# Aladynoulli: A Bayesian Approach to Disease Progression Modeling

## Dynamic Individual Comorbidity Modeling for Genomic Discovery and Clinical Prediction

Sarah Urbut, MD PhD

March 31, 2025

# What happened to Sarah?

- Went back to medical school ... became less nervous
- Got lost in New Haven, hated the pizza
- Detained by a virus
- Found my way to the stethoscopes
- Reacclimated with Bayes Factors

Figure: The good old days

# The call

**THE CALL**


- Short term focus
- No dynamic trajectory
- Missing Genetics ….

Figure: Antiquated Torture Device

# The Challenge: Modeling Disease Progression

- Patients develop multiple diseases over their lifetime
- Disease comorbidities follow patterns, but these vary by individual
- Genetics influences these patterns, but is not destiny
- Standard approaches miss the dynamic nature of these relationships


Figure: My new digs

# Outline

# Clinical Motivation

- Early disease is enriched for all-cause lifetime diagnoses
- Diseases don't occur in isolation
- Traditional risk models focus on short-term risk
- Missing critical information:
  - Long-term trajectories
  - Comorbidity patterns
  - Time-varying effects
  - Underlying genetic factors

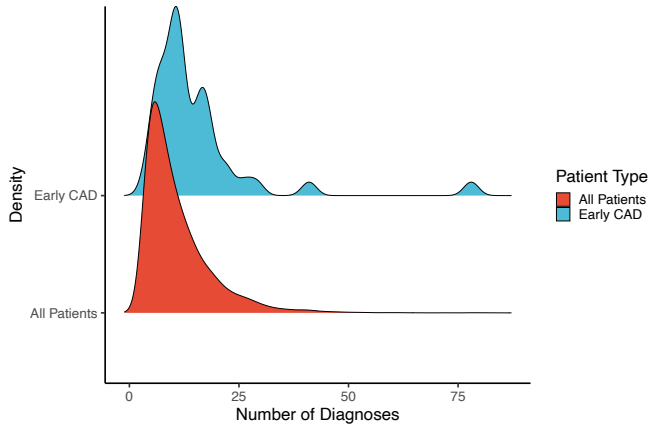



Figure: Too many charts

# Patient Timelines Reveal Diverse Patterns

- No single characteristic pattern of disease co-occurrence
- Temporal dimension adds complexity
- Different patients with the same disease may follow different pathways:

Traditional risk factors leading to mid-life CAD

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| Low HDL | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Hypertension | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| Type 2 Diabetes | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| CAD | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Heart Failure | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

CAD following inflammatory conditions

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| Rheumatoid Arthritis | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| IBD | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Psoriasis | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| CAD | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Heart Failure | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Early CAD without traditional risk factors

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| Low HDL | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hypertension | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Type 2 Diabetes | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CAD | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Heart Failure | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

CAD with psychosocial and behavioral risks

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| Depression | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Tobacco Use | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Alcohol Use | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| CAD | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Heart Failure | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

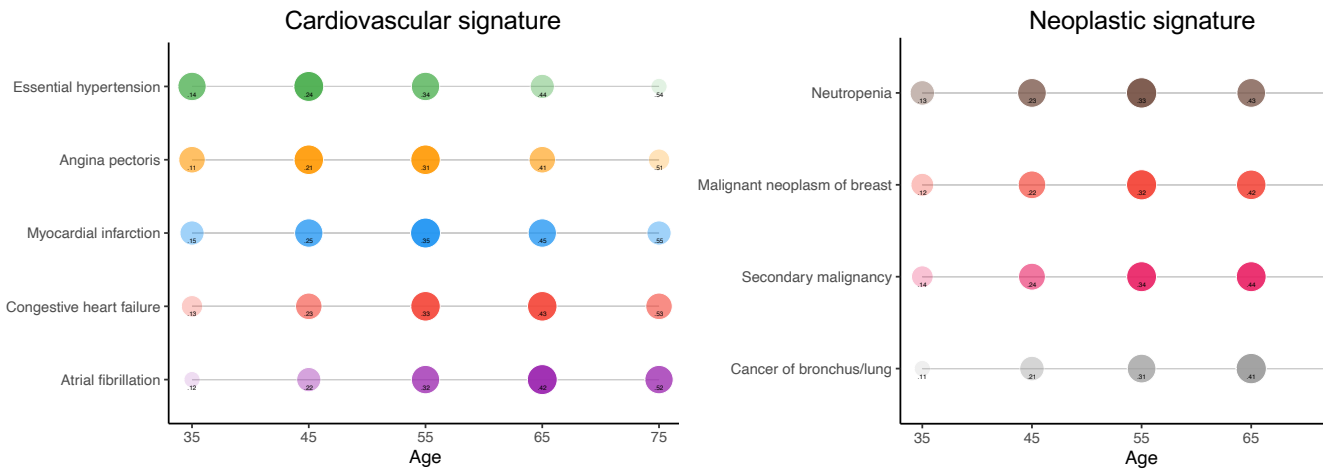
Figure: No two zebras...

- To discover patterns, we need to ask both *what* and *when*

# What are "Signatures"?

Signatures: patterns of disease co-occurrence that vary in time

- Signatures capture which comorbidities tend to occur together and when
- Disease prevalence generally increases with time
- Peak age of onset varies by condition within a signature




***Signatures:*** *patterns of disease co-occurrence that vary in time*

Figure: Time Varying Signatures

# Time-Varying Individual Trajectories

- An individual's signature profile changes dynamically over time
- Characterizing a patient as "cardiovascular" or "neoplastic" depends on when you ask

**An individual's changing signature enrichment**

**Patient A: Metabolic → Cancer**

Classic metabolic syndrome evolving into malignancy

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| Hypertension | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Type 2 Diabetes | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| CAD | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Colon Polyps | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Colon Cancer | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| Metastasis | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

**Patient B: Inflammatory → CVD → Neuro**

Inflammatory disease followed by cardiovascular complications and neurological issues

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| Rheumatoid Arthritis | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| IBD | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| CAD | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Heart Failure | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Depression | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Cognitive Decline | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

**Patient C: Early CVD → GI → Metabolic**

Early cardiovascular disease with later digestive and metabolic complications

| Condition | Age 35 | Age 40 | Age 45 | Age 50 | Age 55 | Age 60 | Age 65 |
|---|---|---|---|---|---|---|---|
| CAD | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Heart Failure | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| GERD | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| IBD | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Type 2 Diabetes | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| Obesity | 0 | 0 | 0 | 0 | 1 | 1 | 1 |

Metabolic CVD Cancer Inflammatory GI Neuro

Figure: Folks Move

# The Role of Genetics

- Genetic factors influence signature enrichment
- Polygenic risk scores show stronger effects at younger ages
- Early disease onset is enriched in high polygenic risk
- But genetics alone tells only part of the story

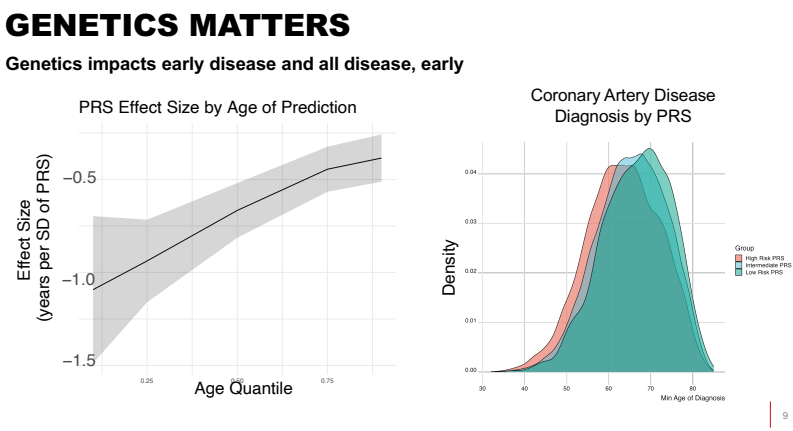



Figure: Genetics matter

# A Model-Based Approach

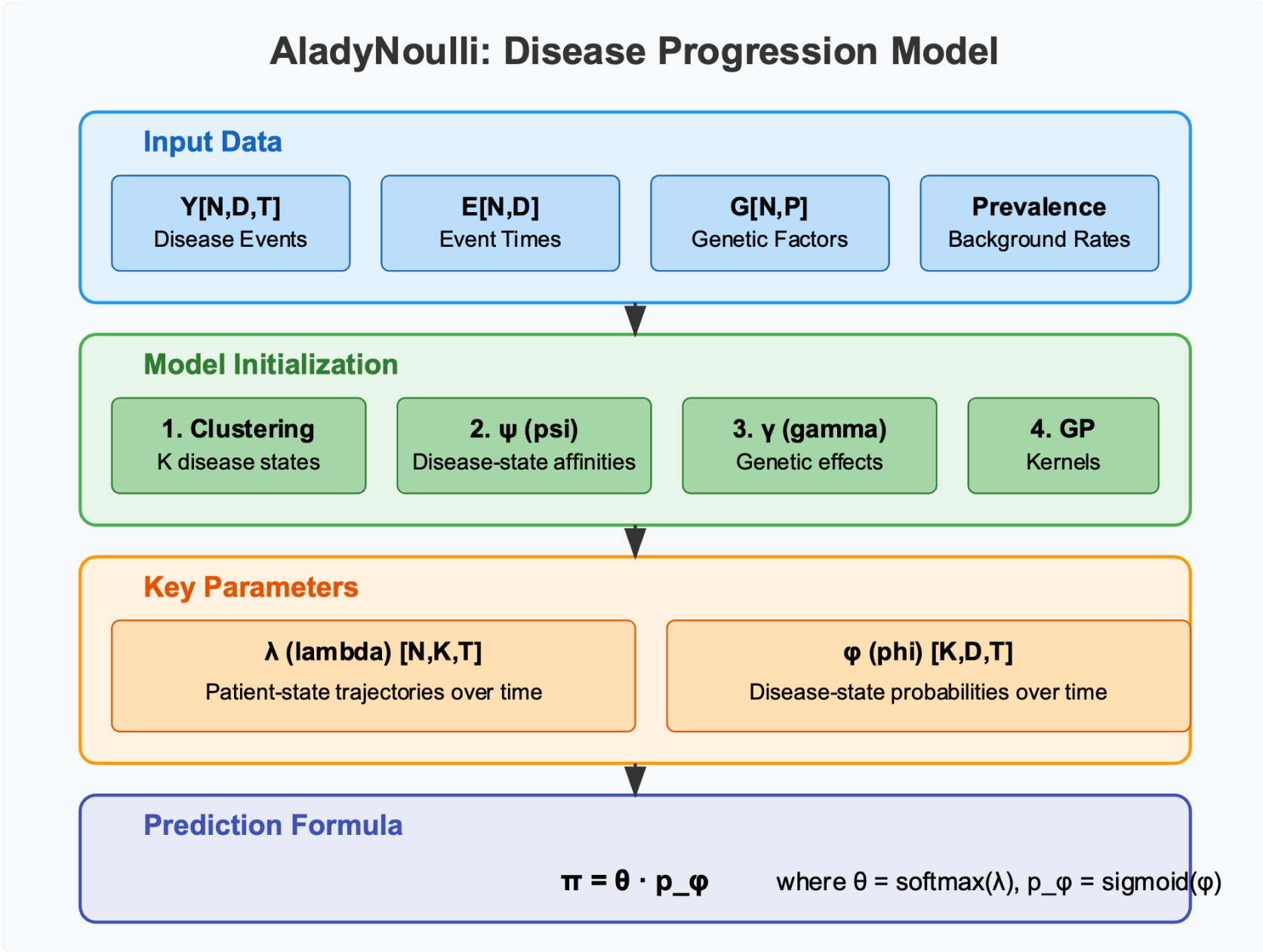
- Unsupervised clustering can miss the generative story
- Need to incorporate:
  - Population-level patterns
  - Individual predilection due to genetics
  - New specific data (diagnoses from EHR)
- Bayes theorem provides the foundation:

$$P(\Pi|\text{data}) \propto P(\text{data}|\Pi) \cdot P(\Pi)$$

# Notation

- **Dimensions:**
  - $N$ individuals, $D$ diseases, $T$ time points, $K$ topics, $P$ genetic variants
- **Data:**
  - $Y_{idt} \in \{0, 1\}$ - Disease diagnosis for individual $i$, disease $d$, time $t$
  - $g_{ip}$ - Genetic covariate $p$ for individual $i$ (PRS)
- **Model Parameters:**
  - $\lambda_{ikt}$ - Logit of topic membership (individual $i$, topic $k$, time $t$)
  - $\theta_{ikt}$ - Topic membership probability ($\theta_{ikt} = \text{softmax}_k(\lambda_{i \cdot t})$)
  - $\phi_{kdt}$ - Logit of disease probability (topic $k$, disease $d$, time $t$)
  - $\psi_{kd}$ - Topic-disease association strength
- **Prediction:**
  - $\pi_{idt} = \kappa \cdot \sum_k \theta_{ikt} \cdot \sigma(\phi_{kdt})$ - Disease probability

# Frame Title


AladyNoulli: Disease Progression Model
Input Data
Y[N,D,T]
Disease Events
E[N,D]
Event Times
G[N,P]
Genetic Factors
Prevalence
Background Rates
Model Initialization
1. Clustering
K disease states
2. ψ (psi)
Disease-state affinities
3. γ (gamma)
Genetic effects
4. GP
Kernels
Key Parameters
λ (lambda) [N,K,T]
Patient-state trajectories over time
φ (phi) [K,D,T]
Disease-state probabilities over time
Prediction Formula
π = θ · p_φ
where θ = softmax(λ), p_φ = sigmoid(φ)

# Individual Risk Profiles ($\lambda$)

- For each individual $i$, signature $k$, and time $t$:

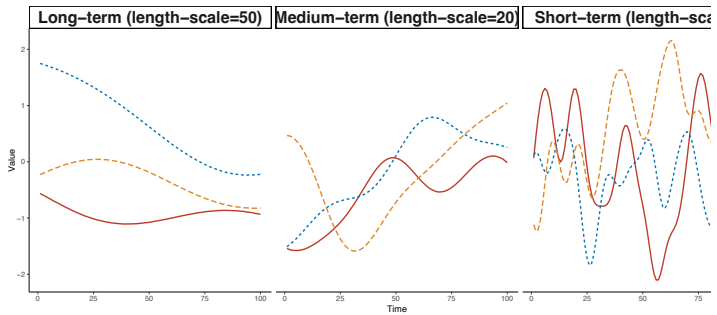
$$\lambda_{ik}(t) \sim \mathcal{GP}(\Gamma_k^T g_i, K_k)$$

- Components:
  - $g_i$: Genetic covariates (PRS)
  - $\Gamma_k$: Genetic effects
  - $K_k$: Temporal covariance matrix
- Clinical meaning:
  - Personal trajectories
  - Genetic influence
  - Smooth evolution over time

# Gaussian Process for Temporal Modeling

- Gaussian processes help with smoothing across time
- The kernel determines how quickly individual predilection can change
- Length scale parameters control the temporal patterns:
  - Long-term (length-scale = 50)
  - Medium-term (length-scale = 20)
  - Short-term (length-scale = 5)

**TRAJECTORIES WITH DIFFERENT LENGTH SCALES REFLECT DIFFERENT TEMPORAL PATTERNS**



# Signature Proportions ($\theta$)

- Convert $\lambda$ to relative weights via softmax transformation:

$$\theta_{ikt} = \frac{\exp(\lambda_{ikt})}{\sum_{j=1}^{K} \exp(\lambda_{ijt})}$$

- Properties:
  - $\theta_{ikt} \in (0, 1)$
  - $\sum_k \theta_{ikt} = 1$
  - Smooth changes over time
- Interpretation:
  - Relative risk weights
  - Competing factors
  - Dynamic profiles

# Disease Signature Loadings ($\phi$)

- For each disease $d$ and signature $k$:

$$\phi_{kd}(t) \sim \mathcal{GP}(\mu_d + \psi_{kd}, K_k)$$

- Components:
  - $\mu_d$: Base disease risk (population prevalence)
  - $\psi_{kd}$: Signature-specific shift
  - $K_k$: Signature covariance
- Clinical meaning:
  - Signature-disease links
  - Disease patterns
  - Time variation

# Disease Probabilities ($\pi$)

- Individual disease risk combines all signatures:

$$\pi_{idt} = \kappa \cdot \sum_{k=1}^{K} \theta_{ikt} \cdot \sigma(\phi_{kdt})$$

- Components:
  - Personal risk profile ($\theta$)
  - Signature contributions ($\phi$)
  - Temporal dynamics
  - Global calibration parameter ($\kappa$)
- Clinical use:
  - Risk prediction
  - Trajectory planning
  - Intervention timing

# Discrete-Time Survival Likelihood

Log-likelihood for a single individual-disease pair:

$$\begin{aligned}\ell_{id} = &\sum_{t<E_{id}} \log(1-\pi_{idt}) \quad \text{(survival until } E_{id}\text{)} \\ &+ Y_{id,E_{id}} \log(\pi_{id,E_{id}}) \quad \text{(event term if } Y_{id,E_{id}}=1\text{)} \\ &+ (1-Y_{id,E_{id}}) \log(1-\pi_{id,E_{id}}) \quad \text{(censoring term)}\end{aligned}$$

Total log-likelihood:

$$\mathcal{L}_{data} = \sum_{i,d} \ell_{id}$$

- $E_{id}$: Event/censoring time
- $Y_{id,E_{id}}$: Event indicator at $E_{id}$ (1=event, 0=censored)

# Challenge 1: Linear vs Binary Factor Models

**Linear Factor Analysis:**

$$Y = LF' + E$$
$$\text{Cov}(Y) = FF' + \sigma^2 I$$

**Binary Factor Model:**

$$\pi_{idt} = \sum_k \theta_{ikt}\sigma(\phi_{kdt})$$
$$\text{Var}(Y_{idt}) = \pi_{idt}(1 - \pi_{idt})$$

**Key Difference:**

- Linear: Mean and variance are independent
- Binary: Variance is constrained by mean

# Challenge 2: Why Linear Factors Separate Naturally

**In Linear Factor Analysis:**

$$Y = LF' + E$$
$$\text{Cov}(Y) = FF' + \sigma^2 I$$

- Each factor explains maximum remaining variance
- Orthogonality emerges naturally from optimization
- No constraints between mean and variance
- Successive factors explain residual covariance

# Challenge 3: Why Binary Factors Don't Separate

**In Our Model:**

$$\text{Var}(Y_{id}) = \pi_{id}(1 - \pi_{id})$$

$$\pi_{id} = \sum_k \theta_{ik}\sigma(\phi_{kd})$$

**Key Issues:**

- Variance tied to mean through binomial structure
- At low prevalence ($\pi \approx 0$): variance $\approx \pi$ (tiny!)
- All factors centered at $\mu_d \rightarrow$ all have same tiny variance
- No natural way to capture distinct patterns
- All factors get "squashed" to similar values by sigmoid

# Challenge 4: The Problem with Centered Topics

Without topic-specific shifts:

$$\phi_{kd} \sim \mathcal{N}(\mu_d, K_{\phi_k})$$

**Issues:**

- All topics "squashed" through sigmoid at similar values
- At low prevalence: $\text{Var}(Y) \approx \pi$ (tiny)
- All factors compete to explain variations around same tiny variance
- Disease covariance dominated by mean structure

# Challenge 5: Scale Invariance Problem

**Logit Mixture Model:**

$$\text{logit}(\pi_{idt}) = \sum_k \theta_{ikt}\phi_{kdt}$$

$$\pi_{idt} = \sigma\left(\sum_k \theta_{ikt}\phi_{kdt}\right)$$

**Scale Invariance Problem:**

$$\begin{aligned}\text{logit}(\pi_{idt}) &= \sum_k \theta_{ikt}\phi_{kdt} \\ &= \sum_k (c\theta_{ikt})(\phi_{kdt}/c)\end{aligned}$$

- **Fundamental issue:** Can arbitrarily scale up $\theta$ and scale down $\phi$ (or vice versa)
- Model is unidentifiable without additional constraints

# Solution: Topic-Specific Shifts

**Our solution:**

- Use mixture of probabilities: $\pi_{idt} = \sum_k \theta_{ikt}\sigma(\phi_{kdt})$
- Introduce $\psi_{kd}$ parameters for topic-disease associations
- Allows distinct probability ranges while maintaining identifiability

**Model Specification:**

$$\phi_{kdt} \sim \mathcal{N}(\mu_{dt} + \psi_{kd}, K_{GP})$$
$$\psi_{kd} \sim [\text{prior distribution}]$$

**Initialization:**

$$\psi_{kd}^{(0)} = \begin{cases} +1 & \text{if } d \in \text{cluster}_k \\ -2 & \text{otherwise} \end{cases}$$

# Benefits of $\psi$ Parameters

- Different baseline levels for each topic
- Creates meaningful variance differences between topics
- Maintains ability to model temporal correlations via GP

For rare disease ($\mu_d = -10$ on logit scale):

- True prevalence: $\text{sigmoid}(-10) \approx 4.5 \times 10^{-5}$
- In-cluster ($\psi = +1$): $\text{sigmoid}(-9) \approx 1.2 \times 10^{-4}$
- Out-cluster ($\psi = -2$): $\text{sigmoid}(-12) \approx 6.1 \times 10^{-6}$

## Key Points:

- $\approx$ 20x difference between in/out cluster
- Strong enough for separation
- Still maintains realistic probabilities

# Solution: Global Calibration with $\kappa$

- **Problem:** $\psi_{kd}$ creates systematic underestimation of disease probabilities
  - Each topic $k$ has positive $\psi_{kd}$ only for its signature diseases
  - Most diseases have negative $\psi_{kd}$ in most topics
  - When averaging across topics: $\pi_{idt} = \sum_k \theta_{ikt} \cdot \sigma(\phi_{kdt})$
  - Result: systematic underestimation of overall disease prevalence
- **Solution:** Global calibration parameter $\kappa$
  - Modified prediction: $\pi_{idt} = \kappa \cdot \sum_k \theta_{ikt} \cdot \sigma(\phi_{kdt})$
  - $\kappa \approx 3$ scales up predictions to match population prevalence
  - Learned during training to optimize likelihood

# Challenge: GP Numerical Stability

- **First Attempt:** Large GP amplitudes for flexibility
  - $\lambda_{\text{amplitude}} \approx 3$, $\phi_{\text{amplitude}} \approx 3$
- **Problems:**
  - Numerical instability in optimization
  - Excessive temporal variation
  - Poor convergence properties
- **Solution:** Reduced GP amplitudes
  - Fixed amplitudes: $\lambda_{\text{amplitude}} = 1$, $\phi_{\text{amplitude}} = 1$
  - Added jitter for numerical stability

# Computational Advantages of Weighted Prior

- Adjusting prior weight ($w_{GP}$) in loss function:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + w_{\text{GP}} \cdot \mathcal{L}_{\text{GP}}$$

- Equivalence:

$$w_{\text{GP}} \cdot (\lambda - \mu)^T K^{-1} (\lambda - \mu) = (\lambda - \mu)^T (w_{\text{GP}}^{-1} K)^{-1} (\lambda - \mu)$$

- Computational advantages:
  - No need to re-invert the kernel matrix when adjusting prior strength
  - Compute $K^{-1}$ once, then scale the result
  - Avoids $O(T^3)$ matrix inversion for each adjustment
- Interpretation:
  - $w_{\text{GP}} \downarrow$: Equivalent to increasing prior variance
  - $w_{\text{GP}} \uparrow$: Equivalent to decreasing prior variance

# Final Model Architecture

$$\pi_{idt} = \kappa \cdot \sum_k \theta_{ikt} \cdot \text{sigmoid}(\phi_{kdt})$$

$$\theta_{ikt} = \text{softmax}_k(\lambda_{i \cdot t})$$

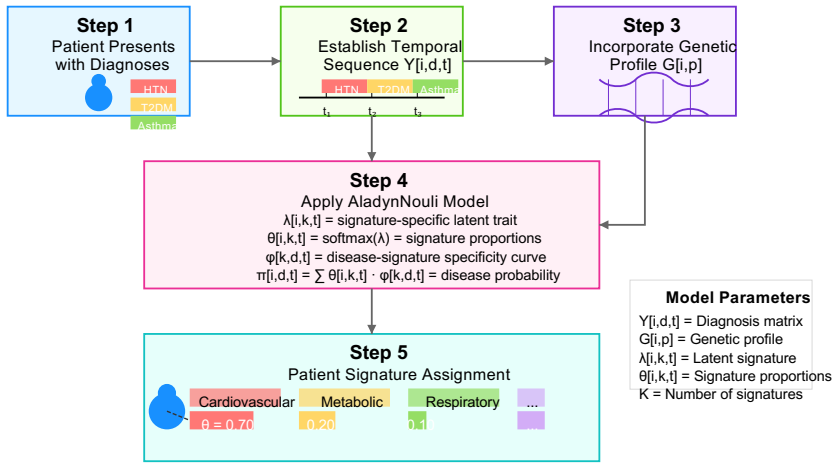
$$\boldsymbol{\lambda}_{ik} \sim \mathcal{N}(r_k + \Gamma_k^\top \mathbf{g}_i, K_\lambda)$$

$$\boldsymbol{\phi}_{kd} \sim \mathcal{N}(\boldsymbol{\mu}_d + \psi_{kd}, K_\phi)$$

**Key Improvements:**

- Topic separation via $\psi_{kd}$ parameters
- Stable GP priors with fixed amplitudes
- Balanced prior weight for flexibility
- Structured initialization from disease clusters

## Patient Journey: From Diagnoses to Signature Assignment

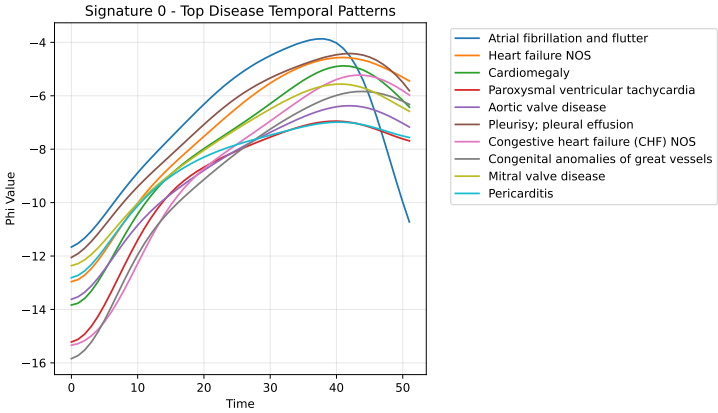


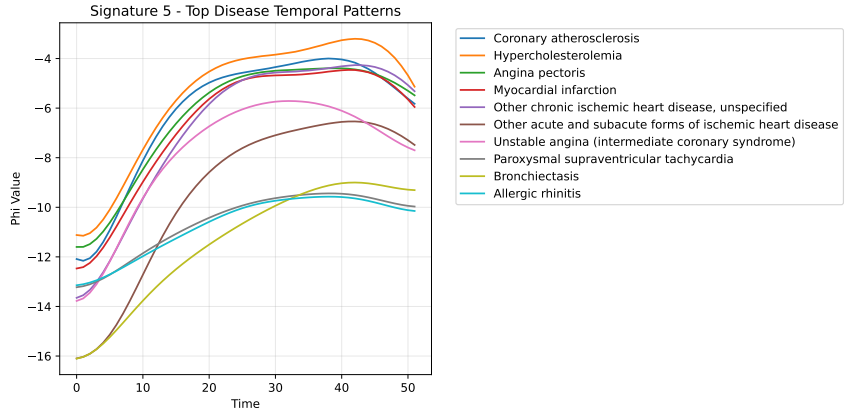

*Model optimizes λ, φ, γ, ψ to maximize likelihood of observed diagnoses*

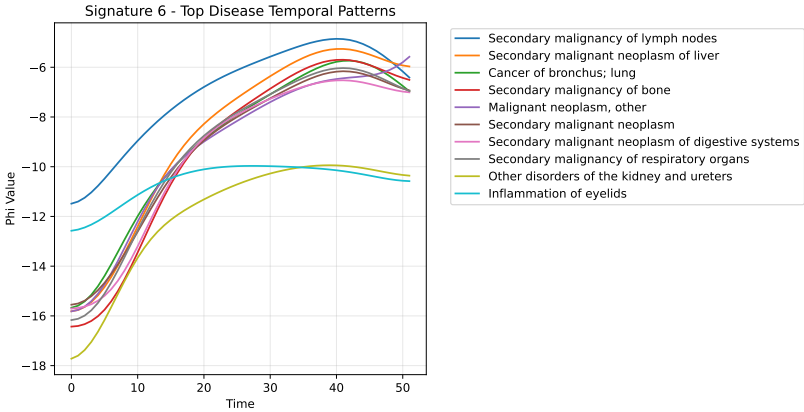
# Real Data Application

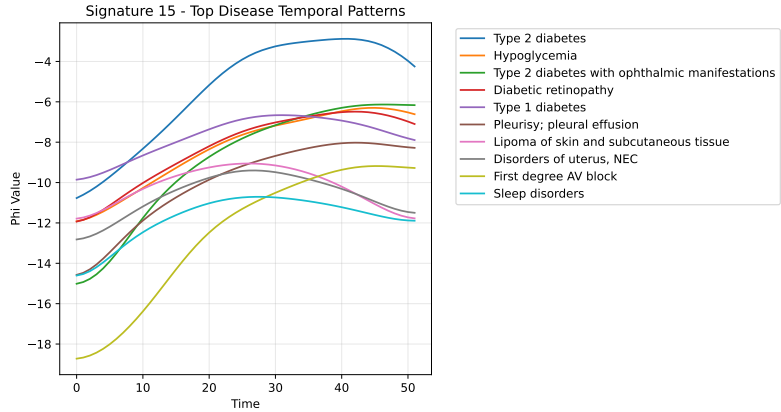
- UK Biobank: 460,000 individuals
- 348 conditions (¿2% lifetime prevalence)
- 51 time points (ages 30-80 years)
- 36 unique externally validated PRS
- Validation in All of Us dataset ( 340,000 individuals)

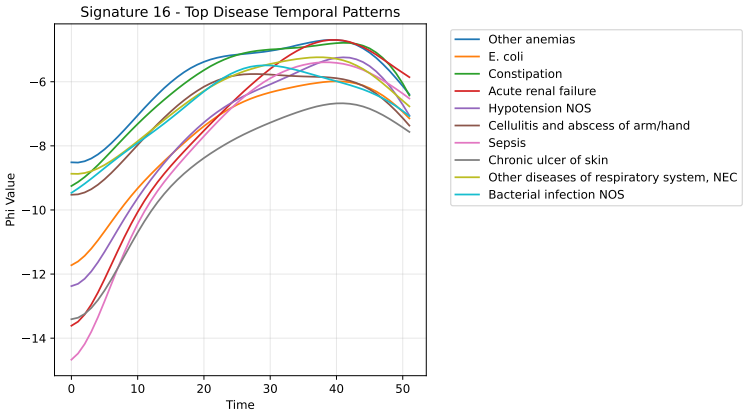
# Disease-specific Signature Patterns

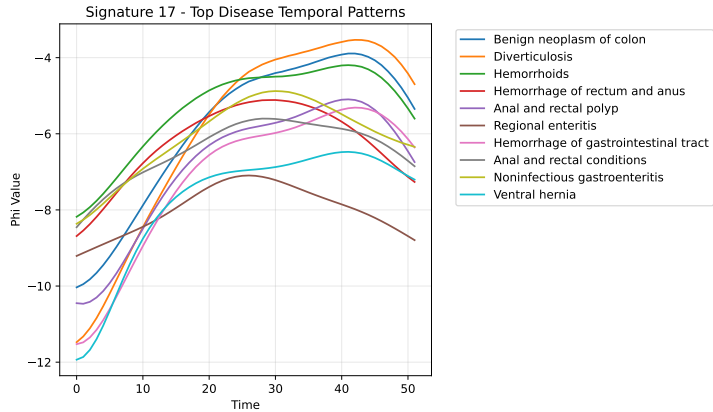


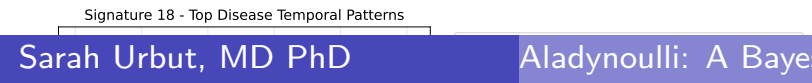



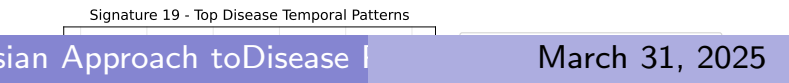



# Case Study: At the patient level

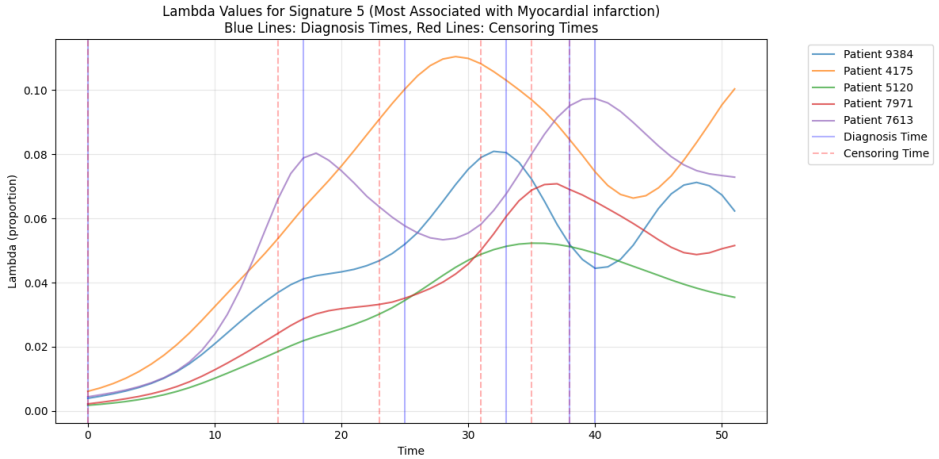



Figure: Enter Caption

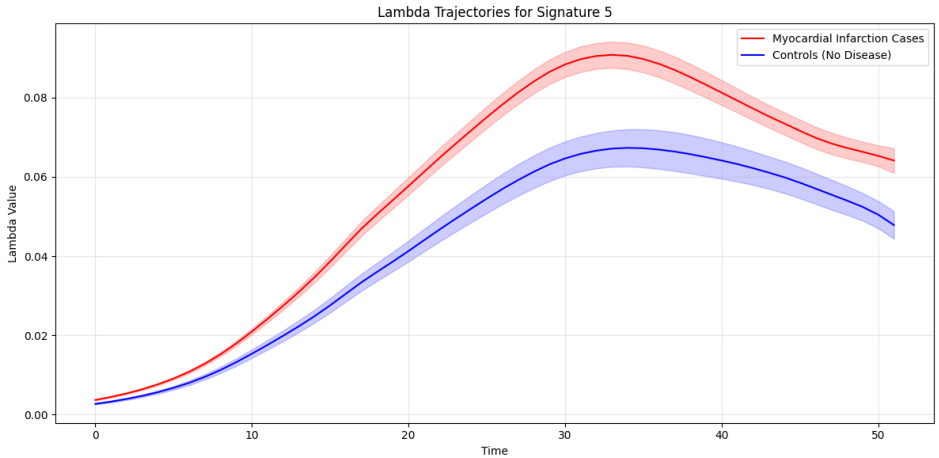



Figure: Enter Caption

# Case Study: Multimorbid Patient

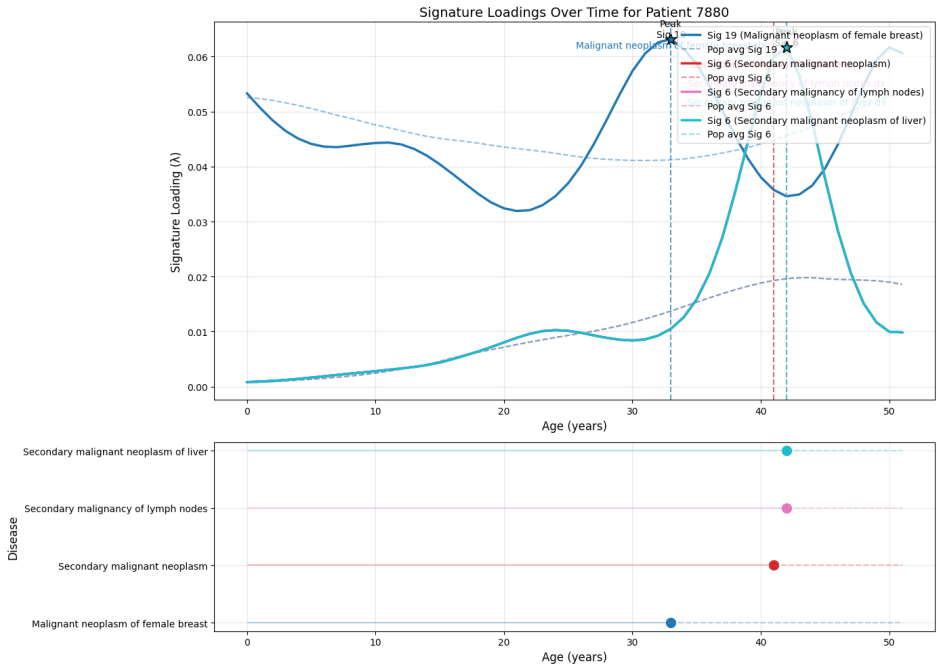



Figure: Enter Caption

# Population Heterogeneity

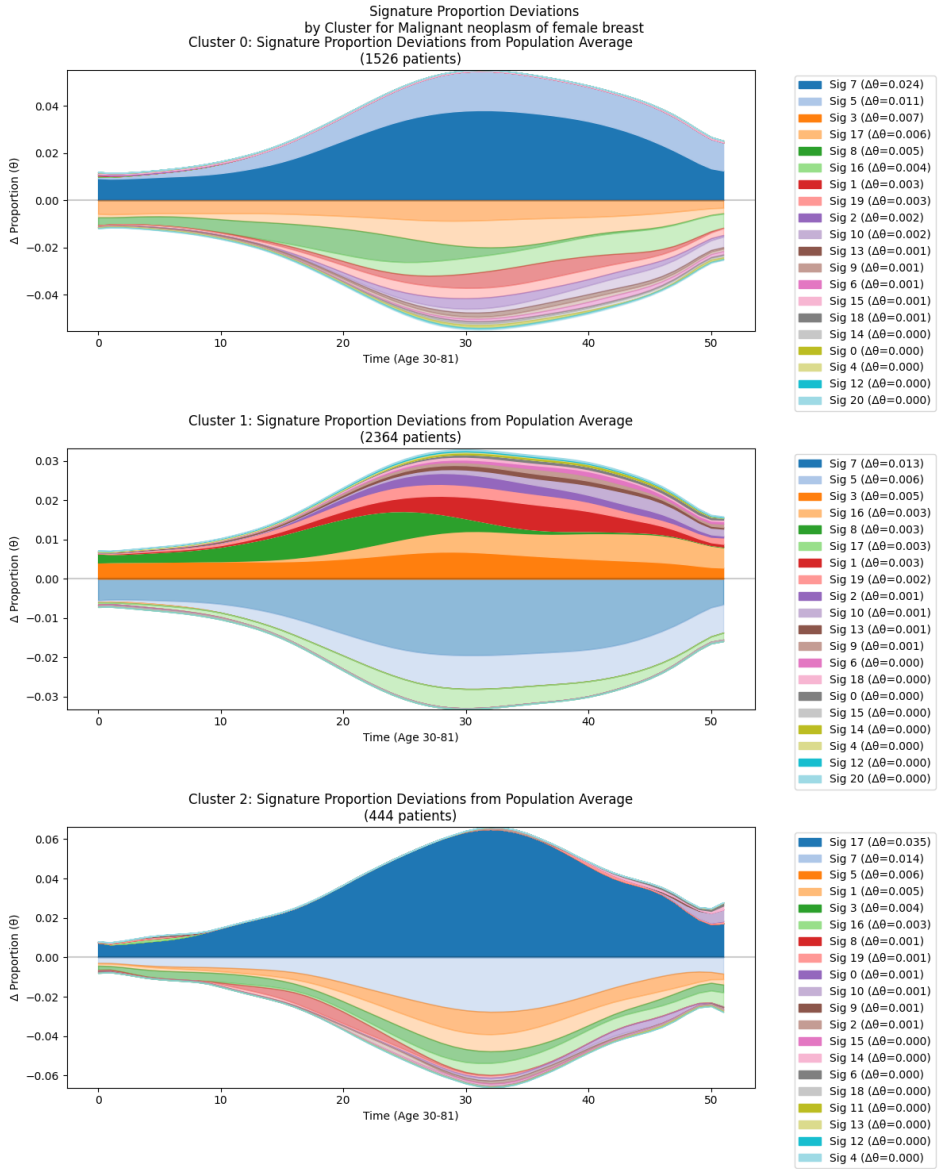



Figure: Enter Caption

# Population Heterogeneity

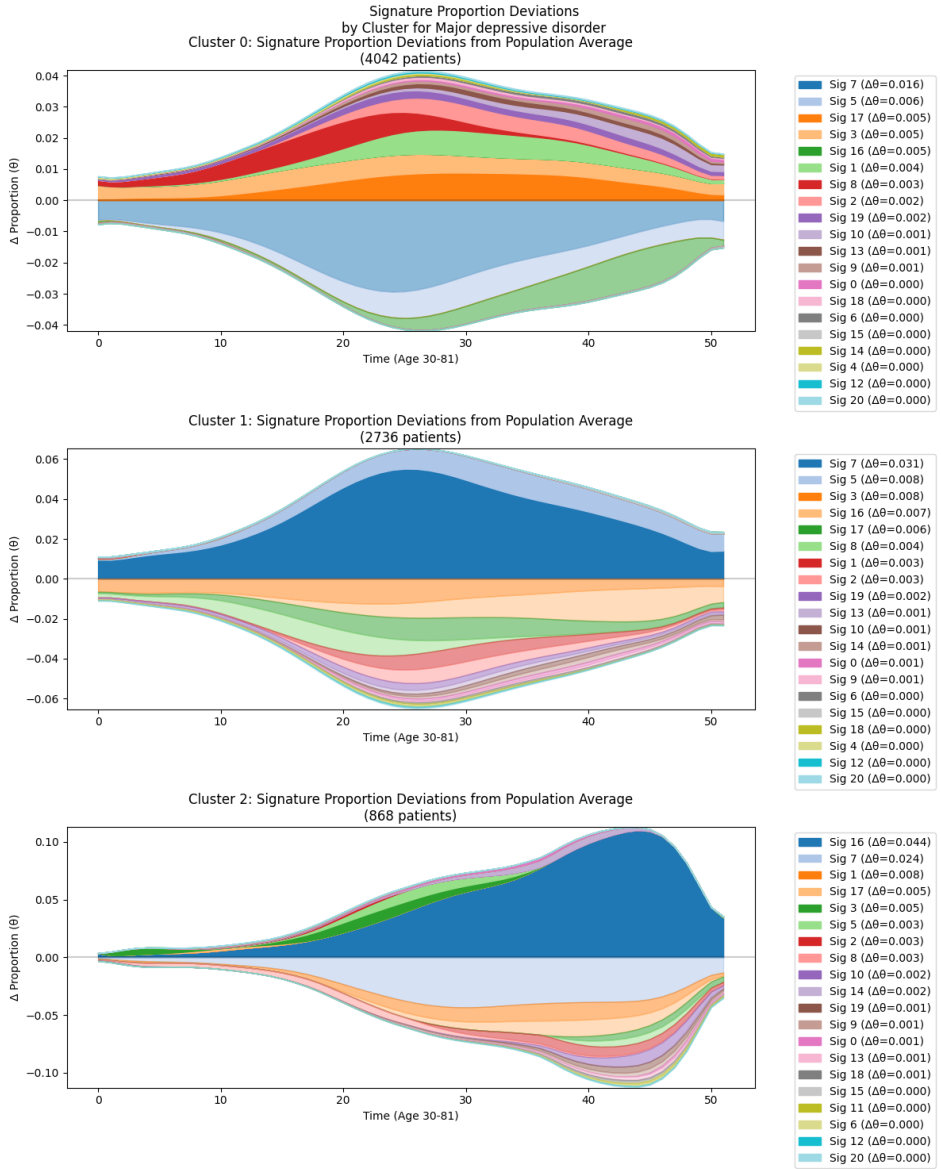



Figure: Enter Caption

# Population Heterogeneity

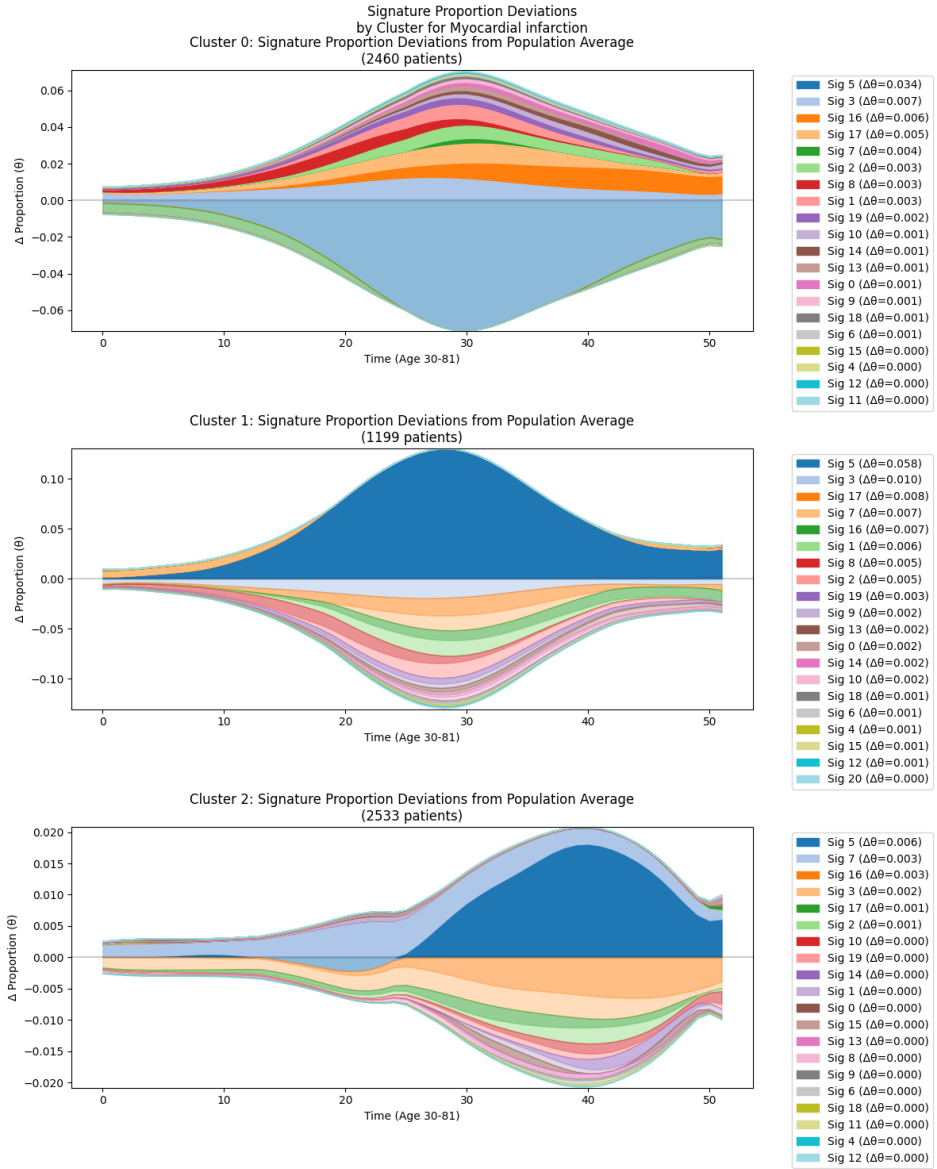



Figure: Enter Caption

# Genetic Factors by Signature

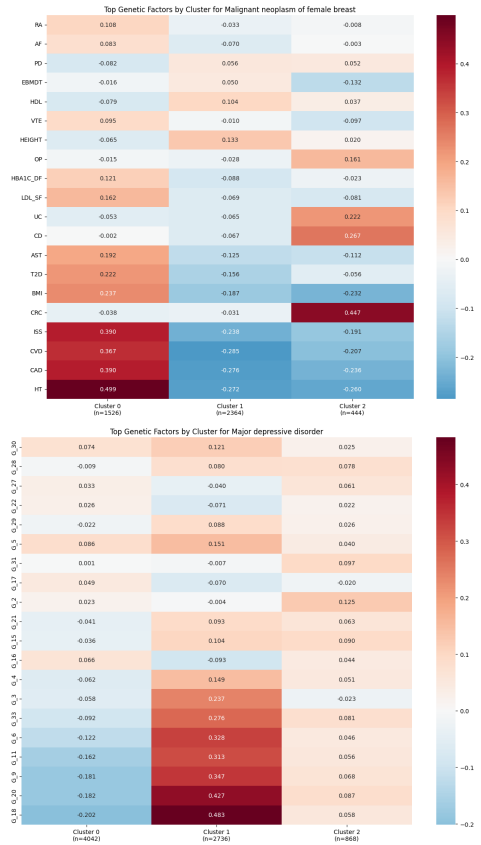



Figure: Enter Caption

- Different genetic factors influence each signature
- Shows underlying biology of disease clusters
- Enables biological interpretation of signatures

# Prediction Performance

[Insert AUC plot here]

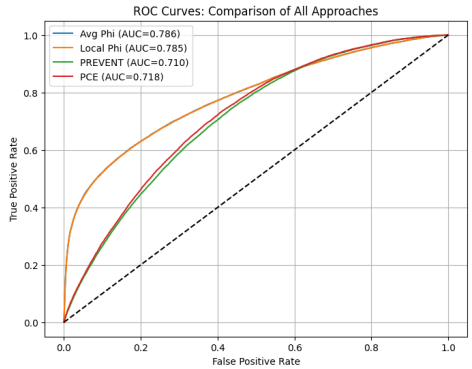



Figure: Enter Caption

- AUC for remaining lifetime risk across all diseases
- Aladynoulli (red) outperforms age and sex-based models (orange)
- Consistent improvement across disease groups
- Borrowing information dynamically improves

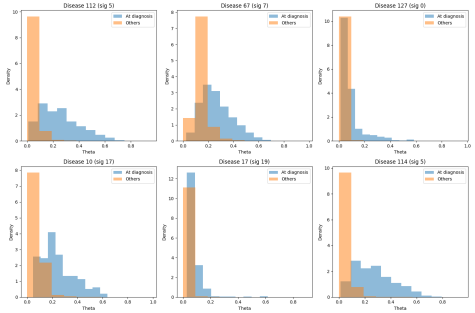


Figure: Enter Caption

# Frame Title

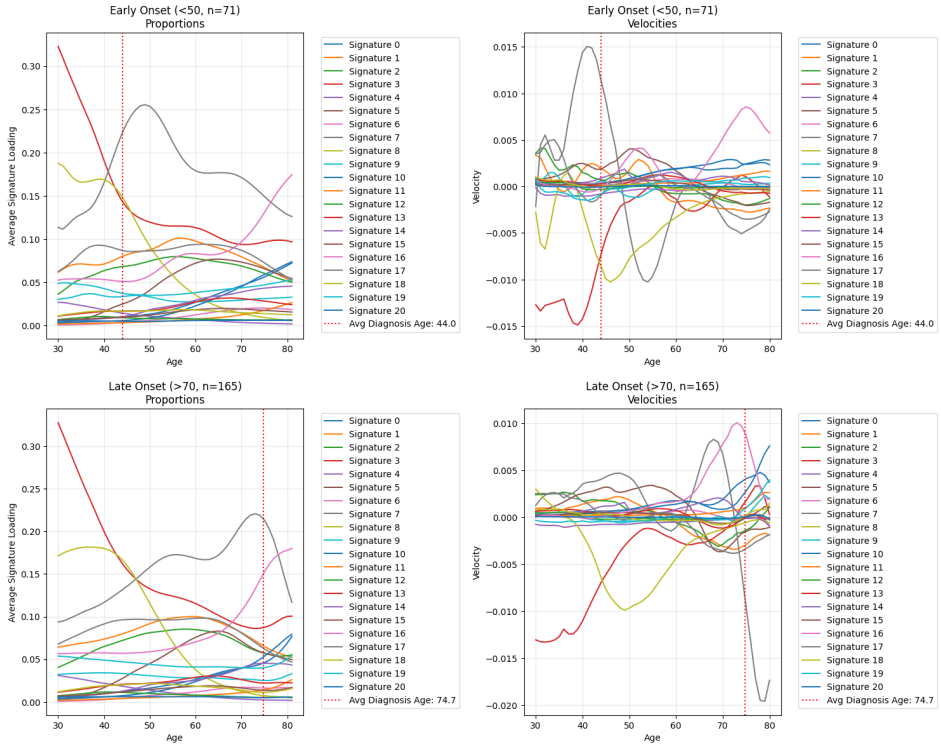



Figure: Enter Caption

# Conclusions

- Aladynoulli provides a comprehensive framework for modeling disease progression
- Key innovations:
  - Topic separation via $\psi$ parameters
  - Stable GP priors with fixed amplitudes
  - Balanced prior weight for flexibility
  - Structured initialization from disease clusters
- Applications:
  - Disease Progression Modeling
  - Genetic Association
  - Risk Prediction
  - Disease Clustering
  - Personalized Medicine

# Resources and Availability

- Paper: `https://www.medrxiv.org/content/10.1101/2024.09.29.24314557V1`
- Github: `github.com/surbut/Aladynoulli`
- Interactive app: `Surbut.shinyapps.io/berndiffapp`

Thank you!