Clustering Interval-Censored Time-Series for Disease Phenotyping

Irene Y. Chen[1], Rahul G. Krishnan[2], David Sontag[1]

[1]MIT CSAIL and IMES [2]University of Toronto
iychen@csail.mit.edu, rahulgk@cs.toronto.edu, dsontag@csail.mit.edu

Abstract

Unsupervised learning is often used to uncover clusters in data. However, different kinds of noise may impede the discovery of useful patterns from real-world time-series data. In this work, we focus on mitigating the interference of interval censoring in the task of clustering for disease phenotyping. We develop a deep generative, continuous-time model of time-series data that clusters time-series while correcting for censorship time. We provide conditions under which clusters and the amount of delayed entry may be identified from data under a noiseless model. On synthetic data, we demonstrate accurate, stable, and interpretable results that outperform several benchmarks. On real-world clinical datasets of heart failure and Parkinson's disease patients, we study how interval censoring can adversely affect the task of disease phenotyping. Our model corrects for this source of error and recovers known clinical subtypes.

Introduction

Cluster analysis of time-series data is a task of interest across a variety of scientific disciplines including biology (Listgarten et al. 2007), meteorology (Camargo et al. 2007), and astrophysics (Rebbapragada et al. 2009). Automating the discovery of latent patterns in real-world data can be challenging due to noise. We focus on mitigating errata in pattern discovery from interval censoring (Miller Jr 2011).

Interval censoring arises when time-series data are only observed within a known interval. The difference between an observed time value and the true time value, known as *left-censorship*, and the lack of a common outcome against which to align, known as *right-censorship*, can lead common techniques for clustering to erroneous conclusions about the underlying patterns. To address this, practitioners must often manually align data to a meaningful start time in order to find non-trivial groups via unsupervised clustering—a process whose difficulty can range from expensive and time-consuming in some problems to infeasible in others. In this work, we develop a machine learning algorithm that clusters time-series data while simultaneously correcting for interval censoring. In doing so, we automate the time-consuming process of manual data alignment and use our method to re-

veal structure that would otherwise not be found by straightforward application of clustering analysis.

This paper is motivated by disease phenotyping in healthcare. Many diseases are biologically heterogeneous despite a common diagnosis—for example autism (Doshi-Velez, Ge, and Kohane 2014), heart failure (Shah et al. 2015), diabetes (Udler et al. 2018), and Parkinson's disease (Fereshtehnejad et al. 2017). The variation in biomarkers (e.g., glucose or creatinine) across patients can stem from different patient subtypes that manifest in distinct disease trajectories. Scientists seek to understand this disease heterogeneity by identifying groups of people whose biomarkers behave similarly. For example, cardiologists use a measurement called ejection fraction as a heuristic to separate heart failure patients into two categories (Owan et al. 2006), with at least one of the two subtypes believed to be heterogeneous (Shah et al. 2015). To better understand patient heterogeneity, cardiologists may turn to longitudinal, observational, and often irregularly-measured patient data for disease phenotype discovery.

Interval censoring in healthcare data presents a significant challenge for disease phenotyping. Left-censorship in clinical datasets can occur when patients have *delayed entry*, meaning data are unavailable before a diagnosis or first hospital visit. Factors including geographic proximity to a hospital (Chan, Hart, and Goodman 2006), financial access to care (Miller and Wherry 2017) or mistrust of the healthcare system (Brandon, Isaac, and LaVeist 2005) can affect when a patient seeks medical help and consequently the beginning of data availability with respect to the underlying progression of their disease. Other datasets align patients by death time, but right-censorship restricts sample size to patients who have died. Since many factors affect mortality, other causes of death may confound results. For heart failure, a chronic disease that progresses over many years with a large range of onset ages and survival outcomes, interval censoring can confound attempts to analyse disease heterogeneity using observational data.

For a simplified illustration of the problem, see Figure 1(a) which depicts the common reality of observational health data whereas Figure 1(b) depicts the idealized latent substructure we would like to identify. Existing subtyping models applied to clinical data assume (potentially erroneously) that patients are aligned at entry into the dataset
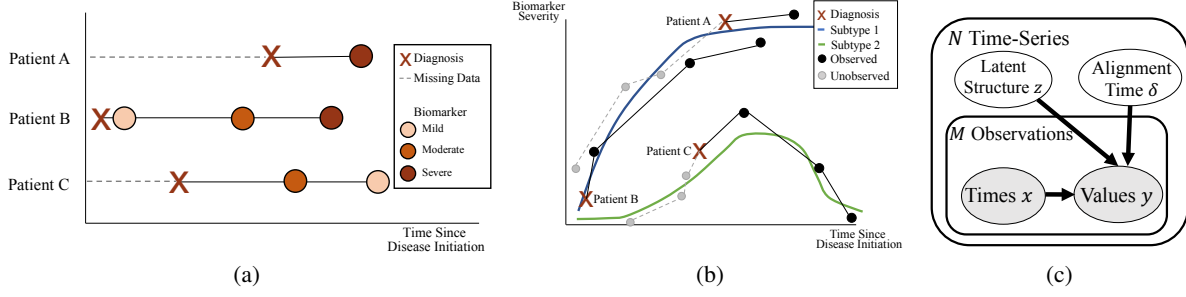
Figure 1: **(a)** Patient data can be interval-censored, meaning longitudinal data can be missing based on entry to the dataset, e.g., first diagnosis, and the data may lack a common outcome against which to align. Patients may enter the dataset at any stage of the disease. **(b)** Interval censoring can make clustering patient time-series data challenging because data may be aligned incorrectly, e.g., first diagnosis. We seek to understand disease heterogeneity by inferring subtypes after correcting for misalignment. **(c)** Graphical model of SubLign.

or study. The problem with such an assumption is that the disparity between true disease stage and observed observation time can result in unsupervised learning algorithms uncovering the wrong, or perhaps less interesting, structure. For example, a naive clustering algorithm might simply return clusters corresponding to the disease stage at entry into the study, which may simply recapitulate some of the biases mentioned earlier.

To that end, our paper makes the following contributions:

1. We introduce and formalize the problem of cluster recovery from interval-censored time-series data.

2. We introduce a practical algorithm, SubLign, based on variational learning of a deep generative model, which:
   - Makes no assumption on the distribution of delayed entry alignment values besides extrema.
   - Operates on multivariate time-series with varying lengths and missing values, both characteristics often found in real-world datasets.
   - Is unsupervised, meaning that neither subtype labels nor alignment values are provided during training.

3. We prove an identifiability result showing that, in a noiseless setting, both the degree of delayed entry from left-censorship and the subtype identity are recoverable.

We show robust quantitative results on synthetic data where, over multiple runs, our method outperforms baselines for subtyping and patient alignment. On two real-world clinical datasets—heart failure and Parkinson's disease—we automatically recover known clinical subtypes without manual alignment or expert knowledge.

Related Work

Learning alignment and clustering has been studied in fields across computer vision, signal processing, and health. Approaches often make assumptions including few discrete time steps (Young et al. 2018; Huopaniemi et al. 2014); a single piecewise linear function (Young et al. 2018) or Gaussian mixture model (Huopaniemi et al. 2014); significantly more samples per object than number of objects (Mattar, Hanson, and Learned-Miller 2012; Gaffney and Smyth

2005); very small windows of potential misalignment (Liu, Tong, and Wheeler 2009; Listgarten et al. 2007); or known lag time (Li et al. 2011). Methods that directly measure similarity between time-series, e.g., dynamic time warping (Cuturi 2011) or methods that aggregate multiple imputation methods (Faucheux et al. 2021) can also be used for clustering time-series data. Our method aims to cluster interval-censored multivariate time series without these constraints.

Clinicians and scientists learn disease subtypes to better understand heterogeneity in disease progression in a process known as disease phenotyping. Existing approaches often rely on the assumption that the observed measurements are aligned — and therefore not censored. Researchers then apply clustering techniques like hierarchical clustering of time series (Doshi-Velez, Ge, and Kohane 2014), affinity clustering (Luo et al. 2020), or matrix factorization (Udler et al. 2018; Perros et al. 2017). Other models define disease subtypes as stages of disease progression (Alaa and van der Schaar 2019). For this work, we define disease subtypes as distinct from disease stage and jointly learn both.

SubLign: Subtype & Align

There are two stages to SubLign. First, we learn a generative model of the observed data which disentangles variation in the observed data due to delayed entry from variation related to subtype identity. Second, we infer subtype representations and (optionally) cluster the representations to obtain the explicit subtype identity for each time-series. Figure 1(c) describes the graphical model, and Algorithm 1 depicts the pseudocode for this procedure.

SubLign Generative Model Consider the following setup. We observe N multivariate time-series (one for each patient), each of length up to M: $[(x_{1,1}, y_{1,1}), \ldots, (x_{1,M}, y_{1,M})], \ldots [(x_{N,1}, y_{N,1}), \ldots, (x_{N,M}, y_{N,M})]$. $y_{i,m} \in \mathbb{R}^D$ is a vector of observations for time-series i at time-stamp $x_{i,m} \in \mathbb{R}^+$. We denote collections of observations as $Y_i = \{y_{i,1}, \ldots, y_{i,M}\}$ and time-stamps as $X_i = \{x_{i,1}, \ldots, x_{i,M}\}$ for patient i.

Figure 1(c) depicts the graphical model corresponding

Algorithm 1: SubLign

1: **Input:** Observation times $X \in \mathbb{R}^{N \times M}$, biomarkers $Y \in \mathbb{R}^{N \times M \times D}$
2: **Output:** τ_k for each subtype and $\hat{\delta}_i$ for each patient
3: **Step** 1**: Learning**
4: **repeat**
5:     Encode time-series: $h_i = \text{RNN}([X_i, Y_i]) \ \forall i \in \{1, \ldots, N\}$
6:     Compute variational distribution $q(Z_i | X_i, Y_i) = \mathcal{N}(\mu(h_i; \phi_2), \Sigma(h_i; \phi_3)) \forall i \in \{1, \ldots, N\}$
7:     **for** patient $i = 1$ **to** N **do**
8:         Run grid-search to find $\hat{\delta}_i = \arg\max_{q(\delta_i)} \mathcal{L}(Y_i | X_i; \gamma, \phi, q(\delta_i))$ (Eq. 3)
9:     **end for**
10:     Update γ, ϕ via stochastic gradient ascent on $\mathcal{L}(Y | X; \gamma, \phi, \hat{\delta})$
11: **until** convergence
12: **Step** 2**: Inference and Clustering**
13: Infer $\mathcal{Z} = \{z_i | z_i = \mu(h_i; \phi_2)\}$ for X_i, Y_i
14: Find K clusters using k-means on $\mathcal{Z}$ and compute cluster centers μ_k
15: Infer parameters of subtype trajectories $\tau_k = g(\mu_k)$
---

to the latent-variable generative model of continuous-time multivariate data:

$$
\begin{aligned}
&\forall i = \{1, \ldots, N\}, \ \forall m \in \{1, \ldots, M\}, \\
&\forall d \in D_{i,m}, \ \delta_i \sim \text{Cat}(\mathcal{D}), z_i \sim \mathcal{N}(\mathbf{0}, \mathbb{I}), \Theta = g(z_i; \gamma), \\
&\overline{y}_{i,m}[d] = f(\kappa(x_{i,m} + \delta_i; \Theta[d])), \ y_{i,m} \sim \mathcal{N}(\overline{y}_{i,m}[d], \mathbf{I})
\end{aligned} \quad (1)
$$

We drop indices denoting patient and dimension where unnecessary. $D_{i,m}$ denotes the set (and $|D_{i,m}|$ denotes the number) of observed biomarkers for patient i at their m-th observation. To accommodate missing data, not all biomarkers are required to be measured at every observation. Each delayed entry value $\delta_i \in \mathbb{R}^+$ has a maximum alignment deviation value δ^+ over all time-series. We discretize the closed interval $[0, \delta^+]$ as $\mathcal{D} = [0, \epsilon, 2\epsilon, \ldots, \delta^+]$ with hyperparameter ϵ and use a categorical distribution over $\mathcal{D}$ with uniform probabilities over each element as our prior over δ_i. Function $g : \mathbb{R}^{N_z} \to \mathbb{R}^{D \times (P+1)}$ has parameters γ and maps from the latent variable to $\Theta \in \mathbb{R}^{D \times (P+1)}$, a matrix of parameters for D polynomials, each of degree P. f is a known link function that describes how observed values relate to observed time-points.

Parameterization We discuss the specific parameterizations of Equation 1. In the context of our motivating application of disease phenotyping, these functions represent common characteristics in the progression of patient biomarkers.

Link function and polynomials: For f and P, we study the following choices: *Sigmoid*: $P = 1$ and $f(x) = \frac{1}{1+\exp(-x)}$ and *Quadratic*: $P = 2$ and $f(x) = x$. The sigmoid function can represent bounded and monotonically increasing clinical variables. The quadratic function represents cases where disease severity, as measured by biomarkers, decreases (likely

in response to therapy) and then increases (once therapy fails), or vice versa. Other choices for P, f are permissible as long as they are differentiable with respect to the model parameters. We allow for the possibility that individual biomarkers have different parameterizations.

Modeling polynomial parameters: We parameterize $g(z_i; \gamma)$ using a two layer neural network with ReLU activation functions with parameters γ. To be concrete, if $D = 1, P = 2$, and f is the sigmoid function, then the outputs of g are $[\beta_0(z), \beta_1(z)]$ and $y = \frac{1}{1+\exp^{-(\beta_0(z)x + \beta_1(z))}}$. Similarly, if $D = 1, P = 1$, and f is the quadratic function then the outputs of g are $[a(z), b(z), c(z)]$ and $y = a(z)x^2 + b(z)x + c(z)$.

Step 1: Learning

We learn the parameters γ of the model in Equation 1 via maximum likelihood estimation. Since the model is a non-linear latent variable model, we maximize a variational lower bound on the conditional likelihood of data given the time-stamps corresponding to observations.

$$
\log \prod_{m=1}^{M} \prod_{d \in D_{i,m}} p(y_{i,m}[d] | x_{i,m}; \gamma) = \log p(Y_i | X_i; \gamma) \quad (2)
$$

$$
\geq \mathcal{L}(Y_i | X_i; \gamma, \phi)
$$

$$
= \mathbb{E}_{q(Z_i | X_i, Y_i; \phi)} \Big[\log \sum_{\delta_i} p(Y_i | X_i, \delta_i, Z_i; \gamma) p(\delta_i) +
$$

$$
+ \log \frac{p(Z_i)}{q(Z_i | X_i, Y_i; \phi)} \Big] \geq \mathcal{L}(Y_i | X_i; \gamma, \phi, q(\delta_i))
$$

$$
= \mathbb{E}_{q(Z_i | X_i, Y_i; \phi)} \Big[\mathbb{E}_{q(\delta_i)} \Big[\log p(Y_i | X_i, \delta_i, Z_i; \gamma)
$$

$$
+ \log \frac{p(\delta_i)}{q(\delta_i)} \Big] + \log \frac{p(Z_i)}{q(Z_i | X_i, Y_i; \phi)} \Big] \quad (3)
$$

The first lower bound uses a variational distribution for Z parameterized via an inference network (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) with parameters ϕ. The second lower bound is a variational distribution over δ. The function $q(\delta_i)$ parameterizes the space of one-hot distributions in $\mathcal{D}$, i.e., a categorical distribution over discrete choices of δ_i.

Our learning algorithm alternates between two steps. We first maximize the lower bound using subgradient ascent. To do so, we solve: $\hat{\delta}_i = \arg\max_{q(\delta_i)} \mathcal{L}(Y_i | X_i; \gamma, \phi, q(\delta_i))$. For our choice of variational distribution, this maximization can be performed via a grid search. We then derive gradients $\nabla_{\gamma, \phi} \mathcal{L}(Y_i | X_i; \gamma, \phi, \hat{\delta}_i)$ to update the generative model and inference network via stochastic gradient ascent.

Step 2: Obtaining Learned Subtypes

After learning the model, we may re-use the inference network to predict the latent variable $z_i = \mu(h_i; \phi_2)$ for each patient in the training set. Combining z_i across all time-series gives us the set $\mathcal{Z}$. When reasonable, we refer to $\mu(h_i; \phi_2)$ as μ_i.

Although latent variable z_i encodes latent structure from each time-series, we may be interested in explicit subtypes

for a given value of K. To obtain discrete subtypes, we can run clustering algorithms on $\mathcal{Z}$ to obtain K cluster centers $\{\mu_1, \ldots, \mu_K\}$. Because we use a Gaussian prior for our biomarker values, measuring distances in the space of the latent variable can be done with the Euclidean norm, making the k-means a reasonable choice of clustering algorithm.

We compute $\{\tau_1, \ldots, \tau_K\}$ where $\tau_k = g(\mu_k)$ as the progression-patterns corresponding to each of the discrete subtypes of the disease. For example, if $f \circ \kappa$ is linear, then we obtain K different slopes and biases, each of which describes how the time-series behaves in that subtype. In practice, K may be chosen based on domain knowledge; alternatively, qualitative results can be assessed for each version of K, e.g., by plotting the corresponding f functions.

Remarks on SubLign

Role of the latent variable: The latent variable z plays an important role in quantifying how each biomarker behaves. Each time-series's latent variable is used to predict the parameters of D polynomial functions and $f \circ \kappa$ maps from observation times onto the observed biomarkers. Time-series whose representation space z are *close* hail from the same subtype, and consequently manifest similar patterns in their biomarkers. This variation in z results in variation in the parameters Θ and therefore in variation among the data as a function of the time-points.

Note this latent representation can be used for other settings beyond disease phenotyping, for example supervised prediction tasks. For example, when $P = 1$ and f is the identity (i.e., a linear function), some time-series might increase (positive slope) versus others that decrease (negative slope). The latent representation z captures subtypes by learning to predict the slope of the function that models variation among time-series.

As an illustration, in Figure 1(b) in the blue phenotype, f is the sigmoid function. Depending on the latent space, we could imagine a one-dimensional z where $z < 0$ represents the curve (and subtype) in blue and $z \geq 0$ to represent the curve in red. The value of δ_i indicates the degree of delayed entry associated with each time-series. The delayed entry from interval censoring is corrected by applying the scalar δ_i element-wise to X_i and then transforming it by f.

Choice of link function: Our current choice of link function f is motivated to mimic degenerative disorders wherein patient biomarkers gradually increase over time (denoting worse outcomes). There is precedence in prior work to restrict function forms that characterize how biomarkers behave to be monotonic (Pierson et al. 2019). However, we emphasize that the model is not restricted to only sigmoid or quadratic functions for the link function f used in the synthetic and clinical experiments. Our work permits alternative choices of f—assuming the choice is smooth and differentiable with respect to model parameters—and allows individual biomarkers to have different parameterizations.

Scalability: The runtime of SubLign is impacted by the grid search over model parameter δ_i. The corresponding lines 7-9 in Algorithm 1 have complexity $\mathcal{O}(NSF)$ where F is the complexity associated with a single forward pass of the inference network and the generative model, N is the number of examples, and $S = \delta^+/\epsilon$ is the number of time steps. The model practitioners may therefore balance computational resources with S. In our experiments, we found comparable performance for S as low as 5.

Real-world clinical data: SubLign is motivated by, and designed to capture, variation in clinical biomarkers while taking into account the challenges of clinical data. Observational healthcare data are often irregularly spaced, and contains missingness. The use of a continuous time model allows us to naturally handle the former issue since we only maximize the likelihood of data corresponding to time-points where they are observed. When a single biomarker is missing while others are observed, it may be marginalized out (by ignoring the corresponding loss term).

Accommodating different kinds of censorship: Equation 1 naturally characterizes delayed entry arising from left-censorship. SubLign can also accommodate right-censorship by reversing the sequence of time-series and applying Algorithm 1 (resulting in our ability to infer the degree of right censorship). When both left-censorship and right-censorship are present, it corrects for left-censorship explicitly (using δ) while right-censorship is implicitly accounted for since we only maximize the likelihood of data up to the point that we observe time-series.

Identifiability under a noiseless model

While SubLign presents a viable, practical model for clustering and aligning censored time-series data, it is worth reflecting upon whether we can ever identify subtype and alignment from observational data. In what follows, we present theoretical conditions that show that there exists conditions under which the problem we study is identifiable.

Identifiability We assume distinct time stamps for the M observations in X_i. The generative process we assume for Y_i, conditional on X_i, is:

$$\forall i = \{1, \ldots, N\}, \ s_i \sim \text{Cat}(K),$$
$$\forall m \in \{1, \ldots, M\}, \ d \in D_{i,m},$$
$$y_{i,m}[d] = f(\kappa(x_{i,m} + \delta_i; \theta^P[s_i, d])) \quad (4)$$

where $s_i \in \{1, \ldots, K\}$ is the subtype for time-series i, $D_{i,m}$ denotes the set of all observations at time-step m for time-series i where $\forall i, m, |D_{i,m}| \leq D$, and δ_i is the delayed entry value. The link function $f : \mathbb{R} \to \mathbb{R}$ has no parameters whereas $\kappa : \mathbb{R} \to \mathbb{R}$ is an unknown polynomial function of degree $P \in \mathbb{Z}^+$. We denote the parameters of κ, for each subtype and dimension (e.g., biomarker), as $\theta^P \in \mathbb{R}^{K \times D \times (P+1)}$. We denote $\theta^P[s_i, d]$ as selecting the (s_i, d)-th vector of size $P + 1$ from the tensor θ^P. Similarly, $\theta^P[s_i]$ selects the s_i-th matrix of size $X \times (P+1)$. We define θ_p as the p-th coefficient of any polynomial function parameter set θ.

By construction, we have that $s_i = s_{i'} \iff \theta^P[s_i] = \theta^P[s_{i'}]$, i.e., for two subtypes, the values of each time-series are described either by $y = f(\kappa(x; \theta^P[s_i]))$ or $y = f(\kappa(x; \theta^P[s_{i'}]))$.

We begin with a set of assumptions for identifiability.

Assumption 1. *f is invertible, and $\kappa(x, \theta) = \theta_0 + \sum_{p=1}^{P} \theta_p x^p$ describes a family of polynomial functions in x with parameter θ and degree $P > 0$. The parameters of each subtype are unique.*

Assumption 2. *$M \geq P + 1$, i.e., for each patient time-series, there exists at least one of the D features where we observe at least $P + 1$ values.*

Assumption 3. *For each subtype s_k, there exists a time-series i whose alignment $\delta_i = 0$, meaning no delayed entry.*

Theorem 1. *Under assumptions 1, 2, 3 for the model in Equation 4, we can identify the time-delays $\delta_1, \ldots, \delta_N$. We can identify the polynomial coefficients θ^P up-to a permutation of its rows and columns and the identity of $s_1, \ldots, s_N$ up-to a permutation over K choices.*

Proof sketch. Consider the case where we have a single biomarker for each patient. The proof is constructive; first we transform the data using the inverse of f resulting in a set of data drawn from polynomial equations. The polynomial coefficients may be estimated from the observed data; we can then find the roots of these polynomials and pick the smallest root. These roots exactly quantify the degree of delayed entry; i.e. they tell us how much each polynomial has been shifted by. We can correct each time-series for this shift, re-estimate polynomial coefficients from the shifted-polynomials and cluster them to reveal the underlying subtype identity for each time-series (and consequently each patient).

Proof. The proof is constructive; i.e. we give an algorithm for the identification of the parameters of the model in Equation 4. The algorithm for identification is presented in Algorithm 2 and proceeds in three steps.

Step 1: The first step transforms the observed biomarkers by applying the inverse of function f, which exists by Assumption 1. This leaves us with data as:

$$f^{-1}(y_{i,m}) = \kappa(x_{i,m} + \delta_i; \theta^P_{s_i}) \ \ \forall i \in N, m \in M$$

i.e. for all bio-markers, across all patients, we have data arising from different polynomial functions.

Step 2: Without loss of generality, the second step uses the first biomarker to identify the values of δ_i for each patient.

a) First, we estimate the polynomial coefficients for each patient separately; we are guaranteed exact recovery of the coefficients by Assumption 2.

b) Next we find the roots for each polynomial. If they are complex, consider their real part, and define ξ_i to be the smallest root of the polynomial. At least one (real or complex) root is guaranteed to exist by the Fundamental Theorem of Algebra for every non-constant polynomial (Assumption 1). Note that the choice of using the smallest root is arbitrary; what matters is that a consistent choice of root is selected for each patient's polynomials.

c) The goal of this step to learn a new polynomial for each patient which is translated to ensure that the root selected in step b) lies at $x = 0$.

Algorithm 2: Procedure for the identification of model parameters

1: **Input:** Observation times $X \in \mathbb{R}^{N \times M}$, biomarkers $Y \in \mathbb{R}^{N \times M \times D}$, polynomial degree P, invertible function f
2: **Output:** $\theta^P, \delta_1, \ldots, \delta_N, s_1, \ldots, s_N$ for each patient
3: **Step 1: Transform the observed biomarkers;** $Q = f^{-1}(Y)$
4: **Step 2: Obtain time-shifts using a single biomarker;**
5: a) For each patient i, estimate the parameters $\dot{\theta^1}_i$ of $\kappa(x; \dot{\theta^1}_i)$ using a single biomarker $((x_{i,1}, q_{i,1}), \ldots, (x_{i,M}, q_{i,M}))$ via polynomial regression,
6: b) Compute up to P roots of polynomial $\kappa(x, \dot{\theta^1}_i)$ for each patient i as $R_i = \{r_1, \ldots, r_P\}$ and set $\xi_i = \min \text{Real}(R_i)$ where Real denotes the real part of (potentially complex) roots.
7: c) Estimate $\tilde{\theta^1}_i$ for polynomials in a *canonical position* using $((x_{i,1} - \xi_i, q_{i,1}), \ldots, (x_{i,M} - \xi_i, q_{i,M}))$ via polynomial regression,
8: d) Cluster $\tilde{\theta^j}_i$ across patients via K-means clustering to yield cluster identities $s_1, \ldots, s_N$
9: e) $\forall k, \quad \eta_k = \min\{\xi_i \,|\, i \text{ s.t. } s_i = k\}$ and $\forall i, \quad \delta_i = \xi_i - \eta_{s_i}$
10: **Step 3: Estimate true polynomial coefficients using shifted observation times;**
11: **for** biomarker $j = 1$ **to** J **do**
12: For each patient, estimate the parameters θ^j_i of $\kappa(x; \hat{\theta^j}_i)$ using $((x_{1,1} - \delta_i, q_{1,1}[j]), \ldots, (x_{1,M} - \delta_i, q_{1,M}[j])$ via polynomial regression,
13: **end for**
14: Return $\theta^P = [\theta^1 | \ldots | \theta^J], \{\delta_1, \ldots, \delta_N\}, \{s_1, \ldots, s_N\}$

To do so, we first shift the observational time-steps by ξ_i, and we re-estimate the coefficients of each *shifted* polynomial.

We make use of the fact that if ξ_i is the smallest complex root of a polynomial $\kappa(x)$ then the polynomial $\kappa(x + \xi_i)$ has its smallest complex root at 0. We can recover the parameters of this polynomial exactly by shifting our observations and re-estimating the coefficients.

This operation recovers the coefficients of every patient's polynomial in its *canonical position* i.e. a translated polynomial whose the smallest root (or its real component) is at $x = 0$.

This step can be viewed as a de-biasing step which allows us to re-estimate $\tilde{\theta}$ without while ignoring the effect that left-censorship has on parameter estimates.

d) We cluster the coefficients estimated in step c). By construction, we know that $s_i = s_{i'} \iff \theta_i = \theta_{i'}$ which guarantees that clustering recovers the true-underlying subtype for each patient (up to a permutation over K choices).

e) Finally we stratify patients by their subtype, and we define δ_i as the difference between their smallest root and the smallest value of ξ_i among all other patients within

that subtype.

By Assumption 3, we know that for each subtype, there exists a patient for whom $\delta_i = 0$, this reference patient will also be the one whose polynomial has the smallest root. We note here that without Assumption 3, we would still have identification of δ_i up to a constant.

Therefore, by shifting each patient's smallest root by their reference patient's smallest root, we can recover the original time-shifts.

Step 3: Given the values of $\delta_1, \ldots, \delta_N$ from Step 2, we can now estimate the true values of the polynomial coefficients exactly in the noiseless setting via polynomial regression. □

Remarks: Theorem 1 describes conditions under which delayed entry and the polynomial parameters of cluster biomarker progression are identifiable. This encouraging result demonstrates scenarios where the parameters of the model in Equation 4 can provably be identified.

On the assumptions for identification: It is possible to relax Assumption 3 to only require the existence of a single time-series from each subtype; this modification only allows identifiability of $\delta_1, \ldots, \delta_N$ up to a translation within each subtype. The above result relies on the existence of at least one biomarker for which there are sufficiently many observations – this is a reasonable assumption in the context of clinical data since there is often a *canonical* biomarker tracked over time for each disease.

On the strategy for identification: We conjecture our analysis for identification of subtype and alignment is of independent interest for identifying causal effects in survival analysis where an important challenge is how to handle confounding that jointly affects both survival time and censorship. Related work in this field (Seaman et al. 2020; Choi and O'Malley 2017) has focused on restricting the class of models used to characterize the survival function. Our work presents distinct parameteric assumptions towards this goal.

Evaluation Setup

Synthetic data. We generate two classes of synthetic datasets from the *sigmoid* and *quadratic* parameterizations in the 'SubLign: Subtype & Align' section. For the sigmoid dataset, we generate data from $K = 2$ subtypes and $\forall i, m, |D_{i,m}| = 3$ biomarker dimensions. For the quadratic dataset, we generate data with $K = 2$ and $\forall i, m, |D_{i,m}| = 1$. See appendix for the data generation process for the six quadratic datasets.

In both synthetic settings, we sample $N = 1000$ patients with $M = 4$ observations, variance $\sigma^2 = 0.25$, and max disease stage $T^+ = 10$. For each patient, we sample subtype $s \sim \text{Bern}(0.5)$. The true disease stage is drawn $t_m \sim \text{Unif}(0, T^+)$ for observation $m \in \{1, \ldots, M\}$. The biomarker values are drawn $y_m \sim N(\lambda_m, \sigma^2)$ where $\lambda_m = \sum_{k \in \{1, \ldots, K\}} \mathbb{1}(s_i = k) f_k(t_m)$. For the sigmoid dataset, the first subtype generating function across three dimensions is $f_1(t) = [\sigma(-4 + t), \sigma(-1 + t), \sigma(-8 + 8t)]$ and the second subtype generating function is $f_2(t) = [\sigma(-1 + t), \sigma(-8 + 8t), \sigma(-25 + 3.5t)]$. The observed disease time x_m is shifted

such that the first patient observation is at time 0. Therefore $x_m = t_m - \zeta$ where $\zeta = \min_{j \in \{1, \ldots, M\}} t_j$ is the earliest true disease time for the patient.

Real-world clinical data. We study two real-world clinical datasets (Parkinson's disease and heart failure) to assess SubLign performance when data have realistic dynamics and are missing over time. Our clinical datasets inherently contain delayed entry due to external factors including patient access to healthcare. We include preprocessing code in the appendix.

Heart failure (HF): We use electronic health records from a large health system in the United States, currently redacted for double blind review. We identify patients who enter the emergency department with a diagnosis of HF and extract echocardiogram values—measurements from an ultrasound of the heart—from the full patient history. We include echocardiogram features that are present in more than 60% of echo studies. Our dataset includes $N = 1534$ patients and $|D_{i,m}| \leq 12; \forall i, m$ features with $M = 38$ maximum observations per patient over a potential span of 10 years in the dataset. We extract 27 baseline features including race, sex, and comorbidities (e.g., renal failure) to validate subtypes only and not for use in the model. The dataset values are linearly scaled such that values are between 0 and 1 with larger values denoting more abnormality.

Parkinson's disease (PD): We use publicly-available data from the Parkinson's Progression Markers Initiative (PPMI), an observational clinical study, totalling $N_t = 423$ PD patients and $N_c = 196$ healthy controls where $N = N_t + N_c$. We extract four biomarker measurements of autonomic, motor, non-motor, and cognitive ability from $N = 619$ total participants with $M = 17$ maximum observations per patient. Our baseline data include 25 features including demographic information and patient history used to validate subtypes. For PD patients, the first recorded visit is within 2 years of the patient's PD diagnosis. Measurements are scaled between 0 and 1 with larger values corresponding to more abnormal values. We use the sigmoid parameterization of the SubLign model for both datasets because HF and PD are chronic and incurable diseases.

Hyperparameters and Baselines

We find optimal hyperparameters via grid search. For both synthetic and clinical experiments, we search over hyperparameters including dimensions of the latent space z (2, 5, 10), the number of hidden units in the RNN (50, 100, 200), the number of hidden units in the multi-layer perceptron (50, 100, 200), the learning rate (0.001, 0.01, 0.1, 1.), regularization parameter (0., 0.1, 1.), and regularization type (L1, L2). We select the hyperparameter configuration with the best validation loss, as measured by Equation 3. Our models are implemented in Python 3.7 using PyTorch (Paszke et al. 2019) and are learned via Adam (Kingma and Ba 2014) on a single NVIDIA k80 GPU for 1000 epochs. We set alignment extrema $\delta^+ = 10$ based on the maximum of the synthetic dataset and the maxima of the HF and PD datasets. We search over 50 time steps with $\epsilon = 0.1$.

For all models, we run for 1000 epochs and use the model

with the best training loss over the 1000 epochs for evaluation. For the sigmoid dataset, the optimal hyperparameters are latent space of dimension 5, 100 hidden units in the RNN, 50 hidden units in the multi-layer perceptron, learning rate of 0.01, and no regularization.

For the Parkinson's disease dataset, we searched on a slightly smaller set of hyperparameters for SubLign and found optimal hyperparameters of $\beta = 0.01$, no regularization, 10 latent dimensions, 10 hidden units for the multi-layer perceptron, 200 units for the recurrent neural network, and learning rate of 0.1.

For the heart failure dataset, we searched on a slightly smaller set of hyperparameters for SubLign and found optimal hyperparameters of $\beta = 0.001$, no regularization, 10 latent dimensions, 20 hidden units for the multi-layer perceptron, 50 units for the recurrent neural network, and learning rate of 0.01.

We compare to seven different baselines. Our greedy baseline, denoted as KMeans+Loss, first clusters the observed values using k-means clustering. Then, using the inferred labels s, we simultaneously learn θ_k for each subtype and δ_i for each patient by minimizing:

$$\arg\min_{\theta,\delta} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{d=1}^{D} \sum_{k=1}^{K} \mathbb{1}[s_i = k][y_{i,j,d} - f(x_{i,j,d} + \delta_i; \theta_k)]^2$$

(5)

using Broyden–Fletcher–Goldfarb–Shanno. This naive clustering based approach (in the space of the original data) attempts to correct for shifts in time. We also compare to:

1. SubNoLign: a modified SubLign with no alignment value. This model is comparable to Zhang et al. (2019), which learns a deep patient representation while controlling for model architecture

2. SuStaIn (Young et al. 2018): a subtype and stage inference algorithm for disease progression,

3. BayLong: a Bayesian model of longitudinal clinical data (Huopaniemi et al. 2014),

4. PAGA (Wolf et al. 2019): a state-of-the-art single-cell trajectory pseudo-time method,

5. Clustering with dynamic time warping (DTW) using kernel methods (Cuturi 2011) and soft-DTW (Cuturi and Blondel 2017),

6. SPARTan: A tensor factorization based approach for phenotyping from time-series data (Perros et al. 2017)

We detail baseline implementations below.

SuStaIn. SuStaIn (Young et al. 2018) is a disease progression algorithm that recovers subtype and stage from cross-sectional data. We transform our longitudinal data by dropping patient affiliation across visits. We transform the data by subtracting the mean for each feature and dividing by the standard deviation for each feature. We assume the Z-scored values have a max of 5. We run for 1,000,000 epochs for the Markov Chain Monte Carlo sampling and 1,000 epochs for optimization. We use an open source implementation by the authors: `https://github.com/ucl-pond/pySuStaIn`

Bayesian approach. The Bayesian approach (Huopaniemi et al. 2014) assumes longitudinal data, but there must be a small number of measured time points. We assume that there are 10 observed time points where observed data can begin as well as a window of 10 time points before the observed window where a patient's values can be aligned to. Because biomarker values are scaled between 0 and 10, we assume that values change between time points based on a Gaussian with $\sigma = 2$ and that subtype means for each time point are drawn from a Gaussian with $\sigma = 5$. We draw 4000 samples and use the maximum a posteriori estimate to determine stage and subtype for test patients. Because we could not find an open-source option, we implemented the algorithm ourselves based on the description in the paper.

PAGA. Partition-based graph abstraction, or PAGA, (Wolf et al. 2019) assumes cross-sectional data, so we create separate visits for each patient visit. For algorithm parameters, we set resolution to 0.05, number of neighbors to 15, and connectivity cutoff of 0.05. We use an open source implementation by the authors: `https://github.com/dynverse/ti_paga/blob/master/run.py`

Dynamic time warping. Dynamic time warping (DTW) defines similarity between time series that can be combined with clustering techniques. DTW methods include using soft-DTW (Cuturi and Blondel 2017) and kernel (Cuturi 2011) before using K-means with the chosen similarity metric. We use open source implementations of DTW algorithms to generate our baseline comparisons: `https://pypi.org/project/dtw-python/`

Tensor factorization. Sparse tensor factorization has been used for disease phenotyping. The decomposition of large and sparse datasets using canonical polyadic decomposition can create an interpretable output for phenotyping. We use the Matlab open source implementation of SPARTan: `https://github.com/kperros/SPARTan` We found these baseline results to yield poor clustering performance despite aggressive hyperparameter tuning. We surmise this is because transforming our data from continuous to discrete time resulted in a very large and extremely sparse matrix factorizing which is a tricky optimization problem.

Evaluation

We evaluate models on 5 trials, each with a different randomized data split and random seed. For each trial, we learn on a training set (60%), find the best performance across all hyperparameters on the validation set (20%), and report the performance metrics on the held-out test set (20%). The same data folds are used across all models for each trial.

We report the performance on the test set over three metrics. *Adjusted Rand index (ARI)* measures whether pairs of samples are correctly assigned in the same or different subtypes (Hubert and Arabie 1985).

The *Swaps metric* reports the number of swaps needed to sort the predicted disease times into the true disease stages, expressed as percentage of total possible swaps. For

true sorted alignment values $a_1, \ldots, a_N$ for N patients, we define the swaps metric $\mathcal{S}$ of proposed alignment values $b_1, \ldots, b_N$ as the number of swaps needed to sort the predicted disease times into the true disease stages, expressed as percentage of total possible swaps.

$$\mathcal{S} = \frac{\sum_{i,j;i<j} \mathbb{1}(a_i < b_i, a_j > b_j)}{N(N-1)/2}$$

The *Pearson correlation coefficient* expresses correlation between the predicted and true disease stage. ARI measures the clustering performance while the Swaps metric and the Pearson correlation coefficient quantify how well the learning algorithm infers the alignment values.

Quantitative metrics on clinical datasets. Because real world data often lack ground truth labels for subtype or alignment, we create two semi-synthetic experiments with clinical datasets. For HF, we evaluate SubLign's ability to infer relative disease stage by introducing additional censoring into the test sets. Specifically, we train SubLign using 80% data (train and validation data) as usual. We then modify the remaining data (20%) by removing the first year of patient observations, creating distorted test set (X', Y'), and by removing the last year of patient observations, creating (X'', Y''). The same amounts of observations are removed from each set to control for length of observations. We infer alignment values using the trained SubLign model: δ' from (X', Y') and δ'' from (X'', Y''). By construction, $\delta' > \delta''$. We report the percentage of patients for which SubLign is able to recover this relationship. For PD, we report the held-out clustering performance for healthy control patients and patients with PD. We use disease status (PD patient or healthy control) as labels and $K = 2$ subtypes.

Missing values. SubLign allows for missing biomarker dimensions and missing patient visits to accommodate the sparsity of clinical data. For missing visits, we adapt the recognition network to handle variable sequence lengths. We mask out missing observations so they have no contribution to the learning stage, except for the recognition network input. We linearly interpolate missing values for each patient only for recognition network input. For baselines that cannot handle missing data, we linearly interpolate missing values for each patient.

Statistical significance. We evaluate held-out performance over 5 trials. Each trial consists of randomized 60/20/20 training/validation/test data folds and a different random seed. In order to compare models across 5 trials, we report the means and standard deviations from the 5 trials. When the reported performance intervals overlap, we compute the statistical significance of the pairwise differences using a t-test and a Benjamini-Hochberg correction of 0.05.

Evaluation and Analysis
Recovering Subtypes with Interval Censoring

SubLign is able to recover subtypes despite interval censoring, outperforming all baselines. For sigmoid synthetic data (Figure 2(a), ARI column), SubLign can recover subtypes (mean ARI of 0.94) better than the KMeans+Loss baseline (0.67) which assumes a greedy approach. Not correcting for alignment time decreases the quality of inferred subtypes as can be seen in SubNoLign (0.81).

Similarly, SubLign recovers known subtypes in the PD dataset with statistically significantly higher ARI over baselines (see Figure 2(b)). When ARI performance intervals overlap, we use a t-test on pairwise differences over trials to compute statistical significance.

Some baselines appear to suffer because SubLign leverages the longitudinal nature of patient data compared to the cross-sectional assumptions of PAGA and SuStaIn. Other baselines have strong priors, i.e., (Huopaniemi et al. 2014), which may explain its poor performance. Dynamic time warping methods appear to perform poorly for datasets with few observations or high missingness rates. Lastly, the tensor factorization method (Perros et al. 2017) fails, likely because transforming data from continuous to discrete time results in a very large and extremely sparse matrix factorization that is a tricky optimization problem.

We include additional results in the appendix, including visualizations of the SubLign subtypes compared to baselines, model misspecification analysis, and experiments varying the level of missingness.

Recovering Known Alignment Values

For the synthetic sigmoid data, SubLign outperforms baselines in inferring alignment values (Figure 2(a), Swaps and Pearson columns). SubLign recovers alignment values better according to the Swaps metric (mean value of 0.09) and the Pearson metric (mean value of 0.85) compared to the next best baselines of KMeans+Loss and SuStaIn. Note that many baselines only recover subtypes and do not learn patient alignment values. Although the real-world clinical datasets do not contain true alignment values, we use the previously described HF setup with an artificially censored test set. We find that SubLign predicts known alignment relationships in an altered dataset. When evaluated on the manipulated test data, SubLign recovers the constructed relationship of $\delta' > \delta''$ with a higher performance ($71\% \pm 2\%$) over five trials compared to K-Means+Loss ($57.8 \pm 4\%$) and SuStaIn ($53.8 \pm 3\%$). See appendix for full results.

Clinical Insights from Correcting for Misalignment

We validate SubLign subtypes learned from the HF and PD datasets using known clinical findings. The appendix includes a table of baseline features, which are not included as input to SubLign, with statistically significant differences in subtypes: 7 features (out of 26) for PD and 11 features (out of 27) for HF.

Heart failure. Cardiologists classify patients into two groups based on ejection fraction: HF with reduced ejection fraction (systolic HF) and HF with preserved ejection fraction (diastolic HF). However, in our HF dataset, over 30% of patients correspond to neither group based on clinical diagnosis. When evaluating SubLign, we set $K = 3$, one more than the number of known groups, to explore a new subtype

MODEL	ARI ↑	SWAPS ↓	PEARSON ↑		MODEL	ARI
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**		**SubLign**	**0.58 ± 0.12**
SubNoLign	0.81 ± 0.21	–	–		SubNoLign	0.42 ± 0.14
KMeans+Loss	0.67 ± 0.04	0.21 ± 0.03	0.49 ± 0.01		KMeans+Loss	0.05 ± 0.04
SuStaIn	0.66 ± 0.02	0.16 ± 0.00	0.30 ± 0.02		SuStaIn	0.12 ± 0.11
BayLong	0.19 ± 0.18	0.48 ± 0.00	0.01 ± 0.02		BayLong	0.04 ± 0.17
PAGA	0.32 ± 0.05	0.52 ± 0.07	0.04 ± 0.20		PAGA	0.02 ± 0.02
Soft-DTW	0.06 ± 0.01	–	–		Soft-DTW	0.46 ± 0.43
Kernel-DTW	0.06 ± 0.07	–	–		Kernel-DTW	0.21 ± 0.36
SPARTan	0.22 ± 0.18	–	–		SPARTan	0.15 ± 0.10
(a)					(b)	

Figure 2: Means and standard deviations over 5 trials for: **(a)** synthetic sigmoid dataset with 1000 patients, 3 dimensions, and 4 observations per patient, **(b)** 619 patients in the PPMI dataset including 423 Parkinson's disease patients and 196 healthy controls. Baseline methods include SuStaIn (Young et al. 2018), BayLong (Huopaniemi et al. 2014), PAGA (Wolf et al. 2019), Soft-DTW (Cuturi 2011), Kernel-DTW (Dhillon, Guan, and Kulis 2004), and SPARTan (Perros et al. 2017).

and to potentially better understand heterogeneity in the existing ejection fraction classifications (Shah et al. 2015).

Without ground truth subtype labels, we observe that SubLign finds systolic HF and diastolic HF as statistically significant baseline features. Of the three subtypes, subtype C corresponds to systolic HF, and subtype A and B correspond to diastolic HF, mirroring known clinical heterogeneity in diastolic HF (Shah et al. 2015). Of the two diastolic HF subtypes, subtype A has a higher proportion of women while subtype B has a higher rate of obese patients, both subgroups with documented heterogeneity in diastolic HF (Duca et al. 2018; Tadic and Cuspidi 2019). In contrast, the subtypes found by the KMeans+Loss baseline do not include known systolic HF and diastolic HF as statistically significant features. See appendix for full results.

Parkinson's disease. Biomarkers used to track PD are self-reported, which can be biased, subjective, and noisy. From these biomarkers, SubLign discovers subtypes that match known clinical findings on two cohorts: 619 combined PD and healthy control patients, and 423 PD patients.

For PD and healthy control patients, we run SubLign with $K = 2$ to uncover characteristics of the two known groups. SubLign subtype A clearly corresponds to healthy controls whereas subtype B designates PD patients. Statistically significant baseline features include all components of the University of Pennsylvania Smell Identification Test (UPSIT), which is a measure of smell dysfunction and highly linked to PD (Haehner et al. 2009), and having a full sibling or biological dad with PD, which aligns with research suggesting PD may be hereditary (Klein and Westenberger 2012).

For PD patients only, we set $K = 3$ in SubLign to explore and discover potential disease heterogeneity. Statistically significant baseline features include race and gender, which parallel recent clinical findings about heterogeneity in PD manifestation (Taylor, Cook, and Counsell 2007; Duca et al. 2018) and indicate new potential areas for future work. See appendix for tables of statistically significant baseline features stratified by the discovered subtypes for HF and PD.

Discussion

We study the task of clustering interval-censored time-series data. We present our method, SubLign, to learn latent representations of disease progression that correct for temporal misalignment in real-world observations and consider conditions for identifiability of subtype and alignment values. Empirically, our method outperforms seven baselines, and analysis of subtypes reveals clinically plausible findings. Better modeling of disease heterogeneity through alignment can help clinicians and scientists to better understand and predict how chronic diseases with many subtypes may progress. We hope that our model—in learning a continuous latent space to model heterogeneity—may be applied to other domains where subtypes and temporal alignment are entangled, for example gene expression analysis (Bar-Joseph et al. 2002) or cancer pathways (Wu, Gao, and Kasabov 2015).

Our model introduces directions for future work. Practically, SubLign assumes that δ and z are marginally independent. Intuitively, this means that, across all subtypes of a disease, the time at which the patient enters the data is independent of any other factor. There are certainly cases where this assumption may easily be violated, and therefore it remains an important area for further improvement e.g., (Kariyawasan et al. 2007).

One area of interest is how to incorporate conditioning δ on features that are predictive to its value (e.g., clinical history) or other complexities, such as differences in treatment effect (Athey 2015). Our results for theoretical and noiseless identification do not naturally extend to the noisy setting since root finding without further assumptions can be sensitive to noise in the coefficients of the polynomials. Alternative strategies for identification that generalize our result are fertile ground for future work.

More broadly, this work contributes to developing clinical models that are robust to real-world factors (Ghassemi et al. 2019), biases that might affect patient interactions with the healthcare system (Chen et al. 2020), and the practicalities of clinical care (Pollard et al. 2019).

Ethical Statement

Acknowledgements

References

Alaa, A. M.; and van der Schaar, M. 2019. Attentive state-space modeling of disease progression. In *Advances in Neural Information Processing Systems*, 11334–11344.

Athey, S. 2015. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 5–6.

Bar-Joseph, Z.; Gerber, G.; Gifford, D. K.; Jaakkola, T. S.; and Simon, I. 2002. A new approach to analyzing gene expression time series data. In *Proceedings of the sixth annual international conference on Computational biology*, 39–48.

Brandon, D. T.; Isaac, L. A.; and LaVeist, T. A. 2005. The legacy of Tuskegee and trust in medical care: is Tuskegee responsible for race differences in mistrust of medical care? *Journal of the National Medical Association*, 97(7): 951.

Camargo, S. J.; Robertson, A. W.; Gaffney, S. J.; Smyth, P.; and Ghil, M. 2007. Cluster analysis of typhoon tracks. Part I: General properties. *Journal of Climate*, 20(14): 3635–3653.

Chan, L.; Hart, L. G.; and Goodman, D. C. 2006. Geographic access to health care for rural Medicare beneficiaries. *The Journal of Rural Health*, 22(2): 140–146.

Chen, I. Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; and Ghassemi, M. 2020. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, 4.

Choi, J.; and O'Malley, A. J. 2017. Estimating the causal effect of treatment in observational studies with survival time endpoints and unmeasured confounding. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 66(1): 159.

Cuturi, M. 2011. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 929–936.

Cuturi, M.; and Blondel, M. 2017. Soft-DTW: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*.

Dhillon, I. S.; Guan, Y.; and Kulis, B. 2004. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 551–556.

Doshi-Velez, F.; Ge, Y.; and Kohane, I. 2014. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, 133(1).

Duca, F.; Zotter-Tufaro, C.; Kammerlander, A. A.; Aschauer, S.; Binder, C.; Mascherbauer, J.; and Bonderman, D. 2018. Gender-related differences in heart failure with preserved ejection fraction. *Scientific reports*, 8(1): 1–9.

Faucheux, L.; Resche-Rigon, M.; Curis, E.; Soumelis, V.; and Chevret, S. 2021. Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures. *Biometrical Journal*, 63(2): 372–393.

Fereshtehnejad, S.-M.; Zeighami, Y.; Dagher, A.; and Postuma, R. B. 2017. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain*, 140(7): 1959–1976.

Gaffney, S. J.; and Smyth, P. 2005. Joint probabilistic curve clustering and alignment. In *Advances in neural information processing systems*, 473–480.

Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A. L.; Chen, I. Y.; and Ranganath, R. 2019. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1(4): e157–e159.

Haehner, A.; Boesveldt, S.; Berendse, H.; Mackay-Sim, A.; Fleischmann, J.; Silburn, P.; Johnston, A.; Mellick, G.; Herting, B.; Reichmann, H.; et al. 2009. Prevalence of smell loss in Parkinson's disease–a multicenter study. *Parkinsonism & related disorders*, 15(7): 490–494.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of classification*, 2(1): 193–218.

Huopaniemi, I.; Nadkarni, G.; Nadukuru, R.; Lotay, V.; Ellis, S.; Gottesman, O.; and Bottinger, E. P. 2014. Disease progression subtype discovery from longitudinal EMR data with a majority of missing values and unknown initial time points. In *AMIA Annual Symposium Proceedings*, volume 2014, 709. American Medical Informatics Association.

Kariyawasan, C.; Hughes, D.; Jayatillake, M.; and Mehta, A. 2007. Multiple myeloma: causes and consequences of delay in diagnosis. *QJM: An International Journal of Medicine*, 100(10): 635–640.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *The 2nd International Conference on Learning Representations (ICLR)*.

Klein, C.; and Westenberger, A. 2012. Genetics of Parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(1): a008888.

Li, G.; Liu, B.; Qin, S. J.; and Zhou, D. 2011. Dynamic latent variable modeling for statistical process monitoring. In *Proc. IFAC World Congress, Milan, Italy*, 12886–12891.

Listgarten, J.; Neal, R. M.; Roweis, S. T.; Puckrin, R.; and Cutler, S. 2007. Bayesian detection of infrequent differences in sets of time series with shared structure. In *Advances in neural information processing systems*, 905–912.

Liu, X.; Tong, Y.; and Wheeler, F. W. 2009. Simultaneous alignment and clustering for an image ensemble. In *2009 IEEE 12th International Conference on Computer Vision*, 1327–1334. IEEE.

Luo, Y.; Eran, A.; Palmer, N.; Avillach, P.; Levy-Moonshine, A.; Szolovits, P.; and Kohane, I. S. 2020. A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nature Medicine*, 26(9): 1375–1379.

Mattar, M. A.; Hanson, A. R.; and Learned-Miller, E. G. 2012. Unsupervised joint alignment and clustering using Bayesian nonparametrics. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 584–593.

Miller, S.; and Wherry, L. R. 2017. Health and access to care during the first 2 years of the ACA Medicaid expansions. *New England Journal of Medicine*, 376(10): 947–956.

Miller Jr, R. G. 2011. *Survival analysis*, volume 66. John Wiley & Sons.

Owan, T. E.; Hodge, D. O.; Herges, R. M.; Jacobsen, S. J.; Roger, V. L.; and Redfield, M. M. 2006. Trends in prevalence and outcome of heart failure with preserved ejection fraction. *New England Journal of Medicine*, 355(3): 251–259.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037.

Perros, I.; Papalexakis, E. E.; Wang, F.; Vuduc, R.; Searles, E.; Thompson, M.; and Sun, J. 2017. SPARTan: Scalable PARAFAC2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 375–384.

Pierson, E.; Koh, P.; Hashimoto, T.; Koller, D.; Leskovec, J.; Eriksson, N.; and Liang, P. 2019. Inferring Multidimensional Rates of Aging from Cross-Sectional Data. *Proceedings of machine learning research*, 89: 97–107.

Pollard, T. J.; Chen, I.; Wiens, J.; Horng, S.; Wong, D.; Ghassemi, M.; Mattie, H.; Lindemer, E.; and Panch, T. 2019. Turning the crank for machine learning: ease, at what expense? *The Lancet Digital Health*, 1(5): e198–e199.

Rebbapragada, U.; Protopapas, P.; Brodley, C. E.; and Alcock, C. 2009. Finding anomalous periodic time series. *Machine learning*, 74(3): 281–313.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, II–1278.

Seaman, S.; Dukes, O.; Keogh, R.; and Vansteelandt, S. 2020. Adjusting for time-varying confounders in survival analysis using structural nested cumulative survival time models. *Biometrics*, 76(2): 472–483.

Shah, S. J.; Katz, D. H.; Selvaraj, S.; Burke, M. A.; Yancy, C. W.; Gheorghiade, M.; Bonow, R. O.; Huang, C.-C.; and Deo, R. C. 2015. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 131(3): 269–279.

Tadic, M.; and Cuspidi, C. 2019. Obesity and heart failure with preserved ejection fraction: a paradox or something else? *Heart Failure Reviews*, 24(3): 379–385.

Taylor, K. S. M.; Cook, J. A.; and Counsell, C. E. 2007. Heterogeneity in male to female risk for Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(8): 905–906.

Udler, M. S.; Kim, J.; von Grotthuss, M.; Bonás-Guarch, S.; Cole, J. B.; Chiou, J.; on behalf of METASTROKE, C. D. A.; the ISGC; Boehnke, M.; Laakso, M.; Atzmon, G.; et al. 2018. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS medicine*, 15(9): e1002654.

Wolf, F. A.; Hamey, F. K.; Plass, M.; Solana, J.; Dahlin, J. S.; Göttgens, B.; Rajewsky, N.; Simon, L.; and Theis, F. J. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1): 59.

Wu, H.; Gao, L.; and Kasabov, N. 2015. Inference of cancer progression from somatic mutation data. *IFAC-PapersOnLine*, 48(28): 234–238.

Young, A. L.; Marinescu, R. V.; Oxtoby, N. P.; Bocchetta, M.; Yong, K.; Firth, N. C.; Cash, D. M.; Thomas, D. L.; Dick, K. M.; Cardoso, J.; et al. 2018. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature communications*, 9(1): 4273.

Zhang, X.; Chou, J.; Liang, J.; Xiao, C.; Zhao, Y.; Sarva, H.; Henchcliffe, C.; and Wang, F. 2019. Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study. *Scientific reports*, 9(1): 1–12.