Aladyn Individual: Bayesian Hierarchical Dynamic Genetic Modeling of Comorbidity Progression

Sarah Urbut[1,2], Yi Ding[3], Xilin Jiang[2,4,5], Whitney Hornsby[1,2]

Alexander Gusev[2,3*], Pradeep Natarajan[1,2*], Giovanni Parmigiani[3,5*]

[1]Cardiology Division, Massachusetts General Hospital, Boston, MA 02114, USA.

[2]Broad Institute, Cambridge, MA 02115, USA. [3]Dana Farber Cancer Institute, Cambridge, MA 02115, USA.

[4]University of Cambridge, GB,UK

[5]Harvard T.H. Chan School of Public Health, Boston, MA 02114, USA.

[*]co-supervised this work, corresponding author: gp@jimmy.harvard.edu

Abstract: Early identification of high-risk individuals through the analysis of their unique disease trajectories has a strong potential to support efficient prevention and clinical management across a range of chronic conditions. In this paper we present a novel approach for dynamic modeling of the evolution of chronic disease risks over time, incorporating individual genetic predispositions. Our approach uses a hierarchical Bayesian topic model including Gaussian Processes to capture age effects. It accounts for genetic predisposition through a time-warping function and topic-dependent genetic scores, enabling both simultaneous learning and updated predictions of complex comorbidity patterns, inclusive of genomic and non-genomic effects. We systematically compare to previous approaches and provide detailed simulations at `https://bookdown.org/sarahmurbut/dynamic_ehr/` and `https://surbut.shinyapps.io/dynamic_ehr`.

Genetic Modeling, Disease Progression, Precision Medicine, Bayesian Inference, Gaussian

Processes

Impact Statement

Our model significantly advances healthcare for aging populations by facilitating the early identification of high-risk individuals through the analysis of their unique disease trajectories within complex comorbidity patterns. Existing models are limited in their capacity to manage diverse comorbidity patterns, particularly those that are time-dependent. We introduce an approach leveraging Bayesian hierarchical modeling to concurrently learn population-level patterns and provide updated real-time predictions across 350 diseases, thereby uncovering and forecasting intricate comorbidity patterns. This methodology paves the way for preventive measures and targeted interventions that enhance health outcomes, mitigate late-stage disease burdens, and foster healthier aging. Furthermore, our model incorporates genetic influences via a genetic predisposition parameter to estimate the lifetime risk of specific diseases and disorders, alongside a time-warping function to facilitate personalized predictions of disease trajectories.

1 Introduction

Understanding the evolution of an individual's disease risk across their lifespan is crucial to advancing personalized medicine. This insight is essential for developing therapeutics tailored to individual patients rather than their diagnosed conditions. Current predictive models, which depend on static health states, often fail to capture the complexities of individual disease progression, particularly among aging populations, intricate disease interactions, and genetic influences. Recent methodologies ([1], [2]) have sought to analyze more sophisticated data types to identify unique disease patterns within extensive healthcare systems. These methodologies use data from large populations to discern diverse patterns of comorbidity, providing insight into the unique trajectories of complex diseases. Nevertheless, these approaches are not without limitations: firstly, temporal analysis of disease patterns typically captures population-level trajectories, overlooking individuals who progress through diseases at varying rates due to underlying genetic factors. Secondly, these methods often aim to classify patterns at the population level rather than provide predictive insights.

Lastly, even the most advanced approaches tend to aggregate an individual's health conditions across all accumulated diagnoses, failing to offer a time-dependent, individual-level profile that may evolve with new diagnoses and treatments. For example, prior research [3] has demonstrated that patients within the top 5% of genetic risk for myocardial infarction experience events nearly a decade earlier than the general population and follow an accelerated pathway through associated comorbidities, for which existing population-based methods offer minimal predictive capability. Our findings reveal significant disparities in disease progression; individuals in the top 5% of genetic risk for myocardial infarction encounter events nearly a decade earlier than those at lower risk (median age of onset 53.7 [53.1-53.9] vs 62.6 years [61.4-63.1]). Our model dynamically updates and predicts accurate diagnoses 79.5% more frequently than existing topic-modeling approaches [1]. The dynamic approach adjusts disease profiles by at least 11.7% annually for 50% of the individuals in the UK Biobank and All of Us cohorts, and 65.4% experience a change greater than 10% in comorbidity profiles. Simulation results corroborate the model's superior accuracy [51.4-41%], precision [50.9 vs 39.9%], and recall [48.4 vs 39.2%] compared to fixed-weight approaches.

This study introduces the Aladyn model, highlighting its potential for incorporating genetic factors into the analysis of disease progression. We present evidence for the model's functionalities, utilizing estimated loadings derived from extant dynamic topic models as a foundational basis. Our primary emphasis is on the innovative method of dynamically updating individual weights, which sets our approach apart from existing methodologies.

2 Methodological Motivation

2.1 Overview

The principal aim of this study is to model temporal variations in disease probabilities across various comorbidity profiles, incorporating underlying genetic factors while considering several essential attributes. Although there are numerous clustering frameworks for electronic healthcare data ([1], [4], [2]), they face challenges in deriving shared biological interpretations, inadequately capture individual-level deviations from population trends, and fail to address temporal variations. Certain frameworks within topic modeling can approximately address this challenge, as documents

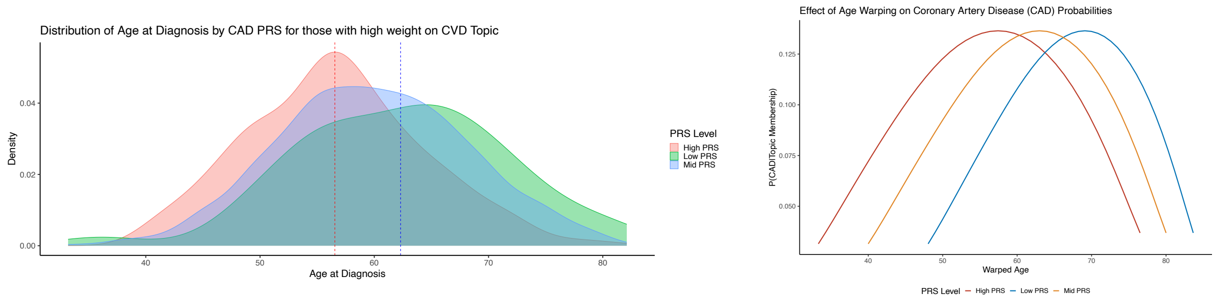

Figure 1: **Differences in age of onset for prevalent conditions**. *As a motivating example, we demonstrate the variation in the age of onset for a variety of common conditions within classically defined categories: this figure illustrates the range of onset ages for common conditions in the UK Biobank, N=421,707, age range 28-81.*

typically contain multiple topics characterized by specific distributions over words [5]. In this study, these genetically informed signatures of shared disease are termed topics. Our research builds upon the foundational principles of dynamic topic models ([6]) and hierarchical Dirichlet processes [7]. However, we introduce a fundamental paradigm shift by a) extending the dimensionality of both topic weights and disease loadings to the individual level, and b) dynamically updating both topics and loadings with accumulated information. Our methodological innovation is motivated mainly by three empirical trends described next.

1. The likelihood of disease manifestation within a defined comorbidity profile, referred to here as a topic, demonstrates temporal variability. Different diseases under the same comorbidity topic or underlying pathological process may have varying ages of onset. For example, coronary artery disease may appear at an intermediate age, whereas heart failure may manifest at more advanced ages, indicating the progression of the underlying disease process. Each disease within a given topic has a specific distribution that indicates its occurrence propensity, and a temporal parameter that denotes its rate of change over time (Figure 1).

2. The variability in disease onset is substantial and is not adequately represented by traditional modeling methodologies. We observe that, even within a certain topic or profile, both the age of onset and the progression rate of diseases differ based on an individual's genetic composition and other contributory factors (Figure 2). Instead of associating these factors

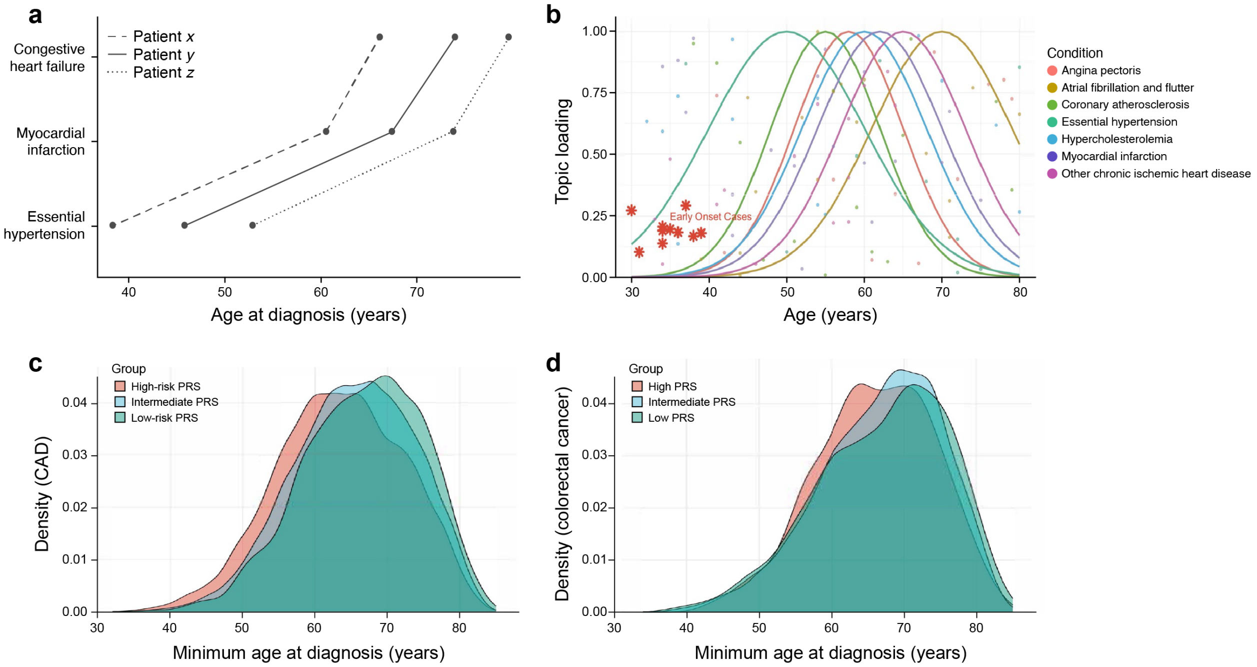

Figure 2: **Individualized timing within population-dictated disease chronology**. *We demonstrate individuals heavily weighted on the cardiovascular topic with similarly ordered but chronologically distinct ages of onset for common cardiovascular conditions. We demonstrate the distribution in age of onset for common conditions, and how early onset individuals are poorly captured. We demonstrate justification for the use of genetics in estimating the warped times.*

post hoc, we incorporate them concurrently.

3. The heterogeneity of profiles contributing to an individual's disease etiology exhibits temporal variability across the lifespan. These fluctuations are observed in alignment with both population-wide and individual-specific trends. Individual-specific trends are influenced by a complex interplay of genetic and non-genetic factors.

Jointly modeling both observed diagnoses and underlying genetics according to these unique time-dependent processes constitutes a novel approach. The integration of innovative diagnostic data to refine the prevalence of underlying genetic themes shows significant potential to improve predictive accuracy and enable novel discoveries. In this article, we first outline a novel framework for describing the evolution of genetically informed disease topics. We discuss the modeling of these comorbidity profiles over time, the incorporation of individual- and population-level trends, the dynamic adjustment of individual time scales within a given topic, and the updating of an individual's profile over time. We introduce a paradigm shift in disease modeling in the following ways:

1. **Individual-Level Time Warping**: Accounting for differences in disease onset and progression speeds via a genetics-driven time warping function.

2. **Genetic Integration**: Polygenic risk scores (PRS) directly influence initial disease risk and a genetic predilection parameter that determines an individual's trajectory adaptability, reflecting underlying genetic influence.

3. **Dynamic Topic Weights**: Bayesian updating of topic weights with each new diagnosis ensures our predictions always reflect the latest health information.

2.2 Connections With Topic Modeling in Natural Language Processing

Within the field of natural language processing, a 'topic' [8] is defined as a pattern of semantically related terms that frequently co-occur within a corpus of documents. For instance, a 'sports' topic may include terms such as 'football,' 'basketball,' and 'soccer,' whereas an 'education' topic might be characterized by terms such as 'class,' 'campus,' 'teacher,' and 'student.' In a similar vein, this study conceptualizes an individual's diagnostic history as document text, where a 'topic' signifies a cluster of interrelated diseases that commonly co-occur within patient histories. For example, a cardiovascular topic is exemplified by a high prevalence of Myocardial Infarction and hypertension diagnoses, supplemented by additional associations with hypertension. Conversely, an endocrine topic may be primarily typified by Diabetes Mellitus and Thyroid disorders. Despite these distinctions, both topics exhibit shared associations with hypertension and hyperlipidemia, potentially due to differing etiological factors. The evolution and composition of each topic are not always evident and can be enhanced through unsupervised learning. However, traditional topic-modeling methodologies fail to adequately address the dynamic progression of diseases within a patient's medical history. For instance, coronary artery disease may manifest at an intermediate age, while heart failure occurs predominantly at more advanced ages, reflecting the progression of the underlying pathological processes. This progression can vary in speed due to both genomic

D	Number of possible diagnoses, indexed by d
M	Number of individual patients, indexed by i
N_i	Total number of diagnoses per patient
T	Number of possible time points, indexed by t
w_{idt}	Observed diagnosis indicator for disease d patient i
K	Number of postulated topics (comorbidity profiles) indexed by k
z_{idt}	Latent (i.e. unobserved) index of the topic for diagnosis d in patient i at time t

Table 1: **Notation for observed data and latent disease topics.**

and non-genomic factors. This study introduces an age-dependent topic modeling framework to capture the varying onset of diseases throughout life. While traditional topic modeling delineates the population-level comorbidity profile, the likelihood of developing specific diseases and their onset can exhibit substantial variability among individuals. This study incorporates both genetic and non-genetic determinants to tailor the disease risk trajectory for individual prediction and pattern discovery, both within and among topics. Genetic factors enter the model in two primary ways: First, an individual with a high genetic predisposition for a particular topic is more likely to demonstrate that topic, although their age-dependent incidence function follows population-level patterns. That is, if cardiovascular disease is uncommon in the population at a young age, an individual with a high predisposition to cardiovascular disease may have a higher than average weighting on this topic, even though their weight of the overarching topic is allocated to alternative topics. Second, the rate of progression of a disease conditional on the membership of a topic can vary by genetic class within a topic.

3 Generative Model

3.1 Population-Level Topic Vocabularies Over Diseases

We first define a model for all diseases invariant to chronic or acute conditions in which a diagnosis may reoccur. In our model, each topic k has an associated vocabulary distribution over diseases. This vocabulary distribution evolves over time and is modeled using a Gaussian Process (GP).

Specifically, for each topic k and each disease d, we define the parameters η_{kdt} that describe the log-odds of disease d occurring within topic k at time t.

The evolution of η_{kdt} over time t is given by the Gaussian Process:

$$\eta_{kdt} \sim \mathcal{GP}(\mu_{kd}(t), \Sigma_\eta) \tag{1}$$

where $\mu_{kd}(t)$ is the mean function (Figure S6), and Σ_η is the covariance matrix that captures the correlation over time. The mean function $\mu_{kd}(t)$ can take various forms, such as linear trends, logistic growth, exponential decay, Gaussian peaks, polynomial trends, or sinusoidal patterns, depending on the expected behavior of the disease d within the topic k.

Given that most diseases exhibit peak activity within a limited number of topics (sparse data rows), we enforce the restriction that diseases remain active only in a small number of topics (Figure S2). This constraint ensures identifiability [8].

The covariance matrix Σ_η is constructed using a covariance function, typically a squared exponential kernel, which ensures that the probability of the disease changes smoothly over time so that the topic-specific probability of a disease is more closely correlated with nearby times.

To map from the log-odds scale to the probability scale, we apply the softmax function to η_{kdt}:

$$\beta_{kdt} = \frac{\exp(\eta_{kdt})}{\sum_{d'=1}^{D} \exp(\eta_{kd't})},$$

where β_{kdt} represents the probability of disease d within the topic k at time t.

3.2 Population-Level Topic Weights

The population-level topic weights describe how the prevalence of each topic k changes over time at the population level. These weights are represented by the parameters α_{kt} for each topic k at time t. Similarly to the disease vocabularies, the topic weights are also modeled using a Gaussian process:

$$\alpha_{kt} \sim \mathcal{GP}(\mu_k(t), \Sigma_\alpha), \tag{2}$$

where $\mu_k(t)$ is the mean function that describes the expected prevalence of topic k over time, and Σ_α is the covariance matrix for topic weights.

3.3 Genetic Novelty: Warping for Topic-specific Disease Probabilities

A critical feature of our model is that genetics influences disease progression in two ways because we observe variation in age of disease onset for a disease even within a given topic. Secondly, we wish to improve our estimation of individual topic predilection by considering genomic predilection to disease topics. There are two critical parameters allowing for the joint incorporation of genetics into our model:

- Time warping parameters ρ's
- Genetic predilection parameters γ's.

Individual genetic factors influence how each person's disease probabilities within topics evolve over time. This is modeled through the genetic warping parameter $\rho_{i,k}$ for each individual i and topic k. The warping function $t' = W(t, \rho_{i,k})$ transforms the original time index t into a new time index t', effectively modulating the speed at which individual i progresses through the disease probabilities of topic k. For instance, for an individual with a genetic warping index of 2 for topic k, the disease probabilities which are supposed to occur at age 60 occur at age 36, and those at age 50 occur at age 25. Critically, this is still taken from the population-level topic probabilities, so there is learning across the population.

3.4 Disease Probability within Topics: Genetics-dependent Warping of Time

For each individual, as above, the warping coefficient ρ_{ik} governs the time scale at which the disease activity operates. Recall that we wish to map the chronologic age for every individual to the warped time, such that the appropriate population level probability values at the relevant warped time are applied to every individual's chronologic time, so that at age 25, we would retrieve the population from age 50 for example.

For each individual i, topic k, and time t:

$$t'_{ik} = \max(1, \tau(t, \rho_{ik}, T))$$

where:

$$\tau(t, \rho_{ik}, T) = \left(\frac{t}{T}\right)^{\frac{1}{\rho_{ik}}} \times T$$

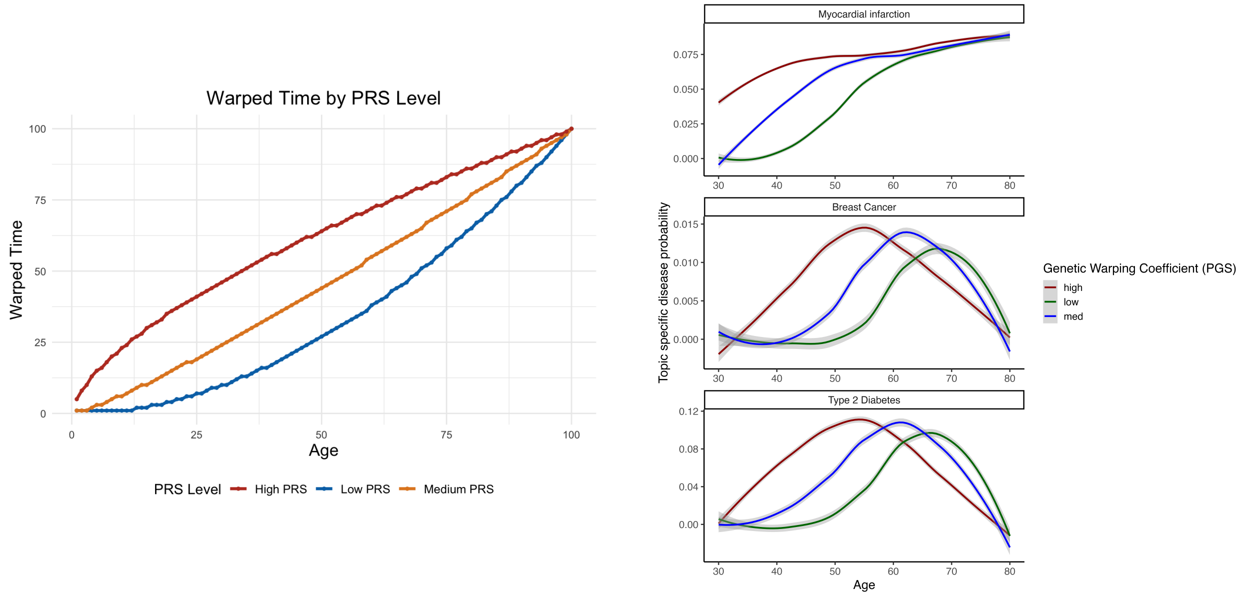

Figure 3: **Warping**. *Here we demonstrate how an individual of low genetic risk will have disease probabilities ascertained from younger ages at present time, and those of high genetic risk will have probabilities of later times pushed forward. We demonstrate this behavior from our model fit on three chosen diseases.*

- T is the maximum time (e.g., age 100).

- ρ_{ik} is the warping coefficient for individual i and topic k, influenced by the genetic scores:

$$\rho_{ik} = \mathbf{w}^T \mathbf{g}_i.$$

  This is a positive number.

- $\mathbf{w}$ is a vector of weights to be learned for each genetic score.

Calendar time t will translate into warped time t' thus retrieving the appropriate population-level value. We use this to determine the index of the probability which is drawn for a given individual.

$$\beta'_{kdt}(i) = \beta_{kdt'_{ik}}$$

This allows for individual-specific variations in the progression of disease probabilities within each topic by pulling the probability from the population-level at the warped time index.

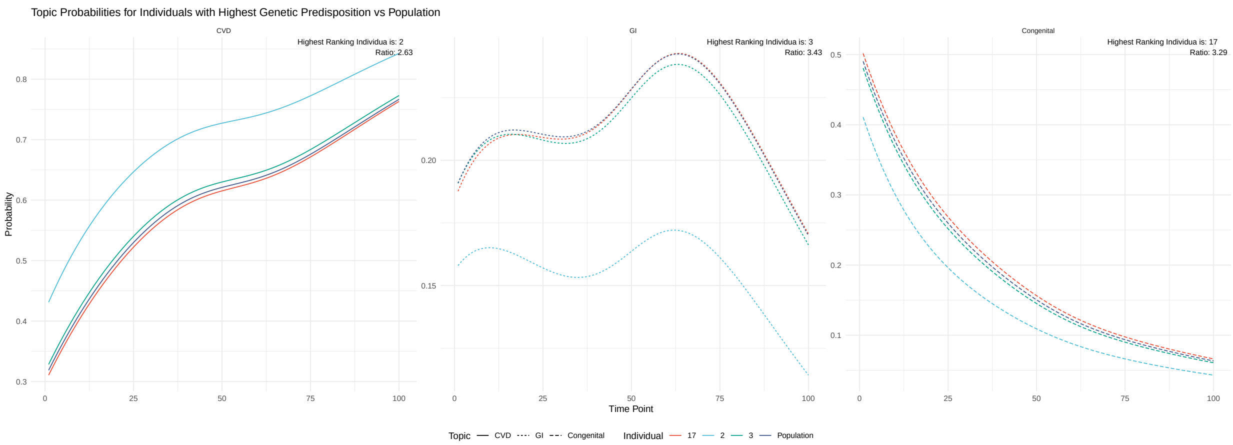

Figure 4: **Gaussian Process for Topic Weights influenced by population and individual level trajectory**. *Here we demonstrate the population level trajectory and the individual predilection for a given topic.*

3.5 Genetic Predilection for Topic Weights

Genetics also influences an individual's predilection for certain topics, which is represented by an adjustment to the population-level mean function of the topic weights. Each individual i has a genetic predilection score γ_{ik} for topic k. This score adjusts the population-level mean function $\mu_k(t)$ to create an individual-specific mean function:

$$\mu_{ik}(t) = \mu_k(t) + \gamma_{ik} \tag{3}$$

Individual-specific topic weights α_{ikt} are then sampled from a Gaussian process with this adjusted mean function:

$$\alpha_{ikt} \sim \mathcal{GP}(\mu_{ik}(t), \Sigma_\alpha). \tag{4}$$

To obtain valid probability distributions for topic proportions at each time point, we apply the softmax function to individual-specific topic weights α_{ikt}. This ensures that the topic proportions sum to one at each time point:

$$\theta_{ikt} = \frac{\exp(\alpha_{ikt})}{\sum_{k'=1}^{K} \exp(\alpha_{ikt})}$$

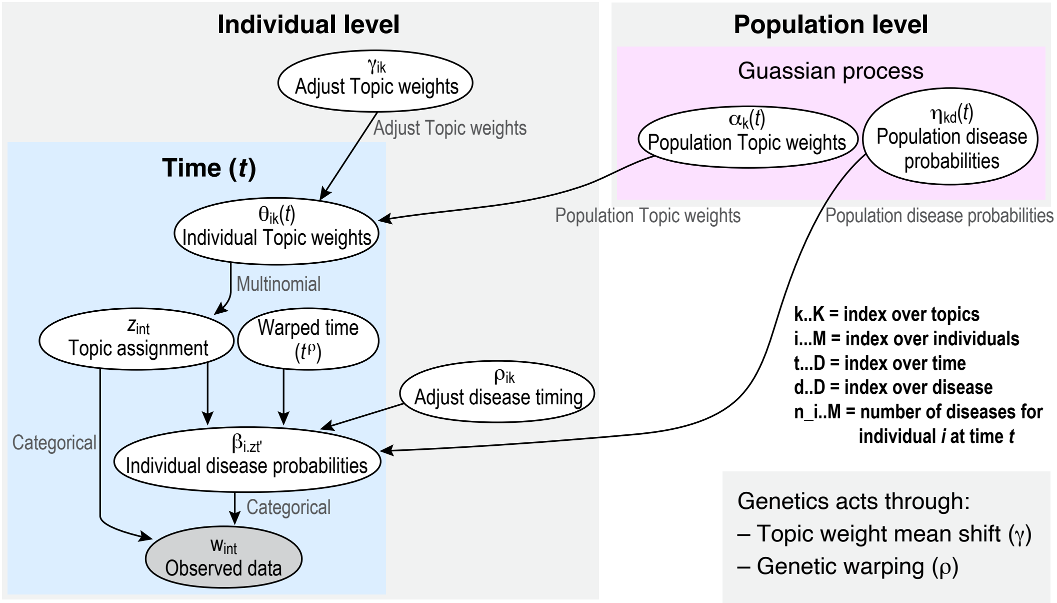
Individual level
Population level
Guassian process
γ_{ik} Adjust Topic weights
Adjust Topic weights
$\alpha_k(t)$ Population Topic weights
$\eta_{kd}(t)$ Population disease probabilities
Time (t)
$\theta_{ik}(t)$ Individual Topic weights
Population Topic weights
Population disease probabilities
Multinomial
z_{int} Topic assignment
Warped time (t^ρ)
ρ_{ik} Adjust disease timing
Categorical
$\beta_{i.zt'}$ Individual disease probabilities
Categorical
w_{int} Observed data
k..K = index over topics
i...M = index over individuals
t...D = index over time
d..D = index over disease
n_i..M = number of diseases for individual *i* at time *t*
Genetics acts through:
– Topic weight mean shift (γ)
– Genetic warping (ρ)

Figure 5: **Plate Diagram Modifying Flow**. *Here we illustrate the flow of our generating model. In brief, our process is governed by two population level parameters, the evolution of topic weights (α) and disease loadings (η). Both arise from topic or disease specific Gaussian process with topic or disease level mean and covariance parameters. The covariance pattern arises from a Gaussian kernel in which adjacent times share probability more closely. Genetics acts in two fundamental ways to influence individual changes: The topic weights for every individual are driven by both the population mean and the individual predilection for a given topic, γ_{ik}. The scale at which an individual progresses through a topic specific disease is governed by a 'warping coefficient', ρ_{ik}. Both topic weights and disease loadings are softmax normalized at each stage to generate a proper probability distribution. Then, at each time point, a latent indicator is chosen according to an individual's time-specific probability over topics, θ_{ik}. This latent indicator then designates the vector of time specific disease probabilities $\beta_{ikt'}$, which will govern the choice of observed diagnosis. The β_{ikt} is generated according to the topic's population-specific vector of disease probabilities for the time governed by the individual's warping parameter.*

3.6 Summary of Generating Model

In summary, the hierarchical structure for the generative process for each individual i over time t can be summarized as follows. We also offer a modified plate diagram describing the evolution (Figure 5).

1. For each topic k and each disease d, the disease vocabulary evolves over time according to a Gaussian Process (GP):

$$\eta_{kd} \sim \text{GP}(\mu_{kd}, \Sigma_{\eta}^2)$$

2. The topic proportions also evolve over time via a GP:

$$\alpha_k \sim \text{GP}(\mu_k, \Sigma_{\alpha}^2)$$

3. For each individual i, the GP from which the topic proportions α_{ik} are drawn varies depending on the genetic 'predilection' γ_{ik}:

$$\mu_{ik} = \mu_k + \gamma_{ik}$$

$$\alpha_{ik} \sim \text{GP}(\mu_{ik}, \Sigma_{\alpha_k}^2)$$

4. The disease vocabulary is adjusted for each individual's 'warped' time scale ρ_{ik}:

$$\eta'_{kdt}(t)(\rho_{ik}) = \eta_{kdt'_{ik}}$$

5. The probability of disease d within topic k at individual-specific 'warped' time t'^{ρ_i} is given by the softmax function over the natural parameter η:

$$\beta'_{kdt}(i) = \frac{\exp(\eta_{kdt'_{ik}})}{\sum_{d'} \exp(\eta_{kdt'_{ik}})}$$

6. Similarly, the topic proportions for individual i at time t are given by the softmax function over α:

$$\theta_{ikt} = \frac{\exp(\alpha_{ikt})}{\sum_{k'} \exp(\alpha_{ik't})}$$

7. For each diagnosis n for individual i at time t, the topic is chosen according to the multinomial distribution based on θ:

$$z_{int} \sim \text{Mult}(\theta_{i,\cdot,t})$$

8. The observed disease w for each diagnosis n is given by:

$$w_{int} | z_{int}, \beta_{\cdot,t} \sim \text{Categorical}(\beta_{kdt'_{ik}})$$

3.7 Simulating Diagnoses

The process of simulating diagnoses for each individual i at each time point t involves several steps:

1. **Simulate Diagnoses Based on Time**: For each time point t from 1 to T, we calculate the expected number of diagnoses λ_{it} using a Poisson distribution scaled by time in which we expect a greater number of diagnoses at later time points:

   $$\lambda_{it} = \frac{\chi \cdot t}{T}$$

   We then sample the number of diagnoses N_i from a Poisson distribution with mean λ_{it}.

2. **Sample Diagnoses and Topics**: For each diagnosis n from 1 to N_i:

   - Sample a topic k from the individual-specific topic proportions θ_{ikt}.
   - Use the time-warping function to adjust the time index for the sampled topic.
   - Sample a disease d from the topic-specific disease probabilities $\beta_{kdt'}$ at the warped time index.

4 Likelihood

4.1 Overview

Our model is trained on data consisting of observed diagnoses by individual and time. The likelihood of the observed data can be derived by considering the probability of observing the diagnoses given the latent variables, which include the topic weights (θ), topic loadings (β), genetic warping parameters (ρ), and genetic predilection parameters (γ). The likelihood is the joint probability of the observed diagnoses given the latent variables. For each individual i, time t, and diagnosis n for M individuals:

$$\mathcal{L} = \prod_{i=1}^{M} \prod_{t=1}^{T} \prod_{n=1}^{N_i(t)} P(w_{int} \mid \theta_{ikt}, \beta_{kdt}, \rho_{i,k}, \gamma_{ik}).$$

Observed Variables	
w_{int}	Observed diagnosis for individual i at diagnosis n and time t
Latent Variables	
θ_{ikt}	Individual-specific topic proportions at time t
β_{kdt}	Topic-specific disease probabilities at time t
$\rho_{i,k}$	Genetic warping parameters for individual i and topic k
γ_{ik}	Genetic predilection parameters for individual i and topic k
z_{int}	Latent topic assignment for each diagnosis w_{int}

Table 2: Observed and Latent Variables

4.2 Likelihood Function

Breaking it down further, we can express the likelihood as follows.

4.2.1 Topic Assignment

The probability of assigning a topic z_{int} given the topic proportions θ_{ikt}:

$$P(z_{int} = k \mid \theta_{ikt}, \gamma_{ik}) = \theta_{ikt}$$

4.2.2 Genetic Warping (ρ)

The warping parameter $\rho_{i,k}$ affects the time index used in the disease probabilities β. We can express this as:

$$P(w_{int}|z_{int} = k, \beta_{kdt}, \rho_{i,k}) = \beta_{kd(W(t,\rho_{i,k}))} \tag{5}$$

Where $W(t, \rho_{i,k})$ is the warping function that transforms the original time t based on the individual's genetic warping parameter.

4.2.3 Genetic Predilection (γ)

The genetic predilection parameter γ_{ik} influences the individual-specific topic proportions θ_{ikt}. We can incorporate this into the likelihood as:

$$P(z_{int} = k|\theta_{ikt}, \gamma_{ik}) = \theta_{ikt} = \text{softmax}(\alpha_{ikt} + \gamma_{ik}) \tag{6}$$

Where α_{ikt} is the population-level topic weight and γ_{ik} is the individual's genetic predilection for topic k.

4.2.4 Diagnosis Probability

Incorporating both genetic parameters, we can express the full likelihood for an individual i as:

$$\mathcal{L}_i = \prod_t \prod_n \left(\sum_k \text{softmax}(\alpha_{ikt} + \gamma_{ik}) \cdot \beta_{kd(W(t,\rho_{i,k}))} \right) \tag{7}$$

This formulation shows how both the topic selection and disease probabilities are influenced by the individual's genetic parameters.

The probability of observing a diagnosis w_{int} given the topic assignment z_{int} and the topic-specific disease probabilities β_{kdt}. This involves the genetic warping function to obtain the appropriate β at the warped time index:

$$P(w_{int} \mid z_{int} = k, \beta_{kdt}, \rho_{i,k}) = \beta_{kd_{W(t,\rho_{i,k})}}.$$

Putting it all together, the joint likelihood for an individual i can be written as:

$$\mathcal{L}_i = \prod_{t=1}^{T} \prod_{n=1}^{N_i(t)} \left(\sum_{k=1}^{K} \theta_{ikt} \beta_{kd_{W(t,\rho_{i,k})}} \right),$$

and the overall likelihood for the entire dataset is then the product over all individuals:

$$\mathcal{L} = \prod_{i=1}^{M} \mathcal{L}_i = \prod_{i=1}^{M} \prod_{t=1}^{T} \prod_{n=1}^{N_i(t)} \left(\sum_{k=1}^{K} \theta_{ikt} \beta_{kd_{W(t,\rho_{i,k})}} \right).$$

5 Posterior Updates for Bayesian Inference

The model specification just described can be used to implement Bayesian inference and evaluate posterior distributions of all the latent variables and unknown parameters. As this is computationally very challenging for high-dimensional real-life EHR data, we implement here a practical approximation that leverages estimates developed in previous work to estimate population parameters, and allows us to focus on novel aspect of time dependencies and warping.

5.1 Updating individual topic probabilities

Considering individual patients, an important step clinically is the evaluation of the posterior probability that patient i's diagnosis w_{int} at time t is manifesting as a result of the action of topic k, based on their previous history of diagnoses. Using Bayes' rule, this posterior probability is updated as follows:

$$P(z_{int} = k \mid w_{int}, \theta_{previous}, \gamma_d, \beta_d) \sim \beta_{kdt} \cdot P(z_{int} = k\theta_{previous}, \gamma_d).$$

where $\theta_{previous} = (\theta_{i\cdot 1}, \ldots, \theta_{i\cdot t-1})$. Also we assume that w_{int} is the sole diagnosis at time t for patient i. Multiple diagnosis updates follow the same logic, not spelled out here for brevity. This allows us to use the new diagnoses at any point in time to compute dynamic updates of an individuals' disease probability.

5.2 Parameter Estimation

In practice, we apply a novel algorithm to update the posterior distribution over topic weights for a given set of diagnoses:

1. **Initialization of $\theta_{ik}(t)$**:

   We first initialize the estimated $\theta_{ik}(t)$ at time $t = 1 : 3$ by drawing from a Dirichlet distribution:

   $$\theta_{ik}(1 : 3) \sim \text{Dirichlet}(\alpha, K)$$

   where α is chosen to be greater than 1 to ensure relatively uniform values across the K topics.

2. **Predicted GP Mean**:

   The predicted GP mean, μ_{ki}, is initialized using the estimated $\theta_{ik}(t)$ values at time $t = 1 : 3$. The population predicted mean is also set to this value.

3. **Gaussian Process Fit from time point 4 to T**:

   A Gaussian Process (GP) is fit using the estimated $\theta_{ik}(t)$ at time points $t = 1 : 3$ to predict $\theta_{ik}(t)$ at time $t = 4$, for both the individual and the population.

4. **Subsequent Time Points**:

   For each subsequent time point t, a Gaussian Process is fit to the estimated $\theta_{ik}(t)$ at time points $t = 1 : (t-1)$. This fit is treated as the prior prediction for $\theta_{ik}(t)$. The GP is fit for both the population (i.e., GP on $E(\theta_{ik}(1 : T-1))$) and the individual (i.e., GP on $\theta_{ik}(1 : T-1)$).

5. **Adjusted GP Mean**:

   The adjusted GP mean at time t is obtained by a weighted average of the population GP mean and the individual GP mean.

6. **Likelihood of New Diagnoses**:

   For patients with new diagnoses at time t, the likelihood of the diagnoses is calculated as:

$$P(z_t|w) = \prod_d P(\beta_{kd}|Z = k)$$

   where β_{kd} represents the disease-specific parameters given the topic k.

7. **Combination of Prior Prediction and Likelihood**:

   The prior prediction is combined with the likelihood by adding the log of the adjusted GP mean and log-likelihood, then exponentiating the result to obtain the final estimate for $\theta_{ik}(t)$:

$$\text{estimate}_{\theta_{ik}(t)} = \exp .$$

Additional Parameters in detailed methods:

5.3 Algorithm for Posterior Updates for θ

Algorithm 1 Posterior Updates for θ Given β and Observed Diagnoses

```
Input: Previous GP predictions of $\theta$, genetic warping parameters $\rho_{i,k}$, genetic predilection parameters $\gamma_{ik}$, observed diagnoses $D$, disease probabilities $\beta$
Output: Updated $\theta$
for each time point $t$ do
    Update population GP means $\mu_{kt}$
    $\mu_{kt} \leftarrow \text{fit\_gp}(1 : (t-1), \text{mean}(\mathit{estimated\_Theta}[:, :, 1 : (t-1)]), \sigma)$
    for each individual $i$ do
        Update individual GP means $\mu_{ikt}$
        $\mu_{ikt} \leftarrow \text{fit\_gp}(1 : (t-1), \mathit{estimated\_Theta}[i, :, 1 : (t-1)], \sigma)$
        Compute individual weight $w_i$
        $\mathit{adjusted\_gp\_mean}[i, :, t] \leftarrow w_i \cdot \mu_{ikt} + (1 - w_i) \cdot \mu_{kt}$
        Retrieve current diagnoses data for individual $i$
        $\mathit{diagnosis\_indices} \leftarrow \text{map\_diagnoses}(\mathit{diagnoses\_codes}, \mathit{list\_of\_diseases})$
        Compute likelihood
        $\mathit{likelihood} \leftarrow \prod(\beta[W(t, \rho_{i,k}), \mathit{diagnosis\_indices}])$
        Update log proportions
        $\mathit{updated\_log\_proportions} \leftarrow \log(\mathit{likelihood}) + \log(\mathit{adjusted\_gp\_mean}[i, :, t])$
        Normalize
        $\mathit{updated\_proportions} \leftarrow \exp(\mathit{updated\_log\_proportions}) / \sum(\exp(\mathit{updated\_log\_proportions}))$
        $\mathit{estimated\_Theta}[i, :, t] \leftarrow \mathit{updated\_proportions}$
    end for
end for
```

5.4 Algorithm for Posterior Updates for β Given θ

Algorithm 2 Posterior Updates for β Given θ

```
Input: Fixed $\theta$, genetic warping parameters $\rho_{i,k}$, observed diagnoses $D$
Output: Updated $\beta$
Initialize diagnosis_counts and count_array with zeros
for each individual $i$ do
    for each topic $k$ do
        for each time point $t$ do
            Compute original time index og_time = $W(t, \rho_{i,k})$
            Retrieve diagnoses at time $t$ for individual $i$
            Update diagnosis_counts and count_array for each diagnosis
        end for
    end for
end for
Compute $\beta$ as the normalized counts
for each topic $k$ do
    for each time point $t$ do
        Normalize $\beta_{kdt}$ using the softmax function
    end for
end for
```

5.4.1 Data

Our primary analyses are in the UK Biobank and AllofUs data sets.

UK Biobank The UK Biobank has collected detailed health and genetic data from approximately 500,000 participants aged 40 to 69 years, recruited between 2006 and 2010. Here, we assemble EHR data for 421,707 participants with at least one EHR diagnosis recorded between the ages of 28 and 81 from 1981 forward [9], [10]. The HESIN EHR data includes coded clinical events, consultations, diagnoses, procedures, and laboratory tests, using coding systems such as READ2, CTV-3, BNF and DM + D. Furthermore, hospital inpatient data, which covers admissions, diagnoses, procedures, and discharge information, is available for the entire cohort and is coded using ICD-9, ICD-10, OPCS-3, and OPCS-4. The UK Biobank also links to national death and cancer registries to provide comprehensive health outcomes data.

All of Us The All of Us Research Program is a diverse biomedical dataset from the United States. Currently, more than 460,000 participants have consented to share their electronic health records (EHR), with approximately 55% of these records already integrated into the program data set [11]. We used data from 239,200 people who contributed both EHR and genomic information [12]. The EHR data in All of Us include information from various health domains such as conditions, procedures, drugs, and measurements, which are standardized using vocabularies such as SNOMED, LOINC, and ICD codes. Here we use ICD10 codes harmonized to the 349 codes used in the UK Biobank.

6 Results

6.1 Model Assessment: Updated Patient Weights

In this section, we evaluate our model fit in comparison to a time-fixed weight approach, which estimates weights only once based on available information. We first demonstrate the comparison between the true weights of the topic and the estimated weights at the population level.

We demonstrate the improvement in predicted theta using our approach to a fixed weight (θ) approach.

6.2 Performance Metrics

To evaluate the performance of our predictive model, we calculate the following metrics: accuracy, precision, recall, and F1-score. These metrics are defined as follows:

Accuracy Accuracy is the ratio of correctly predicted diagnoses to the total number of true diagnoses. It is given by:

$$\text{Accuracy} = \frac{\text{True Positives}}{\text{Total True Diagnoses}}$$

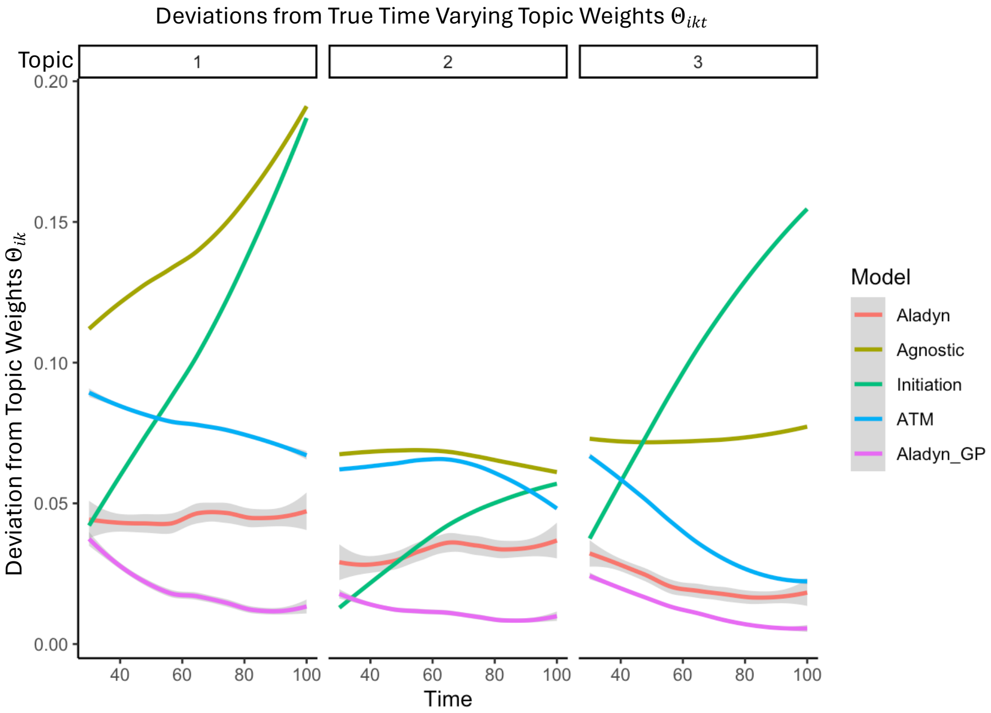

Figure 6: **Delta**. Over 100 time points, we demonstrate the deviations between an approach that uses a flat estimate of θ_{ik} as 1/K (agnostic), an approach that uses the initiation value of θ_{ik} (initiation), an approach 'ATM' that uses the average θ_{ik} over all time points ([1]) and two versions of Aladyn.

Precision Precision is the ratio of correctly predicted diagnoses to the total number of predicted diagnoses. It is given by:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted diagnoses to the total number of actual true diagnoses. It is given by:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score The F1-Score is the harmonic mean of precision and recall. It provides a balance between precision and recall and is given by:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In simulations, we find that the accuracy, precision, and recall are superior with Aladyn as opposed to a fixed weight approach when comparing to true topic identifiers and using simulated diagnoses:

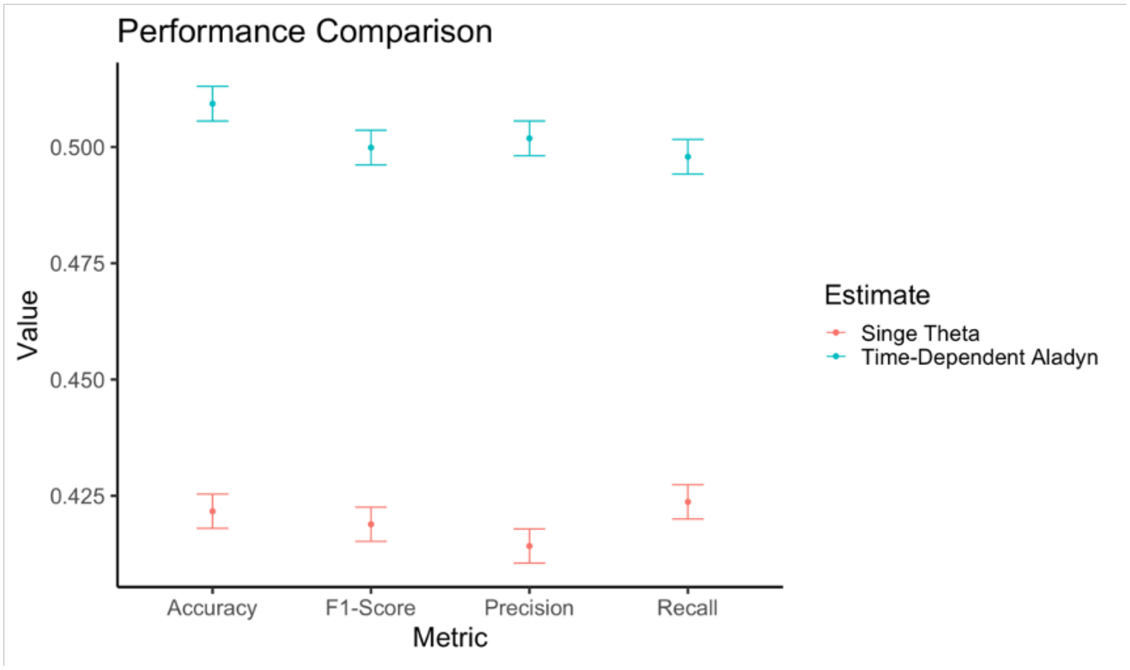

Figure 7: **Performance Assessment**. We demonstrate the Accuracy, Precision, Recall and F1 score of simulations using Aladyn and a fixed weight approach. Terms defined in text.

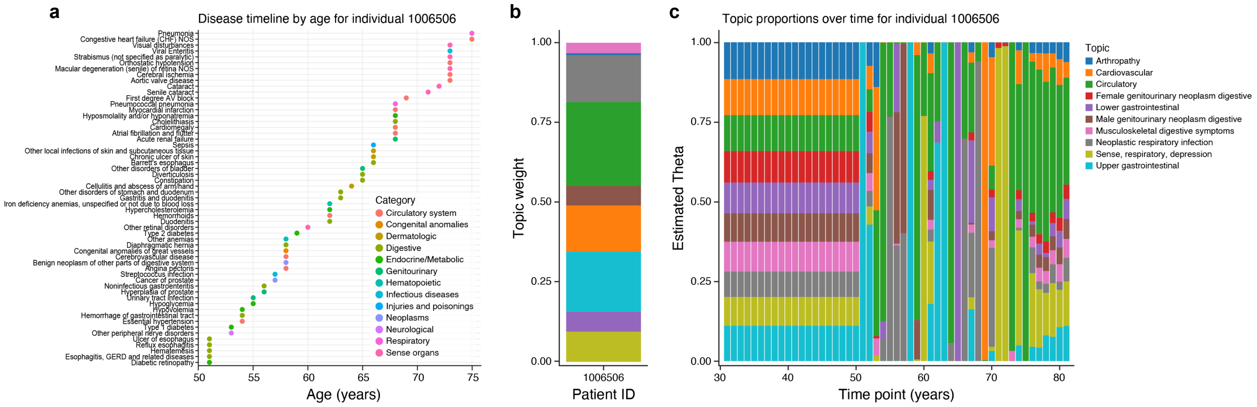

Figure 8: **Time Varying Weights**. *Here we demonstrate the actual age of diagnoses for a sample patient in the UK Biobank. We demonstrate the fit using Aladyn, and using a traditional fixed weight approach according to ATM ([1]).*

6.3 Disease Loadings and Time Varying Weights

We demonstrate the clarity of estimated β for known topics **??**. In brief, as in the algorithm above, we map all simulated disease counts to the unwarped time for an individual, and weight by the expected topic time varying contribution of each individual. We note that this differs fundamentally from existing approaches which only estimate weight once.

Subsequently, we evaluated the results derived from the application of time-varying weights to actual diagnoses (Figure 8). Several key distinctions are observed from existing methodologies. In particular, Aladyn exhibits the proficiency to identify population-level trajectories even amidst sparse population-level data. With the advent of new diagnoses, Aladyn refines the previously estimated weights by incorporating these new diagnoses with the topics that most effectively enhance the likelihood of the specific diagnoses. The ’memory’ of prior diagnoses is sustained through the Gaussian process.

Furthermore, we can see that 65.4% of the population has a shift of more than 10% and 50% of the time, the shift is greater than 11.5% (Figure S7).

6.4 Biological Meaning of Topic Weights

Our findings substantiate the identified warping, as temporal variations in weight reveal that individuals within the top and bottom 20% of the polygenic risk exhibit earlier (later) predictions of

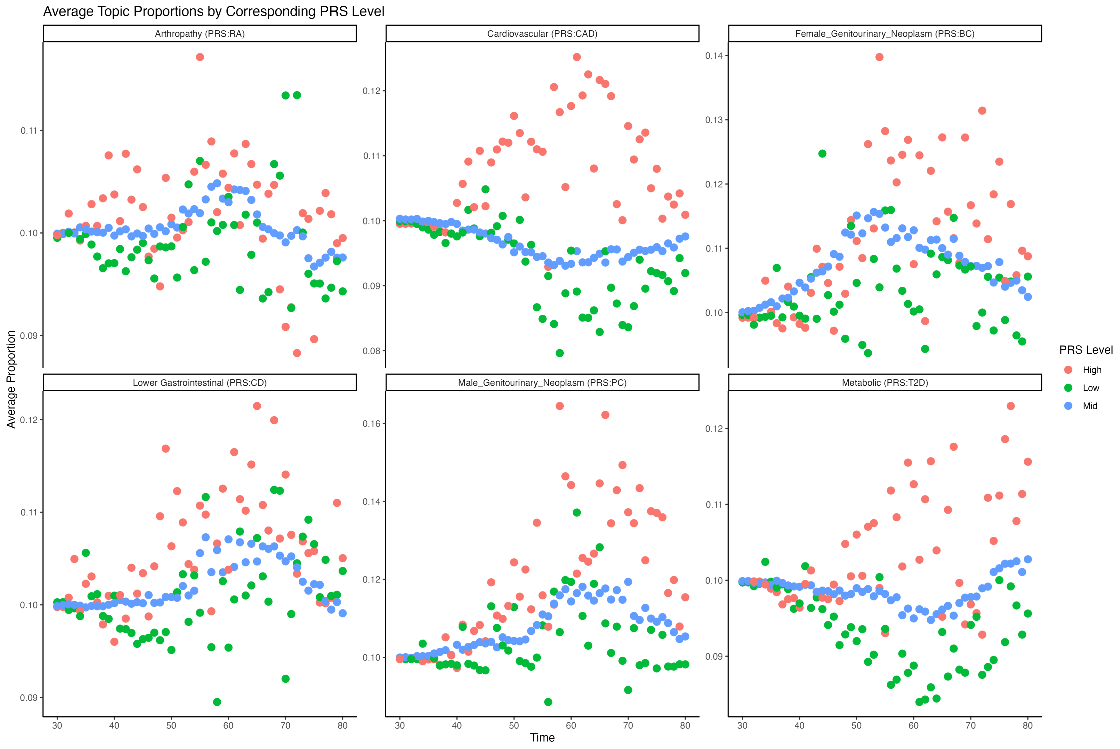

Figure 9: **Marginal Probability**. *Here we show that the marginal probability of disease is shifted for those with varying levels of genetic risk.*

disease onset under our model. Furthermore, we show that individuals with true high genetic risk are highly enriched in topic weights for the expected topics when we use a cardinal polygenic risk score (PRS) as the target PRS from which to assess enrichment.

6.5 Improved Accuracy and Flexibility

Here we show the variation in topic weights when compared with a model that uses time-fixed weights. We first ask, for each condition diagnosed in real data, what proportion of the time did our model Aladyn assign to the diagnosis a time-weighted marginal probability greater than 10%, how often did the competing ATM (age dependent topic model, [1]) do so. Furthermore, we compare the nominal percentage assigned for real diagnoses.

To facilitate exploration of a dynamic time-varying model's, we have developed and interactive web application available at `https://surbut.shinyapps.io/forapp/`. This tool allows users to visualize the important diseases for each topic in the UK Biobank, All of US `https://surbut.`

Model	Accuracy	Proportion Correct	Delta
ATM	0.01	0.21	2.60
Aladyn	0.15	0.79	NA

Table 3: **Comparison of Model Accuracy and Probability of True Disease** Proportion Correct is the proportion of Time Model Assigns True Disease Higher Probability. Delta is the difference in probability assigned by Aladyn vs ATM

shinyapps.io/aouapp/ and Mass General Brigham Biobank https://surbut.shinyapps.io/mgb_topic/.

Furthermore, we have created an app at https://surbut.shinyapps.io/dynamic_ehr/ that allows users to simulate warping, time-varying weights, and unique trajectories in real-time, providing an intuitive understanding of the Aladyn model's functionality.

7 Limitations and Future Work

While the Aladyn model demonstrates promising outcomes, it is crucial to recognize its current limitations and potential areas for future enhancement. Model Implementation: The present iteration of Aladyn utilizes estimated loadings derived from extant dynamic topic models (i.e., [1, 6]). Individualized risk assessment is feasible by using externally estimated parameters and functions pertinent to the general population, with a focus on modeling individual variation. Nevertheless, further research would be beneficial to fully implement the model's proprietary Bayesian learning of these loadings. The incorporation of multiple Gaussian processes and Bayesian updates within the model necessitates significant computational resources. Optimizing these algorithms for large-scale datasets remains an ongoing challenge. Although the data demonstrate superior performance relative to an existing age-dependent topic model, additional validation is recommended against a broader spectrum of state-of-the-art methodologies. Despite utilizing data from both the UK Biobank and All of Us, further evaluation on diverse populations is required to ensure the model's wide-ranging applicability. First, there is a need to develop more efficient computational methods to handle larger datasets. This reflects on the computational challenges associated with the model and aims at improving its performance and scalability. Second, the plan includes conducting comprehensive comparisons with other leading disease progression models. This involves evaluating the performance of the Aladyn model against other models to identify strengths and weaknesses. In addition, there will be investigations into how well the model performs across a broader range of populations and disease types, ensuring its applicability and robustness in various scenarios. These efforts are aimed at addressing current limitations and expanding the model's usability and accuracy.

8 Discussion

This modeling paradigm integrates population trends with individual genetic variations, elucidating commonalities and divergences in disease progression. The described generative model is governed by two population-level parameters: the evolution of topic weights (α) and disease loadings (η). These parameters are inferred from topic- or disease-specific Gaussian processes characterized by their respective mean and covariance structures. The covariance pattern is modulated by a Gaussian kernel, ensuring that temporally adjacent points exhibit higher correlation. Crucially, genetic factors modulate individual variations through topic weights, influenced by both the population mean and individual predisposition (γ_{ik}), and the progression scale of a topic-specific disease, regulated by a 'warping coefficient' (ρ_{ik}). Topic weights and disease loadings undergo softmax normalization to ensure they form a valid probability distribution. At each time point, a latent indicator is selected based on an individual's time-specific topic probabilities (θ_{ik}), which subsequently determines the vector of time-specific disease probabilities ($\beta_{ikt'}$) that dictates the observed diagnostic outcome. The diagnostic outcome is generated in accordance with the topic's population-specific vector of disease probabilities, modulated by the individual's warping parameter. By leveraging Gaussian processes and Bayesian updates, the model provides dynamic, personalized disease predictions. We demonstrate that the integration of genetic data with hierarchical models facilitates the amalgamation of population-level learning with individual-level prediction, thereby enhancing predictive accuracy and enabling the discovery of novel topics. We provide the results of the implementation on the dynamic topic model loadings ([6] in the UKB in supplementary figures 17-26).

The ethical implications of using genetic data for disease prediction are significant and warrant careful consideration. While Aladyn offers powerful predictive capabilities, it's crucial to ensure that its implementation respects patient privacy, avoids genetic discrimination, and considers the psychological impact of early disease predictions. Future work should include collaborations with bioethicists to develop guidelines for the responsible use of such models in clinical settings. To promote transparency and facilitate further research, we have made the simulation code for Aladyn available on GitHub `https://surbut.github.io/dynamic_ehr`. We encourage the scientific community to explore, validate, and build upon our work.

References and Notes

[1] X. Jiang, *et al.*, Age-dependent topic modeling of comorbidities in UK Biobank identifies disease subtypes with differential genetic risk **55** (11), 1854–1865, doi:10.1038/s41588-023-01522-8, https://www.nature.com/articles/s41588-023-01522-8.

[2] H. Estiri, J. G. Klann, S. N. Murphy, A clustering approach for detecting implausible observation values in electronic health records data. *BMC Medical Informatics and Decision Making* **19** (1), 142 (2019), doi:10.1186/s12911-019-0852-6, https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0852-6.

[3] S. M. Urbut, *et al.*, Dynamic Importance of Genomic and Clinical Risk for Coronary Artery Disease Over the Life Course. *medRxiv* (2023), doi:10.1101/2023.11.03.23298055, https://www.medrxiv.org/content/early/2023/11/04/2023.11.03.23298055.

[4] Y. Wang, *et al.*, Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *Journal of Biomedical Informatics* **102**, 103364 (2020), doi:https://doi.org/10.1016/j.jbi.2019.103364, https://www.sciencedirect.com/science/article/pii/S1532046419302849.

[5] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.* **3** (null), 993–1022 (2003).

[6] D. M. Blei, J. D. Lafferty, Dynamic topic models, in *Proceedings of the 23rd international conference on Machine learning - ICML '06* (ACM Press, Pittsburgh, Pennsylvania) (2006), pp. 113–120, doi:10.1145/1143844.1143859, http://portal.acm.org/citation.cfm?doid=1143844.1143859.

[7] L. Ren, D. B. Dunson, L. Carin, The dynamic hierarchical Dirichlet process, in *Proceedings of the 25th international conference on machine learning* (2008), pp. 824–831.

[8] D. Blei, L. Carin, D. Dunson, Probabilistic Topic Models. *IEEE signal processing magazine* **27** (6), 55–65 (2010), doi:10.1109/MSP.2010.938079, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4122269/.

[9] C. Sudlow, *et al.*, UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12** (3), 1–10 (2015), doi: 10.1371/journal.pmed.1001779, `https://doi.org/10.1371/journal.pmed.1001779`.

[10] C. Bycroft, *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562** (7726), 203–209 (2018), doi:10.1038/s41586-018-0579-z, `https://www.nature.com/articles/s41586-018-0579-z`.

[11] null null, The "All of Us" Research Program. *New England Journal of Medicine* **381** (7), 668–676 (2019), doi:10.1056/NEJMsr1809937, `https://www.nejm.org/doi/full/10.1056/NEJMsr1809937`.

[12] The All of Us Research Program Genomics Investigators, *et al.*, Genomic data in the All of Us Research Program. *Nature* **627** (8003), 340–346 (2024), doi:10.1038/s41586-023-06957-x, `https://www.nature.com/articles/s41586-023-06957-x`.

Funding Statement

S.M.U. is supported by T32HG010464 from the National Human Genome Research Institute. A.G. is supported by National Institutes of Health (NIH) grant nos R01CA227237, R01CA244569 and R21HG010748, and awards from the Claudia Adams Barr Foundation, Louis B. Mayer Foundation, Doris Duke Charitable Foundation, Emerson Collective and Phi Beta Psi Sorority. P.N. is supported by grants R01HL1427, R01HL148565, and R01HL148050 from the National Heart, Lung, and Blood Institute, and grant 1U01HG011719 from the National Human Genome Research Institute.

Data Availability Statement

All simulations and a didactic tutorial on our model can be found at: `https://bookdown.org/sarahmurbut/dynamic_ehr/`. An interactive application to simulate warping, time varying weights and unique trajectories in real time can be found at: `https://surbut.shinyapps.io/dynamic_ehr/`.

Ethical Standards

The research meets all ethical guidelines, including adherence to the legal requirements of the study country according to the Institutional Review Board of Massachusetts General Hospital 2021P00228.

Author Contributions Conceptualization: S.U., A.G., G.P.; Methodology: S.U., A.G., G.P..; Data curation: S.U.; Data visualization: S.U.; Writing original draft: S.U., G.P., P.N. All authors approved the final submitted draft.

Supplementary Materials for

Aladyn Individual: Bayesian Hierarchical Dynamic Genetic Modeling of Comorbidity Progression

This PDF file includes:

Materials and Methods

8.1 Updating Genetic Parameters

The genetic parameters ρ and γ are updated using Bayesian inference as follows:

8.1.1 Updating ρ

The posterior distribution of ρ 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 ρ, and $P(\rho)$ is the prior distribution of ρ. We can use a Metropolis-Hastings algorithm to sample from this posterior:

Algorithm 3 Metropolis-Hastings for updating ρ

```
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 γ

Similarly, for γ:

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

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 ρ and γ 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

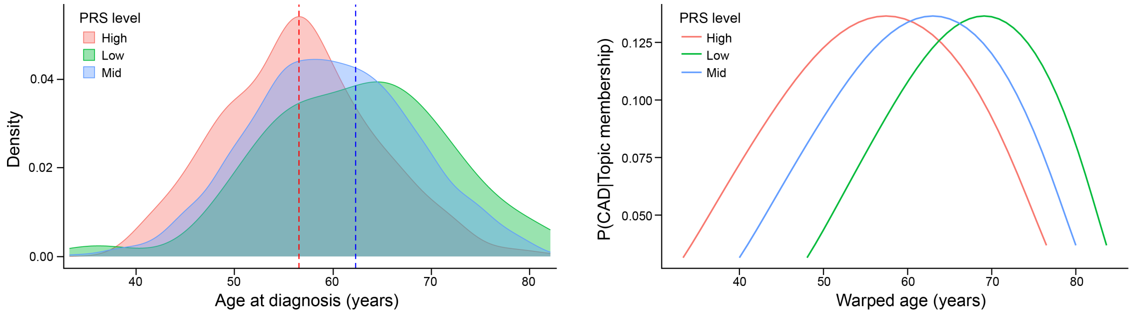

Figure S1: **Age of Onset Captured by Warped Disease Probabilities**. Age of diagnosis for coronary artery disease, tracking closely with underlying genetic risk. We overlay the warped time predicted disease probability of those in the top and bottom deciles of polygenic risk for coronary artery disease.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

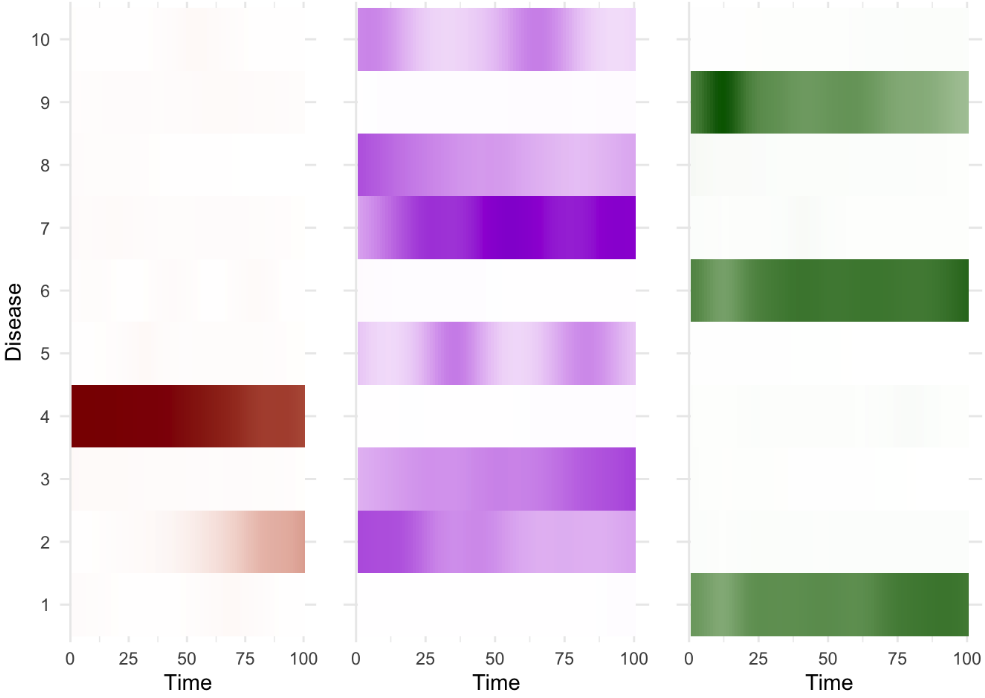

Figure S2: **Topic Specificity**. In both real data and simulations, we recognize that diseases tend to be sparse in the number of topics on which they are loaded.

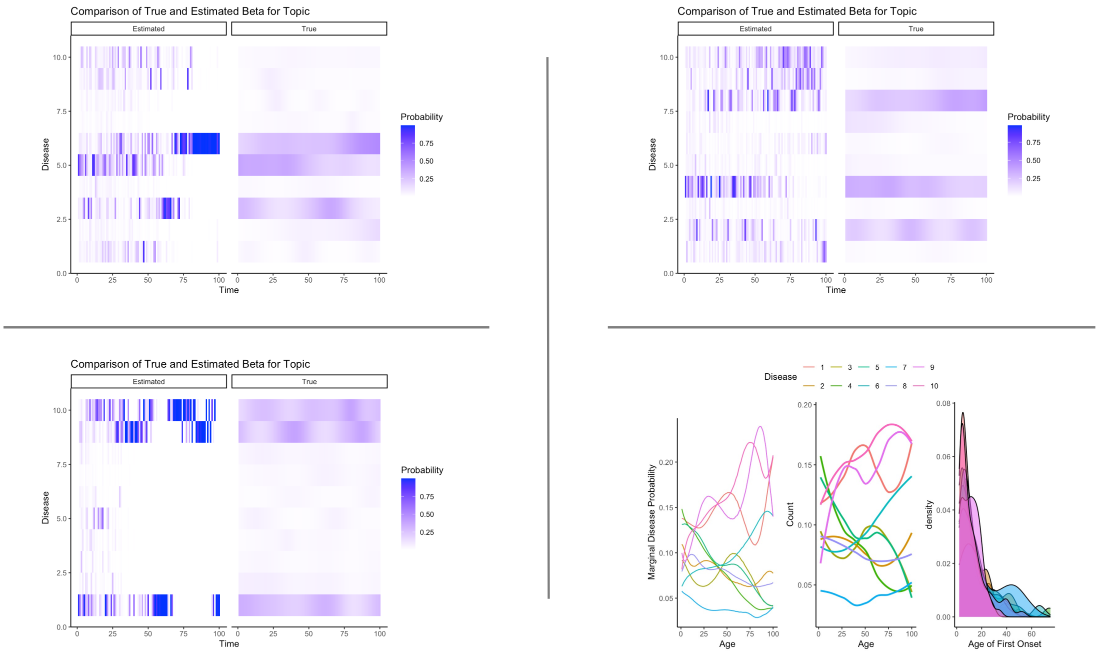

Figure S3: **Estimating Disease Loadings**. Here we demonstrate the approach for estimating disease loadings using counts of disease occurrences mapped to unwarped times. Lower right, we demonstrate the marginal probabilities of each disease, the estimated counts, and the first occurrence. Average Marginal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

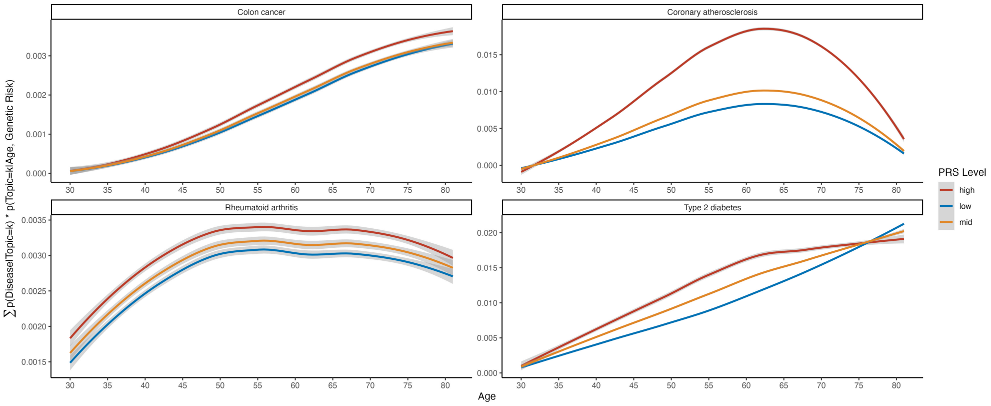

Figure S4: **Genetically Enriched Individuals Show Earlier Onset Disease**. Here we show that the marginal probability of disease is earlier for those with high genetic risk. We use a canonical PRS for each topic.

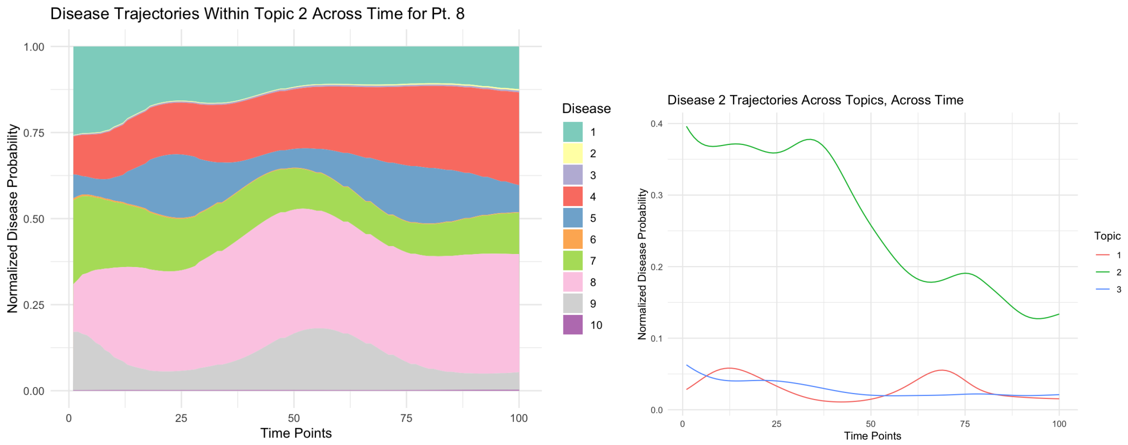

Figure S5: **Topic-specific Disease Probability**. We plot the probabilities of all 10 simulated diseases over time within a given topic, conditional on the time scale of a chosen patient. In B, we demonstrate the trajectory of one chosen disease across all topics. This is simulated data in which the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

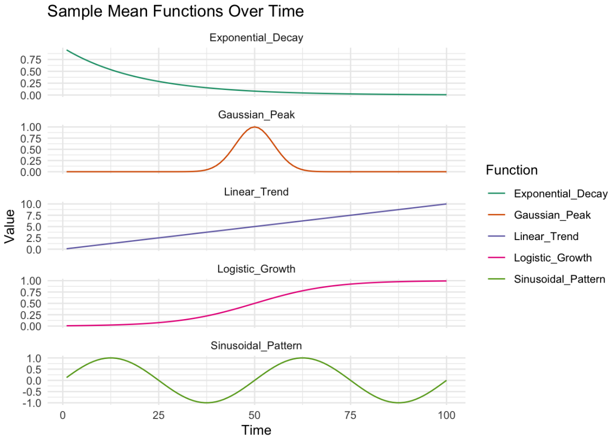

Figure S6: **Sample Mean**. Here we demonstrate several mean functions that govern the process of disease evolution. These are meant to reflect a sampling of biological processes and are learned from the model.

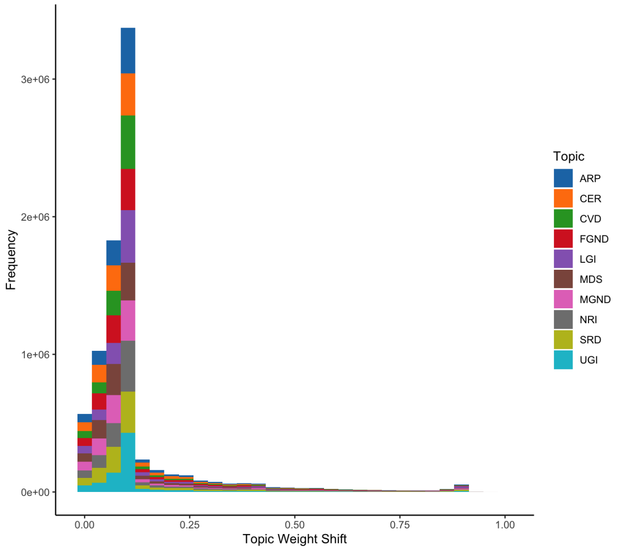

Figure S7: **Change in Topic Weights Over Time**. Here we consider the difference in estimated topic weight under a model with time-varying topic weights in comparison to the time-fixed approach ([1]). Results by topic do not reveal systematic differences. NRI = Neoplastic Respiratory, CVD = Cardiovascular,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

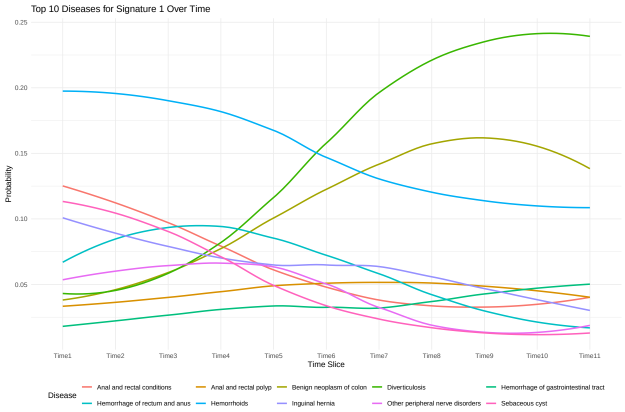

Figure S8: Top 10 diseases for topic 1 in the UK Biobank

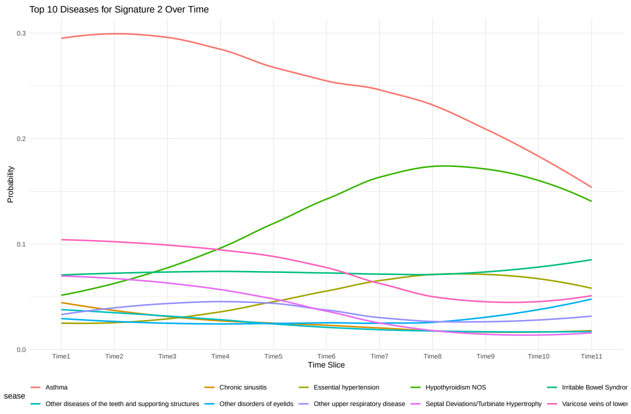

Figure S9: Top 10 diseases for topic 2 in the UK Biobank

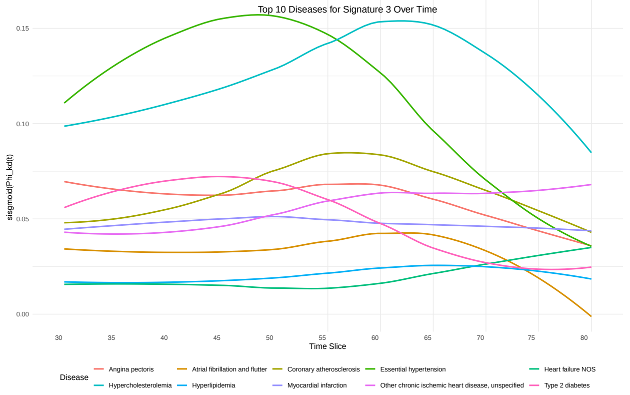

Figure S10: Top 10 diseases for topic 3 in the UK Biobank

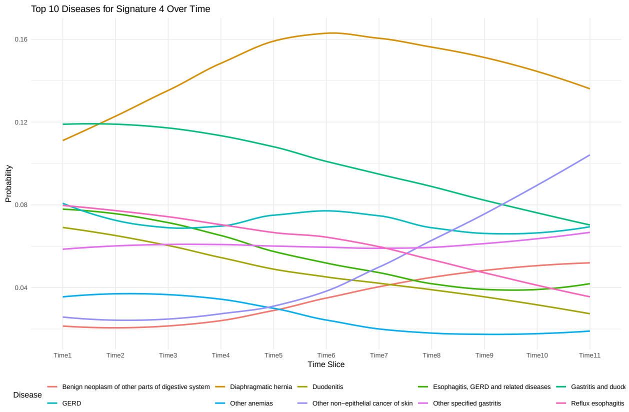

Figure S11: Top 10 diseases for topic 4 in the UK Biobank

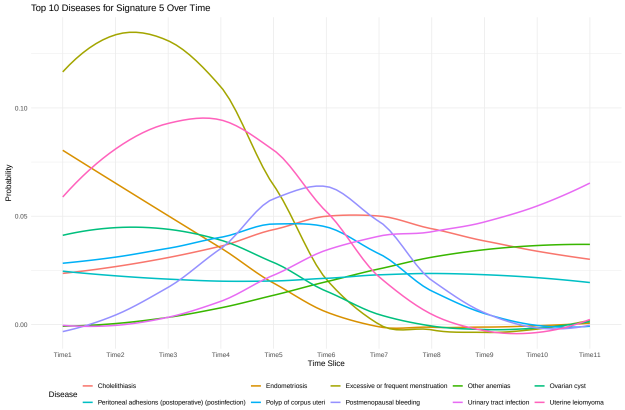

Figure S12: Top 10 diseases for topic 5 in the UK Biobank

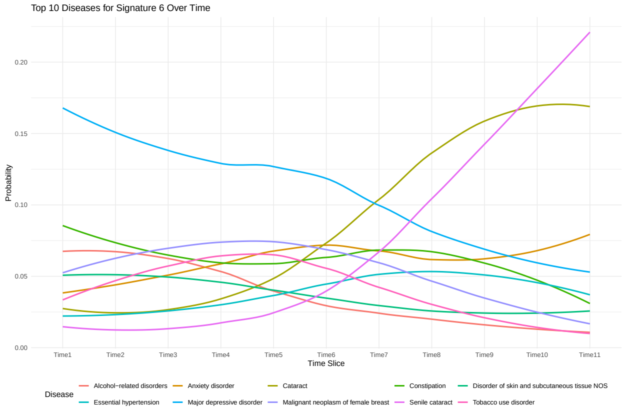

Figure S13: Top 10 diseases for topic 6 in the UK Biobank

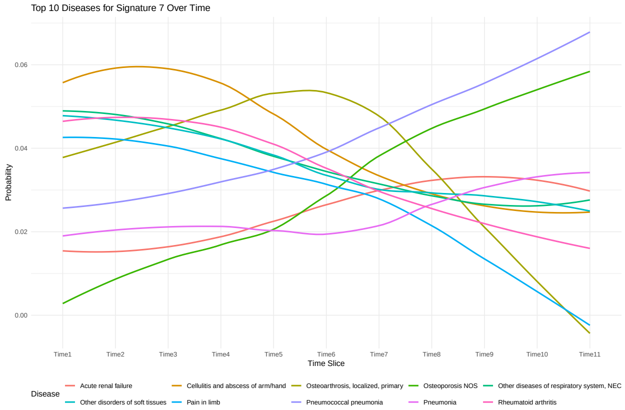

Figure S14: Top 10 diseases for topic 7 in the UK Biobank

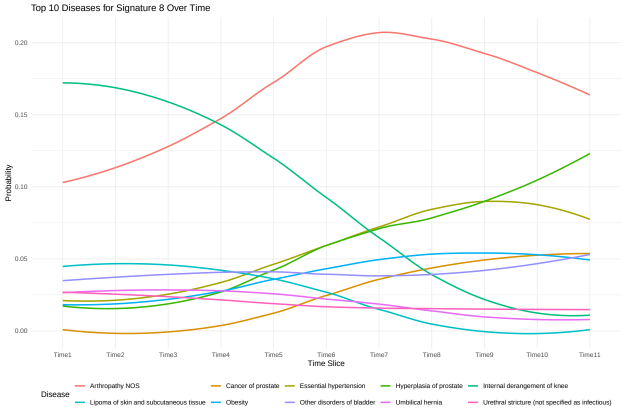

Figure S15: Top 10 diseases for topic 8 in the UK Biobank

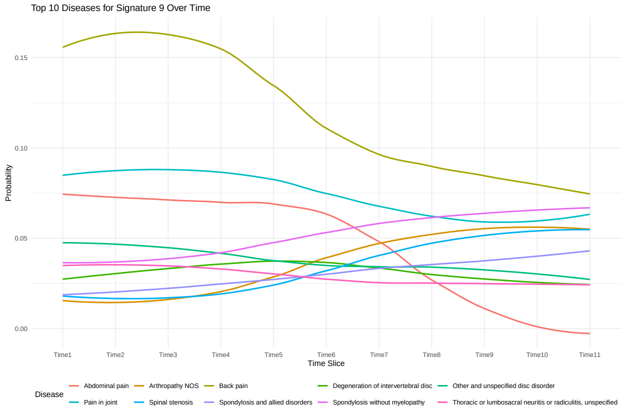

Figure S16: Top 10 diseases for topic 9 in the UK Biobank

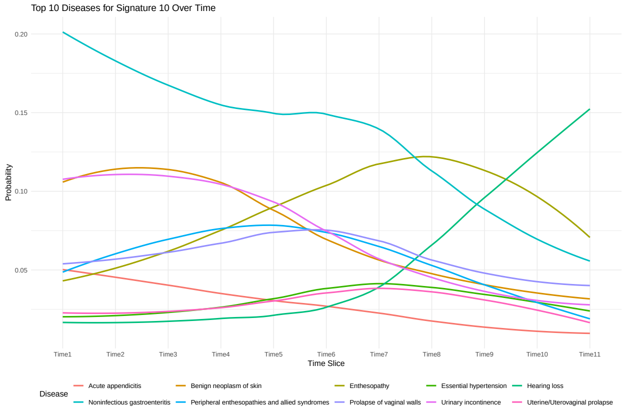

Figure S17: Top 10 diseases for topic 10 in the UK Biobank