (next 24 hours) and to assess the future lab values of ICU patients, providing insights into patient outcomes and treatment efficacy.

1. **Implementation**

Fig 3.Training Model

* The method used in the DT-GPT model involves a combination of advanced machine learning techniques.
* The model leverages the GPT's capacity to process sequential data and predict future values by understanding the patterns and relationships within the data.
* The DT-GPT model is fine-tuned using BioMistral, a large biomedical language model that has been trained on vast amounts of medical literature and clinical data.
* BioMistral is adapted to predict clinical variables, handle missing data, and maintain the relationships between different variables over time.
* The model is designed to make predictions on clinical variables it was not explicitly trained on.This is achieved through zero-shot learning, where the model leverages its understanding of language and data patterns to generalize to new tasks without additional training.
* The DT-GPT model includes a chatbot interface that allows users to interact with the model. Users can ask questions about specific patient data, request predictions, and receive explanations for the model’s decisions.
* The model uses templates to encode patient data from electronic health records into a format compatible with the GPT architecture..
* Both short-term (ICU data) and long-term (NSCLC data) patient trajectories are used to train the model.
* The model is fine-tuned using cross-entropy loss, a common loss function for classification tasks, allowing the model to learn from the data and improve its predictions over time.
* The fine-tuning process involves multiple iterations where the model's parameters are adjusted to minimize the error between predicted and actual clinical values.
* The performance of DT-GPT is benchmarked against traditional machine learning models like LightGBM, linear regression, and time series models.
* The primary evaluation metric used is Mean Absolute Error (MAE), which measures the average difference between the model's predictions and the actual values.
* The DT-GPT model is used to forecast patient trajectories, such as predicting the progression of diseases or the likely outcomes of treatments over time.

1. **Model selection and building**

Various models are deployed like:

1. LightGBM
2. Linear Regression
3. Naïve Bayes Model
4. Temporal Fusion Transformer
5. TiDE (Temporal Dependency Estimation)
6. Time Series LightGBM

**Results**

A number of conventional machine learning models, such as LightGBM, linear regression, and several time series models, were used to compare the performance of the DT-GPT model. In comparison to LightGBM, the DT-GPT model showed better predictive accuracy, obtaining a Mean Absolute Error (MAE) increase of 3.4% in long-term trajectory forecasting for patients with non-small cell lung cancer (NSCLC) and a 1.3% improvement for patients in the intensive care unit (ICU). The NSCLC and ICU datasets showed that the DT-GPT model maintained high R values of 0.98 and 0.99, respectively, suggesting a strong connection between projected and actual values. Furthermore, the model maintained the cross-correlations between clinical variables while handling data problems such missingness and noise successfully.

**Discussion**

To enhance drug trial outcome prediction, several concepts, frameworks, and tools can be employed based on insights from the papers discussed. These include:

* Implement generative AI models, such as transformers, to analyze patient data and generate realistic synthetic patient profiles
* Integrate causal machine learning methodologies to address challenges related to confounding factors and missing data.
* Employ ensemble methods to combine predictions from multiple models.
* Use time series forecasting techniques to predict patient outcomes over different time frames
* Implement advanced data cleaning and preprocessing techniques to handle data heterogeneity and sparsity in EHRs.
* Establish robust validation frameworks, including metrics like Mean Absolute Error (MAE), Precision, Recall, and F1-score should be used to evaluate prediction accuracy.

**Conclusion**

A notable development in predictive modelling for healthcare is the Digital Twin - Generative Pretrained model, which efficiently uses electronic health records (EHRs) to predict clinical variables and improve patient outcomes. The DT-GPT model demonstrated its capacity to manage issues like noise and missing data by outperforming conventional machine learning techniques in its robust prediction of patient trajectories. The model's practical relevance in clinical settings is further highlighted by its ability to deliver predictions with 0% chance and its connection with a chatbot for interactive insights. The potential of the DT-GPT model to enhance patient care by supporting more precise and customised treatment plans could revolutionise healthcare procedures and aid physicians in making decisions.