Neural Pharmacodynamic State Space Modeling

Zeshan Hussain [*1] **Rahul G. Krishnan** [*2] **David Sontag** [1]

Abstract

Modeling the time-series of high-dimensional, longitudinal data is important for predicting patient disease progression. However, existing neural network based approaches that learn representations of patient state, while very flexible, are susceptible to overfitting. We propose a deep generative model that makes use of a novel attention-based neural architecture inspired by the physics of how treatments affect disease state. The result is a scalable and accurate model of high-dimensional patient biomarkers as they vary over time. Our proposed model yields significant improvements in generalization and, on real-world clinical data, provides interpretable insights into the dynamics of cancer progression.

1. Introduction

Clinical biomarkers capture snapshots of a patient's evolving disease state as well as their response to treatment. However, these data can be high-dimensional, exhibit missingness, and display complex nonlinear behaviour over time as a function of time-varying interventions. Good unsupervised models of such data are key to discovering new clinical insights. This task is commonly referred to as disease progression modeling (Alaa & van der Schaar, 2019; Elibol et al., 2016; Liu et al., 2015; Schulam & Saria, 2016; Severson et al., 2020; Venuto et al., 2016; Wang et al., 2014).

Reliable unsupervised models of time-varying clinical data find several uses in healthcare. One use case is enabling practitioners to ask and answer counterfactuals using observational data (Bica et al., 2020a; Pearl et al., 2009; Rubin, 1974). Other use cases include guiding early treatment decisions based on a patient's biomarker trajectory, detecting drug effects in clinical trials (Mould et al., 2007), and clustering patterns in biomarkers that correlate with disease sub-type (Zhang et al., 2019b). To do these tasks well, understanding how a patient's biomarkers evolve over time given a prescribed treatment regimen is vital, since a person's biomarker profile is often the only observed proxy to their true disease state. Like prior work (Alaa & van der Schaar, 2019; Krishnan et al., 2017; Severson et al., 2020), we frame this problem as a conditional density estimation task, where our goal is to model the density of complex multivariate time-series conditional on time-varying treatments.

Representation learning exposes a variety of techniques for good conditional density estimation (Che et al., 2018; Choi et al., 2016; Miotto et al., 2016; Suresh et al., 2017). For sequential data, a popular approach has been to leverage black-box, sequential models (e.g. Recurrent Neural Networks (RNNs)), where a time-varying representation is used to predict clinical biomarkers. Such models are prone to overfitting, particularly on smaller clinical datasets. More importantly, such models often make simplistic assumptions on how time-varying treatments affect downstream clinical biomarkers; for example, one choice is to concatenate treatments to the model's hidden representations (Alaa & van der Schaar, 2019; Krishnan et al., 2017). The assumption here is that the neural network learns how treatments influence the representation. We argue that this choice is a missed opportunity and better choices exist. Concretely, we aim to encourage neural models to learn representations that encode a patient's underlying disease burden by specifying how these representations evolve due to treatment. We develop a new disease progression model that captures such insights by using inductive biases rooted in the biological mechanisms of treatment effect.

Inductive biases have been integral to the success of deep learning in other domains such as vision, text and audio. For example, convolutional neural networks explicitly learn representations invariant to translation or rotation of image data (Jaderberg et al., 2015; LeCun, 2012; Veeling et al., 2018), transformers leverage attention modules (Bahdanau et al., 2014; Vaswani et al., 2017) that mimic how human vision pays attention to various aspects of an image, and modified graph neural networks can explicitly incorporate laws of physics to generalize better (Seo & Liu, 2019). For some learning problems, the physics underlying the domain are often known, e.g. the laws of motion, and may be leveraged in the design of inductive biases (Anderson et al., 2019; Ling

[*]Equal contribution [1]Massachussetts Institute of Technology, CSAIL and IMES, Cambridge, MA [2]Microsoft Research New England, Cambridge, MA. Correspondence to: Zeshan Hussain <zeshanmh@mit.edu>, Rahul G. Krishnan <rahulgk@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

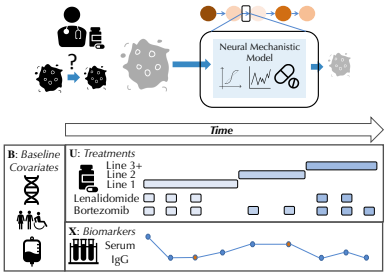

Figure 1. **Inductive Bias Concept (Top):** A clinician often has multiple mechanistic hypotheses as to how the latent tumor burden evolves. Our approach formalizes these hypotheses as neural architectures that specify how representations respond to treatments. **Patient Data (Bottom):** Illustration of data from a chronic disease patient. Baseline (static) data typically consists of genomics, demographics, and initial labs. Longitudinal data typically includes laboratory values (e.g. serum IgG) and treatments (e.g. lenalidomide). Baseline data is usually complete, but longitudinal measurements are frequently missing at various time points.

et al., 2016; Wang et al., 2020). The same does not hold true in healthcare, since exact disease and treatment response mechanisms are not known. However, physicians often have multiple hypotheses of how the disease behaves during treatment. To capture this intuition, we develop inductive biases that allow for a data-driven selection over multiple neural mechanistic models that dictate how treatments affect representations over time.

Contributions: We present a new attention-based neural architecture, $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$, that captures the effect of drug combinations in representation space (Figure 1 [top]). It learns to attend over multiple competing mechanistic explanations of how a patient's genetics, past treatment history, and prior disease state influence the representation to predict the next outcome. The architecture is instantiated in a state space model, $\textbf{SSM}_{\text{PK-PD}}$, and shows strong improvements in generalization compared to several baselines and prior state of the art. We demonstrate the model can provide insights into multiple myeloma progression. Finally, we release a disease progression benchmark dataset called ML-MMRF, comprising a curated, preprocessed subset of data from the Multiple Myeloma Research Foundation CoMMpass study (US National Institutes of Health, and others). Our model code can be found at `https://github.com/clinicalml/ief`, and the data processing code can be found at `https://github.com/clinicalml/ml_mmrf`.

2. Related Work

Much work has been done across machine learning, pharmacology, statistics and biomedical informatics on building models to characterize the progression of chronic diseases. Gaussian Processes (GPs) have been used to model patient biomarkers over time and estimate counterfactuals over a single intervention (Futoma et al., 2016; Schulam & Saria, 2017; Silva, 2016; Soleimani et al., 2017). In each of these cases, the focus is either on a single intervention per time point or on continuous-valued interventions given continuously, both strong assumptions for chronic diseases. To adjust for biases that exist in longitudinal data, Bica et al. (2020a); Lim et al. (2018) use propensity weighting to adjust for time-dependent confounders. However, they concatenate multi-variate treatments to patient biomarkers as input to RNNs; when data is scarce, such approaches have difficulty capturing how the hidden representations respond to treatment.

State space models and other Markov models have been used to model the progression of a variety of chronic diseases, including Cystic Fibrosis, scleroderma, breast cancer, COPD and CKD (Alaa & van der Schaar, 2019; Perotte et al., 2015; Schulam & Saria, 2016; Taghipour et al., 2013; Wang et al., 2014). There has also been much research in characterizing disease trajectories, subtypes, and correlations between risk factors and progression for patients suffering from Alzheimer's Disease (Goyal et al., 2018; Khatami et al., 2019; Marinescu et al., 2019; Zhang et al., 2019a). Like us, the above works pose disease progression as density estimation but in contrast, many of the above models do not condition on time-varying interventions.

3. Background - State Space Models (SSMs)

SSMs are a popular model for sequential data and have a rich history in modeling disease progression.

Notation: $B \in \mathbb{R}^J$ denotes baseline data that are static, i.e. individual-specific covariates. For chronic diseases, these data comprise a J-dimensional vector, including patients' age, gender, genetics, race, and ethnicity. Let $\mathbf{U} = \{U_0, \ldots, U_{T-1}\}$; $U_t \in \mathbb{R}^L$ be a sequence of L-dimensional interventions for an individual. An element of U_t may be binary, to denote prescription of a drug, or real-valued, to denote dosage. $\mathbf{X} = \{X_1, \ldots, X_T\}$; $X_t \in \mathbb{R}^M$ denotes the sequence of real-valued, M-dimensional clinical biomarkers. An element of X_t may denote a serum lab value or blood count, which is used by clinicians to measure organ function as a proxy for disease severity. X_t frequently contains missing data. We assume access to a dataset $\mathcal{D} = \{(\mathbf{X}^1, \mathbf{U}^1, B^1), \ldots, (\mathbf{X}^N, \mathbf{U}^N, B^N)\}$. For a visual depiction of the data, we refer the reader to Figure 1. Unless required, we ignore the superscript denoting the index of the datapoint and denote concatenation with [].

Model: SSMs capture dependencies in sequential data via a time-varying latent state. When this latent state is discrete, SSMs are also known as Hidden Markov Models (HMM).

In our setting, we deal with a continuous latent state. The generative process is:

$$p(\mathbf{X}|\mathbf{U}, B) = \int_Z \prod_{t=1}^{T} p_\theta(Z_t|Z_{t-1}, U_{t-1}, B) p_\theta(X_t|Z_t) dZ$$
$$Z_t|\cdot \sim \mathcal{N}(\mu_\theta(Z_{t-1}, U_{t-1}, B), \Sigma_\theta^t(Z_{t-1}, U_{t-1}, B)),$$
$$X_t|\cdot \sim \mathcal{N}(\kappa_\theta(Z_t), \Sigma_\theta^e(Z_t)) \quad (1)$$

We denote the parameters of a model by θ, which may comprise weight matrices or the parameters of functions that index θ. SSMs make the Markov assumption *on the latent variables*, Z_t, and we assume that relevant information about past medications are captured by the state or contained in U_{t-1}. We set $\Sigma_\theta^t, \Sigma_\theta^e, \kappa_\theta(Z_t)$ to be functions of a concatenation of their inputs, e.g. $\Sigma_\theta^t(\cdot) = \textbf{softplus}(\mathbf{W}[Z_{t-1}, U_{t-1}, B] + \mathbf{b})$. $\Sigma_\theta^t, \Sigma_\theta^e$ are diagonal matrices where the softplus function is used to ensure positivity.

Learning: We maximize $\sum_{i=1}^{N} \log p(\mathbf{X}^i|\mathbf{U}^i, B^i)$ with respect to θ. For a nonlinear SSM, this function is intractable, so we learn via maximizing a variational lower bound on it. To evaluate the bound, we perform probabilistic inference using a structured inference network (Krishnan et al., 2017). The learning algorithm alternates between predicting variational parameters using a bi-directional recurrent neural network, evaluating a variational upper bound, and making gradient updates jointly with respect to the parameters of the generative model and the inference network. When evaluating the likelihood of data under the model, if X_t is missing, it is marginalized out. Since the inference network also conditions on sequences of observed data to predict the variational parameters, we use forward fill imputation where data are missing.

4. Attentive Pharmacodynamic State Space Model

To make the shift from black-box models to those that capture useful structure for modeling clinical data, we begin with a discussion of PK-PD models and some of the key limitations that practitioners may face when directly applying them to modern clinical datasets.

4.1. Limitations of Pharmacokinetic-Pharmacodynamic Modeling

Pharmacology is a natural store of domain expertise for reasoning about how treatments affect disease. We look specifically at pharmacokinetics (PK), which deals with how drugs move in the body, and pharmacodynamics (PD), which studies the body's response to drugs. Consider a classical pharmacokinetic-pharmacodynamic (PK-PD) model used to characterize variation in tumor volume due to chemotherapy (Norton, 2014; West & Newton, 2017). Known as the log-cell kill model, it is based on the hypothesis that a given dose of chemotherapy results in killing a constant fraction of tumor cells rather than a constant number of cells. The original model is an ordinary differential equation but an equivalent expression is:

$$S(t) = S(t-1) \cdot (1 + \rho \log(K/S(t-1)) - \beta_c C(t)), \quad (2)$$

$S(t)$ is the (scalar) tumor volume, $C(t)$ is the (scalar) concentration of a chemotherapeutic drug over time, K is the maximum tumor volume possible, ρ is the growth rate, and β_c represents the drug effect on tumor size. Besides its bespoke nature, there are some key limitations of this model that hinder its broad applicability for unsupervised learning:

Single intervention, single biomarker: The model parameterizes the effect of a *single, scalar* intervention on a single, scalar, time-varying biomarker making it impossible to apply directly to high-dimensional clinical data. Furthermore, the quantity it models, tumor volume, is unobserved for non-solid cancers.

Misspecified in functional form: The log-cell-kill hypothesis, by itself, is not an accurate description of the drug mechanism in most non-cancerous chronic diseases.

Misspecified in time: Patients go through cycles of recovery and relapse during a disease. Even if the hypothesis holds when the patient is sick, it may not hold when the patient is in recovery.

In what follows, we aim to mitigate these limitations to build a practical, scalable model of disease progression.

4.2. Latent Representations of Disease State

Tackling the first limitation, we use nonlinear SSMs in order to model longitudinal, high-dimensional data. Even though tumor volume may not be observed in observational clinical datasets, various proxies (e.g. lab values, blood counts) of the unobserved disease burden often are. We conjecture that the time-varying latent representation, Z_t, implicitly captures such clinical phenotypes from the observations.

To ensure that the phenotypes captured by Z_t vary over time in a manner akin to clinical intuition, we focus the efforts of our design on the transition function, $\mu_\theta(Z_{t-1}, U_{t-1}, B)$, of the state space model. This function controls the way in which the latent state Z_t in an SSM evolves over time (and through it, the data) when exposed to interventions, U_t; this makes the transition function a good starting point for incorporating clinical domain knowledge.

4.3. Neural Attention over Treatment Effect Mechanisms

In order to design a good transition function, we first need to address the second limitation that we may not know the

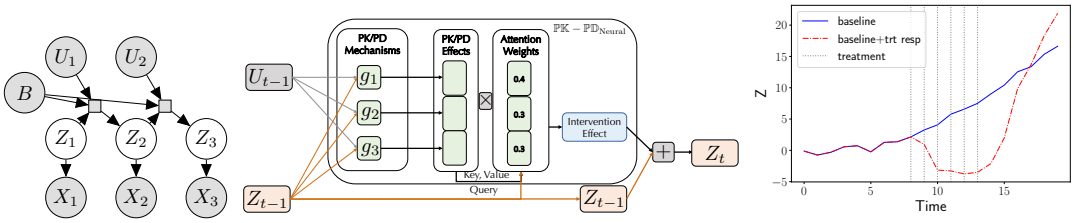

Figure 2. **Unsupervised Models of Sequential Data (Left):** We show a State Space Model (SSM) of **X** (the longitudinal biomarkers) conditioned on B (genetics, demographics) and **U** (binary indicators of treatment and line of therapy). The rectangle depicts the $\mu_\theta(Z_{t-1}, U_{t-1}, B)$. **Neural Architecture for $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$ (Middle):** Illustration of the neural architecture we design; we use a soft-attention mechanism over the neural PK/PD effects using the current patient representation as a query to decide how the masks should be distributed. **Modeling relapse with the neural treatment exponential response (Right):** The curve depicts a single dimension of the representation and vertical lines denote a single treatment. After maintaining the response with treatments, a regression towards baseline (in blue; depicting what would have happened had no treatment been prescribed) occurs when treatment is stopped.

exact mechanism by which drugs affect the disease state. However, we often have a set of reasonable hypotheses about the mechanisms that underlie how we expect the dynamics of the latent disease state to behave.

Putting aside the specifics of what mechanisms we should use for the moment, suppose we are given d mechanism functions, $g_1, \ldots, g_d$, each of which is a neural architecture that we believe captures aspects of how a representation should vary as a response to treatment. How a patient's representation should vary will depend on what state the patient is in. e.g. sicker patients may respond less well to treatment than healthier ones. To operationalize this insight, we make use of an attention mechanism (Bahdanau et al., 2014) to attend to which choice of function is most appropriate.

Attending over mechanisms of effect Attention mechanisms operate by using a "query" to index into a set of "keys" to compute a set of attention weights, which are a distribution over the "values". We propose a soft-attention mechanism to select between $g_1, \ldots, g_d$. At each t, for the query, we have $q = Z_{t-1}W_q$. For the key and value, we have,

$$\tilde{K} = [g_1(Z_{t-1}, U_{t-1}, B); \ldots; g_d(Z_{t-1}, U_{t-1}, B)]^\top W_k$$
$$\tilde{V} = [g_1(Z_{t-1}, U_{t-1}, B); \ldots; g_d(Z_{t-1}, U_{t-1}, B)]^\top W_v.$$

Note that $W_q, W_k, W_v \in \mathbb{R}^{Q\times Q}$ and that $q \in \mathbb{R}^Q$, $\tilde{K} \in \mathbb{R}^{Q\times d}$, and $\tilde{V} \in \mathbb{R}^{Q\times d}$. Then, we have the following,

$$\mu_\theta(Z_{t-1}, U_{t-1}, B) = \left(\sum_{i=1}^{d} \textbf{softmax}\left(\frac{q \odot \tilde{K}}{\sqrt{Q}}\right)_i \odot \tilde{V}_i\right) W_o \tag{3}$$

We compute the attention weights using the latent representation at a particular time point as a "query" and the output of each of $g_1, \ldots, g_d$ as "keys"; see Figure 2 (middle). This choice of neural architecture for μ_θ allows us to parameterize heterogeneous SSMs, where the function characterizing latent dynamics changes over time.

4.4. Lines of Therapy with Local and Global Clocks

Here, we address a third limitation of classical PK-PD models: a proposed drug mechanism's validity may depend on how long the patient has been treated and what stage of therapy they are in. Such stages, or *lines of therapy*, refer to contiguous plans of multiple treatments prescribed to a patient. They are often a unique structure of clinical data from individuals suffering from chronic diseases. For example, first line therapies often represent combinations prioritized due to their efficacy in clinical trials; subsequent lines may be decided by clinician preference. Lines of therapy index treatment plans that span multiple time-steps and are often laid out by clinicians at first diagnosis. We show how to make use of this information within a mechanism function.

To capture the clinician's intention when prescribing treatment, we incorporate line of therapy as a one-hot vector in $U_t[:K]\ \forall t$ (K is the maximal line of therapy). Lines of therapy typically change when a drug combination fails or causes adverse side effects. By conditioning on line of therapy, a transition function (of the SSM) parameterized by a neural network can, in theory, infer the length of time a patient has been on that line. However, although architectures such as Neural Turing Machines (Graves et al., 2014) can learn to count occurrences, they would need a substantial amount of data to do so.

To enforce the specified drug mechanism functions to capture time since change in line of therapy, we use *clocks* to track the time elapsed since an event. This strategy has precedent in RNNs, where Che et al. (2018) use time since the last observation to help RNNs learn well when data is missing. Koutnik et al. (2014) partition the hidden states in RNNs so they are updated at different time-scales. Here, we augment our interventional vector, U_t, with two more dimensions. A global clock, gc, captures time elapsed since $T = 0$, i.e. $U_t[K] = \text{gc}_t = t$. A local clock, lc, captures time elapsed since a line of therapy began; i.e.

$U_t[K+1] = \text{lc}_t = t - p_t$ where p_t denotes the index of time when the line last changed. By using the local clock, $\mu_\theta(Z_{t-1}, U_{t-1}, B)$ can modulate Z_t to capture patterns such as: the longer a line of therapy is deployed, the less or (more) effective it may be.

For the patient in Figure 1, we can see that the first dimension of $\mathbf{U}$ denoting line of therapy would be $[0,0,0,0,1,1,2,2,2]$. Line 0 was used four times, line 1 used twice, line 2 used thrice. Then, $p = [0,0,0,0,4,4,6,6,6,6]$, $gc = [0,1,2,3,4,5,6,7,8,9]$ and $lc = [0,1,2,3,0,1,0,1,2,3]$. To the best of our knowledge, we are the first to make use of lines-of-therapy information and clocks concurrently to capture temporal information when modeling clinical data.

4.5. Neural PK-PD Functions for Chronic Diseases

Having developed solutions to tackle some of the limitations of PK-PD models, we turn to the design of three new mechanism functions, each of which captures different hypotheses a clinician may have about how the underlying disease burden of a patient changes (as manifested in their latent states).

Modeling baseline conditional variation: Biomarkers of chronic diseases can increase, decrease, or stay the same. Such patterns may be found in the dose-response to chemotherapy used in solid cancerous tumors (Klein, 2009). In reality, clinicians find that these changes are often modulated by patient specific features such as age, genetic mutations, and history of illness. Patients who have been in therapy for a long time may find decreased sensitivity to treatments. To capture this variation:

$$g_1(Z_{t-1}, U_{t-1}, B) = Z_{t-1} \cdot \tanh(b_{\text{lin}} + W_{\text{lin}}[U_{t-1}, B]) \tag{4}$$

where $b_{\text{lin}} \in \mathbb{R}^Q, W_{\text{lin}} \in \mathbb{R}^{Q \times (L+J)}$. Here, the effects on the representation are bounded (via the tanh function) but depend on the combination of drugs prescribed and the patient's baseline data, including genetics.

Modeling slow, gradual relapse after treatment: One of the defining features of many chronic diseases is the possibility of a relapse during active therapy. In cancer, a relapse can happen due to cancerous cells escaping the treatment or a variety of other bio-chemical processes, such as increased resistance to treatment due to mutations. The relapse can result in bio-markers reverting to values that they held prior to the start of treatment; for an example of this, see Figure 2 (right). We design the following neural architectures to capture such patterns in a latent representation.

Neural Log-Cell Kill: This architecture is inspired by the classical log cell kill model of tumor volume in solid cell tumors (West & Newton, 2017) but unlike the original model, scales to high-dimensional representations and takes into account *lines of therapy* via the local clock. This allows the model to effectively reset every time a new line of therapy begins. The functional form of the model is,

$$g_2(Z_{t-1}, U_{t-1}, B) = Z_{t-1} \cdot (1 - \rho \log(Z_{t-1}^2) - \beta \exp(-\delta \cdot \text{lc}_{t-1})), \tag{5}$$

where $\beta = \tanh(W_{lc} U_{t-1} + b_{lc})$. $W_{lc} \in \mathbb{R}^{Q \times L}, b_{lc} \in \mathbb{R}^Q, \delta \in \mathbb{R}^Q$ and $\rho \in \mathbb{R}^Q$ are learned. While diseases may not have a single observation that characterizes the state of the organ system (akin to tumor volume), we hypothesize that representations, Z_t, of the observed clinical biomarkers may benefit from mimicking the dynamics exhibited by tumor volume when exposed to chemotherapeutic agents. We emphasize that unlike Equation 2, the function in Equation 5 operates over a *vector valued* set of representations that can be modulated by the patient's genetic markers.

Neural Treatment Exponential: Xu et al. (2016) develop a Bayesian nonparameteric model to explain variation in creatinine, a single biomarker, due to treatment. We design an architecture inspired by their model that scales to high dimensional representations, allows for the representation to vary as a function of the patient's genetics, and makes use of information in the lines of therapy via the clocks.

$$g_3(Z_{t-1}, U_{t-1}, B) = \begin{cases} b_0 + \alpha_{1,t-1}/[1 + \exp(-\alpha_{2,t-1}(\text{lc}_{t-1} - \frac{\gamma_l}{2}))], \\ \text{if } 0 \le \text{lc}_{t-1} < \gamma_l \\ b_l + \alpha_{0,t-1}/[1 + \exp(\alpha_{3,t-1}(\text{lc}_{t-1} - \frac{3\gamma_l}{2}))], \\ \text{if } \text{lc}_{t-1} \ge \gamma_l \end{cases} \tag{6}$$

Despite its complexity, the intermediate representations learned within this architecture have simple intuitive meanings. $\alpha_{1,t-1} = W_d[Z_{t-1}, U_{t-1}, B] + b_d$, where $W_d \in \mathbb{R}^{Q \times (Q+L+J)}, b_d \in \mathbb{R}^Q$ is used to control whether each dimension in Z_{t-1} increases or decreases as a function of the treatment and baseline data. $\alpha_{2,t-1}, \alpha_{3,t-1}$, and γ_l control the steepness and duration of the intervention effect. We restrict these characteristics to be similar for drugs administered under the same line of therapy. Thus, we parameterize: $[\alpha_2, \alpha_3, \gamma_l]_{t-1} = \sigma(W_e \cdot U_{t-1}[0] + b_e)$. If there are three lines of therapy, $W_e \in \mathbb{R}^{3 \times 3}, b_e \in \mathbb{R}^3$ and the biases, $b_0 \in \mathbb{R}^Q$ and $b_l \in \mathbb{R}^Q$, are learned. Finally, $\alpha_{0,t-1} = (\alpha_{1,t-1} + 2b_0 - b_l)/(1 + \exp(-\alpha_{3,t-1}\gamma_l/2))$ ensures that the effect peaks at $t = \text{lc}_t + \gamma_l$. Figure 2 (right) depicts how a single latent dimension may vary over time for a single line of therapy using this neural architecture.

From $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$ to the $\text{SSM}_{\text{PK-PD}}$: When g_1, g_2, g_3, as described in Equations 4, 5, 6, are used in the transition function μ_θ (as defined in Equation 3), we refer to the resulting function as $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$. Moreover, when $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$

is used as the transition function in an SSM, we refer to the resulting model as $\textbf{SSM}_{\text{PK-PD}}$, a heterogeneous state space model designed to model the progression of diseases.

5. Evaluation

5.1. Datasets

We study $\textbf{SSM}_{\text{PK-PD}}$ on three different datasets – two here, and on a third semi-synthetic dataset in the appendix.

Synthetic Data: We begin with a synthetic disease progression dataset where each patient is assigned baseline covariates $B \in \mathbb{R}^6$. B determines how the biomarkers, $X_t \in \mathbb{R}^2$, behave in the absence of treatment. $U_t \in \mathbb{R}^4$ comprises the line of therapy ($K = 2$), the local clock, and a single binary variable indicating when treatment is prescribed. To mimic the data dynamics described in Figure 1, the biomarkers follow second-order polynomial trajectories over time with the underlying treatment effect being determined by the Neural Treatment Exponential (see Equation 6). Biomarker 1 can be thought of as a marker of disease burden, while biomarker 2 can be thought of as a marker of biological function. The full generative process for the data is in the supplementary material. To understand generalization of the model as a function of sample complexity, we train on 100/1000 samples and evaluate on five held-out sets of size 50000.

ML-MMRF: The Multiple Myeloma Research Foundation (MMRF) CoMMpass study releases de-identified clinical data for 1143 patients suffering from multiple myeloma, an incurable plasma cell cancer. All patients are aligned to the start of treatment, which is made according to current standard of care (not random assignment). With an oncologist, we curate demographic and genomic markers, $B \in \mathbb{R}^{16}$, clinical biomarkers, $X_t \in \mathbb{R}^{16}$, and interventions, $U_t \in \mathbb{R}^9$, with one local clock, a three dimensional one-hot encoding for line of therapy, and binary markers of 5 drugs. Our results are obtained using a 75/25 train/test split. To select hyperparameters, we perform 5-fold cross validation on the training set. Finally, there is missingness in the biomarkers, with 66% of the observations missing. We refer the reader to the appendix for more details on the dataset.

5.2. Setup

We learn via: $(\arg\min_\theta - \log p(\mathbf{X}|\mathbf{U}, B; \theta))$ using ADAM (Kingma & Ba, 2014) with a learning rate of 0.001 for 15000 epochs. L1 or L2 regularization is applied in one of two ways: either we regularize all model parameters (including parameters of inference network), or we regularize all weight matrices except those associated with the attention mechanism. We search over regularization strengths of $0.01, 0.1, 1, 10$ and latent dimensions of $16, 48, 64$ and 128. We do model selection using the negative evidence lower bound (NELBO); Appendix B contains details on the derivation of this bound. We use a single copy of g_1, g_2, and g_3 in the transition function. Multiple copies of each function as well as other "mechanistic" functions can be used, highlighting the flexibility of our approach. However, this must be balanced with potentially overfitting on small datasets.

5.3. Baselines

$\textbf{SSM}_{\text{Linear}}$ parametrizes $\mu_\theta(Z_{t-1}, U_{t-1}, B)$ with a linear function. This model is a strong, linear baseline whose variants have been used for modeling data of patients suffering from Chronic Kidney Disease (Perotte et al., 2015).

$\textbf{SSM}_{\text{NL}}$: Krishnan et al. (2017) use a nonlinear SSM to capture variation in the clinical biomarkers of diabetic patients. We compare to their model, parameterizing the transition function with a 2-layer MLP.

$\textbf{SSM}_{\text{MOE}}$: We use an SSM whose transition function is parameterized via a Mixture-of-Experts (MoE) architecture (Jacobs et al., 1991; Jordan & Jacobs, 1994); i.e. g_1, g_2, g_3 are each replaced with a multi-layer perceptron. This baseline does not incorporate any domain knowledge and tests the relative benefits of prescribing the functional forms via mechanisms versus learning them from data.

$\textbf{SSM}_{\text{Attn.Hist.}}$: We implement a variant of the SSM in Alaa & van der Schaar (2019), a state-of-the-art model for disease progression trained via conditional density estimation. The authors use a *discrete* state space for disease progression modeling making a direct comparison difficult. However, $\textbf{SSM}_{\text{Attn.Hist.}}$ preserves the structural modeling assumptions they make. Namely, the transition function of the model attends to a concatenation of previous states and interventions at each point in time. We defer specifics to Appendix B.

In addition, we run two simpler baselines, a First Order Markov Model (FOMM) and Gated Recurrent Unit (GRU) (Cho et al., 2014), on the synthetic data and ML-MMRF but defer those results to Appendix E.

5.4. Evaluation Metrics

NELBO On both the synthetic data and ML-MMRF data, we quantify generalization via the negative evidence lower bound (NELBO), which is a variational upper bound on the negative log-likelihood of the data. A lower NELBO indicates better generalization.

Pairwise Comparisons For a fine-grain evaluation of our models on ML-MMRF, we compare held-out NELBO under $\textbf{SSM}_{\text{PK-PD}}$ versus the corresponding baseline for each patient. For each held-out point, $\Delta_i = 1$ when the NELBO of that datapoint is lower under $\textbf{SSM}_{\text{PK-PD}}$ and $\Delta_i = 0$ when it is not. In Table 1 (bottom), we report $\frac{1}{N}\sum_{i=1}^{N} \Delta_i$, the

proportion of data for which $\mathbf{SSM}_{\text{PK-PD}}$ yields better results.

Counts To get a sense for the number of patients on whom $\mathbf{SSM}_{\text{PK-PD}}$ does much better, we count the number of held-out patients for whom the held-out negative log likelihood (computed via importance sampling) is more than 10 nats lower under $\mathbf{SSM}_{\text{PK-PD}}$ than the corresponding baseline (and vice versa for the baselines).

5.5. Results

We investigate three broad categories of questions.

5.5.1. Generalization under different conditions

$SSM_{PK\text{-}PD}$ generalizes better in setting with few ($\sim$ 100) samples. Table 1 (top) depicts NELBOs on held-out synthetic data across different models, where a lower number implies better generalization. We see a statistically significant gain in generalization for $\mathbf{SSM}_{\text{PK-PD}}$ compared to all other baselines. $\mathbf{SSM}_{\text{NL}}$, $\mathbf{SSM}_{\text{MOE}}$ overfit quickly on 100 samples but recover their performance when learning with 1000 samples.

$SSM_{PK\text{-}PD}$ generalizes well when it is misspecified. Because we often lack prior knowledge about the true underlying dynamics in the data, we study how $\mathbf{SSM}_{\text{PK-PD}}$ performs when it is misspecified. We replace the Neural Treatment Exponential function, g_3, from $\mathbb{PK}\text{-}\mathbb{PD}_{\text{Neural}}$ with another instance of g_1. The resulting model is now misspecified since g_3 is used to generate the data but no longer lies within the model family. We denote this model as (SSM PK-PD w/o TExp). In Table 1 (top), when comparing the sixth column to the others, we find that we outperform all baselines and get comparable generalization to $\mathbf{SSM}_{\text{PK-PD}}$ with the Neural Treatment Exponential function. This result emphasizes our architecture's flexibility and its ability to learn the underlying (unknown) intervention effect through a combination of other, related mechanism functions.

$SSM_{PK\text{-}PD}$ generalizes well on real-world patient data. A substantially harder test of model misspecification is on the ML-MMRF data where we have unknown dynamics that drive the high-dimensional (often missing) biomarkers in addition to combinations of drugs prescribed over time. To rigorously validate whether we improve generalization on ML-MMRF data with $\mathbf{SSM}_{\text{PK-PD}}$, we study model performance with respect to the three metrics introduced in Section 5.4. We report our results in Table 1 (bottom). First, we consistently observe that a high fraction of patient data in the test set are explained better by $\mathbf{SSM}_{\text{PK-PD}}$ than the corresponding baseline (pairwise comparisons). We also note that out of 282 patients in the test set, across all the baselines, we find that the $\mathbf{SSM}_{\text{PK-PD}}$ generalizes better for many more patients (counts). Finally, $\mathbf{SSM}_{\text{PK-PD}}$ has lower NELBO averaged across the entire test set compared to all baselines.

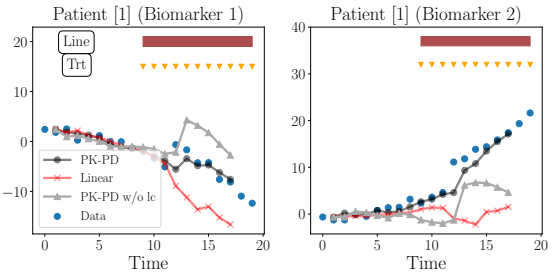

Figure 3. Synthetic: Forward samples (conditioned only on B) from $\mathbf{SSM}_{\text{PK-PD}}$ (o), $\mathbf{SSM}_{\text{Linear}}$ (x), $\mathbf{SSM}_{\text{PK-PD}}$ without local clocks ($\triangle$), for a single patient. Y-axis shows biomarker values.

5.5.2. Model complexity & generalization

The improvements of $SSM_{PK\text{-}PD}$ are consistent taking model sizes into account. We show in Table 1 (bottom) the number of parameters used in each model. We find that more parameters do not imply better performance. Models with the most parameters (e.g. $\mathbf{SSM}_{\text{NL}}$) overfit while those with the lowest number of parameters underfit (e.g. $\mathbf{SSM}_{\text{Linear}}$) suggesting that the gains in generalization that we observe are coming from our parameterization. We experimented with increasing the size of the $\mathbf{SSM}_{\text{Linear}}$ model (via the latent variable dimension) to match the size of the best PK-PD model. We found that doing so *did not outperform* the held-out likelihood of $\mathbf{SSM}_{\text{PK-PD}}$.

When data are scarce, a Mixture of Experts architecture is difficult to learn: How effective are the functional forms of the neural architectures we develop? To answer this question, we compare the held-out log-likelihood of $\mathbf{SSM}_{\text{PK-PD}}$ vs $\mathbf{SSM}_{\text{MOE}}$ in the third column of Table 1 (bottom). In the ML-MMRF data, we find that the $\mathbf{SSM}_{\text{PK-PD}}$ outperforms the $\mathbf{SSM}_{\text{MOE}}$. We suspect this is due to the fact that learning diverse "experts" is hard when data is scarce and supports the hypothesis that the judicious choice of neural architectures plays a vital role in capturing biomarker dynamics.

Can $\mathbb{PK}\text{-}\mathbb{PD}_{Neural}$ be used in other model families? In the supplement, we implement $\mathbb{PK}\text{-}\mathbb{PD}_{\text{Neural}}$ in a first-order Markov model and find similar improvements in generalization on the ML-MMRF dataset. This result suggests that the principle we propose of leveraging domain knowledge from pharmacology to design mechanism functions can allow other kinds of deep generative models (beyond SSMs) to also generalize better when data are scarce.

5.5.3. Visualizing Patient Dynamics

In Figure 4 (right), to further validate our initial hypothesis that the model is using the various neural PK-PD effect func-

Training Set Size	SSM Linear	SSM NL	SSM MOE	SSM Attn. Hist.	SSM PK-PD	SSM PK-PD (w/o TExp)
100	58.57 +/- .06	69.04 +/- .11	60.98 +/- .04	76.94 +/- .02	**55.34 +/- .03**	**58.39 +/- .05**
1000	53.84 +/- .02	44.75 +/- .02	51.57 +/- .03	73.80 +/- .03	**39.84 +/- .02**	**38.93 +/- .01**

Evaluation Metric	SSM PK-PD vs. SSM Linear	SSM PK-PD vs. SSM NL	SSM PK-PD vs. SSM MOE	SSM PK-PD vs. SSM Attn. Hist.
Pairwise Comparison (↑)	0.796 (0.400)	0.760 (0.426)	0.714 (0.450)	0.934 (0.247)
Counts (↑)	PK-PD: 158, LIN: 6	130, 12	94, 8	272, 0
NELBO (↓)	PK-PD: 61.54, LIN: 74.22	61.54, 79.10	61.54, 73.44	61.54, 105.04
# of Model Parameters	PK-PD: 23K, LIN: 7K	23K, 51K	23K, 77K	23K, 17K

Table 1. Top: **Synthetic data:** Lower is better. We report the test NELBO with std. dev. for each SSM model to study generalization in the synthetic setting. *Bottom:* **ML-MMRF:** *Pairwise Comparison*: each number is the fraction (with std. dev.) of test patients for which $\textbf{SSM}_{\text{PK-PD}}$ has a lower NELBO than an SSM that uses a different transition function. *Counts*: We report the number of test patients (out of 282) for which an SSM model (PK-PD or otherwise) has a greater than 10 nats difference in negative log likelihood compared to the alternative model. *NELBO*: We report the test NELBO of each model. Note that we label the metrics associated with $\textbf{SSM}_{\text{PK-PD}}$ and $\textbf{SSM}_{\text{Linear}}$ in the first column, but drop these labels in subsequent columns.

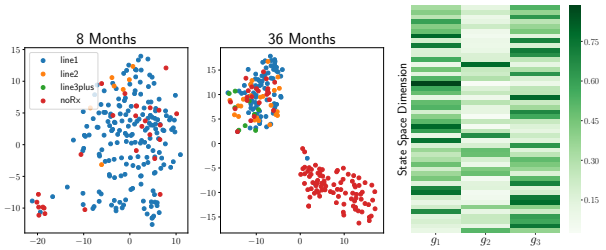

Figure 4. ML-MMRF: (Left two) We visualize the TSNE representations of each test patient's latent state, Z_t, at the start of treatment and three years in. (Right) For $\textbf{SSM}_{\text{PK-PD}}$, we visualize the attention weights, averaged over all time steps and all patients, on each of the neural effect functions across state space dimensions.

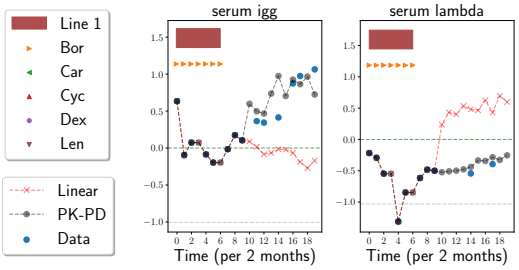

Figure 5. ML-MMRF: Each column is a different biomarker containing forward samples (conditioned on approximately the first 2 years of a patient's data) from $\textbf{SSM}_{\text{PK-PD}}$ (o) and $\textbf{SSM}_{\text{linear}}$ (x) of a single test patient. Blue circles denote ground truth, and the markers above the trajectories represent treatments prescribed across time. Y-axis shows biomarker levels, with the dotted green[gray] line representing the maximum[minimum] healthy value. Car, Cyc, Dex, and Len shown in legend to maintain consistency with plots in Appendix, but are not given in the treatment regimen.

tions, we visualize the attention weights from $\textbf{SSM}_{\text{PK-PD}}$ trained on ML-MMRF averaged across time and all patients. The highest weighted component is the treatment exponential model g_3, followed by the bounded linear model g_1 for many of the latent state dimensions. We also see that several of the latent state dimensions make exclusive use of the neural log-cell kill model g_2.

How do the clocks help model patient dynamics? Figure 3 shows samples from three SSMs trained on synthetic data. $\textbf{SSM}_{\text{PK-PD}}$ captures treatment response accurately while $\textbf{SSM}_{\text{Linear}}$ does not register that the effect of treatment can persist over time. To study the impact of clocks on the learned model, we perform an ablation study on SSMs where the local clock in U_t, used by $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$, is set to a constant. Without clocks (PK-PD w/o lc), the model does not capture the onset or persistence of treatment response.

***SSM**$_{\text{PK-PD}}$ learns latent representations that reflect the patient's disease state:* In ML-MMRF, we restrict the patient population to those with at least $T = 36$ months of data. At two different points during their treatment of the disease, we visualize the result of TSNE (Maaten & Hinton, 2008) applied to their latent representations in Figure 4 (left).

Early in their treatment, the latent representations of these patients appear to have no apparent structure. As time progresses, we find that the dimensions split into two groups. One group, for the most part, is still being treated, while the other is not being treated. A deeper dive into the untreated patients reveals that this cohort has a less severe subtype of myeloma (via a common risk assessment method known as ISS staging). This result suggests that the latent state of $\textbf{SSM}_{\text{PK-PD}}$ has successfully captured the coarse disease severity of patients at particular time points.

*Visualizing patient samples from **SSM**$_{\text{PK-PD}}$:* Figure 5 shows the average of three samples from $\textbf{SSM}_{\text{Linear}}$ and $\textbf{SSM}_{\text{PK-PD}}$ trained on ML-MMRF. We track two biomarkers used by

clinicians to map myeloma progression. $\mathbf{SSM}_{\text{PK-PD}}$ better captures the evolution of these biomarkers conditioned on treatment. For serum IgG, $\mathbf{SSM}_{\text{PK-PD}}$ correctly predicts the relapse of disease after stopping first line therapy, while $\mathbf{SSM}_{\text{Linear}}$ does not. On the other hand, for serum lambda, $\mathbf{SSM}_{\text{PK-PD}}$ correctly predicts it will remain steady.

6. Discussion

$\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$ leverages domain knowledge from pharmacology in the form of treatment effect mechanisms to quantitatively and qualitatively improve performance of a representation-learning based disease progression model. Bica et al. (2020b) note the potential for blending ideas from pharmacology with machine learning: our work is among the first to do so.

We highlight several avenues for future work. For example, in machine learning for healthcare, the model we develop can find use in practical problems such as forecasting high-dimensional clinical biomarkers and learning useful representations of patients that are predictive of outcomes. Doing so can aid in the development of tools for risk stratification or in software to aid in clinical trial design. Applying the model to data from other chronic diseases such as diabetes, congestive heart failure, small cell lung cancer and rheumatoid arthritis, brings opportunities to augment $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$ with new neural PK-PD functions tailored to capture the unique biomaker dynamics exhibited by patients suffering from these diseases.

Finally, we believe $\mathbb{PK}$-$\mathbb{PD}_{\text{Neural}}$ can find use in the design of parameteric environment simulators for different domains. In pharmacology, such simulation based pipelines can help determine effective drug doses (Hutchinson et al., 2019). Our idea of attending over multiple dynamics functions can find use in the design of simulators in domains such as economics, where multiple hypothesized mechanisms are used to explain observed market phenomena (Ghosh et al., 2019).

Acknowledgements

The authors thank Rebecca Boiarsky, Christina Ji, Monica Agrawal, Divya Gopinath for valuable feedback on the manuscript and many helpful discussions; Dr. Andrew Yee (Massachusetts General Hospital) for help in the construction of the ML-MMRF dataset; and both Rebecca Boiarsky and Isaac Lage for their initial analyses and processing of data from the CoMMpass study. These data were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org). This research was generously supported by an ASPIRE award from The Mark Foundation for Cancer Research.

References

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

Anderson, B., Hy, T. S., and Kondor, R. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems*, pp. 14537–14546, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *ICLR*, 2020a.

Bica, I., Alaa, A. M., Lambert, C., and van der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 2020b.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pp. 301–318. PMLR, 2016.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Elibol, H. M., Nguyen, V., Linderman, S., Johnson, M., Hashmi, A., and Doshi-Velez, F. Cross-corpora unsupervised learning of trajectories in autism spectrum disorders. *The Journal of Machine Learning Research*, 17(1):4597–4634, 2016.

Futoma, J., Sendak, M., Cameron, B., and Heller, K. Predicting disease progression with a model for multivariate longitudinal clinical data. In *Machine Learning for Healthcare Conference*, pp. 42–54, 2016.

Ghosh, I., Jana, R. K., and Sanyal, M. K. Analysis of temporal pattern, causal interaction and predictive modeling of financial markets using nonlinear dynamics, econometric models and machine learning algorithms. *Applied Soft Computing*, 82:105553, 2019.

Goyal, D., Tjandra, D., Migrino, R. Q., Giordani, B., Syed, Z., Wiens, J., Initiative, A. D. N., et al. Characterizing heterogeneity in the progression of alzheimer's disease using longitudinal clinical and neuroimaging biomarkers. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:629–637, 2018.

Graves, A., Wayne, G., and Danihelka, I. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Group, I. M. W. Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the international myeloma working group. *British journal of haematology*, 121(5):749–757, 2003.

Hutchinson, L., Steiert, B., Soubret, A., Wagg, J., Phipps, A., Peck, R., Charoin, J.-E., and Ribba, B. Models and machines: How deep learning will take clinical pharmacology to the next level. *CPT: pharmacometrics & systems pharmacology*, 8(3):131–134, 2019.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994.

Khatami, S. G., Robinson, C., Birkenbihl, C., Domingo-Fernández, D., Hoyt, C. T., and Hofmann-Apitius, M. Challenges of integrative disease modeling in alzheimer's disease. *Frontiers in Molecular Biosciences*, 6, 2019.

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

Klein, C. A. Parallel progression of primary tumours and metastases. *Nature Reviews Cancer*, 9(4):302–312, 2009.

Koutnik, J., Greff, K., Gomez, F., and Schmidhuber, J. A clockwork rnn. In *International Conference on Machine Learning*, pp. 1863–1871, 2014.

Krishnan, R. G., Shalit, U., and Sontag, D. Structured inference networks for nonlinear state space models. In *AAAI*, pp. 2101–2109, 2017.

Larson, D., Kyle, R. A., and Rajkumar, S. V. Prevalence and monitoring of oligosecretory myeloma. *The New England journal of medicine*, 367(6):580–581, 2012.

LeCun, Y. Learning invariant feature hierarchies. In *European conference on computer vision*, pp. 496–505. Springer, 2012.

Lim, B., Alaa, A., and van der Schaar, M. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, pp. 7483–7493, 2018.

Ling, J., Kurzawski, A., and Templeton, J. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, 2016.

Liu, Y.-Y., Li, S., Li, F., Song, L., and Rehg, J. M. Efficient learning of continuous-time hidden markov models for disease progression. In *Advances in neural information processing systems*, pp. 3600–3608, 2015.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Marinescu, R. V., Eshaghi, A., Lorenzi, M., Young, A. L., Oxtoby, N. P., Garbarino, S., Crutch, S. J., Alexander, D. C., Initiative, A. D. N., et al. Dive: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage*, 192:166–177, 2019.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

Mould, D., Denman, N., and Duffull, S. Using disease progression models as a tool to detect drug effect. *Clinical Pharmacology & Therapeutics*, 82(1):81–86, 2007.

Norton, L. Cancer log-kill revisited. *American Society of Clinical Oncology Educational Book*, 34(1):3–7, 2014.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. 2019.

Pearl, J. et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Perotte, A., Ranganath, R., Hirsch, J. S., Blei, D., and Elhadad, N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22(4):872–880, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Schulam, P. and Saria, S. Integrative analysis using coupled latent variable models for individualizing prognoses. *The Journal of Machine Learning Research*, 17(1):8244–8278, 2016.

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pp. 1697–1708, 2017.

Seo, S. and Liu, Y. Differentiable physics-informed graph networks. *arXiv preprint arXiv:1902.02950*, 2019.

Severson, K. A., Chahine, L. M., Smolensky, L. A., Ng, K., Hu, J., and Ghosh, S. Personalized input-output hidden markov models for disease progression modeling. *medRxiv*, 2020.

Silva, R. Observational-interventional priors for dose-response learning. In *Advances in Neural Information Processing Systems*, pp. 1561–1569, 2016.

Soleimani, H., Subbaswamy, A., and Saria, S. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press Corvallis, 2017.

Suresh, H., Hunt, N., Johnson, A., Celi, L. A., Szolovits, P., and Ghassemi, M. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pp. 322–337. PMLR, 2017.

Taghipour, S., Banjevic, D., Miller, A., Montgomery, N., Jardine, A., and Harvey, B. Parameter estimates for invasive breast cancer progression in the canadian national breast screening study. *British journal of cancer*, 108(3): 542–548, 2013.

US National Institutes of Health, and others. Relating clinical outcomes in multiple myeloma to personal assessment of genetic profile (com-mpass). *Clinical Trials website. https://clinicaltrials. gov/ct2/show/NCT01454297*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Venuto, C. S., Potter, N. B., Ray Dorsey, E., and Kieburtz, K. A review of disease progression models of parkinson's disease and applications in clinical trials. *Movement Disorders*, 31(7):947–956, 2016.

Wang, R., Walters, R., and Yu, R. Incorporating symmetry into deep dynamics models for improved generalization. *arXiv preprint arXiv:2002.03061*, 2020.

Wang, X., Sontag, D., and Wang, F. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 85–94, 2014.

West, J. and Newton, P. K. Chemotherapeutic dose scheduling based on tumor growth rates provides a case for low-dose metronomic high-entropy therapies. *Cancer research*, 77(23):6717–6728, 2017.

Xu, Y., Xu, Y., and Saria, S. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*, pp. 282–300, 2016.

Zhang, L., Lim, C. Y., Maiti, T., Li, Y., Choi, J., Bozoki, A., Zhu, D. C., of Applied Statistics of the Ministry of Education (KLAS), K. L., and for the Alzheimer's Disease Neuroimaging Initiative. Analysis of conversion of alzheimer's disease using a multi-state markov model. *Statistical methods in medical research*, 28(9): 2801–2819, 2019a.

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