# Supplementary Materials for

# Aladyn Individual: Bayesian Hierarchical Dynamic Genetic Modeling of Comorbidity Progression

**This PDF file includes:**

Materials and Methods

Figures S1 to S3

Tables S1 to S4

# Materials and Methods

## 8.1 Updating Genetic Parameters

The genetic parameters $\rho$ and $\gamma$ are updated using Bayesian inference as follows:

### 8.1.1 Updating $\rho$

The posterior distribution of $\rho$ given the observed data $D$ can be expressed as:

$$P(\rho|D) \propto P(D|\rho) \cdot P(\rho) \tag{8}$$

Where $P(D|\rho)$ is the likelihood of the data given $\rho$, and $P(\rho)$ is the prior distribution of $\rho$.

We can use a Metropolis-Hastings algorithm to sample from this posterior:

**Algorithm 3** Metropolis-Hastings for updating $\rho$

1: Propose a new $\rho^*$ from a proposal distribution $q(\rho^*|\rho)$

2: Calculate the acceptance ratio:

3: $\alpha = \min\left(1, \frac{P(D|\rho^*) \cdot P(\rho^*) \cdot q(\rho|\rho^*)}{P(D|\rho) \cdot P(\rho) \cdot q(\rho^*|\rho)}\right)$

4: Accept $\rho^*$ with probability $\alpha$

### 8.1.2 Updating $\gamma$

Similarly, for $\gamma$:

$$P(\gamma|D) \propto P(D|\gamma) \cdot P(\gamma) \tag{9}$$

We can use a similar Metropolis-Hastings algorithm or, if conjugate priors are used, closed-form updates may be available.

### 8.1.3 Joint Update

In practice, we may want to update $\rho$ and $\gamma$ jointly to account for their potential correlation:

$$P(\rho, \gamma|D) \propto P(D|\rho, \gamma) \cdot P(\rho, \gamma) \tag{10}$$

This can be done using a multivariate proposal distribution in the Metropolis-Hastings algorithm.

## 8.2 Convergence and Practical Considerations

Several practical considerations should be taken into account when implementing these updates:

- Monitor convergence using multiple chains and Gelman-Rubin statistics:

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}} \tag{11}$$

  where $\hat{V}$ is the between-chain variance and $W$ is the within-chain variance.

- Consider adaptive MCMC methods to improve mixing and convergence speed. For example, the adaptive Metropolis algorithm haario2001adaptive can be used to automatically tune the proposal distribution.

- Computational trade-offs and potential parallelization strategies should be considered. For instance, updates for different individuals can be parallelized, as can the likelihood calculations for different topics.

These considerations ensure robust and efficient inference of the genetic parameters within our model.

# 9 Supplementary Figures

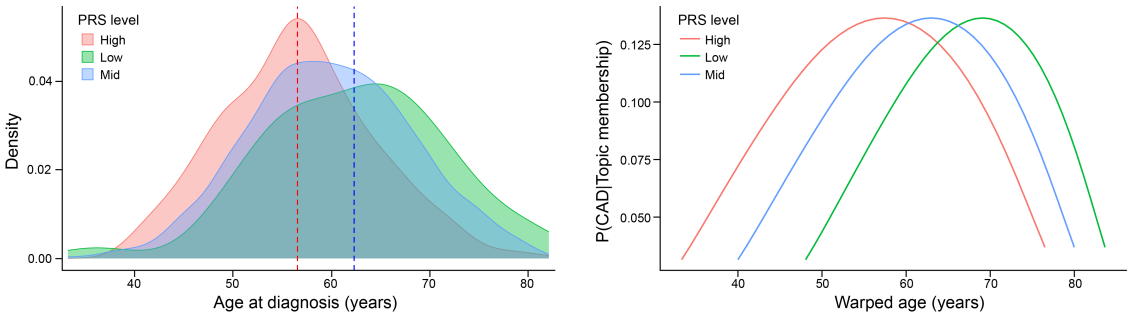


**Figure S1**: **Age of Onset Captured by Warped Disease Probabilities**. Age of diagnosis for coronary artery disease, tracking closely with underlying genetic risk. We overlay the warped time predicted disease probability of those in the top and bottom deciles of polygenic risk for coronary artery disease. In panel at right, we demonstrate the probability of disease ($\beta_{ikt'}$) for individuals in each genetic category.

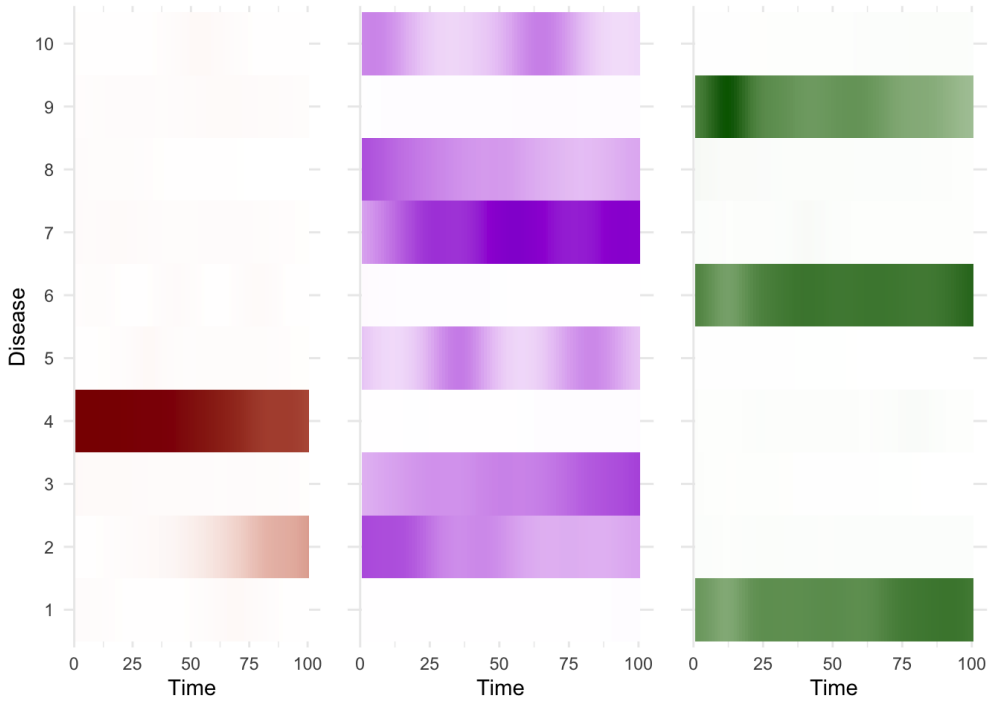


**Figure S2**: **Topic Specificity**. In both real data and simulations, we recognize that diseases tend to be sparse in the number of topics on which they are loaded.

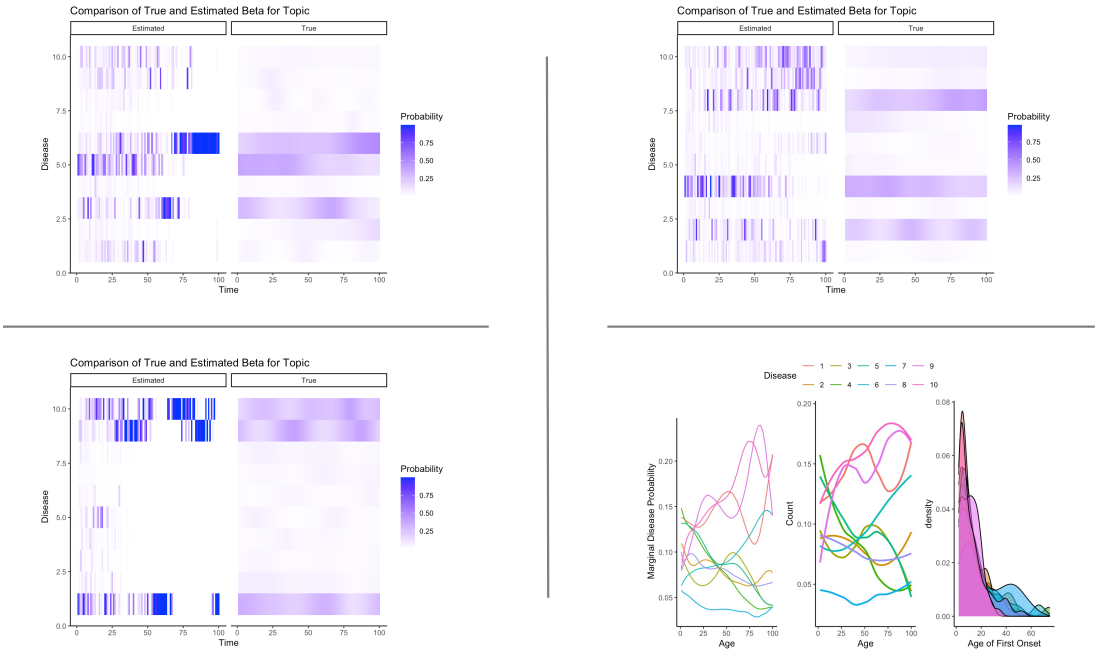


**Figure S3**: **Estimating Disease Loadings**. Here we demonstrate the approach for estimating disease loadings using counts of disease occurrences mapped to unwarped times. Lower right, we demonstrate the marginal probabilities of each disease, the estimated counts, and the first occurrence. Average Marginal probabilities defined as average($\theta_{i,,t} \times \beta_{k,,t}$) where $\beta$ represents the unscaled (population level) disease probabilities across time.

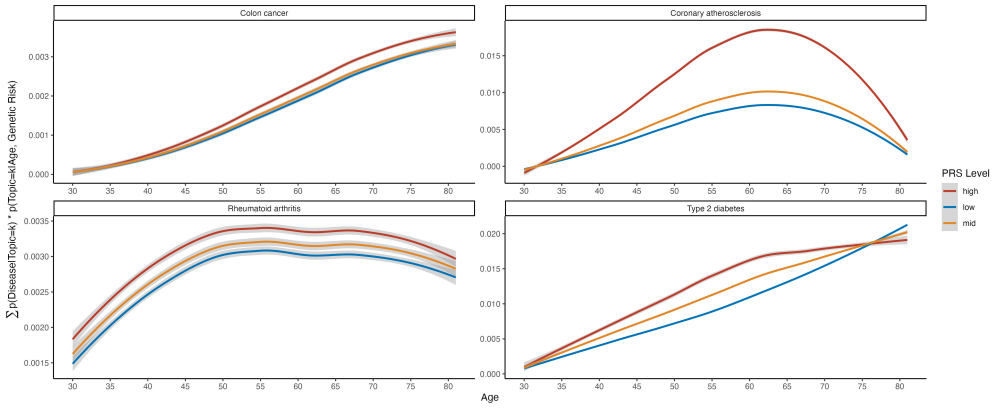


**Figure S4**: **Genetically Enriched Individuals Show Earlier Onset Disease**. Here we show that the marginal probability of disease is earlier for those with high genetic risk. We use a canonical PRS for each topic.

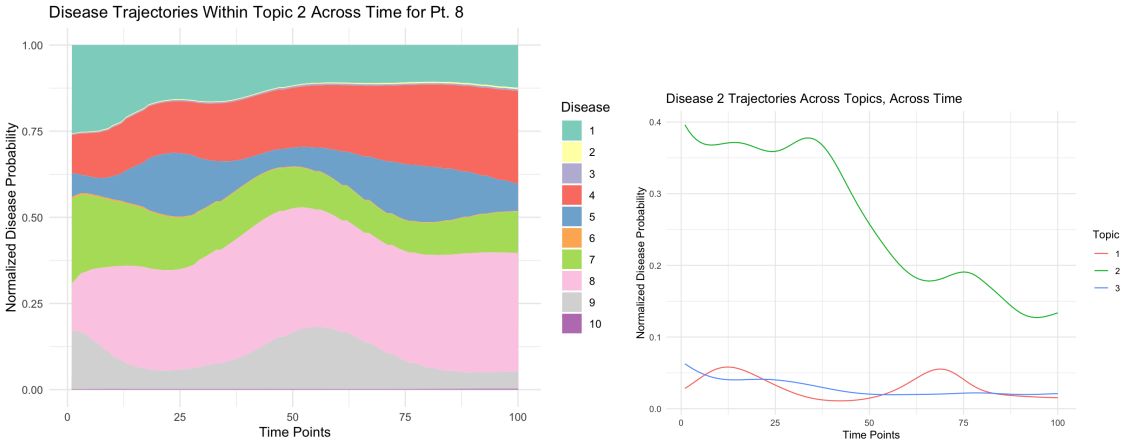


**Figure S5**: **Topic-specific Disease Probability**. We plot the probabilities of all 10 simulated diseases over time within a given topic, conditional on the time scale of a chosen patient. In B, we demonstrate the trajectory of one chosen disease across all topics. This is simulated data in which the diseases are simulated to be topic-specific so that each disease is minimally loaded on a few topics.

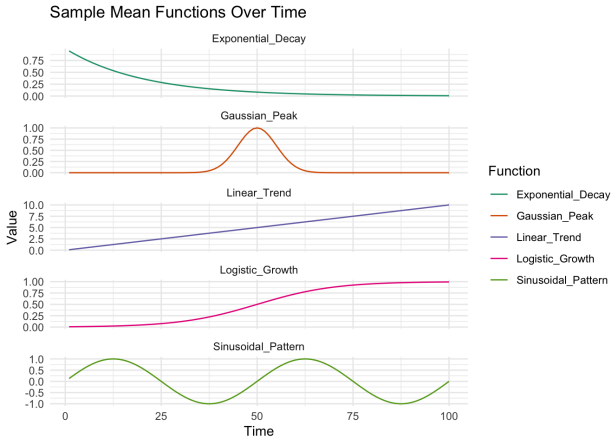


**Figure S6**: **Sample Mean**. Here we demonstrate several mean functions that govern the process of disease evolution. These are meant to reflect a sampling of biological processes and are learned from the model.

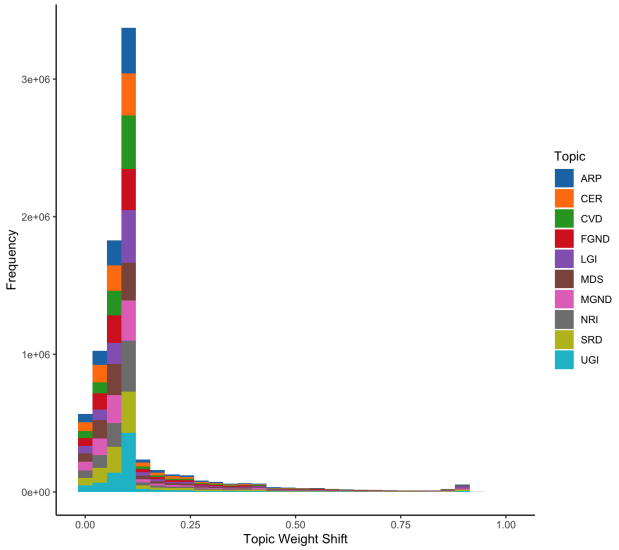


**Figure S7**: **Change in Topic Weights Over Time**. Here we consider the difference in estimated topic weight under a model with time-varying topic weights in comparison to the time-fixed approach ([1]). Results by topic do not reveal systematic differences. NRI = Neoplastic Respiratory, CVD = Cardiovascular, FGND = Female genitourinary, MGND= male genitourinary, CER= Circulatory, UGI=Upper Gastrointestinal, LGI=Lower Gastrointestinal, SRD=Sense respiratory depression, MDS = Musculoskeletal, ARP: Arthropathy.

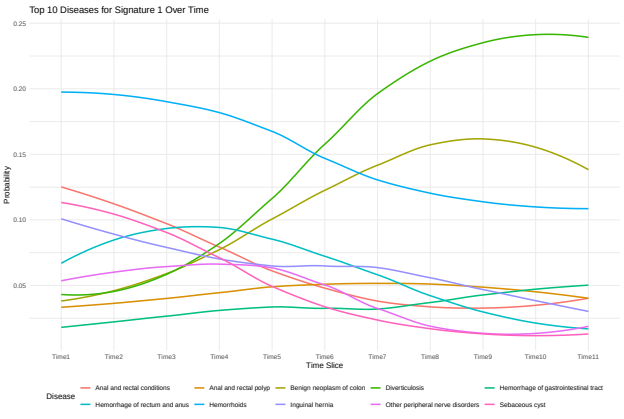


**Figure S8**: Top 10 diseases for topic 1 in the UK Biobank

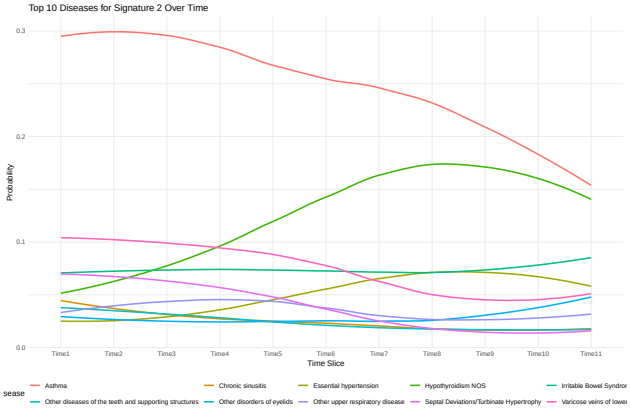


**Figure S9**: Top 10 diseases for topic 2 in the UK Biobank

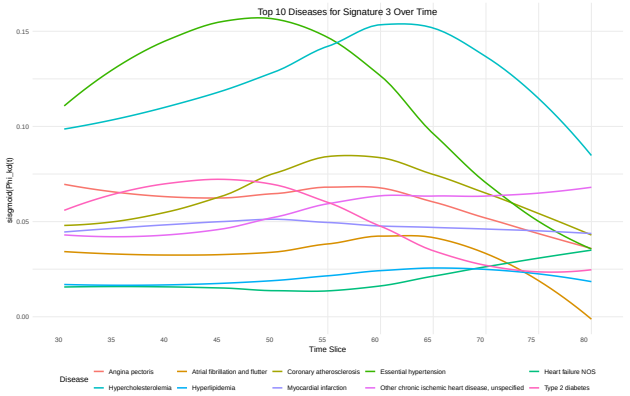


**Figure S10**: Top 10 diseases for topic 3 in the UK Biobank

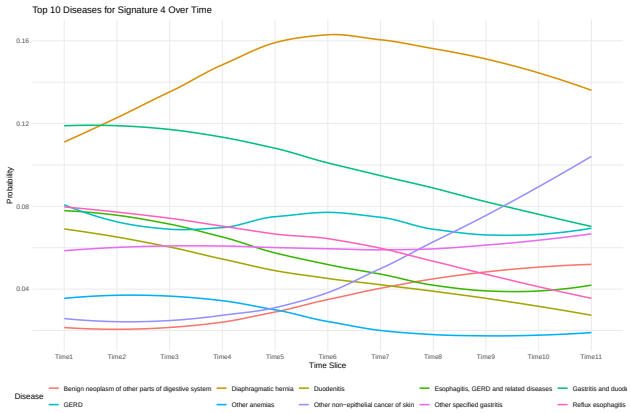


**Figure S11**: Top 10 diseases for topic 4 in the UK Biobank

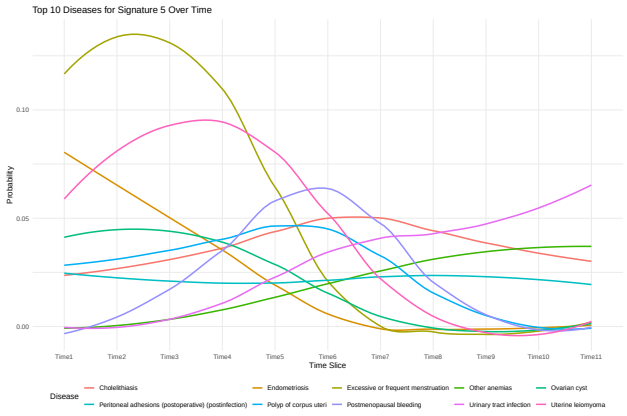


**Figure S12**: Top 10 diseases for topic 5 in the UK Biobank

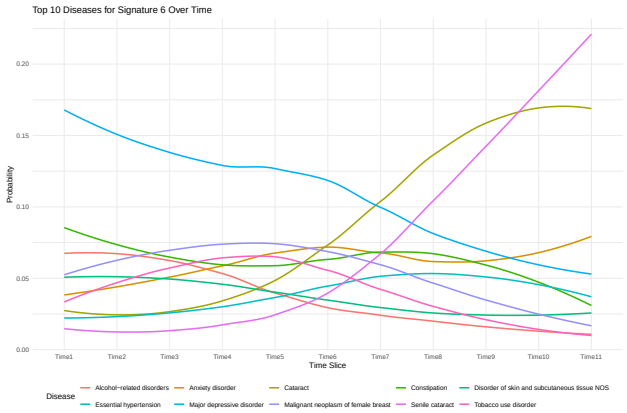


**Figure S13**: Top 10 diseases for topic 6 in the UK Biobank

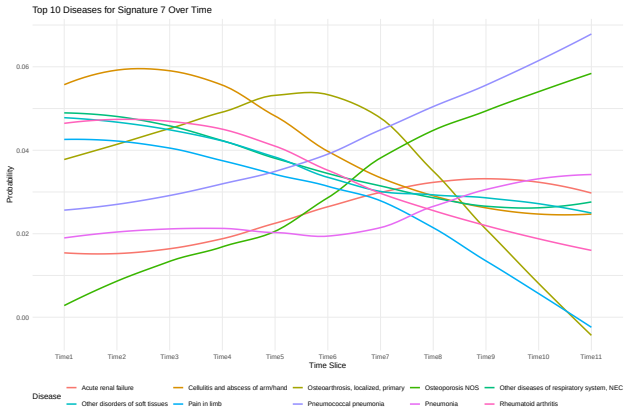


**Figure S14**: Top 10 diseases for topic 7 in the UK Biobank

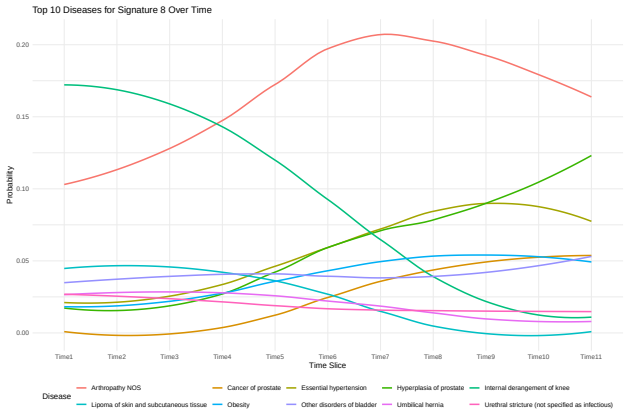


**Figure S15**: Top 10 diseases for topic 8 in the UK Biobank

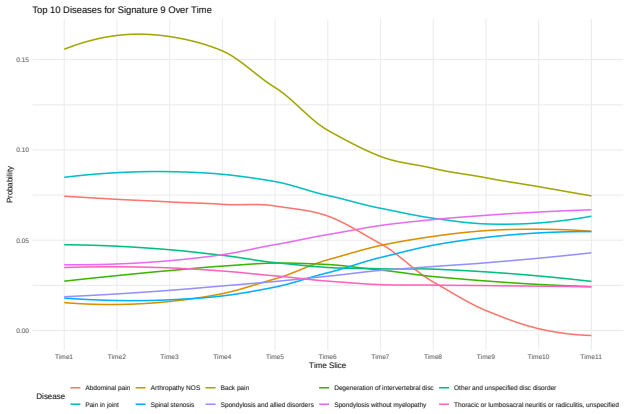


**Figure S16**: Top 10 diseases for topic 9 in the UK Biobank

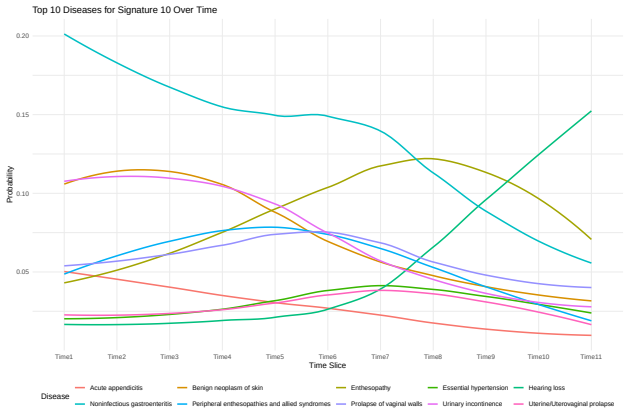


**Figure S17**: Top 10 diseases for topic 10 in the UK Biobank