A Proof of Identifiability

A.1 Proof of Theorem 4.1

Theorem 4.1. *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 A.1. *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:
>
> $$F_i F_j = \tilde{F}_i \tilde{F}_j \;\; \forall i, j, i \neq j \tag{1}$$
> $$(F_i)^2 + \Psi_i = (\tilde{F}_i)^2 + \tilde{\Psi}_i \tag{2}$$
>
> 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:
>
> $$F_0 = \alpha \tilde{F}_0 \implies \alpha = 1 \implies F_i = \tilde{F}_i \;\; \forall i \in [0, d), \tag{5}$$
>
> 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 A.2. *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 A.1, 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:

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

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

$$Z_t = Z_0 + R \cdot t \implies Z_t \sim \mathcal{N}\left(\mu_R \cdot t + \mu_{Z_0}, {\sigma_R}^2 \cdot t^2 + {\sigma_{Z_0}}^2\right)$$

$$X_t = F \cdot Z_t + b + \epsilon_t, \text{ where } \epsilon_t \sim \mathcal{N}(0, \Psi) \tag{6}$$

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 A.3. *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 A.2, we know that $\mu_{Z_0}, \sigma_{Z_0}, \mu_R, \sigma_R$ are identified by $P(X_t \mid A)$. Then

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

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

$$\mathbb{E}_{Z_0,R}[\lambda_t] = \exp(f(\beta_0, \beta_Z, \beta_A, t)) \tag{7}$$

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.

□

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 4.3

Theorem 4.3. *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 $\mathbb{E}[Z_t \mid X_t, A = a] > \mathbb{E}[Z_t \mid X_t]$. 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 4.4

Theorem 4.4. *A model that does not take into account disparities in rate of progression 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 $\mathbb{E}[Z_t \mid X_t, A=a] > \mathbb{E}[Z_t \mid X_t]$. We first show that $P(R \mid X_t, A=a)$ strictly MLRPs $P(R \mid X_t)$ with respect to R:

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

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 4.5

Theorem 4.5. *A model that does not take into account disparities in visit frequency λ_t (conditional on disease severity) will underestimate the disease severity of groups with lower visit frequency and overestimate that of groups with higher visit frequency. Specifically, if it holds for some group a that $\beta_A^{(a)} < \beta_A^{(\tilde{a})}$ for all $\tilde{a} \neq a$, then $\mathbb{E}[Z_t \mid D_t] < \mathbb{E}[Z_t \mid D_t, A=a]$. Similarly, if it holds for some group a that $\beta_A^{(a)} > \beta_A^{(\tilde{a})}$ for all $\tilde{a} \neq a$, then $\mathbb{E}[Z_t \mid D_t] > \mathbb{E}[Z_t \mid D_t, A=a]$.*

We model a patient's visit pattern using an inhomogeneous poisson process characterized by visit rate λ_t, such that $\log(\lambda_t) = g(Z_t) + \beta_A^{(A)}$ for some function of severity $g(Z_t)$ and group-specific adjustments $\beta_A^{(A)}$. In our proof, we assume the large-sample limit in which λ_t can be perfectly estimated from the observed data, and thus treat it as observed; we show empirically that our results hold in finite samples as well. We assume $g(Z_t)$ is a strictly monotonically increasing function of severity.

Proof. We want to show that $\mathbb{E}[Z_t \mid D_t, A=a] > \mathbb{E}[Z_t \mid D_t]$. We do this by calculating each term separately.

We first consider $\mathbb{E}[Z_t \mid D_t, A=a]$. Observing D_t over time gives us an observed value of visit rate λ_t. The strictly monotone assumption of g ensures g is invertible, and the fact that all visit rates λ_t are characterized by $\log(\lambda_t) = g(Z_t) + \beta_A^{(A)}$ ensures that this holds over the entire range of λ_t values. This gives us:

$$\begin{aligned}
\mathbb{E}[Z_t \mid D_t, A=a] &= \mathbb{E}\left[g^{-1}\left(\log(\lambda_t) - \beta_A^{(A)}\right) \;\middle|\; D_t, A=a\right] \\
&= g^{-1}\left(\log(\lambda_t) - \beta_A^{(a)}\right)
\end{aligned}$$

We next consider the case where a model infers severity without taking into account disparities in visit rate conditional on severity. Estimating severity Z_t based solely on visit observations gives:

$$\begin{aligned}
\mathbb{E}[Z_t \mid D_t] &= P(A=a)\cdot\mathbb{E}[Z_t \mid D_t, A=a] + P(A\neq a)\cdot\mathbb{E}[Z_t \mid D_t, A\neq a] \\
&= P(A=a)\cdot\mathbb{E}\left[g^{-1}\left(\log(\lambda_t)-\beta_A^{(A)}\right) \;\middle|\; D_t, A=a\right] \\
&\qquad + P(A\neq a)\cdot\mathbb{E}\left[g^{-1}\left(\log(\lambda_t)-\beta_A^{(A)}\right) \;\middle|\; D_t, A\neq a\right] \\
&< P(A=a)\cdot\mathbb{E}\left[g^{-1}\left(\log(\lambda_t)-\beta_A^{(A)}\right) \;\middle|\; D_t, A=a\right] \\
&\qquad + P(A\neq a)\cdot\mathbb{E}\left[g^{-1}\left(\log(\lambda_t)-\beta_A^{(a)}\right) \;\middle|\; D_t, A=a\right] \qquad (*) \\
&= P(A=a)\cdot\left(g^{-1}\left(\log(\lambda_t)-\beta_A^{(a)}\right)\right) + P(A\neq a)\cdot\left(g^{-1}\left(\log(\lambda_t)-\beta_A^{(a)}\right)\right) \\
&= g^{-1}\left(\log(\lambda_t)-\beta_A^{(a)}\right) \\
&= \mathbb{E}[Z_t \mid D_t, A=a]
\end{aligned}$$

As justification for $(*)$:

$$\begin{aligned}
&\beta_A^{(a)} < \beta_A^{(A)}, \quad \forall A\neq a && \text{(Disparity assumption)} \\
&\implies \log(\lambda_t)-\beta_A^{(a)} > \log(\lambda_t)-\beta_A^{(A)}, \quad \forall A\neq a, \forall\lambda_t \\
&\implies g^{-1}\left(\log(\lambda_t)-\beta_A^{(a)}\right) > g^{-1}\left(\log(\lambda_t)-\beta_A^{(A)}\right), \quad \forall A\neq a, \forall\lambda_t \\
&\qquad\qquad (g \text{ strictly monotonically increasing} \implies g^{-1} \text{ strictly monotonically increasing}) \\
&\implies \mathbb{E}\left[g^{-1}\left(\log(\lambda_t)-\beta_A^{(a)}\right) \;\middle|\; D_t, A=a\right] > \mathbb{E}\left[g^{-1}\left(\log(\lambda_t)-\beta_A^{(A)}\right) \;\middle|\; D_t, A\neq a\right]
\end{aligned}$$

It is clear to see that this argument extends naturally to show that if a group tends to visit the hospital *more frequently* conditional on severity, that their severity will be overestimated: if there exists a group $\tilde{a}$ such that $\beta_A^{(\tilde{a})} > \beta_A^{(A)}$ for all $A\neq\tilde{a}$, then we will see that $\mathbb{E}[Z_t \mid D_t, A=\tilde{a}] < \mathbb{E}[Z_t \mid D_t]$. Thus any model that does not take into account demographic disparities in patient visit rates given their severity will lead to biased estimates of severity. □

C SIMULATIONS

Figure S1 shows the results of 30 simulation runs, where we randomly instantiate the parameters of our model and then generate data to fit on. We generate simulated data for 1000 patients on each run, each of whom is assigned to one group (50% chance of being from either group). We visualize the recovery of each parameter by plotting true parameter values versus recovered posterior mean values, with one dot per run.

To generate data with prevalent disparities, we set our priors to $\mu_{Z_0} \sim \mathcal{N}(0, 2.5)$ and $\sigma_{Z_0} \sim \mathcal{TN}(1, 0.5)$ (normal distribution restricted to positive values) for the non-pinned group; $\mu_R \sim \mathcal{N}(0, 3)$ and $\sigma_R \sim \mathcal{TN}(1, 0.01)$ (normal distribution restricted to positive values) for both groups; $F \sim \mathcal{TN}(1, 1)$ (normal distribution restricted to values above 0.5 to enforce positive constraint) for F_0; $F \sim \mathcal{N}(0, 2)$ for all other features; $b \sim \mathcal{N}(0, 1)$; $\psi \sim \mathcal{TN}(8, 1)$ (normal distribution restricted to positive values); $\beta_0 \sim \mathcal{N}(1.5, 0.1)$; $\beta_Z \sim \mathcal{N}(0.5, 0.1)$; and $\beta_A \sim \mathcal{N}(0, 2)$ for the non-pinned group.

D 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 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.

E Model Evaluation

Fitting model on real data. We fit our model on real data using weakly informative priors: $\mu_{Z_0} \sim \mathcal{N}(0,1)$ and $\sigma_{Z_0} \sim \mathcal{TN}(1,1)$ (normal distribution restricted to positive values) for the non-pinned groups; $\mu_R \sim \mathcal{N}(0,1)$ and $\sigma_R \sim \mathcal{TN}(1.5,1)$ (normal distribution restricted to positive values) for both groups; $b \sim \mathcal{N}(0,1)$; $\psi \sim \mathcal{TN}(1,0.5)$ (normal distribution restricted to positive values); $\beta_0 \sim \mathcal{N}(2.5,1)$; $\beta_Z \sim \mathcal{N}(0,1)$; and $\beta_A \sim \mathcal{N}(0,1)$ for the non-pinned group.

For F, 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.

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.

G SUPPLEMENTARY FIGURES AND TABLES

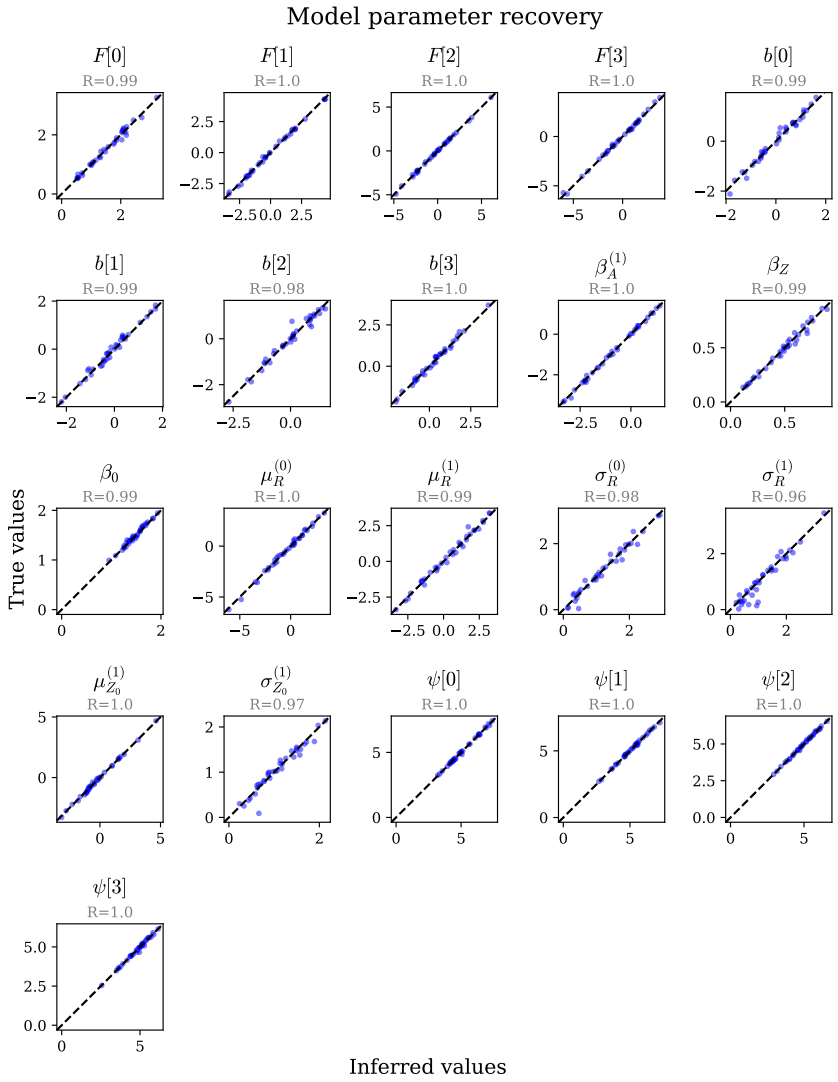

Figure S1: **Parameter recovery from fitting our model to synthetic data.** The priors from which we draw the synthetic data are: $\mu_{Z_0} \sim \mathcal{N}(0, 2)$ and $\sigma_{Z_0} \sim \mathcal{TN}(1, 1)$ (normal distribution restricted to positive values) for the non-pinned group; $\mu_R \sim \mathcal{N}(0, 2)$ and $\sigma_R \sim \mathcal{TN}(1, 1)$ (normal distribution restricted to positive values) for both groups; $F \sim \mathcal{TN}(1, 1)$ (normal distribution restricted to values above 0.5 to enforce positive constraint) for F_0; $F \sim \mathcal{N}(0, 2)$ for all other features; $b \sim \mathcal{N}(0, 1)$; $\psi \sim \mathcal{TN}(5, 1)$ (normal distribution restricted to positive values); $\beta_0 \sim \mathcal{N}(1.5, 0.2)$; $\beta_Z \sim \mathcal{N}(0.5, 0.1)$; and $\beta_A \sim \mathcal{N}(0, 2)$ for the non-pinned group.

	Our model	FA_{visit}	$\text{PCA}_{\text{visit}}$	$\text{FA}_{\text{patient}}$	$\text{PCA}_{\text{patient}}$
RMSE: informative	0.67	0.86	0.77	0.76	0.67
RMSE: all	0.82	0.89	0.77	0.77	0.72

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 root mean squared error (RMSE).

	Our model	Linear regression	Quadratic regression	Latest timestep
RMSE: informative	0.99	1.6	2.3	0.89
RMSE: all	0.98	1.8	2.5	0.98

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 root mean squared error (RMSE).

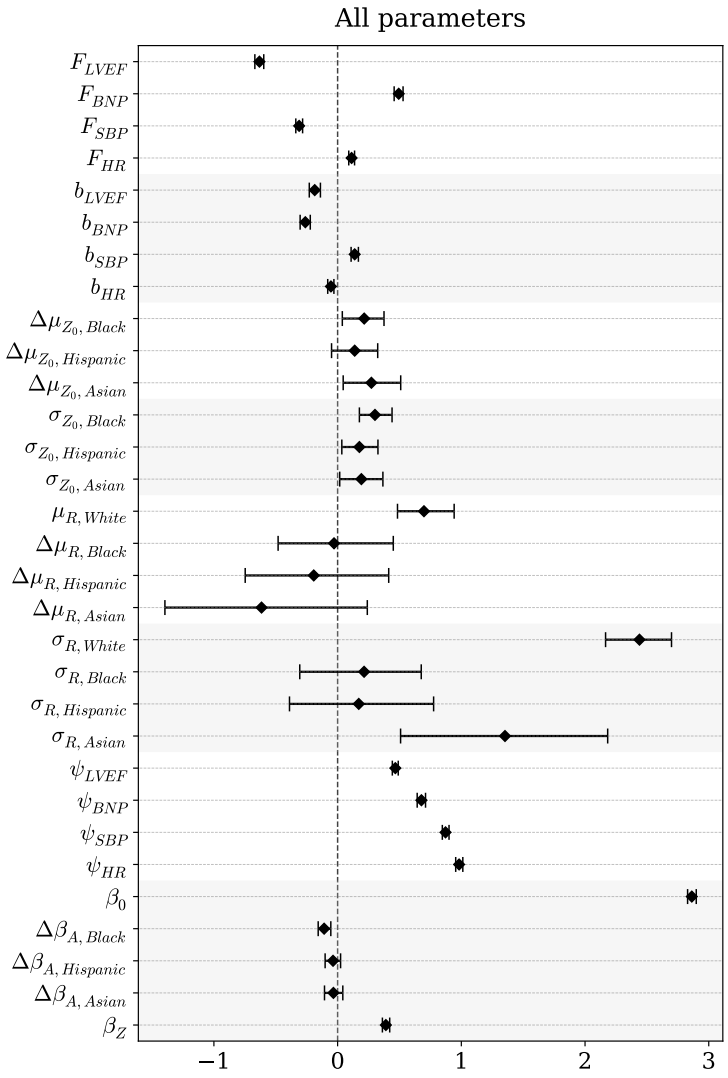

Figure S2: **All parameters learned from fitting model on NYP heart failure cohort.** Parameters of primary interest for interpreting our model (a subset of the parameters shown here) are highlighted in Figure 3.