Evaluating Zero-Shot Foundation Models for Time Series Forecasting in Clinical Settings: A Simulation Study with Electronic Health Records

Gernot Pucher[1,2], Amin Dada[3], Felix Nensa[3], Martin Schuler[4], Christian Reinhardt[1], Jens Kleesiek[3], Christopher M. Sauer[1,2],

[1]Department of Hematology & Stem Cell Transplantation, University Hospital Essen
[2]Laboratory for Clinical Research and Real-World Evidence, Institute for AI in Medicine, University Hospital Essen
[3] Institute for AI in Medicine, University Hospital Essen
[4]Department of Medical Oncology, University Hospital Essen

gernot.pucher@uk-essen.de, amin.dada@uk-essen.de, felix.nensa@uk-essen.de, martin.schuler@uk-essen.de, christian.reinhardt@uk-essen.de, jens.kleesiek@uk-essen.de, christopher.sauer@uk-essen.de

Abstract

Longitudinal healthcare data offer significant potential for advancing clinical decision-making through time series forecasting. Despite the development of high-performing task-specific models, their clinical implementation is often limited by challenges in generalizability, data sharing, and resource constraints. Foundation models, which demonstrate zero-shot capabilities and reduced dependency on task-specific data, present a promising alternative. This study evaluates the zero-shot forecasting performance of three foundation models—Chronos, Time-LLM, and Time-MoE—compared to optimized task-specific models, using electronic health records from a German university hospital for training and two external validation datasets. Three clinical use cases with diverse temporal and predictive properties were analyzed. In this study, task-specific models, particularly deep learning models, outperformed zero-shot models in accuracy across most scenarios. However, zero-shot models demonstrated competitive performance, particularly in external validation datasets, underscoring their strong generalization potential. These findings suggest that the ease of implementation and transferability of zero-shot foundation models make them a viable option for clinical scenarios where retraining is impractical.

Introduction

Electronic Health Records (EHRs) are systematic digital repositories of patient information collected during institutionalized episodes of care (Jensen et al., 2012). Their widespread adoption across hospitals has created ample opportunities for secondary data analysis, offering the potential to enhance point-of-care clinical decision-making (Nair et al., 2016). One such application are time series prediction tasks, which leverage sequential health measurements of both numerical and categorical data types (Morid et al. 2023; Zammel et al. 2024).

Despite the potential of large clinical datasets to develop predictive models with strong reported performance, a substantial gap persists between research findings and the clinical implementation of these models (Markowetz 2024). A key challenge is the limited generalizability and transportability of models across diverse patient cohorts and hospital settings (Chekroud et al. 2024; Yang et al. 2022). This issue is partly attributable to challenges in sharing healthcare data and the substantial effort required to prepare such data for analytical purposes, as they are primarily structured for clinical documentation (De Kok et al. 2023; Johnson et al. 2023). Consequently, the development of clinical models and the effective handling of EHRs demand specialized expertise and the collaboration of multidisciplinary teams (Efthimiou et al. 2024; Sauer et al. 2022). Clinical prediction models are therefore primarily developed and implemented in centers of excellence, limiting their potential impact on the broader healthcare system (Markowetz 2024).

Foundation models hold the promise of overcoming the resource-intensive need to retrain or fine-tune models for unseen data distributions or new tasks and reducing reliance on the availability of large, task-specific labeled datasets (Wornow et al. 2023). In the domain of time series forecasting, numerous implementations of foundation models have emerged, some demonstrating forecasting capabilities comparable to those of specialized, task-specific time series models, even without explicit training on the target dataset (Liang et al. 2024). If such zero-shot performance could be replicated consistently in hospital settings, foundation models have the potential to disrupt the current task-specific paradigm of narrow clinical models and enhance the relevance of AI in bedside decision support (Moor et al. 2023). However, a thorough comparison of the zero-shot time series forecasting capabilities of foundation models with optimized task-specific models in real-world clinical scenarios,

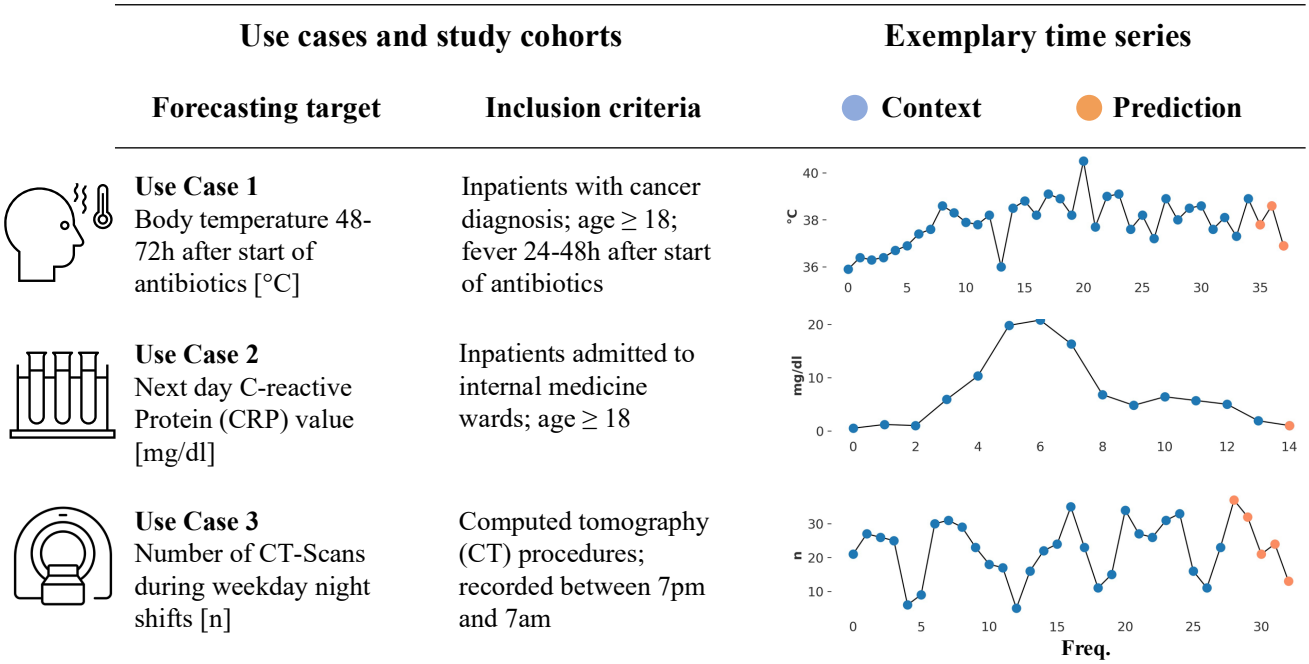

Figure 1: Clinical use cases included in this simulation study, with exemplary time series.

along with an analysis of their differences in generalizability and transferability, has not yet been reported.

To address this gap, we conducted a simulation study utilizing tabular EHRs from a large German university hospital for training, as well as from two smaller German hospitals for external validation. The aim of our study was twofold. First, we evaluated the zero-shot forecasting capabilities of three time series foundation models in comparison to various optimized task-specific models across three distinct clinical use cases, each characterized by different time series properties. Second, we examined the models' ability to transfer and generalize their performance using the external validation datasets from different hospital sites and patient cohorts.

Methods

Datasets

The dataset used for training and internal validation of the models was extracted from the existing HL7® FHIR® (Fast Healthcare Interoperability Resources) infrastructure at University Medicine Essen (UME), encompassing data from the Essen University Hospital (Brehmer et al. 2024). This database integrates relevant patient information from various primary databases into a unified and standardized data structure. For the training cohort (TC), data from the University Hospital Essen recorded between 2020-01-01 and 2023-12-31 was used. For external validation, two additional datasets were utilized: a prospective hold-out dataset from the University Hospital Essen (VC1) recorded between 2024-01-01 and 2024-10-01 to evaluate the models' generalization to new unseen data of the same target population; and an aggregate dataset from two smaller nearby hospitals, the Ruhrland Clinic and St. Josef Hospital (VC2), which was also extracted through the same FHIR interface to evaluate the transportability of models to different patient cohorts and data distributions. Data from FHIR interfaces was extracted using the open-source Python package FHIR-PYrate (Hosch et al. 2023).

Duplicate entries, implausible values, and time series shorter than the sum of the defined minimum context length and prediction length (see Table 1) were subsequently removed. The remaining time series were then resampled.

Clinical use cases

Three distinct use cases were defined by medical experts to comprehensively evaluate the models' performance across clinically relevant scenarios with varying temporal dynamics and prediction requirements. Hereby we intended to increase the applicability of the findings beyond a specific context, aiming to address a broader range of real-world clinical challenges. Figure 1 provides an overview of the different clinical use cases with exemplary time series: In use case one (UC1), the body temperature (°C) of inpatients with a cancer diagnosis was predicted for 48 to 72 hours following the administration of the first dose of antibiotics after they had developed a fever. In use case two (UC2), the next-day C-reactive Protein (CrP) laboratory value (mg/dl) was predicted for patients admitted to internal medicine wards.

In use case three (UC3), the number of computed tomography (CT) scans conducted during night shifts from Monday to Friday was predicted.

Appropriate context lengths, resampling frequencies, prediction horizons, and covariates of time series for each use case were determined based on the raw data properties and aligned with the clinical requirements of the respective use cases (Table 1).

Properties	**UC1**	**UC2**	**UC3**
Resampled Frequency	8H	1D	1D
Prediction length	3x8H	1x1D	5x1D
Context (min-max)	6-21x8H	3-14x1D	28x1D
Covariates	16	None	1

Table 1: Main time series modelling properties of the three clinical use cases.

Model development and evaluation

As zero-shot foundation models, we utilized the transformer-based models Chronos-t5-large (Ansari et al. 2024), Time-LLM with Llama-2-7b (Jin et al. 2023) and Time-MoE-large (Shi et al. 2024). These models demonstrated promising performance in previous zero-shot forecasting benchmarks, and their training corpora included longitudinal healthcare data.

For comparison with the zero-shot foundation models, we included a diverse set of model architectures selected based on their reported applicability in short-time forecasting tasks using structured EHRs (Bhatti et al. 2023; Morid et al. 2023; Olsavszky et al. 2020; Park et al. 2022). As task-specific global forecasting models, we implemented the feed-forward neural network model N-BEATS (Oreshkin et al. 2020), the gradient boosting model LightGBM (Ke et al. 2017), and an AutoML approach employing a weighted ensemble of the deep learning models Tempural Fusion Transformer, TiDE, DeepAR and PatchTST (Shchur et al. 2023). AutoARIMA (Mélard & Pasteels 2000) was implemented as a representative of local forecasting models, while a Naïve Average model served as the baseline.

The hyperparameters of N-BEATS and LightGBM were determined using Bayesian Optimization. The AutoML weighted ensemble and AutoARIMA were fitted using the configuration and training presets for optimal quality of the utilized AutoGluon (Shchur et al. 2023) library. To handle variable-length sequences, pre-padding was applied. Missing values were imputed through forward filling. Irregular time series were resampled to regular intervals as shown in Table 1.

The internal validation of task-specific models was performed using nested 10-fold cross-validation with 50 trials to prevent data leakage during hyperparameter optimization. Bootstrapping with 1,000 iterations was applied to the aggregated test-fold results to calculate confidence intervals for the performance metrics. For external validation, the models were retrained and optimized on the full training dataset, and performance metrics were calculated using bootstrapping.

	Dataset	Data points	Number of time series	Target values, mean (SD)	Context length, mean (SD)	Imputed data/series, mean (SD)
UC1	TC	6,698	428	37.5 (0.9)	15.5 (5.0)	3.3 (5.0)
	VC1	1,470	92	37.6 (0.9)	15.8 (4.9)	3.8 (5.8)
	VC2	1,012	80	37.5 (0.8)	12.4 (4.0)	3.5 (4.2)
UC2	TC	70,909	5,748	4.6 (5.7)	9.5 (3.4)	1.8 (2.3)
	VC1	28,454	2,931	4.5 (5.8)	8.7 (3.1)	1.2 (2.1)
	VC2	31,184	2,904	4.2 (4.8)	9.7 (3.2)	2.5 (3.3)
UC3	TC	14,880	465	15.2 (8.1)		
	VC1	1,120	35	20.7 (8.6)	28 (0)	None
	VC2	7,776	243	2.6 (2.6)		

Table 2: Dataset properties of the extracted datasets by use case. TC: Dataset used for training task-specific models. VC1: Prospective out-of-sample dataset from the same study population as the training dataset (generalization). VC2: Dataset from a different study population as the training dataset (transportability).

This study was implemented using Python 3.9, with the primary libraries for model development including Darts (v0.30.1), AutoGluon (v1.1.1), Neuralforecast (v1.7.6) and Transformers (v4.40).

Results

Data extraction

Across all use cases, a total of 2,360,017 data points, including 94,480 distinct hospital stays and 66,073 unique patients, were initially extracted from the FHIR database. After data cleaning, applying inclusion criteria and resampling, a combined number of 12,926 time series with 163,503 data points were available for model training and validation.

Table 2 highlights the differences between the use cases and the training and validation datasets in terms of the number of observations and time series, mean target values, context lengths, and imputed data points. UC2 included the largest dataset, with a total of 11,583 time series. While the context lengths were variable in UC1 and UC2, they were fixed at 28 days for UC3. To meet the resampling frequency, an average of 1.8/9.3 and 3.4/15.1 data points per time series were imputed for UC1 and UC2, respectively. No imputations were required for UC3, as continuous dates were used to align the counted number of CT-procedures.

Model performance

The provided plot (Figure 2) shows a comparison of models evaluated across use cases and datasets. The performance metric used is the Mean Absolute Scaled Error (MASE), which scales the model-specific mean absolute error (MAE) by the MAE of the Naïve Average baseline model, with lower values indicating better performance.

In UC1, task-specific models consistently performed best and with comparable performances across datasets. N-BEATS achieved the lowest MASE values in TC (0.74, CI: 0.70–0.78) and VC2 (0.62, CI: 0.56–0.68), while performing competitively in VC1 (0.76, CI: 0.69–0.82). Among the zero-shot models, Time-MoE showed performances in VC1 (0.79, CI: 0.71–0.87) and Time-LLM in VC2 (0.68, CI: 0.61–0.77) which are comparable to the task-specific models.

Among all models in UC2, N-BEATS showed the lowest MASE across all datasets, with TC MASE of 0.54 (CI: 0.51–0.57), VC1 MASE of 0.56 (CI: 0.54–0.59), and VC2 MASE

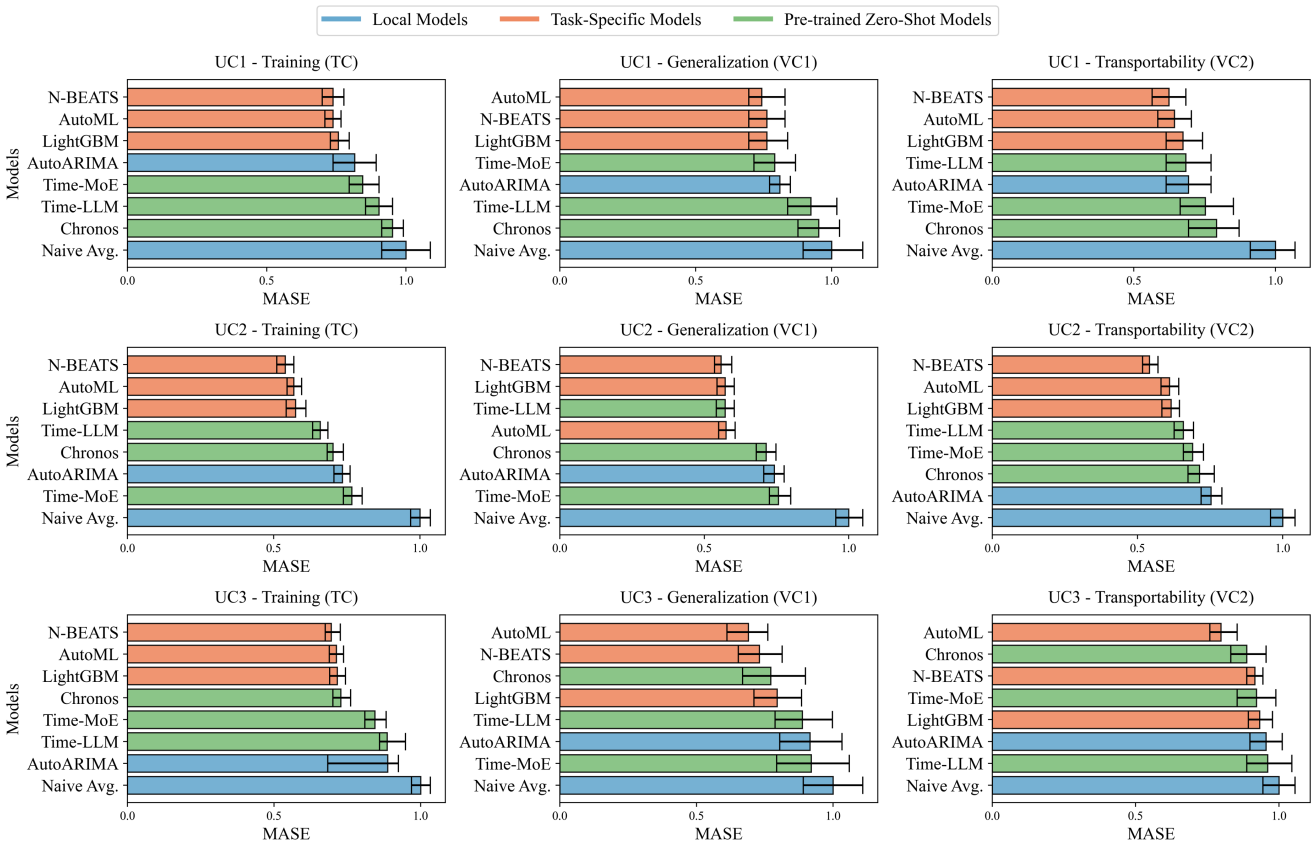

Figure 2: Comparison of model performances across use cases and datasets. Colors represent model types. Mean absolute scaled error (MASE) was calculated as the model-specific mean absolute error by the mean absolute error of the Naïve Average baseline model. Lower values indicate better performance.

of 0.54 (CI: 0.52–0.57). LightGBM and AutoML also performed well, with TC MASE values of 0.57 (CI: 0.54–0.61) and 0.57 (CI: 0.55–0.59), respectively, and comparable performance in the validation phases. The zero-shot models showed mixed performances. Time-LLM achieved competitive performance in VC1 with a MASE of 0.57 (CI: 0.54–0.60) but had slightly higher MASE in TC (0.66, CI: 0.63–0.69) and VC2 (0.66, CI: 0.63–0.69).

The results for UC3 show a competitive performance of the zero-shot model Chronos, particularly in TC (0.73, CI: 0.70–0.76) and VC1 (0.77, CI: 0.67–0.90). The best overall performance was achieved by AutoML, with a TC MASE of 0.71 (CI: 0.69–0.74), the lowest VC1 MASE of 0.69 (CI: 0.61–0.76), and the lowest VC2 MASE of 0.80 (CI: 0.76–0.85).

Figure 3 presents the trends of relative Mean Absolute Errors (rMAE), defined as the ratio of a model's MAE to the MAE of the Naïve Average baseline model, aggregated across Training (TC), Generalization (VC1), and Transportability (VC2) datasets. Lower rMAE values indicate better performance and greater added value compared to the baseline.

In the TC dataset, the task-specific models demonstrated superior performance in terms of rMAE, with N-BEATS achieving 0.66 (CI: 0.59–0.73), AutoML at 0.67 (CI: 0.61–0.74), and LightGBM at 0.68 (CI: 0.62–0.76). In contrast, the zero-shot models exhibited higher rMAE values: Chronos at 0.79 (CI: 0.73–0.88), Time-LLM at 0.82 (CI: 0.74–0.91), and Time-MoE also at 0.82 (CI: 0.74–0.91).

For the VC1 dataset, the mean rMAE values for task-specific models slightly increased compared to TC, with N-BEATS at 0.68 (CI: 0.57–0.82), LightGBM at 0.71 (CI: 0.59–0.85), and AutoML remaining stable at 0.67 (CI: 0.56–0.80). Similarly, the zero-shot models showed slight variations in their mean values compared to the TC dataset: Chronos increased slightly to 0.81 (CI: 0.68–0.98), Time-MoE remained nearly unchanged at 0.82 (CI: 0.68–0.99), and Time-LLM showed a slight improvement, decreasing to 0.79 (CI: 0.66–0.96).

In VC2, the mean rMAE values for task-specific models further increased compared to TC, with N-BEATS at 0.69 (CI: 0.62–0.78), LightGBM at 0.74 (CI: 0.66–0.84), and AutoML at 0.68 (CI: 0.61–0.78). The zero-shot models demonstrated slight improvements in this dataset. Time-LLM decreased in mean rMAE to 0.77 (CI: 0.67–0.89), Time-MoE to 0.79 (CI: 0.69–0.91), and Chronos remained relatively stable at 0.80 (CI: 0.68–0.98).

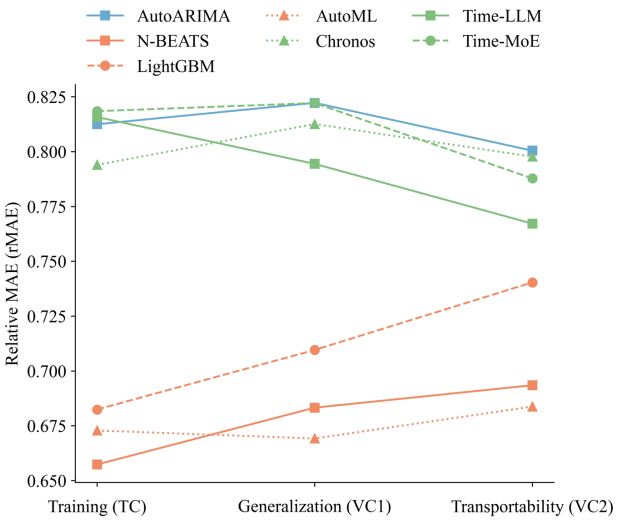

Figure 3: Ratio of a model's MAE to the MAE of the Naïve Average baseline model (rMAE), aggregated over TC (Training), VC1 (Generalization) and VC2 (Transportability) datasets. Lower values indicate better performance compared to the baseline model.

Discussion

In this simulation study, the optimized task-specific models either outperformed or were on par with the zero-shot foundation models across all three clinical use cases and datasets. Notably, the deep learning model N-BEATS demonstrated stable and reliable performance across datasets, emphasizing the effectiveness of models explicitly trained to capture task-specific temporal patterns in clinical data. Dataset-specific trends revealed only subtle differences in model performance. For the UC1 and UC2 datasets, task-specific models maintained a noticeable edge for the training datasets (TC). However, this advantage diminished in the prospective validation cohort (VC1) and the validation cohort from different hospital sites (VC2), suggesting performance degradation of task-specific models when applied to a different patient cohort. In most scenarios, at least one zero-shot model could approach the accuracy of task-specific models, particularly in the validation datasets. An overall trend of converging performance between task-specific and zero-shot models from the training datasets (TC) to datasets validating model transportability (VC2) was observed, highlighting the potential of foundation models to be used for forecasting unseen clinical data. However, smaller samples sizes in the validation cohorts compared to the training cohorts resulted in wider confidence intervals, increasing statistical uncertainty and warranting a cautious interpretation of performance differences. The foundation models included in this study demonstrated varying forecasting capabilities depending on the use case, with each model outper-

forming the others in at least one specific scenario. This variability in their generalization capabilities indicates the potential of leveraging ensemble approaches or model selection strategies tailored to the specific clinical use case or data characteristics.

Our findings are consistent with benchmarks reported in literature, where zero-shot foundation models have demonstrated scores comparable to the best performing optimized task-specific models (Gruver et al. 2023; Woo et al., 2024). As in our use cases, short time series are prevalent in EHRs due to patients spending limited continuous time in hospital. Especially in such scenarios with limited learning context, zero-shot performance of foundation models has been found to have a larger scope for improvement(Ansari et al. 2024; Goswami et al. 2024).

This study has several limitations. Although multiple use cases and hospitals sites were included, our findings are not necessarily generalizable to other healthcare challenges and settings. Moreover, our focus was limited to evaluating the zero-shot forecasting performance of foundation models. This decision was intentional, as we aimed to assess whether foundation models could be applied to forecasting tasks in clinical settings without requiring the resources and specialized expertise necessary for model adjustments. Nonetheless, the exploration of fine-tuning and few-shot learning remains an area for future research. Lastly, although we selected eight different models for this study, we recognize that many other model architectures were not examined, and that their inclusion could have led to different results.

The zero-shot forecasting capabilities of foundation models were generally inferior to the forecasting performance of locally optimized task-specific models in the predefined real-world clinical use cases. However, in some scenarios, their performance approached that of task-specific models, particularly in validation datasets where generalization is critical. Combined with their ease of implementation and superior generalizability across different settings, foundation models may offer unique advantages over traditional use-case-specific models. In our opinion, foundation models may therefore be a noteworthy alternative whenever retraining is technically, logistically, or financially impractical. In the medical field, common barriers include a lack of historical training data, expert knowledge, or financial resources. By eliminating the need for retraining at individual sites, the zero-shot capabilities of foundation models present a potentially more cost-effective solution for the healthcare system overall.

References

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). *Chronos: Learning the Language of Time Series* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2403.07815

Bhatti, A., Thangavelu, N., Hassan, M., Kim, C., Lee, S., Kim, Y., & Kim, J. Y. (2023). *Interpreting Forecasted Vital Signs Using N-BEATS in Sepsis Patients* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2306.14016

Brehmer, A., Sauer, C. M., Salazar Rodríguez, J., Herrmann, K., Kim, M., Keyl, J., Bahnsen, F. H., Frank, B., Köhrmann, M., Rassaf, T., Mahabadi, A.-A., Hadaschik, B., Darr, C., Herrmann, K., Tan, S., Buer, J., Brenner, T., Reinhardt, H. C., Nensa, F., … Kleesiek, J. (2024). Establishing Medical Intelligence—Leveraging Fast Healthcare Interoperability Resources to Improve Clinical Management: Retrospective Cohort and Clinical Implementation Study. *Journal of Medical Internet Research*, *26*, e55148. https://doi.org/10.2196/55148

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science*, *383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

De Kok, J. W. T. M., De La Hoz, M. Á. A., De Jong, Y., Brokke, V., Elbers, P. W. G., Thoral, P., Castillejo, A., Trenor, T., Castellano, J. M., Bronchalo, A. E., Merz, T. M., Faltys, M., Collaborator group, Casares, C., Jiménez, A., Requejo, J., Gutiérrez, S., Curto, D., Rätsch, G., … Borrat, X. (2023). A guide to sharing open healthcare data under the General Data Protection Regulation. *Scientific Data*, *10*(1), 404. https://doi.org/10.1038/s41597-023-02256-2

Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing clinical prediction models: A step-by-step guide. *BMJ*, e078276. https://doi.org/10.1136/bmj-2023-078276

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). *MOMENT: A Family of Open Time-series Foundation Models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2402.03885

Gruver, N., Finzi, M., Qiu, S., & Wilson, A. G. (2023). *Large Language Models Are Zero-Shot Time Series Forecasters* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2310.07820

Hosch, R., Baldini, G., Parmar, V., Borys, K., Koitka, S., Engelke, M., Arzideh, K., Ulrich, M., & Nensa, F. (2023). FHIR-PYrate: A data science friendly Python package to query FHIR servers. *BMC Health Services Research*, *23*(1), 734. https://doi.org/10.1186/s12913-023-09498-1

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405. https://doi.org/10.1038/nrg3208

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., & Wen, Q. (2023). *Time-LLM: Time Series Forecasting by Reprogramming Large Language Models* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2310.01728

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*,

10(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., & Wen, Q. (2024). *Foundation Models for Time Series Analysis: A Tutorial and Survey*. https://doi.org/10.48550/ARXIV.2403.14735

Markowetz, F. (2024). All models are wrong and yours are useless: Making clinical prediction models impactful for patients. *Npj Precision Oncology*, *8*(1), 54. https://doi.org/10.1038/s41698-024-00553-6

Mélard, G., & Pasteels, J.-M. (2000). Automatic ARIMA modeling including interventions, using time series expert software. *International Journal of Forecasting*, *16*(4), 497–508. https://doi.org/10.1016/S0169-2070(00)00067-4

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

Morid, M. A., Sheng, O. R. L., & Dunbar, J. (2023). Time Series Prediction Using Deep Learning Methods in Healthcare. *ACM Transactions on Management Information Systems*, *14*(1), 1–29. https://doi.org/10.1145/3531326

Nair, S., Hsu, D., & Celi, L. A. (2016). Challenges and Opportunities in Secondary Analyses of Electronic Health Record Data. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records*. Springer. http://www.ncbi.nlm.nih.gov/books/NBK543649/

Olsavszky, V., Dosius, M., Vladescu, C., & Benecke, J. (2020). Time Series Analysis and Forecasting with Automated Machine Learning on a National ICD-10 Database. *International Journal of Environmental Research and Public Health*, *17*(14), 4979. https://doi.org/10.3390/ijerph17144979

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting* (arXiv:1905.10437; Version 4). arXiv. https://doi.org/10.48550/arXiv.1905.10437

Park, J., Artin, M. G., Lee, K. E., Pumpalova, Y. S., Ingram, M. A., May, B. L., Park, M., Hur, C., & Tatonetti, N. P. (2022). Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer. *Journal of Biomedical Informatics*, *131*, 104095. https://doi.org/10.1016/j.jbi.2022.104095

Sauer, C. M., Chen, L.-C., Hyland, S. L., Girbes, A., Elbers, P., & Celi, L. A. (2022). Leveraging electronic health records for data science: Common pitfalls and how to avoid them. *The Lancet Digital Health*, *4*(12), e893–e898. https://doi.org/10.1016/S2589-7500(22)00154-6

Shchur, O., Turkmen, C., Erickson, N., Shen, H., Shirkov, A., Hu, T., & Wang, Y. (2023). *AutoGluon-TimeSeries: AutoML for Probabilistic Time Series Forecasting* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2308.05566

Shi, X., Wang, S., Nie, Y., Li, D., Ye, Z., Wen, Q., & Jin, M. (2024). *Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts* (arXiv:2409.16040). arXiv. https://doi.org/10.48550/arXiv.2409.16040

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). *Unified Training of Universal Time Series Forecasting Transformers* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2402.02592

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *Npj Digital Medicine*, *6*(1), 135. https://doi.org/10.1038/s41746-023-00879-8

Yang, J., Soltan, A. A. S., & Clifton, D. A. (2022). Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening. *Npj Digital Medicine*, *5*(1), 69. https://doi.org/10.1038/s41746-022-00614-9

Zammel, Z., Khabou, N., Souifi, L., & Bouassida Rodriguez, I. (2024). Time Series Prediction Models in Healthcare: Systematic Literature Review: *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, 1286–1293. https://doi.org/10.5220/0012465000003636