# Evaluating Time-Series Foundation Models as Zero-Shot Surrogates for Mechanistic Virtual Patients

**Ahmad Wisnu Mulyadi***, **Charlie George Barker***, **Sanjana Balaji Kuttae**, **Lilija Wehling**, **Thomas Rückle**, **Gurdeep Singh**†
Virtual Patient Engine, BioMed X GmbH, Heidelberg, Germany
{mulyadi, barker, balaji-kuttae, wehling, tr, gsingh}@bmedx.com

**Nicolas Boucher**[1], **Firas Abdessalem**[1], **Sven Jager**[2], **Anastasios Siokis**[4], **Sommer Anjum**[4], **Mohammed H. Mosa**[3], **Thomas Klabunde**[4], **Tommaso Andreani**[4†]
[1]Digital R&D, Accelerator, Sanofi Digital, Paris, France
[2]Digital R&D, Data & Computational Science, Sanofi Digital, Frankfurt am Main, Germany
[3]Therapeutic Area Type 1/17 Immunology & Inflammation, Sanofi R&D, Frankfurt am Main, Germany
[4]Translational Medicine Unit (TMU) - Disease Modeling, Sanofi R&D, Frankfurt am Main, Germany
{Nicolas.Boucher, Firas.Abdessalem, Sven.Jager, Anastasios.Siokis, Sommer.Anjum, Mohammed.Mosa, Thomas.Klabunde, Tommaso.Andreani}@sanofi.com

## Abstract

Mechanistic models such as Quantitative Systems Pharmacology (QSP) models are widely used to simulate the behavior of virtual patients (VPs) under different therapeutic conditions, supporting hypothesis generation and trial design. However, large-scale VP simulations are computationally expensive and require expert calibration. Recent advances in time-series foundation models (TS-FM) have demonstrated strong generalization across diverse temporal domains in a zero shot manner. In this study, we explore whether these models can act as zero-shot surrogates for QSP-based VP simulations. Using simulation outputs from 5 representative QSP models across multiple treatment scenarios and VP parameterizations, we benchmark 3 TS-FMs on their ability to predict future system trajectories. Our results show that TS-FM capture certain pharmacodynamic patterns and VP-level variability without fine-tuning, although performance varies between biological systems. This work highlights both the promise and the current limitations of using TS-FMs to accelerate VP-based *in silico* experimentation.

*These authors contributed equally to this work.
†Correspondence: gsingh@bmedx.com; tommaso.andreani@sanofi.com

European Conference on Neural Information Processing Systems (EurIPS) Workshop: SimBioChem 2025.

## 1 Introduction

Virtual patients (VPs) provide a computational analog of clinical variability which can be used to predict individual responses to therapy[23]. These simulations are often based on Quantitative Systems Pharmacology (QSP) models, mechanistic systems of ordinary differential equations (ODE) that integrate molecular, cellular, physiological and biological pathways to capture disease progression and treatment effects. QSP models allow mechanistic interrogation of predictions, which is essential for explaining patient-specific responses and informing novel drug development[11, 20]. However, scaling QSP simulations to large VP cohorts, necessary to capture population-level variability over long durations, remains computationally intensive[10].

To address this, researchers have explored surrogate modelling approaches, which aim to replicate QSP model outputs with substantially lower computational cost[10]. For example, neural ordinary differential equations (NODEs) have been proposed as trainable surrogates that preserve the mechanistic structure while enabling accelerated simulation[7]. Yet, NODEs and related approaches typically require model-specific training, limiting their scalability and generalizability across diverse therapeutic contexts[16]. Recent efforts to create "digital twin" style forecasting include Delphi-2M[19] and DT-GPT[17], which predict clinical trajectories using large-scale EHR (Electronic Health Records) or biomarker data. Delphi-2M generates synthetic pathways to disease progression in populations based on medical history and lifestyle factors, while DT-GPT extends LLM (Large Language Model) architectures to predict clinical time series, supporting zero-shot predictions for some unseen variables. However, both models lack mechanistic grounding, provide limited interpretability, and cannot reliably extrapolate to novel perturbations (e.g., first-in-class therapeutics).

An intriguing possibility is whether time series foundation models (TS-FMs) can serve as surrogate generators for virtual patients. If viable, such surrogates could accelerate ensemble simulations or augment scarce mechanistic model runs. These models, such as Chronos[2], TiRex[4], and Moirai[24], are pre-trained across millions of heterogeneous time series data and have demonstrated strong zero-shot forecasting performance on benchmarks spanning traffic, energy, climate, and finance data[1]. Furthermore, TS-FMs are architecture-agnostic surrogates that can be applied to unseen systems without retraining, offering lightweight deployment and rapid inference, properties that make them attractive candidates for approximating QSP-generated virtual patient trajectories.

Using five open-source and proprietary QSP models spanning multiple therapeutic areas, we benchmark the Chronos family, TiRex, and $\text{Moirai}_{\text{Small}}$ on their ability to reproduce virtual patient trajectories. We deliberately focus on *off-the-shelf* TS-FMs without mechanistic priors to establish a baseline for their intrinsic generalization ability; if successful, future work can explore hybrid models that incorporate mechanistic knowledge for greater interpretability and robustness. To our knowledge, this is the first systematic investigation of time-series foundation models as surrogates for virtual patient simulations.

## 2 Proposed Method

**QSP-derived Simulated Data** Let $\{\mathbf{X}^r\}_{r=1}^{R}$ denote the set of simulated data produced from a given QSP model with $\mathbf{X}^r$ denoting a cohort of virtual patients under particular treatment regimen $\{1, R\} \in \mathcal{R}$. Such a cohort $\mathbf{X}^r \in \mathbb{R}^{P\times S\times T}$ constitutes time-series data for virtual patients $\{1, P\} \in \mathcal{P}$, comprising bio-signals from species $\{1, S\} \in \mathcal{S}$ over the temporal span $\{1, T\} \in \mathcal{T}$. For the sake of clarity, we omit these indices when unambiguous. Each time series is further partitioned into context and future-horizon segments, *i.e.*, $\mathbf{X}_{1:T} = (\mathbf{X}^c, \mathbf{X}^h)$, where $\mathbf{X}^c = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t=c}]$ and $\mathbf{X}^h = [\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_{T=c+h}]$, respectively. The context data $\mathbf{X}^c$ are used as input to the forecasting model, while the horizon signals $\mathbf{X}^h$ serve as the ground truth for performance evaluation.

**Time-series Forecasting** The objective of this work is to systematically benchmark a set of time-series foundation models on its zero-shot forecasting performances over QSP model-simulated data. Leveraging an optimized TS-FM $f$, such a forecasting process can be formulated as follows

$$\widetilde{\mathbf{X}}^h = f_\theta(\mathbf{X}^c), \tag{1}$$

where $\widetilde{\mathbf{X}}^h$ denotes the predicted future horizon signals and $\theta$ as the frozen parameters of the forecasting model. Note that we simplify the forecasting process of the model $f_\theta$ in Eq. (1). In practice,

time-series forecasting methods can be broadly categorized into two input paradigms: univariate and multivariate forecasting [15], denoted as $f_\theta^U$ and $f_\theta^M$, respectively. In our formulation, we treat the species in each QSP model as variables (or channels) that span across timesteps. Consequently, the forecasting equation can be reformulated to reflect whether the model incorporates covariates as auxiliary features when predicting target signals, as expressed in the following Eqs. (2)–(3).

$$\widetilde{\mathbf{X}}_s^h = f_\theta^U(\mathbf{X}_s^c) \tag{2}$$

$$\widetilde{\mathbf{X}}_s^h = f_\theta^M(\mathbf{X}_{1:S}^c) \tag{3}$$

Since some of the baselines belong to the probabilistic forecasting models, we make use of the median of their predictive distributions as the final forecasted outcome. Finally, we assess the performance of the forecasting model by comparing the predicted signals $\widetilde{\mathbf{X}}^h$ against the ground-truth horizon signals $\mathbf{X}^h$ using a set of widely adopted forecasting evaluation metrics.

## 3 Experimental Results

**Dataset and Data Preprocessing** We began with two proprietary QSP models representing asthma and Inflammatory Bowel Disease (IBD). To broaden disease coverage and include publicly available systems, we incorporated three additional QSP models obtained from the BioModels Database [18]. These models were manually verified to confirm their biological relevance and suitability for virtual patient simulations. The final set comprised five QSP models (two proprietary and three from BioModels) spanning multiple therapeutic areas (see Appendix A).

For each selected model, we extracted simulation time frames and other parameters from the associated publications and reproduced the simulations in COPASI version 4.44 (Build 295)[13]. All models were executed under their reported conditions to generate corresponding time-course datasets.

To place all simulated time course data on a common scale suitable for multivariate modeling, we applied global z-score normalization across all VPs, treatments, and time points. For each species in the QSP model, the global mean and standard deviation were computed over the entire dataset. We then standardized each VP's time-course matrix according to the equation:

$$z_{s,t}^{(p)} = \frac{x_{s,t}^{(p)} - \mu_s}{\sigma_s} \tag{4}$$

where $x_{s,t}^{(p)}$ denotes the value of species $s$ at time $t$ for virtual patient $p$, while $\mu_s$ and $\sigma_s$ indicate the global mean and standard deviation, respectively. This normalization preserved the temporal dynamics of each species while ensuring all species contributed on a comparable numerical scale, thereby preventing high-magnitude variables from dominating the analysis. We excluded species with zero variance, as these were invariant across all simulations and uninformative for downstream inference. The normalized simulated biomarker trajectories were used to evaluate the predictive performance of selected TS-FMs.

**Baselines** We selected several top-performing open-source TS-FMs based on their forecasting accuracy on standardized datasets. The selected baselines comprised *(i) univariate:* the Chronos family[2], TiRex [4]; *(ii) multivariate:* $\text{Moirai}_{\text{Small}}$[24].

- **Chronos [2]** is a Transformer-based family of models that quantize real-valued time series into discrete tokens, enabling autoregressive sequence modelling. We evaluated all three variants (small, medium, and large), which differ in model depth and hidden dimension. Pre-training on millions of heterogeneous time series allows Chronos to capture complex temporal dependencies, and forecasts are generated by autoregressively sampling future tokens and mapping them back to numerical values.
- **TiRex [4]** is an xLSTM-based model with extended recurrent units and multi-head gating. Pretraining employs Contiguous Patch Masking (CPM) to improve long-horizon coherence. The model produces probabilistic forecasts at each timestep, effectively capturing long-term dependencies in nonlinear time series.

- **Moirai$_{\text{Small}}$ [24]** is a multivariate masked-encoder Transformer. Input sequences are embedded as temporal patches, and the model predicts randomly masked patches during pretraining. This design enables any-variate forecasting, captures cross-variable correlations, and handles heterogeneous and irregularly sampled data.

**Evaluation metrics** We evaluated the selected TS-FMs in zero-shot inference mode using the following metrics:

(i) Mean Absolute Error (MAE) measures the average magnitude of errors between predicted and observed values, without considering their direction.

(ii) Mean Squared Error (MSE) computes the average of the squared differences between predicted and actual values. By squaring the errors, MSE penalizes larger deviations more heavily, making it sensitive to outliers.

(iii) Root Mean Squared Error (RMSE) represents the square root of the MSE.

(iv) Symmetric Mean Absolute Percentage Error (SMAPE) measures the relative prediction error by comparing the absolute difference between predicted and actual values to their mean magnitude. It is defined as:

$$\text{SMAPE} = \frac{2}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

In this study, SMAPE is expressed as a unitless ratio ranging from 0 to 2, where lower values indicate higher predictive accuracy.

We aggregated the evaluation metrics for each species across the horizon segments for all virtual patients. We then calculated the mean and standard deviation across the treatment regimens for each QSP model. We computed the MAE, MSE, and RMSE using z-score–normalized values between the predicted and ground-truth horizon signals. For SMAPE, we reconverted the z-scores using the global mean and standard deviation to compare the original (unnormalized) values. This approach ensured that the evaluation metrics reflected both the normalized and original value scales. All experiments were performed on a single NVIDIA A100 GPU equipped with 80 GB of memory.

**Forecasting Results** We evaluated the predictive performance of three TS-FMs, Chronos (transformer-based), Moirai$_{\text{Small}}$(transformer-based) and TiRex (xLSTM-based) on five QSP models. These QSP models represent diverse disease areas including asthma, Inflammatory Bowel Disease (IBD), HER2-positive breast cancer, osteoarthritic pain, and acute liver failure (Appendix A). The TS-FMs were assessed under varying context-horizon splits (C:H = 75%:25%, 50%:50%, 25%:75%), capturing short-, medium-, and long-term forecasting scenarios (Appendix B). Standard error metrics including MAE, RMSE, MSE, and SMAPE were used to quantify predictive fidelity. Across all experiments, the smaller Chronos model (Chronos$_{\text{Small}}$) mostly outperformed the larger variants (Chronos$_{\text{Base/Large}}$), including for long-horizon forecasts. This suggests that, in our setting, increased model size did not always translate into improved performances. TiRex exhibited competitive performance, frequently surpassing the Chronos family and Moirai$_{\text{Small}}$ in low-context, high-horizon scenarios, suggesting that xLSTM-based architectures are robust in sequential prediction under limited contextual data.

Predictive accuracy exhibited clear dependence on the context-horizon configuration (Figure 1). In high-context, short-horizon settings (C:75%, H:25%), all models achieved low error metrics, reflecting ease of prediction when ample past information is available. In balanced scenarios (C:50%, H:50%), error metrics increased moderately, consistent with the increased difficulty of mid-horizon forecasts, with TiRex occasionally outperforming on RMSE and SMAPE. Low-context, long-horizon predictions (C:25%, H:75%) were challenging across all models. Chronos$_{\text{Base/Large}}$ maintained an advantage over smaller transformers, whereas TiRex showed relative robustness, demonstrating the ability of xLSTM architectures to generalize under limited context, long-horizon forecasting conditions. Full metrics for predictive accuracies are listed in Appendix B.

Performance of TS-FMs varied across QSP models, reflecting differences in system dynamics (Figure 1). Predictions for asthma and Acute Liver Failure (ALF) achieved the lowest errors and SMAPE, consistent with predictable, structured dynamics. Asthma exhibits quasi-periodic inflammatory patterns [12], while ALF follows largely deterministic cascades of hepatocellular injury and regeneration [3].

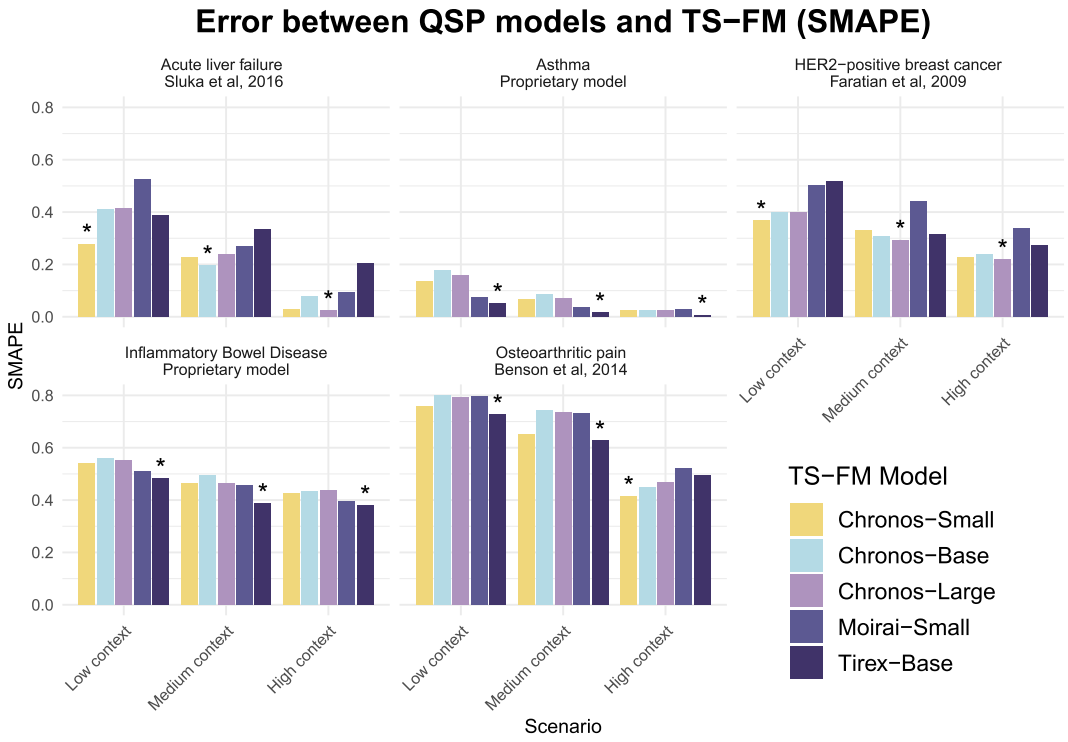


Figure 1: **Comparison of TS-FM model predictive performance across QSP disease models and simulation contexts.** Each panel represents a QSP model, showing the symmetric SMAPE (Symmetric Mean Absolute Percentage Error) between the TS-FM model and the corresponding QSP simulation across low, medium, and high context scenarios. Bars indicate average error for each TS-FM architecture (color-coded), and an asterisk (*) marks the best-performing model (lowest SMAPE) within each scenario. SMAPE values are on a scale of 0-2. Lower values indicate more accurate predictions.

HER2-positive breast cancer and IBD showed intermediate performance, reflecting moderate-to-high stochasticity [5, 22]. The osteoarthritic (OA) pain model had the lowest performance, likely due to the highly stochastic pain trajectories influenced by joint degeneration, inflammation, and psychosocial factors [8].

Visualizing trajectories of the inflammatory biomarker calprotectin in the IBD model (Figure 2) further illustrates these trends. Across varied context sizes, the predicted trajectories of all TS-FMs largely encompassed the ground-truth dynamics within their confidence intervals, even under low-context conditions. Representative virtual patients were identified via hierarchical clustering of trajectories using dynamic time warping distance, ensuring interpretable exemplars of system behaviour (Appendix C). Among models, $\text{Chronos}_{\text{Small}}$ tended to overestimate oscillatory behaviour, producing confident yet exaggerated outputs. $\text{Moirai}_{\text{Small}}$ exhibited more stochastic transitions, generating abrupt deviations from the ground truth. In contrast, TiRex demonstrated stable and realistic dynamics consistent with its overall quantitative performance. These qualitative patterns reinforce the quantitative results, suggesting that xLSTM-based architectures may better capture the nonlinear yet structured feedback characteristic of biological systems.

## 4 Limitations and future works

Our empirical evaluation suggests that state-of-the-art TS-FMs such as the Chronos family, $\text{Moirai}_{\text{Small}}$ and TiRex hold promise as fast surrogate engines for virtual patient trajectories, especially in interpolation settings. However, in their present form they suffer from three key limitations: *(i)* they remain black-box forecasters that do not explicitly encode mechanistic insights or inherent causal structure of system biological models; *(ii)* they are unlikely to extrapolate reliably under dramatic

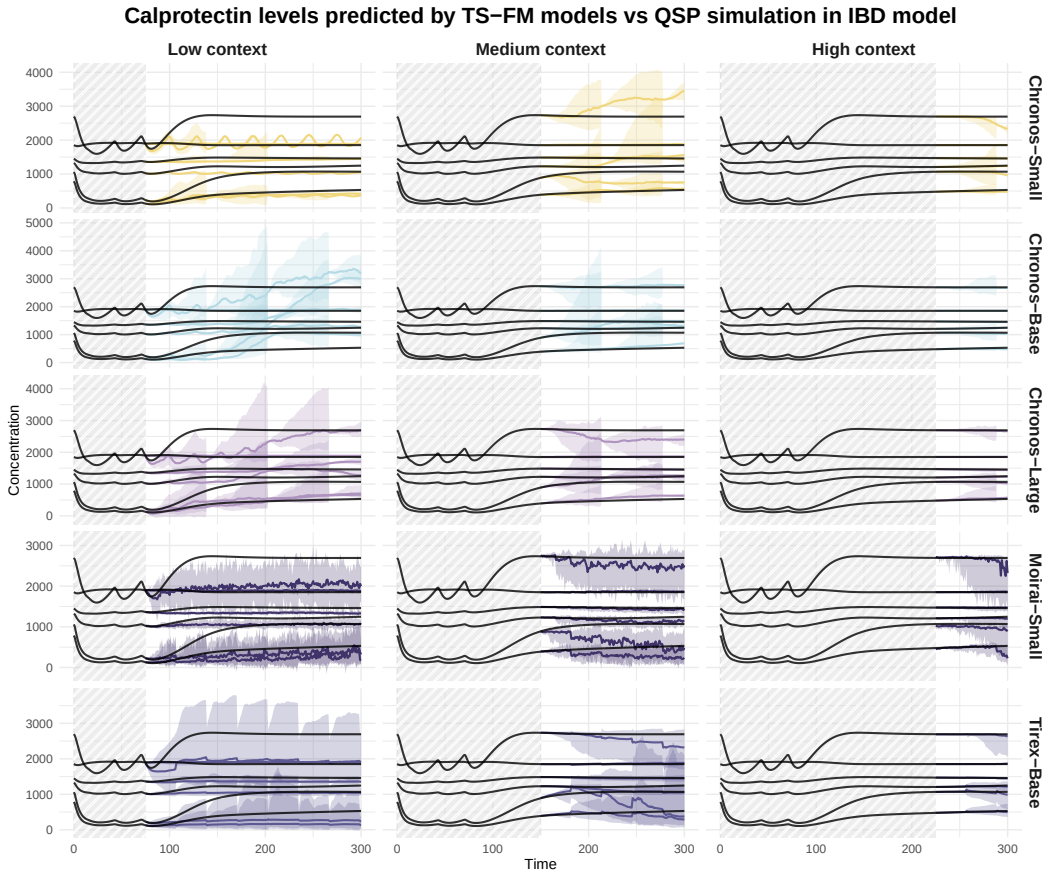


Figure 2: **Comparison of TS-FM predictions versus QSP simulations for calprotectin dynamics in the IBD model across contexts.** Each panel shows the predicted and simulated trajectories of calprotectin following drug treatment across three training contexts (Low: C25H75, Medium: C50H50, High: C75H25) and four TS-FM model architectures ($\text{Chronos}_{\text{Base}}$, $\text{Chronos}_{\text{Large}}$, $\text{Chronos}_{\text{Small}}$, $\text{Moirai}_{\text{Small}}$, Tirex-base). Black lines represent the QSP simulation (ground truth) of 6 representative virtual patients, while colored lines and shaded regions indicate the TS-FM median predictions and corresponding 80% confidence intervals. The grey boxes represent the contexts.

domain shifts, such as first-in-class therapeutics or novel biological perturbations; and *(iii)* their performance varies across dynamic systems based on stochasticity.

A promising direction to address these limitations is to fine-tune TS-FMs on QSP-derived trajectories or patient-specific data to enhance their fidelity and robustness. Beyond fine-tuning, future research should explore hybrid architectures that augment TS-FMs with knowledge graphs (KGs) derived from QSP models or curated systems biology ontologies. By embedding these KGs, which encode causal and mechanistic relationships among pathways, cytokines, cell types, and clinical endpoints, directly into TS-FM backbones, models could better capture underlying biological structure while retaining their statistical generalization capacity. Such integration offers a path toward combining the expressive power of foundation models with the mechanistic rigour of systems pharmacology. Additionally, based on GIFT-EVAL benchmark[1], expanding the set of TS-FMs evaluated could increase model diversity.

The limitation arising from the observed variability of TS-FM performance across different QSP models may reflect underlying biological and patient heterogeneity. Differences in biomarker relevance, the coexistence of responders and non-responders, and variations in response timing can all affect predictive accuracy. Future work should explicitly model these sources of heterogeneity to improve predictive reliability beyond what is constrained by stochasticity in QSP simulations. Moreover, incorporating a broader set of QSP models from repositories such as BioModels and

related databases could also expand disease coverage, enabling more comprehensive virtual patient analyses.

Finally, recent advances in Time-Series Language Models (TSLMs), such as OpenTSLM[14], demonstrate the feasibility of integrating temporal modalities into large language model architectures to enable reasoning, forecasting, and extrapolation over time-dependent data. Building upon these advances, future virtual patient surrogate architectures may fuse TS-FM forecasting ability with TSLM-style reasoning over mechanistic knowledge graphs and patient-specific endotypes, resulting in models that are not only computationally efficient, but also interpretable, mechanistically grounded, and robust to novel perturbations.

## 5 Conclusion

We systematically evaluated the potential of time-series foundation models (TS-FMs) as surrogates for virtual patient simulations generated by quantitative systems pharmacology (QSP) models. Using five QSP models spanning diverse therapeutic areas, we demonstrated that *off-the-shelf* TS-FMs, including the Chronos family, TiRex, and $\text{Moirai}_{\text{Small}}$, can effectively reproduce virtual patient trajectories, particularly in high-context, short- to mid-horizon forecasting scenarios. Transformer-based models scaled to larger architectures ($\text{Chronos}_{\text{Base/Large}}$) consistently outperformed their smaller variant for long-horizon predictions, while xLSTM-based TiRex exhibited robustness under low-context, high-horizon conditions. Performance variability across QSP models reflected the underlying system dynamics, with predictable systems (e.g., asthma, acute liver failure) yielding lower errors and highly stochastic systems (e.g., osteoarthritic pain) posing greater challenges.

Looking forward, these findings open the door to hybrid approaches that combine TS-FMs with mechanistic knowledge or patient-specific data, bridging statistical forecasting with biological interpretability. Such integrations could enable scalable, computationally efficient, and clinically relevant virtual patient simulations, supporting drug development, personalized therapy planning, and predictive modeling of disease trajectories.


### Acknowledgments

We acknowledge Henrik Cordes from Sanofi for collaborative in-depth discussion in performing the benchmark. We thank Douglas McCloskey for discussions on time-series foundation models during his time at BioMed X GmbH. We thank Vultr[3] for graciously providing free credits to perform some experiments.

### Funding

This study was funded by Sanofi.


### Conflict Interests

NB, FA, SJ, AS, SA, MHM, TK, and TA are Sanofi employees and may hold shares and/or stock options in the company.

## References


[1] Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-eval: A benchmark for general time series forecasting model evaluation.

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series.

[3] Daniel J Antoine, James W Dear, Philip Starkey Lewis, Vivien Platt, Judy Coyle, Moyra Masson, Ruben H Thanacoody, Alasdair J Gray, David J Webb, Jonathan G Moggs, D Nicholas Bateman, Christopher E Goldring, and B Kevin Park. Mechanistic biomarkers provide early and sensitive detection of


[3] https://www.vultr.com/

acetaminophen-induced acute liver injury at first presentation to hospital. *Hepatology (Baltimore, Md.)*, 58:777–87, 2013.

[4] Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-shot forecasting across long and short horizons. In *1st ICML Workshop on Foundation Models for Structured Data*.

[5] Yael Bar, Geoffrey Fell, Aylin Dedeoglu, Natalie Moffett, Neelima Vidula, Laura Spring, Seth A Wander, Aditya Bardia, Naomi Ko, Beverly Moy, Leif W Ellisen, and Steven J Isakoff. Dynamic her2-low status among patients with triple negative breast cancer (tnbc) and the impact of repeat biopsies. *NPJ breast cancer*, 11:27, 2025.

[6] N Benson, E Metelkin, O Demin, G L Li, D Nichols, and P H van der Graaf. A systems pharmacology perspective on the clinical development of fatty acid amide hydrolase inhibitors for pain. *CPT: pharmacometrics  systems pharmacology*, 3:e91, 2014.

[7] Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. Low-dimensional neural ODEs and their application in pharmacokinetics. 51(2):123–140.

[8] L A Deveza, L Melo, T P Yamato, K Mills, V Ravi, and D J Hunter. Knee osteoarthritis phenotypes and their relevance for outcomes: a systematic review. *Osteoarthritis and cartilage*, 25:1926–1941, 2017.

[9] Dana Faratian, Alexey Goltsov, Galina Lebedeva, Anatoly Sorokin, Stuart Moodie, Peter Mullen, Charlene Kay, In Hwa Um, Simon Langdon, Igor Goryanin, and David J Harrison. Systems biology reveals new strategies for personalizing cancer medicine and confirms the role of pten in resistance to trastuzumab. *Cancer research*, 69:6713–20, 2009.

[10] Igor Goryanin, Irina Goryanin, and Oleg Demin. Revolutionizing drug discovery: Integrating artificial intelligence with quantitative systems pharmacology. 30(9):104448.

[11] Gabriel Helmlinger, Victor Sokolov, Kirill Peskov, Karen M. Hallow, Yuri Kosinsky, Veronika Voronova, Lulu Chu, Tatiana Yakovleva, Ivan Azarov, Daniel Kaschek, Artem Dolgun, Henning Schmidt, David W. Boulton, and Robert C. Penland. Quantitative systems pharmacology: An exemplar model-building workflow with applications in cardiovascular, metabolic, and oncology drug development. 8(6):380–395.

[12] Stephen T Holgate. Innate and adaptive immune responses in asthma. *Nature medicine*, 18:673–83, 2012.

[13] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. COPASI—a COmplex PAthway SImulator. 22(24):3067–3074.

[14] Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Tarlochan Singh Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Aalami, Filipe Barata, and Paul Schmiedmayer. OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data.

[15] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565.

[16] Idris Bachali Losada and Nadia Terranova. Bridging pharmacology and neural networks: A deep dive into neural ordinary differential equations. 13(8):1289–1296.

[17] Nikita Makarov, Maria Bordukova, Papichaya Quengdaeng, Daniel Garger, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P. Menden. Large language models forecast patient health trajectories enabling digital twins. 8(1):588. Publisher: Nature Publishing Group.

[18] Rahuman S. Malik-Sheriff, Mihai Glont, Tung V. N. Nguyen, Krishna Tiwari, Matthew G. Roberts, Ashley Xavier, Manh T. Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, Emma L. Fairbanks, Johannes P. Meyer, Chinmay Arankalle, Thawfeek M. Varusai, Vincent Knight-Schrijver, Lu Li, Corina Dueñas-Roca, Gaurhari Dass, Sarah M. Keating, Young M. Park, Nicola Buso, Nicolas Rodriguez, Michael Hucka, and Henning Hermjakob. BioModels-15 years of sharing computational models in life science. 48:D407–D415.

[19] Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. pages 1–9. Publisher: Nature Publishing Group.

[20] Gurdeep Singh, Lilija Wehling, Ahmad Wisnu Mulyadi, Rakesh Hadne Sreenath, Thomas Klabunde, Tommaso Andreani, and Douglas McCloskey. Talk2Biomodels and Talk2KnowledgeGraph: AI agent-based application for prediction of patient biomarkers and reasoning over biomedical knowledge graphs. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

[21] James P Sluka, Xiao Fu, Maciej Swat, Julio M Belmonte, Alin Cosmanescu, Sherry G Clendenon, John F Wambaugh, and James A Glazier. A liver-centric multiscale modeling framework for xenobiotics. *PloS one*, 11:e0162428, 2016.

[22] Joana Torres, Saurabh Mehandru, Jean-Frédéric Colombel, and Laurent Peyrin-Biroulet. Crohn's disease. *Lancet (London, England)*, 389:1741–1755, 2017.

[23] Ken Wang, Neil John Parrott, and Thierry Lavé. Embracing the future of medicine with virtual patients. 30(3):104322.

[24] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*.

# A Summary of QSP models used for benchmark

Table 1: Summary of QSP models used for benchmark. Each entry lists reference and model details, simulation settings, and BioModels identifiers where available, respectively.

| # | Reference | # Treatment(s) | # Virtual Patient(s) | # Species | # Parameters | Simulation Time | Disease | BioModels ID |
|---|---|---|---|---|---|---|---|---|
| 1 | Proprietary Model | 5 | 1000 | 16 | – | 364 days | Asthma | – |
| 2 | Proprietary Model | 6 | 71 | 265 | – | 300 days | IBD | – |
| 3 | Faratian et al., 2009[9] | 2 | 1 | 55 | 114 | 60 mins | Cancer (HER2-positive breast cancer) | BIOMD0000000424 |
| 4 | Benson et al., 2014[6] | 7 | 1 | 39 | 155 | 350 hrs | Osteoarthritic Pain | BIOMD0000000512 |
| 5 | Sluka et al., 2016[21] | 1 | 1 | 7 | 9 | 40 mins | Acute Liver Failure | BIOMD0000000624 |

## B Performance of time series foundation models on selected QSP models

Table 2: Forecasting performances of time series foundation models on the proprietary QSP model of Asthma. C and H refer to the context and horizon window percentage over total timestep, respectively.

| Scenario | Baseline Model | MAE | MSE | RMSE | SMAPE |
|---|---|---|---|---|---|
| C: 75%, H: 25% | $\text{Chronos}_{\text{Small}}$[2] | 0.0042 ± 0.0041 | 0.0029 ± 0.0057 | 0.0071 ± 0.0091 | 0.0257 ± 0.0129 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.0048 ± 0.0050 | 0.0026 ± 0.0051 | 0.0068 ± 0.0082 | 0.0263 ± 0.0131 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0026 ± 0.0014 | 0.0002 ± 0.0002 | 0.0033 ± 0.0022 | 0.0230 ± 0.0120 |
| | TiRex[4] | 0.0044 ± 0.0072 | 0.0034 ± 0.0067 | 0.0063 ± 0.0106 | 0.0067 ± 0.0060 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.0184 ± 0.0297 | 0.0426 ± 0.0841 | 0.0229 ± 0.0374 | 0.0280 ± 0.0190 |
| C: 50%, H: 50% | $\text{Chronos}_{\text{Small}}$[2] | 0.0315 ± 0.0414 | 0.0754 ± 0.1344 | 0.0392 ± 0.0510 | 0.0664 ± 0.0297 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.0224 ± 0.0128 | 0.0234 ± 0.0263 | 0.0304 ± 0.0191 | 0.0845 ± 0.0423 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0209 ± 0.0153 | 0.0240 ± 0.0316 | 0.0282 ± 0.0221 | 0.0688 ± 0.0284 |
| | TiRex[4] | 0.0134 ± 0.0210 | 0.0284 ± 0.0562 | 0.0187 ± 0.0299 | 0.0163 ± 0.0137 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.0233 ± 0.0249 | 0.0395 ± 0.0650 | 0.0285 ± 0.0306 | 0.0364 ± 0.0192 |
| C: 25%, H: 75% | $\text{Chronos}_{\text{Small}}$[2] | 0.0500 ± 0.0421 | 0.1102 ± 0.1804 | 0.0605 ± 0.0498 | 0.1361 ± 0.0546 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.0611 ± 0.0391 | 0.1176 ± 0.1625 | 0.0759 ± 0.0472 | 0.1759 ± 0.0755 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0606 ± 0.0465 | 0.1204 ± 0.1850 | 0.0741 ± 0.0555 | 0.1604 ± 0.0580 |
| | TiRex[4] | 0.0287 ± 0.0251 | 0.0474 ± 0.0790 | 0.0365 ± 0.0332 | 0.0517 ± 0.0340 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.0389 ± 0.0301 | 0.0514 ± 0.0572 | 0.0457 ± 0.0355 | 0.0754 ± 0.0534 |

Table 3: Forecasting performance of time series foundation models on the proprietary QSP model of IBD. C and H refer to the context and horizon window percentage over total timestep, respectively.

| Scenario | Baseline Model | MAE | MSE | RMSE | SMAPE |
|---|---|---|---|---|---|
| C: 75%, H: 25% | $\text{Chronos}_{\text{Small}}$[2] | 0.0754 ± 0.0592 | 0.1592 ± 0.1472 | 0.0941 ± 0.0719 | 0.4247 ± 0.0260 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.0798 ± 0.0757 | 0.2462 ± 0.4092 | 0.1008 ± 0.0949 | 0.4353 ± 0.0325 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0949 ± 0.0982 | 0.4383 ± 0.7836 | 0.1159 ± 0.1217 | 0.4361 ± 0.0333 |
| | TiRex[4] | 0.0594 ± 0.0539 | 0.0729 ± 0.0829 | 0.0789 ± 0.0675 | 0.3798 ± 0.0492 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.0779 ± 0.0768 | 0.1011 ± 0.1113 | 0.1013 ± 0.0898 | 0.3968 ± 0.0443 |
| C: 50%, H: 50% | $\text{Chronos}_{\text{Small}}$[2] | 0.1254 ± 0.1059 | 0.1960 ± 0.1978 | 0.1601 ± 0.1287 | 0.4625 ± 0.0376 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.1894 ± 0.1288 | 0.4883 ± 0.4556 | 0.2445 ± 0.1608 | 0.4956 ± 0.0592 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.1270 ± 0.0824 | 0.2200 ± 0.1702 | 0.1636 ± 0.0998 | 0.4627 ± 0.0433 |
| | TiRex[4] | 0.1139 ± 0.1239 | 0.1933 ± 0.2570 | 0.1446 ± 0.1487 | 0.3876 ± 0.0697 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.1639 ± 0.1694 | 0.2986 ± 0.3785 | 0.1963 ± 0.1938 | 0.4557 ± 0.0984 |
| C: 25%, H: 75% | $\text{Chronos}_{\text{Small}}$[2] | 0.2008 ± 0.1450 | 0.3739 ± 0.4040 | 0.2492 ± 0.1813 | 0.5423 ± 0.0821 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.2625 ± 0.1615 | 0.6878 ± 0.5638 | 0.3362 ± 0.2067 | 0.5614 ± 0.0991 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.2466 ± 0.1495 | 0.6019 ± 0.5046 | 0.3054 ± 0.1882 | 0.5524 ± 0.0879 |
| | TiRex[4] | 0.2198 ± 0.1697 | 0.4880 ± 0.4601 | 0.2690 ± 0.2118 | 0.4836 ± 0.1074 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.2335 ± 0.1576 | 0.5126 ± 0.4316 | 0.2772 ± 0.1919 | 0.5088 ± 0.0864 |

Table 4: Forecasting performance of time series foundation models on the open-source QSP model of HER2-positive breast cancer (BIOMD0000000424). C and H refer to the context and horizon window percentage over total timestep, respectively.

| Scenario | Baseline Model | MAE | MSE | RMSE | SMAPE |
|---|---|---|---|---|---|
| C: 75%, H: 25% | $\text{Chronos}_{\text{Small}}$[2] | 0.0361 ± 0.0153 | 0.0126 ± 0.0092 | 0.0449 ± 0.0189 | 0.2260 ± 0.0386 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.0387 ± 0.0236 | 0.0168 ± 0.0156 | 0.0493 ± 0.0310 | 0.2402 ± 0.0206 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0327 ± 0.0173 | 0.0123 ± 0.0100 | 0.0394 ± 0.0215 | 0.2196 ± 0.0411 |
| | TiRex[4] | 0.0248 ± 0.0092 | 0.0057 ± 0.0047 | 0.0288 ± 0.0109 | 0.2740 ± 0.0267 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.1097 ± 0.0522 | 0.0537 ± 0.0398 | 0.1231 ± 0.0586 | 0.3379 ± 0.0795 |
| C: 50%, H: 50% | $\text{Chronos}_{\text{Small}}$[2] | 0.1170 ± 0.0526 | 0.0739 ± 0.0453 | 0.1427 ± 0.0678 | 0.3288 ± 0.0314 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.1112 ± 0.0509 | 0.0876 ± 0.0634 | 0.1359 ± 0.0633 | 0.3061 ± 0.0097 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0965 ± 0.0425 | 0.0629 ± 0.0435 | 0.1205 ± 0.0563 | 0.2936 ± 0.0040 |
| | TiRex[4] | 0.0846 ± 0.0325 | 0.0377 ± 0.0263 | 0.1015 ± 0.0408 | 0.3140 ± 0.0454 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.2058 ± 0.0614 | 0.1683 ± 0.0943 | 0.2420 ± 0.0747 | 0.4429 ± 0.0276 |
| C: 25%, H: 75% | $\text{Chronos}_{\text{Small}}$[2] | 0.1948 ± 0.0880 | 0.1941 ± 0.1136 | 0.2351 ± 0.1057 | 0.3683 ± 0.0508 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.2249 ± 0.1009 | 0.2403 ± 0.1478 | 0.2700 ± 0.1199 | 0.4004 ± 0.0728 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.1994 ± 0.0914 | 0.1759 ± 0.1052 | 0.2396 ± 0.1122 | 0.3982 ± 0.0671 |
| | TiRex[4] | 0.3300 ± 0.0737 | 0.6500 ± 0.1661 | 0.4209 ± 0.0843 | 0.5181 ± 0.0172 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.3544 ± 0.0985 | 0.3708 ± 0.1591 | 0.4138 ± 0.1105 | 0.5007 ± 0.0403 |

Table 5: Forecasting performance of time series foundation models on the open-source QSP model of Osteoarthritic pain (BIOMD0000000512). C and H refer to the context and horizon window percentage over total timestep, respectively.

| Scenario | Baseline Model | MAE | MSE | RMSE | SMAPE |
|---|---|---|---|---|---|
| C: 75%, H: 25% | $\text{Chronos}_{\text{Small}}$[2] | 0.1004 ± 0.0616 | 0.0319 ± 0.0245 | 0.1238 ± 0.0751 | 0.4153 ± 0.0806 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.1120 ± 0.0666 | 0.0385 ± 0.0344 | 0.1387 ± 0.0777 | 0.4473 ± 0.0774 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.1241 ± 0.0667 | 0.0432 ± 0.0300 | 0.1556 ± 0.0795 | 0.4666 ± 0.0907 |
| | TiRex[4] | 0.2198 ± 0.1898 | 0.2332 ± 0.2127 | 0.2820 ± 0.2425 | 0.4956 ± 0.1878 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.2630 ± 0.2477 | 0.2671 ± 0.3434 | 0.3190 ± 0.2965 | 0.5198 ± 0.1820 |
| C: 50%, H: 50% | $\text{Chronos}_{\text{Small}}$[2] | 0.4175 ± 0.2870 | 0.4545 ± 0.4544 | 0.4963 ± 0.3413 | 0.6527 ± 0.1572 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.6657 ± 0.5419 | 1.4997 ± 1.7180 | 0.8288 ± 0.6666 | 0.7418 ± 0.2542 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.7386 ± 0.6492 | 2.4638 ± 3.5164 | 0.9006 ± 0.7770 | 0.7353 ± 0.2255 |
| | TiRex[4] | 0.4368 ± 0.3238 | 0.5358 ± 0.4813 | 0.5227 ± 0.3741 | 0.6268 ± 0.2015 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.6728 ± 0.4547 | 1.0207 ± 0.8413 | 0.7522 ± 0.5042 | 0.7331 ± 0.2400 |
| C: 25%, H: 75% | $\text{Chronos}_{\text{Small}}$[2] | 0.7093 ± 0.2639 | 1.1106 ± 0.6097 | 0.8764 ± 0.3530 | 0.7602 ± 0.1029 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.8825 ± 0.4047 | 3.4756 ± 4.3879 | 1.1361 ± 0.5790 | 0.8018 ± 0.1485 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.8219 ± 0.3624 | 1.8939 ± 1.6302 | 1.0236 ± 0.4592 | 0.7942 ± 0.1680 |
| | TiRex[4] | 0.7186 ± 0.3549 | 1.2354 ± 0.6449 | 0.8984 ± 0.4381 | 0.7270 ± 0.1864 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.8511 ± 0.3454 | 1.4294 ± 0.7262 | 1.0156 ± 0.4180 | 0.7974 ± 0.1736 |

Table 6: Forecasting performance of time series foundation models on the open-source QSP model of Acute Liver Failure (BIOMD0000000624). C and H refer to the context and horizon window percentage over total timestep, respectively.

| Scenario | Baseline Model | MAE | MSE | RMSE | SMAPE |
|---|---|---|---|---|---|
| C: 75%, H: 25% | $\text{Chronos}_{\text{Small}}$[2] | 0.0331 ± 0.0000 | 0.0024 ± 0.0000 | 0.0446 ± 0.0000 | 0.0287 ± 0.0000 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.0301 ± 0.0000 | 0.0019 ± 0.0000 | 0.0412 ± 0.0000 | 0.0774 ± 0.0000 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.0332 ± 0.0000 | 0.0032 ± 0.0000 | 0.0481 ± 0.0000 | 0.0242 ± 0.0000 |
| | TiRex[4] | 0.0688 ± 0.0000 | 0.0085 ± 0.0000 | 0.0885 ± 0.0000 | 0.2040 ± 0.0000 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.0374 ± 0.0000 | 0.0023 ± 0.0000 | 0.0447 ± 0.0000 | 0.0945 ± 0.0000 |
| C: 50%, H: 50% | $\text{Chronos}_{\text{Small}}$[2] | 0.1124 ± 0.0000 | 0.0342 ± 0.0000 | 0.1548 ± 0.0000 | 0.2287 ± 0.0000 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.2171 ± 0.0000 | 0.0705 ± 0.0000 | 0.2496 ± 0.0000 | 0.1963 ± 0.0000 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.3102 ± 0.0000 | 0.1683 ± 0.0000 | 0.3683 ± 0.0000 | 0.2375 ± 0.0000 |
| | TiRex[4] | 0.3011 ± 0.0000 | 0.1359 ± 0.0000 | 0.3566 ± 0.0000 | 0.3345 ± 0.0000 |
| | $\text{Moirai}_{\text{Small}}$[24] | 0.3166 ± 0.0000 | 0.2044 ± 0.0000 | 0.3688 ± 0.0000 | 0.2702 ± 0.0000 |
| C: 25%, H: 75% | $\text{Chronos}_{\text{Small}}$[2] | 0.5170 ± 0.0000 | 0.3825 ± 0.0000 | 0.5843 ± 0.0000 | 0.2788 ± 0.0000 |
| | $\text{Chronos}_{\text{Base}}$[2] | 0.8271 ± 0.0000 | 0.9704 ± 0.0000 | 0.9433 ± 0.0000 | 0.4125 ± 0.0000 |
| | $\text{Chronos}_{\text{Large}}$[2] | 0.8266 ± 0.0000 | 0.8229 ± 0.0000 | 0.9051 ± 0.0000 | 0.4160 ± 0.0000 |
| | TiRex[4] | 1.1806 ± 0.0000 | 2.4308 ± 0.0000 | 1.5135 ± 0.0000 | 0.3892 ± 0.0000 |
| | $\text{Moirai}_{\text{Small}}$[24] | 1.1475 ± 0.0000 | 1.6292 ± 0.0000 | 1.2724 ± 0.0000 | 0.5263 ± 0.0000 |

## C Selection of representative virtual patients

To select representative Virtual Patient (VP) trajectories for visualising, temporal clustering was performed using dynamic time warping (DTW) followed by hierarchical clustering on DTW distances. DTW allows flexible alignment of time series that differ in timing or duration, providing a robust measure of trajectory similarity. The optimal number of clusters was determined using the elbow method, which balances within-cluster homogeneity against model complexity. For each cluster, a representative (medoid) trajectory was identified (the VP whose time course was most central within its cluster based on DTW distance). These representative profiles summarize the dominant temporal patterns observed across the virtual population.