Learning Disease Progression Models That Capture Health Disparities

Erica Chiang ERICACHIANG@CS.CORNELL.EDU
Divya Shanmugam DIVYAS@CORNELL.EDU
Cornell Tech, Cornell University, USA

Ashley N. Beecy ASB9028@NYP.ORG
NewYork-Presbyterian, Weill Cornell Medical College, USA

Gabriel Sayer GTS2102@CUMC.COLUMBIA.EDU
NewYork-Presbyterian, Columbia University Irving Medical Center, USA

Deborah Estrin DESTRIN@CORNELL.EDU
Nikhil Garg NGARG@CORNELL.EDU
Cornell Tech, Cornell University, USA

Emma Pierson EMMAPIERSON@BERKELEY.EDU
University of California, Berkeley, USA

Abstract

Disease progression models are widely used to inform the diagnosis and treatment of many progressive diseases. However, a significant limitation of existing models is that they do not account for health disparities that can bias the observed data. To address this, we develop an interpretable Bayesian disease progression model that captures three key health disparities: certain patient populations may (1) start receiving care only when their disease is more severe, (2) experience faster disease progression even while receiving care, or (3) receive follow-up care less frequently conditional on disease severity. We show theoretically and empirically that failing to account for any of these disparities can result in biased estimates of severity (e.g., underestimating severity for disadvantaged groups). On a dataset of heart failure patients, we show that our model can identify groups that face each type of health disparity, and that accounting for these disparities while inferring disease severity meaningfully shifts which patients are considered high-risk.

Data and Code Availability This paper uses data from the NewYork-Presbyterian (NYP)/Weill Cornell Medical Center's electronic health record (EHR) system, which is not publicly available. Code for our model and all synthetic experiments can be found at https://github.com/erica-chiang/progression-disparities.

1. Introduction

In many settings, observed data is used to model the progression of a latent variable over time. Models of human aging use a person's physical and biological characteristics to model progression of their latent "biological age" (Pierson et al., 2019); models of infrastructure deterioration use inspection results to model progression of a system's latent overall health (Madanat et al., 1995); and disease progression models, which we focus on in this paper, use observed symptoms to model progression of a patient's latent severity of a chronic disease (Wang et al., 2014). Disease progression models can help predict a patient's disease trajectory and thus personalize care, detect diseases at earlier stages, and guide drug development and clinical trial design (Mould et al., 2007; Romero et al., 2015). They have been applied to a wide variety of progressive diseases such as Parkinson's disease (Post et al., 2005), Alzheimer's disease (Holford and Peace, 1992) and cancer (Gupta and Bar-Joseph, 2008).

For the benefits of these models to apply to all patients equitably, it is crucial that they accurately describe progression for all patient populations. However, disease progression models have typically failed to account for the fact that systemic disparities in the healthcare process can bias the observed data that they are trained on. For example, disparities have been shown to arise along axes such as socioeconomic

status (Weaver et al., 2010; Miller and Wherry, 2017), race (Yearby, 2018), and proximity to care (Chan et al., 2006; Reilly, 2021). Accounting for such disparities is important because it can meaningfully shift estimates of disease progression. For intuition, imagine learning that a patient in the emergency room traveled three hours to get there; if their symptoms are ambiguous, this contextual information may increase our estimate of how severe their underlying condition is. Disease progression models have historically been unable to capture this type of context and, as we show, this can lead to biased estimates of severity. To address this, we propose a method for learning disease progression models that interpretably capture three well-documented health disparities:

1. **Disparities in initial severity.** Certain patient groups may start receiving care only when their disease is more severe (Hu et al., 2024).
2. **Disparities in disease progression rate.** Certain patient groups may experience faster disease progression, even while receiving care (Diamantidis et al., 2021).
3. **Disparities in visit frequency.** Certain patient groups may visit healthcare providers for follow-up care less frequently, even at the same disease severity (Nouri et al., 2023).

A core technical challenge we address is designing a model that is flexible enough to capture all three disparities but still identifiable. Identifiability is necessary for accurate estimates of disparities and disease progression. As such, our key contributions are: (1) we develop an interpretable Bayesian model of disease progression that accounts for multiple types of disparities but remains provably identifiable from the observed data; (2) we prove and show empirically that failing to account for any of these three disparities leads to biased estimates of severity; and (3) we characterize fine-grained disparities in a heart failure dataset. Our model reveals that non-white patients have more severe heart failure and face multiple types of health disparities: Black and Asian patients tend to start receiving care at more severe stages of heart failure than do White patients, and Black patients see healthcare providers for heart failure 10% less frequently than do White patients at the same disease severity level. Accounting for these disparities meaningfully shifts our estimates of disease severity, increasing the fraction of Black and Hispanic patients identified as high-risk.

While we ground this work in healthcare, our method for learning progression models that account for disparities applies naturally to many other progression model settings where disparities are of interest, including infrastructure deterioration (Madanat et al., 1995) and human aging (Pierson et al., 2019).

2. Related Work

Disease progression modeling. Disease progression models have been developed for many chronic diseases, including Parkinson's disease (Post et al., 2005), Alzheimer's disease (Holford and Peace, 1992), diabetes (Perveen et al., 2020), and cancer (Gupta and Bar-Joseph, 2008). A key feature of the progression models we consider, common in the machine learning literature, is that a latent severity Z_t progresses over time and gives rise to a set of observed symptoms X_t. Models in this family include variants of hidden Markov models (HMMs) (Wang et al., 2014; Liu et al., 2015; Alaa and Hu, 2017; Sukkar et al., 2012; Jackson et al., 2003) and recurrent neural networks (RNNs) (Choi et al., 2016b; Lipton et al., 2017; Lim and van der Schaar, 2018; Choi et al., 2016a; Ma et al., 2017; Kwon et al., 2019; Alaa and van der Schaar, 2019). This existing literature has not focused on modeling disparities; we extend it by proposing a new approach to disease progression modeling that can interpretably characterize and account for multiple types of health disparities.

Health disparities. Disparities have been documented in many parts of the healthcare process. Factors such as distance from hospitals (Reilly, 2021), distrust of the healthcare system (LaVeist et al., 2009), or lack of insurance (Venkatesh et al., 2019) can result in underutilization of health services; biases in the judgements of healthcare providers can lead minority groups to receive later screening (Lee et al., 2021), fewer referrals (Landon et al., 2021), or generally worse care (Schäfer et al., 2016); and issues such as limited health literacy or trust can create disparities in follow-through for appointments or the effectiveness of at-home care (Davis, 1968; Brandon et al., 2005).

The existing literature has shown that disparities emerge along the three axes that we capture in this paper: (1) how severe a patient's disease becomes before they start to receive care (Chen et al., 2021; Iqbal et al., 2015; Hu et al., 2024); (2) how quickly their latent severity progresses even while receiving care

(Diamantidis et al., 2021; Suarez et al., 2018); and (3) how likely they are to visit a healthcare provider at a given severity level (Nouri et al., 2023). Our goal is to show how accounting for disparities along all three of these axes improves the severity estimates of disease progression models, while also learning more fine-grained descriptions of disparities.

Capturing disparities with machine learning. We build upon a large body of past work that uses machine learning as a tool to capture and address health disparities, including models that estimate the relative prevalence of underreported medical conditions (Shanmugam et al., 2021), improve risk prediction for patients with missing outcome data (Balachandar et al., 2023), evaluate the impact of race corrections in risk prediction (Zink et al., 2023), assess disparate impacts of AI in healthcare (Chen et al., 2019), and quantify disparities in the performance of clinical prediction tasks (Zhang et al., 2020). The closest work to our own is Chen et al. (2021), which develops a clustering algorithm that accounts for the fact that some patients do not come in (and are therefore not observed) until later in their disease progression. While their work addresses one form of data bias that can arise due to health disparities, it differs from our own in two ways: it does not specifically document or study health disparities, and it focuses on clustering patients as opposed to modeling disease severity or progression. Our work proposes a model for capturing three types of health disparities in the disease progression setting in order to learn precise descriptions of multiple disparities and make severity estimates that exhibit less bias than existing disease progression models.

3. Model

We build on a standard setup for disease progression modeling, in which each patient has an underlying latent disease severity Z_t that progresses over time and gives rise to a set of observed features X_t.

We characterize each patient's severity $Z_t \in \mathbb{R}$ at time t by their *initial severity* Z_0 at their first observation (which we denote as $t = 0$) and their *rate of progression* R after that point:

$$Z_t = Z_0 + R \cdot t$$

If a patient visits a healthcare provider at time t, we observe some recorded set of *features* $X_t \in \mathbb{R}^d$ (e.g., lab results, imaging, symptoms). At any given visit, a clinician does not necessarily observe or record all features—we model the features that *are* observed as a noisy function of the patient's latent severity Z_t:

$$X_t = f(Z_t) + \epsilon_t$$

$$\epsilon_t \sim \mathcal{N}(0, \Psi)$$

where the diagonal covariance matrix $\Psi \in \mathbb{R}^{d \times d}$ parameterizes feature-specific noise (accounting for both measurement error and variation in how the patient's physical state can fluctuate day-to-day). In our experiments, we specifically instantiate f as a linear function $f(Z_t) = F \cdot Z_t + b$, where $F \in \mathbb{R}^d$ is a feature-specific scaling factor and $b \in \mathbb{R}^d$ is a feature-specific intercept, but our approach extends to more general parametric forms for f. We constrain the first feature $F_0 > 0$ using domain knowledge; this restriction is necessary for identifiability because it restricts the mapping between features and severity (Shapiro, 1985). We also observe a set of time values at which a patient visits a healthcare provider; we encode this with a binary indicator $D_t \in \{0, 1\}$ that is equal to 1 if a patient has a visit at time t and 0 otherwise.

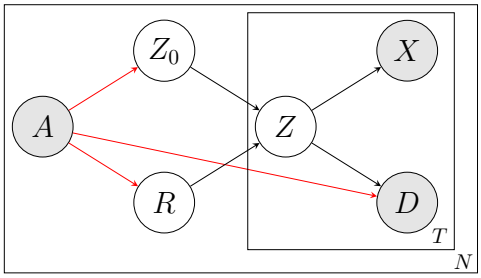

Figure 1: **Disease progression generative model.** Plate diagram captures N patients over T timesteps. Shaded nodes indicate observed features: demographics $A^{(i)}$, visit indicator $D_t^{(i)}$, and symptoms $X_t{}^{(i)}$ (only observed when $D_t^{(i)} = 1$). Unshaded nodes indicate latent variables: a patient's initial severity $Z_0{}^{(i)}$, rate of progression $R^{(i)}$, and severity $Z_t{}^{(i)}$. Red arrows indicate dependencies capturing health disparities.

Capturing disparities. Our model captures the three types of health disparities discussed in §2 by allowing model parameters to vary as a function of a patient's demographic feature vector A. For expositional clarity, we describe a setup where A encodes a single categorical label (e.g., a patient's race group), but our approach naturally extends to multiple categorical groupings or to continuous features.

1. **Disparities in initial severity.** Underserved patients may start receiving care only when their disease is more severe. We capture this by learning group-specific distributions of Z_0, a patient's disease severity at their first visit. For one group $A = a_0$, we pin Z_0 to be drawn from a unit normal distribution; this is a standard and necessary identifiability condition since it fixes the scale of Z_t (Shapiro, 1985). For other groups a, we model

$$Z_0 \sim \mathcal{N}\left(\mu_{Z_0}^{(a)}, {\sigma_{Z_0}^{(a)}}^2\right)$$

where $\mu_{Z_0}^{(a)}$ and $\sigma_{Z_0}^{(a)}$ are learned group-specific parameters.

2. **Disparities in disease progression rate.** Underserved patients may experience faster disease progression even while receiving care. We capture this by learning group-specific distributions of disease progression rate R:

$$R \sim \mathcal{N}\left(\mu_R^{(a)}, {\sigma_R^{(a)}}^2\right)$$

where $\mu_R^{(a)}$ and $\sigma_R^{(a)}$ are learned group-specific parameters for each group a.

3. **Disparities in visit frequency.** Underserved patients may visit healthcare providers for follow-up care less frequently at a given disease severity. We capture this by modeling patient visits as generated by an inhomogeneous Poisson process, parameterized by a time-varying rate parameter λ_t that depends on both Z_t and A:

$$\log(\lambda_t) = \beta_0 + \beta_Z \cdot Z_t + \beta_A^{(a)}$$

where β_Z and β_0 are learned parameters for the entire population and $\beta_A^{(a)}$ is a learned group-specific parameter for each group a (we pin $\beta_A^{(a_0)} = 0$ for reference).

Overall, our model parameters (on which we place weakly informative priors) are the parameters shared across groups $\{F,\ b,\ \Psi,\ \beta_0,\ \beta_Z\}$, and the group-specific parameters $\{\mu_{Z_0}^{(a)},\ \sigma_{Z_0}^{(a)},\ \mu_R^{(a)},\ \sigma_R^{(a)},\ \beta_A^{(a)}\}$. We learn posterior distributions over these parameters from our observed data $\{X_t, D_t, A\}$ using Hamiltonian Monte Carlo, a standard algorithm for Bayesian inference (Betancourt, 2018), as implemented in Stan (Carpenter et al., 2017). Figure 1 summarizes the data generating process and Table 1 summarizes the notation for our model.

Notation	Meaning
X_t	Observed features at time t
D_t	Binary visit indicator for time t
A	Demographic features
Z_t	Disease severity at time t
Z_0	Initial severity
R	Disease progression rate
F	Severity-feature matrix
b	Feature intercepts
Ψ	Feature covariance matrix
μ_{Z_0}, σ_{Z_0}	Group-specific mean and sd of Z_0
μ_R, σ_R	Group-specific mean and sd of R
λ_t	Visit rate at time t
β_0	Visit rate intercept
β_Z	Visit rate Z_t coefficient
β_A	Visit rate A coefficient

Table 1: **Summary of notation.** Observed data are listed above the double horizontal line.

Model discussion. Our model makes several common assumptions. First, we model event frequency with a Poisson process; this is a common approach, including in work that seeks to capture disparities in event frequency (Liu et al., 2024; Kurashima et al., 2018). Second, we model progression as linear over time, a common approach for learning interpretable characterizations of trajectories (Holford and Peace, 1992; Kimko, 2000; Pierson et al., 2019). Finally, our assumption of a linear relationship between the latent state Z_t and feature values X_t is also standard; for example, factor analysis makes this assumption. While our linearity assumptions may limit our model's ability to capture some nuances of disease progression, they allow the model to interpretably capture progression trends over time; interpretability is especially valuable in our setting, allowing us to directly compare quantities like initial severity and progression rate across patient subgroups.

4. Theoretical analysis

In this section, we prove two main theoretical results. First, we show that our model is *identifiable*, a necessary condition for its parameters to be estimated from the observed data. Interpreting these parameter estimates is what allows us to quantify disparities. Second, we prove that failing to account for disparities produces *biased estimates of severity*. We sum-

marize proof strategies in the main text and provide formal proofs in Appendices A and B.

4.1. Identifiability

We show that our model is identifiable, meaning different sets of parameters yield different observed data distributions (Bellman and Åström, 1970), which is necessary to correctly estimate model parameters from the observed data. Learning a model of progression that is *flexible* enough to characterize multiple disparities but *still identifiable* is a core challenge our work addresses. In fact, if we added one more dependence on A—in particular, adding an arrow from A to X in Figure 1—the model would no longer be identifiable (without a shared interpretation across groups of how features map to severity, it would be impossible to identify disparities in disease progression).

Theorem 1 *All model parameters are identified by the observed data distribution* $P(X_t, D_t \mid A)$.

As mentioned in §3, the distribution of initial severity Z_0 is pinned to a unit normal for one demographic group a_0. This pinned distribution reduces the number of unknown latent parameters for group a_0, allowing us to show that $\{F, b, \Psi\}$ are identified by $P(X_t \mid A = a_0)$. Having identified these, we show that the parameters $\{\mu_{Z_0}^{(a)}, \sigma_{Z_0}^{(a)}, \mu_R^{(a)}, \sigma_R^{(a)}\}$ are identified by $P(X_t \mid A = a)$ for all groups a. Finally, we show that given the previously identified parameters, $\{\beta_0, \beta_Z\}$ are identified by $P(D_t \mid A = a_0)$ and $\{\beta_A^{(a)}\}$ is identified by $P(D_t \mid A = a)$ for all other groups a. We provide a full proof in §A.1.

4.2. Bias in models that do not account for disparities

Next we show that, when any of the health disparities we discuss are present, a model that does not account for group-specific disparities will produce *biased estimates* of severity—i.e., $\mathbb{E}[Z_t \mid X_t, D_t] \neq \mathbb{E}[Z_t \mid X_t, D_t, A = a]$. These theoretical results hold whenever the model dependencies are encoded by the graph in Figure 1, a more general assumption than our full parametric model. For each proof, we analyze the effect of one disparity—e.g., for disparities in initial severity, we assume that $P(Z_0 \mid A)$ differs across groups—while keeping other distributions constant across groups. The results hold in the presence of multiple disparities as long as the existing disparities disfavor or favor the same group, so as to not cancel each other out in their effects.

To quantify disparities, we use the strict Monotone Likelihood Ratio Property (MLRP) to reason about the probability density functions of initial severity and progression rate for certain groups, relative to the overall population (Karlin and Rubin, 1956):

Definition 2 *Two distributions characterized by probability density functions* $f(x)$ *and* $g(x)$ *have the strict monotone likelihood ratio property in* x *if* $\frac{f(x)}{g(x)}$ *is a strictly increasing function of* x.

Intuitively, this means that as some variable x (Z_0 or R, in our case) gets larger, it is more likely to be drawn from f than g. The MLRP is a widely-used assumption across many settings (Gaebler and Goel, 2024; Anwar and Fang, 2006; Chemla and Hennessy, 2019); the normal, exponential, binomial, and Poisson families all have this property. For brevity, we say "$f(x)$ strictly MLRPs $g(x)$" to mean that $f(x)$ and $g(x)$ satisfy the strict MLRP in x. We now prove for each disparity that any model failing to account for the disparity will produce biased estimates of severity.

Theorem 3 *A model that does not take into account disparities in initial disease severity* Z_0 *will underestimate the disease severity of groups with higher initial severity and overestimate that of groups with lower initial severity. Specifically, if* $P(Z_0 \mid A = a)$ *strictly MLRPs* $P(Z_0)$ *for some group* a*, then* $\mathbb{E}[Z_t \mid X_t] < \mathbb{E}[Z_t \mid X_t, A = a]$. *Similarly, if* $P(Z_0)$ *strictly MLRPs* $P(Z_0 \mid A = a)$ *for some group* a*, then* $\mathbb{E}[Z_t \mid X_t] > \mathbb{E}[Z_t \mid X_t, A = a]$.

We prove this by showing that $P(Z_0 \mid X_t, A = a)$ strictly MLRPs $P(Z_0 \mid X_t)$, which implies that $\mathbb{E}[Z_t \mid X_t, A = a] > \mathbb{E}[Z_t \mid X_t]$; §B.1 provides a full proof.

Theorem 4 *Suppose disease severity progresses linearly at some rate* R. *A model that does not take into account disparities in* R *will underestimate the disease severity of groups with higher progression rates and overestimate that of groups with lower progression rates. Specifically, if* $P(R \mid A = a)$ *strictly MLRPs* $P(R)$ *for some group* a*, then* $\mathbb{E}[Z_t \mid X_t] < \mathbb{E}[Z_t \mid X_t, A = a]$. *Similarly, if* $P(R)$ *strictly MLRPs* $P(R \mid A = a)$ *for some group* a*, then* $\mathbb{E}[Z_t \mid X_t] > \mathbb{E}[Z_t \mid X_t, A = a]$.

We use a similar proof technique as for Theorem 3 and provide a full proof in §B.2.

Finally, we analyze disparities in visit frequency. For this portion of the theoretical analysis, we consider discrete, non-infinitesimal time intervals (with

the i-th interval starting at time t_i, where $t_0 = 0$) and whether or not a patient visited at any point during each interval. We introduce a new random variable E_{t_i} to indicate whether the patient has any visits during the time interval starting at t_i (i.e., whether $D_t = 1$ for some t in $[t_i, t_{i+1})$). We show that conditioned on E_t, the group with lower visit frequency has higher expected severity at the beginning of the interval.

Theorem 5 *A model that does not take into account disparities in visit frequency conditional on disease severity will underestimate the disease severity of groups with lower visit frequency conditional on severity and overestimate the disease severity of groups with higher visit frequency conditional on severity. Specifically, assume that $P(E_t = 1 \mid Z_t)$ is strictly monotone increasing in Z_t, $\lim_{Z_t \to -\infty} P(E_t = 1 \mid Z_t) = 0$, and $\lim_{Z_t \to \infty} P(E_t = 1 \mid Z_t) = 1$. Then if some group a has a lower probability of visiting a healthcare provider at any given severity level—that is, $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z - \alpha(z))$ for all z, where $\alpha(z)$ is a positive function of z—then $\mathbb{E}[Z_t \mid E_t] < \mathbb{E}[Z_t \mid E_t, A = a]$. Similarly, if $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z + \alpha(z))$ for all z, where $\alpha(z)$ is a positive function of z, then $\mathbb{E}[Z_t \mid E_t] > \mathbb{E}[Z_t \mid E_t, A = a]$.*

We prove this by directly reasoning about the expected value of Z_t when conditioning on group versus not; we provide a full proof in §B.3.

Overall, these results convey the importance of accounting for disparities in disease progression models: it is fundamentally not possible to make well-calibrated estimates of severity without accounting for group differences in initial severity, progression rate, and visit frequency.

5. Synthetic experiments

In this section, we validate our model and theoretical results in synthetic data simulations. We generate synthetic datasets according to the modeling assumptions in §3 (with parameter values for each dataset drawn randomly from each parameter's prior distribution). For each dataset, we generate simulated data for two separate groups, differing in their distributions of initial severity, progression rate, and visit frequency (characterized by different μ_{Z_0}, μ_R, and β_A, respectively).

5.1. Identifiability and Severity Estimation

We first verify Theorem 1 in simulations, showing that when we fit our model on synthetic data, it accurately recovers the true data-generating parameters. We do this by examining the concordance between the model's estimated parameters and the true, latent parameter values, a common approach in past work (Chang et al., 2021; Pierson et al., 2019). We find high correlation between the true parameters and our model's posterior mean estimates (mean Pearson's r 0.96 across all parameters; median 0.99), and good calibration (mean linear regression slope 1.0; median 1.0 when fit without an intercept term). We provide scatterplots of true versus estimated values for all parameters in Appendix C. We also see that our model's mean *severity estimates* for each group are highly correlated and well-calibrated with ground truth, despite underlying differences in group severity distributions and visit rates (Figure 2).

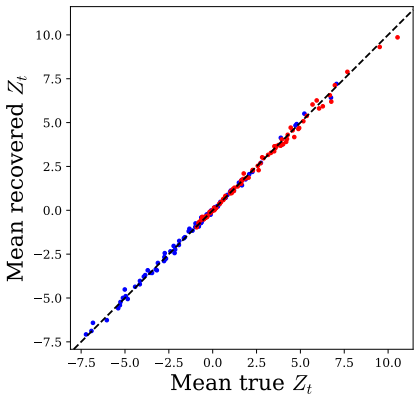

Figure 2: **Well-calibrated severity estimates.** Each dot shows the mean true vs. mean recovered severity values for one group in a given simulation trial. Groups depicted in red tend to be underserved compared to groups depicted in blue. Our full model produces accurate and well-calibrated severity estimates (estimates lie near dotted $y = x$ line).

5.2. Bias in models that do not account for disparities

We now demonstrate in simulation that failing to account for disparities can lead to biased severity estimates, consistent with Theorems 3, 4, and 5. In each trial, we use the same data to fit four models: our full model, which accounts for all disparities, plus three ablated models that each fail to account for one of the

disparities (initial severity, progression rate, visit frequency). To characterize the resulting bias of failing to account for each type of disparity, we compute the average error in severity estimates (mean inferred estimate minus mean true severity) of each model, broken down by group. For each ablated model and trial, we define the "underserved group" to be the one that is underserved with respect to the specific disparity that the model fails to capture. When evaluating our full model, we define the "underserved group" to be the one with higher initial severity.

As seen in Table 2, the models that do not account for disparities produce biased estimates: while our full model achieves average error across all trials of -0.02 for underserved patient groups and 0.01 for other patient groups, the ablated models all have negative error for underserved patients (underestimated severity) and positive error for other patients (overestimated severity). The ablated models also produce severity estimates that are less correlated with true severity.

6. Modeling health disparities in heart failure progression

Finally, we fit our model on a real-world dataset of heart failure patients in the NewYork-Presbyterian hospital system. Heart failure is a progressive disease that affects many people, requires both specialty and preventive care (Colucci et al., 2020), and has known health disparities (Lewsey and Breathett, 2021), making it a natural application setting for our model. In §6.1 we summarize the dataset, and in §6.2 we confirm that our model can learn meaningful low-dimensional representations of disease severity by evaluating its reconstruction and predictive performance compared to standard baselines. In §6.3 we present our main results: we interpret our model's learned parameters to provide precise descriptions of health disparities in our setting, and we show that (as our theory predicts) failing to account for these disparities meaningfully shifts severity estimates.

6.1. Data

Our data comes from the NewYork-Presbyterian (NYP)/Weill Cornell Medical Center's electronic health record (EHR) system from 2012-2020. We analyze a cohort of $N = 2,942$ patients who (1) have a specific subtype of heart failure (heart failure with reduced ejection fraction), to ensure our cohort can be described by a single progression model, and (2) are likely to receive most of their cardiology care in the NYP system, to ensure we can reasonably estimate when they receive care.

Observed feature data X_t for each patient includes four types of measurements: left ventricle ejection fraction (LVEF), brain natriuretic peptide (BNP), systolic blood pressure (SBP), and heart rate (HR). LVEF and BNP have strong clinical associations with heart failure severity (in terms of both underlying physiological health and observed symptoms) (Murphy et al., 2020). SBP and HR are less informative (more prone to fluctuation and changes not related to heart failure), but they are still expected to show general trends over time as a patient's heart failure progresses. Since we must pin the sign of at least one scaling factor F for identifiability, and decreasing LVEF is strongly associated with increasing severity in the heart failure subtype we study, we pin the sign of the scaling factor between severity Z_t and LVEF values ($F_{\text{LVEF}} < 0$).

Measurements close in time are often from the same hospital visit, so we combine measurements within the same week (which has the additional benefits of increasing the speed of model fitting and allowing us to focus on capturing longer-term changes in disease severity). Specifically, for each week, we average together all measurements of the same type and treat any measurements as if they occurred at the beginning of the week. We then capture disparities across four self-reported race/ethnicity groups: White non-Hispanic patients, Black non-Hispanic patients, Hispanic patients, and Asian non-Hispanic patients (which we will hereby describe as White, Black, Hispanic, and Asian subgroups). A full description of our data processing can be found in Appendix D.

6.2. Model validation

We first confirm that our model accurately fits the data: we verify that the model's inferred parameters are consistent with medical knowledge (§6.2.1) and compare the model's reconstruction and predictive performance to standard baselines (§6.2.2). We then show in §6.3, as our primary result, that our model provides insight into disparities in disease progression.

6.2.1. Consistency with medical knowledge

Figure 3 plots our model's inferred parameters, all of which are consistent with existing medical knowl-

		Model that fails to account for disparities in...		
	Full model	*Initial severity*	*Progression rate*	*Visit frequency*
Underserved group bias	-0.02	-0.89	-0.04	-0.37
Non-underserved group bias	0.01	+1.02	+0.20	+0.33
Underserved group correlation	1.00	0.72	0.99	0.94
Non-underserved group correlation	1.00	0.73	0.90	0.97

Table 2: **Failing to account for disparities produces biased estimates of severity Z_t.** We compare severity estimates from our full model to three ablated models that each fail to account for one of the three health disparities. While our full model produces accurate, well-calibrated severity estimates, each ablated model underestimates severity for the underserved group and overestimates it for the other group. The ablated model estimates are also *less correlated* with the true severity values.

edge.[1] Specifically, (1) the model correctly learns that BNP and HR tend to increase with heart failure severity ($F_{\text{BNP}}, F_{\text{HR}} > 0$), while SBP tends to decrease ($F_{\text{SBP}} < 0$) (Murphy et al., 2020); (2) the model learns larger variance parameters for SBP and HR values (ψ), correctly inferring that these features are less informative about heart failure progression than BNP and LVEF (Murphy et al., 2020); and (3) the model estimates that $\beta_Z > 0$, meaning it learns that patients with higher disease severity tend to see healthcare providers more frequently, as expected.

6.2.2. Reconstruction and predictive performance

We next evaluate the model's ability to reconstruct and predict patient features X_t. Because the model represents each patient visit in terms of a scalar severity Z_t, we do not expect the model to perfectly reconstruct or predict the multi-dimensional X_t; rather, we hope for predictions that correlate significantly with X_t. Consistent with this, when fit on 3 years of data per patient, our model's predicted feature values correlate with true values both in- and out-of-sample. As we would hope, the model best represents the features that are known to be most informative about heart failure progression—LVEF ($r = 0.81$ in-sample, $r = 0.51$ out-of-sample) and BNP ($r = 0.62$ in-sample, $r = 0.31$ out-of-sample)—as opposed to the less-informative features SBP ($r = 0.42$ in-sample, $r = 0.24$ out-of-sample) and HR ($r = 0.17$ in-sample,

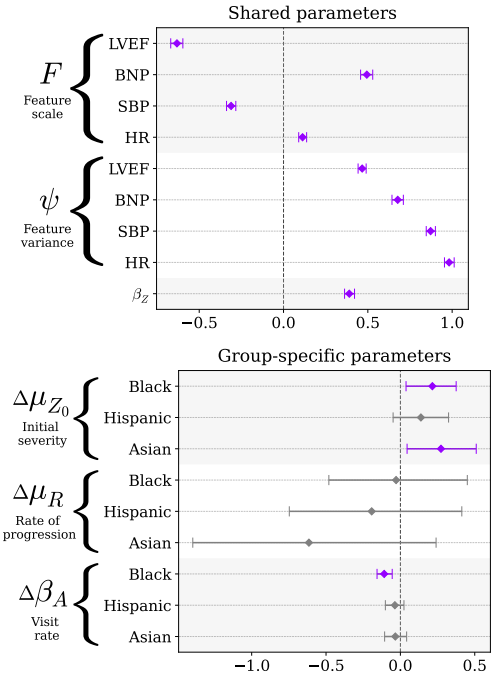

Figure 3: **Inferred model parameters with 95% confidence intervals.** Shared parameters (top) are consistent with medical knowledge of heart failure progression. Group-specific parameters (bottom) are plotted as differences compared to White patients, so confidence intervals that are non-overlapping with 0 (colored in purple) indicate significant racial/ethnic differences in parameters.

1. For succinctness, Figure 3 plots only the model parameters of primary interest for interpreting our model (omitting, for example, estimated intercepts for each feature); a similar coefficient plot with all learned parameters is shown in Figure S5.

$r = 0.03$ out-of-sample); all p-values besides HR out-of-sample < 0.001.

To provide a more detailed assessment of performance, we evaluate our model's ability to *reconstruct* patient features X_t in-sample and *predict* X_t out-of-sample, in comparison to seven standard baselines. While predicting and reconstructing X_t is not the primary goal of our model, the model performs generally well relative to these baselines, validating its ability to meaningfully represent the data.

All of the baselines are designed to reconstruct or predict only the feature values X_t; our model can also predict the occurrence of patient visits (D_t), but in order to provide a direct comparison of reconstruction and predictive performance, we compare only the feature prediction aspect of our model (so we do not fit any models using D_t data) in this subsection. We use mean absolute percentage error (MAPE) to report a normalized measure of error across features.

Reconstruction performance. We compare our model's reconstruction performance to that of two standard *dimensionality reduction baselines*: principal component analysis (PCA) and factor analysis (FA). We compare our model to two variants of each. First, we compare our model to PCA and FA fit at the *visit level*: one component per patient visit, analogous to our model's Z_t. Second, we compare our model to PCA and FA fit at the *patient level*: two components for each patient, to capture the trajectory of feature values as we do with Z_0 and R. We describe the implementation of these baselines with more detail in Appendix E. Because both PCA and FA require input vectors of consistent size, all models are fit on feature values from the first three visits per patient. Compared to all baselines, we achieve equivalent or better reconstruction performance across all features, and better performance on the more informative features (Table S1).

Predictive Performance. We also compare our model's predictive performance to that of three standard *timeseries forecasting baselines*: (1) a linear regression for each patient and feature; (2) a quadratic regression for each patient and feature; and (3) predicting values equal to those at the last timestep in training data. For this comparison, all models are fit on feature values from the first three years of data per patient, and we evaluate predictive performance on all remaining visits. While prediction is not the primary goal of our model (and models with relatively low predictive performance can still provide useful insights on disparities (Pierson et al., 2021)), these results serve as an additional validation of our model's ability to meaningfully represent the data. Our model outperforms both linear regression and quadratic regression on all features. Our model has slightly higher MAPE than latest timestep, which is a widely-used, strong baseline for pure predictive performance (Hyndman, 2018); latest timestep does not, however, provide any insight into disparities or even patterns of progression over time (Table S2).

6.3. Analysis of disparities

We now discuss three main findings from fitting our model on the heart failure data. We learn that (1) Black patients tend to have higher disease severity than White patients; (2) our model learns precise descriptions of health disparities and finds that disparities of multiple types exist in our setting; and (3) failing to account for the existing disparities meaningfully shifts severity estimates for all racial/ethnic groups. This analysis is descriptive and does not require evaluating held-out performance, so models are fit on all available data.

Black patients have higher disease severity. As seen in Figure 4, our model infers that Black patients have significantly higher disease severity than White patients ($p < 0.05$, computed by cluster bootstrapping at the patient-level; Hispanic and Asian patients also have higher inferred mean severity than White patients, but the differences are not statistically significant).

Model parameters capture fine-grained disparities. As seen in Figure 3 (bottom), our model infers that Black and Asian patients have significantly higher initial severity than do White patients (inferred average initial severity μ_{Z_0} for Black and Asian patient groups is greater than for White patients by 0.22 and 0.27, respectively). To contextualize the magnitude of these disparities, if all patients progressed at the average learned progression rate across the entire population, Black patients' first heart failure visit would occur 3.0 years later in their disease progression than White patients', and Asian patients' first visit would occur 3.8 years later. We also observe that β_A for Black patients is significantly lower than that of White patients, indicating that Black patients visit healthcare providers 10% less frequently than White patients with the same disease severity. We describe these calculations in Appendix F.

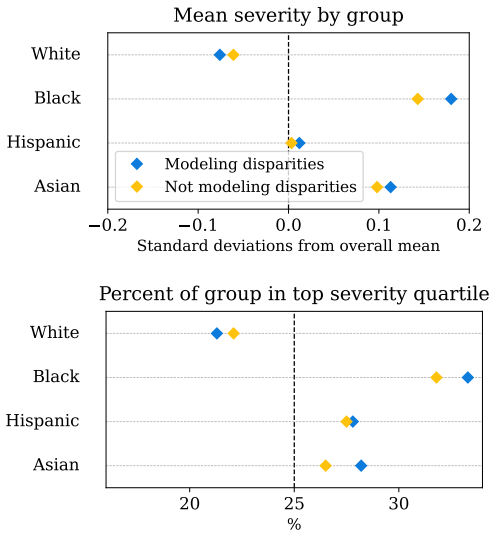

Figure 4: **Accounting for disparities leads to less biased severity estimates.** We visualize the improvement of our full model (blue) over one that does not account for disparities but is otherwise the same (yellow) in two ways. On the top, we show each group's average difference from the overall mean severity, normalized by the overall standard deviation of severity. On the bottom, we capture the portion of each group that is identified as "high-risk" (top quartile of disease severity).

Accounting for disparities increases estimated severity for all non-white patient groups. We compare severity estimates from our model to those of an ablated model that does not account for disparities (but is otherwise identical) and find that this meaningfully shifts severity estimates (Figure 4 top): while both models learn that non-white patients tend to have higher severity, the ablated model produces higher severity estimates for White patients and lower estimates for other groups ($p < 0.001$ for all groups, computed by cluster bootstrapping at the patient-level). This is consistent with our theoretical results.

To highlight some implications of these shifted severity estimates, we look at each model's ranking of patient severity values and profile of "high-risk" visits: visits where inferred severity lies in the top quartile (25%) of all visits. The ablated model is less likely to rank Black or Hispanic patient visits as high risk (Figure 4 bottom; $p < 0.05$, computed by cluster bootstrapping at the patient-level), skewing the demographics of the high-risk patient cohort *away* from groups that we know to have higher disease severity.

7. Discussion

In this paper, we formalize three specific types of health disparities that bias observed health data: underserved patients may (1) first receive care only when their disease is more severe, (2) progress faster even while receiving care, or (3) receive care less frequently even at the same disease severity. We prove that failing to account for any of these disparities while learning disease progression can lead to biased estimates of severity, and we develop a disease progression model to capture all three disparities while provably retaining interpretability and identifiability.

Our model can be used to make less biased severity estimates from patient health data *and* to learn fine-grained descriptions of disparities in observational health data. Using a real-world heart failure dataset, we show that accounting for health disparities does indeed meaningfully shift severity estimates (by increasing the proportion of non-white patients identified as high-risk) and validate the model's ability to identify groups that face each type of health disparity. We thus urge future work in disease progression modeling to account for disparities in healthcare; we lay a foundation for doing so by developing a method to (1) make disease severity estimates that are accurate across diverse populations of patients and (2) learn interpretable estimates of distinct disparities that can inform future public health interventions.

There are several natural directions for future work. First, beyond heart failure, our approach could be applied to many other progressive diseases, including Parkinson's (Post et al., 2005), Alzheimer's (Holford and Peace, 1992), diabetes (Perveen et al., 2020), and cancer (Gupta and Bar-Joseph, 2008), where it is possible that disparities manifest differently. Future work should similarly validate findings across multiple sites to assess generalizability of findings. To improve prediction and allow our model to more accurately capture rich medical data sources, another interesting technical direction is to extend our model to use additional data modalities (e.g., medical images) or more flexible progression models (e.g., non-linear trajectories), while retaining its interpretability and identifiability. Finally, our approach generalizes naturally to progression model settings beyond healthcare where disparities are of interest, including infrastructure deterioration (Madanat et al., 1995) and human aging (Pierson et al., 2019); these would be interesting domains for future work.

Institutional Review Board (IRB) This study was approved by the institutional review boards at Weill Cornell Medicine (IRB #22-06024904) and Columbia University Irving Medical Center (IRB #AAAU1701). A waiver for informed consent was obtained.

Acknowledgments

EC is supported by NSF DGE #2139899. NG is supported by NSF CAREER #2339427, and Cornell Tech Urban Tech Hub, Meta, and Amazon research awards. EP is supported by a Google Research Scholar award, NSF CAREER #2142419, a CIFAR Azrieli Global scholarship, a gift to the LinkedIn-Cornell Bowers CIS Strategic Partnership, the Abby Joseph Cohen Faculty Fund, an AI2050 Early Career Fellowship, and the Survival and Flourishing Fund. This work was partially supported by funding from NewYork-Presbyterian for the NYP-Cornell Cardiovascular AI Collaboration, as well as funding from the MacArthur Foundation. Thank you to Gabriel Agostini, Sidhika Balachandar, Sarah Cen, Serina Chang, Evan Dong, Albert Gong, Sophie Greenwood, Evelyn Horn, Rishi Jha, Chris Kelsey, Konwoo Kim, Mitchel Lang, Benjamin Lee, Zhi Liu, Linda Lu, Rajiv Movva, Chidozie Onyeze, Marios Papachristou, Tom Reilly, Jeffrey Ruhl, Shuvom Sadhuka, and Naomi Tesfuzigta for valuable conversations and feedback on this paper.

References

Ahmed M Alaa and Scott Hu. Learning from Clinical Judgments: Semi-Markov-Modulated Marked Hawkes Processes for Risk Prognosis. 2017.

Ahmed M. Alaa and Mihaela van der Schaar. Attentive State-Space Modeling of Disease Progression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Shamena Anwar and Hanming Fang. An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence. *American Economic Review*, 96(1):127–151, February 2006.

Sidhika Balachandar, Nikhil Garg, and Emma Pierson. Domain constraints improve risk prediction when outcome data is missing. 2023.

R. Bellman and K.J. Åström. On structural identifiability. *Mathematical Biosciences*, 7(3-4):329–339, April 1970.

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo, July 2018.

Dwayne T. Brandon, Lydia A. Isaac, and Thomas A. LaVeist. 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):951–956, July 2005.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. *Stan* : A Probabilistic Programming Language. *Journal of Statistical Software*, 76 (1), 2017.

Leighton Chan, L. Gary Hart, and David C. Goodman. Geographic Access to Health Care for Rural Medicare Beneficiaries. *The Journal of Rural Health*, 22(2):140–146, April 2006.

Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, January 2021.

Gilles Chemla and Christopher A. Hennessy. Controls, belief updating, and bias in medical rcts. *Journal of Economic Theory*, 184:104929, 2019.

Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA Journal of Ethics*, 21(2):E167–179, February 2019.

Irene Y. Chen, Rahul G. Krishnan, and David Sontag. Clustering Interval-Censored Time-Series for Disease Phenotyping, December 2021.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. 2016a.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR workshop and conference proceedings*, 56:301–318, August 2016b.

Wilson S Colucci, SS Gottlieb, and SB Yeon. Overview of the management of heart failure with reduced ejection fraction in adults. *U: UpToDate, Gottlieb SS ed. UpToDate [Internet]. Waltham, MA: UpToDate*, 2020.

Milton S. Davis. Physiologic, Psychological and Demographic Factors in Patient Compliance with Doctors' Orders. *Medical Care*, 6(2):115–122, 1968.

Clarissa Jonas Diamantidis, Lindsay Zepel, Virginia Wang, Valerie A. Smith, Sarah Hudson Scholle, Loida Tamayo, and Matthew L. Maciejewski. Disparities in Chronic Kidney Disease Progression by Medicare Advantage Enrollees. *American Journal of Nephrology*, 52(12):949–957, 2021.

Johann D. Gaebler and Sharad Goel. A Simple, Statistically Robust Test of Discrimination, July 2024. arXiv:2407.06539 [stat].

A. Gupta and Z. Bar-Joseph. Extracting Dynamics from Static Cancer Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(2):172–182, April 2008.

Stefanie Hendricks, Iryna Dykun, Bastian Balcer, Matthias Totzeck, Tienush Rassaf, and Amir A. Mahabadi. Higher BNP/NT-pro BNP levels stratify prognosis equally well in patients with and without heart failure: a meta-analysis. *ESC Heart Failure*, 9(5):3198–3209, October 2022.

N H Holford and K E Peace. Methodologic aspects of a population pharmacodynamic model for cognitive effects in Alzheimer patients treated with tacrine. *Proceedings of the National Academy of Sciences*, 89(23):11466–11470, December 1992.

Xiao Hu, John W Melson, Stacey S Pan, Yana V Salei, and Yu Cao. Screening, Diagnosis, and Initial Care of Asian and White Patients With Lung Cancer. *The Oncologist*, 29(4):332–341, April 2024.

RJ Hyndman. *Forecasting: principles and practice*. OTexts, 2018.

Javaid Iqbal, Ophira Ginsburg, Paula A. Rochon, Ping Sun, and Steven A. Narod. Differences in Breast Cancer Stage at Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. *JAMA*, 313(2):165, January 2015.

Christopher H. Jackson, Linda D. Sharples, Simon G. Thompson, Stephen W. Duffy, and Elisabeth Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, July 2003.

Samuel Karlin and Herman Rubin. The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, 27(2):272–299, 1956.

H Kimko. Prediction of the outcome of a phase 3 clinical trial of an antischizophrenic agent (quetiapine fumarate) by simulation with a population pharmacokinetic and pharmacodynamic model. *Clinical Pharmacology & Therapeutics*, 68(5):568–577, November 2000.

Ben Klemens. When Do Ordered Prior Distributions Induce Ordered Posterior Distributions? *SSRN Electronic Journal*, 2007.

Takeshi Kurashima, Tim Althoff, and Jure Leskovec. Modeling interdependent and periodic real-world action sequences. In *Proceedings of the 2018 world wide web conference*, pages 803–812, 2018.

Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, January 2019.

Bruce E. Landon, Jukka-Pekka Onnela, Laurie Meneades, A. James O'Malley, and Nancy L. Keating. Assessment of Racial Disparities in Primary Care Physician Specialty Referrals. *JAMA Network Open*, 4(1):e2029238, January 2021.

Thomas A. LaVeist, Lydia A. Isaac, and Karen Patricia Williams. Mistrust of Health Care Organizations Is Associated with Underutilization of Health Services. *Health Services Research*, 44(6):2093–2105, December 2009.

Richard J. Lee, Ravi A. Madan, Jayoung Kim, Edwin M. Posadas, and Evan Y. Yu. Disparities in Cancer Care and the Asian American Population. *The Oncologist*, 26(6):453–460, June 2021.

Sabra C. Lewsey and Khadijah Breathett. Racial and ethnic disparities in heart failure: current state and future directions. *Current Opinion in Cardiology*, 36(3):320–328, May 2021.

Bryan Lim and Mihaela van der Schaar. Disease-Atlas: Navigating Disease Trajectories with Deep Learning, July 2018.

Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. Learning to Diagnose with LSTM Recurrent Neural Networks, March 2017.

Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, and James M. Rehg. Efficient Learning of Continuous-Time Hidden Markov Models for Disease Progression. *Advances in Neural Information Processing Systems*, 28:3599–3607, 2015.

Zhi Liu, Uma Bhandaram, and Nikhil Garg. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 4 (1):57–65, 2024.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. 2017.

Samer Madanat, Rabi Mishalani, and Wan Hashim Wan Ibrahim. Estimation of Infrastructure Transition Probabilities from Condition Rating Data. *Journal of Infrastructure Systems*, 1 (2):120–125, June 1995.

Sarah Miller and Laura R. Wherry. Health and Access to Care during the First 2 Years of the ACA Medicaid Expansions. *New England Journal of Medicine*, 376(10):947–956, March 2017.

D R Mould, N G Denman, and S Duffull. Using Disease Progression Models as a Tool to Detect Drug Effect. *Clinical Pharmacology & Therapeutics*, 82 (1):81–86, July 2007.

Sean P. Murphy, Nasrien E. Ibrahim, and James L. Januzzi. Heart Failure With Reduced Ejection Fraction: A Review. *JAMA*, 324(5):488, August 2020.

Sarah Nouri, Courtney R. Lyles, Elizabeth B. Sherwin, Magdalene Kuznia, Anna D. Rubinsky, Kathryn E. Kemper, Oanh K. Nguyen, Urmimala Sarkar, Dean Schillinger, and Elaine C. Khoong. Visit and Between-Visit Interaction Frequency Before and After COVID-19 Telehealth Implementation. *JAMA Network Open*, 6(9):e2333944, September 2023.

Sajida Perveen, Muhammad Shahbaz, Muhammad Sajjad Ansari, Karim Keshavjee, and Aziz Guergachi. A Hybrid Approach for Modeling Type 2 Diabetes Mellitus Progression. *Frontiers in Genetics*, 10:1076, January 2020.

Emma Pierson, Pang Wei Koh, Tatsunori Hashimoto, Daphne Koller, Jure Leskovec, Nicholas Eriksson, and Percy Liang. Inferring Multidimensional Rates of Aging from Cross-Sectional Data, March 2019.

Emma Pierson, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, January 2021.

Teun M. Post, Jan I. Freijer, Joost DeJongh, and Meindert Danhof. Disease System Analysis: Basic Disease Progression Models in Degenerative Disease. *Pharmaceutical Research*, 22(7):1038–1049, July 2005.

Megan Reilly. Health Disparities and Access to Healthcare in Rural vs. Urban Areas. *Theory in Action*, 14(2):6–27, April 2021.

K Romero, K Ito, Ja Rogers, D Polhamus, R Qiu, D Stephenson, R Mohs, R Lalonde, V Sinha, Y Wang, D Brown, M Isaac, S Vamvakas, R Hemmings, L Pani, Lj Bain, B Corrigan, and Alzheimer's Disease Neuroimaging Initiative* for the Coalition Against Major Diseases**. The future is now: Model-based clinical trial design for Alzheimer's disease. *Clinical Pharmacology & Therapeutics*, 97(3):210–214, March 2015.

Rasmus Rørth, Pardeep S. Jhund, Mehmet B. Yilmaz, Søren Lund Kristensen, Paul Welsh, Akshay S. Desai, Lars Køber, Margaret F. Prescott, Jean L. Rouleau, Scott D. Solomon, Karl Swedberg, Michael R. Zile, Milton Packer, and John J.V. McMurray. Comparison of bnp and nt-probnp in patients with heart failure and reduced ejection fraction. *Circulation: Heart Failure*, 13(2): e006541, 2020.

Gráinne Schäfer, Kenneth M. Prkachin, Kimberley A. Kaseweter, and Amanda C. De C Williams. Health

care providers' judgments in chronic pain: the influence of gender and trustworthiness. *Pain*, 157 (8):1618–1625, August 2016.

Divya Shanmugam, Kaihua Hou, and Emma Pierson. Quantifying disparities in intimate partner violence: a machine learning method to correct for underreporting. 2021.

Alexander Shapiro. Identifiability of factor analysis: Some results and open problems. *Linear Algebra and its Applications*, 70:1–7, 1985.

Jonathan Suarez, Jordana B. Cohen, Vishnu Potluri, Wei Yang, David E. Kaplan, Marina Serper, Siddharth P. Shah, and Peter Philip Reese. Racial Disparities in Nephrology Consultation and Disease Progression among Veterans with CKD: An Observational Cohort Study. *Journal of the American Society of Nephrology*, 29(10):2563–2573, October 2018.

R. Sukkar, E. Katz, Yanwei Zhang, D. Raunig, and B. T. Wyman. Disease progression modeling using Hidden Markov Models. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2845–2848, San Diego, CA, August 2012. IEEE.

Arjun K. Venkatesh, Shih-Chuan Chou, Shu-Xia Li, Jennie Choi, Joseph S. Ross, Gail D'Onofrio, Harlan M. Krumholz, and Kumar Dharmarajan. Association Between Insurance Status and Access to Hospital Care in Emergency Department Disposition. *JAMA Internal Medicine*, 179(5):686, May 2019.

Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD 2014, pages 85–94, New York New York USA, August 2014. ACM. ISBN 978-1-4503-2956-9.

Kathryn E. Weaver, Julia H. Rowland, Keith M. Bellizzi, and Noreen M. Aziz. Forgoing medical care because of cost: Assessing disparities in healthcare access among cancer survivors living in the United States. *Cancer*, 116(14):3493–3504, July 2010.

Ruqaiijah Yearby. Racial Disparities in Health Status and Access to Healthcare: The Continuation of Inequality in the United States Due to Structural Racism. *The American Journal of Economics and Sociology*, 77(3-4):1113–1152, May 2018.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. 2020.

Anna Zink, Ziad Obermeyer, and Emma Pierson. Race Corrections in Clinical Algorithms Can Help Correct for Racial Disparities in Data Quality, April 2023.

Appendix A. Proof of Identifiability

A.1. Proof of Theorem 1

Theorem 1 *All model parameters are identified by the observed data distribution* $P(X_t, D_t \mid A)$.

Proof We want to show that each unique set of parameter assignments leads to a different distribution over the observed data. To do this, we divide our argument into four lemmas:

Lemma 6 *Parameters* F, b, Ψ *are identified by* $P(X_t \mid A = a_0)$.

Proof

We want to show that if two parameter sets $\{F, b, \Psi\}$ and $\{\tilde{F}, \tilde{b}, \tilde{\Psi}\}$ yield the same observed data distribution $P(X_0 \mid A = a_0)$, the parameter sets must be identical.

We first note that at $t = 0$, we have $Z_t = Z_0 \sim \mathcal{N}(0, 1)$ for group a_0. Then the mapping between severity and features

$$X_0 = F \cdot Z_0 + b + \epsilon_t$$
$$\epsilon_t \sim \mathcal{N}(0, \Psi)$$

captures a factor analysis model with factor loading matrix F and diagonal covariance matrix Ψ. At $t = 0$, the feature distribution for group a_0 has the standard factor analysis distribution (Shapiro, 1985):

$$X_0 \sim \mathcal{N}(b, FF^T + \Psi).$$

Assuming the two sets of parameters map to distributions of X_0 with the same mean, it must hold that $b = \tilde{b}$. Thus, parameter b is identified by data distribution $P(X_0 \mid A = a_0)$.

Further, the covariance matrix of X_0 induced by each set of parameters must be the same: $F(F)^T + \Psi = \tilde{F}(\tilde{F})^T + \tilde{\Psi}$. Element-wise equality of the covariance matrix gives us the following, where subscripts i refer to the i-th element of each parameter vector:

$$F_i F_j = \tilde{F}_i \tilde{F}_j \quad \forall i, j, i \neq j \tag{1}$$

$$(F_i)^2 + \Psi_i = (\tilde{F}_i)^2 + \tilde{\Psi}_i \tag{2}$$

Using the equality constraint (1) for multiple pairs of indices, we have that for all assignments of distinct indices i, j, k:

$$(F_i F_j = \tilde{F}_i \tilde{F}_j) \wedge (F_j F_k = \tilde{F}_j \tilde{F}_k) \implies \frac{\tilde{F}_i}{F_i} = \frac{\tilde{F}_k}{F_k} \tag{3}$$

$$F_i F_k = \tilde{F}_i \tilde{F}_k \implies \frac{F_i}{\tilde{F}_i} = \frac{\tilde{F}_k}{F_k} \tag{4}$$

Together, equations 3 and 4 give us:

$$\frac{\tilde{F}_i}{F_i} = \frac{F_i}{\tilde{F}_i} \implies (\tilde{F}_i)^2 = (F_i)^2 \implies F_i = \alpha \tilde{F}_i$$

where $\alpha \in \{-1, +1\}$. Since we have fixed $F_0 > 0$ for *all* factor loading matrices F, the sign of α is fixed:

$$F_0 = \alpha \tilde{F}_0 \implies \alpha = 1 \implies F_i = \tilde{F}_i \quad \forall i \in [0, d), \tag{5}$$

meaning we have identified F.

Lastly, using equations (2) and (5) we get $F_i = \tilde{F}_i \implies \Psi_i = \tilde{\Psi}_i$. We have now shown that if two parameter sets induce the same distribution of X at time $t = 0$, they must have the same exact value assignments. Therefore F, b, Ψ are identified by $P(X_t \mid A = a_0)$. ■

Lemma 7 *Global parameters F, b, Ψ and parameters $\mu_{Z_0}^{(a)}, \sigma_{Z_0}^{(a)}, \mu_R^{(a)}, \sigma_R^{(a)}$ for each group a are identified by $P(X_t \mid A)$.*

Proof By Lemma 6, we know that F, b, Ψ are identified by $P(X_0 \mid A = a_0)$. We want to show that for any group a, if two parameter sets $\{\mu_{Z_0}^{(a)}, \sigma_{Z_0}^{(a)}, \mu_R^{(a)}, \sigma_R^{(a)}\}$ and $\{\tilde{\mu}_{Z_0}^{(a)}, \tilde{\sigma}_{Z_0}^{(a)}, \tilde{\mu}_R^{(a)}, \tilde{\sigma}_R^{(a)}\}$ yield the same observed data distribution $P(X_t \mid A = a)$, the parameter sets must be identical. In this proof we consider an arbitrary group a and omit the (a) superscript for brevity.

We model the following:

$$\begin{aligned} Z_0 &\sim \mathcal{N}\left(\mu_{Z_0}, {\sigma_{Z_0}}^2\right) \\ R &\sim \mathcal{N}\left(\mu_R, {\sigma_R}^2\right) \\ Z_t = Z_0 + R \cdot t &\implies Z_t \sim \mathcal{N}\left(\mu_R \cdot t + \mu_{Z_0}, {\sigma_R}^2 \cdot t^2 + {\sigma_{Z_0}}^2\right) \\ X_t &= F \cdot Z_t + b + \epsilon_t, \text{ where } \epsilon_t \sim \mathcal{N}(0, \Psi) \end{aligned} \tag{6}$$

We see that equation (6) captures a factor analysis model with factor loading matrix F and diagonal covariance matrix Ψ, meaning

$$X_t \sim \mathcal{N}(b + F(\mu_R \cdot t + \mu_{Z_0}), F({\sigma_R}^2 \cdot t^2 + {\sigma_{Z_0}}^2)F^T + \Psi).$$

Recalling that $F_0 > 0$, we first consider $t = 0$, where $X_0 \sim \mathcal{N}(b + F\mu_{Z_0}, F({\sigma_{Z_0}}^2)F^T + \Psi)$. In order for the two parameter sets to map to distributions of X_0 with the same mean, it must be the case that

$$b + F\mu_{Z_0} = b + F\tilde{\mu}_{Z_0} \implies \mu_{Z_0} = \tilde{\mu}_{Z_0}.$$

Further, for the two parameter sets to map to distributions with the same covariance matrix, it must hold that

$$F({\sigma_{Z_0}}^2)F^T + \Psi = F({\tilde{\sigma}_{Z_0}}^2)F^T + \Psi \implies \sigma_{Z_0} = \tilde{\sigma}_{Z_0}$$

since we know $\sigma_{Z_0}, \tilde{\sigma}_{Z_0} > 0$. So we have identified μ_{Z_0} and σ_{Z_0}. We next consider any time $t \neq 0$. For the two parameter sets to map to distributions of X_t with the same mean, given that we have already shown μ_{Z_0} must equal $\tilde{\mu}_{Z_0}$, it must hold that

$$b + F(\mu_R \cdot t + \mu_{Z_0}) = b + F(\tilde{\mu}_R \cdot t + \tilde{\mu}_{Z_0}) \implies \mu_R = \tilde{\mu}_R.$$

For the two parameter sets to map to distributions with the same covariance matrix, given that we have already shown σ_{Z_0} must equal $\tilde{\sigma}_{Z_0}$, it must hold that

$$F({\sigma_R}^2 \cdot t^2 + {\sigma_{Z_0}}^2)F^T + \Psi = F({\tilde{\sigma}_R}^2 \cdot t^2 + {\tilde{\sigma}_{Z_0}}^2)F^T + \Psi \implies \sigma_R = \tilde{\sigma}_R$$

since $\sigma_R, \tilde{\sigma}_R > 0$. Thus we have shown that for any group a, group-specific values of $\mu_{Z_0}, \sigma_{Z_0}, \mu_R, \sigma_R$ are identified by $P(X_t \mid A = a)$.

■

Lemma 8 *Global parameters β_0, β_Z and the parameter $\beta_A^{(a)}$ for each group a are identified by $P(D_t \mid A)$.*

Proof We want to show that if two parameter sets $\{\beta_0, \beta_Z, \beta_A^{(a)}\}$ and $\{\tilde{\beta}_0, \tilde{\beta}_Z, \tilde{\beta}_A^{(a)}\}$ yield the same observed data distribution $P(D_t \mid A = a)$, the parameter sets must be identical. Unless otherwise specified, we consider an arbitrary group a and omit the (a) superscript for brevity. We also assume $\mu_R \neq 0$, since in general the severity of a progressive disease should change over time and it does not make sense to learn progression in the case that it does not.

Each event when a patient visits the hospital ($D_t = 1$) is generated by an inhomogeneous Poisson process parameterized by λ_t, where $\log(\lambda_t) = \beta_0 + \beta_Z \cdot Z_t + \beta_A$.

In order for two data distributions to have identical $P(D_t \mid A = a)$ they must have identical expected rates $\mathbb{E}_{Z_0,R}[\lambda_t]$: $\mathbb{E}_{Z_0,R}[\lambda_t]$ is the expected rate of events (across the population) at time t—if two distributions have a different expected rate of events at any time t, then $P(D_t \mid A = a_0)$ must differ at that point in time as well. Thus if two sets of parameters $\{\beta_0, \beta_Z, \beta_A\}$ and $\{\tilde{\beta}_0, \tilde{\beta}_Z, \tilde{\beta}_A\}$ yield the same observed data distribution $P(D_t \mid A = a)$, they must also generate the same observed values $\mathbb{E}_{Z_0,R}[\lambda_t]$ at all timesteps t. We finish the proof by showing that this holds only if $\{\beta_0, \beta_Z, \beta_A\} = \{\tilde{\beta}_0, \tilde{\beta}_Z, \tilde{\beta}_A\}$.

$$\mathbb{E}_{Z_0,R}[\lambda_t] = \int\int \lambda_t \cdot P(Z_0) \cdot P(R)\ dZ_0 dR$$

By Lemma 7, we know that $\mu_{Z_0}, \sigma_{Z_0}, \mu_R, \sigma_R$ are identified by $P(X_t \mid A)$. Then

$$P(Z_0) = \frac{1}{\sqrt{2\pi(\sigma_{Z_0})^2}} \exp\left(-\frac{(Z_0 - \mu_{Z_0})^2}{2(\sigma_{Z_0})^2}\right)$$
$$P(R) = \frac{1}{\sqrt{2\pi(\sigma_R)^2}} \exp\left(-\frac{(R - \mu_R)^2}{2(\sigma_R)^2}\right)$$

$$\mathbb{E}_{Z_0,R}[\lambda_t] = \exp(f(\beta_0, \beta_Z, \beta_A, t)) \tag{7}$$
$$\text{where } f(\beta_0, \beta_Z, \beta_A, t) = \left(\frac{(\beta_Z\sigma_R)^2}{2}\right)t^2 + (\beta_Z\mu_R)t + \left(\beta_0 + \frac{(\beta_Z\sigma_{Z_0})^2}{2} + \beta_Z\mu_{Z_0} + \beta_A\right)$$

The expression in 7 must be equal for $\{\beta_0, \beta_Z, \beta_A\}$ and $\{\tilde{\beta}_0, \tilde{\beta}_Z, \tilde{\beta}_A\}$ at all timesteps t. Since exp is an injective function, this means that $f(\beta_0, \beta_Z, \beta_A, t) = f(\tilde{\beta}_0, \tilde{\beta}_Z, \tilde{\beta}_A, t)$ for all t. By equality of polynomials, each of the individual polynomial coefficients must be equal must be equal for this to hold.

We first consider the case for group a_0, since we pin $\beta_A^{(a_0)}$ at 0 as a reference for all other groups. Given that we have already identified $\mu_{Z_0}^{(a_0)}, \sigma_{Z_0}^{(a_0)}, \mu_R^{(a_0)}, \sigma_R^{(a_0)}$,

$$\left(\beta_0 + \frac{(\beta_Z\sigma_{Z_0})^2}{2} + \beta_Z\mu_{Z_0}\right) = \left(\tilde{\beta}_0 + \frac{(\tilde{\beta}_Z\sigma_{Z_0})^2}{2} + \tilde{\beta}_Z\mu_{Z_0}\right) \implies \beta_0 = \tilde{\beta}_0$$

Now we return to our analysis of any arbitrary group a. Given that we have already identified $\mu_{Z_0}, \sigma_{Z_0}, \mu_R \neq 0, \sigma_R$,

$$\beta_Z\mu_R = \tilde{\beta}_Z\mu_R \implies \beta_Z = \tilde{\beta}_Z$$
$$\left(\beta_0 + \frac{(\beta_Z\sigma_{Z_0})^2}{2} + \beta_Z\mu_{Z_0} + \beta_A\right) = \left(\tilde{\beta}_0 + \frac{(\tilde{\beta}_Z\sigma_{Z_0})^2}{2} + \tilde{\beta}_Z\mu_{Z_0} + \tilde{\beta}_A\right) \implies \beta_A = \tilde{\beta}_A$$

Thus we have shown that β_0, β_Z, and $\beta_A^{(a)}$ for any group a are identified by $P(D_t \mid Z_t, A)$.

■

By showing that each parameter of the model is uniquely recovered from the observed data, we have proved that our model is identifiable.

■

Appendix B. Proofs of Bias

In this section, in order to capture the effect of failing to account for one disparity at a time, we consider the setting where everything between two groups is the same except for disparity of focus. It is clear to see from our analysis that these results hold even more generally—as long as all existing disparities disfavor or favor the same group (e.g. a disadvantaged group with respect to one disparity is not advantaged with respect to another, in which case the effects could cancel each other out), our proofs of bias will hold. Throughout our proofs, we assume that all PDFs and conditional PDFs have positive support over their entire domain, and that all PDFs are differentiable, a very reasonable assumption over our setting.

B.1. Theorem 3

Theorem 3 *A model that does not take into account disparities in initial disease severity Z_0 will underestimate the disease severity of groups with higher initial severity and overestimate that of groups with lower initial severity. Specifically, if $P(Z_0 \mid A = a)$ strictly MLRPs $P(Z_0)$ for some group a, then $\mathbb{E}[Z_t \mid X_t] < \mathbb{E}[Z_t \mid X_t, A = a]$. Similarly, if $P(Z_0)$ strictly MLRPs $P(Z_0 \mid A = a)$ for some group a, then $\mathbb{E}[Z_t \mid X_t] > \mathbb{E}[Z_t \mid X_t, A = a]$.*

Proof We want to show that $\mathbb{E}[Z_t \mid X_t, A = a] > \mathbb{E}[Z_t \mid X_t]$ when $P(Z_0 \mid A = a)$ strictly MLRPs $P(Z_0)$. We first show that $P(Z_0 \mid X_t = x, A = a)$ strictly MLRPs $P(Z_0 \mid X_t)$ with respect to Z_0:

$$\begin{aligned}
\frac{\partial}{\partial Z_0}\left(\frac{P(Z_0 \mid X_t, A = a)}{P(Z_0 \mid X_t)}\right) &= \frac{\partial}{\partial Z_0}\left(\frac{\frac{P(X_t \mid Z_0, A=a)P(Z_0 \mid A=a)}{P(X_t \mid A=a)}}{\frac{P(X_t \mid Z_0)P(Z_0)}{P(X_t)}}\right) && \text{(Bayes Rule)} \\
&= \frac{\partial}{\partial Z_0}\left(\frac{\frac{P(Z_0 \mid A=a)}{P(X_t \mid A=a)}}{\frac{P(Z_0)}{P(X_t)}}\right) && (X_t \perp A \mid Z_0, R) \\
&= \frac{P(X_t)}{P(X_t \mid A = a)} \cdot \frac{\partial}{\partial Z_0}\left(\frac{P(Z_0 \mid A = a)}{P(Z_0)}\right) \\
&> 0 && \text{(Disparity assumption)}
\end{aligned}$$

Since MLRP implies first-order stochastic dominance (FOSD) (Klemens, 2007), this proves that $P(Z_0 \mid X_t, A = a)$ strictly FOSDs $P(Z_0 \mid X_t)$ and thus that $\mathbb{E}[Z_0 \mid X_t, A = a] > \mathbb{E}[Z_0 \mid X_t]$. By linearity of expectation,

$$\begin{aligned}
&\mathbb{E}[Z_0 \mid X_t, A = a] + \mathbb{E}[f(R,t) \mid X_t, A = a] > \mathbb{E}[Z_0 \mid X_t] + \mathbb{E}[f(R,t) \mid X_t], \quad \forall t \geq 0 \\
&\implies \mathbb{E}[Z_t \mid X_t, A = a] > \mathbb{E}[Z_t \mid X_t]
\end{aligned}$$

It is clear to see that this argument extends naturally to show that if a group tends to come in at *earlier* disease stages than the rest of the population, that their severity will be overestimated: If there exists a group $\tilde{a}$ such that $P(Z_0)$ strictly MLRPs $P(Z_0 \mid A = \tilde{a})$ with respect to Z_0 and $\mathbb{E}[R \mid X_t] \geq \mathbb{E}[R \mid X_t, A = \tilde{a}]$, then we will see that $\mathbb{E}[Z_t \mid X_t, A = \tilde{a}] < \mathbb{E}[Z_t \mid X_t]$. Hence any model that does not take into account

demographic disparities in initial disease severity levels at a patient's first visit will lead to biased estimates of severity. ■

B.2. Proof of Theorem 4

Theorem 4 *Suppose disease severity progresses linearly at some rate R. A model that does not take into account disparities in R will underestimate the disease severity of groups with higher progression rates and overestimate that of groups with lower progression rates. Specifically, if $P(R \mid A = a)$ strictly MLRPs $P(R)$ for some group a, then $\mathbb{E}[Z_t \mid X_t] < \mathbb{E}[Z_t \mid X_t, A = a]$. Similarly, if $P(R)$ strictly MLRPs $P(R \mid A = a)$ for some group a, then $\mathbb{E}[Z_t \mid X_t] > \mathbb{E}[Z_t \mid X_t, A = a]$.*

R is a patient's linear rate of progression, so we model a patient's severity over time as $Z_t = f(R, t) + Z_0$, where f is linearly increasing in R.

Proof We want to show that $\mathbb{E}[Z_t \mid X_t, A = a] > \mathbb{E}[Z_t \mid X_t]$ when $P(R \mid A = a)$ strictly MLRPs $P(R)$. We first show that $P(R \mid X_t, A = a)$ strictly MLRPs $P(R \mid X_t)$ with respect to R:

$$\begin{aligned}
\frac{\partial}{\partial R}\left(\frac{P(R \mid X_t, A = a)}{P(R \mid X_t)}\right) &= \frac{\partial}{\partial R}\left(\frac{\frac{P(X_t \mid R, A=a)P(R \mid A=a)}{P(X_t \mid A=a)}}{\frac{P(X_t \mid R)P(Z_t = z_t)}{P(X_t)}}\right) && \text{(Bayes Rule)} \\
&= \frac{\partial}{\partial R}\left(\frac{\frac{P(R \mid A=a)}{P(X_t \mid A=a)}}{\frac{P(R)}{P(X_t)}}\right) && (X \perp A \mid Z_0, R) \\
&= \frac{P(X_t)}{P(X_t \mid A = a)} \cdot \frac{\partial}{\partial R}\left(\frac{P(R \mid A = a)}{P(R)}\right) \\
&> 0 && \text{(Disparity assumption)}
\end{aligned}$$

Since MLRP implies FOSD (Klemens, 2007), this also implies that $P(R \mid X_t, A = a)$ strictly FOSDs $P(R \mid X_t)$. It follows directly that $\mathbb{E}[R \mid X_t, A = a] > \mathbb{E}[R \mid X_t]$. By linearity of expectation,

$$\begin{aligned}
&\mathbb{E}[f(R, t) + Z_0 \mid X_t, A = a] > \mathbb{E}[f(R, t) + Z_0 \mid X_t], \quad \forall t > 0 \\
&\implies \mathbb{E}[Z_t \mid X_t, A = a] > \mathbb{E}[Z_t \mid X_t]
\end{aligned}$$

It is clear to see that this argument extends naturally to show that if a group tends to progress *more slowly* than the rest of the population, that their severity will be overestimated: if there exists a group $\tilde{a}$ such that $P(R)$ strictly MLRPs $P(R \mid A = \tilde{a})$ with respect to R and $\mathbb{E}[Z_0 \mid X_t] \geq \mathbb{E}[Z_0 \mid X_t, A = \tilde{a}]$, then we will see that $\mathbb{E}[Z_t \mid X_t, A = \tilde{a}] < \mathbb{E}[Z_t \mid X_t]$. Thus any model that does not take into account demographic disparities in patient progression rates will lead to biased estimates of severity. ■

B.3. Proof of Theorem 5

Theorem 5 *A model that does not take into account disparities in visit frequency conditional on disease severity will underestimate the disease severity of groups with lower visit frequency conditional on severity and overestimate the disease severity of groups with higher visit frequency conditional on severity. Specifically, assume that $P(E_t = 1 \mid Z_t)$ is strictly monotone increasing in Z_t, $\lim_{Z_t \to -\infty} P(E_t = 1 \mid Z_t) = 0$, and $\lim_{Z_t \to \infty} P(E_t = 1 \mid Z_t) = 1$. Then if some group a has a lower probability of visiting a healthcare provider at any given severity level—that is, $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z - \alpha(z))$ for all z, where $\alpha(z)$ is a positive function of z—then $\mathbb{E}[Z_t \mid E_t] < \mathbb{E}[Z_t \mid E_t, A = a]$. Similarly, if $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z + \alpha(z))$ for all z, where $\alpha(z)$ is a positive function of z, then*

$\mathbb{E}[Z_t \mid E_t] > \mathbb{E}[Z_t \mid E_t, A = a]$.

Proof Recall that for a given patient, E_t corresponds to the event where some visit occurs during the timestep starting at time t, and $P(E_t = 1 \mid Z_t = z)$ indicates the probability that the patient visits during the time period if their severity is z at the beginning of this time period. We will first show that, when there is some positive $\alpha(z)$ for all z such that $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z - \alpha(z))$, it holds that $\mathbb{E}[Z_t \mid E_t = 1, A = a] > \mathbb{E}[Z_t \mid E_t = 1]$. We will then show that this argument holds when conditioning on $E_t = 0$ as well—i.e., $\mathbb{E}[Z_t \mid E_t = 0, A = a] > \mathbb{E}[Z_t \mid E_t = 0]$.

We first compute $\mathbb{E}[Z_t \mid E_t = 1]$. Define $p(z) \coloneqq P(Z_t = z)$ and $F(z) \coloneqq P(E_t = 1 \mid Z_t = z)$. By Bayes' rule we have:

$$P(Z_t = z \mid E_t = 1) = \frac{p(z)F(z)}{\int_{-\infty}^{\infty} p(z)F(z)\ dz}$$

By assumption, $F(z)$ is strictly monotone increasing in z, $\lim_{z \to -\infty} F(z) = 0$, and $\lim_{z \to \infty} F(z) = 1$. These are the properties of a CDF, so we can write $F(z)$ in terms of its corresponding PDF $f(z)$: i.e., $F(z) = \int_{-\infty}^{z} f(x)dx$. This yields:

$$P(Z_t = z \mid E_t = 1) = \frac{p(z) \int_{-\infty}^{z} f(x)\ dx}{\int_{-\infty}^{\infty} p(z) \int_{-\infty}^{z} f(x)\ dx\ dz}$$

Then we can write the expectation $\mathbb{E}[Z_t \mid E_t = 1]$ as:

$$\begin{aligned}\mathbb{E}[Z_t \mid E_t = 1] &= \int_{-\infty}^{\infty} P(Z_t = z \mid E_t = 1) z\ dz \\ &= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{z} p(z)f(x)z\ dx\ dz}{\int_{-\infty}^{\infty} \int_{-\infty}^{z} p(z)f(x)\ dx\ dz}\end{aligned}$$

Graphically, this corresponds to taking the expectation of z when points are sampled from the blue region in Figure S1, where the probability of sampling each point is proportional to $p(z)f(x)$.

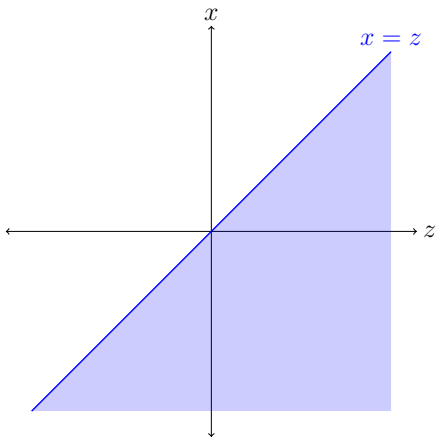

Figure S1: We can calculate $\mathbb{E}[Z_t \mid E_t = 1]$ by taking the expectation over the blue region, with each point having probability $p(z)f(x)$.

We next consider $\mathbb{E}[Z_t \mid E_t = 1, A = a]$, which yields an analogous expression. Define $p_a(z) \coloneqq P(Z_t = z \mid A = a)$ and $F_a(z) = P(E_t = 1 \mid Z_t = z, A = a)$. Since all groups have the same severity distribution, we see

that $p_a(z) = p(z)$. Further, $F_a(z) = P(E_t = 1 \mid Z_t = z - \alpha(z)) = F(z - \alpha(z))$ by our disparity assumption. By Bayes' rule we have:

$$\begin{aligned} P(Z_t \mid E_t = 1, A = a) &= \frac{p_a(z)F_a(z)}{\int_{-\infty}^{\infty} p_a(z)F_a(z)\ dz} \\ &= \frac{p(z)F(z - \alpha(z))}{\int_{-\infty}^{\infty} p(z)F(z - \alpha(z))\ dz} \end{aligned}$$

As before, we write $F(z)$ in terms of its corresponding PDF $f(z)$, yielding:

$$P(Z_t \mid E_t = 1, A = a) = \frac{p(z)\int_{-\infty}^{z-\alpha(z)} f(x)\ dx}{\int_{-\infty}^{\infty} p(z)\int_{-\infty}^{z-\alpha(z)} f(x)\ dx\ dz}$$

Finally, we can write the expectation $\mathbb{E}[Z_t \mid E_t = 1, A = a]$ as:

$$\begin{aligned} \mathbb{E}[Z_t \mid E_t = 1, A = a] &= \int_{-\infty}^{\infty} P(Z_t \mid E_t = 1, A = a) z\ dz \\ &= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{z-\alpha(z)} p(z)f(x)z\ dx\ dz}{\int_{-\infty}^{\infty}\int_{-\infty}^{z-\alpha(z)} p(z)f(x)\ dx\ dz} \end{aligned}$$

Now the region of z and x that we integrate over corresponds to the red region in Figure S2; the crosshatched blue region corresponds to the region that we integrate over in our calculation of $\mathbb{E}[Z_t \mid E_t = 1]$ but *not* in our calculation of $\mathbb{E}[Z_t \mid E_t = 1, A = a]$ (such that the solid blue region in Figure S1 is the combination of the blue crosshatched and red regions in Figure S2). Note that visualization of the red region assumes constant $\alpha(z)$, but the logical argument applies to any function $\alpha(z) > 0$.

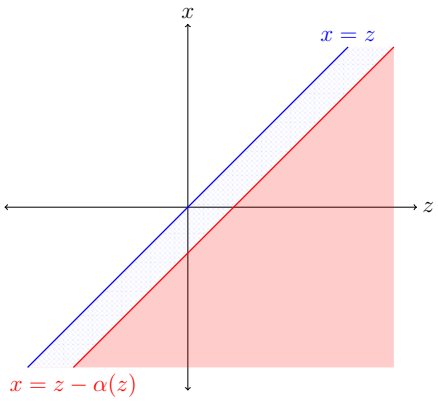

Figure S2: We can calculate $\mathbb{E}[Z_t \mid E_t = 1, A = a]$ by taking the expectation over the red region, with each point having probability $p(z)f(x)$. The crosshatched blue region provides a comparison to the integration space for $\mathbb{E}[Z_t \mid E_t = 1]$.

We see that at each value of x, the blue crosshatched region adds strictly positive weight to lower values of z; thus, the expectation of z over the blue crosshatched region *plus* red region must be lower than the expectation of z over the red region itself. We therefore conclude that $\mathbb{E}[Z_t \mid E_t = 1] < \mathbb{E}[Z_t \mid E_t = 1, A = a]$.

The reasoning when conditioning on $E_t = 0$ is analogous. Since $P(E_t = 0 \mid Z_t = z) = 1 - F(z)$, we get the following expressions:

$$\mathbb{E}[Z_t \mid E_t = 0] = \frac{\int_{-\infty}^{\infty} \int_{z}^{\infty} p(z) f(x) z \; dx \; dz}{\int_{-\infty}^{\infty} \int_{z}^{\infty} p(z) f(x) \; dx \; dz}$$

$$\mathbb{E}[Z_t \mid E_t = 0, A = a] = \frac{\int_{-\infty}^{\infty} \int_{z-\alpha(z)}^{\infty} p(z) f(x) z \; dx \; dz}{\int_{-\infty}^{\infty} \int_{z-\alpha(z)}^{\infty} p(z) f(x) \; dx \; dz}$$

Graphically, $\mathbb{E}[Z_t \mid E_t = 0]$ corresponds to taking the expectation of z when points are sampled from the blue region in Figure S3, where the probability of sampling each point is proportional to $p(z)f(x)$. $\mathbb{E}[Z_t \mid E_t = 0, A = a]$ corresponds to taking the expectation over the blue plus crosshatched red regions.

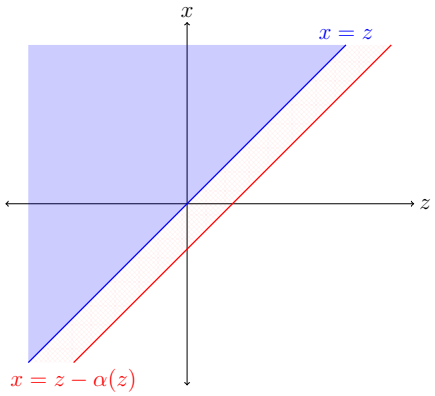

Figure S3: We can calculate $\mathbb{E}[Z_t \mid E_t = 0]$ by taking the expectation over the blue region and $\mathbb{E}[Z_t \mid E_t = 0, A = a]$ by taking the expectation over the blue plus crosshatched red regions, with each point having probability $p(z)f(x)$.

We thus see that at each value of x, the red crosshatched region adds strictly positive weight to higher values of z; thus, the expectation of z over the red crosshatched region *plus* blue region must be higher than the expectation of z over the blue region itself. We therefore conclude that $\mathbb{E}[Z_t \mid E_t = 0, A = a] > \mathbb{E}[Z_t \mid E_t = 0]$.

It is clear to see that this argument extends naturally to show that if a group tends to come in *more frequently* than the rest of the population, their severity will be overestimated: if for some group $\tilde{a}$ there is some positive $\alpha(z)$ for all z such that $P(E_t = 1 \mid Z_t = z, A = \tilde{a}) = P(E_t = 1 \mid Z_t = z + \alpha(z))$, it will hold that $\mathbb{E}[Z_t \mid E_t, A = \tilde{a}] < \mathbb{E}[Z_t \mid E_t]$. Hence any model that does not take into account demographic disparities in visit frequency will lead to biased estimates of severity.

■

We finally show that our model's specific parameterization of visit rate satisfies the properties we assume and will, therefore, exhibit the bias we characterize in this Theorem. As described in §3, we model a patient's visit rate using an inhomogeneous Poisson process characterized by visit rate $\lambda_t = \exp(\beta_0 + \beta_Z \cdot Z_t + \beta_A^{(a)})$; for simplicity, we consider the two-group case and pin $\beta_A^{(\tilde{a})} = 0$ for one group; without loss of generality, we assume that $\beta_A^{(a)} < 0$ (group a has a lower visit rate at the same severity). We show that when $\beta_Z > 0$, this parameterization satisfies each of the more general assumptions in Theorem 5—namely, $P(E_t = 1 \mid Z_t)$ is strictly monotone increasing in Z_t; $\lim_{Z_t \to -\infty} P(E_t = 1 \mid Z_t) = 0$; $\lim_{Z_t \to \infty} P(E_t = 1 \mid Z_t) = 1$; and $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z - \alpha(z))$—and the theorem's results thus apply to our specific parameterization.

$P(E_t = 1 \mid Z_t)$ is strictly monotone increasing in Z_t because the rate $\lambda_t = \exp(\beta_0 + \beta_Z \cdot Z_t + \beta_A^{(a)})$ is monotone increasing for $\beta_Z > 0$. Further, $\lim_{Z_t \to -\infty} \lambda(Z_t) = 0$, meaning the expected number of visits in any discrete time period limits to 0 and the probability of an event in any $[t_i, t_{i+1})$ limits to 0; similarly, $\lim_{Z_t \to \infty} \lambda(Z_t) = \infty$, meaning that the probability of an event in $[t_i, t_{i+1})$ limits to 1.

Finally, we want to show that $P(E_t = 1 \mid Z_t = z, A = a) = P(E_t = 1 \mid Z_t = z - \alpha(z))$ for all z, where $\alpha(z)$ is a positive function of z. For any z, let $F_a(z) = P(E_t = 1 \mid Z_t = z, A = a)$; $0 \leq F_a(z) \leq 1$. Because $P(E_t = 1 \mid Z_t = z)$ is continuous in z and increases from 0 to 1, by the intermediate value theorem there must exist some value $z - \alpha(z)$ for any z such that $P(E_t = 1 \mid Z_t = z - \alpha(z)) = F_a(z) = P(E_t = 1 \mid Z_t = z, A = a)$. Because $\beta_A^{(a)} < 0$, $P(E_t = 1 \mid Z_t = z, A = a) < P(E_t = 1 \mid Z_t = z)$. So $\alpha(z)$ must be positive, as desired.

Appendix C. Simulations

Our simulations show that, on synthetic data, our model accurately recovers true data-generating parameters, learns severity estimates that are well-calibrated with ground truth, and produces less biased estimates of severity than models that do not account for disparities. We describe the simulations in detail below, and all associated code can be found at https://github.com/erica-chiang/progression-disparities.

Data generation. We generate synthetic datasets by drawing parameter values for each dataset from the prior distributions assumed by our model. We generate simulated data for 1000 patients in each dataset, each of whom is assigned to one of two demographic groups (50% chance of being from either group). Our model priors are as follows (where the normal distribution is represented as $\mathcal{N}(\mu, \sigma)$, and $\mathcal{TN}(\mu, \sigma, a)$ indicates a normal distribution with a lower bound of a).

As described in the main text, we pin values $\mu_{Z_0} = 0$, $\sigma_{Z_0} = 1$, and $\beta_A = 0$ for one group, for identifiability. Then for the non-pinned group:

$$\begin{aligned} \mu_{Z_0} &\sim \mathcal{N}(0, 4) \\ \sigma_{Z_0} &\sim \mathcal{TN}(1, 0.1, 0) \\ \beta_A &\sim \mathcal{N}(0, 2) \end{aligned}$$

The remaining group-independent priors are:

$$\begin{aligned} \mu_R &\sim \mathcal{N}(1, 4) \\ \sigma_R &\sim \mathcal{TN}(0.1, 0.4, 0) \\ F_0 &\sim \mathcal{TN}(1, 1, 0.5), \text{ to enforce positive constraint} \\ F_i &\sim \mathcal{N}(0, 2), \text{ for } i > 0 \\ b &\sim \mathcal{N}(0, 1) \\ \psi &\sim \mathcal{TN}(5, 1, 0) \\ \beta_0 &\sim \mathcal{N}(1.5, 0.1) \\ \beta_Z &\sim \mathcal{TN}(0.5, 0.1, 0.1) \end{aligned}$$

Parameter recovery. We fit our full model on 100 synthetic datasets and compare the true data-generating values and recovered values of each parameter in our model. In Figure S4, we visualize the recovery of each parameter by plotting true parameter values versus recovered posterior mean values, with one dot per run.

Severity recovery. We also compare the latent severity values of each patient at each timepoint to the recovered posterior mean values of severity for each patient. We examine the correlation between true and recovered values across both groups.

Appendix D. NewYork-Presbyterian (NYP) Heart Failure Data Processing

This study was conducted in accordance with Health Insurance Portability and Accountability Act (HIPAA) guidelines and with Institutional Review Board (IRB) approval.

Cohort filtering. We analyze patients with *heart failure with reduced ejection fraction* (HFrEF) whom we identify, following clinical guidance, by filtering the available NYP data for patients who have at least one LVEF measurement below 50% and who have been recorded as receiving a diuretic prescription. To ensure we have relatively complete records for each patient, we then filter for patients who are likely to receive most of their cardiology care within the NYP system, by filtering for patients whose home zipcode is in the New York metropolitan area and who have at least two LVEF or BNP records at least 6 months apart within our data. Lastly, NYP switched electronic health record (EHR) systems, introducing inconsistencies in record-keeping across sites and years; to ensure our records are consistently recorded, we analyze data from Weill Cornell Medical Center, one of NYP's two largest sites, between January 1, 2012 (the start of reliable record-keeping) to October 2, 2020 (NYP Cornell's transition to a new EHR). This ensures records are consistently recorded in our data.

Feature processing. We convert pBNP to BNP with the conversion pBNP = 6.25 * BNP (Rørth et al., 2020) and then log-transform BNP values to get one combined $\log_2$(BNP) feature (Hendricks et al., 2022). We then normalize (z-score) all feature values so that each feature has mean 0 and variance 1. Because patient blood pressure and heart rate are much more likely to be measured at hospital visits unrelated to heart failure (while visiting another specialist in the NYP medical system), we limit patient observations to visits where a patient had one measurement of at least one of LVEF and BNP.

We encode demographic categories by making A a one-hot encoding of race/ethnicity groups. Lastly, we describe the time scale of our model. As mentioned in §6, we discretize time in 1-week bins; if a patient has multiple measurements of one feature within a timestep, we average all measurements within that timestep. Discretizing time in this way allows us to capture more long-term changes rather than acute changes in patient status. We normalize time so that the total time range in our model is 0 to 1. The longest patient trajectory in our data is 446 weeks (timesteps), so we normalize timestep values so that they range from 0 to 1; we therefore have fractional, discrete time values, each representing one week as 1/446 units of time.

Appendix E. Model Evaluation

Fitting model on real data. We fit our model on real data using weakly informative priors: $\mu_{Z_0} \sim \mathcal{N}(0,1)$, $\sigma_{Z_0} \sim \mathcal{TN}(1,1,0)$, and $\beta_A \sim \mathcal{N}(0,1)$ for the non-pinned groups; $\mu_R \sim \mathcal{N}(0,1)$ and $\sigma_R \sim \mathcal{TN}(1.5,1,0)$ for all groups; $b \sim \mathcal{N}(0,1)$; $\psi \sim \mathcal{TN}(1,0.5,0)$; $\beta_0 \sim \mathcal{N}(2.5,1)$; $\beta_Z \sim \mathcal{N}(0,1)$. For F, b, and Ψ, we set model priors using Factor Analysis: at $t = 0$, we have $Z_t = Z_0 \sim \mathcal{N}(0,1)$ for group a_0, meaning the mapping between severity and features

$$X_0 = F \cdot Z_0 + b + \epsilon_t$$

$$\epsilon_t \sim \mathcal{N}(0, \Psi)$$

captures a factor analysis model with factor loading matrix F and diagonal covariance matrix Ψ. We run factor analysis using feature measurements from the *first timestep* of all White patients (our a_0 group) and use the estimates of F from Factor Analysis as the mean of our priors on F. We define the variance of our priors on F to be 1, and we pin the sign of F_{LVEF} to be negative for identifiability. Since we have no inherent value scale for what F values should be, Factor Analysis allows us to fit the model on more substantiated priors for feature scaling factors.

We then fit the model and get the parameter estimates from 1000 samples. For any time t, we can calculate an estimate of Z_t and X_t for each sample, based on the sample's parameter estimates; we then take the average over all samples to get a patient's estimate of Z_t and X_t. In order to get actual feature value estimates, we can linearly transform X_t to undo the normalization for each feature and recover an estimate

of each feature value at t. We can then use our model's estimates of Z_t and predicted feature values to analyze and evaluate our model's behavior.

Comparison to baselines. We filter out patients who do not have at least three visits (since several of the baselines we fit require this many visits per patient, as we describe below), leaving a total of 1834 patients: 1118 White, 347 Black, 216 Hispanic, and 153 Asian patients.

To evaluate our model's ability to reconstruct feature values, we compare our model to PCA and FA. PCA and FA require consistent dimensionality of the input data, so we fit all models on the first three visits for each patient. We train two variants of both PCA and FA: the first attempts to reconstruct patient *visits* from a single latent dimension (analogous to Z in our model), taking as input the X_t vector at one visit (4 features total) and representing it with a single latent component. The second variant attempts to reconstruct *patient trajectories* from two latent dimensions (analogous to Z_0 and R in our model), taking as input a concatenated vector of features X_t from the first three visits (12 features total) and representing it with two latent components. We impute missing values as the overall mean of the data for both PCA and FA, since these methods cannot naturally handle missing data.

To evaluate our model's ability to predict future feature values, we compare our model to last time-step, logistic regression, and quadratic regression. Unlike PCA and FA, these methods do not require consistent dimensionality in the input data, so we fit the models to the first three years of observed data. Last-timestep predicts all future feature values to be equal to the most recent feature value observed in the training data for that patient; if there is no observed feature value, the baseline predicts the population mean. Linear regression regresses values on time for each patient and each feature to predict future feature values. For patients with fewer than 2 observations for a given feature value, we use the population mean for the preceding or subsequent timestep. Quadratic regression follows a similar approach. Because linear regression and quadratic regression can overfit the data and make unrealistic predictions, we clip their predicted feature values to a range determined by that observed within the training data.

Ablated Model. We compare our full model to an ablated version of the model that does not account for any of our three disparities. We do this by removing all group-specific parameters from the model, while leaving everything else the same: we learn one value of μ_R and σ_R and exclude β_A from the model. Since the distribution of Z_0 must be fixed for at least one group for identifiability (to fix the scale of Z_t), the distribution is pinned for all groups. Factor Analysis for model priors on F is also fit on all patients rather than only on white patients.

Appendix F. Disparities Estimates

We first describe our calculations for §6.3 to estimate how much later Black and Asian patients start receiving care for heart failure compared to White patients. Our model learns the following:

$$\mu_{Z_0}^{(\text{Black})} = \mu_{Z_0}^{(\text{White})} + 0.22$$
$$\mu_{Z_0}^{(\text{Asian})} = \mu_{Z_0}^{(\text{White})} + 0.27$$

The learned average rate of progression across all patients is 0.62. This means that Black patients come in $0.22/0.62 = 0.35$ units of time later in their disease progression than White patients, and Asian patients come in $0.27/0.62 = 0.44$ units of time later than White patients. Given that one unit of time is the longest patient trajectory, 8.5 years, this leads us to 3.0 and 3.8 years for Black and Asian patients, respectively.

Next we describe our calculations to estimate how much less frequently Black patients visit the hospital than White patients at the same disease severity. Our model learns that

$$\beta_A^{(\text{Black})} = \beta_A^{(\text{White})} - 0.11$$

At the same disease severity Z_t, Black patients will have a visit rate of

$$\begin{aligned}\lambda_t &= \exp(\beta_0 + \beta_Z \cdot Z_t + (\beta_A^{(\text{White})} - 0.11)) \\ &= \exp(\beta_0 + \beta_Z \cdot Z_t + \beta_A^{(\text{White})}) \cdot \exp(-0.11) \\ &= 0.897 \cdot \exp(\beta_0 + \beta_Z \cdot Z_t + \beta_A^{(\text{White})})\end{aligned}$$

So at the same disease severity, we estimate that Black patients have a visit rate that is 90% that of a White patient's visit rate.

Appendix G. Supplementary Figures and Tables

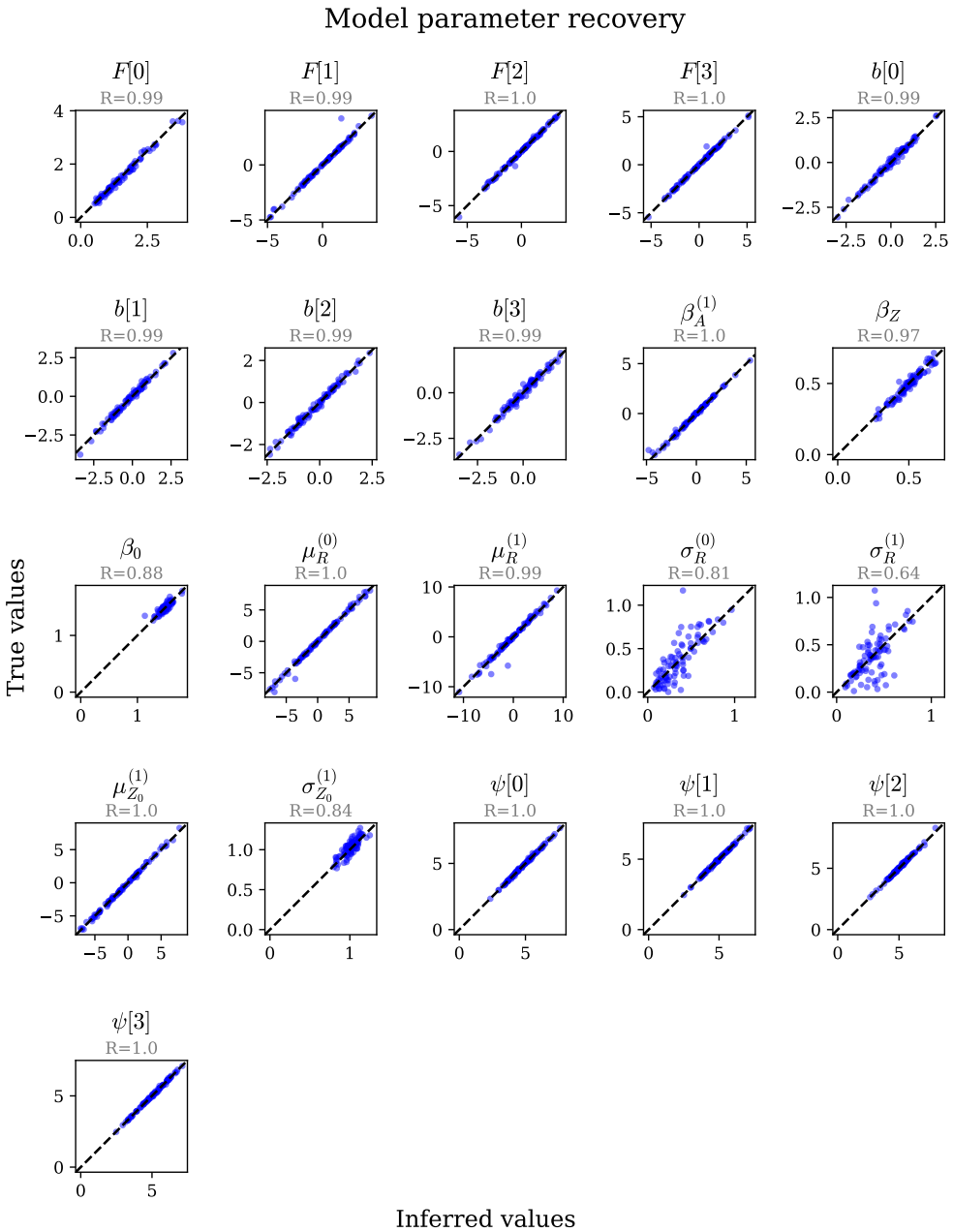

Figure S4: **Parameter recovery from fitting our model to synthetic data.**

	Our model	**FA$_{\text{visit}}$**	**PCA$_{\text{visit}}$**	**FA$_{\text{patient}}$**	**PCA$_{\text{patient}}$**
MAPE: informative	20%	28%	23%	25%	21%
MAPE: all	16%	19%	17%	18%	16%

Table S1: **Our model compared to standard baselines for reconstruction performance.** We compare to factor analysis and principal component analysis fit at the patient visit level (FA_{visit}, $\text{PCA}_{\text{visit}}$) and at the trajectory level ($\text{FA}_{\text{patient}}$, $\text{PCA}_{\text{patient}}$). Models are fit on the first 3 visits from each patient and evaluated on same data using mean absolute percentage error (MAPE). We report aggregate performance for features that are more informative of heart failure progression (LVEF and BNP), along with performance for all features (LVEF, BNP, systolic blood pressure, heart rate).

	Our model	**Linear regression**	**Quadratic regression**	**Latest timestep**
MAPE: informative	28%	39%	59%	22%
MAPE: all	21%	32%	49%	18%

Table S2: **Our model compared to standard baselines for predictive performance.** We compare to linear regression, quadratic regression, and latest timestep prediction, each fit at the patient feature level. Models are fit on data from the first 3 years of each patient's disease trajectory and evaluated on visits after 3 years using mean absolute percentage error (MAPE). We report performance for features that are more informative of heart failure progression (LVEF and BNP), along with performance for all features (LVEF, BNP, systolic blood pressure, heart rate).

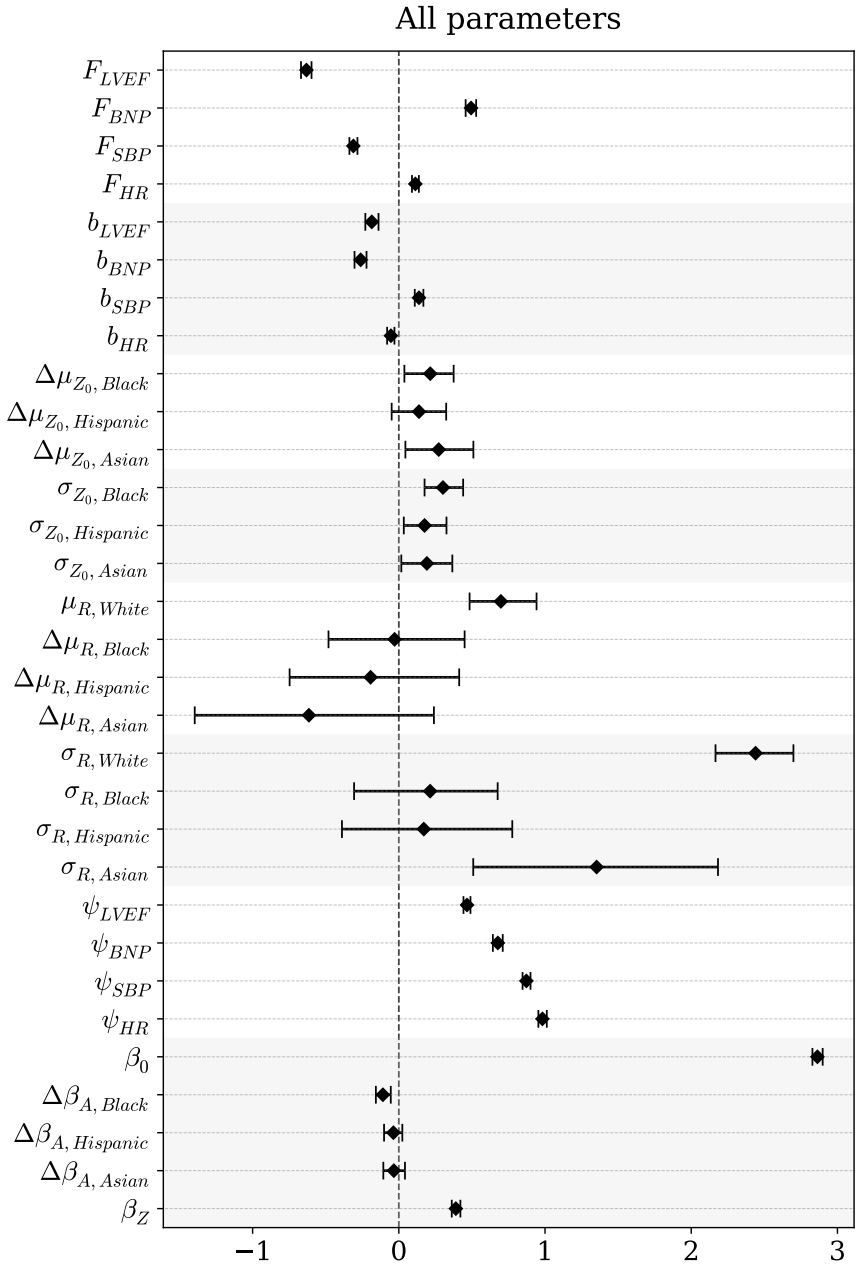

Figure S5: **All parameters learned from fitting model on heart failure cohort.** Parameters of primary interest for interpreting our model (a subset of the parameters shown here) are highlighted in Figure 3.