Generative Time Series Models with Interpretable Latent Processes for Complex Disease Trajectories

Cécile Trottet*
University of Zurich
cecileclaire.trottet@uzh.ch

Manuel Schürch*
University of Zurich
manuel.schuerch@uzh.ch

Ahmed Allam
University of Zurich

Amina Mollaysa
University of Zurich

Imon Barua
University of Oslo

Liubov Petelytska
University Hospital Zurich

Oliver Distler
University Hospital Zurich

Anna-Maria Hoffmann-Vold
University Hospital Zurich

Michael Krauthammer
University of Zurich

Abstract

We propose a deep generative time series approach using latent temporal processes for modeling and holistically analyzing complex disease trajectories and demonstrate its effectiveness in modeling systemic sclerosis. We aim to find meaningful temporal latent representations of an underlying generative process that explain the observed disease trajectories in an interpretable and comprehensive way. To enhance the interpretability of these latent temporal processes, we develop a semi-supervised approach for disentangling the latent space using established medical concepts. We show that the learned temporal latent processes can be utilized for further data analysis, including finding similar patients and clustering the disease into new sub-types. Moreover, our method enables personalized online monitoring and prediction of multivariate time series including uncertainty quantification.

1 Introduction

Understanding clinical trajectories of complex diseases is crucial for improving diagnosis, treatment, and patient outcomes. However, modeling such multivariate time series data poses significant challenges due to the high dimensionality of clinical measurements, low signal-to-noise ratio, sparsity, and the complex interplay of various, potentially unobserved, factors. We present a deep generative temporal model that captures both the joint distribution of all the observed longitudinal clinical variables and of the latent temporal variables (Figure 1a), suited for the holistic analysis of temporal disease trajectories. We propose a semi-supervised approach for disentangling the latent space using known medical concepts to enhance the interpretability and allowing for the discovery of novel medically-driven patterns in the data. We demonstrate the effectiveness of our method in modeling the progression of systemic sclerosis (SSc), a severe and yet only partially understood autoimmune disease. SSc triggers the immune system to attack the body's connective tissue, causing severe damage to multiple organs. We seek to understand the evolution of SSc by modeling the patterns of organ involvement and progression and aim to learn temporal hidden representations that distinctly capture the disentangled medical concepts related to each organ.

There are various extensions to Kingma and Welling's seminal work on Variational Autoencoders (VAEs) [16], explicitly modeling time in the latent space such as RNN-VAE [8], GP-VAE [5, 9], or

*These authors contributed equally to this work.

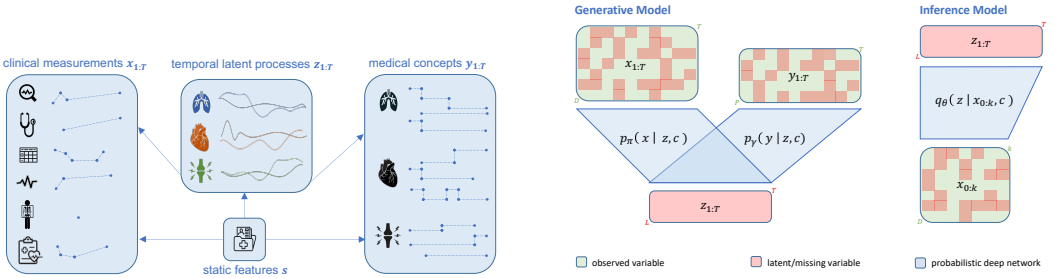
1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

longitudinal VAE [24]. While these approaches have showcased remarkable efficacy in modeling time series, the interpretability of the resulting latent processes remains limited for complex data. Thus, there is ongoing research in designing temporal generative models with disentangled latent factors, such as disentangled sequential VAE [14] and disentangled GP-VAE [2]. However, learning interpretable and disentangled latent representations is highly difficult for complex data without any inductive bias [17], leading researchers to focus on weakly supervised latent representation learning [18, 35, 22]. In a similar spirit, we tackle the *temporal* semi-supervised guidance of the latent space by providing sparse labels representing established medical domain knowledge concepts.

2 Methodology

We analyze patient histories that consist of two main types of data: raw temporal clinical measurements $\boldsymbol{x} = \boldsymbol{x}_{1:T} \in \mathbb{R}^{D\times T}$, such as blood pressure, and sparse medical concept labels $\boldsymbol{y} = \boldsymbol{y}_{1:T} \in \mathbb{R}^{P\times T}$, describing higher-level medical definitions related to the disease, for instance, the medical definition of severity staging of the heart involvement (Figure 1). The medical concept definitions are typically derived from multiple clinical measurements using logical operations. For example, a patient may be classified as having "severe heart involvement" if certain conditions are satisfied, for instance, $\boldsymbol{x}^{(i)} > \varepsilon$ AND $\boldsymbol{x}^{(j)} = 1$. Both the raw measurements and labels are irregularly sampled, and we denote by $\boldsymbol{\tau}_{1:T} \in \mathbb{R}^T$ the vector of sampling time-points. Moreover, static information denoted as $\boldsymbol{s} \in \mathbb{R}^S$ is present, alongside additional temporal covariates such as medications $\boldsymbol{p}_{1:T} \in \mathbb{R}^{P\times T}$ for each patient.

We condition our generative model on the context variable $\boldsymbol{c} = \{\boldsymbol{\tau}, \boldsymbol{p}, \boldsymbol{s}\}$ to be able to generate latent processes under certain conditions, for instance when a specific medication is administered. Furthermore, in the next sections, we introduce our approach to learning multivariate latent processes denoted as $\boldsymbol{z} = \boldsymbol{z}_{1:T} \in \mathbb{R}^{L\times T}$, responsible for generating both the raw clinical measurement processes $\boldsymbol{x}_{1:T}$ and the medical labels $\boldsymbol{y}_{1:T}$. In particular, we use the different temporal medical concepts to disentangle the L dimensions of the latent processes by allocating distinct dimensions to represent different medical concepts. We assume a dataset $\{\boldsymbol{x}^i_{1:T_i}, \boldsymbol{y}^i_{1:T_i}, \boldsymbol{c}^i_{1:T_i}\}_{i=1}^N$ of N patients, and omit the dependency to i and the time index when the context is clear.



(a) Temporal generative model for systemic sclerosis. (b) Semi-supervised temporal latent variable model.

Figure 1: Modeling approach

2.1 Generative Model

We propose the probabilistic conditional generative latent variable model $p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c}) = p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})p_\phi(\boldsymbol{z}|\boldsymbol{c})$, with learnable prior network $p_\phi(\boldsymbol{z}|\boldsymbol{c})$, likelihood network $p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, and guidance network $p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})$, where $\psi = \{\gamma, \pi, \phi\}$ are learnable parameters (Figure 1b). We learn the prior network for $\boldsymbol{z}$, i.e. $p_\phi(\boldsymbol{z}|\boldsymbol{c}) = \prod_{t=1}^T \prod_{l=1}^L \mathcal{N}\left(\boldsymbol{z}_t^l | \mu_\phi^l(\boldsymbol{c}_t), \sigma_\phi^l(\boldsymbol{c}_t)\right)$ conditioned on $\boldsymbol{c}$, so that time-varying or demographic effects can be learned in the prior (Appendix D.2.1). The means $\mu_\phi^l(\boldsymbol{c}_t)$ and variances $\sigma_\phi^l(\boldsymbol{c}_t)$ are deep neural networks and we assume a factorized Gaussian prior distribution per time and latent dimensions.

The probabilistic likelihood network maps $\boldsymbol{z}$ and $\boldsymbol{c}$ to $\boldsymbol{x}$, i.e. $p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) = \prod_{t=1}^T \prod_{d\in\mathcal{G}} \mathcal{N}(x_t^d | \mu_\pi^d, \sigma_\pi^d) \prod_{d\in\mathcal{K}} \mathcal{C}(x_t^d | p_\pi^d)$, where we assume time- and feature-wise conditional independence. We assume either Gaussian $\mathcal{N}$ or categorical $\mathcal{C}$ likelihoods for the observed

variables $\boldsymbol{x}$, where $\mathcal{G}$ and $\mathcal{K}$ are the corresponding indices. The mean $\mu_\pi^d = \mu_\pi^d(\boldsymbol{z}_t, \boldsymbol{c}_t)$, variance $\sigma_\pi^d = \sigma_\pi^d(\boldsymbol{z}_t, \boldsymbol{c}_t)$, and category probability vector $p_\pi^d = p_\pi^d(\boldsymbol{z}_t, \boldsymbol{c}_t)$ are deep parametrized functions. We propose a semi-supervised approach to disentangle $\boldsymbol{z}$ with respect to defined medical concepts $\boldsymbol{y}$. In particular, we assume $p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) = \prod_{t=1}^T \prod_{g=1}^G \prod_{j \in \nu(g)} \mathcal{C}(y_t^j | h_\gamma^j(\boldsymbol{z}_t^{\varepsilon(g)}, \boldsymbol{c}_t))$ where $h_\gamma^j(\boldsymbol{z}_t^{\varepsilon(g)}, \boldsymbol{c}_t)$ is a deep parametrized probability vector, and $\nu(g)$ and $\varepsilon(g)$ correspond to the indices of the gth guided medical concept, and the indices in the latent space defined for guided concept g, respectively.

2.2 Probabilistic inference

We approximate $p_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c})$, with an amortized variational distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ (Appendix B.1). We assume $q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) = \prod_{t=1}^T \prod_{l=1}^L \mathcal{N}(z_t^l | \mu_\theta^l(\boldsymbol{x}_{0:k}, \boldsymbol{c}), \sigma_\theta^l(\boldsymbol{x}_{0:k}, \boldsymbol{c}))$ with variational parameters θ and $0 \leq k \leq T$. Note that only the measurements $\boldsymbol{x}_{0:k}$ until observation k are part of the variational distribution, and not the medical concepts $\boldsymbol{y}$, since the latter are sparse and medically defined, in contrast to the raw clinical measurements. If $0 \leq k < T$, we forecast the future latent variables $\boldsymbol{z}_{k+1:T}$ from $\boldsymbol{x}_{0:k}$. We apply amortized variational inference [3] by maximizing a lower bound $\log p_\psi(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c}) \geq \mathcal{L}(\psi, \theta; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c})$ of the intractable marginal log likelihood. For a fixed k, this leads to the objective function

$$\begin{aligned}\mathcal{L}_k(\psi, \theta; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}) = &\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})} \left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right] \\ &+ \alpha\, \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})} \left[\log p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})\right] \\ &- \beta\, KL\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) \,||\, p_\phi(\boldsymbol{z}|\boldsymbol{c})\right],\end{aligned}$$

where we introduce weights α and β inspired by the disentangled $\beta-$VAE [11]. The first term $\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})} \left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right]$ is unsupervised, whereas $\alpha \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})} \left[\log p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})\right]$ is supervised and $\beta KL\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) || p_\phi(\boldsymbol{z}|\boldsymbol{c})\right]$ is a regularization term, ensuring that the posterior is close to the prior with respect to the Kullback-Leibler (KL) divergence. Since all dimensions in $\boldsymbol{z}$ are connected to all the measurements $\boldsymbol{x}$, the potential correlations between clinically measured variables can be exploited in an unsupervised fashion while disentangling $\boldsymbol{z}$ using the guidance networks for $\boldsymbol{y}$. The expectation $\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})}$ is approximated with Monte-Carlo (MC) sampling (Appendix B.1).

Given a dataset with N iid patients, the optimal parameters are given by $\psi^*, \theta^* = \text{argmax}_{\psi, \theta} \sum_{i=1}^N \sum_{k=0}^{T_i} \mathcal{L}_k(\psi, \theta; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{c}^i)$, which is computed with stochastic optimization using mini-batches of patients and different values of k (Appendix B.1.2). The predictive distributions $q_*(\boldsymbol{y}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) = \int p_{\gamma^*}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) q_{\theta^*}(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) d\boldsymbol{z}$ and $q_*(\boldsymbol{x}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) = \int p_{\pi^*}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) q_{\theta^*}(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) d\boldsymbol{z}$ are approximated with a two-stage MC sampling (Appendix B.1.3). The former can be used to automatically label and forecast $\boldsymbol{y}$ based on the partially observed $\boldsymbol{x}$, whereas the latter corresponds to the reconstruction and forecasting of partially observed trajectories.

2.3 Patient Similarity and Clustering

The learned $q_{\theta^*}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}, \boldsymbol{c}_{1:T})$ can map any observed patient trajectory $\mathcal{T}_i = \{\boldsymbol{x}_{1:T_i}^i, \boldsymbol{c}_{1:T_i}^i\}$ to its latent trajectory $\mathcal{H}_i = \mathbb{E}_{q_\theta(\boldsymbol{z}_{1:T_i}^i | \boldsymbol{x}_{1:T_i}^i, \boldsymbol{c}_{1:T_i}^i)} \left[\boldsymbol{z}_{1:T_i}^i\right]$. Through our semi-supervised approach, $\mathcal{H}_i$ captures the important elements from $\boldsymbol{x}_{1:T_i}^i$ and $\boldsymbol{y}_{1:T_i}^i$, without explicitly depending on $\boldsymbol{y}_{1:T_i}^i$. Since defining a similarity metric between two trajectories in the original space is challenging, due to the missingness and high dimensionality, we instead define it in the latent space. To measure the similarity between latent trajectories, we employ the *dynamic-time-warping (dtw)* metric to account for the different trajectory lengths and misalignments [20].

2.4 Modeling Systemic Sclerosis

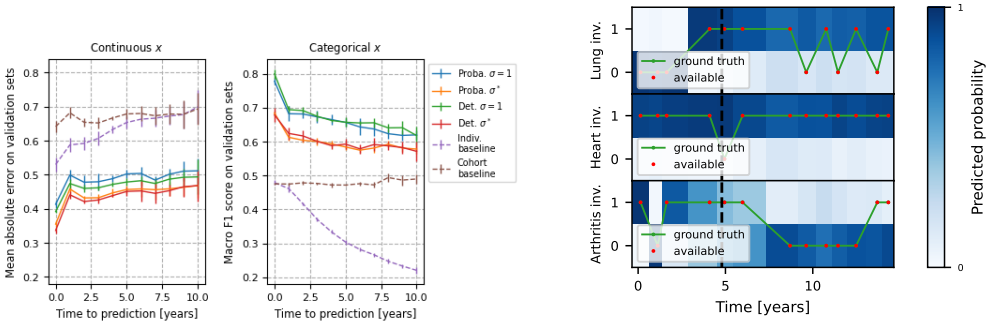
We aim to model the overall SSc disease trajectories as well as the distinct organ involvement trajectories for patients from the European Scleroderma Trials and Research (EUSTAR) database (Appendix E.1). We focus on the involvement of the lung, heart, and joints (arthritis) in SSc. Each organ has two related medical concepts: *involvement* and *stage*. Based upon the medical definitions provided in Appendix E.2, for each of the three organs $\mathcal{O} := \{lung, heart, joints\}$, we create labels signaling the organ involvement (yes/no) and severity stage $(1-4)$, respectively. We write $o(m)$, $m \in M := \{involvement, stage\}$, $o \in \mathcal{O}$ to refer to the corresponding medical concept for organ o. For each organ, we guide a distinct subset of latent processes (non-overlapping subsets and each

dimension in $\boldsymbol{z}$ is guided), leading to $p_\gamma(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c}) = \prod_{t=1}^{T}\prod_{o\in\mathcal{O}}\prod_{m\in M} p_\gamma(\boldsymbol{y}_t^{\nu(o(m))}|\boldsymbol{z}_t^{\varepsilon(o(m))},\boldsymbol{c}_t)$. The implementation details are in Appendix C.2.

3 Experiments and Results

3.1 Model Evaluation

We assessed the model's performance at predicting the $\boldsymbol{x}$ trajectories in probabilistic and deterministic settings, and with either learning the likelihood network variance σ^* or setting $\sigma = 1$ (Figure 2a). All



(a) Performance for $\boldsymbol{x}$ prediction.

(b) Probabilities of organ involv. for p_{idx}.

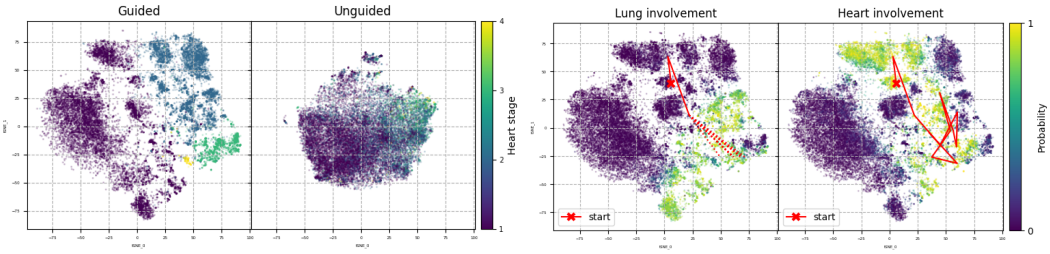
Figure 2: Prediction performance and online monitoring.

of the models outperform non-ML-driven individualized or cohort-based baselines. For a variable x^j, the individualized baseline predicts the patient's last available measurement for x^j as its future value. The cohort baseline samples a value from the empirical Gaussian/Categorical distribution of x^j. In Appendix D.1, we provide further details on the inference process and evaluation results. In particular we compare our model to an unguided baseline, assess the performance for $\boldsymbol{y}$ prediction, and compute the coverage and calibration of the predictions. The probabilistic model with learned σ^* strikes the best balance between predictive capabilities, coverage, and generative ability. In the following, we explore further capabilities of this model.

Online Prediction with Uncertainty To illustrate how the model allows a holistic understanding of a patient's disease course, we follow an index patient p_{idx} with a complex disease trajectory and various impacted organs. Our model forecasts the high-dimensional distribution of $\boldsymbol{x}_{1:T}$ and $\boldsymbol{y}_{1:T}$ given the past measurements $\boldsymbol{x}_{0:k}$. For example, the heatmaps in Figure 2b show the predicted probabilities of organ involvement at a given time (values after the dashed line are forecasted). We provide additional plots for $\boldsymbol{x}$ and $\boldsymbol{y}$ prediction in Appendix D.1.1.

3.2 Cohort Analysis

3.2.1 Latent Space and Medical Concepts



(a) Guided versus unguided latent spaces.

(b) Probabilities of lung and heart involvement.

Figure 3: $t-$SNE visualizations of latent spaces.

We can analyze $\boldsymbol{z}$ to find cohort-level disease patterns, i.e. global trends in the cohort. Figure 3a compares the distribution of medical concept ground truth labels (*heart stage*) in guided versus unguided models (i.e. without training guidance networks). The guided model clearly increases the disentanglement with respect to medical concepts, thus enhancing the interpretability of $\boldsymbol{z}$.

Figure 3b shows $\boldsymbol{z}$ overlaid with different predicted probabilities of organ involvement. The red line highlights the trajectory of p_{idx} in $\boldsymbol{z}$ with respect to different medical concepts. The second panel (solid line) shows the complete reconstructed trajectory of p_{idx} in $\boldsymbol{z}$, and in the first panel we sample forecasted $\boldsymbol{z}$ trajectories (dotted lines), providing estimates of future disease phases.

3.2.2 Clustering and Similarity of Patient Trajectories

We clustered and retrieved similar latent trajectories using *k-means* and *k-nn* with the *dtw* similarity measure [30]. Figure 4 shows the three mean cluster trajectories in $\boldsymbol{z}$. The first and second cluster

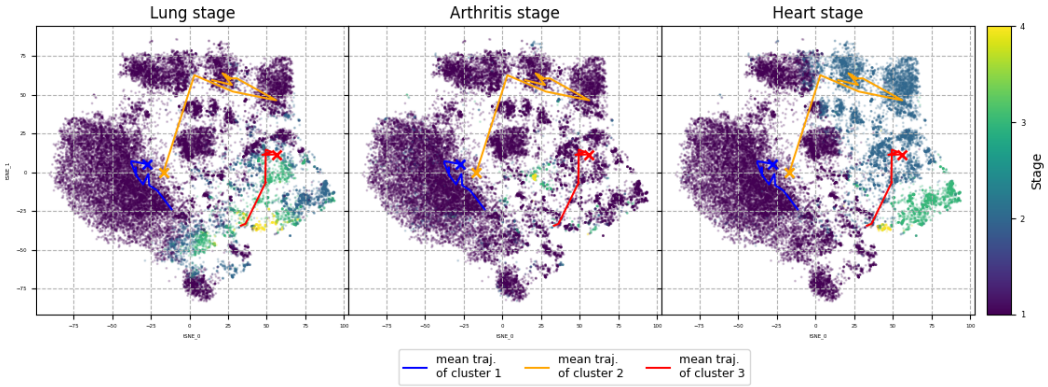


Figure 4: Mean cluster trajectories (start at the cross **x**) overlaid with predicted organ stages.

trajectories start close to each other with the second progressing to regions with higher heart stages. The third cluster contains the most severely progressing patients. In Appendix D.2.3, we compute the medical concept probabilities for the mean cluster trajectories and compare the $\boldsymbol{y}$ and $\boldsymbol{z}$ trajectories of p_{idx} and its 3 nearest neighbors.

4 Conclusion

We present a novel deep semi-supervised generative latent variable approach to model complex disease trajectories. With the guidance networks, we propose a method to augment unsupervised deep generative models with established medical concepts and achieve more interpretable and disentangled latent processes. Our non-discriminative approach effectively addresses desiderata for healthcare models such as forecasting, uncertainty quantification, dimensionality reduction, and interpretability. Furthermore, we empirically show that our model is suited for a real-world use case, and enables a holistic understanding of the patients' disease course. The disentangled latent space facilitates comprehensive trajectory visualizations, clustering, and forecasting. Both our presented experiments and modeling approaches hold the potential to be extended and adapted in many ways. In future work, we intend to extend our framework to handle continuous time (Appendix B.2), include medications for generating future hypothetical conditional trajectories (Appendix B.3), and also include guidance networks to model additional disease dynamics like long-term outcomes.

References

[1] Ahmed Allam, Stefan Feuerriegel, Michael Rebhan, and Michael Krauthammer. Analyzing patient trajectories with artificial intelligence. *Journal of medical internet research*, 23(12): e29812, 2021.

[2] Simon Bing, Vincent Fortuin, and Gunnar Rätsch. On disentanglement in gaussian process variational autoencoders. *arXiv preprint arXiv:2102.05507*, 2021.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[4] Francesco Bonomi, Silvia Peretti, Gemma Lepri, Vincenzo Venerito, Edda Russo, Cosimo Bruni, Florenzo Iannone, Sabina Tangaro, Amedeo Amedei, Serena Guiducci, et al. The use and utility of machine learning in achieving precision medicine in systemic sclerosis: A narrative review. *Journal of Personalized Medicine*, 12(8):1198, 2022.

[5] Francesco Paolo Casale, Adrian Dalca, Luca Saglietti, Jennifer Listgarten, and Nicolo Fusi. Gaussian process prior variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

[6] Clément Chadebec, Louis Vincent, and Stephanie Allassonniere. Pythae: Unifying generative autoencoders in python - a benchmarking use case. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21575–21589. Curran Associates, Inc., 2022.

[7] Li-Fang Cheng, Bianca Dumitrascu, Gregory Darnell, Corey Chivers, Michael Draugelis, Kai Li, and Barbara E Engelhardt. Sparse multi-output gaussian processes for online medical time series prediction. *BMC medical informatics and decision making*, 20(1):1–23, 2020.

[8] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

[9] Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pages 1651–1661. PMLR, 2020.

[10] Alexandru Garaiman, Farhad Nooralahzadeh, Carina Mihai, Nicolas Perez Gonzalez, Nikitas Gkikopoulos, Mike Oliver Becker, Oliver Distler, Michael Krauthammer, and Britta Maurer. Vision transformer assisting rheumatologists in screening for capillaroscopy changes in systemic sclerosis: an artificial intelligence model. *Rheumatology*, page keac541, 2022.

[11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[13] Anna-Maria Hoffmann-Vold, Yannick Allanore, Margarida Alves, Cathrine Brunborg, Paolo Airó, Lidia P Ananieva, László Czirják, Serena Guiducci, Eric Hachulla, Mengtao Li, et al. Progressive interstitial lung disease in patients with systemic sclerosis-associated interstitial lung disease in the eustar database. *Annals of the rheumatic diseases*, 80(2):219–227, 2021.

[14] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *Advances in neural information processing systems*, 30, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. *The Journal of Machine Learning Research*, 21(1):8629–8690, 2020.

[18] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[19] Florian MP Meier, Klaus W Frommer, Robert Dinser, Ulrich A Walker, Laszlo Czirjak, Christopher P Denton, Yannick Allanore, Oliver Distler, Gabriela Riemekasten, Gabriele Valentini, et al. Update on the profile of the eustar cohort: an analysis of the eular scleroderma trials and research group database. *Annals of the rheumatic diseases*, 71(8):1355–1360, 2012.

[20] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[21] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[22] Emanuele Palumbo, Sonia Laguna, Daphné Chopard, and Julia E Vogt. Deep generative clustering with multimodal variational autoencoders. 2023.

[23] Pavlin G Poličar, Martin Stražar, and Blaž Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *BioRxiv*, page 731877, 2019.

[24] Siddharth Ramchandran, Gleb Tikhonov, Kalle Kujanpää, Miika Koskinen, and Harri Lähdesmäki. Longitudinal variational autoencoder. In *International Conference on Artificial Intelligence and Statistics*, pages 3898–3906. PMLR, 2021.

[25] Margherita Rosnati and Vincent Fortuin. Mgp-atttcn: An interpretable machine learning model for the prediction of sepsis. *Plos one*, 16(5):e0251248, 2021.

[26] Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective vae training with calibrated decoders. In *International Conference on Machine Learning*, pages 9179–9189. PMLR, 2021.

[27] Manuel Schürch, Dario Azzimonti, Alessio Benavoli, and Marco Zaffalon. Recursive estimation for sparse gaussian process regression. *Automatica*, 120:109127, 2020.

[28] Manuel Schürch, Dario Azzimonti, Alessio Benavoli, and Marco Zaffalon. Correlated product of experts for sparse gaussian process regression. *Machine Learning*, pages 1–22, 2023.

[29] Manuel Schürch, Xiang Li, Ahmed Allam, Giulia Rathmes, Amina Mollaysa, Claudia Cavelti-Weder, and Michael Krauthammer. Generating personalized insulin treatments strategies with deep conditional generative time series models. *arXiv preprint arXiv:2309.16521*, 2023.

[30] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *The Journal of Machine Learning Research*, 21(1):4686–4691, 2020.

[31] Jakub M. Tomczak. Deep Generative Modeling. *Deep Generative Modeling*, pages 1–197, 1 2022. doi: 10.1007/978-3-030-93158-2.

[32] Cécile Trottet, Ahmed Allam, Raphael Micheroli, Aron Horvath, Michael Krauthammer, and Caroline Ospelt. Explainable Deep Learning for Disease Activity Prediction in Chronic Inflammatory Joint Diseases. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[34] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[35] Jiageng Zhu, Hanchen Xie, and Wael Abd-Almageed. Sw-vae: Weakly supervised learn disentangled representation via latent factor swapping. In *European Conference on Computer Vision*, pages 73–87. Springer, 2022.

A Data and Code Availability

The dataset used is owned by a third party, the EUSTAR group, and may be obtained by request after the approval and permission from EUSTAR. The code builds upon the pythae library [6] and is publicly available at `https://github.com/uzh-dqbm-cmi/eustar_dgm4h` with examples from artificially generated data.

B Details and Extensions for Generative Model

In this section, we provide more details and several possible extensions to the main temporal generative model presented in Section 2.1.

B.1 Inference

In this section, we explain the inference process of the proposed generative model $p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c}) = p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})p_\phi(\boldsymbol{z}|\boldsymbol{c})$ in more detail. We are particularly interested in the posterior of the latent variables $\boldsymbol{z}$ given $\boldsymbol{y}$, $\boldsymbol{x}$, and $\boldsymbol{c}$, that is,

$$p_\psi(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{c}) = \frac{p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c})}{p_\psi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{c})} = \frac{p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c})}{\int p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c})d\boldsymbol{z}},$$

which is in general intractable due to the marginalization of the latent process in the marginal likelihood $p_\psi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{c}) = \int p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c})d\boldsymbol{z}$. Therefore, we resort to approximate inference, in particular, amortized variational inference (VI) [3], where a variational distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ close to the true posterior distribution $p_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}) \approx q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ is introduced. The similarity between these distributions is usually measured in terms of KL divergence [21], therefore, we aim to find parameters satisfying

$$\theta^*, \psi^* = \underset{\theta,\psi}{\operatorname{argmin}} \, KL\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) \,||\, p_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c})\right].$$

This optimization problem is equivalent [21] to maximizing a lower bound $\mathcal{L}(\psi, \theta; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}) \leq p_\psi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{c})$ to the intractable marginal likelihood, that is,

$$\theta^*, \psi^* = \underset{\theta,\psi}{\operatorname{argmax}} \, \mathcal{L}(\psi, \theta; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}).$$

In particular, this lower bound equals

$$\mathcal{L} = \int q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) \log \frac{p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c})}{q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})} d\boldsymbol{z} = \int q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) \log \frac{p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})p_\phi(\boldsymbol{z}|\boldsymbol{c})}{q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})} d\boldsymbol{z},$$

which can be rearranged to

$$\mathcal{L} = \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{c})}\left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right] + \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{c})}\left[\log p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})\right] - KL\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) \,||\, p_\phi(\boldsymbol{z}|\boldsymbol{c})\right].$$

For the Gaussian prior and approximate posterior , the KL-term can be computed analytically and efficiently [31]. On the other hand, the expectations $\mathbb{E}_{q_\theta}$ can be approximated with a few Monte-Carlo samples $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^s, \ldots, \boldsymbol{z}^S \sim q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ leading to

$$\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{c})}\left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})\right] \approx \frac{1}{S}\sum_{s=1}^{S} \log p_\pi(\boldsymbol{x}|\boldsymbol{z}^s, \boldsymbol{c})p_\gamma(\boldsymbol{y}|\boldsymbol{z}^s, \boldsymbol{c}).$$

B.1.1 Partially Observed Data

The measurements $\boldsymbol{x} \in \mathbb{R}^{D\times T}$ and the concepts $\boldsymbol{y} \in \mathbb{R}^{P\times T}$ contain many missing values. We define the indices $\boldsymbol{o}_x \in \mathbb{R}^{D\times T}$ and $\boldsymbol{o}_y \in \mathbb{R}^{P\times T}$ for which the observations are actually measured. Therefore, we compute the lower bound only on the observed variables, i.e. $\log p_\psi(\boldsymbol{x}^{\boldsymbol{o}_x}, \boldsymbol{y}^{\boldsymbol{o}_y}|\boldsymbol{c}) \geq \mathcal{L}(\psi, \theta; \boldsymbol{x}^{\boldsymbol{o}_x}, \boldsymbol{y}^{\boldsymbol{o}_y}, \boldsymbol{c})$, as is similarly done by Fortuin et al. [9], Ramchandran et al. [24]. This then leads for instance to

$$\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{c})}\left[\log p_\pi(\boldsymbol{x}^{\boldsymbol{o}_x}|\boldsymbol{z},\boldsymbol{c}) p_\gamma(\boldsymbol{y}^{\boldsymbol{o}_y}|\boldsymbol{z},\boldsymbol{c})\right],$$

where the related log-likelihood $\log p_\pi(\boldsymbol{x}^{\boldsymbol{o}_x}|\boldsymbol{z},\boldsymbol{c}) = \log \prod_{t,d\in\boldsymbol{o}_x} p_\pi(x_t^d|\boldsymbol{z}_t,\boldsymbol{c}_t) = \sum_{t,d\in\boldsymbol{o}_x} \log p_\pi(x_t^d|\boldsymbol{z}_t,\boldsymbol{c}_t)$ is only summed over the actually observed measurements. The same can be derived for the medical concepts $\boldsymbol{y}^{\boldsymbol{o}_y}$.

B.1.2 Lower Bound for N Samples

Given a dataset with N iid patients $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^N = \{\boldsymbol{x}_{1:T_i}^i, \boldsymbol{y}_{1:T_i}^i, \boldsymbol{c}_{1:T_i}^i\}_{i=1}^N$, the lower bound to the marginal log-likelihood is

$$\log p_\psi(\mathcal{D}) = \log \prod_{i=1}^N p_\psi(\mathcal{D}_i) \geq \sum_{i=1}^N \mathcal{L}(\psi, \theta; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{c}^i),$$

which is maximized through stochastic optimization with mini-batches. Moreover, suppose we have $T+1$ iid copies of the whole dataset $\{\mathcal{D}^k\}_{k=0}^T$, then

$$\log p_\psi(\{\mathcal{D}^k\}_{k=0}^T) = \log \prod_{i=1}^N \prod_{k=0}^T p_\psi(\mathcal{D}_i^k) \geq \sum_{i=1}^N \sum_{k=0}^T \mathcal{L}_k(\psi, \theta; \boldsymbol{x}^{i,k}, \boldsymbol{y}^{i,k}, \boldsymbol{c}^{i,k}),$$

where $\mathcal{L}_k(\psi, \theta; \boldsymbol{x}^{i,k}, \boldsymbol{y}^{i,k}, \boldsymbol{c}^{i,k})$ is the lower bound obtained by plugging in the corresponding approximate posterior $q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})$.

B.1.3 Predictive Distributions

The predictive distributions for the measurement $\boldsymbol{x}_{1:T}$ and concept trajectories $\boldsymbol{y}_{1:T}$ can be obtained via a two-stage Monte-Carlo approach. For instance, we can sample from the distribution of the measurements

$$q_*(\boldsymbol{x}_{1:T}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) = \int p_{\pi^*}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T}, \boldsymbol{c}) q_{\theta^*}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) d\boldsymbol{z}$$

by first sampling from the latent trajectories

$$\boldsymbol{z}_{1:T}^1, \ldots, \boldsymbol{z}_{1:T}^s, \ldots \boldsymbol{z}_{1:T}^S \sim q_{\theta^*}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{0:k}, \boldsymbol{c})$$

given the current observed measurements $\boldsymbol{x}_{1:k}$. In a second step, for each of the samples, we compute

$$\boldsymbol{x}_{1:T}^1, \ldots, \boldsymbol{x}_{1:T}^u, \ldots \boldsymbol{x}_{1:T}^U \sim p_{\pi^*}(\boldsymbol{x}_{1:T}|\boldsymbol{z}_{1:T}^s, \boldsymbol{c})$$

to represent the overall uncertainty of the measurement distribution.

B.2 Different Prior

The factorized prior can be extended to continuous time with Gaussian processes (GPs, [34, 27, 28]), as introduced by [5, 9] in the unsupervised setting. In particular, we can replace

$$p_\phi(\boldsymbol{z}|\boldsymbol{c}) = p_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{c}_{1:T}) = \prod_{t=1}^T \prod_{l=1}^L p_\phi(\boldsymbol{z}_t^l|\boldsymbol{c}_t) = \prod_{t=1}^T \prod_{l=1}^L \mathcal{N}\left(\boldsymbol{z}_t^l|\mu_\phi^l(\boldsymbol{c}_t), \sigma_\phi^l(\boldsymbol{c}_t)\right),$$

with

$$p_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{c}_{1:T}) = \prod_{l=1}^L \mathcal{GP}\left(\boldsymbol{z}^l|m_\phi^l(\boldsymbol{c}), k_\phi^l(\boldsymbol{c}, \boldsymbol{c}')\right)$$

with a mean function $m_\phi^l(\boldsymbol{c})$ and kernel $k_\phi^l(\boldsymbol{c}, \boldsymbol{c}')$, to take into account all the probabilistic correlations occurring in continuous time. This leads to a *stochastic* dynamic process, which theoretically matches the assumed disease process more adequately than a deterministic one. A further advantage is the incorporation of prior knowledge via the choice of the particular kernels for each latent process so that different characteristics such as long and small lengthscales, trends, or periodicity can be explicitly enforced in the latent space.

B.3 Conditional Generative Trajectory Generation

Our generative approach is also promising for conditional generative trajectory sampling, in a similar spirit as [29]. In particular, if we use medications as additional covariates $\boldsymbol{p} = \boldsymbol{p}_{1:T} = \{\boldsymbol{p}_{0:k}, \boldsymbol{p}_{k+1:T}\}$ in our approximate posterior distribution $q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) = q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{p}_{0:k}, \boldsymbol{p}_{k+1:T})$ with $\boldsymbol{c} = \{\boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{p}\}$, the model can be used to sample future hypothetical trajectories $\boldsymbol{x}_{k+1:T}$ with

$$q_*(\boldsymbol{x}_{k+1:T}|\boldsymbol{x}_{0:k}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{p}_{0:k}, \boldsymbol{p}_{k+1:T})$$
$$= \int p_{\pi^*}(\boldsymbol{x}_{k+1:T}|\boldsymbol{z}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{p}_{0:k}, \boldsymbol{p}_{k+1:T}) q_{\theta^*}(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{p}_{0:k}, \boldsymbol{p}_{k+1:T}) d\boldsymbol{z}$$

based on future query medications $\boldsymbol{p}_{k+1:T}$.

C Modeling approach

C.1 Analyzing Disease Trajectories with ML

Recently, extensive research has focused on modeling and analyzing clinical time series with machine learning – we refer to Allam et al. [1] for a recent overview. However, most approaches focus on deterministic time series forecasting, and only a few focus on interpretable representation learning with deep models [32] or on online uncertainty quantification with generative models [27, 7, 25].

Furthermore, prior research on data-driven analysis of systemic sclerosis is limited. In their recent review, Bonomi et al. [4] discuss the existing studies applying machine learning for precision medicine in systemic sclerosis. However, all of the listed studies are limited by the small cohort size (maximum of 250 patients), making the use of deep learning models challenging. Deep models were only used for analyzing imaging data (mainly nailfold capillaroscopy, Garaiman et al. [10]). Furthermore, most existing works solely focus on the involvement of a single organ in SSc, namely interstitial lung disease (ILD), and on forecasting methods. To the best of our knowledge, our work is the first attempt at such a comprehensive and large-scale (N=5673 patients) ML analysis of systemic sclerosis involving multiple organs and a wide range of observed clinical variables together with a systematic integration of medical knowledge.

C.2 Model Architecture

As shown in Figure 1b, our model combines several deep probabilistic networks. We implemented a temporal network with fully connected and LSTM layers [12] for $q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})$ and multilayer perceptrons for the prior $p_\phi(\boldsymbol{z}|\boldsymbol{c})$, guidance $p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})$ and likelihood $p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$ networks. By adapting our framework, we recover well-established temporal latent variable models. For instance, if we discard the guidance networks, the model becomes similar to a deterministic RNN-AE, or probabilistic RNN-VAE if we learn the latent space distribution. Furthermore, the likelihood variance can either be learned, or kept constant as is common practice [26]. We evaluated the predictive performance of the guided model in the probabilistic and deterministic settings, with or without learning the likelihood variance. Many further architectural choices could be explored, such as a temporal likelihood network or a GP prior, but they are beyond this paper's scope.

We describe the architecture and inputs/outputs of the different neural networks in our final model for SSc. For a patient with measurement time points $\boldsymbol{\tau}_{1:T}$ of the complete trajectory, the model input at time $t \in \boldsymbol{\tau}$ are the static variables $\boldsymbol{s}$, the clinical measurements $\boldsymbol{x}_{0:t}$, and the trajectory time points $\boldsymbol{\tau}$. Thus for SSc modeling, we have that $\boldsymbol{c} = \{\boldsymbol{\tau}, \boldsymbol{s}\}$. The model $\mathcal{M}$ outputs the distribution parameters of the clinical measurements and the organ labels for all trajectory time points $\boldsymbol{\tau}$. Without loss of generality, we assume that $\boldsymbol{x}^{1:M}$ are continuous variables and $\boldsymbol{x}^{M+1:D}$ categorical, so that the model can be described as

$$\mathcal{M} : (\boldsymbol{c}, \boldsymbol{x}_{0:t}) \longrightarrow \left(\hat{\boldsymbol{\mu}}_{1:T}^{x^{1:M}}(t), \hat{\boldsymbol{\sigma}}_{1:T}^{x^{1:M}}(t), \hat{\boldsymbol{\pi}}_{1:T}^{x^{M+1:D}}(t), \hat{\boldsymbol{\pi}}_{1:T}^{y}(t) \right).$$

We explicitly include the dependencies to t to emphasize that the parameters of the whole trajectory are estimated given the information up to time t.

- **Prior network**: The prior is a multilayer perceptron (MLP). It takes as input $\boldsymbol{c}$ and outputs the estimated mean and variance of the prior latent distribution $\hat{\mu}_{1:T}^{prior}$ and $\hat{\sigma}_{1:T}^{prior}$.

- **Encoder network** (posterior): The encoder contains LSTM layers followed by fully connected feed-forward layers. It takes as input $\boldsymbol{x}_{0:t}$ and $\boldsymbol{c}$ and outputs the estimated mean and standard deviation of the posterior distribution of the latent variables $\hat{\mu}^{post}_{1:T}(t)$ and $\hat{\sigma}^{post}_{1:T}(t)$, from which we sample the latent variables $\boldsymbol{z}_{1:T}(t)$ (complete temporal latent process) given the information up to t.
- **Decoder network** (likelihood): The decoder is an MLP and takes as input the sampled latent variables $\boldsymbol{z}_{1:T}(t)$ and $\boldsymbol{c}$ and outputs the estimated means and standard deviations $\hat{\mu}^{x^{1:M}}_{1:T}(t)$ and $\hat{\sigma}^{x^{1:M}}_{1:T}(t)$ of the distribution of the continuous clinical measurements and class probabilities $\hat{\pi}^{x^{M+1:D}}_{1:T}(t)$ of the categorical measurements.
- **Guidance networks**: For each organ, we define one MLP guidance network per related medical concept (involvement and stage). A guidance network for organ $o \in \mathcal{O} := \{lung, heart, joints\}$ and related medical concept $m \in \{inv, stage\}$, takes as input the sampled latent variables $\boldsymbol{z}^{\epsilon(o(m))}_{1:T}(t)$ and outputs the predicted class probabilities $\hat{\pi}^{y^{\nu(o(m))}}_{1:T}(t)$ of the labels, where $\nu(o(m))$ are the indices in y related to the medical concept $o(m)$, and $\epsilon(o(m))$ the indices in the latent space.

C.3 Training Objective

We follow the notation introduced in Section 2 and Appendix B. To train the model to perform forecasting, for each patient, we augment the data by assuming $T+1$ *iid* copies of the data x and y (see also B.1.2) and recursively try to predict the last $T-t$, $t = 0, ..., T$ clinical measurements and medical concepts. The total loss for a patient p is

$$\mathcal{L}_p = \sum_{t=0}^{T} \mathcal{L}(t), \tag{1}$$

with

$$\begin{aligned}\mathcal{L}(t) = NLL\left(\hat{\mu}^{x^{1:M}}(t), \hat{\sigma}^{x^{1:M}}(t), \boldsymbol{x}^{1:M}\right) + CE\left(\hat{\pi}^{x^{M+1:D}}(t), \boldsymbol{x}^{M+1:D}\right) \\ + \alpha * CE\left(\hat{\pi}^{y}(t), \boldsymbol{y}\right) + \beta * KL\left(\hat{\mu}^{prior}, \hat{\sigma}^{prior}, \hat{\mu}^{post}(t), \hat{\sigma}^{post}(t)\right),\end{aligned}$$

where NLL, CE and KL are the negative log-likelihood, cross-entropy and KL divergence, respectively. Further, α and β are hyperparameters weighting the guidance and KL terms.

C.3.1 Model Optimization

We only computed the loss with respect to the available measurements. We randomly split the set of patients $\mathcal{P}$ into a train set $\mathcal{P}_{train}$ and test set $\mathcal{P}_{test}$ and performed $5-$fold CV with random search on $\mathcal{P}_{train}$ for hyperparameter tuning. Following the principle of empirical risk minimization, we trained our model to minimize the objective loss over $\mathcal{P}_{train}$, using the Adam [15] optimizer with mini-batch processing and early stopping.

C.3.2 Architecture and Hyperparameters

We tuned the dropout rate and the number and size of hidden layers using 5-fold CV, and used a simple architecture for our final model. The posterior network contains a single lstm layer with hidden state of size 100, followed by two fully connected layers of size 100. The likelihood network contains two separate fully connected layers of size 100, learning the mean and variances of the distributions separately. The guidance networks contain a single fully connected layer of size 40 and the prior network a single fully connected layer of size 50. We used batch normalization, ReLU activations, and a dropout rate of 0.1. We set $\alpha = 0.2$ and $\beta = 0.01$.

D Results

D.1 Model Evaluation

We discuss the evaluation results for unguided models, medical concept prediction, and uncertainty quantification. In Figure 5, we compare the performance of the clinical measurement $\boldsymbol{x}$ prediction of the different guided models versus their unguided counterparts (with the same number of latent processes). Note that these unguided models are optimal baselines for $\boldsymbol{x}$ prediction since they are not trained to predict $\boldsymbol{y}$, too. As Figure 5 shows, the unguided models usually outperform the guided models, but the difference is not significant for the probabilistic models. Unsurprisingly, the best performing model is a deterministic unguided model, i.e. not trained to learn the $\boldsymbol{z}$ and $\boldsymbol{y}$ distributions.

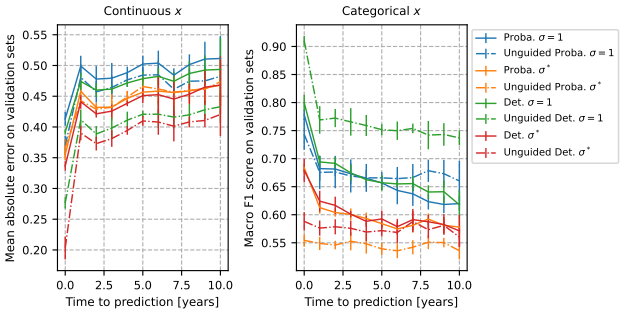


Figure 5: Performance for x prediction, guided versus unguided models.

Figure 6 shows the macro F_1 scores for the medical concepts $\boldsymbol{y}$ prediction of the different models. The models with fixed likelihood variance generally slightly outperform the models with learned variance. All of the models outperform the individualized and cohort baselines.

To evaluate the uncertainty quantification of the models, we computed the coverage of the continuous predictions and calibration of the predicted probabilities for categorical measurements. The coverage is the probability that the confidence interval (CI) predicted by the model contains the true data point. Since the likelihood distribution is Gaussian, the 95% CI is $\mu_{pred} \pm 1.96\sigma_{pred}$. To achieve perfect coverage of the 95% CI, the predictions should fall within the predicted CI 95% of the time. We computed the coverage over all forecasted data points. For continuous $\boldsymbol{x}$ forecasting, both probabilistic models achieve coverage of $92 \pm 1\%$ and of $98 \pm 0\%$ for the deterministic models, thus all slightly diverging from the optimal 95%. For categorical measurements, the calibration curve is computed to assess the reliability of the predicted class probabilities. They are computed in the following way. We grouped all of the forecasted probabilities (for one-hot encoded vectors) into $n = 20$ bins dividing the 0-1 interval. Then, for each bin, we compared the observed frequency of ground truth positives (aka “fraction of positive") with the average predicted probability within the bin. Ideally, these two quantities should be as close as possible, i.e. close to the line of “perfect calibration" in Figure 7a and Figure 7b. The calibration curves in Figure 7a and Figure 7b show that all of the models are well calibrated both in their categorical $\boldsymbol{x}$ and medical concept $\boldsymbol{y}$ forecasts (averaged over all forecasted data points in the respective validation sets).

D.1.1 Online Prediction with Uncertainty

We provide additional online prediction results for the index patient p_{idx}. Figures 8a and 8b show the evolution in the predicted mean and 95% CI of the Forced Vital Capacity (FVC)[1] and DLCO(SB)[2] for p_{idx}. The values after the dashed line are forecasted. As more prior information becomes available to the model, the forecast becomes more accurate and the CI shrinks.

Figure 8c shows predicted probabilities of organ stages at a given time point. The intensity of the heatmap reflects the predicted probability.

[1]FVC is the amount of air that can be exhaled from the lungs.

[2]DLCO(SB) stands for single breath (SB) diffusing capacity of carbon monoxide (DLCO).

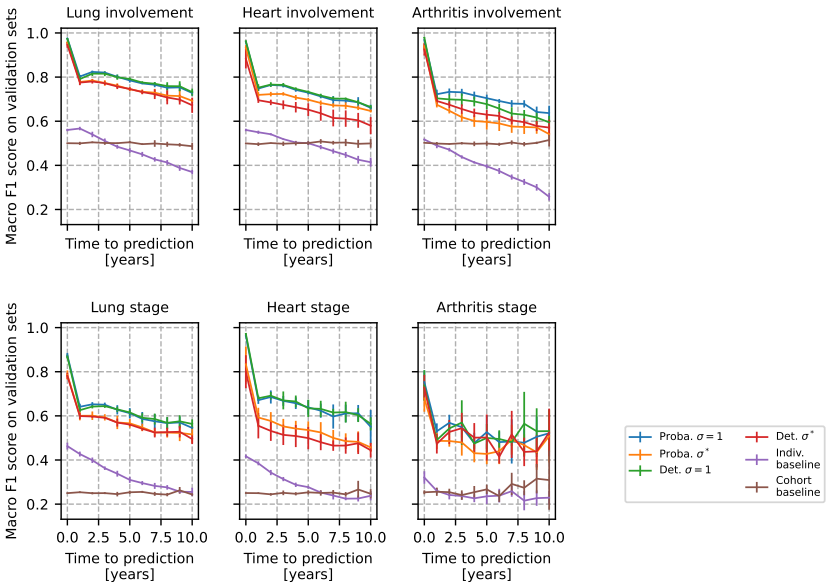
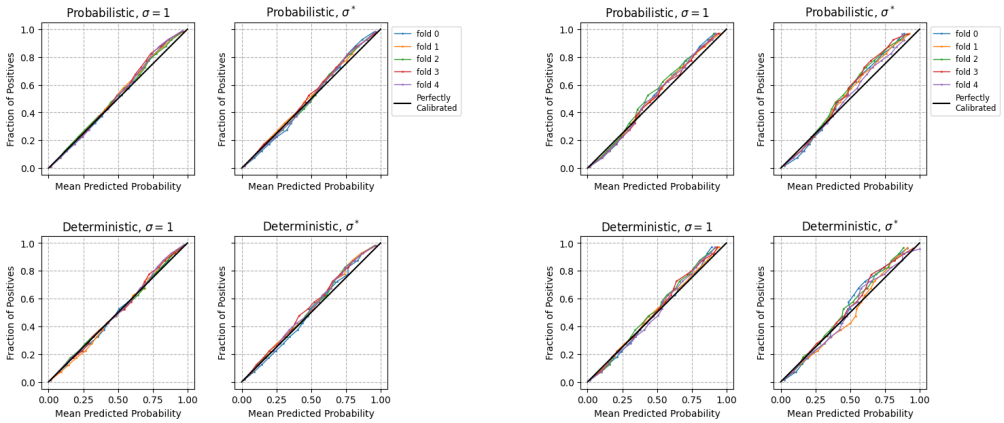


Figure 6: Performance for y prediction.



(a) Categorical clinical measurements x

(b) Medical concepts y

Figure 7: Calibration curves

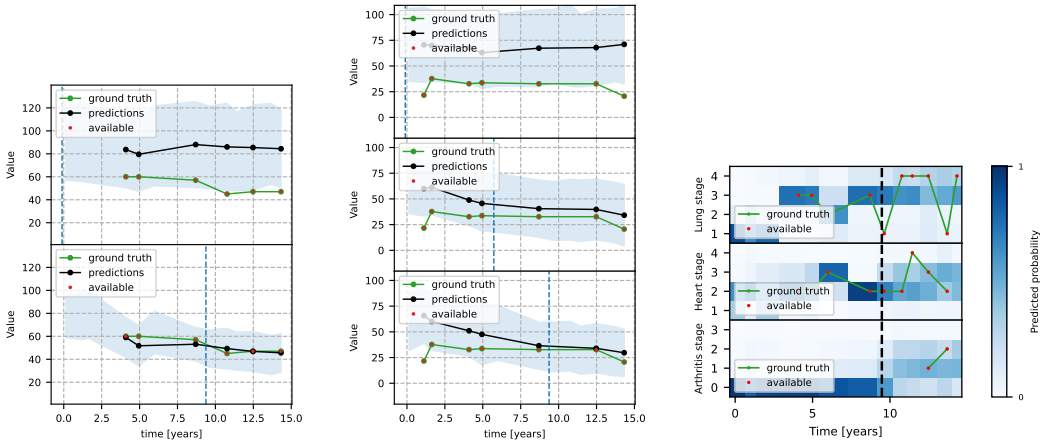
D.2 Cohort Analysis

We present here additional experiments to gain insights into cohort patterns.

D.2.1 Prior z Distributions

By learning $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\tau})$, we estimate the average prior disease trajectories in the cohort. This allows the comparison of trajectories, conditioned only on the simple subset of variables $\boldsymbol{s}$ and $\boldsymbol{\tau}$ and thus without facing any confounding in the trajectories, for instance, due to past clinical measurements $\boldsymbol{x}$. For example, in Figure Figure 9a we overlaid the predicted prior trajectories of Forced Vital Capacity (FVC)[3] for a subset of patients in $\mathcal{P}_{test}$ with a static variable corresponding to the SSc subtype. Overall, the FVC values are predicted to remain quite stable over time, but with different average values depending on the SSc subtype. In Figure Figure 9b, the prior predicted N-terminal pro

[3]FVC is the amount of air that can be exhaled from the lungs. Low levels indicate lung malfunction.

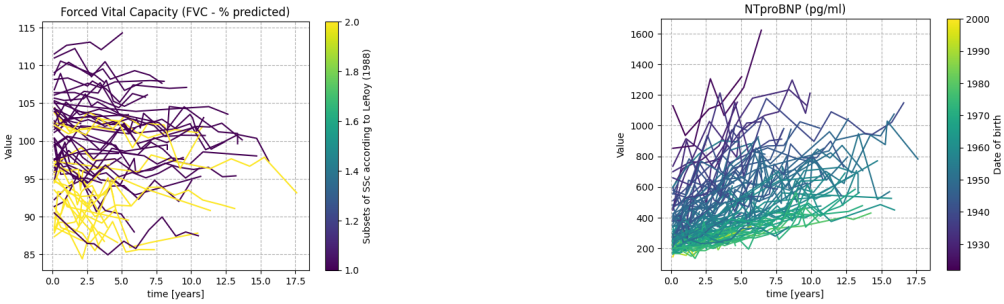


(a) FVC of p_{idx}: predicted mean and 95% CI

(b) DLCO(SB) of p_{idx}: predicted mean and 95% CI

(c) Probabilities of organ stages for p_{idx}.

Figure 8: Online prediction of clinical variables and medical concepts.



(a) Prior FVC trajectories overlaid with different SSC subsets.

(b) Prior predicted trajectories conditioned on static variables.

Figure 9: Combined figure with two subfigures.

b-type natriuretic peptide (NTproBNP)[4] trajectories overlaid with age, show that the model predicts an overall increase in NTproBNP over time, and steeper for older patients.

D.2.2 Latent Space and Medical Concepts

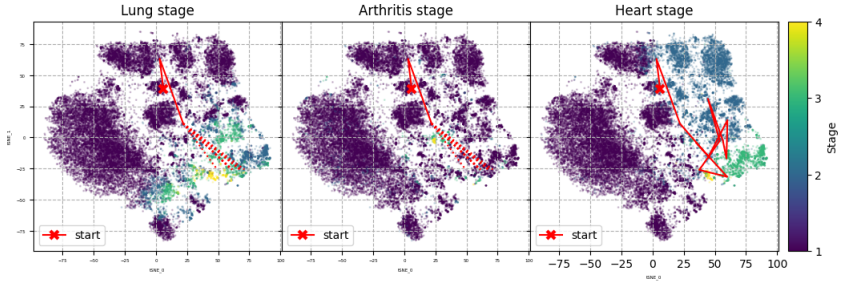


Figure 10: Predicted organ stages in the latent space. The red line highlights the trajectory of p_{idx}.

t−SNEs: The t-SNE [33] graphs were obtained by computing the two-dimensional t-SNE projection of the latent variables $\boldsymbol{z}_{1:T} \mid (\boldsymbol{x}_{1:T}, \boldsymbol{c})$ (i.e. only using reconstructed $\boldsymbol{z}$) of a subset of

[4]They are substances produced by the heart. High levels indicate potential heart failure.

$\mathcal{P}_{train}$ and then transforming and plotting the projected latent variables (reconstructed or forecasted) from patients in $\mathcal{P}_{test}$ [23].

In Figure 3b, we showed the trajectory of p_{idx} overlaid with the predicted organ involvement probabilities. In Figure 10, we additionally show the trajectory overlaid with the organ stages, showing for instance in the first panel that the model predicts an increase in the lung stage and in the last panel that p_{idx} undergoes many different heart stages throughout the disease course.

D.2.3 Clustering of Patient Trajectories and Trajectory Similarity

We discuss additional results obtained through clustering and similarity analysis of latent trajectories (subsubsection 3.2.2). In Figure 12, we show the different predicted probabilities of the medical concepts $\boldsymbol{y}$ for the mean trajectories within the three found clusters. This reveals which medical concepts are most differentiated by the clustering algorithm. For instance, cluster one exhibits low probabilities of organ involvement, while cluster two shows increasing probabilities of heart involvement and low probabilities of lung involvement. In contrast, cluster three shows increasing probabilities for both heart and lung involvement.

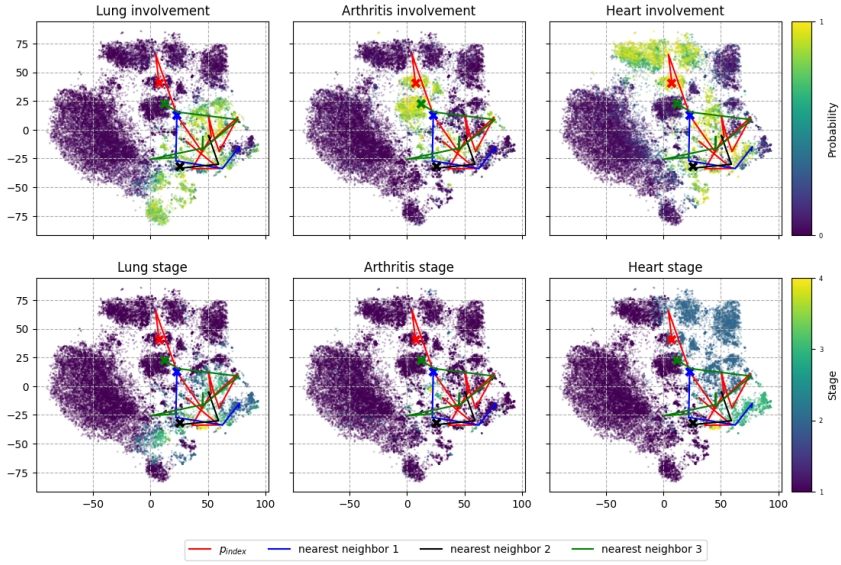


Figure 11: Trajectory of p_{idx} and its 3 nearest neighbors in the latent space.

Additionally, we apply a *k-nn* algorithm with the *dtw* distance in the latent space to find patients with similar trajectories to p_{idx}. Figure 11 shows the trajectory of p_{idx} and its three nearest neighbors in the latent space. We can see that the nearest neighbors also have an evolving disease, going through various organ involvements and stages. Similarly, in Figure 13, the medical concept trajectories of p_{idx} and its nearest neighbors reveal consistent patterns.

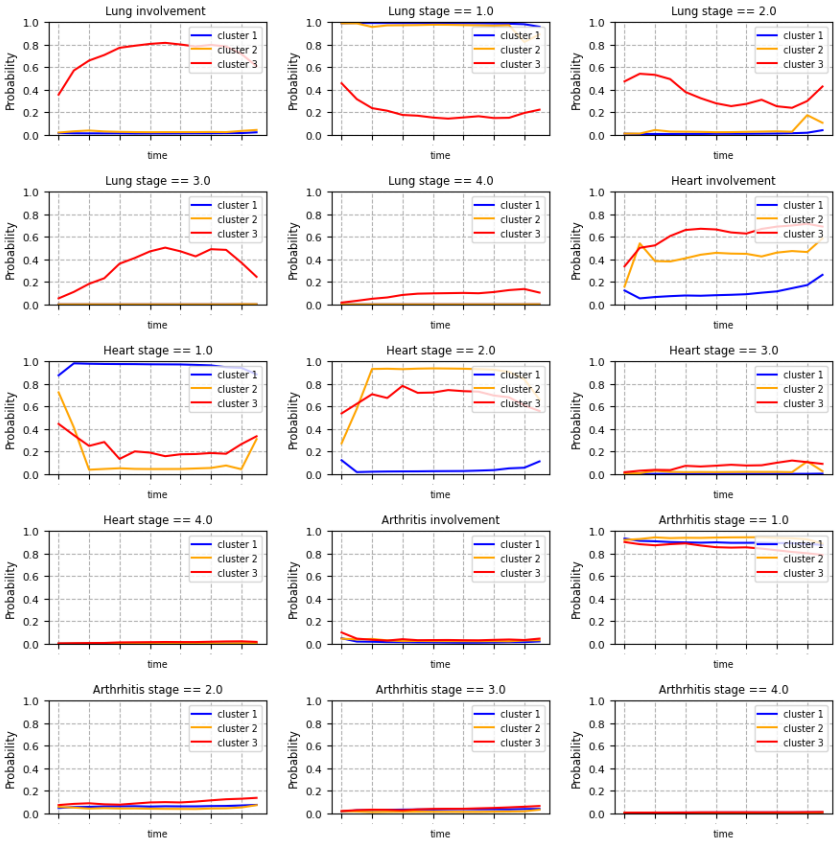


Figure 12: Medical concept trajectories for cluster means.

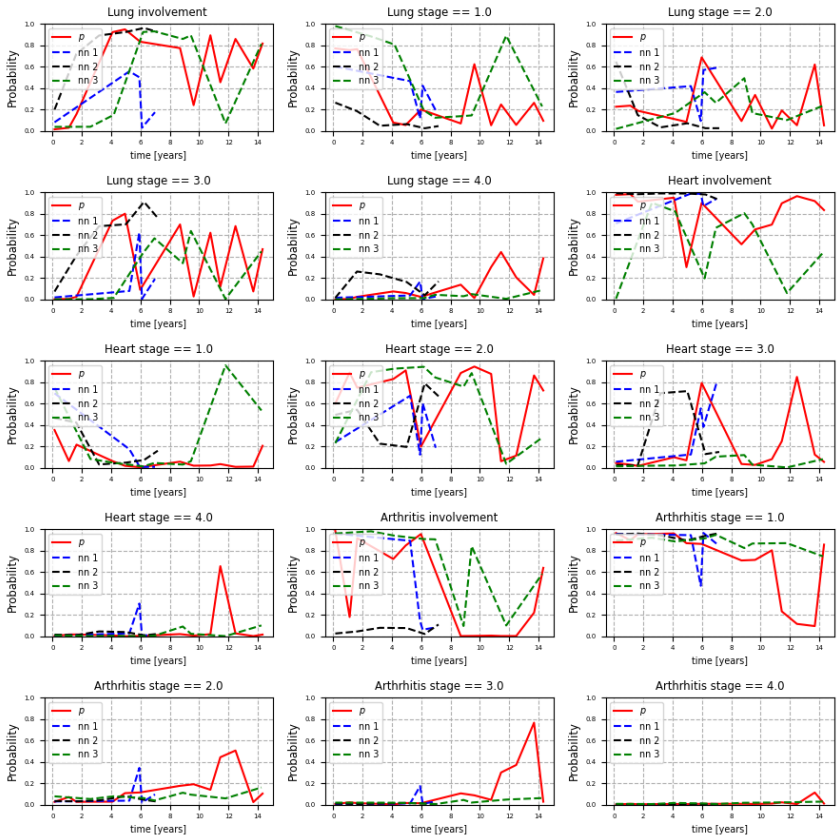


Figure 13: Medical concept trajectories for p_{idx} and its 3 nearest neighbors.

E Clinical Insights for Systemic Sclerosis

In this paper, we present a general approach for modeling and analyzing complex disease trajectories, for which we used the progression of systemic sclerosis as an example. The focus of this paper is on the machine learning methodology, while clinically relevant insights and data analysis regarding systemic sclerosis will be discussed in a clinical follow-up paper where our model will be applied to investigate the involvement of multiple organs.

Since there is ongoing research and discussion towards finding optimal definitions of the medical concepts (involvement, stage, progression) for all impacted organs in SSc, we used preliminary definitions for three organs as a proof of concept.

E.1 Dataset

The European Scleroderma Trials and Research group (EUSTAR) maintains a registry dataset of about 20'000 patients extensively documenting organ involvement in SSc. It contains around 30 demographic variables, and 500 temporal clinical measurement variables documenting the patients' overall and organ-specific disease evolution. For a detailed description of the database, we refer the reader to Meier et al. [19], Hoffmann-Vold et al. [13].

For our analysis, we included 5673 patients with enough temporality (i.e. at least 5 medical visits). We used 10 static variables related to the patients' demographics and 34 clinical measurement variables, mainly related to the monitoring of the lung, heart, and joints in SSc. In future work, we plan to include more patients and more clinical measurements for analyzing all involved organs.

E.2 Medical Concepts Definitions

Defining the organ involvement and stages in SSc is a challenging task as varying and sometimes contradicting definitions are used in different studies. However, there is ongoing research to find the most accurate definitions. Since this work is meant as a proof of concept, we used the following preliminary definitions of involvement and stage for the lung, heart, and joints (arthritis). The medical concepts are defined for the variables of the EUSTAR database. There are 4 stages of increasing severity for each organ. If multiple definitions are satisfied, the most severe stage is selected. Furthermore, there is missingness in the labels due to incomplete clinical measurements. Our modeling approach thus also could be used to label the medical concepts when missing.

We use the following abbreviations:

- Interstitial Lung Disease: ILD
- High-resolution computed tomography: HRCT
- Forced Vital Capacity: FVC
- Left Ventricular Ejection Fraction: LVEF
- Brain Natriuretic Peptide: BNP
- N-terminal pro b-type natriuretic peptide: NTproBNP
- Disease Activity Score 28: DAS28

E.2.1 Lung

Involvement At least one of the following must be present:

- ILD on HRCT
- FVC $< 70\%$

Severity staging

1. FVC $> 80\%$ or Dyspnea stage of 2
2. ILD extent $< 20\%$ or $70\% <$ FVC $\leq 80\%$ or Dyspnea stage of 3
3. ILD extent $> 20\%$ or $50\% \leq$ FVC $\leq 70\%$ or Dyspnea stage of 4
4. FVC$< 50\%$ or Lung transplant or Dyspnea stage of 4

E.2.2 Heart

Involvement At least one of the following must be present:

- LVEF $< 45\%$
- Worsening of cardiopulmonary manifestations within the last month
- Abnormal diastolic function
- Ventricular arrhythmias
- Pericardial effusion on echocardiography
- Conduction blocks
- BNP > 35 pg/mL
- NTproBNP> 125 pg/mL

Severity staging

1. Dyspnea stage of 1
2. Dyspnea stage of 2
3. Dyspnea stage of 3
4. Dyspnea stage of 4

E.2.3 Arthritis

Involvement At least one of the following must be present:

- Joint synovitis
- Tendon friction rubs

Severity staging

1. DAS28 < 2.7
2. $2.7 \leq$ DAS28 ≤ 3.2
3. $3.2 <$ DAS28 ≤ 5.1
4. DAS28 > 5.1