Supplementary information

Supplementary Table 1. Dataset details.

	Flatiron Health - NSCLC	**ICU - MIMIC-IV**	**Alzheimer's Disease - ADNI**
Data Type	Real World Data	Real World Data	Observational Study
# Patients	16496	35131	1140
# Time Points	773607	1686288	5700
# Input Variables	320	300	17
# Output Variables	6	3	3
Avg. % Missing in Input	94.4%	98.1%	3.9%
Avg. % Missing in Output	74.5%	35.1%	19.5%
Female/Male/NA %	51.0/49.0/0.0	39.7/51.1/9.2	56.8/43.2/0.0
Avg./Std. Age at Start	67.5/10.2	64.0/16.3	73.9/7.0
Avg./Std. Length of Full Patient Trajectory	160.5/328.6 days	46.78/0.87 hours	727.2/38.05
Avg./Std. Nr of Total Events per Patient Trajectory	35.4/33.1	47.3/2.3	4.24/0.45
Avg./Std. average length between events	5.8/17.0 weeks	1.0/0.2 hours	226.9/26.8 days
Time Point Resolution	Weekly	Hourly	6 months
Input Time Horizon	Unlimited	24 hours	Baseline measurements
Forecast Time Horizon	Up to 13 weeks	Up to 24 hours	Up to 24 months

Supplementary Table 1: Dataset details.

Supplementary Table 2. Details on the NSCLC output variables.

Variable	LOINC	Impact	Mean	Standard Deviation
Leukocytes [10^9/L]	26464-8	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.	7.18	3.36
Lymphocytes/Leukocytes [%]	26478-8	This ratio is often used to monitor the immune status and inflammatory response.	19.98	10.14
Neutrophils [10^9/L]	26499-4	Chemotherapy can lead to neutropenia, resulting in a reduced neutrophil count. Neutropenia increases the risk of infections.	5.17	3.22
Lymphocytes [10^9/L]	26474-7	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.	1.33	0.78
Lactate Dehydrogenase [U/L]	2532-0	Elevated levels of lactate dehydrogenase (LDH) can be observed, indicating tissue damage or tumor burden.	253.77	139.47
Hemoglobin [g/dL]	718-7	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.	11.84	1.87

Supplementary Table 2: Details on the NSCLC output variables.

Supplementary Table 3. Details on the ICU output variables.

Variable	Mean	Standard Deviation
Respiratory Rate [insp/min]	19.57	5.35
Oxygen Saturation [%]	96.84	2.63
Magnesium [mg/dL]	2.09	0.34

Supplementary Table 3: Details on the ICU output variables.

Supplementary Table 4. An overview of key parameters of the ICU dataset.

Parameter of ICU Dataset	**Value**
% of patients with diagnosed renal problems	46.9%
% of patients intubated and ventilated	26.8%
% of patients intubated	29.2%
% of patients ventilated	48.0%

Supplementary Table 4: An overview of some key parameters of the ICU dataset. We highlight the clinical relevance of the forecasting setup on the ICU patients by showing that almost half of all patients have kidney issues, where magnesium helps to monitor the patients' kidney status. Both respiratory rate and oxygen saturation are affected by intubation and ventilation, which is seen in half of the patients in the dataset.

Supplementary Table 5. Details on the Alzheimer's disease output variables.

Variable	Mean	Standard Deviation
Mini Mental State Examination (MMSE) [0-30]	26.8	3.57
Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS11) [0-70]	11.05	7.74
Clinical Dementia Rating Scale Sum of Boxes (CDRSB) [0-18]	1.92	2.30

Supplementary Table 5: Details on the Alzheimer's disease output variables.

Supplementary Note 1. Variable selection strategies.

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

For the MIMIC-IV dataset, the three variables O2 saturation pulse oximetry, respiratory rate and magnesium were selected since at least 50% of the patients have at least one measurement, and they have the highest temporal variability. The temporal variability was measured by the R2 of the copy forward model, with lower values showing higher variability. The exact values were -0.18 for magnesium, -0.07 for respiratory rate and -0.05 for 02 saturation pulse oximetry.

For the ADNI dataset, we selected three variables for complete cognitive scores that summarize results for separate cognitive tests (**Supplementary Table 5**) .

Supplementary Table 6. Results for the hemoglobin variable in the NSCLC dataset.

	Hemoglobin Forecasting Metrics					**Hemoglobin Classification Metrics**				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑	AUC low↑	AUC high↑	AUC weighted↑	Trend decrease↑	Trend increase↑
DT-GPT	**0.440**	**0.822**	**0.629**	**7.577**	**0.797**	**0.793**	0.850	**0.793**	**0.638**	**0.704**
Copy forward	0.698	1.308	-	11.605	0.629	*0.771*	*0.899*	*0.770*	0.500	0.500
Linear Regression	0.486	0.911	0.696	8.377	0.759	0.726	0.599	0.726	0.616	0.652
LightGBM	0.453	*0.848*	*0.649*	*7.792*	*0.780*	0.738	0.500	0.738	0.617	0.659
RNN	0.529	0.992	0.759	9.072	0.689	0.691	0.500	0.691	0.602	0.550
LSTM	0.526	0.985	0.753	9.012	0.693	0.717	0.500	0.716	0.598	0.569
Transformer	0.469	0.928	0.710	8.483	0.745	0.733	0.500	0.733	0.612	0.632
Temporal Fusion Transformer	0.469	0.879	0.672	8.056	0.773	0.750	0.500	0.750	0.578	0.647
TiDE	0.464	0.869	0.665	7.994	0.768	0.733	0.550	0.733	0.629	0.658
TCN	0.660	1.236	0.945	11.188	0.501	0.550	0.500	0.550	0.603	0.599
PatchTST	0.684	1.281	0.980	11.415	0.628	0.767	**0.900**	0.767	0.574	0.641
LLMTime	0.736	1.377	1.053	12.124	0.594	0.758	0.594	0.756	0.567	0.601
Time-LLM	0.665	1.246	0.953	11.169	0.612	0.741	0.650	0.741	0.604	0.611

Supplementary Table 6: Results for the hemoglobin variable in the NSCLC dataset.

Supplementary Table 7. Results for the leukocytes variable in the NSCLC dataset.

	Leukocytes Forecasting Metrics					**Leukocytes Classification Metrics**				
Model	Scaled MAE ↓	MAE ↓	MASE ↓	SMAPE ↓	Spearman correlation ↑	AUC low↑	AUC high↑	AUC weighted↑	Trend decrease↑	Trend increase↑
DT-GPT	**0.687**	**2.308**	**0.709**	**33.290**	**0.547**	**0.657**	*0.578*	**0.614**	**0.638**	**0.684**
Copy forward	0.969	3.257	-	43.301	0.333	0.557	**0.616**	0.554	0.500	0.500
Linear Regression	0.782	2.629	0.807	37.714	0.441	0.547	0.565	0.545	0.621	0.651
LightGBM	0.727	2.443	0.750	35.384	*0.508*	0.560	0.574	0.558	0.631	0.651
RNN	0.806	2.711	0.832	38.768	0.389	0.528	0.570	0.533	0.567	0.678
LSTM	0.781	2.625	0.806	37.813	0.397	0.540	0.553	0.537	0.624	0.608
Transformer	0.749	2.517	0.773	36.329	0.447	0.568	0.541	0.551	0.593	0.653
Temporal Fusion Transformer	*0.719*	*2.418*	*0.743*	*35.113*	0.504	0.561	0.534	0.547	0.581	0.615
TiDE	0.737	2.477	0.760	35.743	0.488	*0.574*	0.564	*0.561*	0.612	0.639
TCN	0.857	2.882	0.885	41.424	0.256	0.500	0.505	0.500	0.596	0.589
PatchTST	0.968	3.254	0.999	43.123	0.349	0.557	0.635	0.559	0.579	0.645
LLMTime	0.923	3.104	0.953	42.342	0.331	0.557	0.601	0.548	0.588	0.591
Time-LLM	0.894	3.006	0.923	41.668	0.339	0.542	0.605	0.543	0.601	0.595

Supplementary Table 7: Results for the leukocytes variable in the NSCLC dataset.

Supplementary Table 8. Results for the lymphocytes/leukocytes variable in the NSCLC dataset.

	Lymphocytes/ Leukocytes Forecasting Metrics					**Lymphocytes/ Leukocytes Classification Metrics**				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑	AUC low↑	AUC high↑	AUC weighted↑	Trend decrease↑	Trend increase↑
DT-GPT	**0.643**	**6.519**	**0.878**	**37.197**	**0.583**	**0.718**	0.532	**0.696**	**0.649**	**0.638**
Copy forward	0.731	7.421	-	44.156	0.499	0.677	**0.573**	0.660	0.500	0.500
Linear Regression	0.668	6.779	0.913	39.155	0.545	0.703	0.526	0.684	0.633	0.641
LightGBM	*0.644*	*6.537*	*0.881*	*37.656*	**0.583**	*0.717*	0.512	*0.695*	0.628	0.630
RNN	0.671	6.812	0.918	39.107	0.527	0.697	0.509	0.678	0.610	0.629
LSTM	0.665	6.743	0.909	38.668	0.540	0.693	0.507	0.674	0.639	0.616
Transformer	0.683	6.927	0.933	39.896	0.538	0.702	0.507	0.683	0.638	0.609
Temporal Fusion Transformer	0.651	6.604	0.890	37.974	0.563	0.708	0.502	0.689	0.579	0.575
TiDE	0.655	6.645	0.895	38.347	*0.567*	0.710	0.508	0.689	0.635	0.633
TCN	0.752	7.630	1.028	43.412	0.401	0.641	0.500	0.626	0.597	0.595
PatchTST	0.719	7.294	0.983	43.048	0.508	0.682	0.578	0.665	0,575	0.622
LLMTime	0.725	7.355	0.991	43.592	0.504	0.679	*0.557*	0.661	0.590	0.603
Time-LLM	0.684	6.945	0.936	40.076	0.519	0.688	*0.557*	0.670	0.614	0.598

Supplementary Table 8: Results for the lymphocytes/leukocytes variable in the NSCLC dataset.

Supplementary Table 9. Results for the lymphocytes variable in the NSCLC dataset.

	Lymphocytes Forecasting Metrics					Lymphocytes Classification Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑	AUC low↑	AUC high↑	AUC weighted↑	Trend decrease↑	Trend increase↑
DT-GPT	**0.434**	**0.342**	**0.764**	**29.770**	**0.734**	**0.788**	0.578	**0.788**	**0.635**	**0.656**
Copy forward	0.569	0.448	-	37.098	0.603	0.722	*0.672*	0.717	0.500	0.500
Linear Regression	0.506	0.398	0.889	34.023	0.656	0.741	0.571	0.736	0.618	0.628
LightGBM	*0.456*	*0.359*	*0.802*	*31.194*	*0.712*	*0.760*	0.558	*0.755*	0.619	0.636
RNN	0.511	0.402	0.898	35.401	0.622	0.722	0.571	0.718	0.512	0.656
LSTM	0.595	0.389	0.870	34.329	0.632	0.733	0.558	0.728	0.577	0.609
Transformer	0.503	0.396	0.884	34.529	0.670	0.717	0.577	0.713	0.598	0.646
Temporal Fusion Transformer	0.463	0.364	0.813	31.929	0.696	0.760	0.545	0.754	0.557	0.595
TiDE	0.465	0.366	0.817	32.076	0.704	0.758	0.545	0.752	0.630	0.642
TCN	0.606	0.476	1.065	40.687	0.477	0.619	0.500	0.617	0.604	0.600
PatchTST	0.560	0.440	0.984	36.689	0.610	0.727	**0.659**	0.722	0.576	0.644
LLMTime	0.601	0.472	1.056	36.997	0.600	0.728	0.642	0.722	0.582	0.579
Time-LLM	0.544	0.428	0.956	36.155	0.607	0.705	0.622	0.701	0.624	0.610

Supplementary Table 9: Results for the lymphocytes variable in the NSCLC dataset.

Supplementary Table 10. Results for the neutrophils variable in the NSCLC dataset.

	Neutrophils Forecasting Metrics					Neutrophils Classification Metrics				
Models	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑	AUC low↑	AUC high↑	AUC weighted↑	Trend decrease↑	Trend increase↑
DT-GPT	**0.701**	**2.265**	**0.720**	**43.473**	**0.499**	**0.528**	0.635	0.629	**0.648**	**0.683**
Copy forward	0.974	3.144	-	54.267	0.323	*0.524*	**0.648**	**0.641**	0.500	0.500
Linear Regression	0.778	2.512	0.799	47.873	0.410	0.508	0.608	0.605	0.611	0.642
LightGBM	0.734	2.369	0.753	45.696	0.467	0.502	0.618	0.614	0.611	0.639
RNN	0.801	2.588	0.823	49.210	0.370	0.503	0.614	0.610	0.564	0.681
LSTM	0.764	2.466	0.784	47.376	0.379	0.501	0.594	0.590	0.608	0.625
Transformer	0.741	2.393	0.761	46.242	0.417	0.503	0.573	0.570	0.607	0.632
Temporal Fusion Transformer	*0.717*	*2.315*	*0.736*	*45.030*	*0.476*	0.500	0.565	0.562	0.573	0.606
TiDE	0.740	2.389	0.760	45.907	0.452	0.508	0.598	0.594	0.617	0.651
TCN	0.832	2.688	0.855	51.678	0.254	0.500	0.512	0.511	0.594	0.592
PatchTST	0.959	3.097	0.985	53.837	0.321	0.521	*0.644*	*0.637*	0.565	0.635
LLMTime	0.900	2.906	0.924	52.386	0.314	0.519	0.612	0.608	0.588	0.594
Time-LLM	0.878	2.834	0.901	52.187	0.312	0.511	0.630	0.624	0.602	0.593

Supplementary Table 10: Results for the neutrophils variable in the NSCLC dataset.

Supplementary Table 11. Results for the lactate dehydrogenase variable in the NSCLC dataset.

	Lactate Dehydrogenase Forecasting Metrics					**Lactate Dehydrogenase Classification Metrics**				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE ↓	Spearman correlation↑	AUC low↑	AUC high↑	AUC weighted↑	Trend decrease↑	Trend increase↑
DT-GPT	**0.418**	**58.269**	**0.966**	**21.037**	*0.738*	0.753	**0.793**	**0.791**	**0.641**	**0.648**
Copy forward	0.433	60.350	-	*21.903*	**0.739**	*0.782*	*0.784*	*0.778*	0.500	0.500
Linear Regression	0.475	66.227	1.097	24.988	0.683	0.530	0.738	0.736	0.564	0.551
LightGBM	*0.425*	*59.278*	*0.982*	21.940	0.732	0.633	0.773	0.773	0.553	0.540
RNN	0.433	60.424	1.001	21.880	0.712	0.690	0.773	*0.778*	0.551	0.573
LSTM	0.441	61.521	1.019	22.595	0.693	0.640	0.765	0.764	0.592	0.535
Transformer	0.514	71.699	1.188	27.135	0.650	0.500	0.639	0.648	0.523	0.565
Temporal Fusion Transformer	0.480	66.938	1.109	25.356	0.646	0.500	0.691	0.702	0.531	0.563
TiDE	0.453	63.126	1.046	23.544	0.668	0.661	0.748	0.752	0.561	0.541
TCN	0.731	101.890	1.688	37.336	0.137	0.500	0.500	0.500	0.549	0.537
PatchTST	0.447	62.362	1.033	22.717	0.699	0.746	0.776	0.773	0.549	0.538
LLMTime	0.437	60.937	1.010	22.507	0.729	**0.793**	0.775	0.771	0.577	0.534
Time-LLM	0.443	61.844	1.025	22.625	0.681	0.693	0.768	0.765	0,582	0.548

Supplementary Table 11: Results for the lactate dehydrogenase variable in the NSCLC dataset.

Supplementary Table 12. Results for the magnesium variable in the ICU dataset.

	Magnesium Forecasting Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑
DT-GPT	**0.505**	**0.175**	**0.741**	**8.274**	**0.609**
Copy forward	0.681	0.236	-	11.192	0.462
Linear Regression	0.606	0.210	0.890	9.938	0.463
LightGBM	*0.520*	*0.180*	*0.763*	*8.514*	*0.583*
RNN	0.597	0.207	0.877	9.773	0.459
LSTM	0.567	0.197	0.833	9.320	0.469
Transformer	0.537	0.186	0.789	8.804	0.546
Temporal Fusion Transformer	0.537	0.186	0.788	8.799	0.555
TiDE	0.534	0.185	0.785	8.766	0.549
TCN	0.612	0.212	0.899	9.988	0.395
PatchTST	0.671	0.233	0.987	11.048	0.464
LLMTime	0.759	0.263	1.114	12.159	0.422
Time-LLM	0.664	0.230	0.976	10.964	0.462

Supplementary Table 12: Results for the magnesium variable in the ICU dataset.

Supplementary Table 13. Results for the respiratory rate variable in the ICU dataset.

	Respiratory Rate Forecasting Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑
DT-GPT	0.636	3.406	0.827	17.623	**0.562**
Copy forward	0.769	4.117	-	21.144	0.470
Linear Regression	0.680	3.631	0.882	18.816	0.509
LightGBM	**0.634**	**3.393**	**0.824**	*17.569*	**0.562**
RNN	0.647	3.446	0.842	17.921	0.558
LSTM	0.643	3.451	0.837	17.817	0.552
Transformer	0.644	3.403	0.838	17.843	0.549
Temporal Fusion Transformer	*0.635*	*3.402*	0.827	17.589	**0.562**
TiDE	*0.635*	3.818	0.826	17.601	*0.559*
TCN	0.713	3.398	0.927	19.713	0.407
PatchTST	*0.635*	3.398	*0.825*	*17.568*	**0.562**
LLMTime	0.686	3.671	0.892	18.526	0.538
Time-LLM	0.655	3.905	0.852	18.110	0.547

Supplementary Table 13: Results for the respiratory rate variable in the ICU dataset.

Supplementary Table 14. Results for the oxygen saturation variable in the ICU dataset.

	Oxygen Saturation Forecasting Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑
DT-GPT	**0.635**	**1.672**	**0.851**	**1.739**	**0.576**
Copy forward	0.746	1.964	-	2.045	0.484
Linear Regression	0.681	1.793	0.913	1.863	0.525
LightGBM	0.644	1.696	0.863	1.763	*0.573*
RNN	0.674	1.773	0.903	1.843	0.550
LSTM	*0.642*	*1.690*	*0.861*	*1.757*	0.566
Transformer	0.651	1.713	0.872	1.781	0.559
Temporal Fusion Transformer	0.644	1.695	0.863	1.762	**0.576**
TiDE	0.652	1.716	0.874	1.783	0.570
TCN	0.726	1.911	0.973	1.985	0.440
PatchTST	0.646	1.701	0.866	1.769	0.556
LLMTime	0.688	1.810	0.922	1.873	0.528
Time-LLM	0.665	1.749	0.891	1.820	0.545

Supplementary Table 14: Results for the oxygen saturation variable in the ICU dataset.

Supplementary Table 15. Results for the CDRSB variable in the Alzheimer's disease dataset.

	CDRSB Forecasting Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑
DT-GPT	**0.417**	**0.765**	**0.774**	**41.794**	*0.902*
Copy forward	0.539	0.988	-	*45.519*	0.855
Linear Regression	*0.449*	*0.823*	*0.833*	78.650	0.900
LightGBM	0.455	0.836	0.845	80.400	0.897
RNN	0.463	0.850	0.860	78.414	0.891
LSTM	0.468	0.859	0.869	77.957	0.891
Transformer	0.485	0.889	0.900	91.937	0.885
Temporal Fusion Transformer	0.451	0.829	0.839	78.464	**0.908**
TiDE	0.498	0.914	0.925	80.357	0.881
TCN	-	-	-	-	-
PatchTST	0.540	0.990	1.002	85.056	0.845
LLMTime	0.539	0.988	1.000	*45.519*	0.855
Time-LLM	0.540	0.991	1.002	85.060	0.848

Supplementary Table 15: Results for the CDRSB variable in the Alzheimer's disease dataset.

Supplementary Table 16. Results for the ADAS11 variable in the Alzheimer's disease dataset.

	ADAS11 Forecasting Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑
DT-GPT	*0.458*	*3.031*	*0.882*	31.297	*0.852*
Copy forward	0.519	3.437	-	33.102	0.809
Linear Regression	**0.457**	**3.024**	**0.880**	31.391	**0.859**
LightGBM	0.462	3.059	0.890	31.135	0.849
RNN	0.465	3.079	0.896	**30.898**	0.846
LSTM	0.475	3.147	0.916	31.558	0.848
Transformer	0.481	3.181	0.926	32.367	0.837
Temporal Fusion Transformer	0.466	3.083	0.897	*31.023*	0.847
TiDE	0.506	3.347	0.974	33.578	0.837
TCN	-	-	-	-	-
PatchTST	0.506	3.350	0.975	32.703	0.814
LLMTime	0.503	3.333	0.970	32.652	0.815
Time-LLM	0.506	3.350	0.975	32.704	0.815

Supplementary Table 16: Results for the ADAS11 variable in the Alzheimer's disease dataset.

Supplementary Table 17. Results for the MMSE variable in the Alzheimer's disease dataset.

	MMSE Forecasting Metrics				
Model	Scaled MAE ↓	MAE↓	MASE↓	SMAPE↓	Spearman correlation↑
DT-GPT	*0.535*	*1.574*	*0.818*	*6.973*	*0.782*
Copy forward	0.654	1.925	-	8.305	0.683
Linear Regression	0.551	1.620	0.842	7.107	0.771
LightGBM	0.540	1.589	0.826	7.008	0.760
RNN	0.545	1.603	0.833	7.070	*0.782*
LSTM	0.545	1.603	0.833	7.071	**0.783**
Transformer	0.553	1.626	0.845	7.143	0.757
Temporal Fusion Transformer	**0.520**	**1.530**	**0.795**	**6.775**	*0.782*
TiDE	0.578	1.700	0.883	7.456	0.744
TCN	-	-	-	-	-
PatchTST	0.654	1.925	1.000	8.305	0.678
LLMTime	0.654	1.925	1.000	8.305	0.683
Time-LLM	0.654	1.925	1.000	8.307	0.675

Supplementary Table 17: Results for the MMSE variable in the Alzheimer's disease dataset.

Supplementary Note 2. Evaluation results.

In **Supplementary Tables 6-17**, we show the performance of the models across the three datasets. It is interesting to note that LightGBM often performs better than more complex models, which we hypothesize is due to the high dimensional noisy data, though this has also been observed in the literature[1,2]. In the Alzheimer's disease dataset, we see that for ADAS11 linear regression has a slightly lower scaled MAE, and for MMSE TFT has the lowest MAE. We believe this is due to convergence issues since the dataset has less than a thousand training patients.

In the majority of cases DT-GPT outperforms the baselines across all metrics, both in forecasting and classification. In the classification metrics, DT-GPT exhibits improved performance over the baselines in trend detection, which is an especially clinically relevant component of trajectories.

Forecasting evaluation

To statistically quantify the differences between DT-GPT and the second-best performing model, which is LightGBM for the NSCLC and ICU datasets and TFT for the ADNI dataset, respectively, a one-tailed Wilcoxon signed-rank test with the following hypotheses was conducted:

- Null hypothesis (H_0): The distribution of the difference in errors between DT-GPT and the second-best performing model is symmetric around zero (i.e. there is no systematic difference between models),
- Alternative hypothesis (H_1): The distribution of error between DT-GPT and second-best performing model differences is stochastically less than a symmetric distribution around zero (i.e. DT-GPT has systematically lower errors than the second-best performing model).

First, the difference of paired-data MAEs calculated for each patient, aggregated over all variables and time points,was tested. The difference in error distribution was significant at the confidence level $\alpha = 0.05$ in favor of DT-GPT for the NSCLC (p-value = 9.6162×10^{-17}) and ICU (p-value = 0.00043) datasets. While the difference was not statistically significant for the ADNI dataset (p-value = 0.16358), DT-GPT still had a lower overall MAE, indicating robustness across diverse clinical settings (**Supplementary Table 18**).

Next, a one-tailed Wilcoxon signed-rank test was performed for each variable separately using paired-data MAEs calculated for each patient aggregated over all time points. The Benjamini-Hochberg correction procedure with false discovery rate (FDR) threshold of 0.05 was applied to adjust for multiple testing. DT-GPT was shown to have significantly systematically lower errors than the second-best performing model on 9 out of 12 variables (on all 6 variables for the NSCLC, 2 variables for the ICU and 1 variable for the ADNI dataset, respectively; **Supplementary Table 19**; **Supplementary Figure 1**).

Furthermore, bootstrapping on the test set was used to determine confidence intervals for the relative MAE improvement of DT-GPT to the second best performing model, which is given by 1 - (MAE DT-GPT / MAE second best performing model). On each dataset, MAEs were calculated for each variable separately and then averaged across each variable to give each variable the same importance independent of the number of values observed. Sampling with replacement was performed for n = 10,000

times and 95% confidence intervals (CI) were determined by the 2.5th and 97.5th quantiles. Bootstrapping showed DT-GPT to have a mean relative MAE improvement of 3.40% with CI [2.79%, 4.02%] for NSCLC and 1.20% with CI [0.79%, 1.62%] for the ICU dataset, respectively. However, on the ADNI dataset, the mean relative MAE improvement of 1.95% had CI of [-0.77%, 4.74%] (**Supplementary Figure 1**).

The results of statistical testing and bootstrapping are coherent, whereby DT-GPT significantly outperforms LightGBM on the NSCLC and ICU datasets. While the performance improvement on the ADNI dataset is not significant, DT-GPT still performs competitively (**Supplementary Table 18**). This result is likely influenced by the relatively small size of the ADNI training dataset (1,140 patients only), which may limit the advantages offered by DT-GPT's foundation model architecture. Notably, even in this constrained setting, DT-GPT maintains performance on par with strong baselines. Furthermore, it is important to recognize that our analysis focused on comparison against only the second best performing model. Overall, DT-GPT was compared against 14 baseline models, all of which exhibit lower performance, underscoring that DT-GPT achieves state-of-the-art results across diverse clinical tasks.

Interestingly, we observe that hemoglobin, lymphocytes, and lactate dehydrogenase exhibit relatively high Spearman correlations (above 0.7)[3], indicating a strong positive fit. In contrast, variables with lower correlations often display spikes in their time series. In these cases, Spearman correlation may not be the most appropriate measure, as the magnitude of the spikes is crucial; when calculating the correlation, differing spike magnitudes can lead to different ranks even if the overall spike is accurately captured by the model. Additionally, Spearman correlation does not capture temporal dependencies. While DT-GPT generally outperforms other models in terms of Spearman correlation (**Supplementary Tables 6-17**), we believe that future work can further enhance model performance in this area and incorporate clinical trajectory-specific metrics.

Classification metrics

DT-GPT generally performs better in clinically relevant classification metrics (**Methods**) than baselines, however the evaluation results suggest that further model optimization is necessary. Taking bleeding as an example, while DT-GPT is able to forecast hemoglobin values below the reference range (10 g/dL for women and 12 g/dL for men) with ROC AUC score of 0.793, it fails to predict the low values in hemoglobin (< 7.5 g/dL) associated with a bleeding with the corresponding ROC AUC score of 0.506. We note, however, that only 1.2% of all hemoglobin measurements in the NSCLC dataset are below 7.5 g/dL, and the best ROC AUC score achieved by the copy forward baseline is 0.528. Moreover, all models were unable to detect high values in leukocytes (> 11.0 10^9/L) associated with an infection, with the best performing baseline copy forward achieving ROC AUC of 0.616 and the second best model DT-GPT of 0.578, respectively.

On the other hand, DT-GPT is capable of capturing trends in white blood cell dynamics by having the highest ROC AUC values for increasing trend detection over the period of three weeks that can be associated with a progressive chronic inflammation or infection. We obtain ROC AUC of 0.684, 0.656 and 0.683 for leukocytes, lymphocytes and neutrophils, respectively.

Finally, DT-GPT is able to detect values of lactate dehydrogenase (> 222 U/L) that are associated with disease progression of NSCLC patients[4] with ROC AUC of 0.793, followed by the copy forward baseline with ROC AUC of 0.784.

Supplementary Table 18. Wilcoxon signed-rank test shows that DT-GPT significantly outperforms the second best performing model LightGBM on the NSCLC and ICU datasets.

Dataset	Test statistics	p-value	Sample size
NSCLC*	1,576,194	9.6162×10^{-17}	2,773
ICU*	2,885,993	0.00043	3,513
ADNI	12,076	0.16358	228

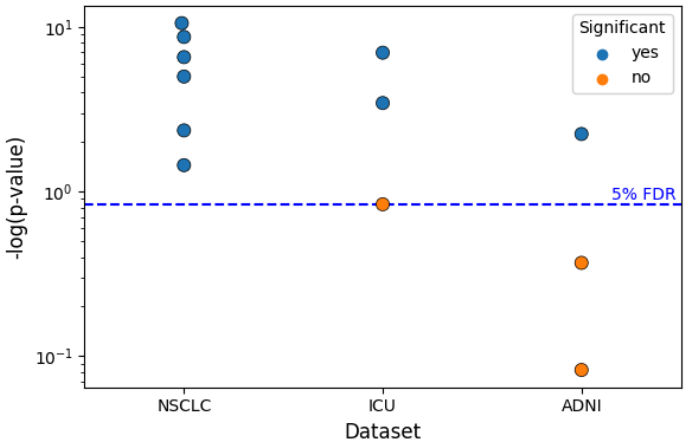
Supplementary Table 18: Comparison of DT-GPT against second best performing model across datasets. One-tailed Wilcoxon signed-rank test revealed that DT-GPT has significantly lower errors than the second-best performing model (LightGBM for the NSCLC and ICU datasets, TFT for the ADNI dataset) at the confidence level of $\alpha = 0.05$. Asterisks indicate datasets with significant test results.

Supplementary Table 19. Wilcoxon signed-rank test shows that DT-GPT significantly outperforms the second best performing model on 9 out of 12 variables from NSCLC, ICU and ADNI datasets.

Dataset	Variable	Test statistics	p-value	Adjusted p-value	Sample size
NSCLC*	Neutrophils*	630,730	2.99×10^{-11}	3.58×10^{-10}	1,754
NSCLC*	Leukocytes*	851,376	2.10×10^{-9}	1.26×10^{-8}	2,003
ICU*	O2 Saturation*	2,758,322	1.09×10^{-7}	4.35×10^{-7}	3,503
NSCLC*	Lymphocytes*	743,642	2.84×10^{-7}	8.51×10^{-7}	1,853
NSCLC*	Hemoglobin*	1,726,761	1.03×10^{-5}	2.46×10^{-5}	2,760
ICU*	Magnesium*	2,641,014	0.00036	0.00072	3,365
NSCLC*	Lymphocytes/ Leukocytes*	873,201	0.00447	0.76682	1,936
ADNI*	CDRSB*	10,538	0.00583	0.00874	228
NSCLC*	Lactate dehydrogenase*	41,096	0.03593	0.0479	427
ICU	Respiratory Rate	2,988,284	0.14584	0.17501	3,493
ADNI	ADAS11	12,872	0.42798	0.46688	228
ADNI	MMSE	13,993	0.82709	0.82709	228

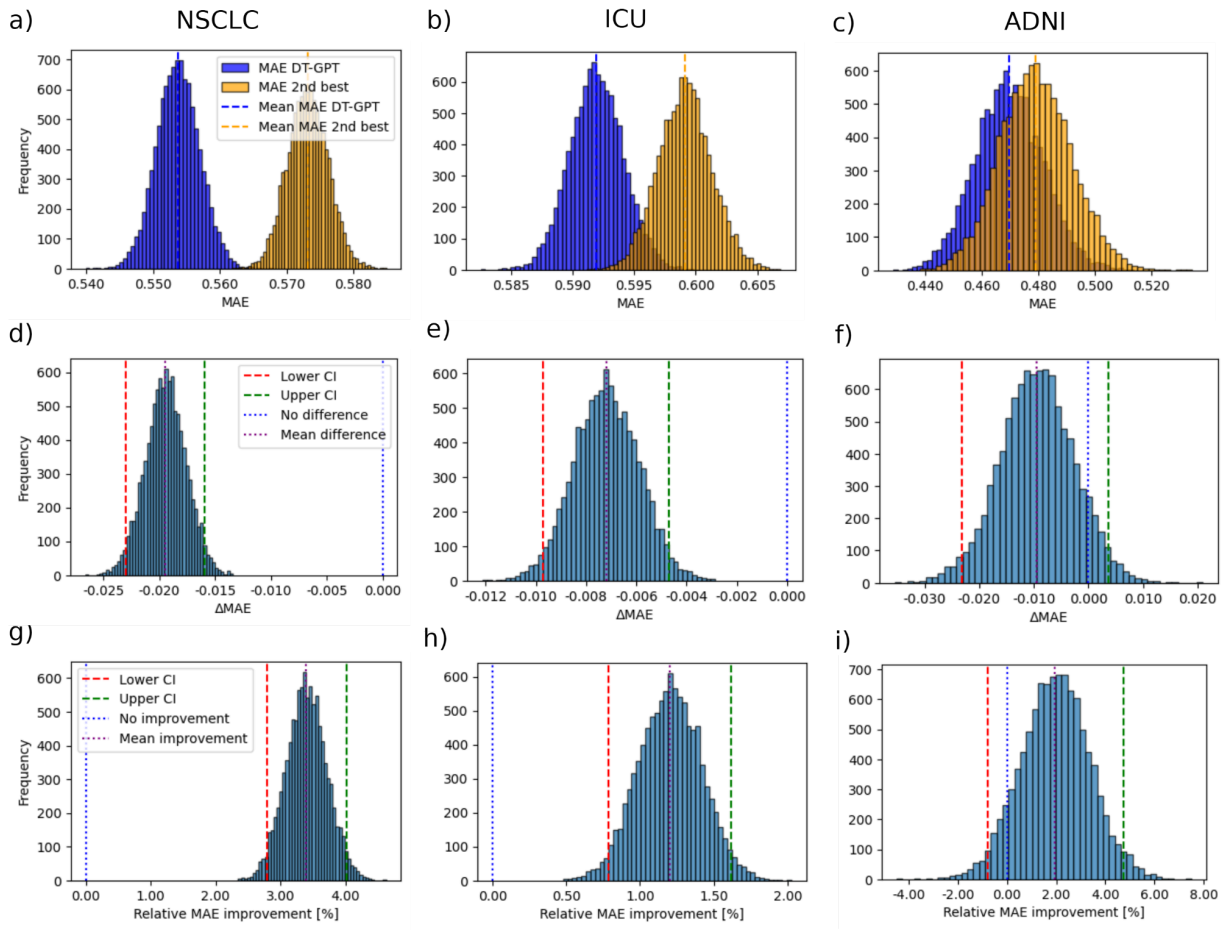
Supplementary Table 19: Comparison of DT-GPT against second best performing model on clinical variable level. One-tailed Wilcoxon signed-rank test revealed that DT-GPT significantly outperformed the second-best performing model (LightGBM for the NSCLC and ICU datasets, TFT for the ADNI dataset) at the confidence level of $\alpha = 0.05$ for 9 out of 12 variables. Asterisks indicate variables with significant test results.

Supplementary Figure 1. DT-GPT significantly outperforms the second best performing model on 9 out of 12 variables from the NSCLC, ICU and ADNI datasets.



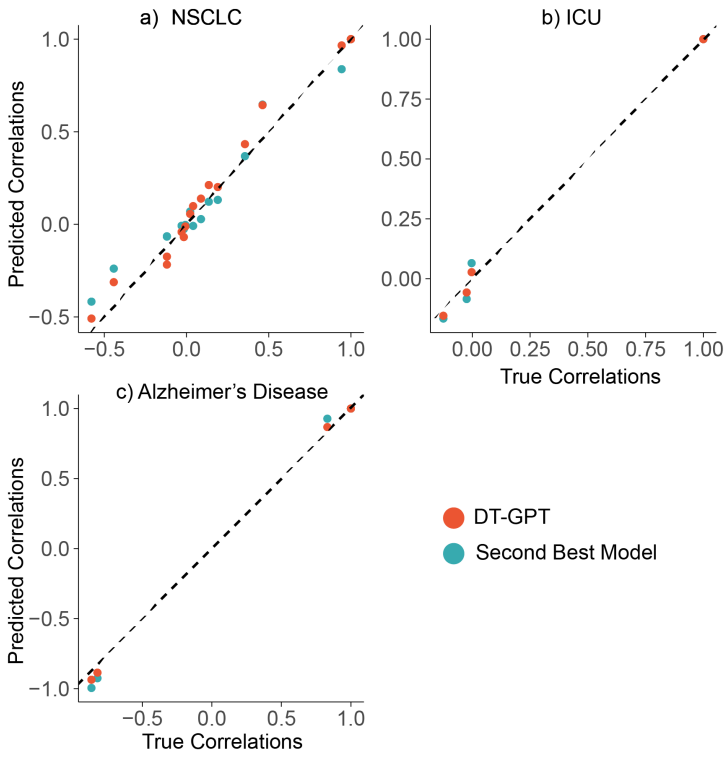
Supplementary Figure 1: DT-GPT significantly outperforms the second best performing model on 9 out of 12 variables from the NSCLC, ICU and ADNI datasets. P-values for one-tailed Wilcoxon signed-rank test with Benjamini-Hochberg correction (false discovery rate (FDR) of 5%) are reported, whereby DT-GPT has significantly systematically lower errors than LightGBM on all 6 variables for the NSCLC dataset, on 2 out of 3 variables for the ICU dataset and significantly lower errors than TFT on 1 out of 3 variables for the ADNI dataset. We note that the ADNI dataset with 1,140 patients only is the smallest dataset and benefits less from the foundation model strengths of DT-GPT.

Supplementary Figure 2. Bootstrapping showed significant relative MAE improvement of DT-GPT over the second best performing model for NSCLC and ICU datasets.



Supplementary Figure 2: Bootstrapping showed significant relative MAE improvement of DT-GPT over the second best performing model for NSCLC and ICU datasets. To analyze the robustness of the improvements of DT-GPT over the second best model (2nd best) for all three datasets, we apply paired bootstrapping on the respective test sets. The overall frequency of the paired MAE for the NSCLC, ICU and Alzheimer's disease (ADNI) datasets are depicted in **a), b)** and **c),** respectively. In **d), e)** and **f)**, the histogram of MAE differences of DT-GPT to the second best performing model (Δ MAE) are shown. Finally, in **g), h)** and **i),** relative MAE improvement is evaluated from DT-GPT to the second best performing model.

Supplementary Figure 3. DT-GPT achieved high inter-variable correlations.



Supplementary Figure 3: **DT-GPT achieves high inter-variable correlations. a)** Inter-variable correlations of non-small cell lung cancer (NSCLC) target variables for DT-GPT and second best baseline LightGBM. The closer the points are to the dotted diagonal, the closer they are to the observed inter-variable correlation. Each dot represents one variable pair, e.g. neutrophils and lymphocytes, with the x value representing the true correlation, and the y value the correlation in the predictions. **b)** Inter-variable correlations of intensity care unit (ICU) target variables for DT-GPT and second best baseline LightGBM. **c)** Inter-variable correlations of Alzheimer's disease target variables for DT-GPT and second best baseline TFT.

Supplementary Table 20. Ablation study using 1140 patients on NSCLC and ICU datasets.

Scaled Mean Absolute Error	Model	Non-Small Cell Lung Cancer (NSCLC) - 1140 patient train subset/full test set						Intensive Care Unit - 1140 train subset/full test set			Alzheimer's Disease (reference)		
		Hemo-globin	Leuko-cytes	Lympho-cytes/ Leuko-cytes	Lympho-cytes	Neutro-phils	Lactate Dehydro-genase	Magne-sium	Resp. Rate	Oxygen Saturation	MMSE	CDSRB	ADAS11
Channel-Independent Input	Copy Forward	0.698	0.969	0.731	0.569	0.974	**0.433**	0.681	0.769	0.746	0.654	0.539	0.519
	PatchTST	0.643	0.885	0.697	0.531	0.862	0.435	0.672	**0.644**	**0.661**	0.654	0.540	0.506
	Time-LLM	0.658	0.869	0.687	0.542	0.848	0.463	0.655	0.660	0.679	0.654	0.540	0.506
	LLMTime	0.736	0.923	0.725	0.601	0.900	0.437	0.759	0.686	0.688	0.654	0.539	0.503
Channel-Dependent Input	BioMistral-7B	0.984	1.097	0.756	0.997	1.953	0.600	0.790	0.770	0.945	2.064	0.883	0.728
	Qwen3-32B	0.670	0.942	0.736	0.573	0.937	0.453	0.709	0.720	0.791	0.686	0.546	0.555
	TCN	0.693	0.893	0.776	0.649	0.866	0.705	0.660	0.753	0.758	*Not Applicable*		
	Linear Regression	0.582	0.893	0.738	0.629	0.887	0.493	1.440	1.548	1.627	0.551	0.449	**0.457**
	RNN	0.557	0.842	0.690	0.552	0.820	0.474	0.578	0.694	0.688	0.545	0.463	0.465
	Transformer	0.577	0.873	0.761	0.614	0.828	0.599	0.680	0.722	0.757	0.553	0.485	0.481
	LSTM	0.551	0.890	0.681	0.554	0.855	0.465	0.596	0.659	0.663	0.545	0.468	0.475
	Temporal Fusion Transformer	0.557	0.856	0.740	0.554	0.821	0.694	0.675	0.725	0.720	**0.520**	0.451	0.466
	TiDE	0.536	0.819	0.712	0.542	0.829	0.511	0.624	0.686	0.697	0.578	0.498	0.506
	LightGBM	0.495	0.801	**0.677**	0.509	0.801	0.487	0.581	0.683	0.683	0.540	0.455	0.462
	DT-GPT (ours)	**0.490**	**0.747**	0.684	**0.476**	**0.724**	0.531	**0.563**	0.667	0.682	0.535	**0.417**	0.458

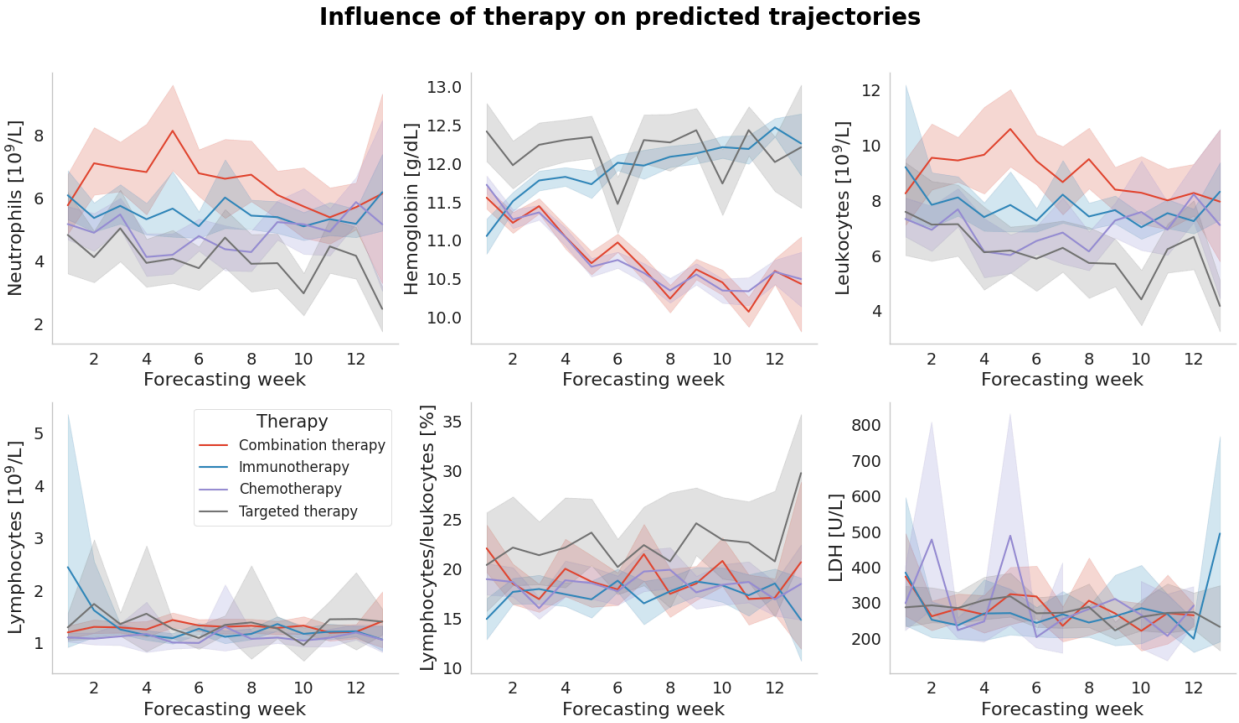
Supplementary Table 20: Comparison of models using a 1140 randomly subsampled training set and evaluated on the standard, full test set. To compare the effect of dataset size on model performance consistently across datasets, we randomly subsample 1,140 patients from the NSCLC and ICU training datasets, stratified by age and gender. We then process the data as described previously, train the models and evaluate them on the full dataset. Looking at the performance, we see it dropping across all datasets, as expected, with DT-GPT outperforming the baselines in six out of 12 variables.

Supplementary Table 21. The percentage of trajectories explained by the most important variables and patient baseline characteristics for the NSCLC forecasting.

Variable	**% of predicted trajectories it explains according to DT-GPT**
therapy	87.0
ECOG	55.6
leukocytes	44.5
age	36.2
alanine aminotransferase	26.0
hemoglobin	21.8
neutrophils	21.3
lymphocytes/100 leukocytes	21.0
body weight and body height	19.4
body height	19.4
albumin	16.5
gender	15.8
lymphocytes	15.8
lactate dehydrogenase 2	12.3
alkaline phosphatase	9.5
ferritin	7.2

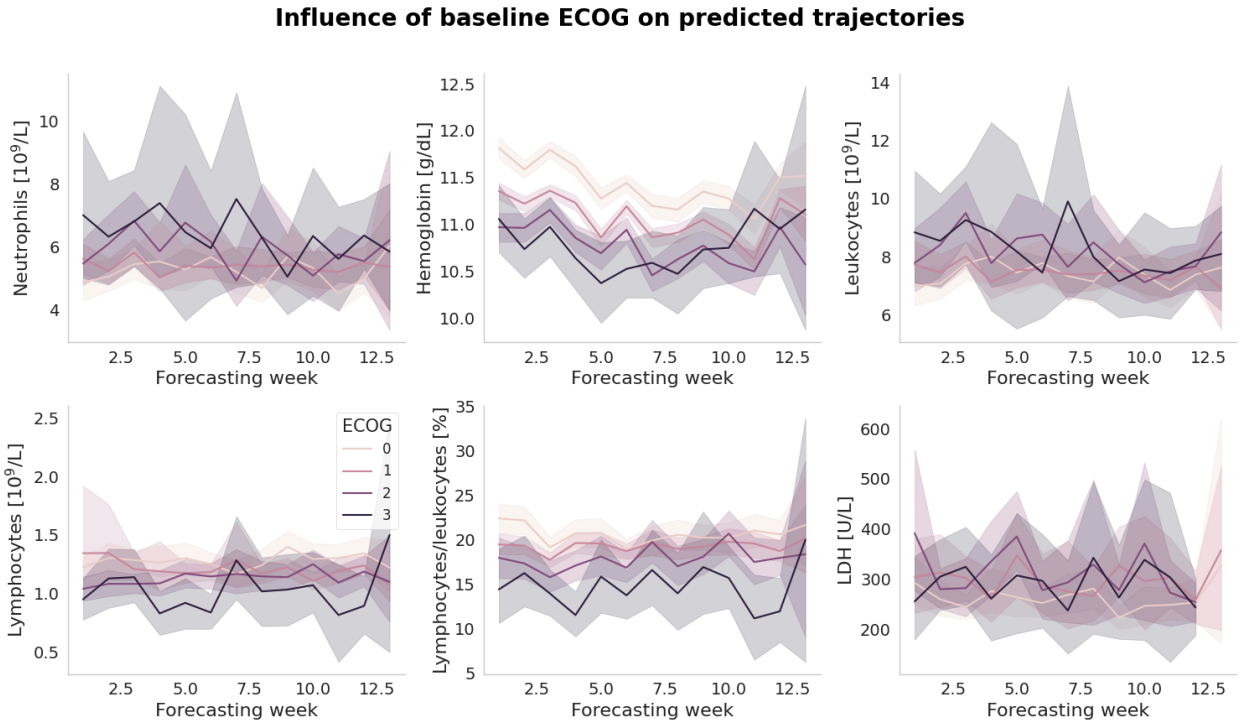
Supplementary Table 18: The percentage of trajectories explained by the most important variables and patient baseline characteristics for the NSCLC forecasting.

Supplementary Figure 4. Influence of the therapy type on the predicted NSCLC trajectories.



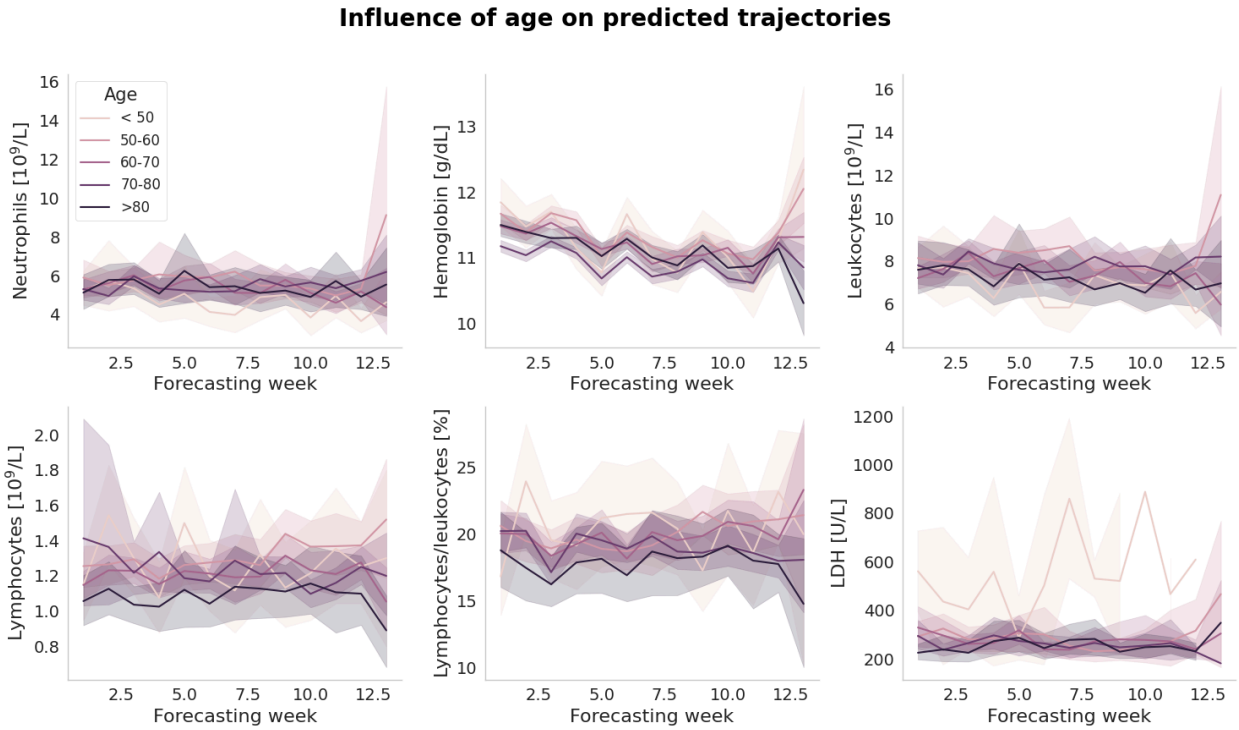
Supplementary Figure 4: Influence of the therapy type on the predicted NSCLC trajectories. The most important variable, therapy particularly influences the predicted dynamics of neutrophils, hemoglobin, leukocytes and lymphocytes to leukocytes ratio. Here, lines represent average trajectories with 95% confidence intervals calculated through bootstrapping.

Supplementary Figure 5. Influence of the baseline ECOG values on the predicted NSCLC trajectories.



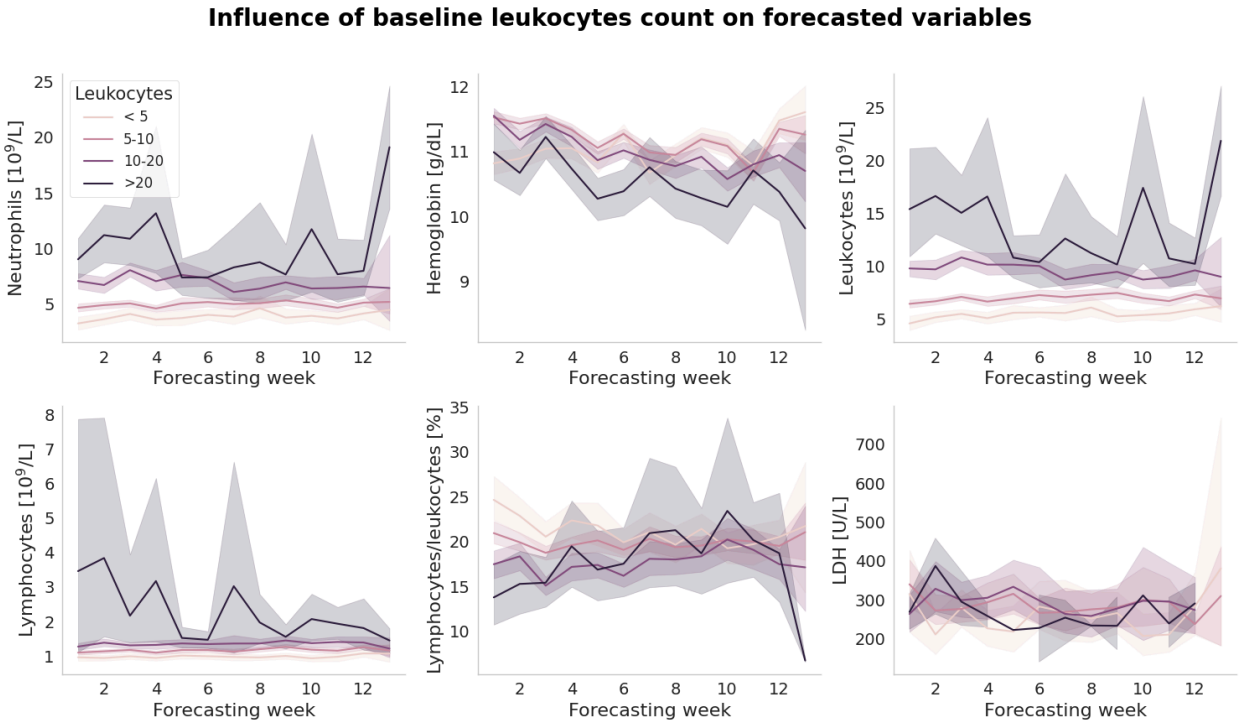
Supplementary Figure 5: Influence of the baseline ECOG values on the predicted NSCLC trajectories. The second most important variable, ECOG, particularly influences the predicted dynamics of hemoglobin and lymphocytes to leukocytes ratio. Here, lines represent average trajectories with 95% confidence intervals calculated through bootstrapping.

Supplementary Figure 6. Influence of the age on the predicted NSCLC trajectories.



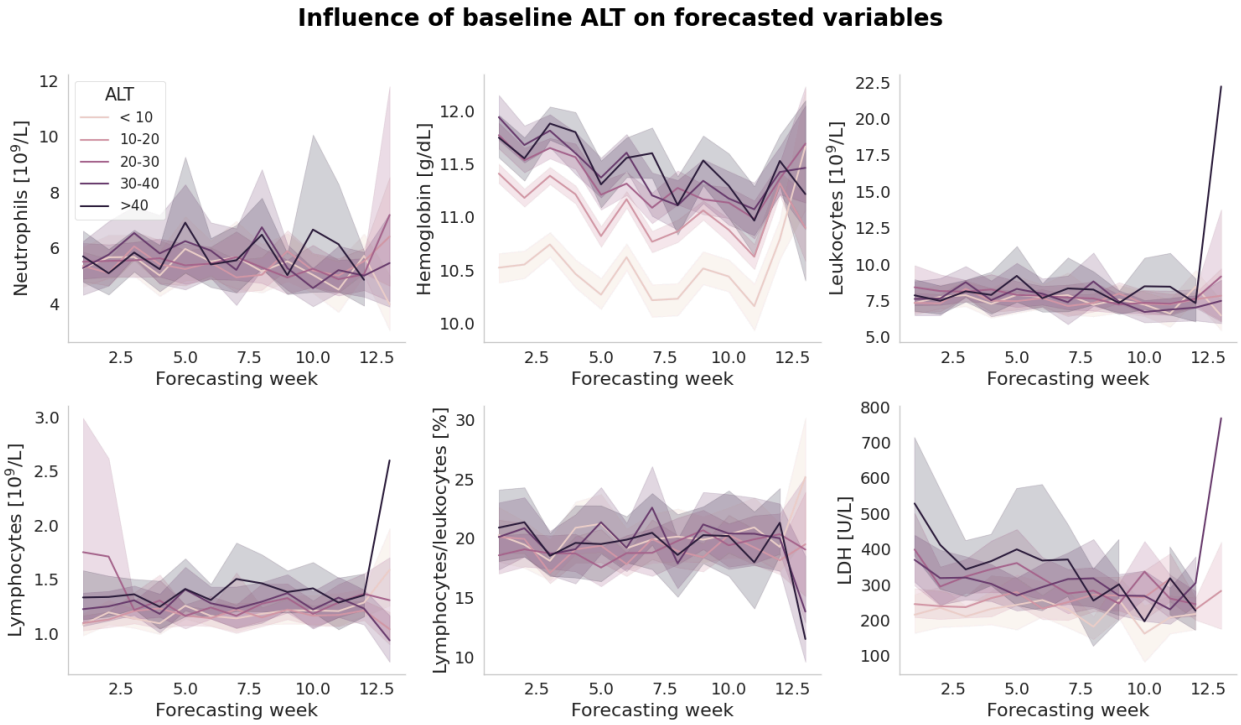
Supplementary Figure 6: Influence of the age on the predicted NSCLC trajectories. The third most important variable, age, particularly influences the predicted dynamics of lactate dehydrogenase (LDH), whereby younger patients (less than 50 years old) have on average higher LDH values. Here, lines represent average trajectories with 95% confidence intervals calculated through bootstrapping.

Supplementary Figure 7. Influence of the baseline leukocytes count on the predicted NSCLC trajectories.



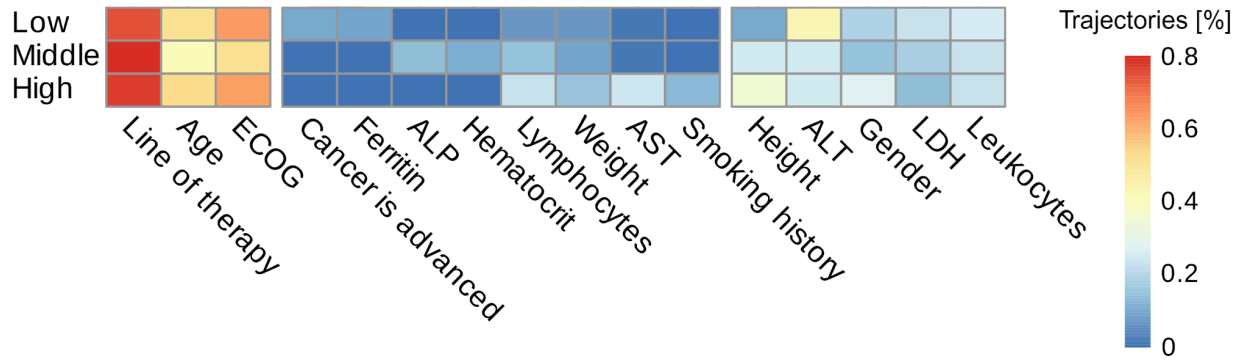
Supplementary Figure 7: Influence of the baseline leukocytes count on the predicted NSCLC trajectories. The fourth most important variable, leukocytes, particularly influences the predicted dynamics of neutrophils, hemoglobin, leukocytes and lymphocytes to leukocytes ratio. Here, lines represent average trajectories with 95% confidence intervals calculated through bootstrapping.

Supplementary Figure 8. Influence of the baseline alanine aminotransferase (ALT) values on the predicted NSCLC trajectories.



Supplementary Figure 8: Influence of the baseline alanine aminotransferase (ALT) values on the predicted NSCLC trajectories. The fifth most important variable, alanine aminotransferase (ALT), particularly influences the predicted dynamics of neutrophils, hemoglobin. Here, lines represent average trajectories with 95% confidence intervals calculated through bootstrapping.

Supplementary Figure 9. DT-GTP considers different variables when predicting low, middle and high level hemoglobin trajectories.



Supplementary Figure 9: DT-GPT considers different variables when predicting low, middle and high level hemoglobin trajectories. Complex and non-trivial explanatory abilities of DT-GPT are highlighted in an example of hemoglobin trajectories. A set of obtained important variables for predicted hemoglobin trajectories is not constant across all trajectories, but is correlated with the predicted values. We performed quantile-based clustering of predicted hemoglobin trajectories by their mean values over time and assign each trajectory to the low level group (mean value less than 9.76 g/dL; 0.15 quantile), to the high level group (mean value greater than 13.31 g/dL; 0.85 quantile) and to the middle level group (mean value between 9.76 and 13.31 g/dL). For each of the groups we consider 10 most important variables separately and determine their intersection to define a final set of 16 important variables. The fractions of important variables explaining hemoglobin trajectories (e.g., relative frequency of a variable to be in the set of important variables as outputted by DT-GPT for the trajectories in each group) are different for each of the hemoglobin group. While therapy, age and ECOG are present in each of the groups, the ferritin and "cancer is advanced" variables have a slightly higher prevalence in the lower group whilst gender is more frequent for the high hemoglobin level group. Chi-squared test for independence rejected the null hypothesis (p-value $<$ 2.2e-16) indicating that DT-GPT might put more weight on different variables when predicting hemoglobin trajectories of different levels. However, our analysis only reveals correlations between hemoglobin level and important variables, thus causal claims are out of scope. (Abbreviations: ECOG - Eastern Cooperative Oncology Group performance status scale, LDH - Lactate dehydrogenase, ALT - Alanine aminotransferase, ALP - Alkaline phosphatase , AST - Aspartate aminotransferase).

Supplementary Note 3. Analysis of prediction reasoning.

We visualize and investigate the influence of the five most important variables, therapy (**Supplementary Figure 3**), ECOG (**Supplementary Figure 4**), age (**Supplementary Figure 5**), leukocytes (**Supplementary Figure 6**) and alanine aminotransferase (**Supplementary Figure 7**) on the dynamics of predicted NSCLC output variables. While the variables were marked as important for the prediction of six output variables simultaneously, we note that most of them particularly influence dynamics in only a subset of output variables. For instance, therapy is correlated with neutrophils, hemoglobin and leukocytes trajectories (**Supplementary Figure 3**), ECOG with hemoglobin and lymphocytes to leukocytes ratio trajectories (**Supplementary Figure 4**), and age with lactate dehydrogenase (**Supplementary Figure 5**), respectively.

For the therapy, we consider 10 most frequent therapies and group them into therapy group as follows: carboplatin & paclitaxel, carboplatin & pemetrexed and pemetrexed for the chemotherapy group (450 patients), pembrolizumab, nivolumab and durvalumab for the immunotherapy group (598 patients), carboplatin & pembrolizumab & pemetrexed, docetaxel & ramucirumab, pembrolizumab & pemetrexed for the combination therapy group (434 patients) and osimertinib for the target therapy group (112 patients), respectively. For the ECOG variable, we consider values of 0 (765 patients), 1 (1299 patients), 2 (384 patients) and 3 (85 patients). For the age, we define the following groups based on the age histogram: younger than 50 years old (124 patients), between 50 and 60 years old (474 patients), between 60 and 70 years old (911 patients), between 70 and 80 years old (981 patients) and older than 80 years old (261 patients). For leukocytes, we consider the last observed value within 13 weeks of medical history prior to the start of the treatment as a baseline value, and based on the histogram of leukocytes baseline values combined with the reference intervals for the leukocytes define four groups: less than 5 10^9/L (315 patients), between 5 and 10 10^9/L (1093 patients), between 10 and 20 10^9/L (488 patients) and more than 20 10^9/L (66 patients). Similarly, for the alanine aminotransferase (ALT) we consider the last observed value within 13 weeks of medical history prior to the start of the treatment as a baseline value, and based on the histogram of ALT baseline values combined with the reference intervals for the ALT define four groups: less than 10 U/L (365 patients), between 10 and 20 U/L (1090 patients), between 20 and 30 U/L (602 patients), between 30 and 40 U/L (249 patients) and more than 40 U/L (246 patients). For each of the most important variables, only predicted trajectories of patients with available variable values and ground truth for output variables were analyzed.

Since we ask DT-GPT to explain all output variables simultaneously, the observed important variables do not lead to a "perfect" separation of predicted trajectories conditioned on the variable value. For instance, if we perform forecasting of hemoglobin trajectories only, further variables are said to be important by DT-GPT (**Supplementary Figure 8**). Furthermore, the choice of groups in the following analysis can be made arbitrary and might influence the results. A better grouping approach would include other variables such as age, gender or other demographics or laboratory test data. Specifically, hemoglobin and ALT have different reference intervals for males and females. Such interactions were outside of the scope of this analysis. Finally, we note that we can establish only the correlation between obtained important variables and predicted trajectories; the causal relationships are more complex and are subject to further investigation.

Supplementary Table 22. Zero-shot performance results.

Variable	LOINC	MAE LightGBM	MAE DT-GPT
ferritin	2276-4	0.07	0.03
lactate dehydrogenase	14804-9	0.55	0.14
erythrocytes	789-8	0.15	0.27
carcinoembryonic ag	2039-6	0.29	0.29
erythrocytes 2	26453-1	0.26	0.37
neutrophils.segmented	30451-9	0.71	0.45
lymphocytes 3	731-0	0.48	0.47
neutrophils.band form	26507-4	0.65	0.55
neutrophils 2	751-8	0.70	0.58
lymphocytes/100 leukocytes	736-9	0.66	0.60
lymphocytes	732-8	0.68	0.60
granulocytes	30394-1	0.78	0.67
leukocytes	6690-2	0.71	0.69
lymphocytes/100 leukocytes 2	737-7	0.75	0.70
bilirubin.non-glucuronidated	1971-1	0.57	0.71
monocytes	26484-6	0.72	0.71
granulocytes 2	20482-6	0.76	0.72
alkaline phosphatase	6768-6	0.46	0.78

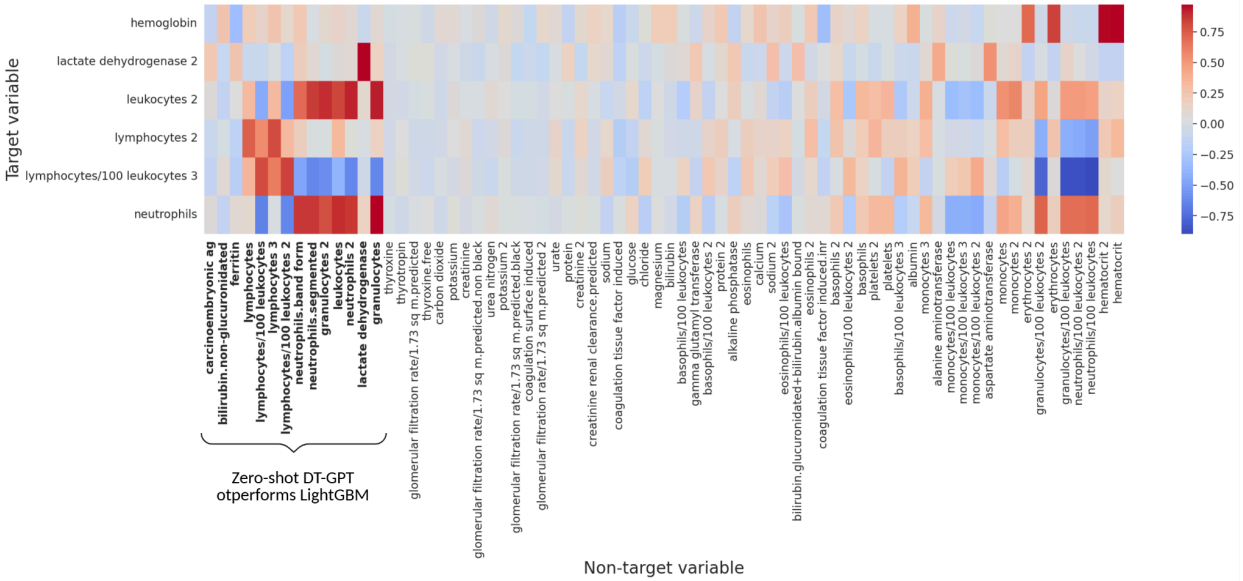
Variable	LOINC	MAE LightGBM	MAE DT-GPT
coagulation tissue factor induced	5902-2	0.58	0.81
urea nitrogen	3094-0	0.57	0.84
glomerular filtration rate/1.73 sq m.predicted 2	69405-9	0.40	0.88
protein	2888-6	0.60	0.93
monocytes/100 leukocytes 2	5905-5	0.78	0.93
platelets 2	777-3	0.64	0.97
glomerular filtration rate/1.73 sq m.predicted.black	48643-1	0.39	1.05
platelets	26515-7	0.64	1.07
hematocrit 2	20570-8	0.49	1.10
glomerular filtration rate/1.73 sq m.predicted.non black	48642-3	0.39	1.12
calcium	17861-6	0.64	1.14
creatinine renal clearance.predicted	35591-7	0.33	1.16
potassium 2	6298-4	0.71	1.17
coagulation tissue factor induced.inr	38875-1	0.33	1.27
urate	3084-1	0.51	1.28
bilirubin.glucuronidated+bilirubin.albumin bound	1968-7	0.60	1.32
monocytes 2	742-7	0.75	1.34
monocytes 3	743-5	0.81	1.37

Variable	LOINC	MAE LightGBM	MAE DT-GPT
eosinophils 2	26449-9	0.46	1.38
alanine aminotransferase	1742-6	0.64	1.38
glucose	2345-7	0.66	1.38
monocytes/100 leukocytes	26485-3	0.76	1.41
sodium 2	2947-0	0.67	1.57
coagulation surface induced	14979-9	0.80	1.57
monocytes/100 leukocytes 3	744-3	0.71	1.64
eosinophils/100 leukocytes 2	714-6	0.65	1.65
hematocrit	4544-3	0.50	1.76
granulocytes/100 leukocytes 2	19023-1	0.72	1.83
carbon dioxide	2028-9	0.61	1.87
glomerular filtration rate/1.73 sq m.predicted	98979-8	0.57	1.88
bilirubin	1975-2	0.57	1.99
protein 2	2885-2	0.56	2.00
albumin	1751-7	0.53	2.02
potassium	2823-3	0.70	2.24
sodium	2951-2	0.65	2.47
creatinine 2	38483-4	0.40	2.52
granulocytes/100 leukocytes	30395-8	0.71	2.59
aspartate aminotransferase	1920-8	0.60	2.6

Variable	LOINC	MAE LightGBM	MAE DT-GPT
chloride	2075-0	0.63	2.69
eosinophils/100 leukocytes	26450-7	0.57	2.82
creatinine	2160-0	0.37	2.84
magnesium	19123-9	0.60	2.96
neutrophils/100 leukocytes	26511-6	0.71	3.04
basophils	704-7	0.60	3.06
neutrophils/100 leukocytes 2	770-8	0.73	3.09
eosinophils	712-0	0.54	3.12
basophils/100 leukocytes 3	706-2	0.69	3.16
basophils/100 leukocytes	707-0	0.73	3.21
basophils 2	26444-0	0.64	3.34
basophils/100 leukocytes 2	30180-4	0.70	3.35
gamma glutamyl transferase	2324-2	0.37	3.57

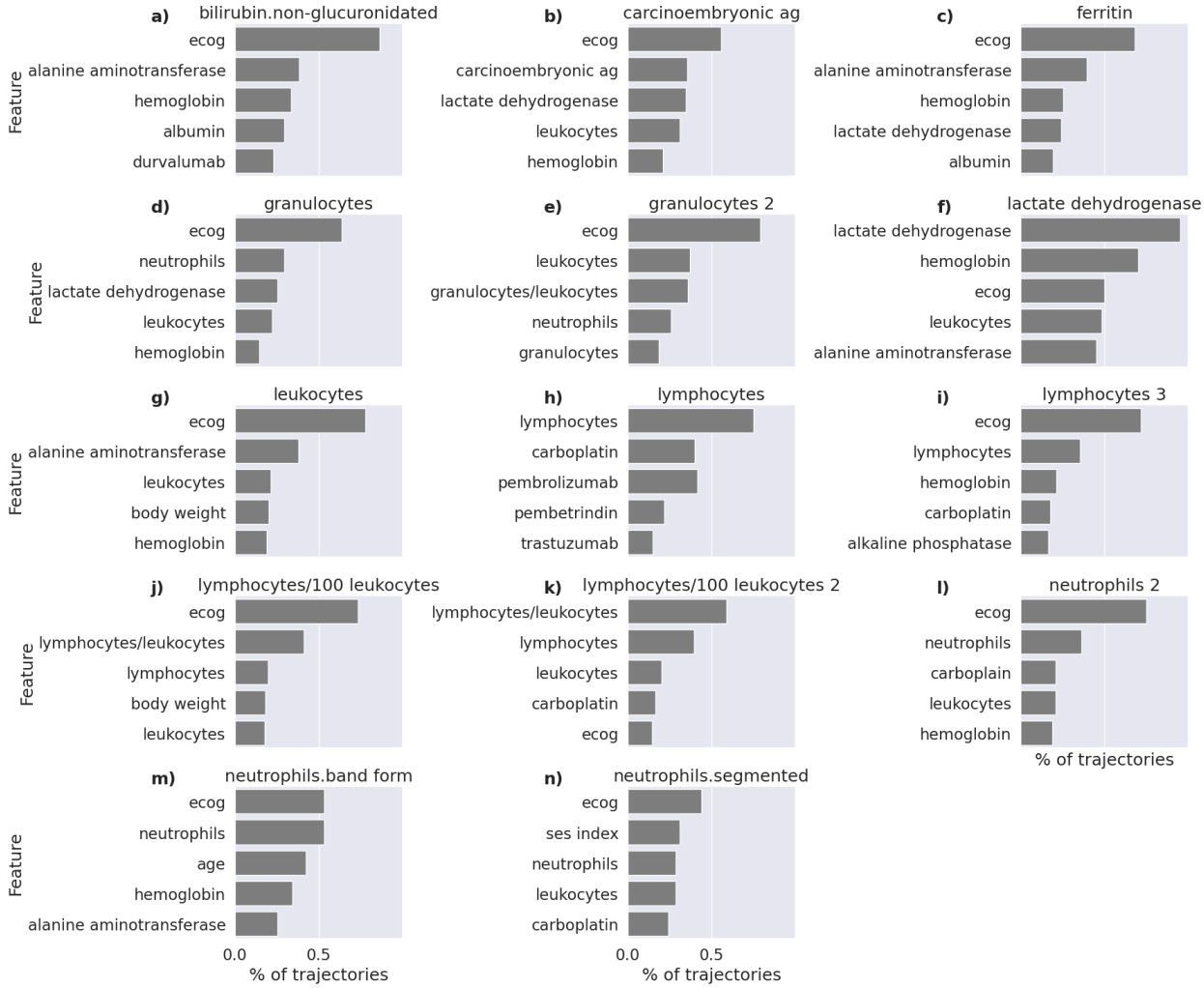
Supplementary Table 22: Zero-shot performance results.

Supplementary Figure 10: Correlations between target and non-target variables.



Supplementary Figure 10. Non-target variables, on which DT-GPT outperforms LightGBM, mostly exhibit higher correlations with target variables than other zero-shot variables. 11 of 13 non-target variables with better (or equal, including CEA) DT-GPT performance have high Spearman correlation coefficient values ($|\rho| > 0.7$) with at least one of the target variables. Only carcinoembryonic ag, bilirubin.non-glucuronidated and ferritin have low Spearman correlation coefficient values ($|\rho| < 0.33$) with all of the target variables. Correlations between variables were calculated on the training set. LOINC codes uniquely identifying each non-target variable can be found in **Supplementary Table 22**.

Supplementary Figure 11. Analysis of feature importance for relevant zero shot variables.



Supplementary Figure 11: Interpretability approach applied to the non-target variables reveals general and latent clinical knowledge. (a)-(m): five most important features and percentage of predicted trajectories (% of trajectories) explained by them for each non-target variable on which DT-GPT outperforms LightGBM. While ECOG is selected as an important feature for almost all variables, most variables include key/target variables that are strongly correlated with variables in the list of important features (e.g., neutrophils 2 and neutrophils have Spearman correlation coefficient of 0.866). LOINC codes uniquely identifying each non-target variable can be found in Supplementary Table 22.

Supplementary Table 23. Zero shot latent knowledge overview.

Zero-shot Variable	Key Model -Identified Features in Patient's History	Direct Clinical Interpretation	Hypotheses of Latent Knowledge (Second/Third Order Insight)	Refs.
Ferritin	ECOG, ALT, Hemoglobin, LDH, Albumin	Ferritin is an acute phase reactant whose levels rise with inflammation, tumor burden, and cell turnover. It is a prognostic marker in NSCLC, correlating with poor patient condition (ECOG), anemia of chronic inflammation (low Hemoglobin), liver stress (ALT), and cell death (LDH).	The model has potentially learnt the components of a composite biomarker: The ferritin-to-hemoglobin ratio, which captures the interplay between inflammation and anemia, is a validated independent prognostic factor in advanced NSCLC.	5–8
Carcinoembryonic Ag (CEA)	ECOG, LDH, Hemoglobin, Leukocytes, CEA	CEA is a tumor marker associated with high tumor burden. Elevated CEA correlates with poor performance status (ECOG), systemic inflammation (Leukocytes), high cell turnover (LDH) and anemia (low Hemoglobin).	The model has potentially learnt the prognostic value of CEA and Hemoglobin, reflecting a holistic "Advanced Disease Syndrome." The prognostic significance of an elevated CEA is possibly modulated by the patient's hematological status (anemia). However, the combined prognostic factor is not yet robustly validated.	9
Bilirubin (non-glucuronidated)	ECOG, Albumin, ALT, Hemoglobin, Durvalumab	Unconjugated bilirubin is a product of hemoglobin breakdown, is transported in the blood bound to albumin, and is conjugated by the liver (function assessed by ALT). Its level reflects the balance of production, transport, and hepatic clearance.	The model has potentially learnt part of the components of the ALBI (Albumin-Bilirubin) score, an emerging prognostic marker for survival in NSCLC patients receiving immunotherapy, such as possibly durvalumab. Note that, in the ALBI score, total bilirubin is used.	10–12

Supplementary Table 23: Analysis of the most important features of the zero-shot model and highlighting the model's latent potentially learnt clinical knowledge. We exemplify the potentially learnt latent clinical knowledge by analyzing three non-trivially zero-shot variables for which DT-GPT outperformed, or performed on par with,

LightGBM. To find non-trivial variables, we first ranked variables that had the same or lower MAE than LightGBM by their maximum absolute Spearman correlation to the trained variables, and selected the lowest three. By selecting the lowest, we focus on variables that are not spuriously correlated with those that the model learnt, but rather those for which the model could have leveraged its wider pre-trained and fine-tuned knowledge. In this table, we briefly explore related literature between the three variables with the lowest correlation and the most important features. The results highlight the potential of latent clinical knowledge captured by the model but further investigation is required to validate it. The ALBI Score was reported as a prognostic factor in NSCLC patients on immunotherapy. Additionally, the CEA analysis shows the potential of a relationship between CEA and hemoglobin. Finally, the ferritin analysis finds that the model is possibly considering the potential prognostic factor ferritin-to-hemoglobin ratio.

Supplementary Note 4. 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 or date. For each patient visit and for each observed value, we note down the 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 in the 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, for each variable, if the value was observed in the patient's history, it will be forward propagated, ensuring that we have the information even if the context length is normally not long enough. Here, we present synthetic examples of both the manual template 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: {"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."}

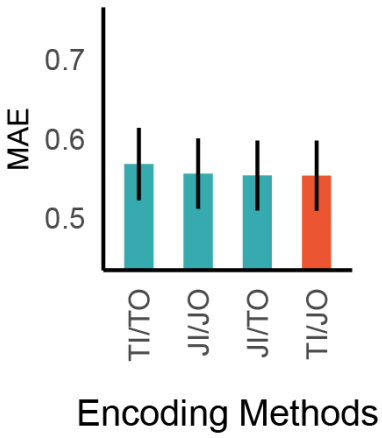
Manual Template Output (Synthetic Patient)

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", "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"]}

Supplementary Figure 12. DT-GPT performance with respect to different encoding methods.



Supplementary Figure 12: DT-GPT performance with respect to different encoding methods. The performance is measured by scaled mean absolute error (MAE). Here, abbreviations are as follows: TI - "Text Input", TO - "Text Output", JI - "JSON Input" and JO - "JSON Output". DT-GPT is stable with respect to different data encoding strategies though 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, JI/TO reaches 0.554 ± 0.04, TI/JO attains 0.554 ± 0.04. Error bars refer to the standard error of the encoding experiment results aggregated across all variables.

Supplementary Note 5. Hyperparameters for model fine-tuning and inference time.

We initially experimented with applying the loss to the entire sequence, which would also allow generating synthetic patients, however the models hallucinated to an unusable 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 3,400 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 4,000 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 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.

Supplementary Note 6. Baseline details.

The LightGBM, TFT, TCN, RNN, LSTM, Transformer and TiDE baseline models are implemented in the Darts library and the default hyperparameters were 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, and the baseline measurements in the Alzheimer's disease 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. We apply the filtering based on three standard deviations, as with DT-GPT, 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, with 1 for every future date that will be measured and 0 for those that are imputed. Using TCN is not feasible for the Alzheimer's disease dataset, since it requires a longer input sequence than output in the Darts implementation.

PatchTST and Time-LLM are implemented in the NeuralForecast library. For Time-LLM, the Llama2-7B model is used as the base LLM, and both models were explored on 91-day lookback for NSCLC, ensuring a comprehensive patient view. All other parameters are set as default, and are applied the same procedure as the Darts models, including imputation.

For LLMTime, following the original paper, Llama2-70B is used as the prediction model. Since the model can deal with missing data, no imputation is performed. All further aspects are implemented as in the original implementation.

Supplementary Note 7. Chatbot and zero-shot 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.

…

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 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”]]}

Supplementary Note 8. Interpretation of forecasting and classification metrics.

Forecasting metrics

MAE measures the average magnitude of the errors and operates on the original scale of the variable enabling intuitive understanding of model performance. $scaled\ MAE$, in turn, is scaled independently and allows comparison across all variables, hence can be used as a primary score to benchmark different models and select the best performing approach. $MASE$ scales MAE by the mean absolute error of a naive forecast, which is, in a multi timestep forecasting setting such as ours, equivalent to the copy forward forecast. $SMAPE$ quantifies the accuracy of a forecast by calculating the average of the absolute percentage errors between predicted and actual values, adjusted to be symmetric and, similarly to MAE, gives an intuitive understanding of model performance. Finally, Spearman correlation coefficient $Spearman\ \rho$ assesses the strength and direction of monotonic relationships, providing insights into how well the forecasted values preserve the order of the actual values. In the context of clinical signals, a Spearman correlation coefficient above 0.4 is in some cases interpreted as "moderate" and above 0.7 as "strong" correlation[3]. Furthermore, the Spearman correlation was shown to be extremely sensitive to fluctuations within a narrow range of values[13], which is typical of laboratory measurement data, and hence should be interpreted only in combination with other statistical properties (e.g. Kolmogorov-Smirnov test).

Classification metrics

Classification metrics quantify model performance with respect to different properties of a clinical trajectory. AUC_{low}, AUC_{high} and $AUC_{weighted}$ assess the model's ability to detect absolute deviation from the variable normal range as defined for all patients. For some of the variables (e.g. hemoglobin) only AUC_{low} and might be interesting, whereas for other (e.g. lactate dehydrogenase) only increase, i.e., AUC_{high} is relevant. The metrics $AUC_{trend\downarrow}$ and $AUC_{trend\uparrow}$ assess the model's ability to detect directed changes in variable value that might be associated with chronic conditions and require closer patient observation.

Supplementary Note 9. Research in Context.

Evidence before this study

Digital Twins (DTs) for forecasting patient-specific clinical trajectories, events and outcomes are being increasingly realized by the means of generative artificial intelligence (AI). We conducted a comprehensive search using Google Scholar and Scopus for studies and reviews on methods for predicting longitudinal patient trajectories, published in English between Jan 1, 2019, and March 31, 2024. The search employed the following keyword combinations: “forecasting” and “patient trajectory”, “forecasting” and “patient”, “forecasting” and “clinical”, “prediction” and “patient trajectory” and “time”, ”forecasting and electronic health records”, “prediction” and “patient” and “time”, “prediction” and “clinical” and “time”, “prediction” and “patient trajectory” and “longitudinal”, “prediction” and “patient” and “longitudinal”, “prediction” and “clinical” and “longitudinal”, “foundation model” and “electronic health records”, “large language model” and “electronic health records”. We restricted the search results to the studies describing generative AI-based methods. The identified studies introduce high-performing clinical forecasting methods but their validation is limited to a single application only with a specific disease indication or dataset. Additionally, these methods typically require extensive data preprocessing, such as imputation of missing values and normalization. While large language models and foundation models offer a more general setup applicable to various research questions and datasets, existing methods predominantly focus on single time-point predictions rather than longitudinal predictions. Moreover, the interpretability of existing models is limited.

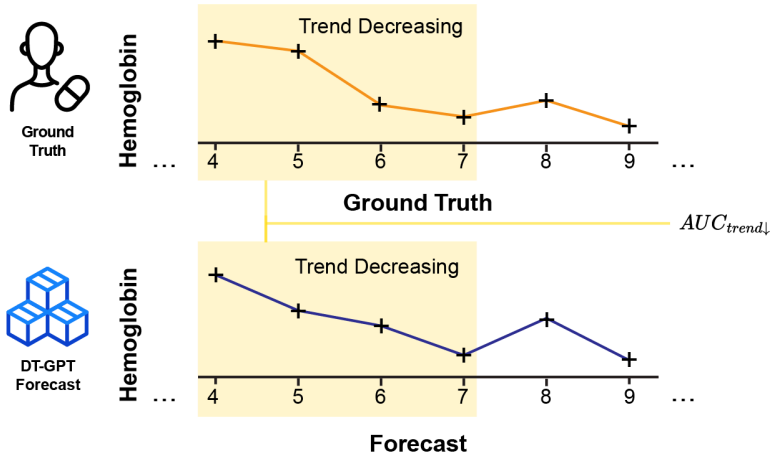
Added value of this study

This study introduces the Digital Twin - Generative Pretrained Transformer (DT-GPT) model, a novel method to fine-tune large language models (LLMs) to forecast multi-variable patient trajectories, combining the advantages of existing methods while overcoming the limitations of data heterogeneity and sparsity associated with electronic health records (EHRs). DT-GPT achieves state-of-the-art forecasting performance on long-term US nationwide non-small cell lung cancer datasets, a short-term intensive care unit dataset and an Alzheimer’s disease dataset, demonstrating its applicability to multiple disease conditions with various time horizons and both regular and irregular time sampling. Furthermore, DT-GPT learns relationships and preserves cross-correlation between variables, enabling zero-shot (i.e., without any training) prediction of clinical variables previously not trained on. Finally, the interactive interface provides preliminary prediction explainability through chatbot functionality.

Implications of all the available evidence

Generative AI-based models enhance the capabilities of patient DTs for treatment selection, patient monitoring, and clinical trial support by creating state-of-the-art patient trajectory predictions. DT-GPTshows that LLMs are able to advance the development of DTs by reducing the need for extensive data preprocessing and enabling interaction with the model through a human-interpretable interface. We anticipate that LLM-based DTs will be easily accessible to clinicians, allowing efficient simulations of patient trajectories under various scenarios to support clinical decision-making.

Supplementary Figure 13. Illustration of the trend definition used for evaluating time-series forecasts.



Supplementary Figure 13: **Illustration of the trend definition used for evaluating time-series forecasts.** This example shows how a 'decreasing trend' label is assigned to a forecasted hemoglobin value at week 7. Using a 3-week trend window ($s = 3\,weeks$) and weekly resolution, the label requires a consistent week-on-week decrease in the forecast across the entire relevant period (weeks 4 to 7). This specific prediction ('decreasing trend') is then evaluated against the actual trend observed in the patient's data over the same interval (weeks 4 to 7) to calculate metrics such as $AUC_{trend\downarrow}$.

Supplementary references

1. Nespoli, L. & Medici, V. Multivariate Boosted Trees and Applications to Forecasting and Control. *Journal of Machine Learning Research* **23**, 1–47 (2022).
2. Ke, G. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Preprint at (2017).
3. Schober, P., Boer, C. & Schwarte, L. A. Correlation Coefficients. *Anesthesia Analg.* **126**, 1763–1768 (2018).
4. Farhana, A. & Lappin, S. L. *Biochemistry, Lactate Dehydrogenase*. (2023).
5. Lee, S., Eo, W., Jeon, H., Park, S. & Chae, J. Prognostic Significance of Host-related Biomarkers for Survival in Patients with Advanced Non-Small Cell Lung Cancer. *J. Cancer* **8**, 2974–2983 (2017).
6. Lee, S., Jeon, H. & Shim, B. Prognostic Value of Ferritin-to-Hemoglobin Ratio in Patients with Advanced Non-Small-Cell Lung Cancer. J. Cancer **10**, 1717–1725 (2019).
7. Milman, N. & Pedersen, L. M. The serum ferritin concentration is a significant prognostic indicator of survival in primary lung cancer. *Oncol. Rep.* **9**, 193–8 (2001).
8. Senjo, H., Higuchi, T., Okada, S. & Takahashi, O. Hyperferritinemia: causes and significance in a general hospital. *Hematology* **23**, 817–822 (2018).
9. Tomita, M., Shimizu, T., Hara, M., Ayabe, T. & Onitsuka, T. Impact of preoperative hemoglobin level on survival of non-small cell lung cancer patients. *Anticancer Res.* **28**, 1947–50 (2008).
10. Matsukane, R. *et al.* Prognostic significance of pre-treatment ALBI grade in advanced non-small cell lung cancer receiving immune checkpoint therapy. *Sci. Rep.* **11**, 15057 (2021).
11. Takada, K. et al. Assessment of the albumin-bilirubin grade as a prognostic factor in patients with non-small-cell lung cancer receiving anti-PD-1-based therapy. *ESMO Open* **7**, 100348 (2022).
12. Jiang, J., Li, H., Chen, L. & Qiu, X. Prognostic value of albumin-bilirubin grade in lung cancer: a meta-analysis. *J. Cardiothorac. Surg.* **19**, 685 (2024).
13. Ye, J., Xiao, C., Esteves, R. M. & Rong, C. Time Series Similarity Evaluation Based on Spearman's Correlation Coefficients and Distance Measures. *Lect. Notes Comput. Sci.* 319–331 (2016) doi:10.1007/978-3-319-28430-9_24.