medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Large Language Models forecast Patient Health Trajectories enabling Digital Twins

Nikita Makarov[1,2,3,†], Maria Bordukova[1,2,3,†],
Raul Rodriguez-Esteban[4,#], Fabian Schmich[1,#], Michael P. Menden[2,5,#]

Abstract

Background: Generative artificial intelligence (AI) facilitates the development of digital twins, which enable virtual representations of real patients to explore, predict and simulate patient health trajectories, ultimately aiding treatment selection and clinical trial design, among other applications. Recent advances in forecasting utilizing generative AI, in particular large language models (LLMs), highlights untapped potential to overcome real-world data (RWD) challenges such as missingness, noise and limited sample sizes, thus empowering the next generation of AI algorithms in healthcare.

Methods: We developed the Digital Twin - Generative Pretrained Transformer (DT-GPT) model, which leverages biomedical LLMs using rich electronic health record (EHR) data. Our method eliminates the need for data imputation and normalization, enables forecasting of clinical variables, and prediction exploration via a chatbot interface. We analyzed the method's performance on RWD from both a long-term US nationwide non-small cell lung cancer (NSCLC) dataset and a short-term intensive care unit (MIMIC-IV) dataset.

Findings: DT-GPT surpassed state-of-the-art machine learning methods in patient trajectory forecasting on mean absolute error (MAE) for both the long-term (3.4% MAE improvement) and the short-term (1.3% MAE improvement) datasets. Additionally, DT-GPT was capable of preserving cross-correlations of clinical variables (average R^2 of 0.98), and handling data missingness as well as noise. Finally, we discovered the ability of DT-GPT both to provide insights into a forecast's rationale and to perform zero-shot forecasting on variables not used during the fine-tuning, outperforming even fully trained, leading task-specific machine learning models on 14 clinical variables.

Interpretation: DT-GPT demonstrates that LLMs can serve as a robust medical forecasting platform, empowering digital twins that are able to virtually replicate patient characteristics beyond their training data. We envision that LLM-based digital twins will enable a variety of use cases, including clinical trial simulations, treatment selection and adverse event mitigation.

[1] Data & Analytics, Pharmaceutical Research and Early Development, Roche Innovation Center Munich (RICM), Penzberg, Germany
[2] Computational Health Center, Helmholtz Munich, Munich, Germany
[3] Department of Biology, Ludwig-Maximilians University Munich, Munich, Germany
[4] Data & Analytics, Pharmaceutical Research and Early Development, Roche Innovation Center Basel (RICB), Basel, Switzerland
[5] Department of Biochemistry and Pharmacology, University of Melbourne, Melbourne, Australia

[†] Equal contribution
[#] Correspondence: raul.rodriguez-esteban@roche.com, fabian.schmich@roche.com, michael.menden@helmholtz-munich.de

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1. Introduction

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 [1]. Digital twins are virtual representations of patients that leverage a patient's medical history to generate detailed multi-variable forecasts of future health states [2]. 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 [3].

Generative artificial intelligence (AI) holds promises for creating digital twins due to its potential to produce synthetic yet realistic data, but this area of application is still in its infancy [4]. Generative AI methods for predicting patient trajectories include recurrent neural networks [5,6,7,8], transformers [9,10] and stable diffusion [11]. These often fall short in terms of handling missing data, interpretability and performance. The challenges are partially addressed by causal machine learning [12,13,14], however these algorithms face limitations related to small datasets or being confined to simulations [15].

Recent breakthroughs in generative AI have been achieved with foundation models, which are pre-trained AI models adaptable for various specific tasks involving different types of data. Most foundation models for patient forecasting, e.g. EHRShot [43], focus on single-point predictions [16] rather than comprehensive longitudinal patient trajectories, which are needed for clinical decision-making. Less explored remain text-focused Large Language Models (LLMs), which have demonstrated forecasting capabilities [17,18], including the ability of zero-shot forecasting, i.e. forecasting without any prior specific training in the task [19,20], thus highlighting their remarkable generalizability.

We propose the creation of digital twins based on LLMs that leverage data from electronic health records (EHRs). EHRs are a key source of training data for machine learning models in healthcare [21], as they record patient characteristics such as demographics, diagnoses and lab results over time. However, they pose specific challenges such as data heterogeneity, rare events, sparsity and quality issues [16]. There have been developments in machine learning to overcome these challenges, especially for data sparsity [8,11], usually by adapting the model's architecture, resulting in increased model complexity and the introduction of further assumptions on the data.

We hypothesize that LLMs will empower digital twins and overcome the above outlined challenges of patient trajectory forecasting. 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 unseen clinical variables, and iii) preliminary interpretability utilizing chatbot functionalities. We analyze the performance of the model by forecasting laboratory values on both a long-term (up to 13 week) scale for non-small cell lung cancer (NSCLC) patients, as well predicting short-term (next 24 hours) values for ICU patients. We anticipate that DT-GPT will pave the way for AI-based digital twins in healthcare.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

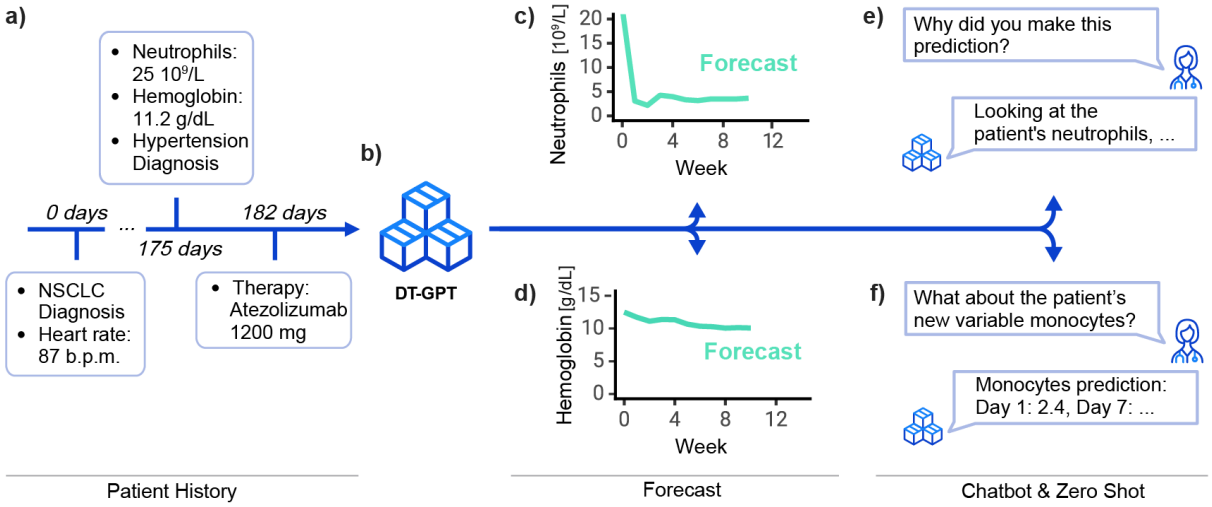

Figure 1: The LLM-based DT-GPT framework enables forecasting patient trajectories, identifying key covariates, and zero-shot predictions. Here exemplified, **a**) a sparse synthetic patient timeline, which **b**) DT-GPT utilizes for generating longitudinal clinical variable forecasts, e.g., **c**) neutrophil and **d**) hemoglobin blood levels. DT-GPT can **e**) chat and respond to inquiries about important covariates, as well as **f**) perform zero-shot forecasting on clinical variables previously not used during training.

2. Methods

DT-GPT is a method that utilizes pre-trained LLMs fine-tuned on clinical data (**Fig. 2a)**. Notably, DT-GPT is agnostic about the underlying LLM and can be applied without architectural changes to any general-purpose or specialized text-focused LLM.

2.1 Datasets and data preparation

We trained and evaluated DT-GPT for forecasting patients' laboratory values across two independent datasets, namely long-term and short-term trajectories of non-small cell lung cancer (NSCLC) and intensive-care unit (ICU) patients, respectively. For the US-based NSCLC dataset, we used the nationwide Flatiron Health EHR-derived de-identified database. The data are de-identified and subject to obligations to prevent re-identification and protect patient confidentiality. The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction [22,23]. During the study period, the de-identified data originated from approximately 280 cancer clinics (~800 sites of care).

The study included 16,496 patients diagnosed with NSCLC from 01 January 1991 to 06 July 2023. The majority of patients in the database originate from community oncology settings; relative community/academic proportions may vary depending on study cohort. Patients with a birth year of 1938 or earlier may have an adjusted birth year in Flatiron Health datasets due to patient de-identification requirements. To harmonize the data, we aggregated all values in a week based on the last observed value. We focused on the 50 most common diagnoses and 80 most common laboratory measurements, complemented by the Eastern Cooperative Oncology Group (ECOG) score, metastases, vitals, drug administrations, response and mortality covariates totaling 773,607 patient-days across 320 variables.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

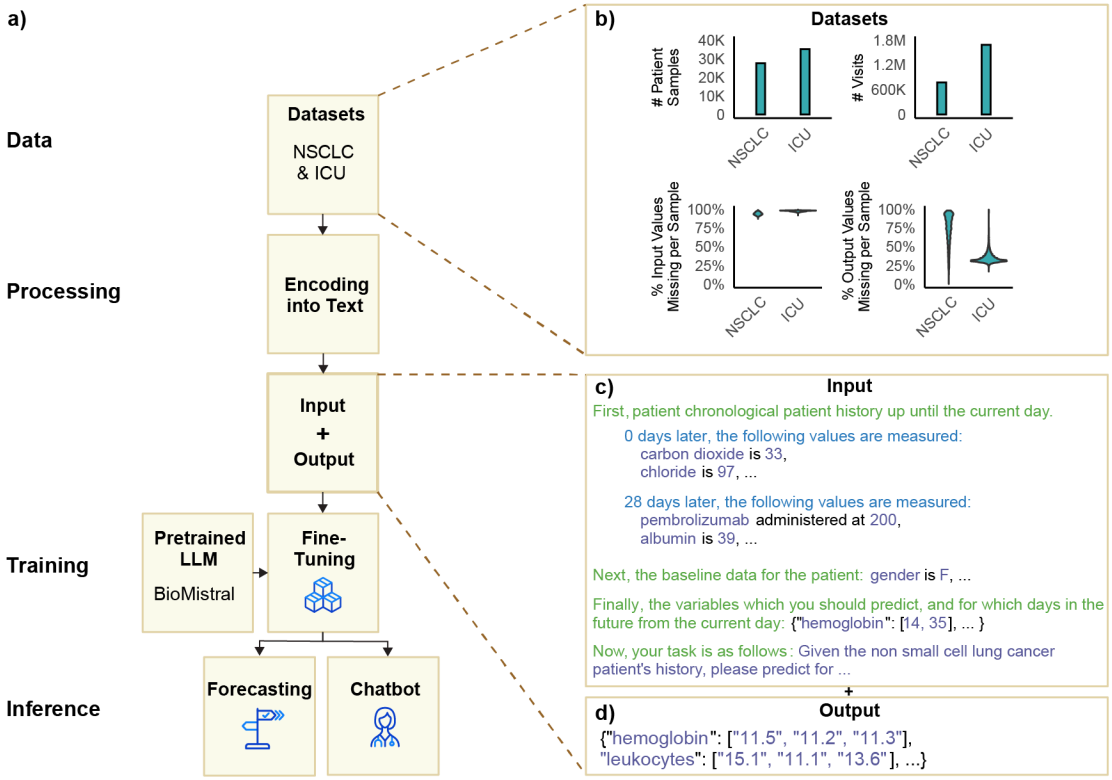

Figure 2: The DT-GPT framework transforms EHRs into text and subsequently fine-tunes an LLM on this data. a) Overview of the pipeline: datasets are split and encoded into input/output text based on landmark timepoints, then used to fine-tune an LLM, e.g. BioMistral. The model output is evaluated for trajectory forecasting whilst zero-shot predictions and feature importances are explored via a chat interface. **b)** Sample size, visit frequency, and missing data for each dataset. **c)** Input and **d)** output encoded examples, emphasizing the chronological encoding of observations.

For every NSCLC patient, we split their trajectory into input and output based on the start date of each line of therapy to create each patient sample. All variables up to the start date were considered as input. The task was to predict the weekly values of the following variables up to 13 weeks after the start date: hemoglobin, leukocytes, lymphocytes/leukocytes, lymphocytes, neutrophils, lactate dehydrogenase. The variables were selected because they were measured frequently and reflect important NSCLC treatment response characteristics **(Appendix A1)**.

To highlight the generalizability of DT-GPT, we analyzed intensive-care unit (ICU) trajectories from the publicly-accessible Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset [24]. We employed an established processing pipeline [25], resulting in 300 input covariates across 1,686,288 time points from 35,131 patients. Here, the task was to predict a patient's future hourly lab variables given their first 24 hours in the ICU. Specifically, the patient history was considered as the first 24 hours for all variables, and the task was to forecast the future 24 hourly values for the following variables: O2 saturation pulse oximetry, respiratory rate and magnesium. These were selected for being clinically relevant, having the top three variables with the highest variation across time, and with at least 50% of patients having at least one measurement. The resulting variables present both an increased forecasting challenge, as well as have wide representation across the patient population.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Both datasets were randomly split at the patient level into 80% training, 10% validation, and 10% test set. Thus, each set comprised disjoint sets of patients to avoid data leakage. The test sets were solely used for final evaluation and estimating generalizability (**Fig. 2b**; **Appendix A1)**.

2.2 Encoding

We encoded patient trajectories by using templates that converted their medical histories based on EHRs into a text format that the LLM could ingest, similar to [18] and [19] (**Fig. 2c,d**; **Appendix A4)**. The input template was structured into four components: 1) Patient history, 2) demographic data, 3) forecast dates and 4) prompt. The key component was the patient history, containing a chronological description of patient visits without including any missing values. The output trajectories were encoded based on templates containing only the respective output variables for the forecasted time points. We used a manually developed template for the input encoding and a JSON-format encoding for the output. Examples can be found in **Appendix A4** and **A10**.

2.3 LLMs and fine-tuning

For our experiments, we utilized the biomedical LLM BioMistral 7B DARE [26], since it is provided with an open source license and based on an established LLM. Furthermore, BioMistral is instruction tuned and through its biomedical specialization incorporates compressed representations of vast amounts of biomedical knowledge [26]. We further fine tuned the LLM using the standard cross entropy loss, masked so that the gradient was only computed on the output text. Following existing literature [20,27], we performed 30 predictions for each patient sample during evaluation, then took the mean of each time point as the final prediction. All hyperparameters used in fine-tuning are shown in **Appendix A5**.

2.4 Chatbot and zero-shot learning

We employed the DT-GPT model to run a chatbot on the patient history, allowing for prediction explanation and zero-shot forecasting of new variables. The process involved two steps: (1) running DT-GPT on patient history to generate forecasting results, and (2) concatenating a new task-specific prompt surrounded by the respective instruction indication tokens to the DT-GPT chat history and getting a response. In the prediction explanation task, the prompt asked for the most important variables influencing the predicted trajectory. In forecasting new variables, a prompt specifying the output format and days to predict was used. We note that DT-GPT was not optimized for these tasks and no response examples were given, i.e. zero-shot learning. Example prompts and chatbot interactions for both settings are provided in **Appendix A6** and **Fig. 5a,d**.

2.5 Baseline models

We employed five multi-step, multivariate baselines, ranging from a simple baseline to state-of-the-art forecasting models. Specifically, we used a naïve model that repeats the last observed value, a linear regression model, a time series LightGBM model, a Temporal Fusion Transformer [10,28] as well as a TiDE model [29]. These models were selected due to being able to handle future covariates and have been shown to achieve state-of-the-art results in both medical and standard time series forecasting [30,31]. The hyperparameters and training details are shown in **Appendix A7**.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

2.6 Evaluation

For evaluation of the forecasted patient trajectories, we used the mean absolute error (MAE) as our primary metric. We first standardized and calculated the pairwise error between the forecasted value and the true value of a given sample and time step, averaged over all patients and timepoints and then averaged again over all variables [29]. All error bars refer to the standard error [32] of the model aggregated across all variables. For the evaluation of the effects of RWD missingness, misspellings as well as chatbot exploration and zero shot forecasting, we use a randomly selected 200 patient sample subset of the test set.

3. Results

DT-GPT achieved state-of-the-art forecasting performance, being stable to common RWD challenges and able to forecast zero shot lab variables. Additionally, we explored how DT-GPT was able to provide preliminary insights into important features used in its predictions.

3.1 DT-GPT achieved state-of-the-art forecasting performance

DT-GPT achieved the lowest mean absolute error (MAE) on both the NSCLC and ICU datasets. On the NSCLC dataset, DT-GPT achieved an average MAE of 0.55 ± 0.04, whilst LightGBM, the second best model, achieved an average MAE of 0.57 ± 0.05 (**Fig. 3a**), showing a relative improvement of 3.4%. On the ICU dataset, DT-GPT achieved an average MAE of 0.59 ± 0.03, whilst the second best model, LightGBM, performed at 0.60 ± 0.03, equivalent to a 1.3% improvement (**Fig. 3b**; **Appendix A8**).

DT-GPT forecasts preserved inter-variable relationships (**Fig. 3c**). The correlations between the variables forecasted by DT-GPT aligned with the correlations of the variables in the test datasets with an R^2 of 0.98 and 0.99, whilst those of LightGBM achieved an R^2 of 0.97 and 0.99, on the NSCLC and ICU datasets respectively. We also saw that DT-GPT outperformed LightGBM in the majority of timepoints in both datasets, demonstrating that the improvement was consistent across time (**Fig. 3d,e**).

DT-GPT can be further improved by exploring alternative trajectory aggregation methods. To inspect both low and high MAE predictions from DT-GPT, we visualized two sample forecasts for the variable neutrophils (**Fig. 3f,g**) picked from both the low and high end of the distribution (**Fig. 3h**). Note that the final prediction was derived by averaging 30 generated trajectories and that, even in poor performing cases, individual non-averaged forecasted trajectories were sometimes able to capture parts of the true trajectory well. To explore the importance of trajectory aggregation, we calculated the error given an optimal aggregation. To this end, we selected the individual trajectories with the lowest MAE and recalculated the hypothetical MAE on the NSCLC dataset, achieving a 26% improvement in error to 0.40 ± 0.02, without any further model training. Note that this is a theoretical lower bound. Finally, we noted that in the distribution of MAE for neutrophils across all patients, most of the errors were left skewed, indicating that the high errors came from a small number of outlier patients (**Fig. 3h**).

3.2 DT-GPT is robust to common RWD challenges

DT-GPT is flexible and robust to common practical data challenges, exhibiting desired properties in a variety of ablation studies on the NSCLC dataset. First, DT-GPT performance was competitive with

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

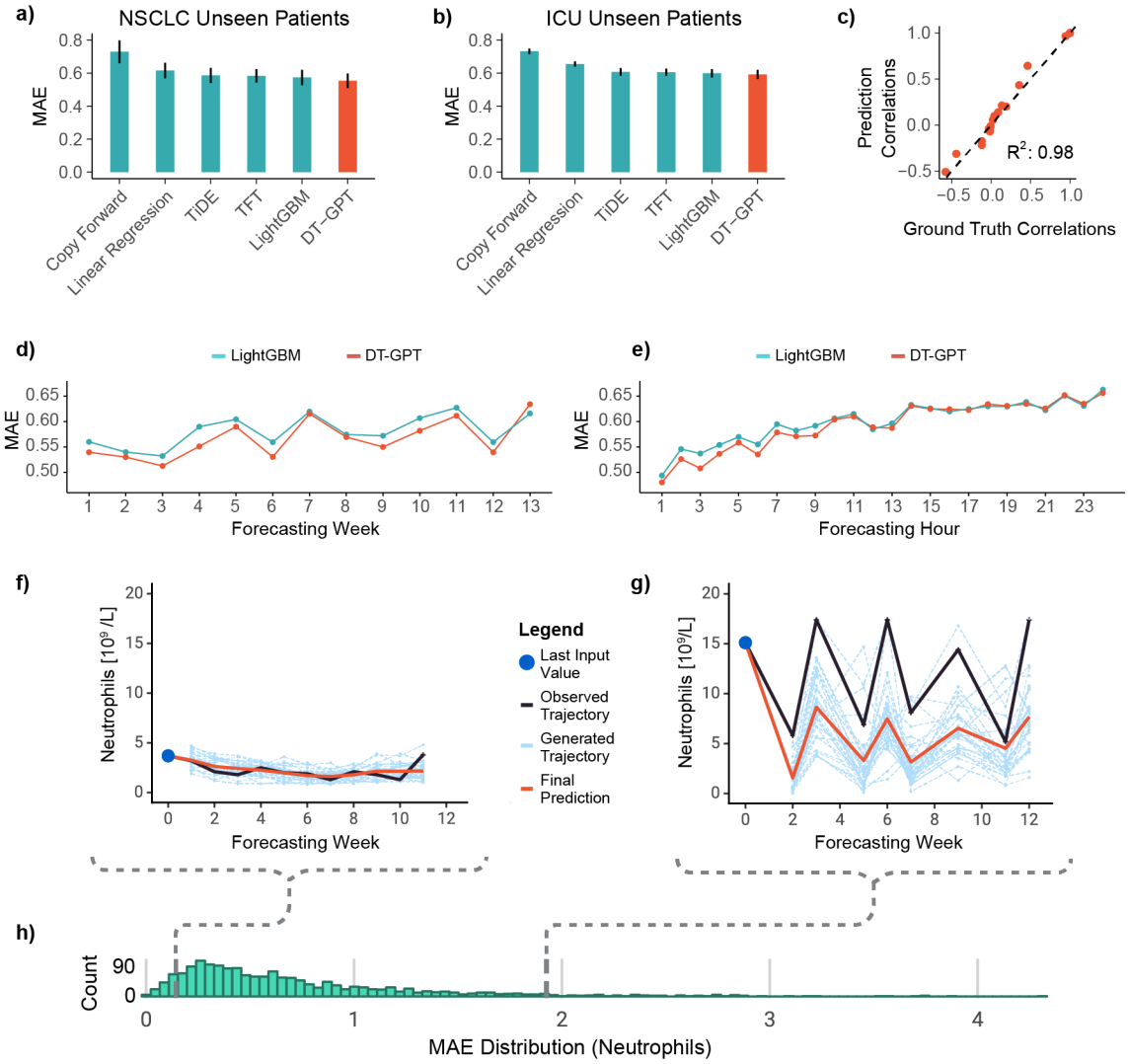

Figure 3: DT-GPT outperformed the baselines in the majority of cases on both the long-term NSCLC and the short-term ICU datasets. a) and **b)** show a comparison of all baseline models on the NSCLC and ICU datasets, respectively, with error bars showing the standard error across all variables. **c)** DT-GPT is able to capture inter-variable correlations, with each dot corresponding to a pair of variables (e.g. neutrophils and leukocytes), the x-axis being the correlations observed in the ground truth and the y-axis being the correlations in the predictions. **d)** and **e)** show a comparison of MAE versus forecasted time point for DT-GPT and LightGBM on the NSCLC and ICU datasets, respectively, with the x-axis showing relative time points and the y-axis the corresponding MAE. **f)** and **g)** portray forecasts with low and high error, respectively. **h)** Histogram of MAE distribution for the variable neutrophils.

baselines after training with data corresponding to 5,000 patients and it further improved with the number of patients in the training dataset (**Fig. 4a**; **Fig. 3a**). Additionally, DT-GPT could handle increased input missingness, with performance degradation only showing after more than 20% of the input is masked, on top of the 94.4% initial missingness of the NSCLC dataset (**Fig. 4b**). Thirdly, DT-GPT was stable to misspellings in the input, only significantly degrading in performance after 25 misspellings per patient sample (**Fig. 4c**). Note that misspellings cannot be handled by most established machine learning methods and either require completely dropping the value or manually curating the data.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

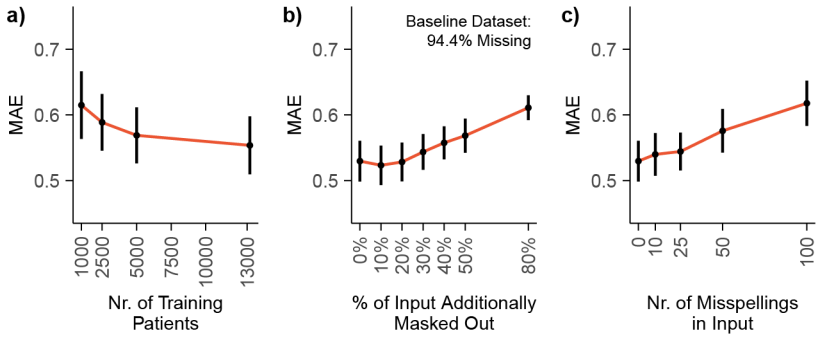

Figure 4: DT-GPT is robust to common RWD issues in the long-term NSCLC dataset. a) Mean absolute error (MAE) according to the number of patients in the training set. Assessing impact on MAE based on **b)** added missingness, on top of the baseline 94.4% missingness of the NSCLC dataset, and **c)** injected misspellings in the input.

3.3 DT-GPT enables prediction insights and zero-shot forecasting

DT-GPT preserved its chatting capability, supporting user interaction after fine-tuning on the forecasting task, and exhibiting preliminary interpretability. For each patient sample, we obtained 10 predicted trajectories and a set of variables explaining those predictions (**Fig. 5a**). We extracted explanatory variables from 1,885 out of 2,000 chatbot responses. The most important variables are line of therapy, ECOG and leukocytes (**Fig. 5b**). We observed consistent trends across chemotherapies and immunotherapies, which impact the white blood cell dynamics [33]. An extended list of the most important variables and baseline characteristics for forecasting is provided in **Appendix A11**.

Important variables are specific with respect to a patient and predicted patient trajectories. DT-GPT is a non-deterministic algorithm, thus multiple generations of a patient trajectory may lead to diverse results, however, these diverse predictions are driven by distinctly different variables and baseline characteristics. For instance, in the case of outlier trajectories as shown in an example in **Fig. 5c**, DT-GPT considered partially disjunct feature sets when predicting extremely lower and higher than average predicted values, e.g. ferritin and lactate dehydrogenase were associated with hemoglobin prediction, which is concordant with established literature [34,35].

DT-GPT supports zero-shot forecasting of 69 non-target clinical variables which are observed in patient medical histories, but were not subject to model fine tuning. In our experiments, we forecasted each non-target variable separately (**Fig. 5d**) and extracted 5,625 trajectories from 5,628 forecasting results. We compared the performance of zero-shot DT-GPT with a supervised LightGBM model, which was trained on each non-target variable separately using data from over 13,000 patients. Notably, LightGBM is a fully supervised model, thus is anticipated to perform better, whilst being unable to perform zero-shot predictions, and surprisingly, is outperformed in some instances by zero-shot DT-GPT.

Zero-shot DT-GPT outperforms the leading fully trained machine learning model on 14 out of 69 non-target variables (**Fig. 5e,f**). The variables with improved performance can be described as closely related to the target variables (**Fig. 5f**). For instance, *segmented neutrophils*, *band form neutrophils* and *neutrophils by automated count* have different LOINC codes from the trained variable (30451-9, 26507-4, 751-8 respectively), but are correlated with the target *neutrophils* variable (LOINC 26499-4).

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

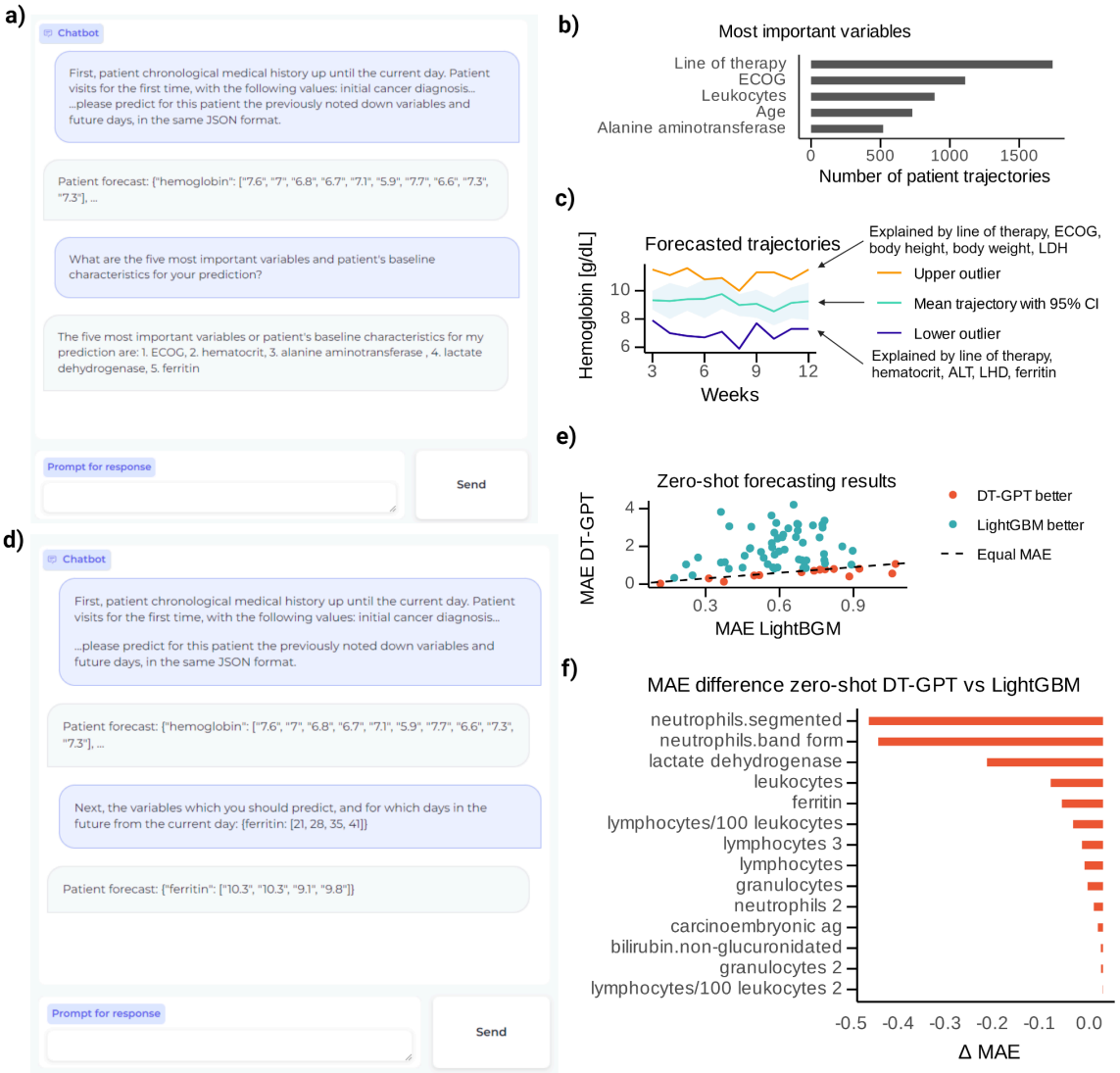

Figure 5: DT-GPT preserves its chatting capability after the fine-tuning, allowing preliminary explainability and zero-shot forecasting. a) Chatbot interaction example of explaining predictions. **b)** Five most important features and counts from forecasting 200 test patient samples with 10 predicted trajectories each. **c)** Different forecasted trajectories are explained by different sets of important variables: hematocrit and ferritin for low-level outlier (purple line) in hemoglobin trajectory, whilst ECOG, patient weight and height are important for the higher-value outlier (orange line). **d)** Chatbot interaction example for the forecasting of an unseen variable. **e)** DT-GPT outperforms LightGBM models on 14 out of 69 unseen non-target variables. We note that LightGBM models were trained on more than 13,000 patient data, whilst DT-GPT fully performs zero-shot predictions. **f)** DT-GPT is superior for variables related to the target variables used during fine tuning. (Abbreviations: ECOG - Eastern Cooperative Oncology Group performance status scale, LDH - Lactate dehydrogenase, ALT - Alanine aminotransferase, CI - confidence interval).

A table containing MAE values for DT-GPT and the LightGBM baseline is provided in **Appendix A12**.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

4. Discussion

Our main finding is that a simple yet effective method allows training LLMs on EHRs to generate detailed patient trajectories that preserve inter variable correlations. This method achieves novel zero-shot performance, potentially making DT-GPT a digital twin platform for use cases such as treatment selection and clinical trial support.

Building on past LLM research in general forecasting [19,20], DT-GPT outperforms existing baselines in NSCLC and ICU datasets. These findings align with recent LLM forecasting developments [17,18], demonstrating that clinically specific adjustments enable accurate predictions. Additionally, DT-GPT's generative nature allows for multiple trajectory simulations per patient, offering insights into possible patient scenarios and uncertainty estimates.

The positive performance of LLMs for patient forecasting may stem from parallels between natural language and biomedical data, such as non-random missingness. For example, a doctor might skip measuring blood pressure if a patient appears healthy, indicating information even when missing. However, it could also be an artifact of the data processing. Natural language implicitly handles such ambiguity; unspoken words can still convey meaning or none at all. Recent advancements suggest that LLMs can capture these complex relationships [36].

DT-GPT addresses EHR challenges including noise, sparsity and lack of data normalization [16]. Unlike most established machine learning models that require data normalization and imputation, DT-GPT operates without these requirements. Its robustness to data sparsity and misspellings demonstrates its capability to handle incomplete, noisy medical data typical in real-world datasets. Moreover, EHR data often contain mixed data types; for instance, drug information may vary in encoding, such as being the dosage used or noted only as "administered", both which DT-GPT handles without additional preprocessing. Overall, DT-GPT simplifies and streamlines data preparation, thus enabling faster deployment across diverse datasets.

DT-GPT offers realistic prediction explanations and supports zero-shot forecasting, predicting patient variables not used during fine-tuning. This capability bridges the gap between medical expert and model, enabling the exploration of prediction rationales and alternative patient scenarios efficiently. This flexibility reduces the steps required to address new patient forecasts. Moreover, DT-GPT's zero-shot forecasting ability demonstrates its capacity to predict variables not used in training by learning their dynamics and adapting to new tasks. The unexpected finding that DT-GPT can outperform a supervised and fully trained machine learning model on specific variables, implicates the pioneering role of LLM-based models in real world dataset analyses.

A challenge of LLM-based models is the restricted number of simultaneously forecasted variables. The current constraint on the number of forecasted variables is due to the limited sequence length of the LLMs used in fine-tuning. Advances in extending the context length will enable modeling of additional patient variables [37]. Furthermore, we anticipate that transitioning from zero-shot to few-shot learning, where the model receives further training on a small subset of data, would enable a wider span of forecasted variables and extend DT-GPT's applicability to broader clinical challenges such as disease progression or survival prediction.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Another established shortcoming of LLM-based models is their tendency to hallucinate, i.e. the explainability results not being mathematically rigorous and may not necessarily provide true answers. This is critical for the medical domain and an active field of research in explainable AI [38], which we believe will be the focus of the next generation of LLM-based models.

Finally, we observe that high error predictions often occur due to the high variance between the multiple generated trajectories of each patient sample, with the mean aggregation into the final prediction not capturing key dynamics. It is thus an open challenge to develop improved aggregation methods, for example by using a second LLM or by having a human expert select the most realistic trajectory.

In conclusion, DT-GPT serves as a digital twin forecasting platform, enabling accurate and stable predictions, exploratory interpretability via a natural-language interface, and forecasting of patient variables not used in fine-tuning. DT-GPT exhibits true digital twin behaviors, potentially reproducing all aspects of the patients it represents, and surpassing traditional AI methods optimized for individual variables. We believe patient-level digital twins will impact clinical trials by supporting biomarker exploration, trial design, and interim analysis. Additionally, digital twins will assist doctors in treatment selection and patient monitoring. Overall, we envision LLM-powered digital twins becoming integral to healthcare systems.

Acknowledgements

We would like to thank Anton Kraxner for providing crucial insights into NSCLC and which variables to select, as well as Ginte Kutkaite, Hugo Loureiro, Franziska Braun, Rudolf Kinder and Venus So for their valuable input and discussions.

Source code and data access

The data that support the findings of this study were originated by and are the property of Flatiron Health, Inc., which has restrictions prohibiting the authors from making the data set publicly available. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to PublicationsDataAccess@flatiron.com. The Medical Information Mart for Intensive Care IV (MIMIC-IV) is available publicly online. The source code is in the process of being released, and will be shared openly in the future.

Funding

F. Hoffmann-La Roche, Helmholtz Association, Munich School for Data Science - MUDS, and European Union’s Horizon 2020 Research and Innovation Programme.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Bibliography

1. Schachter AD, Ramoni MF. Clinical forecasting in drug development. Nat Rev Drug Discov. 2007;6(2):107–8.
2. Boulos MNK, Zhang P. Digital Twins: From Personalised Medicine to Precision Public Health. J Pers Med. 2021;11(8):745.
3. Armeni P, Polat I, Rossi LMD, Diaferia L, Meregalli S, Gatti A. Digital Twins in Healthcare: Is It the Beginning of a New Era of Evidence-Based Medicine? A Critical Review. J Pers Med. 2022;12(8):1255.
4. Bordukova M, Makarov N, Rodriguez-Esteban R, Schmich F, Menden MP. Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. Expert Opin Drug Discov. 2024;19(1):33–42.
5. Nguyen M, He T, An L, Alexander DC, Feng J, Yeo BTT, Initiative for the ADN. Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage. 2020;222:117203.
6. Fox I, Ang L, Jaiswal M, Pop-Busui R, Wiens J. Deep Multi-Output Forecasting: Learning to Accurately Predict Blood Glucose Trajectories. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018. (Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining).
7. Ullah U, Xu Z, Wang H, Menzel S, Sendhoff B, Bäck T. Exploring Clinical Time Series Forecasting with Meta-Features in Variational Recurrent Models. 2020 Int Jt Conf Neural Netw (IJCNN). 2020;00:1–9.
8. Jung W, Mulyadi AW, Suk HI. Unified Modeling of Imputation, Forecasting, and Prediction for AD Progression. Lect Notes Comput Sci. 2019;168–76.
9. Wu F, Zhao G, Zhou Y, Qian X, Baedorf-Kassis E, Lehman L wei H. Forecasting Treatment Outcomes Over Time Using Alternating Deep Sequential Models. IEEE Trans Biomed Eng. 2023;PP(99):1–10.
10. Phetrittikun R, Suvirat K, Pattalung TN, Kongkamol C, Ingviya T, Chaichulee S. Temporal Fusion Transformer for forecasting vital sign trajectories in intensive care patients. 2021 13th Biomed Eng Int Conf (BMEiCON). 2021;00:1–5.
11. Chang P, Li H, Quan SF, Lu S, Wung SF, Roveda J, Li A. TDSTF: Transformer-based Diffusion probabilistic model for Sparse Time series Forecasting. arXiv. 2023;246:108060.
12. Seedat N, Imrie F, Bellot A, Qian Z, Schaar M van der. Continuous-Time Modeling of Counterfactual Outcomes Using Neural Controlled Differential Equations. In: International Conference on Machine Learning. 2022. p. 19497–521.
13. Melnychuk V, Frauen D, Feuerriegel S. Causal Transformer for Estimating Counterfactual Outcomes. In: International Conference on Machine Learning. 2022. p. 15293–329. (PMLR).
14. Hess K, Melnychuk V, Frauen D, Feuerriegel S. Bayesian Neural Controlled Differential Equations for Treatment Effect Estimation. arXiv. 2023;
15. Kaddour J, Lynch A, Liu Q, Kusner MJ, Silva R. Causal Machine Learning: A Survey and Open Problems. arXiv. 2022;
16. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. npj Digit Med. 2023;6(1):135.
17. Liang Y, Wen H, Nie Y, Jiang Y, Jin M, Song D, Pan S, Wen Q. Foundation Models for Time Series Analysis: A Tutorial and Survey. arXiv. 2024;
18. Xue H, Salim FD. PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting. IEEE Transactions on Knowledge and Data Engineering. 2023;

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

19. Liu H, Zhao Z, Wang J, Kamarthi H, Prakash BA. LSTPrompt: Large Language Models as Zero-Shot Time Series Forecasters by Long-Short-Term Prompting. arXiv. 2024;
20. Gruver N, Finzi M, Qiu S, Wilson AG. Large Language Models Are Zero-Shot Time Series Forecasters. In: Advances in Neural Information Processing Systems. 2023.
21. Loureiro H, Kolben TM, Kiermaier A, Rüttinger D, Ahmidi N, Becker T, Bauer-Mehren A. Correlation Between Early Trends of a Prognostic Biomarker and Overall Survival in Non–Small-Cell Lung Cancer Clinical Trials. JCO Clin Cancer Inform. 2023;7(7):e2300062.
22. Ma X, Long L, Moon S, Adamson BJS, Baxi SS. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR. medRxiv. 2023;2020.03.16.20037143.
23. Birnbaum B, Nussbaum N, Seidl-Rathkopf K, Agrawal M, Estevez M, Estola E, Haimson J, He L, Larson P, Richardson P. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. arXiv. 2020;
24. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman L wei H, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1.
25. Gupta M, Gallamoza B, Cutrona N, Dhakal P, Poulain R, Beheshti R. An Extensive Data Processing Pipeline for MIMIC-IV. In: Machine Learning for Health. 2022. p. 311–25. (PMLR).
26. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv. 2024;
27. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: The Eleventh International Conference on Learning Representations. 2022.
28. Lim B, Arık SÖ, Loeff N, Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. Int J Forecast. 2021;37(4):1748–64.
29. Das A, Kong W, Leach A, Mathur S, Sen R, Yu R. Long-term Forecasting with TiDE: Time-series Dense Encoder. arXiv. 2023;
30. Nespoli L, Medici V. Multivariate Boosted Trees and Applications to Forecasting and Control. ournal of Machine Learning Research, 23. 2022;(246):1–47.
31. Ke G. LightGBM: A Highly Efficient Gradient Boosting Decision Tree [Internet]. 2017 [cited 2024 Apr 4]. Available from: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
32. Altman DG, Bland JM. Standard deviations and standard errors. BMJ. 2005;331(7521):903.
33. Sibille A, Henket M, Corhay JL, Alfieri R, Louis R, Duysinx B. White Blood Cells in Patients Treated with Programmed Cell Death-1 Inhibitors for Non-small Cell Lung Cancer. Lung. 2021;199(5):549–57.
34. Lee S, Jeon H, Shim B. Prognostic Value of Ferritin-to-Hemoglobin Ratio in Patients with Advanced Non-Small-Cell Lung Cancer. J Cancer. 2019;10(7):1717–25.
35. Miller RF, Lipman MCI, Morris A. Clinical Respiratory Medicine (Fourth Edition). Sect 5: Infect Dis. 2012;(Proc Am Thorac Soc82011):346–73.
36. Sravanthi SL, Doshi M, Kalyan TP, Murthy R, Bhattacharyya P, Dabre R. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. arXiv. 2024;
37. Ding Y, Zhang LL, Zhang C, Xu Y, Shang N, Xu J, Yang F, Yang M. LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens. arXiv. 2024;
38. Luo H, Specia L. From Understanding to Utilization: A Survey on Explainability for Large Language Models. arXiv. 2024;

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

39. Herzen J, Lässig F, Piazzetta SG, Neuer T, Tafti L, Raille G, Pottelbergh TV, Pasieka M, Skrodzki A, Huguenin N, Dumonal M, Kościsz J, Bader D, Gusset F, Benheddi M, Williamson C, Kosinski M, Petrik M, Grosch G. Darts: User-Friendly Modern Machine Learning for Time Series. Journal of Machine Learning Research. 2022;23(124):1–6.

40. Makridakis S, Spiliotis E, Assimakopoulos V. M5 accuracy competition: Results, findings, and conclusions. Int J Forecast. 2022;38(Machine Learning 45 2001):1346–64.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Appendix

A1 Dataset details

	Flatiron Health - NSCLC	**ICU - MIMIC-IV**
# Patients	16496	35131
# Time Points	773607	1686288
# Input Variables	320	300
# Output Variables	6	3
Avg. % Missing in Input	94.4%	98.1%
Avg. % Missing in Output	74.5%	35.1%
Female/Male/NA %	51.0/49.0/0.0	39.7/51.1/9.2
Avg./Std. Age at Start	67.5/10.2	64.0/16.3
Avg./Std. Length of Full Patient Trajectory	160.5/328.6 days	46.78/0.87 hours
Avg./Std. Nr of Total Events per Patient Trajectory	35.4/33.1	47.3/2.3
Avg./Std. average length between events	5.82/17.00 weeks	1.02/0.24 hours
Time Point Resolution	Weekly	Hourly
Input Time Horizon	Unlimited	24 hours
Forecast Time Horizon	Up to 13 weeks	Up to 24 hours

Table A1.1 Dataset details.

The outlier processing method is outlined in **Appendix A2**. The splitting into training/validation/test datasets is performed randomly for the MIMIC-IV dataset, whilst stratified by group stage, smoking status, number of observations per visit and number of drug visits for the NSCLC dataset to ensure a balanced evaluation.

For the NSCLC dataset, we selected the number of laboratory variables to incorporate all features that were already used in linear prognostic models [21], as well to have enough data with variables having over 2000 patients. The number of diagnoses was chosen to include key information, as well to have enough patient data for useful model training, having at least 1700 observations. With the clinical importance of variables shown in **Table A1.2**.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

For the MIMIC-IV dataset, we define high variability as measured by the R^2 of using the last observed value for the full forecast. The lower this value is, the higher the variability.

Variable	**Impact**
Leukocytes	NSCLC treatment, particularly chemotherapy, can cause leukopenia, leading to decreased leukocyte counts. This reduction in leukocytes can increase the risk of infections due to a compromised immune system.
Lymphocytes/Leukocytes	This ratio is often used to monitor the immune status and inflammatory response.
Neutrophils	Chemotherapy can lead to neutropenia, resulting in a reduced neutrophil count. Neutropenia increases the risk of infections.
Lymphocytes	Lymphocyte counts often decrease during NSCLC treatment due to the immunosuppressive effects of chemotherapy. This reduction can impair the body's ability to fight infections and may affect the overall immune response.
Lactate Dehydrogenase	Elevated levels of lactate dehydrogenase (LDH) can be observed, indicating tissue damage or tumor burden.
Hemoglobin	Hemoglobin levels may decrease, leading to anemia, as a side effect of chemotherapy or due to the cancer itself. Anemia can cause symptoms such as fatigue, weakness, and shortness of breath, impacting the patient's quality of life.

Table A1.2 Details on the NSCLC output variables.

A2 Outlier processing

To remove outliers, all target values more than three standard deviations were initially filtered out. Since there was still too much noise in the data, a second round of the filter was applied, though this time clipping the values, since high values could be outliers but still provide useful information.

A3 Base pretrained LLM

We focus on taking biomedical LLMs as the base LLMs, since they have been shown to contain biomedical knowledge, potentially improving the performance of the overall model. A number of different biomedical LLMs have been proposed, however, we examined BioMistral 7B DARE [26]. We take this model since it has an open source license, is based on popular existing LLMs and has been shown to perform well on biomedical tasks. Additionally, following [20], the model tokenizes numbers into individual digits. Here we use the 7B parameter version since it allows for a larger amount of experimentation. However, since the method is agnostic to the underlying LLM, DT-GPT can be equally developed based on larger models as well.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A4 Forecasting prompt examples

We structure the template in four components:

1. The patient's history is noted down chronologically, using relative dating to prevent overfitting on time. For each patient visit and for each observed value, we note down the patient variable's name and value, whilst omitting any missing variables.
2. Next, we include the patient's baseline data, such as age and cancer stage
3. Since we do not impute target values, we include information about which variables should be output at which future time points.
4. Finally, we add a short prompt.

The target variables are also converted based on templates, containing only the respective target values. To reduce the amount of tokens required, the output is formatted so the target variable is provided followed by the list of values corresponding to the days that we want to output.

Note that for the MIMIC dataset, we forward propagate all observed values, to ensure that for each variable, if the value was observed in the input, it will also be encoded. Here we present synthetic examples of both the manual and JSON input as well as output.

Manual Template Input (Synthetic Patient)

First, patient chronological patient history up until the current day. Patient visits for the first time, with the following values: advanced cancer diagnosis is non small cell NSCLC, initial cancer diagnosis is non small cell NSCLC.
14 days after previous visit, patient visits again, with the following values: ECOG is 0, alanine aminotransferase is 21, albumin is 42, calcium is 9.4, aspartate aminotransferase is 29, bilirubin is 0.5, carbon dioxide is 24, carcinoembryonic ag is 78.2, hematocrit 2 is 45.5, creatinine is 0.8, glucose is 123, lactate dehydrogenase 2 is 196, basophils 2 is 0, eosinophils 2 is 0.2, eosinophils/100 leukocytes is 3.6, erythrocytes 2 is 4.6, leukocytes 2 is 6.3, lymphocytes 2 is 2.5, lymphocytes/100 leukocytes 3 is 39.9, monocytes is 0.5, monocytes/100 leukocytes is 8.1, neutrophils is 3, platelets is 231, protein 2 is 68, basophils/100 leukocytes 2 is 0.7, granulocytes is 3, granulocytes/100 leukocytes is 47.7, urea nitrogen is 15, glomerular filtration rate/1.73 sq m.predicted.non black is 103, glomerular filtration rate/1.73 sq m.predicted.black is 125, alkaline phosphatase is 49, hemoglobin is 15.5, body height is 191.8, body weight is 116.4.

…

14 days after previous visit, patient visits again, with the following values: Dehydration is diagnosed, Adverse effect of antineoplastic and immunosuppressive drugs, initial encounter is diagnosed, cisplatin is 60, pemetrexed is 1225, ECOG is 0, alanine aminotransferase is 21, albumin is 41, … glomerular filtration rate/1.73 sq m.predicted.non black is 80, glomerular filtration rate/1.73 sq m.predicted.black is 109, alkaline phosphatase is 44, hemoglobin is 14.5, body height is 191.8, body weight is 117.8.
Next, the baseline data for the patient: birth year is 1948, gender is M, ses index is 2, is cancer advanced is True, histology is Non-squamous cell carcinoma, cancer stage is Stage IIIB, smoking status is No history of smoking, ethnicity is Not Hispanic or Latino, Current line of therapy is Cisplatin,Pemetrexed, Current line number is 1.
Finally, the variables which you should predict, and for which days in the future from the current day:

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

{"hemoglobin": [14, 21, 28, 42, 49, 56, 63, 70, 77], "leukocytes 2": [14, 21, 28, 42, 49, 56, 63, 70, 77], "lymphocytes 2": [14, 21, 28, 42, 49, 56, 63, 70, 77], "lymphocytes/100 leukocytes 3": [14, 21, 28, 42, 49, 56, 63, 70, 77], "neutrophils": [14, 21, 28, 42, 49, 56, 63, 70, 77]}
Now, your task is as follows: Given the non small cell NSCLC patient's history, please predict for this patient the previously noted down variables and future days, in the same JSON format.

JSON Input (Synthetic Patient)

{"Patient history, with each visit in chronological order and relative days to previous visit": {"0 days": {"initial cancer diagnosis": "non small cell NSCLC"}, "28 days": {"body height": "172.2", "body weight": "64.4", "oxygen saturation": "98"}, "126 days": {"creatinine": "1.2"}, "14 days": {"body weight": "70.4", "oxygen saturation": "99"}, "14 days": {"body height": "172.2", "body weight": "64.7", "oxygen saturation": "95"}, "70 days": {"creatinine": "1.5"}, "14 days": {"body height": "170.2", "body weight": "68.2", "oxygen saturation": "95"}, "21 days": {"body weight": "69.2", "oxygen saturation": "98"},

…

"14 days": {"body weight": "64.9"}, "7 days": {"Nausea with vomiting, unspecified": "diagnosed", "carboplatin": "140", "paclitaxel": "88", "alanine aminotransferase": "13", "albumin": "40", "calcium": "8.8", "aspartate aminotransferase": "29", "bilirubin": "0.4", "carbon dioxide": "20", … "neutrophils": "5.9", "neutrophils/100 leukocytes": "79", "platelets": "176", "potassium": "4.6", "protein 2": "69", "sodium": "136", "basophils/100 leukocytes 2": "0.1", "urea nitrogen": "15", "glomerular filtration rate/1.73 sq m.predicted.non black": "63", "alkaline phosphatase": "119", "hemoglobin": "16.4"}}, "Baseline data": {"birth year": 1941, "gender": "M", "ses index": "4", "is cancer advanced": true, "histology": "Non-squamous cell carcinoma", "cancer stage": "Stage IA2", "smoking status": "History of smoking", "ethnicity": "Not Hispanic or Latino", "line of therapy": "Carboplatin,Paclitaxel", "line number": 1}, "Output variables": {"Variables to predict for respective days": {"hemoglobin": [7, 14, 21, 28, 35, 42, 49, 56], "lactate dehydrogenase 2": [56], "leukocytes 2": [7, 14, 21, 28, 35, 42, 49, 56], "lymphocytes 2": [7, 14, 21, 28, 35, 42, 49, 56], "lymphocytes/100 leukocytes 3": [7, 14, 21, 28, 35, 42, 49, 56], "neutrophils": [7, 14, 21, 28, 35, 42, 49, 56]}}, "Prompt": "Given the non small cell NSCLC patient's history, please predict for this patient the previously noted down variables and future days, in the same JSON format."}

Manual Template Output (Synthetic Patient)

hemoglobin starts at 15.5 decreases to 14.4 increases to 14.5 decreases to 13.6 increases to 14.1 increases to 14.8 decreases to 14.4 decreases to 13.8.
lactate dehydrogenase 2 starts at 232.
leukocytes 2 starts at 6 increases to 7.7 decreases to 3.1 decreases to 2.3 increases to 3.1 increases to 6 decreases to 3.6 increases to 3.7.
lymphocytes 2 starts at 0.6 increases to 0.9 decreases to 0.4 decreases to 0.3 stays at 0.3 increases to 0.5 decreases to 0.4 stays at 0.4.
lymphocytes/100 leukocytes 3 starts at 9.5 increases to 12.3 increases to 13.1 stays at 13.1 decreases to 11.1 decreases to 8.1 increases to 10.2 increases to 12.1.
neutrophils starts at 5.1 increases to 6.2 decreases to 2.5 decreases to 1.8 stays at 1.8 increases to 4.7 decreases to 2.9 decreases to 2.6.

JSON Output (Synthetic Patient)

{"hemoglobin": ["13.9", "12.8", "13.4", "13.7", "12.9", "13.1", "12.9", "12.9", "12.8"], "leukocytes 2": ["2.5", "5.2", "2.3", "5", "1.8", "5.2", "4.3", "1.7", "2.8"], "lymphocytes 2": ["1.2", "1.5", "0.8", "1.4",

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

"0.6", "0.9", "1", "0.7", "0.8"], "lymphocytes/100 leukocytes 3": ["47.7", "28.8", "36", "27.7", "32.4", "17.1", "23.1", "39.4", "28.9"], "neutrophils": ["1", "3.3", "1.2", "2.9", "0.9", "3.6", "2.7", "0.6", "1.6"]}

A5 Fine-tuning & inference details

We initially experimented with applying the loss to the entire sequence, which would also allow generating synthetic patients, however the models hallucinated to an usable point. Instead we employed a masking such that the gradient is only computed for the tokens that need to be forecast. For the training, we set the learning rate to 10^{-5}, a warm up ratio of 0.1, batch size of 1, employ a cosine learning rate scheduler, with a weight decay 0.1 and the optimizer being AdamW. During training, we limit the input sequence length to 3400 tokens due to memory constraints. The optimal epoch was identified based on the loss on the validation set, with the training taking around 20 hours on a single NVIDIA A100 80GB GPU. For all evaluations, we run the model 30 times on each patient sample, and a maximum final sequence length of 4000 tokens. We used nucleus sampling with top p set to 0.9 and temperature set to 1.0.

For the chatbot prediction explainability and zero-shot non-target variable forecasting, we used the same nucleus sampling parameters (top p = 0.9 and temperature = 1). The maximum sequence length was set to 200 tokens for the explainability task and 120 tokens for the zero-shot forecasting task, respectively. The numbers were selected to cover the desired output sequence length and prevent severe hallucinations. For the zero-shot forecasting, we run DT-GPT 10 times on each patient sample and use mean aggregation to obtain the final prediction.

In the context of patient digital twins, it is crucial to differentiate between simulation and forecasting. Simulations represent realistic patient trajectories, whereas forecasts predict the trajectories that are most likely to happen. Ideally, simulations should be able to cover the distribution of all possible patient trajectories.

A6 Chatbot prompt examples

A two-step chatbot interaction example for the prediction explainability task is provided below.

Original input prompt (Synthetic Patient)
First, patient chronological patient history up until the current day. Patient visits for the first time, with the following values: metastasis Adrenal is Adrenal, metastasis Bone is Bone, metastasis Liver is Liver. 21 days after previous visit, patient visits again, with the following values: advanced cancer diagnosis is non small cell NSCLC, initial cancer diagnosis is non small cell NSCLC. 14 days after previous visit, patient visits again, with the following values: Type 2 diabetes mellitus without complications is diagnosed, Hyperlipidemia, unspecified is diagnosed. …

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

14 days after previous visit, patient visits again, with the following values: carboplatin is 900, paclitaxel protein-bound is 255, ECOG is 1, alanine aminotransferase is 15, …, hemoglobin is 13.1, body height is 160.2, body weight is 91.4, oxygen saturation is 97. Next, the baseline data for the patient: birth year is 1954, gender is M, ses index is 5 - Highest SES, is cancer advanced is True, histology is Non-squamous cell carcinoma, cancer stage is Stage IVB, smoking status is History of smoking, ethnicity is Not Hispanic or Latino, Current line of therapy is Carboplatin,Paclitaxel Protein-Bound, Current line number is 1. Finally, the variables which you should predict, and for which days in the future from the current day: {hemoglobin: [21, 35], leukocytes 2: [21, 35], lymphocytes 2: [21, 35], lymphocytes/100 leukocytes 3: [21, 35], neutrophils: [21, 35]} Now, your task is as follows: Given the non small cell NSCLC patient's history, please predict for this patient the previously noted down variables and future days, in the same JSON format. <patient_prediction>
DT-GPT forecasting results response
{"hemoglobin": ["11", "9.8"], "leukocytes 2": ["6.4", "4.7"], "lymphocytes 2": ["0.7", "0.4"], "lymphocytes/100 leukocytes 3": ["11.5", "8.8"], "neutrophils": ["5.4", "4.1"]}
Prediction explainability prompt
What are the five most important variables or patient's baseline characteristics for your prediction?
DT-GPT explainability response
The five most important variables or patient's baseline characteristics for my prediction are: 1. ECOG 2. alanine aminotransferase 3. albumin 4. hemoglobin 5. leukocytes 2.

A two-step chatbot interaction example for the zero-shot forecasting task is provided below.

Original input prompt (Synthetic Patient)
First, patient chronological patient history up until the current day. Patient visits for the first time, with the following values: metastasis Adrenal is Adrenal, metastasis Bone is Bone, metastasis Liver is Liver. 21 days after previous visit, patient visits again, with the following values: advanced cancer diagnosis is non small cell NSCLC, initial cancer diagnosis is non small cell NSCLC. 14 days after previous visit, patient visits again, with the following values: Type 2 diabetes mellitus without complications is diagnosed, Hyperlipidemia, unspecified is diagnosed. … 14 days after previous visit, patient visits again, with the following values: carboplatin is 900, paclitaxel protein-bound is 255, ECOG is 1, alanine aminotransferase is 15, …, hemoglobin is 13.1, body height is 160.2, body weight is 91.4, oxygen saturation is 97. Next, the baseline data for the patient: birth year is 1954, gender is M, ses index is 5 - Highest SES, is cancer advanced is True, histology is Non-squamous cell carcinoma, cancer stage is Stage IVB, smoking status is History

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

of smoking, ethnicity is Not Hispanic or Latino, Current line of therapy is Carboplatin,Paclitaxel Protein-Bound, Current line number is 1. Finally, the variables which you should predict, and for which days in the future from the current day: {hemoglobin: [21, 35], leukocytes 2: [21, 35], lymphocytes 2: [21, 35], lymphocytes/100 leukocytes 3: [21, 35], neutrophils: [21, 35]} Now, your task is as follows: Given the non small cell NSCLC patient's history, please predict for this patient the previously noted down variables and future days, in the same JSON format. <patient_prediction>
DT-GPT forecasting results response
{"hemoglobin": ["11", "9.8"], "leukocytes 2": ["6.4", "4.7"], "lymphocytes 2": ["0.7", "0.4"], "lymphocytes/100 leukocytes 3": ["11.5", "8.8"], "neutrophils": ["5.4", "4.1"]}
Non-target variable forecasting prompt
Next, the variables which you should predict, and for which days in the future from the current day: {calcium: [21, 35]} Now, your task is as follows: Given the non small cell NSCLC patient's history, please predict for this patient the previously noted down variables and future days, in the same JSON format. <patient_prediction>
DT-GPT non-target forecasting results response
{"calcium": ["9.4", "10.3"]]}

A7 Baselines & metric details

The baseline models are implemented in the Darts library [39] and the default hyperparameters are used. For the input time horizon, both 35 and 91 days were explored for the Flatiron Health NSCLC dataset, whilst the full 24 hours was used for the MIMIC dataset. Since the models cannot natively deal with missing data, we employ linear interpolation with forward and backward passes on the input data, and linear interpolation only with forward pass on the target data. As for DT-GPT, we apply the filtering based on three standard deviations, and then apply standardization or one hot encoding. To ensure fairness between the baseline models and DT-GPT, we also provide the baselines with an indicator variable, having 1 for every future date which will be measured and 0 for those which are imputed.

A8 Results tables

In **Tables A8.1** and **A8.2** we show the performance of the models across the two datasets. It is interesting to note that LightGBM performs better than more complex models, which we hypothesize is due to the high dimensional noisy data, though this has also been observed in literature [30,31,40].

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

	NSCLC											
Model	**Hemoglobin**		**Leukocytes**		**Lymphocytes/Leukocytes**		**Lymphocytes**		**Neutrophils**		**Lactate Dehydrogenase**	
	MAE	**Corr.**	**MAE**	**Corr.**	**MAE**	**Corr.**	**MAE**	**Corr.**	**MAE**	**Corr.**	**MAE**	**Corr.**
Copy Forward	0.698	0.629	0.969	0.333	0.731	0.499	0.569	0.603	0.974	0.323	0.433	**0.739**
Linear Regr.	0.486	0.759	0.782	0.441	0.668	0.545	0.506	0.656	0.778	0.410	0.475	0.683
TFT	0.469	0.774	0.719	0.504	0.651	0.563	0.463	0.696	0.717	0.476	0.480	0.646
TiDE	0.464	0.768	0.737	0.488	0.655	0.567	0.465	0.704	0.740	0.452	0.453	0.668
Light GBM	0.453	0.780	0.727	0.508	**0.644**	0.583	0.456	0.712	0.734	0.467	0.425	0.732
DT-GPT (ours)	**0.440**	**0.796**	**0.689**	**0.540**	0.650	**0.587**	**0.437**	**0.733**	**0.699**	**0.496**	**0.417**	0.731

Table A8.1 Performance of all models on the NSCLC dataset. “Corr.” means Spearman Correlation (higher is better), “MAE” is Mean Absolute Error (lower is better), with the best performance highlighted in bold and ranked by the average MAE.

	Intensive Care Unit					
Model	**Magnesium**		**Respiratory Rate**		**02 Saturation**	
	MAE	**Corr.**	**MAE**	**Corr.**	**MAE**	**Corr.**
Copy Forward	0.681	0.462	0.769	0.470	0.746	0.484
Linear Regression	0.606	0.463	0.680	0.509	0.681	0.525
TiDE	0.534	0.549	0.635	0.559	0.652	0.570
Temporal Fusion Transformer	0.537	0.555	0.635	**0.562**	0.644	**0.576**
LightGBM	0.520	0.583	**0.634**	**0.562**	0.644	0.573
DT-GPT (ours)	**0.505**	**0.609**	0.636	**0.562**	**0.635**	**0.576**

Table A8.2 Performance of all models on the ICU dataset. “Corr.” means Spearman Correlation (higher is better), “MAE” is Mean Absolute Error (lower is better), with the best performance highlighted in bold and ranked by the average MAE.

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A9 Ablation study details

The misspelling algorithm randomly performs either perturbation, insertion, deletion or replacement, using all ASCII letters & digits, applied to the entire input text. This includes dates, variable names, values, baseline information and prompts. One operation is considered one misspelling.

A10 Encoding method comparison

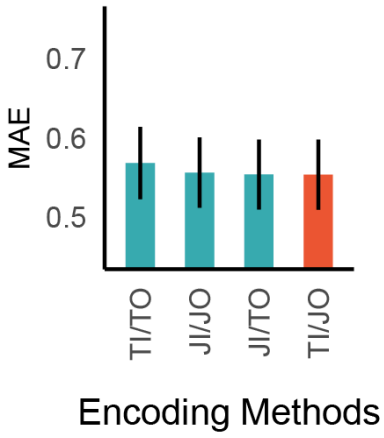

Figure A10.1 Different encoding methods and their respective performances on mean absolute error (MAE), with the following abbreviations: TI - "Text Input", TO - "Text Output", JI - "JSON Input" and JO - "JSON Output".

DT-GPT is stable with respect to different data encoding strategies (**Fig. A10.1)**, though using Text In, JSON Out (TI/JO) and JSON In, TEXT Out (JI/TO) perform best, with TI/JO being marginally more efficient. Specifically Text In, Text Out (TI/TO) achieves an average MAE of 0.568 ± 0.05, JSON In, JSON Out (JI/JO) reaches 0.556 ± 0.04, JSON In, Text Out (JI/TO) reaches 0.554 ± 0.04, Text In, JSON Out (TI/JO) attains 0.554 ± 0.04.

A11 Most important variables and patient baseline characteristics for the forecasting

For each of 200 patient samples in the test set sample, we obtain 10 predicted trajectories and 5 variables or patient baseline characteristics explaining those trajectories. In total, we observe 2000 trajectories and present aggregated counts of variables in the table below (**Tab. A11.1)**.

Variable	**Number of predicted trajectories it explains**
Line of therapy	1740
ECOG	1110

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

leukocytes	890
age	730
alanine aminotransferase	520
hemoglobin	507
neutrophils	463
Lymphocytes/100 leukocytes	420
Body weight and body height	403
albumin	311
gender	274
lymphocytes	256
lactate dehydrogenase 2	224
alkaline phosphatase	184
ferritin	137

Table A11.1 Key variables and their respective counts, as extracted from DT-GPT.

A12 Performance results of zero-shot DT-GPT and LightGBM baselines for non-target forecasting task

From the original 80 lab variables, we could evaluate the zero shot performance on 69 (**Tab. A12.1)**. The difference is due to the 6 variables used in training and a further 5 which had too few samples on either the input or output.

Variable	**MAE LightGBM**	**MAE DT-GPT**
neutrophils.segmented	1.057	0.560
neutrophils.band form	0.883	0.406
lactate dehydrogenase	0.374	0.128
leukocytes	0.924	0.813
ferritin	0.116	0.029
lymphocytes/100 leukocytes	0.689	0.625
lymphocytes 3	0.518	0.473

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Variable	MAE LightGBM	MAE DT-GPT
lymphocytes	0.497	0.457
granulocytes	0.740	0.707
neutrophils 2	0.820	0.801
carcinoembryonic ag	0.313	0.302
bilirubin.non-glucuronidated	0.785	0.780
granulocytes 2	1.071	1.067
lymphocytes/100 leukocytes 2	0.764	0.763
monocytes/100 leukocytes 2	0.892	1.032
monocytes	0.707	0.848
erythrocytes	0.173	0.333
platelets 2	0.696	0.862
erythrocytes 2	0.247	0.467
platelets	0.578	0.820
urea nitrogen	0.592	0.882
coagulation tissue factor induced	0.574	0.880
monocytes 2	0.784	1.098
protein	0.394	0.813
alkaline phosphatase	0.450	0.875
monocytes/100 leukocytes	0.782	1.270
monocytes 3	0.702	1.194
glomerular filtration rate/1.73 sq m.predicted 2	0.552	1.055
potassium 2	0.709	1.256
calcium	0.691	1.266
glucose	0.678	1.309
glomerular filtration rate/1.73 sq m.predicted.black	0.362	1.143

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Variable	MAE LightGBM	MAE DT-GPT
glomerular filtration rate/1.73 sq m.predicted.non black	0.378	1.164
coagulation surface induced	0.780	1.597
creatinine renal clearance.predicted	0.221	1.042
urate	0.536	1.393
eosinophils/100 leukocytes 2	0.897	1.767
coagulation tissue factor induced.inr	0.580	1.560
eosinophils 2	0.458	1.503
monocytes/100 leukocytes 3	0.598	1.735
glomerular filtration rate/1.73 sq m.predicted	0.269	1.408
sodium 2	0.855	1.995
bilirubin.glucuronidated+bilirubin.albumin bound	0.525	1.710
carbon dioxide	0.620	1.828
bilirubin	0.570	1.948
hematocrit 2	0.479	1.887
potassium	0.694	2.196
albumin	0.569	2.177
granulocytes/100 leukocytes 2	0.762	2.474
alanine aminotransferase	0.592	2.410
magnesium	0.667	2.492
protein 2	0.593	2.443
hematocrit	0.611	2.482
chloride	0.615	2.617
eosinophils/100 leukocytes	0.579	2.726
sodium	0.674	2.836
basophils/100 leukocytes 3	0.775	2.996
neutrophils/100 leukocytes	0.636	2.956

medRxiv preprint doi: https://doi.org/10.1101/2024.07.05.24309957; this version posted July 7, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Variable	MAE LightGBM	MAE DT-GPT
neutrophils/100 leukocytes 2	0.736	3.113
granulocytes/100 leukocytes	0.774	3.159
basophils	0.675	3.164
basophils/100 leukocytes 2	0.672	3.195
creatinine 2	0.485	3.039
basophils/100 leukocytes	0.782	3.367
aspartate aminotransferase	0.587	3.241
creatinine	0.396	3.065
eosinophils	0.567	3.637
gamma glutamyl transferase	0.362	3.825
basophils 2	0.657	4.208

Table A12.1 Zero shot performance of DT-GPT on variables that were previously not trained.