Semi-Supervised Generative Models for Disease Trajectories: A Case Study on Systemic Sclerosis

Cécile Trottet* CECILECLAIRE.TROTTET@UZH.CH
Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

Manuel Schürch* MANUEL.SCHUERCH@UZH.CH
Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

Ahmed Allam
Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

Imon Barua
Department of Rheumatology, Oslo University Hospital, University of Oslo, Oslo, Norway

Liubov Petelytska
Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland, Department of Internal Medicine #3, Bogomolets National Medical University, Kyiv, Ukraine

David Launay
Hôpital Huriez, CHU Lille, Lille University, Lille, France

Paolo Airò
Rheumatology and Clinical Immunology Unit, ASST Spedali Civili of Brescia, University of Brescia, Brescia, Italy

Radim Bečvář
Institute of Rheumatology, Department of Rheumatology, 1st Medical School, Charles University, Prague, Czech Republic

Christopher Denton
Centre for Rheumatology Royal Free, University College London Medical School, London, United Kingdom

Mislav Radic
Division of Rheumatology and Clinical Immunology, Department of Internal Medicine, University of Split, School of Medicine, University Hospital Center Split, Split, Croatia

Oliver Distler
Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

Anna-Maria Hoffmann-Vold
Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland, Department of Rheumatology, Oslo University Hospital, Oslo, Norway

Michael Krauthammer
Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

the EUSTAR collaborators

* These authors contributed equally.

Abstract

We propose a deep generative approach using latent temporal processes for modeling and holistically analyzing complex disease trajectories, with a particular focus on Systemic Sclerosis (SSc). We aim to learn temporal latent representations of the underlying generative process that explain the observed patient 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 knowledge. By combining the generative approach with medical definitions of different characteristics of SSc, we facilitate the discovery of new aspects of the disease.

We show that the learned temporal latent processes can be utilized for further data analysis and clinical hypothesis testing, including finding similar patients and clustering SSc patient trajectories into novel sub-types. Moreover, our method enables personalized online monitoring and prediction of multivariate time series with uncertainty quantification.

1. Introduction

Understanding and analyzing clinical trajectories of complex diseases, such as Systemic Sclerosis (SSc), is crucial for improving diagnosis, treatment, and patient outcomes (Allam et al., 2021). 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 influencing the disease progression (Allam et al., 2021). Therefore, our primary goal is to develop a machine learning (ML) model suited for the holistic analysis of temporal disease trajectories. Moreover, we aim to uncover meaningful temporal latent representations capturing the complex interactions within the raw data while also providing interpretable insights, and potentially revealing novel medical aspects of clinical disease trajectories. To achieve these goals, we present a deep generative temporal model that captures both the joint distribution of all observed longitudinal clinical variables and latent temporal variables (Figure 1).

Since inferring interpretable temporal representations in a fully unsupervised way is very challenging (Locatello et al., 2020a), we propose a semi-supervised approach for disentangling the latent space using known medical knowledge to enhance the interpretability. Combining an unsupervised latent generative model with known medical labels facilitates the discovery of novel medically-driven patterns in the data.

Deep probabilistic generative models (Tomczak (2022)) provide a more holistic approach to modeling complex data than deterministic discriminative models. By learning the joint distribution over all observed variables, they model the underlying data-generating mechanism. In contrast, discriminative models only learn the conditional distribution of the target variable given the input variables.

While our method is general and can be applied to a wide range of high-dimensional clinical datasets, in this paper, we demonstrate its effectiveness 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 tissues, causing severe damage to the skin and multiple other internal organs. We seek to understand the evolution of SSc by modeling the patterns of organ involvement and progression. In doing so, we aim

to learn temporal hidden representations that distinctly capture the disentangled medical disease processes related to each organ.

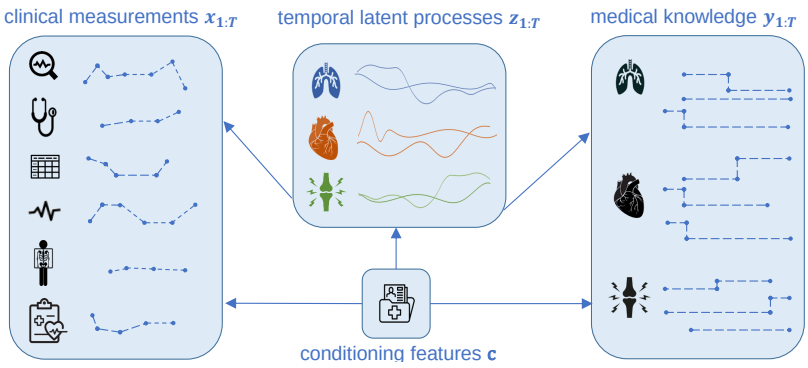


Figure 1: Temporal generative model for systemic sclerosis. The latent temporal process $\boldsymbol{z}$ generates the observed $\boldsymbol{x}$ and $\boldsymbol{y}$ trajectories conditioned on data $\boldsymbol{c}$.

Our approach offers several contributions:

- **Interpretable Temporal Latent Processes:** Our generative model allows the non-linear projection of patient trajectories onto lower-dimensional temporal latent processes, providing useful representations for the visualization and understanding of complex medical time series data.
- **Semi-Supervised Guided Latent Processes:** To achieve more interpretable latent temporal spaces, we propose a semi-supervised approach for disentangling the latent space with respect to medical knowledge. By combining the generative approach with medical domain knowledge, new aspects of the disease can be discovered.
- **Online Prediction with Uncertainty Quantification:** Our deep generative probabilistic model facilitates personalized online monitoring and reliable predictions of multivariate time series data with uncertainty quantification.
- **Facilitating Clinical Hypothesis Testing:** The learned temporal latent processes can be inspected for further data analysis and clinical hypothesis testing, such as finding similar patients and clustering the disease trajectories into new sub-types.
- **Large-Scale Analysis of Systemic Sclerosis :** We demonstrate the potential of our ML model for comprehensively analyzing SSc for the first time at a large scale including multiple organs and various observed clinical variables.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work offers contributions at the intersection of machine learning methodology and clinical practice, by proposing a new deep generative approach to model patient disease trajectories and conducting a large-scale ML analysis of organ involvement in SSc. We

propose an approach to augment deep unsupervised generative models with medical knowledge, resulting in models with more interpretable and disentangled latent processes, suited for further downstream tasks like clustering. Our work also highlights the challenges of developing a holistic end-to-end approach to process high-dimensional, sparse, and temporal clinical data. Moreover, we demonstrate that our approach facilitates meaningful phenotyping of SSc and offers a novel methodology for understanding the disease's progression across multiple organs, setting a foundational baseline for future machine learning research on organ-specific disease modeling in SSc. While our expertise is suited to the modeling of SSc, we are confident that enriching deep generative models with targeted clinical knowledge holds potential for interdisciplinary collaborations between ML researchers and clinical experts to study various complex chronic diseases.

2. Related Work

2.1. Generative Latent Variable Models

Learning latent representations from raw data has a long tradition in statistics and ML with foundational research such as principal component analysis (Hotelling, 1933), factor analysis (Lawley and Maxwell, 1962) or independent component analysis (Comon, 1994), which all can be used to project high-dimensional tabular data to a latent space. For temporal data, models with latent processes such as hidden Markov models (Baum and Petrie, 1966) and Gaussian processes (Williams and Rasmussen, 2006) have extensively been used for discrete and continuous time applications, respectively. Conceptually, all these models can be viewed as probabilistic generative models with latent variables (e.g. Murphy (2022)), however, these models only learn linear or simple relationships between the input data and the latent space.

In their seminal work on Variational Autoencoders (VAEs), Kingma and Welling (2013) proposed a powerful generalization for latent generative models. The key idea is to use deep neural networks as function approximators to learn the moments of the data distribution, enabling the representation of arbitrarily complex distributions. The parameters of the neural networks are inferred using amortized variational inference (VI) (Blei et al., 2017), a powerful Bayesian inference method for approximating intractable probability distributions. There are various successors building and improving on the original model, for instance, conditional VAE (Sohn et al., 2015), LVAE (Sønderby et al., 2016), or VQ-VAE (Van Den Oord et al., 2017). Moreover, there are also several extensions that explicitly model time in the latent space such as RNN-VAE (Chung et al., 2015), GP-VAE (Casale et al., 2018; Fortuin et al., 2020), or longitudinal VAE (Ramchandran et al., 2021).

While these approaches have showcased remarkable efficacy in generating diverse objects such as images or modeling time series, the interpretability of the resulting latent spaces or processes remains limited for complex data. Moreover, the true underlying distributions of known processes often cannot be recovered, and instead become *entangled* within a single latent factor (Bengio et al., 2013). Thus, there is ongoing research in designing generative models with disentangled latent factors, such as $\beta-$VAE (Higgins et al., 2016), factorVAE (Kim and Mnih, 2018), TCVAE (Chen et al., 2018) or temporal versions including disentangled sequential VAE (Hsu et al., 2017) and disentangled GP-VAE (Bing et al., 2021).

However, learning interpretable and disentangled latent representations is highly difficult or even impossible for complex data without any inductive bias (Locatello et al., 2020a). Hence, purely unsupervised modeling falls short, leading researchers to focus on weakly supervised latent representation learning instead (Locatello et al., 2020b; Zhu et al., 2022; Palumbo et al., 2023). In a similar spirit, we tackle the *temporal* semi-supervised guidance of the latent space by using sparse labels representing established medical domain knowledge. We model the progression of complex diseases in an unsupervised way using the raw temporal clinical measurements, while augmenting the model with temporal medical labels.

2.2. 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. (2021) for an overview. However, most approaches focus on deterministic time series forecasting, and only a few focus on interpretable representation learning with deep models (Trottet et al., 2023) and irregularly sampled times (Chen et al., 2023) or on online uncertainty quantification with generative models (Schürch et al., 2020; Cheng et al., 2020; Rosnati and Fortuin, 2021).

A few approaches aim at uncovering disease stages from electronic health records in a fully unsupervised way (Yang et al., 2014; Wang et al., 2014; Alaa and van der Schaar, 2019) or with a self-supervised approach (Raghu et al., 2023). However, as motivated in the previous section, and by Chen et al. (2021), we rather develop a semi-supervised approach to model latent disease stages using sparse medical labels.

Recent approaches for clustering time series (Lee and Van Der Schaar, 2020; Srivastava and Rajan, 2023; Qin et al., 2023) focus on learning predictive embeddings for future events. However, these techniques are not designed for semi-supervised environments with high-dimensional, multi-labeled, and sparse temporal data. While Noroozizadeh et al. (2023) and Holland et al. (2023) leverage contrastive learning to cluster time series, we rather adopt a generative approach to fully model the complete patient trajectory.

Furthermore, prior research on data-driven analysis of systemic sclerosis is limited. In their recent review, Bonomi et al. (2022) 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 related to nailfold capillaroscopy (Garaiman et al., 2022). Furthermore, most existing works solely focus on the involvement of a single organ in SSc, namely interstitial lung disease (ILD), and on forecasting methods (Bonomi et al., 2022). To the best of our knowledge, our work is the first attempt at 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 the latest medical knowledge.

3. Methods

We analyze patient medical 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 knowledge labels $\boldsymbol{y} = \boldsymbol{y}_{1:T} \in \mathbb{R}^{P \times T}$, describing the medical definitions of selected aspects

of the disease, such as the medical definition of severity staging of the heart involvement in SSc (Figure 1). The medical knowledge definitions (Appendix A.2) are typically derived from multiple clinical measurements using logical operations. For example, a patient may be classified as having "lung involvement" if certain conditions are satisfied, for instance, $\boldsymbol{x}^{(i)} > \varepsilon$ OR $\boldsymbol{x}^{(j)} = 1$. Both the raw measurements and labels are irregularly sampled, and we denote by $\boldsymbol{\tau}_{1:T} \in \mathbb{R}^T$ the vector of observation time-points of $\boldsymbol{x}$ and $\boldsymbol{y}$. Moreover, there is non-temporal information denoted by $\boldsymbol{s} \in \mathbb{R}^S$ such as patient demographics, 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 take into account the heterogeneous patient preconditions. Furthermore, in the next sections, we introduce our approach to learning unobserved multivariate latent processes denoted as $\boldsymbol{z} = \boldsymbol{z}_{1:T} \in \mathbb{R}^{L \times T}$, responsible for generating both the raw clinical measurements $\boldsymbol{x}_{1:T}$ and the medical labels $\boldsymbol{y}_{1:T}$. Specifically, we use the different temporal medical labels to disentangle the L dimensions of the latent processes by allocating distinct dimensions to represent different medical knowledge labels.

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. Note that the measurements and medical labels are often partially observed, see more details in Appendix B.1.1. A table of the main introduced symbols is provided in Table 2 in the appendix.

3.1. Generative Model

We propose the probabilistic conditional generative latent variable model

$$p_\psi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{c}) = p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) p_\phi(\boldsymbol{z}|\boldsymbol{c}),$$

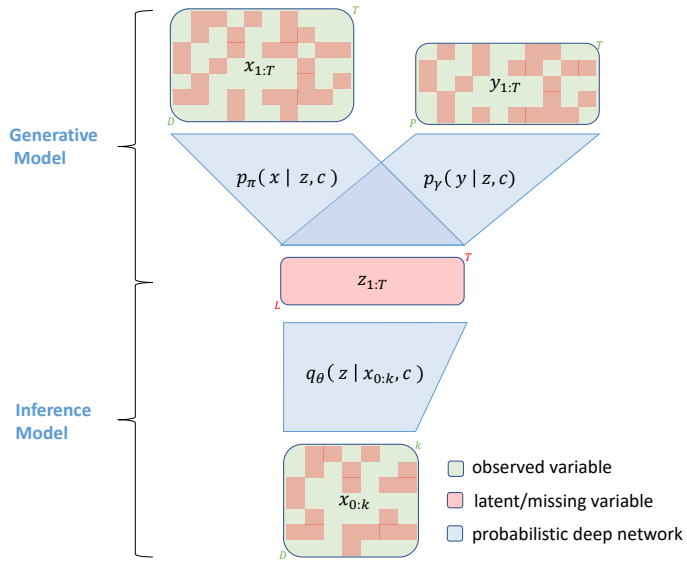
with learnable prior network $p_\phi(\boldsymbol{z}|\boldsymbol{c})$, measurement likelihood network $p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, and guidance networks $p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})$, where $\psi = \{\gamma, \pi, \phi\}$ are learnable parameters (2(a)subfigure). We assume conditional independence of $\boldsymbol{x}$ and $\boldsymbol{y}$ given $\boldsymbol{z}$ and $\boldsymbol{c}$. Although the measurements and the medical labels are conditionally independent, the marginal distribution $p_\psi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{c}) = \int p_\psi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}|\boldsymbol{c}) d\boldsymbol{z}$ allows arbitrarily rich correlations among the observed variables. For the sake of brevity, we do not include the time index explicitly.

3.2. Prior of Latent Process

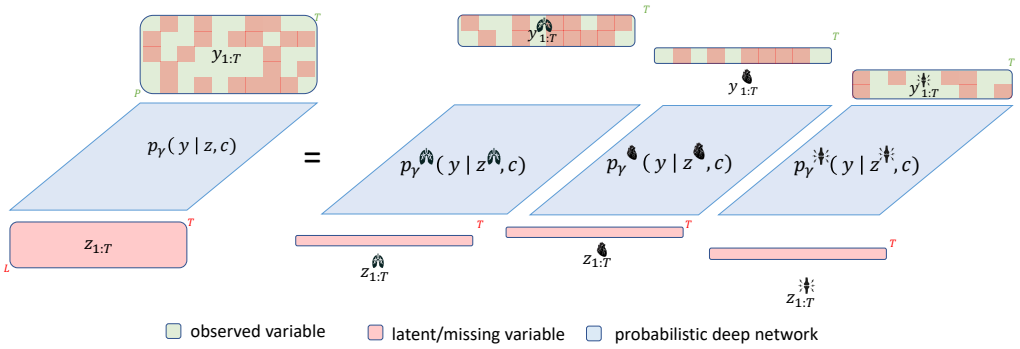
We use a learnable prior network for the latent temporal variables $\boldsymbol{z} \in \mathbb{R}^{L \times T}$, that is,

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

conditioned on the context variables $\boldsymbol{c} = \{\boldsymbol{\tau}, \boldsymbol{p}, \boldsymbol{s}\}$, 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 parametrized by deep neural networks. We assume a factorized Gaussian prior distribution per time and latent dimensions, however, many interesting extensions including continuous-time priors are straightforward (Appendix B.2).



(*a*) Model architecture.



(*b*) Guidance networks.

Figure 2: Semi-supervised temporal latent variable model. The left panel shows the model architecture with the inference and generative components, and the right panel describes the guidance networks. We have independent guidance networks for each medical label, taking as input a subset of the latent dimensions and predicting the corresponding medical label.

3.3. Likelihood of Measurements

The probabilistic likelihood network maps the latent temporal processes $\boldsymbol{z} \in \mathbb{R}^{L \times T}$ together with the context variables $\boldsymbol{c}$ to the clinical measurements $\boldsymbol{x} \in \mathbb{R}^{D \times T}$, i.e. we assume the following factorization

$$p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) = \prod_{t=1}^{T} \prod_{d \in \mathcal{G}} \mathcal{N}(x_t^d | \mu_\pi^d, \sigma_\pi^d) \prod_{d \in \mathcal{K}} \mathcal{C}(x_t^d | p_\pi^d),$$

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 moments of these distributions are parametrized by deep neural networks, i.e. 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 $p_\pi^d = p_\pi^d(\boldsymbol{z}_t, \boldsymbol{c}_t)$. Although the likelihood is a parametric distribution, the posterior distribution can be arbitrarily complex after marginalizing out the latent process $\boldsymbol{z}$.

3.4. Semi-Supervised Guidance Network

We propose a semi-supervised approach to disentangle the latent process $\boldsymbol{z}$ with respect to defined medical labels $\boldsymbol{y} = \boldsymbol{y}_{1:T} \in \mathbb{R}^{P \times T}$. In particular, we assume

$$p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) = \prod_{t=1}^{T} \prod_{g=1}^{G} \mathcal{C}(y_t^{\nu(g)} | h_\gamma^{\nu(g)}(\boldsymbol{z}_t^{\varepsilon(g)}, \boldsymbol{c}_t)),$$

where $|G|$ is the number of different medical labels. We assume categorical distributions for all medical labels, but the extension to continuous labels is straightforward. $h_\gamma^{\nu(g)}(\boldsymbol{z}_t^{\varepsilon(g)}, \boldsymbol{c}_t)$ is a deep parametrized category probability matrix, and $\nu(g)$ and $\varepsilon(g)$ correspond to the indices of the gth guided medical label, and the indices in the latent space defined for guided label g, respectively (Figure 2(*b*)subfigure).

3.5. Posterior of Latent Process

We are mainly interested in the posterior distribution $p_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c})$ of the latent process given the observations, which we approximate with an amortized variational distribution (Section 3.6, Appendix B.1)

$$q_\theta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c}) \approx p_\psi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}).$$

We use the amortized variational distribution

$$q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) = \prod_{t=1}^{T} \prod_{l=1}^{L} \mathcal{N}(z_t^l | \mu_\theta^l(\boldsymbol{x}_{0:k}, \boldsymbol{c}), \sigma_\theta^l(\boldsymbol{x}_{0:k}, \boldsymbol{c}))$$

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 labels $\boldsymbol{y}$. If $k = T$, there is no forecasting, whereas for $0 \leq k < T$, we can also forecast the future latent variables $\boldsymbol{z}_{k+1:T}$ from the first measurements $\boldsymbol{x}_{0:k}$.

3.6. Probabilistic Inference

Since exact inference with the marginal likelihood $p_\psi(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{c}) = \int p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) p_\phi(\boldsymbol{z}|\boldsymbol{c}) d\boldsymbol{z}$ is not feasible (Appendix B.1), we apply amortized variational inference (Blei et al., 2017) 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 following objective function

$$\begin{aligned} \mathcal{L}_k(\psi, \theta; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}) = & \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})} \left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right] \\ & + \alpha \, \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})} \left[\log p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})\right] \\ & - \beta \, KL \left[q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) \, || \, p_\phi(\boldsymbol{z}|\boldsymbol{c})\right], \end{aligned} \tag{1}$$

where we introduce weights α and β inspired by the disentangled $\beta-$VAE (Higgins et al., 2016). The first term $\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})} \left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})\right]$ is unsupervised, whereas the second

$$\alpha \mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})} \left[\log p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})\right]$$

is supervised and $\beta KL \left[q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})||p_\phi(\boldsymbol{z}|\boldsymbol{c})\right]$ 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 the latent space $\boldsymbol{z}$ are connected to all the measurements $\boldsymbol{x}$ through the likelihood network, all the potential correlations between clinical measurement variables can be exploited in an unsupervised fashion while disentangling the latent variables using the guidance networks for $\boldsymbol{y}$. The expectation over the variational distribution $\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})}$ is approximated with a few Monte-Carlo samples (Appendix B.1).

Given a dataset with N iid patients $\{\boldsymbol{x}^i_{1:T_i}, \boldsymbol{y}^i_{1:T_i}, \boldsymbol{c}^i_{1:T_i}\}_{i=1}^N$, the optimal parameters are obtained by the maximization task

$$\psi^*, \theta^* = \underset{\psi, \theta}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{k=0}^{T_i} \mathcal{L}_k(\psi, \theta; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{c}^i),$$

which is solved with stochastic optimization using mini-batches of patients and different values for k (Appendix B.1.2). Since real-world time series data often contains many missing values, the objective function can be adapted accordingly (Appendix B.1.1).

3.7. Online Prediction with Uncertainty Quantification

Our model can be used for online monitoring and continuous prediction of high-dimensional medical label and clinical measurement distributions based on an increasing number of available past clinical observations $\boldsymbol{x}_{0:k}$ for $k = 0, 1, \ldots, T$. The distributions

$$\begin{aligned} q_*(\boldsymbol{y}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) &= \int p_{\gamma^*}(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) q_{\theta^*}(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) d\boldsymbol{z} \\ q_*(\boldsymbol{x}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) &= \int p_{\pi^*}(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c}) q_{\theta^*}(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c}) d\boldsymbol{z} \end{aligned}$$

are approximated with two-stage Monte-Carlo sampling (Appendix B.1.3). The former can be used to automatically label and forecast the multiple medical labels based on the raw and partially observed measurements, whereas the latter corresponds to the reconstruction and forecasting of partially observed clinical measurement trajectories. Note that these distributions represent a complex class of potentially multi-modal distributions.

3.8. Patient Similarity and Clustering

The learned posterior network $q_{\theta^*}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}, \boldsymbol{c}_{1:T})$ can be used to map any observed patient trajectory $\mathcal{T}_i = \{\boldsymbol{x}^i_{1:T_i}, \boldsymbol{c}^i_{1:T_i}\}$ to their latent trajectory

$$\mathcal{H}_i = h(\mathcal{T}_i) = \mathbb{E}_{q_\theta(\boldsymbol{z}^i_{1:T_i}|\boldsymbol{x}^i_{1:T_i}, \boldsymbol{c}^i_{1:T_i})}\left[\boldsymbol{z}^i_{1:T_i}\right]$$

by taking the mean of the latent process. These temporal latent trajectories $\{\mathcal{H}_i\}_{i=1}^N$ of the N patients in the cohort are used to define a patient similarity over the partially observed and high-dimensional original disease trajectories $\{\mathcal{T}_i\}_{i=1}^N$. Through our semi-supervised generative approach, the latent trajectories effectively capture the important components from $\boldsymbol{x}^i_{1:T_i}$ and $\boldsymbol{y}^i_{1:T_i}$, without explicitly depending on $\boldsymbol{y}^i_{1:T_i}$. Indeed, all the information related to the medical labels is learned by θ.

Since defining a patient similarity measure between two trajectories $\mathcal{T}_i$ and $\mathcal{T}_j$ in the original space is very challenging, due to the missingness and high dimensionality of the variables, we instead define it in the latent space, setting

$$d_{\mathcal{T}}\left(\mathcal{T}_i, \mathcal{T}_j\right) = d_{\mathcal{H}}\left(\mathcal{H}_i, \mathcal{H}_j\right).$$

To measure the similarity $d_{\mathcal{H}}\left(\mathcal{H}_i, \mathcal{H}_j\right)$ between latent trajectories, we employ the *dynamic-time-warping (dtw)* measure to account for the different lengths of the trajectories as well as the potentially misaligned disease progressions in time (Müller, 2007). We then utilize the similarity measure to cluster the disease trajectories and identify similar patient trajectories as discussed in Section 5.3.2.

3.9. Deep Probabilistic Networks

As shown in Figures 2(a)subfigure and 2(b)subfigure, our model combines several deep probabilistic networks. For the posterior $q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k}, \boldsymbol{c})$, we implemented a temporal network with fully connected and LSTM layers (Hochreiter and Schmidhuber, 1997) and multilayer perceptrons (MLPs) 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. Implementation details are provided in Appendix C.1.

By omitting the guidance $p_\gamma(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c})$ or likelihood networks $p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, we recover well-established temporal latent variable models. Specifically, removing the guidance networks transforms the model into a deterministic predictive LSTM-Autoencoder, or probabilistic predictive LSTM-VAE if we learn the latent space distribution. Moreover, if we exclude the likelihood network $p_\pi(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, the model operates in a fully supervised setting, focusing solely on optimizing the latent space for the prediction of the medical labels $\boldsymbol{y}$. Many further architectural choices could be explored, such as a temporal likelihood network or a Gaussian process prior (Appendix B.2), but they are beyond the scope of this paper.

4. Cohort

We evaluate our model on the European Scleroderma Trials and Research (EUSTAR) database. The EUSTAR database extensively documents organ involvement in SSc for about 20'000 patients. For a detailed description of the database, we refer the reader to Meier et al. (2012); Hoffmann-Vold et al. (2021). We use this database because this work

is part of a broader initiative aiming to find the optimal medical definitions for organ involvement in SSc, leveraging data from the EUSTAR registry.

We included 5673 patients with at least 5 and at most 15 medical visits. We used 6 static variables related to the patients' demographics and almost 40 clinical measurement variables, mainly related to the lung, heart, and joint monitoring in SSc (Appendix A.3).

4.1. Data extraction and Feature Choices

The clinical measurement variables and patient demographics are directly available in the EUSTAR database. We provide a list of the used clinical measurement variables in Appendix A.3. They were selected based upon clinical relevance for modeling SSc. Each medical label is based on multiple EUSTAR variables (cf definitions in A.2) and created by using logical operations. For instance, the lung is involved if ILD on HRCT[1] = YES OR FVC[2] $< 70\%$.

4.2. Missing values

Missingness is a common issue in medical records. We used mean value imputation for missing clinical measurements. However, we did not train our model to reconstruct these missing measurements, i.e. the imputed values are not part of the optimized loss (cf Appendix B.1.1), thus mitigating the bias induced by the missingness. Additionally, given that the medical labels often rely on multiple EUSTAR variables, they are even sparser due to a propagation of the missingness. The advantage of our semi-supervised approach is that it relies solely on available labels for guidance, without necessitating any label imputation.

The code and examples using an artificial dataset are available as supplementary material.

5. Results on the EUSTAR Database

5.1. Study Design: Modeling Systemic Sclerosis

We aim to model the overall SSc disease trajectories as well as the distinct organ involvement trajectories for patients from the EUSTAR database.

We focus on the involvement of three important organs in SSc, namely the lung, heart, and arthritis in the joints. Each organ has two related medical knowledge labels: *involvement* and *stage*. Based upon the medical definitions provided in Appendix A.2, for each of the three organs, we created labels signaling the organ involvement (yes/no) and severity stage $(1-4)$, respectively. We write $o(m)$, $m \in \{involvement, stage\} := \mathcal{M}$, $o \in \mathcal{O} := \{lung, heart, joints\}$, to refer to the corresponding medical label for organ o. We project the $D = 34$ and $P = 11$ input features to a latent process $\boldsymbol{z}$ of dimension $L = 21$.

For each organ, we guide a distinct subset of 7 latent processes (non-overlapping subsets), thus all of the dimensions in $\boldsymbol{z}$ are guided (2(b)subfigure). Following the notations from

1. Interstitial Lung Disease on High-Resolution computed Tomography
2. Forced Vital Capacity

Section 3.4, we assume the following guidance network structure

$$p_\gamma(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c}) = \prod_{t=1}^{T}\prod_{o\in\mathcal{O}}\prod_{m\in\mathcal{M}} p_\gamma(\boldsymbol{y}_t^{\nu(o(m))}|\boldsymbol{z}_t^{\varepsilon(o(m))},\boldsymbol{c}_t),$$

where $\nu(o(m))$ and $\varepsilon(o(m))$ are the corresponding indices of the dimensions in the output and latent process, respectively.

5.1.1. Evaluation

Given the conditioning data $\boldsymbol{c}$ and the clinical measurements $\boldsymbol{x}_{0:k}$ up to a given time-step k, our model learns the optimal parameters of the variational distribution of the complete latent trajectory $q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k},\boldsymbol{c})$, of the likelihood $p_\pi(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})$ and of the guidance networks $p_\gamma(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c})$. Thus given $\boldsymbol{x}_{0:k}$ and $\boldsymbol{c}$, our model predicts the complete trajectories of both $\boldsymbol{x}$ and $\boldsymbol{y}$.

We aim to propose a holistic model that prioritizes versatility over achieving cutting-edge predictive performance. Thus, we evaluate the predictive trade-offs of our approach versus well-established deterministic and probabilistic temporal deep latent variable models optimized separately for each of the predictive tasks, i.e. predicting only $\boldsymbol{x}$ or $\boldsymbol{y}$ in a fully supervised way.

We evaluate the interpretability and disentanglement of the latent processes in our model against fully unsupervised methods. Furthermore, we evaluate and discuss the clinical relevance of the trajectory clusters identified for SSc patients. Lastly, we follow an index patient to showcase how our model enhances the understanding of their disease course, including online patient monitoring, patient trajectory sampling, and visualizations in the latent space. We refer the reader to the Appendix D.1 for additional results related to patient similarity, adjustment of uncertainty quantification to out-of-distribution data, and sampling of prior trajectories.

5.2. Predictive Performance Evaluation

5.2.1. Baselines

The temporal baselines follow a similar encoder-decoder architecture as our model and are optimized to predict either $\boldsymbol{x}$ or $\boldsymbol{y}$ as targets. Similarly to our model, their temporal encoders take as input $\boldsymbol{x}_{0:k}$ and $\boldsymbol{c}$ and learn the distribution of the latent variables $\boldsymbol{z}$. The predictive decoders take as input a sampled $\boldsymbol{z}$ and predict the future targets. We implemented probabilistic and deterministic versions of each baseline. The encoders of the probabilistic models learn the mean and variance of the distributions of the latent variables, and the encoders of the deterministic models only learn their mean.

Similarly to our model, the temporal encoders contain LSTM and fully connected layers, and the decoders are MLPs. We denote LSTM-MLP-x and LSTM-MLP-y for the deterministic supervised models trained to predict $\boldsymbol{x}$ or $\boldsymbol{y}$, respectively, and LSTM-MLP-x* and LSTM-MLP-y* for the probabilistic variants. We expect these models to generally outperform our approach, since they have a similar model capacity but learn simpler tasks. Their training objective can be expressed as simplified versions of our model's objective (Equation (1)). The associated loss functions can be found in Appendix D.1.1.

In addition to these temporal deep learning models, we also evaluated our approach against a non-temporal MLP baseline taking as input the conditioning data and the last available values of each clinical measurement $\boldsymbol{x}$ before the prediction time-point. We denote this baseline as MLP-xy, as it predicts both $\boldsymbol{x}$ and $\boldsymbol{y}$. Lastly, we also implemented a naive cohort baseline drawing a value from the empirical distribution of the variable in the cohort (assuming a Gaussian distribution for continuous variables and a categorical distribution otherwise). We used 5−fold cross-validation to select the hyperparameters that achieved the lowest validation loss for each model. Details about the inference process are provided in subsection 3.6 and Appendix C.2.1.

5.2.2. Results

In Figure 3, we report the predictive performance of the different models for the prediction of future $\boldsymbol{x}$ and $\boldsymbol{y}$ versus time to prediction. We report the average macro F1 score for categorical variables and the mean absolute error (MAE) for continuous variables.

In the first panel of Figure 3, we evaluate the models' performance for the prediction of medical labels $\boldsymbol{y}$. Both of the task-specific models, i.e. the LSTM-MLP-y (yellow) and LSTM-MLP-y* (orange), slightly outperform our model (red), as expected since they only have to learn one category of outcomes. Furthermore, our model outperforms the MLP-xy (grey) and naive (green) baselines. We report the performance results separately for each medical label in Appendix D.1.

The last two panels of Figure 3 show the prediction performance for categorical and continuous $\boldsymbol{x}$ versus time to prediction. For categorical $\boldsymbol{x}$, our model (red) performs similarly or outperforms all of the models, except for the first time-step, where the MLP-xy baseline (grey) performs the best. For continuous features, the LSTM-MLP-x (purple) outperforms our model in terms of MAE. There is no significant difference between our model and the LSTM-MLP-x* (brown), even though our model also learns the distribution of $\boldsymbol{y}$. Lastly, our model greatly outperforms the MLP-xy (grey) and naive (green) baselines.

As expected, the task-specific deterministic models generally slightly outperform our model, when allowed a similar capacity in the latent space, since they have to learn simpler tasks and fewer variables. In Figure 8 in Appendix D.1, we show that by decreasing the capacity of the LSTM-MLP-x, via reduction of the latent space dimension, we recover a similar performance to our multi-task holistic model.

To evaluate the uncertainty quantification, we computed the coverage of the forecasted 95% confidence intervals (CI) for continuous variables and the calibration for categorical variables. Furthermore, we computed the average ratio between CI length and feature range versus time to prediction. CIs are on average wider for long-term predictions and out-of-distribution data points (Figure 10 in Appendix D.1). For continuous $\boldsymbol{x}$ forecasting, our model and the LSTM-MLP-x* achieve coverage of $92 \pm 1\%$ both, and the LSTM-MLP-x of $98 \pm 0\%$, thus all slightly diverging from the optimal 95%. All of the models have accurate calibration for categorical $\boldsymbol{x}$ and $\boldsymbol{y}$ forecasting, as shown in Figure 9 in Appendix D.1.

5.2.3. Online Prediction with Uncertainty Quantification

To illustrate how the model allows a holistic understanding of a patient's disease course, we follow an index patient p_{idx} throughout the experiments. This patient has a complex

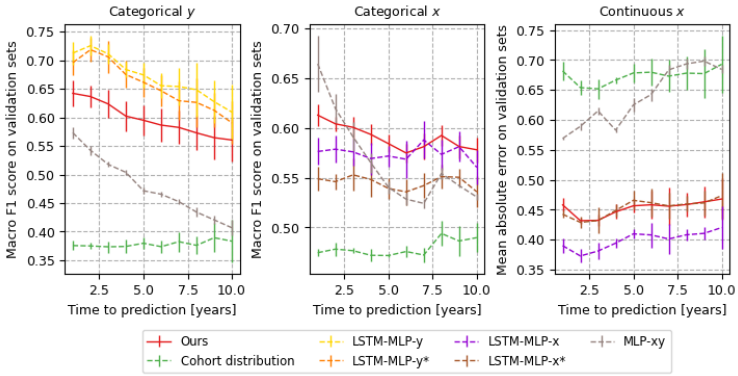
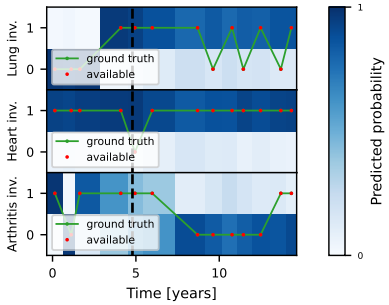
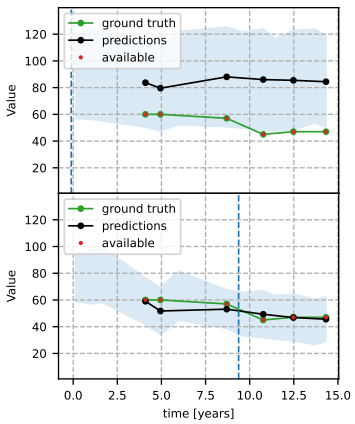


Figure 3: Model performances versus time to prediction for $\boldsymbol{x}$ and $\boldsymbol{y}$ forecasting.



(*a*) Predicted probabilities of organ involvement. The heatmap reflects the predicted probabilities.



(*b*) FVC: predicted mean and 95% CIs at two different time points

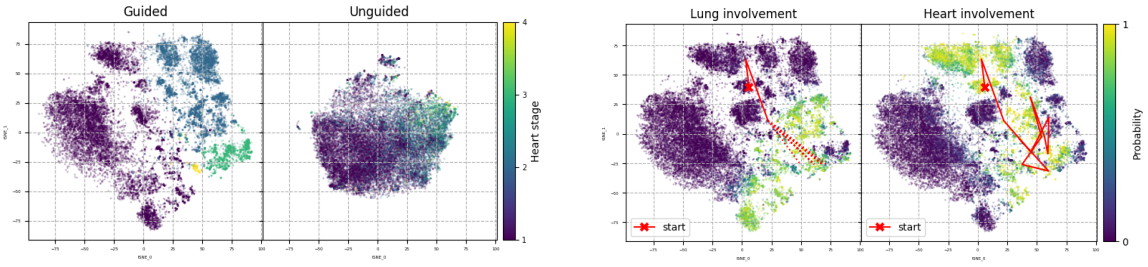
Figure 4: Online monitoring for p_{idx}. The model uses information prior to the dashed line as input and predicts the values after.

disease trajectory, with varying organ involvement and stages. We can use our model to forecast the high-dimensional distributions of $\boldsymbol{x}_{1:T}$ and $\boldsymbol{y}_{1:T}$ given the past measurements $\boldsymbol{x}_{0:k}$, as described in subsection 3.7. The plots in Figure 4 show the predicted probabilities of organ involvement and predicted values of Forced Vital Capacity (FVC)[3] at different time points for p_{idx}. The plots are overlaid with the ground truth labels in green. In particular, Figure 4(b)subfigure shows how the predictions become more accurate when more prior information is available to the model. We provide online prediction plots for additional $\boldsymbol{x}$ and $\boldsymbol{y}$ in Appendix D.1.4.

In the next sections, we explore further applications and results of our model. While the performance was computed on validation sets, the subsequent results are derived from applying our model to a separate withheld test set. Furthermore, all of the $t-$SNE projections (Van der Maaten and Hinton, 2008) of the test set were obtained following the procedure described in Appendix D.2.2.

5.3. Results: Cohort Analysis

By learning the joint distribution $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, our model allows us to analyze disease patterns in the cohort through the analysis of the latent process $\boldsymbol{z}$. Furthermore, by learning $p(\boldsymbol{z}|\boldsymbol{c})$, we estimate the average prior disease trajectories in the cohort. We analyze these prior trajectories in Appendix D.2.1.



(a) Guided versus unguided latent spaces, overlaid with heart stage medical label.

(b) Probabilities of lung and heart involvement in the latent space.

Figure 5: Analysis of latent spaces.

5.3.1. Latent Space and Medical Labels

We aim to provide a method achieving semi-supervised disentanglement in the latent space. In 5(a)subfigure, we compare the distribution of the ground truth medical labels (here *heart stage*) in a guided versus an unguided model (i.e. without training any guidance networks). The guided model clearly provides higher medical knowledge label disentanglement than the unguided model and thus enhances the interpretability of the different subspaces in $\boldsymbol{z}$.

In 5(b)subfigure, we visualize the latent space overlaid with the different predicted probabilities of organ involvement. In red, we draw the latent space trajectory of p_{idx}, thus getting an understandable overview of their trajectory with respect to the different medical

3. FVC is the amount of air that can be exhaled from the lungs.

labels. The solid line highlights the reconstructed trajectory, whereas the dotted lines are forecasted sampled trajectories.

In the first panel of 5(*b*)subfigure, we leverage the model's generative abilities to sample forecasted $\boldsymbol{z}$ trajectories (dotted lines), providing estimates of future disease stages. The model forecasts that p_{idx} will move towards a region with higher probabilities of lung and heart involvement. All of the sampled trajectories converge towards the same region in this case. The second panel is overlaid with the complete reconstructed trajectory of p_{idx} in the latent space. The disentanglement in the latent space enables a straightforward overview of the past and future patient trajectory. Additionally, Figure 14 in the appendix shows the patient trajectory overlaid with the predicted organ stages.

5.3.2. Clustering and Similarity of Patient Trajectories

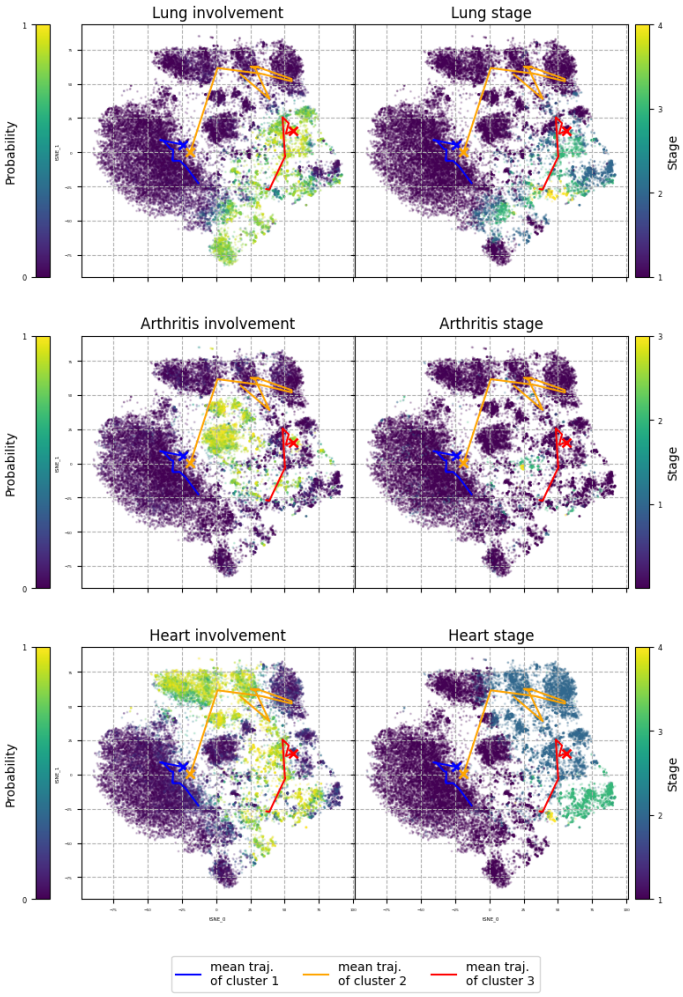
As described in subsection 3.8, we compute the dynamic-time-warping similarity measure for the latent trajectories $\mathcal{H}_i = h(\mathcal{T}_i)$, and subsequently apply *k-means* or *k-nn* to respectively cluster the multivariate time series $\{\mathcal{H}_i\}_{i=1}^N$ or find similar patient trajectories. We used the library implemented by Tavenard et al. (2020). We focus here on the trajectory clustering results and refer the reader to Appendix D.2.3 for the patient similarity/nearest neighbor analysis.

Table 1: Prevalence of cutaneous involvement and gender in the clusters versus cohort prevalence. The arrows indicate the direction of the relative change compared to the cohort prevalence.

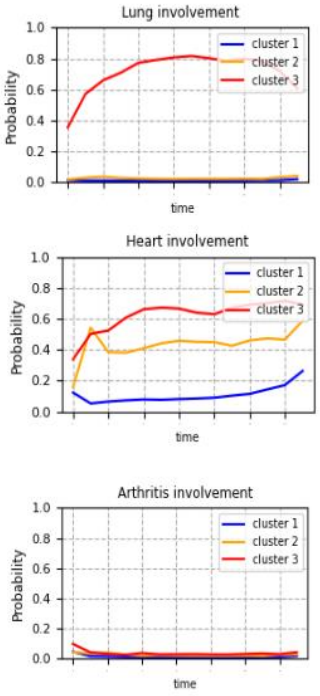
	Diffuse Cutaneous Involvement of SSc	Male
Percentage in Cohort	**33%**	**13%**
Percentage in mild severity cluster	26% ↓	13%
Percentage in medium severity cluster	31%	9% ↓
Percentage in high severity cluster	46% ↑	21% ↑

6(*a*)subfigure shows the three mean cluster trajectories in the latent space overlaid with the predicted medical labels. Moreover, we computed the predicted organ involvement probabilities for the cluster mean trajectories using the guidance networks (6(*b*)subfigure). The first found cluster corresponds to patients with no or little organ involvement. The second mean trajectory starts close to the first but progresses towards regions with heart involvement. The third cluster contains the most severely progressing patients. The identified clusters show distinct patterns of organ involvement and disease severity (mild-medium-high), showing that the model separates disease trajectories into further subtypes.

Clustering Evaluation We evaluated our clustering approach both quantitatively and clinically. The optimal number of clusters k, was identified using the elbow method as shown in Figure 15 in the Appendix. We compared two methods: clustering latent trajectories z against direct clustering of raw trajectories x. As discussed in Appendix D.2.3, clustering latent trajectories achieves higher separation with respect to the medical labels compared to clustering the raw data. Lastly, contrarily to traditional clustering approaches, our



(*a*) Mean cluster trajectories in the latent space (starting at the cross $\mathbf{x}$), overlaid with predicted probabilities of organ involvement and severity stages.



(*b*) Probabilities of organ involvement for cluster means.

Figure 6: Clustering of latent trajectories and predicted probabilities of organ involvement for the mean cluster trajectories.

method also supports *predictive clustering*. Indeed, we can compare the cluster assignment of a forecasted trajectory (illustrated by the dotted line in Figure 5) to the final cluster assignment based on the complete patient history encoded in the model. This approach achieved a macro F_1 score of 0.78, indicating effective patient assignment to severity clusters early in their diagnosis.

We further evaluated the clinical relevance of the found clusters. In Table 1, we compare the prevalence of SSc subtypes (limited versus diffuse cutaneous SSc) and gender between the clusters. For instance, the cluster with the mildest severity contains a higher proportion of patients with limited cutaneous involvement at baseline, while the cluster with the highest severity includes a significantly larger number of patients with diffuse cutaneous SSc, showing that the model can separate the trajectories based upon the widely accepted SSc subtypes (Bains, 2017). Furthermore, the most severe cluster exhibits a significantly higher proportion of male patients in comparison to the rest of the cohort. Recent studies also found that males tend to experience more frequent and severe lung and heart involvement, and concurrent organ involvement tends to result in poorer overall outcomes (Peoples et al., 2016; Becker et al., 2019).

6. Discussion

In this paper, we present a novel deep semi-supervised generative latent variable approach to model complex disease trajectories. By introducing the guidance networks, we propose a method to augment unsupervised deep generative models with established medical knowledge and achieve more interpretable and disentangled latent processes.

Our non-discriminative approach effectively addresses important 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, namely the modeling of systemic sclerosis, and enables a holistic understanding of the patients' disease course. The disentangled latent space facilitates comprehensive trajectory visualizations, straightforward analysis, and forecasting of patient trajectories. Most importantly, learning medically informed latent processes allows the discovery of novel clinically meaningful disease subtypes. We showed that the cluster separation is driven by clinically relevant features that have also been recognized as important predictors of SSc trajectories in recent studies.

Limitations and Future Work While we have presented the benefits of proposing a multi-task "holistic" model, this approach also has limitations. Naturally, the model is less performant at specific tasks compared to fine-tuned models, for instance, fully supervised predictive models for prediction. However, our modular approach could be adapted to excel in specific settings by removing certain components of the model.

Our current approach holds the potential to be extended and adapted in several ways. We included only the most pertinent experiments and opted for a simple architecture suited to the modeling of systemic sclerosis. For instance, the model could be explicitly trained to reconstruct missing values, akin to denoising autoencoders. 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), include more

organs in the modeling of SSc, and also include guidance networks to model additional disease dynamics like long-term outcomes.

Acknowledgments

The authors thank the patients and caregivers who made the study possible, as well as all involved clinicians from the EUSTAR who collected the data. This work was funded by the Swiss National Science Foundation (project number 201184).

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 (Chadebec et al., 2022). The code and examples using some artificial data are available at https://github.com/uzh-dqbm-cmi/eustar_mlhc.

References

Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems*, 32, 2019.

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

Pooja Bains. Classification criteria of systemic sclerosis: Journey so far. *Our Dermatology Online*, 8(2):220–223, 2017.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

Mike Becker, Nicole Graf, Rafael Sauter, Yannick Allanore, John Curram, Christopher P Denton, Dinesh Khanna, Marco Matucci-Cerinic, Janethe de Oliveira Pena, Janet E Pope, et al. Predictors of disease worsening defined by progression of organ damage in diffuse systemic sclerosis: a european scleroderma trials and research (eustar) analysis. *Annals of the rheumatic diseases*, 78(9):1242–1248, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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

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.

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.

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.

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.

Irene Y Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *Annual review of biomedical data science*, 4:393–415, 2021.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Xingyu Chen, Xiaochen Zheng, Amina Mollaysa, Manuel Schürch, Ahmed Allam, and Michael Krauthammer. Dynamic local attention with hierarchical patching for irregular clinical time series, 2023.

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.

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.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

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.

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.

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.

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

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.

Robbie Holland, Oliver Leingang, Christopher Holmes, Philipp Anders, Rebecca Kaye, Sophie Riedl, Johannes C Paetzold, Ivan Ezhov, Hrvoje Bogunović, Ursula Schmidt-Erfurth, et al. Clustering disease trajectories in contrastive feature space for biomarker discovery in age-related macular degeneration. *arXiv preprint arXiv:2301.04525*, 2023.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

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.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

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

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

David N Lawley and Adam E Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

Changhee Lee and Mihaela Van Der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In *International conference on machine learning*, pages 5767–5777. PMLR, 2020.

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, 2020a.

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, 2020b.

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.

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

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

Shahriar Noroozizadeh, Jeremy C Weiss, and George H Chen. Temporal supervised contrastive learning for modeling patient risk progression. In *Machine Learning for Health (ML4H)*, pages 403–427. PMLR, 2023.

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

Christine Peoples, Thomas A Medsger Jr, Mary Lucas, Bedda L Rosario, and Carol A Feghali-Bostwick. Gender differences in systemic sclerosis: relationship to clinical features, serologic status and outcomes. *Journal of scleroderma and related disorders*, 1(2): 204–212, 2016.

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

Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. T-phenotype: Discovering phenotypes of predictive temporal patterns in disease progression. In *International Conference on Artificial Intelligence and Statistics*, pages 3466–3492. PMLR, 2023.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Sequential multi-dimensional self-supervised learning for clinical time series. In *International Conference on Machine Learning*, pages 28531–28548. PMLR, 2023.

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.

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

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

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

Manuel Schürch, Xiang Li, Ahmed Allam, Giulia Hofer, Amina Mollaysa, Claudia Cavelti-Weder, and Michael Krauthammer. Generating personalized insulin treatments strategies with conditional generative time series models. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023b.

Manuel Pascal Schürch. Contributions to scalable gaussian processes. 2022.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016.

Shivin Srivastava and Vaibhav Rajan. Expertnet: A deep learning approach to combined risk modeling and subtyping in intensive care units. *IEEE Journal of Biomedical and Health Informatics*, 2023.

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.

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

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.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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

Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94, 2014.

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

Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd international conference on World wide web*, pages 783–794, 2014.

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.

Appendix A. Systemic Sclerosis

A.1. 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 knowledge labels (involvement, stage, progression) for all impacted organs in SSc, we used preliminary definitions for three organs.

A.2. Medical Labels 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 labels 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 labels 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

A.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

A.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

A.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

A.3. Model variables

Our model uses as temporal input features the following variables related to each organ and collected during medical visits:

- Lung: Forced Vital Capacity, DLCO/SB, DLCOc/VA, Lung fibrosis, Dyspnea (NYHA-stage), Worsening of cardiopulmonary manifestations within the last month, HRCT: Lung fibrosis, Ground glass opacification, Honey combing, Tractions, Reticular changes, PAPsys (mmHg), TAPSE: tricuspid annular plane systolic excursion in cm, Right ventricular area (cm^2) (right ventricular dilation), Tricuspid regurgitation velocity (m/sec), Pulmonary wedge pressure (mmHg), Pulmonary resistance, 6 Minute walk test (distance in m)
- Heart: Left ventricular ejection fraction, Worsening of cardiopulmonary manifestations within the last month, Diastolic function abnormal, Ventricular arrhythmias, Arrhythmias requiring therapy, Pericardial effusion on echo, Conduction blocks, NT-proBNP (pg/ml), Auricular Arrhythmias, BNP (pg/ml), Cardiac arrhythmias, Dyspnea (NYHA-stage)
- Arthritis: Joint synovitis, Joint polyarthritis, Swollen joints, Tendon friction rubs, DAS 28 (ESR, calculated), DAS 28 (CRP, calculated)

Moreover, we use the following the following (static) demographic variables:

- Demographics: Sex, Height, Race, Subset of SSc according to LeRoy, Date of birth, Onset of first non-Raynaud's of the disease.

Appendix 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 3.1.

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) (Blei et al., 2017), 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 (Murphy, 2022), therefore, we aim to find parameters satisfying

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

Table 2: Table of symbols

Symbol	Description	Domain
$\boldsymbol{x} = \boldsymbol{x}_{1:T}$	Clinical measurements (e.g., blood pressure).	$\mathbb{R}^{D\times T}$
$\boldsymbol{y} = \boldsymbol{y}_{1:T}$	Medical knowledge labels (e.g., disease severity).	$\mathbb{R}^{P\times T}$
$\boldsymbol{\tau} = \boldsymbol{\tau}_{1:T}$	Observation time-points.	$\mathbb{R}^{T}$
$\boldsymbol{s}$	Non-temporal info (e.g. patient demographics).	$\mathbb{R}^{S}$
$\boldsymbol{p} = \boldsymbol{p}_{1:T}$	Additional temporal covariates (e.g. medications).	$\mathbb{R}^{P\times T}$
$\boldsymbol{c}$	Context variables $(\boldsymbol{\tau}, \boldsymbol{p}, \boldsymbol{s})$.	
$\boldsymbol{z} = \boldsymbol{z}_{1:T}$	Multivariate latent processes	$\mathbb{R}^{L\times T}$
$p_\phi(\boldsymbol{z}\|\boldsymbol{c})$	Prior network	
ϕ	Prior network parameters	
$p_\pi(\boldsymbol{x}\|\boldsymbol{z}, \boldsymbol{c})$	Likelihood network	
π	Likelihood network parameters	
$p_\gamma(\boldsymbol{y}\|\boldsymbol{z}, \boldsymbol{c})$	Guidance network	
γ	Guidance network parameters	
$\psi = \{\gamma, \pi, \phi\}$	Parameters of the generative model.	
$q_\theta(\boldsymbol{z}\|\boldsymbol{x}_{0:k}, \boldsymbol{c})$	Variational distribution	
θ	Variational parameters	
α, β	Weights in the training objective for balancing terms.	$\mathbb{R}_+$
$\mathcal{T}$	Observed patient trajectory (clinical measurements over time).	Sequence in $\mathbb{R}^{D\times T}$
$\mathcal{H}$	Latent patient trajectory (latent space representation).	Sequence in $\mathbb{R}^{L\times T}$

This optimization problem is equivalent (Murphy, 2022) 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

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

which can be rearranged to

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

For the Gaussian prior and approximate posterior described in Section 3.2 and 3.5, respectively, the KL-term can be computed analytically and efficiently (Tomczak, 2022). 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

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

B.1.1. Partially Observed Data

The measurements $\boldsymbol{x} \in \mathbb{R}^{D\times T}$ and the labels $\boldsymbol{y} \in \mathbb{R}^{P\times T}$ 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. (2020); Ramchandran et al. (2021). 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 labels $\boldsymbol{y}^{\boldsymbol{o}_y}$.

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 (subsection 3.6). Moreover, suppose we have $T+1$ iid copies of the whole dataset $\{\mathcal{D}^k\}_{k=0}^T$, then

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

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 label trajectories $\boldsymbol{y}_{1:T}$ in subsection 3.7 can be obtained via a two-stage Monte-Carlo approach. For instance, we can sample from the distribution of the measurements

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

by first sampling from the latent trajectories

$$\boldsymbol{z}^1_{1:T}, \dots, \boldsymbol{z}^s_{1:T}, \dots \boldsymbol{z}^S_{1:T} \sim q_{\theta^*}(\boldsymbol{z}_{1:T}|\boldsymbol{x}_{0:k}, \boldsymbol{c})$$

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

$$\boldsymbol{x}^1_{1:T}, \dots, \boldsymbol{x}^u_{1:T}, \dots \boldsymbol{x}^U_{1:T} \sim p_{\pi^*}(\boldsymbol{x}_{1:T}|\boldsymbol{z}^s_{1:T}, \boldsymbol{c})$$

to represent the overall uncertainty of the measurement distribution.

B.2. Different Prior

The factorized prior described in subsection 3.2 can be extended to continuous time with Gaussian processes (GPs) (Williams and Rasmussen, 2006; Schürch et al., 2020, 2023a; Schürch, 2022), as introduced by Casale et al. (2018); Fortuin et al. (2020) in the unsupervised setting. In particular, we can replace

$$\begin{aligned} p_\phi(\boldsymbol{z}|\boldsymbol{c}) = p_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{c}_{1:T}) &= \prod_{t=1}^{T}\prod_{l=1}^{L} p_\phi(\boldsymbol{z}^l_t|\boldsymbol{c}_t) \\ &= \prod_{t=1}^{T}\prod_{l=1}^{L} \mathcal{N}\left(\boldsymbol{z}^l_t|\mu^l_\phi(\boldsymbol{c}_t), \sigma^l_\phi(\boldsymbol{c}_t)\right), \end{aligned}$$

with

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

with a mean function $m^l_\phi(\boldsymbol{c})$ and kernel $k^l_\phi(\boldsymbol{c}, \boldsymbol{c}')$, 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 proposed by Schürch et al. (2023b). 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

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

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

Appendix C. Model Implementation

C.1. Model Architecture

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

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

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 $\hat{\mu}^{prior}_{1:T}$ and $\hat{\sigma}^{prior}_{1:T}$.
- **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 label (involvement and stage). A guidance network for organ $o \in \mathcal{O} := \{lung, heart, joints\}$ and related medical label $m \in \{inv, stage\}$, 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 label $o(m)$, and $\epsilon(o(m))$ the indices in the latent space.

C.2. Training Objective

We follow the notation introduced in Section 3 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 labels. The total loss for a patient p is

$$\mathcal{L}_p = \sum_{t=0}^{T} \mathcal{L}(t), \tag{2}$$

where

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

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.2.1. 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 (Kingma and Ba, 2014) optimizer with mini-batch processing and early stopping.

C.2.2. 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 $\beta = 0.01$.

Appendix D. Results

D.1. Model Evaluation

D.1.1. Baselines

As discussed in subsubsection 5.2.1, we evaluated our model against temporal latent variable models optimized to predict either only $\boldsymbol{x}$ or $\boldsymbol{y}$ in a fully supervised way. We implemented probabilistic and deterministic variants of these models. Following the notations introduced in subsection 3.1 for the encoder networks (posterior), decoder and guidance networks, we can rewrite the objectives of the baselines as variants of the objective of our model objective described in Equation (1).

For the prediction of the clinical measurements, the LSTM-MLP-x* optimizes the objective

$$\begin{aligned}\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k},\boldsymbol{c})}\left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})\right] \\ -\beta\ KL\left[q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k},\boldsymbol{c})\ ||\ p_\phi(\boldsymbol{z}|\boldsymbol{c})\right],\end{aligned}$$

and the LSTM-MLP-x

$$\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k},\boldsymbol{c})}\left[\log p_\pi(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})\right]$$

respectively. Similarly, for the prediction of the medical labels, the LSTM-MLP-y* optimizes

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

and the LSTM-MLP-y

$$\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x}_{0:k},\boldsymbol{c})}\left[\log p_\gamma(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{c})\right].$$

D.1.2. Results

Figure 7 shows the prediction performance of the different models for each of the medical labels. The supervised models generally slightly outperform our model, and all temporal models greatly outperform the MLP and cohort baselines.

As discussed in the model evaluation, given the same model capacity, the LSTM-MLP baselines are expected to outperform our approach since they learn simpler tasks and fewer variables. Figure 8 shows the effect of reducing the dimension of the latent space of the deterministic LSTM-MLP-x baseline. Contrarily to our model, this baseline learns neither the variance of the latent variables nor the distribution of $\boldsymbol{y}$. For continuous $\boldsymbol{x}$, the LSTM-MLP-x with five dimensions in the latent space performs similarly to our model.

D.1.3. Uncertainty quantification

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. Figure 10 shows the average ratio between CI length and feature range versus time to prediction. CIs are on average wider for long-term predictions and out-of-distribution data points, showing that the model predicts higher uncertainty for data points that are more difficult to predict. 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 9. The calibration curves in Figure 9 show that all of the temporal models are well calibrated both in their categorical $\boldsymbol{x}$ and medical label $\boldsymbol{y}$ forecasts (averaged over all forecasted data points in the respective validation sets).

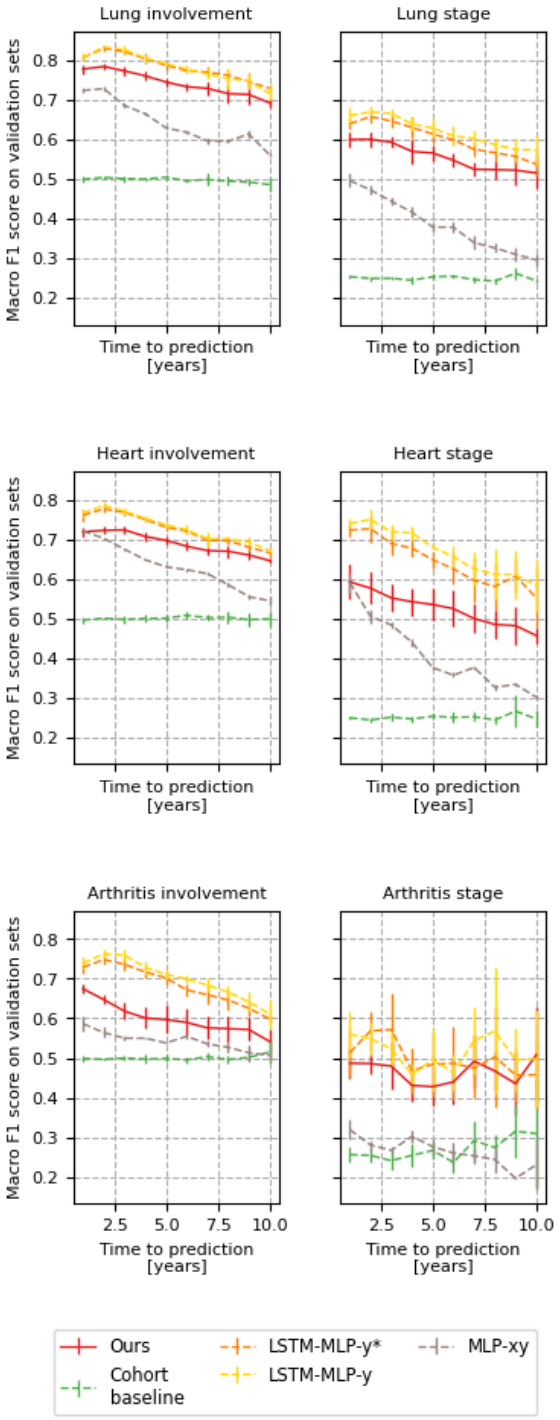


Figure 7: Performance for $\boldsymbol{y}$ prediction.

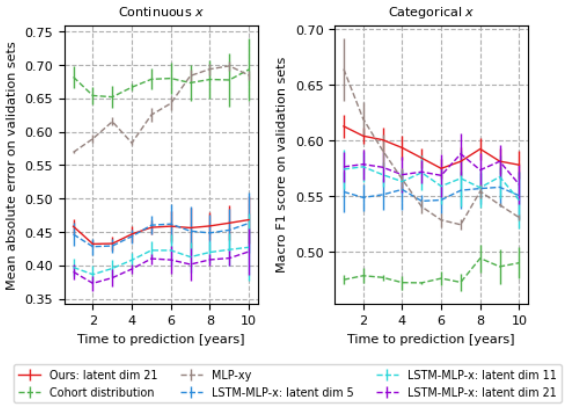


Figure 8: Effect of latent space dimension.

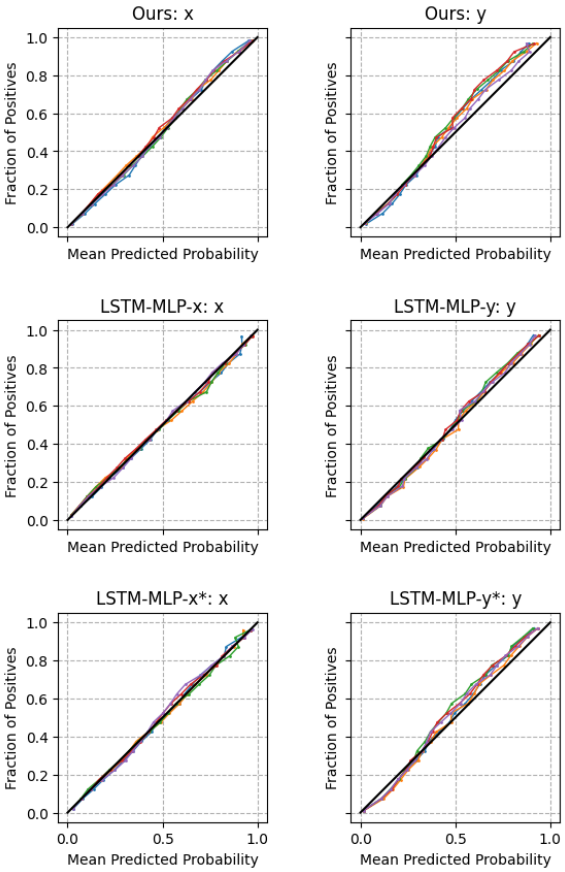


Figure 9: Calibration curves for our model and the LSTM-MLPs.

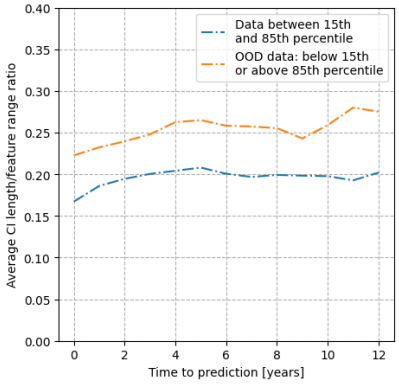


Figure 10: Average ratio between CI length and feature range versus time to prediction. Out-of-distribution data points have wider CIs on average.

D.1.4. Online Prediction with Uncertainty

We provide additional online prediction results for the index patient p_{idx}.

Figure 11(a)subfigure shows the evolution in the predicted mean and 95% CI of DLCO(SB)[4] 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. Moreover, in Figure 12 we contrast the predicted uncertainty for a patient with an out-of-distribution (OOD) number of swollen joints (i.e. an unusually high number of swollen joints), and for the index patient. The model predicts significantly larger CIs for the OOD data point.

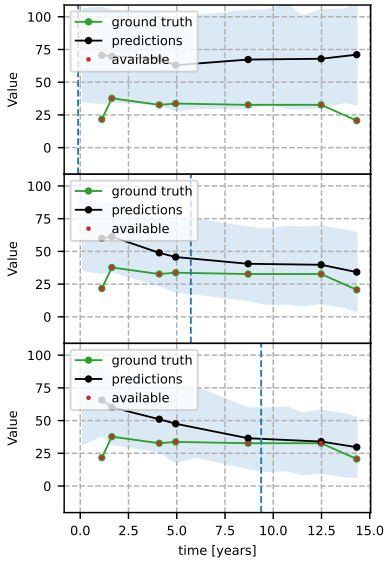
D.2. Cohort Analysis

We present here additional cohort-level experiments using our model.

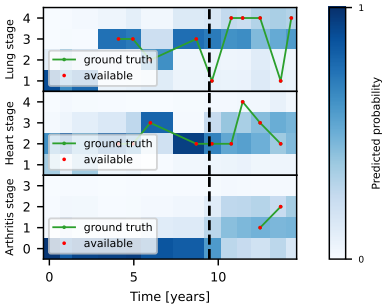
D.2.1. Prior z Distributions

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 13(a)subfigure we overlaid the predicted prior trajectories of Forced Vital Capacity (FVC)[5] 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 13(b)subfigure, the prior predicted N-terminal pro b-type natriuretic peptide (NTproBNP)[6] trajectories overlaid with age, show that the model predicts an overall increase in NTproBNP over time, and steeper for older patients.

4. DLCO(SB) stands for single breath (SB) diffusing capacity of carbon monoxide (DLCO).
5. FVC is the amount of air that can be exhaled from the lungs. Low levels indicate lung malfunction.
6. They are substances produced by the heart. High levels indicate potential heart failure.

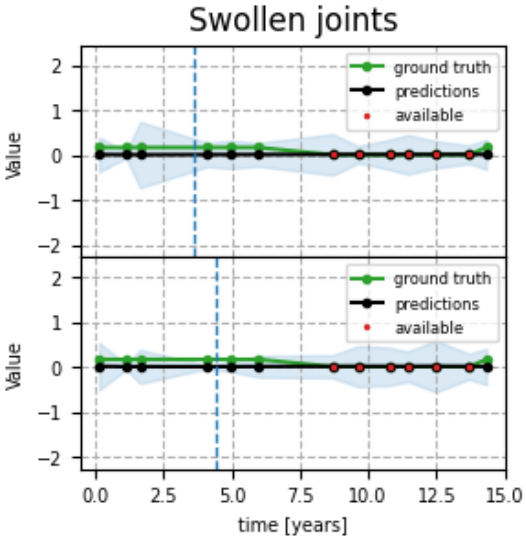


(*a*) DLCO(SB) of p_{idx}: predicted mean and 95% CI.

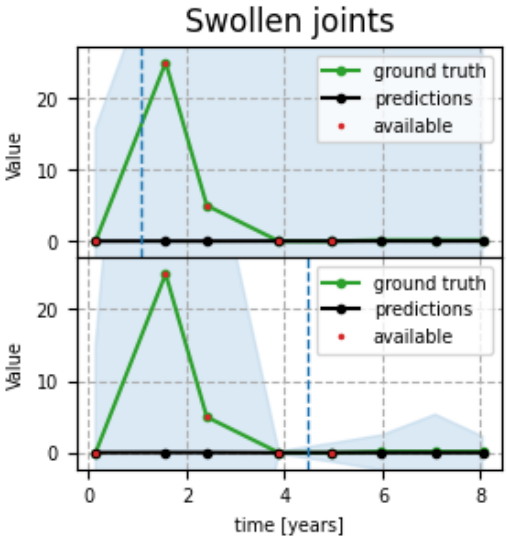


(*b*) Probabilities of organ stages for p_{idx}. The intensity of the heatmap reflects the predicted probability.

Figure 11: DLCO(SB) and organ stage probabilities for p_{idx}.

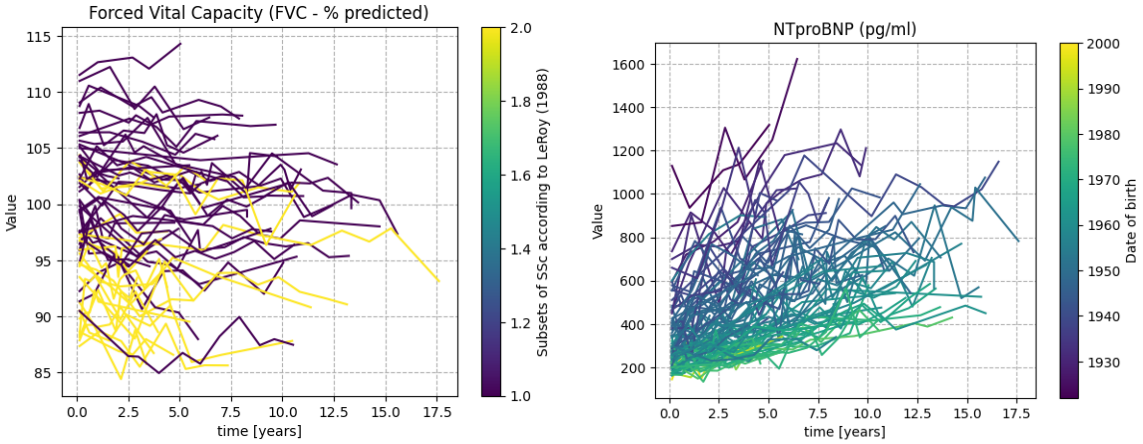


(*a*) Swollen joints for index patient



(*b*) Swollen joints for out of distribution patient

Figure 12: Comparison between in and out-of-distribution predictions of swollen joints.



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

(*b*) Prior natriuretic peptides trajectories overlaid with date of birth.

Figure 13: Prior predicted $\boldsymbol{x}$ trajectories conditioned on time and static variables.

D.2.2. Latent Space and Medical labels

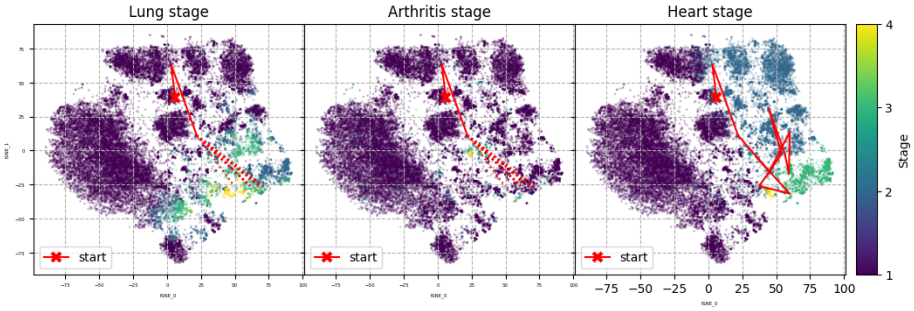


Figure 14: Predicted organ stages in the latent space. The red line highlights the trajectory of p_{idx}.

t−SNEs: The *t*-SNE (Van der Maaten and Hinton, 2008) 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 $\mathcal{P}_{train}$ and then transforming and plotting the projected latent variables (reconstructed or forecasted) from patients in $\mathcal{P}_{test}$ (Poličar et al., 2019).

In 5(*b*)subfigure, we showed the trajectory of p_{idx} overlaid with the predicted organ involvement probabilities. In 14, 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 5.3.2). In 18(a)subfigure, we show the different predicted probabilities of the medical labels $\boldsymbol{y}$ for the mean trajectories within the three found clusters. This reveals which medical labels 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. We compared our approach of clustering latent trajectories z to clustering the raw trajectories x directly. In Figure 16, we compare the average medical label trajectories in the clusters obtained using both approaches. We see that clustering latent trajectories achieves more separation with respect to the medical labels than clustering the raw data. This indicates that our approach is better suited to uncover new subtypes with respect to medical knowledge. Furthermore, in Table 1, we compare the prevalence of SSc subtypes (limited versus diffuse cutaneous SSc) and gender between the clusters. For instance, the most severe cluster contains an increased proportion of males compared to the cohort prevalence.

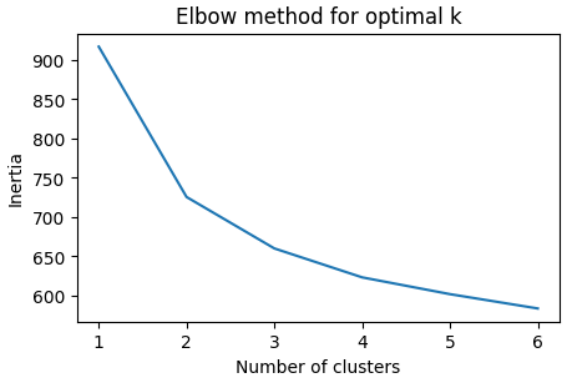


Figure 15: Clustering: elbow plot for choice of optimal k. We set k to 3.

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 17 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 18(b)subfigure, the medical label trajectories of p_{idx} and its nearest neighbors reveal consistent patterns.

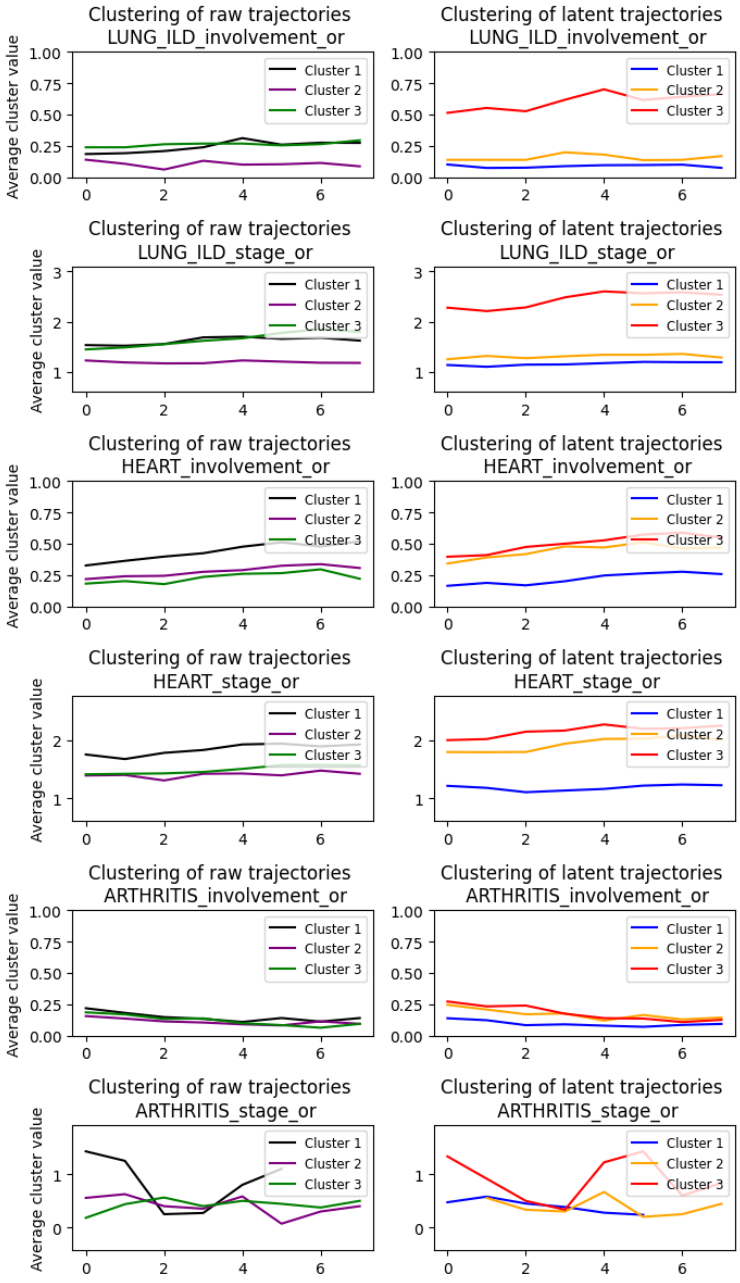


Figure 16: Cluster separation with respect to medical labels. Comparison between clustering of raw x trajectories versus clustering of latent trajectories z. Our approach, where we cluster the latent trajectories, shows a higher separation with respect to the medical labels.

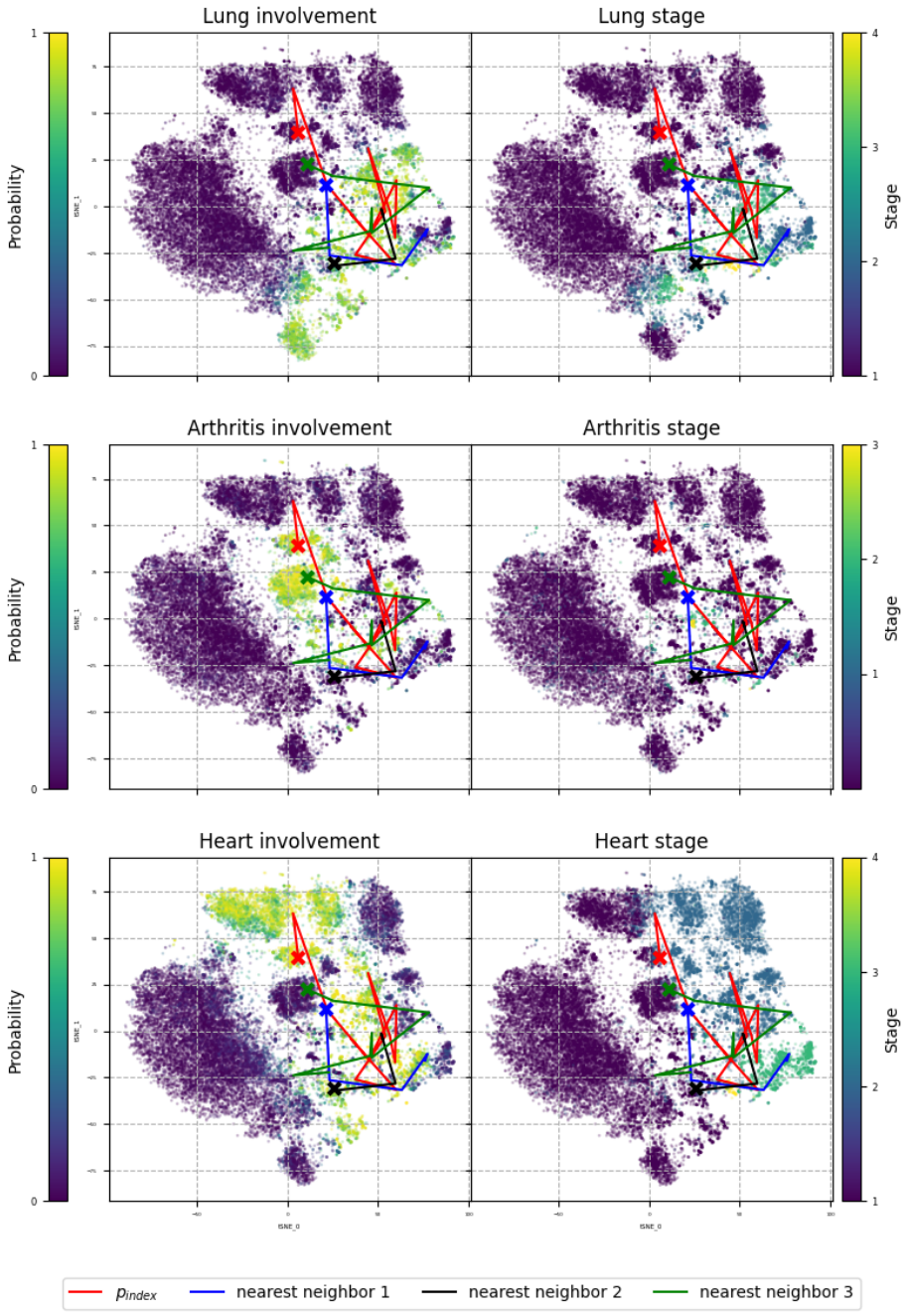
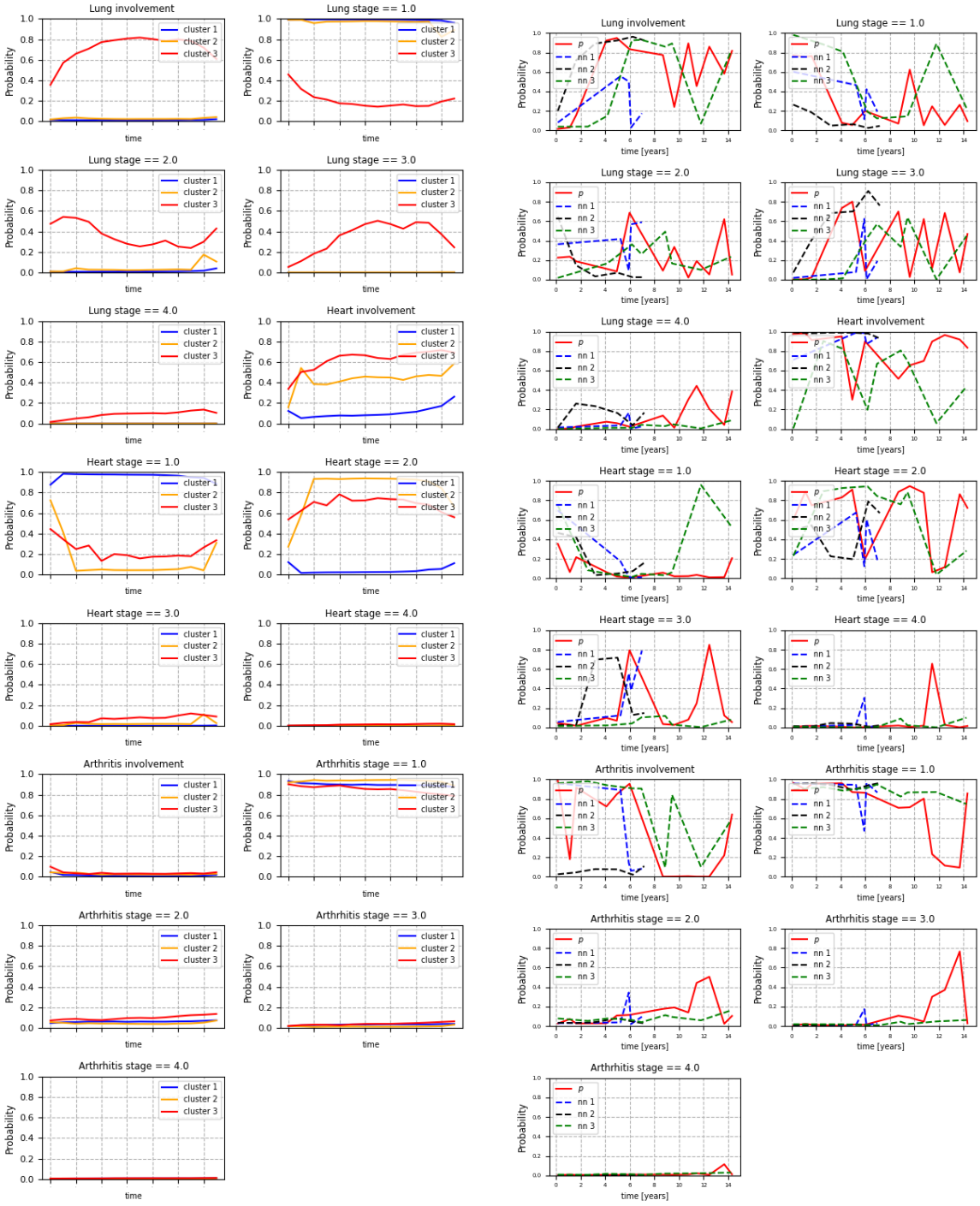


Figure 17: Trajectory of p_{idx} and their 3 nearest neighbors in the latent space.



(*a*) Medical label trajectories for cluster means.

(*b*) Medical label trajectories for p_{idx} and its 3 nearest neighbors.

Figure 18: Medical label trajectories.