Automated generation of personalized trajectories of aging phenotypes with DyViA-GAN

Saumyadipta Pyne*[1,†], Deep Ray[2,†], and Meghana S. Ray[1,3]

[1]Health Analytics Network, Columbia, MD, USA
[2]Department of Mathematics, University of Maryland, College Park, MD, USA
[3]Heedlab, Columbia, MD, USA

April 24, 2026

Abstract

With a general increase in human lifespan, the need for technological advances to develop strategies for healthy aging has assumed great importance. In the present study, our goal is to predict the progression of selected aging phenotypes in a given healthy individual as one continues aging past 65 years. Therefore, we developed a novel framework called Dynamic Views of Aging with conditional Generative Adversarial Networks (or DyViA-GAN) which is capable of predicting the plausible personalized trajectories of a selected aging phenotype conditioned on the available measurements of the phenotype at a few initial time instances, and additional covariates. Given the prevalence of osteoporosis in the aging population, we selected femoral neck Bone Mineral Density (BMD) of a healthy individual as the phenotype of interest, and baseline individual Body Mass Index (BMI) as covariate. We trained DyViA-GAN on a publicly available longitudinal dataset of a large cohort of mostly white women in the United States of age 65 years or above. Thus, it generated, for each individual, continuous phenotype trajectories, along with a corresponding region of acceptable predictions, for an age range of 66 to 89 years, for eight different combinations both with and without involving the covariate. The prediction results were subjected to rigorous quality-control and multiple comparative analyses. Our results clearly demonstrate the potential of generative deep learning frameworks in healthspan research.

Keywords: GAN, Deep Learning, BMD, trajectory prediction, aging phenotype, osteoporosis

[†]These authors contributed equally to this paper.

*Corresponding author, Email: spyne@healthanalytics.net

1 Introduction

Forecasting the trajectories of diseases has been a cornerstone of the field of epidemiology involving numerous approaches ranging from the classical Susceptible-Infectious-Removed (SIR) and related models to more recent ones based on time series and predictive analytics, genomics, etc. [42]. In fact, modeling and prediction of infectious diseases have been among the key drivers of methodological advances in biostatistics and public health over the past century [44, 45]. Since the COVID-19 pandemic, novel approaches have emerged for outbreak forecasting including artificial intelligence (AI)-driven deep learning of spatio-temporal patterns, prediction by real and synthetic data fusion, e.g., [57, 18]. However, with a general increase in human lifespan over the past decades, some researchers in the field of disease prediction have now shifted their attention from the modeling of infectious disease outbreaks in a population to that of different chronic conditions and their progression in individuals in the form of personalized trajectories. It has led to such advances as deep aging clocks and various means of biological age assessment, e.g., [52, 16, 27, 48, 43, 41].

In the present study, our goal is to predict the progression of selected aging phenotypes in a given healthy individual as one continues aging past 65 years. Initial measurements of such phenotypes may be taken annually. With data from these initial time points as input, in combination with her other static or dynamic risk factors as covariates, we aim to build a reliable generative AI model that is capable of forecasting future values of the same phenotypes for this individual. Thus, our model will produce a set of plausible trajectories (univariate or multivariate depending on the selection of phenotypes), which could provide timely insights to inform and aid clinical decision-making that might extend her *healthspan*. For example, trajectories of aging phenotypes associated with one's cognition or bone health can help one strategically build preparedness or adopt preventive measures against any future onset of dementia or possible fractures.

In particular, we focus on osteoporosis, which accounts for approximately 1.5 million fractures just in the U.S. each year [10]. According to its National Institute of Aging (NIA)*, osteoporosis is a "silent disease" since it does not have any markers of progression until a bone breaks – usually in the hip, spine. or wrist. In severe osteoporosis cases, a simple movement such as a cough or minor bump can result in a broken bone, i.e., a fracture. Additionally, individuals with osteoporosis may also have a longer duration of recovery and may experience chronic pain. In older adults, hip and spine fractures, in particular, can have serious consequences such as loss of mobility and independence. In a single typical year (2019), there were 318,797 emergency department visits, 290,130 hospitalizations, and 7,731 deaths related to hip fractures alone among adults of age 65 years or older in the U.S. [37]. Thus, osteoporosis acts as a key determinant of one's (bone) healthspan over the course of aging.

Osteoporosis is defined as a systemic skeletal disease that has both the following characteristics: low bone mass and microarchitectural deterioration of bone tissues [39]. In clinical practice, osteoporosis is generally diagnosed when the measure of bone mineral density (BMD) of an individual is below the mean BMD of a reference population by at least 2.5 times the standard deviation, i.e., T-score ≤ -2.5 [30]. Low BMD or osteopenia is determined by a T-score between -1 and -2.5. As individuals age, BMD decreases and, consequently, osteoporosis becomes more prevalent in older adults [21, 46, 31]. A study of BMD of the femoral neck or lumbar spine estimated 43.4 and 10.2 million U.S. adults of age 50 years and above as having low bone mass and osteoporosis respectively.[55] While several factors may be involved in the etiology of

*https://www.nia.nih.gov/health/osteoporosis/osteoporosis

osteoporosis, historically, the body mass index (BMI) has been linked to bone health as a protective factor [2, 46]. Paganini-Hill et al. using data from 8600 postmenopausal women reported that high BMI was associated with a significant reduction in hip fracture risk independently of other potential confounders [40]. Their findings were later corroborated by other studies [11, 14].

The longest running cohort study in the U.S. on this topic is the National Institute of Aging (NIA) sponsored Study of Osteoporotic Fractures (SOF), a multi-site study that ran from 1986 to 2017 and recruited over 10,000 participants at 4 different sites. Initially the study enrolled white women, and from 1997, it began to recruit African American women. Exclusion criteria included bilateral hip prostheses or inability to walk without assistance. Study participants attended a series of clinical visits with physical and mental evaluations including imaging to measure bone density, physical examinations, and questionnaires. In addition to clinical visits, participants were contacted via telephone or mail every four months to complete follow-up assessments for information on any falls, fractures, and vital status [1].

In a seminal paper in 1995, the SOF researchers established the dual energy X-ray absorptiometry (DEXA) scan as a non-invasive, painless, safe, and accurate measurement of BMD [11]. Subsequently, the U.S. Preventive Services Task Force osteoporosis management guidelines recommended routine BMD screening DEXA scans for women 65 years of age or older [38, 29]. The SOF study also showed that women who fell and broke their hip were five times more likely to die in their first year of fracture than those who did not break their hip [1]. Notably, beyond screening, DEXA scans can also serve to monitor osteoporosis progression, leading to opportunities for intervention and prevention of fractures. Therefore, femoral neck BMD data, as measured by DEXA, is an ideal initial choice of an aging phenotype to build a model for automated generation of personalized trajectories using a suitable computational framework.

Deep learning frameworks, such as Long Short-Term Memory Networks (LSTMs), Gated Recurrent Units (GRUs) and transformers networks, have been successful in applications involving time series forecasting [20, 50, 58]. The training and forecasting accuracy of these models typically usually rely on access to training data recorded over many time-points. Many of the previous studies that applied deep learning for producing disease trajectories used one or more recurrent neural networks (RNN) for predicting the ICD code for an individual's next visit to the clinic [8, 9, 47]. Some have used generative adversarial networks (GANs) to predict such codes for multiple subsequent visits but not for tracking any actual phenotype of an individual [49]. Other deep learning applications have focused on specific health conditions or diseases, e.g., cystic fibrosis [32], Alzheimer's disease [17]. Measures of DNA methylation biomarkers that estimate epigenetic aging serve a somewhat distinct purpose [24, 4]. Yet other platforms such as the Danish Disease Trajectory Browser (DTB) use statistical analysis of population-scale medical data to produce disease trajectories but these are not personalized predictions [51]. Various applications of generative AI and deep neural networks in aging research are described in reviews [54, 59] and the references therein.

To address such issues, we work with (a) individual longitudinal measurements of the selected phenotype that are available (b) only at a few (typically, 3-4) time-points. The aforementioned deep learning models are not appropriate to predict the phenotype trajectories under such conditions. We instead appeal to deep generative models which are suitable for solving probabilistic problems. Among existing generative frameworks, conditional Generative Adversarial Networks (cGANs) are popular in learning the underlying probability distribution of data conditioned on known values of other associated variables or parameters [35]. cGANs have been successfully used in studies of medical images such as for modeling the dynamics of cardiac

aging [6], and Alzheimer's disease progression based on MRI scans [26].

In this study, we proposed a novel cGAN framework called Dynamic Views of Aging with GAN (or DyViA-GAN) which is capable of predicting the plausible personalized trajectories of a selected aging phenotype conditioned on the available measurements of the phenotype at a few initial time instances, and additional covariates. For this purpose, we selected the BMD of the femoral neck of a healthy individual as the phenotype of interest and the baseline individual BMI as the covariate. Thus, DyViA-GAN generated personalized trajectories, for eight different combinations both with and without involving the covariate. The publicly available dataset used for training our model, due to the SOF study, is described in the Dataset section. Then we provided the details of the cGAN framework including the network architectures, and its model nomenclature for different data and covariate combinations in the Methods section. The predictions by DyViA-GAN were subjected to rigorous quality-control and multiple comparative analyses. The Results illustrate the corresponding cGAN-based generation of personalized trajectories of the selected phenotype. We end with Discussion on some of the limitations of our study and the scope for future work.

2 Methods

2.1 Dataset

We worked with an anonymized dataset for 9,704 mostly Caucasian women participants in the now concluded Study of Osteoporotic Fractures (SOF) study. The dataset is publicly available from the 'SOF Online' website (`https://sofonline.ucsf.edu/`). Attendance during the study varied as not all participant were available for all the visits, leading to significant gaps in the dataset. To ensure maximal number of participants in the training set, we considered participant data from Visit 2 (1989-90), Visit 4 (1992-94), Visit 5 (1995-96) and Visit 8 (2002-04), and considered the femoral neck BMD (denoted by FND) as measured by DEXA scans using Hologic QDR 1000 workstations. In accordance with best practices we exclude observations for patients over the age of 90 years to preserve patient anonymity [36]. We used longitudinally-adjusted scan measurements obtained by re-analyzing the previous scans and adjusting the regions-of-interest or deleting bone segments to match later scans. The T-scores of BMD were based on the parameters of a healthy reference population recommended by the International Society for Clinical Densitometry (ISCD) [33, 53].

In total, we have 2,113 de-identified participant records which were stratified based on the individual BMI values measured during Visit 1 (1986-87). We considered three strata: underweight and healthy participants (UWH) with BMI < 25, overweight participants (OW) with $25 \leq$ BMI < 30, and obese participants (OB) with BMI ≥ 30. Details of the samples and stratification can be found in Table 1. We also used the recorded age of the various participants during each of the four visits, with the ages rounded down to half years.

2.2 DyViA-GAN

The proposed DyViA-GAN to predict personalized plausible trajectories of the selected phenotype comprises two neural networks, namely the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. We describe the algorithm for the most general setting where the phenotype of interest is denoted by ρ and additional known covariates is denoted by Q. Here ρ and Q may be a combination of different data modalities, such anthropometry, images, lifestyle, cognitive functions, physical performance, and biomarkers. We assume access to a dataset from which we can extract the measured values of the desired phenotype(s) ρ at two sets of time-points, $\{t_i\}_{i\in\mathcal{I}}$ and $\{t_j\}_{j\in\mathcal{J}}$,

Table 1: Description of data (FND T-scores) used for training and testing. Stratification based on underweight and healthy participants (UWH), overweight participants (OW), and obese participants (OB).

	Strata	Number of Participants	Min. FND	Max. FND	Mean FND (± SD)
Training	UWH	800	-4.63	4.94	-2.02 (± 0.87)
	OW	720	-5.07	2.51	-1.69 (± 0.87)
	OB	360	-3.97	1.70	-1.35 (± 0.92)
Test	UWH	98	-4.28	1.58	-2.04 (± 0.94)
	OW	95	-4.03	5.42	-1.70 (± 1.10)
	OB	40	-3.09	1.42	-1.44 (± 0.96)

where $\mathcal{I}$ and $\mathcal{J}$ are two disjoint (non-overlapping) time indexing sets. The corresponding phenotype values are denoted by $\{\rho_i\}_{i\in\mathcal{I}}$ and $\{\rho_j\}_{j\in\mathcal{J}}$, respectively. Furthermore, the time-point sets $\{t_i\}_{i\in\mathcal{I}}$ and $\{t_j\}_{j\in\mathcal{J}}$ may be different and be unstructured across various records in the dataset.

The generator $\mathcal{G}$ is fed as input the vector $Y = (Q, \{\rho_i\}_{i\in\mathcal{I}}, \{t_i\}_{i\in\mathcal{I}})$, a new time-point t, and a random latent variable Z which is sampled from an uncorrelated standard Gaussian distribution of dimension N_z. The output of the generator is a predicted value of the phenotype in question at the time-point t.

We use the notations $X = \{\rho_j\}_{j\in\mathcal{J}}$ and $T_X = \{t_j\}_{j\in\mathcal{J}}$ to denote the "true" paired data associated with a given vector Y. The corresponding predictions by the generator are denoted by $\widehat{X} = \{\hat{\rho}_j\}_{j\in\mathcal{J}}$ where $\hat{\rho}_j$ is the prediction at time-point t_j. The scalar-valued discriminator network $\mathcal{D}$ is fed as input either a "true" data tuple (X, Y, T_X) or a "fake" tuple $(\widehat{X}, Y, T_X)$. The role of the $\mathcal{D}$ is to distinguish between true and fake tuples, while the role of the $\mathcal{G}$ is to generate samples indistinguishable from those drawn from the true (unknown) conditional probability distribution of X given (Y, T_X), i.e., $P_{X|(Y,T_X)}$.

The two network are trained in an adversarial fashion by solving the following optimization problem associated with a least-squares GAN [34]

$$\begin{aligned} \mathcal{D}^* &= \arg\min_{\mathcal{D}} \; \mathop{\mathbb{E}}_{P_{X,Y,T_X}} \left[\left(\mathcal{D}(X, Y, T_X) - 1 \right)^2 \right] + \mathop{\mathbb{E}}_{P_{\widehat{X},Y,T_X}} \left[\left(\mathcal{D}(\widehat{X}, Y, T_X) \right)^2 \right] \\ \mathcal{G}^* &= \arg\min_{\mathcal{G}} \; \mathop{\mathbb{E}}_{P_{\widehat{X},Y,T_X}} \left[\left(\mathcal{D}(\widehat{X}, Y, T_X) - 1 \right)^2 \right] \end{aligned} \tag{1}$$

where P_{X,Y,T_X} is the joint probability distribution of the true data, while $P_{\widehat{X},Y,T_X}$ is joint probability distribution for the same data but with $\widehat{X}$ generated by $\mathcal{G}$.

Once the DyViA-GAN is trained, the optimized generator $\mathcal{G}^*$ is used to generate an ensemble of possible values of phenotype ρ at any point t given Y for an individual. This is achieved by drawing K random samples $\{Z^{(k)}\}_{k=1}^{K}$ of the latent variable and evaluating $\mathcal{G}^*(Y, t, Z^{(k)})$ for $1 \leq k \leq K$. By allowing t to vary in the entire temporal window of interest, we can generate an ensemble of plausible continuous trajectories of phenotype for an individual. These trajectories can then be used to compute the point-wise empirical mean

and variance of the predicted trajectories (given a Y) for any t as

$$\overline{\rho}^*(t) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{G}^*(Y, t, Z^{(k)}), \quad \mathrm{Var}[\rho^*](t) = \frac{1}{K}\sum_{k=1}^{K} \left(\mathcal{G}^*(Y, t, Z^{(k)}) - \overline{\rho}^*(t)\right)^2 \tag{2}$$

A schematic of the training and evaluation of DyViA-GAN is shown in Figure 1.

For the study on the SOF dataset considered in the present work, ρ is taken to be FND, while Q is chosen as the BMI measurement taken during Visit 1. Regression of the mean FND T-score against BMI in the data analyzed by this study shows a statistically significant positive association ($\beta = 0.066$, $\mathrm{R}^2 = 0.116$, p-value=0.017). Historically, it was shown that "A high body mass index protects against femoral neck osteoporosis in healthy elderly subjects [2]." As noted in subsection 2.1, we have FND values during Visits 2,3,5 and 8. Thus the index sets $\mathcal{I}$ and $\mathcal{J}$ used to construct Y, X, T_X will be chosen as subset of $\{2, 4, 5, 8\}$ for training purposes. In particular, we set $\mathcal{I} = \{2, 4\}$, while $\mathcal{J} = \{5\}$ or $\mathcal{J} = \{5, 8\}$. We scale the BMI by 100 before feeding it to the generator, and scale the age of the participants in accordance to

$$t \to \frac{t - t_{\min}}{t_{\max} - t_{\min}}, \quad t_{\min} = 66,\ t_{\max} = 89, \tag{3}$$

where $[t_{\min}, t_{\max}]$ serves as the maximal age range (in years). These age limits were chosen based on the data available in the SOF dataset. The above scaling (3) ensures that $t \in [0, 1]$ when fed as input to the network. We also remark here that the ages $\{t_2, t_4, t_5, t_8\}$ differ across participants in the dataset, thus adding an additional layer of complexity to the trajectory prediction problem.

2.3 Network architectures

Let $|\mathcal{I}|$ and $|\mathcal{J}|$ denote the size of the index sets. For the choices made in the present work, $|\mathcal{I}| = 2$, while $|\mathcal{J}| = 1$ or 2. The generator $\mathcal{G}$ is taken to be a fully-connected network with input dimension $2|\mathcal{I}| + N_z + 1$ (or $2|\mathcal{I}| + N_z + 2$ if BMI is used as a covariate) with a scalar-valued output. We used four hidden layers of width 50 each with the hat activation function [23, 56]. To to ensure that the generator's prediction is constrained to pass through the given initial measurements corresponding to $\mathcal{I}$, i.e., at t_2 and t_4, the output of the generator is transformed as follows

$$\mathcal{G}(Y, t, Z) \to [\mathcal{G}(Y, t, Z) \times (t_4 - t) + \mathrm{FND}_4)]\,(1 - \phi(t; \sigma)) + [m \times (t - t_4) + \mathrm{FND}_4]\,\phi(t; \sigma) \tag{4}$$

where

$$m = \frac{\mathrm{FND}_4 - \mathrm{FND}_2}{t_4 - t_2}, \quad \phi(t; \sigma) = \frac{1}{1 + \exp((t - t_4)/\sigma)}.$$

The function $\phi(t; \sigma)$ in the above transformation smoothly blends the linear curve passing through (t_2, FND_2) and (t_4, FND_4) to the network output. In this work, we choose $\sigma = 10^{-2}$.

The discriminator $\mathcal{D}$ is a also taken to be fully-connected network with input dimension $2|\mathcal{I}| + 2|\mathcal{J}|$ (or $2|\mathcal{I}| + 2|\mathcal{J}| + 1$ if BMI is used as a covariate) with a scalar-valued output. We used four hidden layers of width 50 with the hat activation function. Other hyperparameter values for DyViA-GAN are listed in Table 2.

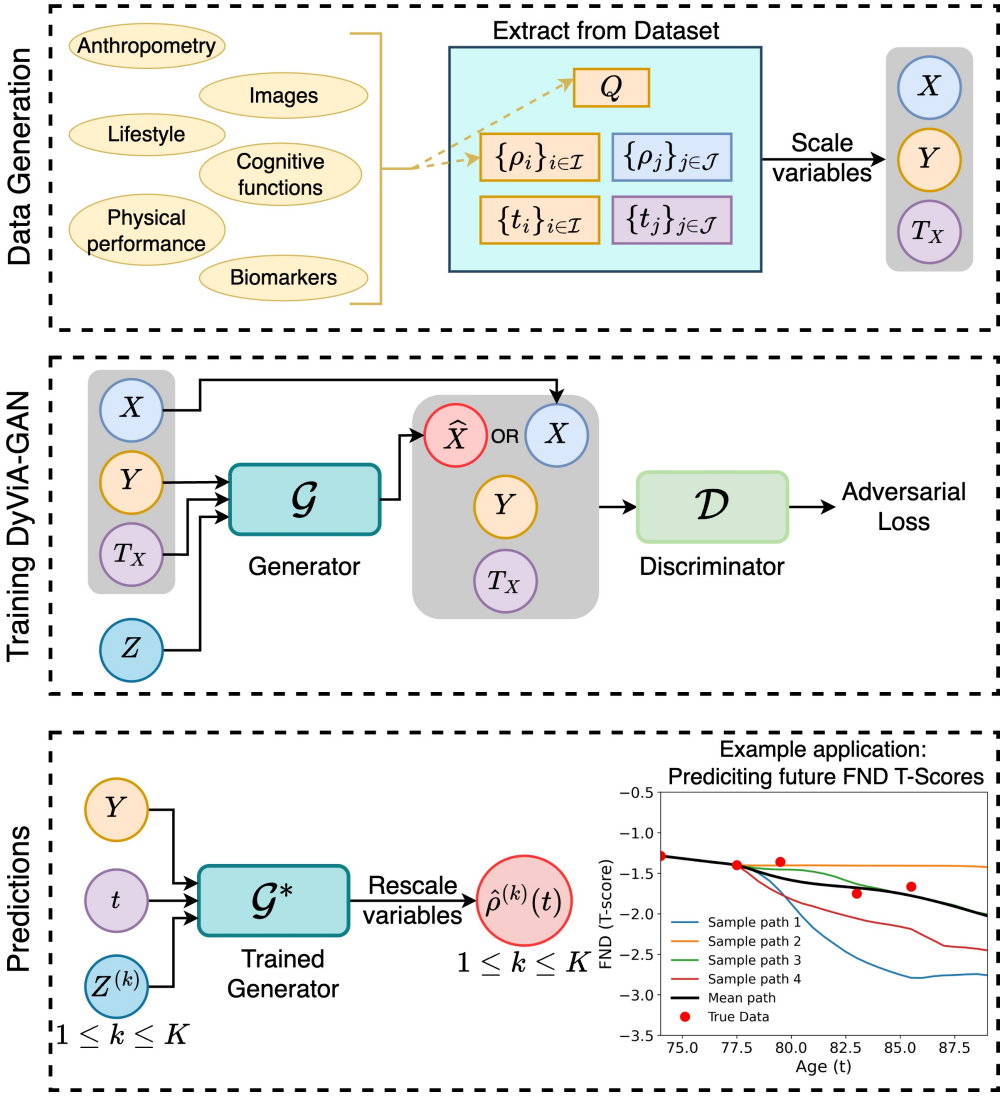


Figure 1: Schematic depicting the construction of the datasets, the inputs and outputs of $\mathcal{G}$ and $\mathcal{D}$ when training the DyViA-GAN, and how the trained generator $\mathcal{G}^*$ is used to predict plausible phenotype trajectories. **Top Panel:** The phenotype measurements, the time-points of the measurements, and the additional covariates Q are extracted from the dataset to construct the tuple (X, Y, T_X) after appropriate scaling. **Middle Panel:** During training phase, Y, Z and T_X are fed to $\mathcal{G}$ to obtain the prediction vector $\widehat{X}$ at the time-points in T_X. The discriminator $\mathcal{D}$ is fed either a true tuple (X, Y, T_X) or a fake tuple $(\widehat{X}, Y, T_X)$. Both network are trained simultaneously according to Equation 1. **Bottom Panel:** The trained generator $\mathcal{G}^*$ is fed the conditional data Y for a participant, any time-point t, and K samples of the latent variable. The output is an ensemble of K personalized trajectories $\hat{\rho}^{(k)}(t)$ for $1 \leq k \leq K$ after rescaling the variables. As an example, we consider the FND prediction problem, with 4 sample trajectories and the mean trajectory (black curve) shown, which can be compared to the known phenotype measurements (red dots).

Table 2: Common hyperparameter values used for the DyViA-GAN models.

Hyperparameter	Value/Choice
Latent Dimension N_z	20
Optimizer	Adam
Learning rate for $\mathcal{G}$	10^{-3}
Learning rate for $\mathcal{D}$	10^{-3}
Number of $\mathcal{D}$ updates per $\mathcal{G}$ update	4
ℓ^2 regularization parameter	10^{-7}
Batch size	ALL: 200, OWH: 200, OW: 180, OB: 90
Total epochs	2000

2.4 DyViA-GAN model nomenclature

We trained multiple version of DyViA-GAN on different data and covariate combinations. For easier reference, we use the nomenclature Model-$\langle$DT$\rangle$ for the models trained on different data types determined by tag $\langle$DT$\rangle$. The tag value $\langle$DT$\rangle$ = ALL implies that the model is trained on unstratified training data with 2100 samples (see Table 1), $\langle$DT$\rangle$ = UWH implies the networks are trained on the UWH strata samples with 900 samples, $\langle$DT$\rangle$ = OW implies training on the OW strata with 800 training samples, while $\langle$DT$\rangle$ = OB implies training on the OB strata with 400 training samples. Furthermore, the model name ends with an additional tag -Q if the covariate Q is fed as input to the generator. For example, Model-OWH denotes the model trained (and tested) on the OWH strata, while Model-OB-Q denotes the model trained (and tested) on OB strata with Q used as an input for the generator. In all, we consider 8 DyViA-GAN models: 1) Model-ALL, 2) Model-ALL-Q, 3) Model-UWH, 4) Model-UWH-Q, 5) Model-OW, 6) Model-OW-Q, 7) Model-OB, and 8) Model-OB-Q. Furthermore, we consider two variants of the model, one with $\mathcal{J} = \{5\}$ and another with $\mathcal{J} = \{5, 8\}$. Thus there are 8 DyViA-GAN models for each variant.

2.5 Evaluation of model performance

Using the trained generator $\mathcal{G}^*$, we quantify the performance of the various models via the following metrics:

- **Root mean squared error (RMSE)**: This will be evaluated either across all time-points not fed as input to the generator (see Table 3), or at specific time-points (see Table 4).
- **Maximal score:** For a particular participant, we define the score $\mathcal{S}(t)$ at any time-point t using the empirical measures computed in (2)

$$\mathcal{S}(t) = \frac{|\rho(t) - \overline{\rho}^*(t)|}{\sqrt{\mathrm{Var}[\rho^*](t)}} \tag{5}$$

  which is the normalized distance of the true value of the phenotype at t from the mean predicted value at that time-point. The maximal score $\mathcal{S}_M$ is define as the maximum over of an individuals score over

all time-points not shown to the generator, and at which the true value is known. In the present context, $\mathcal{S}_M = \max_{t \in \{t_5, t_8\}} S(t)$, with the empirical statistics computed using $K = 1000$ random samples of Z. Thus, predictions with lower maximal scores are preferable. We say that the *score is acceptable* for a sample if $\mathcal{S}_M \leq \epsilon_s$ where ϵ_s is a tunable parameter.

- **Filtering plausible trajectories:** In order to eliminate outliers from the predicted plausible trajectories for an individual, we perform an additional post-processing filtering. For each predicted plausible trajectory $\hat{\rho}_k,\ 1 \leq k \leq K$, consider the mean path $\bar{\bar{\rho}}[k]$ computed using the remaining $K-1$ trajectories, i.e., excluding the k-th trajectory. We then retain the k-th trajectory if $\text{dist}(\hat{\rho}_k, \bar{\bar{\rho}}[k]) < \epsilon_f$, where dist() is a meaningful function to measure the distance between trajectories, while ϵ_f is a tunable parameter. In the present work, we take dist() to be the Dynamic Time Warping (DTW) distance, which is a commonly used elastic shape-based measure of similarity between a pair of time series [15].

We remark that the parameters ϵ_s used to define acceptable scores and ϵ_f used filter trajectories are both are problem-dependent and tunable by the practitioner based on domain expertise. In the present work, we choose $\epsilon_s = 2.5$ and $\epsilon_f = 2.0$ motivated by the ablation study presented in Appendix B.

3 Results

We ran DyViA-GAN on the individual samples of available (or "true") FND measurements from the described dataset to generate their respective personalized continuous trajectories of FND based on the stated models.

Using the criteria discussed in (subsection 2.5), we evaluate and compare the performance of the various designed models. Although Model-ALL and Model-ALL-Q are trained on unstratified data, we also evaluate the trained generators on stratified data for a more fine-grained analysis. The RMSE and maximal scores of the other models trained on stratified data are only evaluated on the corresponding strata.

3.1 Score-based comparison

When comparing the two variants of the model, we observe from Table 3 that a larger percentage of training and test samples have acceptable scores with models trained using $\mathcal{J} = \{5\}$, as compared to models trained using $\mathcal{J} = \{5, 8\}$. This indicates that the plausible trajectories generated using the first variant have a larger spread about the predicted mean trajectory, thus having a better chance of capturing dynamics with larger temporal variations. We also observe that feeding the covariate Q (i.e., the BMI) as a generator input improves the performance of the models (both variants) trained on stratified data. For instance, the percentage of test samples with acceptable scores with models trained on OW strata increases from 82.11% to 90.53% for the first variant of the model (see top subtable in Table 3) and from 72.63% to 90.53% for the second variant of the model (see bottom subtable in Table 3).

3.2 RMSE-based comparison

If we consider the mean predicted trajectory as a single representation of the phenotype evolution, then the RMSE is a more appropriate metric to measure model performance. We observe from Table 4 that the models trained with $\mathcal{J} = \{5\}$ (top subtable) consistently lead to higher RMSE when compared to the respective models trained with $\mathcal{J} = \{5, 8\}$ (bottom subtable) . Thus, based on the RMSE metric, the later variant of the DyViA-GAN models perform significantly better. In general, the inclusion of the BMI as a

Table 3: Percentage (%) of samples with acceptable scores ($\mathcal{S}_{\mathcal{M}} \leq 2.5$) for 8 different DyViA-GAN models using $\mathcal{J} = \{5\}$ (top subtable) and $\mathcal{J} = \{5, 8\}$ bottom subtable). The percentages are also listed for each data strata for models trained on unstratified samples. For models trained on stratified data, the percentages are listed only for the corresponding strata. The notation - indicates no scores are evaluated for a particular strata with a given model.

Models using $\mathcal{J} = \{5\}$								
	Training				Test			
Strata / Model	UWH	OW	OB	ALL	UWH	OW	OB	ALL
Model-ALL	89.12	86.81	88.61	88.14	87.76	83.16	90.00	86.27
Model-ALL-Q	89.75	86.94	88.61	88.46	87.76	84.21	90.00	86.70
Model-UWH	85.00	-	-	-	77.55	-	-	-
Model-UWH-Q	91.62	-	-	-	87.76	-	-	-
Model-OW	-	84.72	-	-	-	82.11	-	-
Model-OW-Q	-	90.56	-	-	-	90.53	-	-
Model-OB	-	-	79.17	-	-	-	80.00	-
Model-OB-Q	-	-	84.17	-	-	-	85.00	-

Models using $\mathcal{J} = \{5, 8\}$								
	Training				Test			
Strata / Model	UWH	OW	OB	ALL	UWH	OW	OB	ALL
Model-ALL	82.62	78.33	81.11	80.69	82.65	80.00	87.50	82.40
Model-ALL-Q	80.00	76.11	79.44	78.40	79.59	78.95	87.50	80.69
Model-UWH	79.88	-	-	-	77.55	-	-	-
Model-UWH-Q	81.50	-	-	-	79.59	-	-	-
Model-OW	-	72.78	-	-	-	72.63	-	-
Model-OW-Q	-	90.56	-	-	-	90.53	-	-
Model-OB	-	-	72.22	-	-	-	77.50	-
Model-OB-Q	-	-	73.33	-	-	-	80.00	-

generator input seems to marginally deteriorate the RMSE – with the exception of the $\mathcal{J} = \{5, 8\}$ variant of Model-ALL-Q where the error reduces when BMI is included.

3.3 Checking and filtering predicted trajectories

With the goal of eliminating outliers among the predicted trajectories, we filter the trajectories generated by the DyViA-GAN models using DTW with a filtering parameter $\epsilon_f = 2.0$. We demonstrate the qualitative performance of the trained models with this filtering on two test participants chosen from each strata UWH, OW and OB. For each participant, we generate $K = 1000$ trajectories using each model, and then compute the personalized mean trajectory only using the filtered trajectories. The filtered trajectories and corresponding mean trajectories are shown in Figure 2 (OWH), Figure 3 (OW) and Figure 4 (OB).

In addition, we have access to the adjusted phenotype value at Visit 7 (1999) for these six test participants. This serves as held-out data point for both variants of DyViA-GAN, as data from this visit was not used during the training phase. Note that Visit 8 is also a held-out data point for the models trained with $\mathcal{J} = \{5\}$. We depict the generator inputs using blue dots, the generator predictions using green triangles, and mark the unseen/held-out time-points using red (unfilled) triangles in Figure 2, Figure 3 and Figure 4. Appendix A contains additional plots depicted a few sample trajectories for each of these test participants for better clarity. We make the following observations from the plots in Figure 2, Figure 3 and Figure 4:

- The mean trajectories pass through the true data at t_2, t_4, which is a feature of the constraint Equation 4 applied on the generator output.
- With the filtering threshold $\epsilon_f = 2.0$, a significantly larger number of plausible trajectories are accepted with the $\mathcal{J} = \{5, 8\}$ model variants as compared to the $\mathcal{J} = \{5\}$ variants. This can be easily inferred by the observing the denseness of the filtered trajectories shown in the figures.
- The filtered trajectories with the $\mathcal{J} = \{5, 8\}$ models are significantly more nonlinear with richer structure, while those obtained using the $\mathcal{J} = \{5\}$ models appear to be mostly (piecewise) linear (also see Supplementary Material).
- The (filtered) mean with the $\mathcal{J} = \{5, 8\}$ models leads to better approximation of the data at unseen data at future time point t_7.
- The mean trajectory with all models generally predict a downward trend in the FND values as time progresses, which is consistent with clinical observations in aging patients.

3.4 Benchmarking against other methods

We compare the proposed DyViA-GAN against other standard approaches. In particular, we consider a standard cubic smoothing spline model, linear mixed-effect models (LMMs), and Gaussian process (GP) regression. Among these, the later two approaches are well suited to for longitudinal data with unstructured time-points.

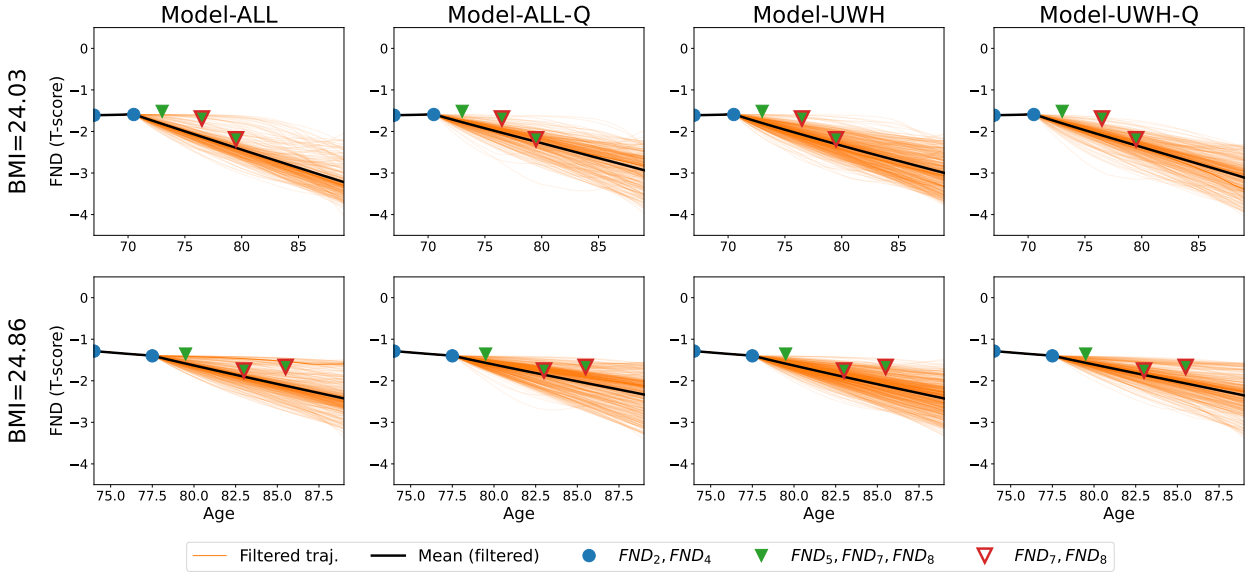
3.4.1 Cubic splines

We use a standard cubic smoothing spline model [19] as implemented by the 'splinefun' function in R. It has been noted that the advantage of such a model (say, over a full ARIMA model) is that not only it provides

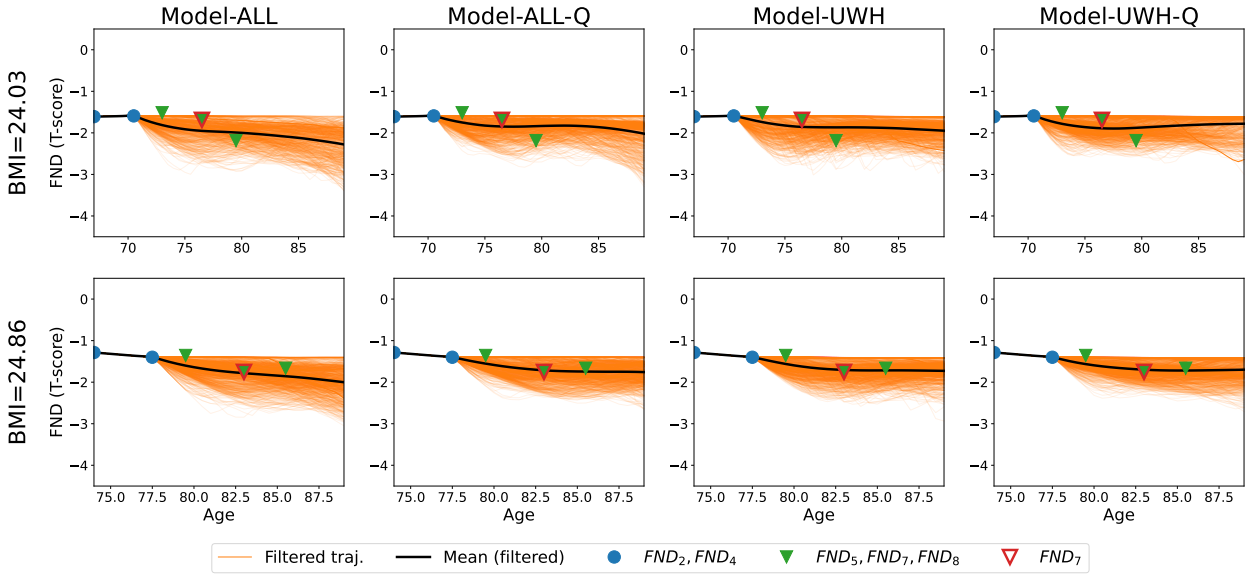
Table 4: RMSE for 8 different DyViA-GAN models using $\mathcal{J} = \{5\}$ (top subtable) and $\mathcal{J} = \{5, 8\}$ bottom subtable). The errors are also listed for each data strata for models trained on unstratified samples. For models trained on stratified data, the errors are listed only for the corresponding strata. The notation - indicates no errors are evaluated for a particular strata with a given model.

Models using $\mathcal{J} = \{5\}$								
	Training				Test			
Strata / Model	UWH	OW	OB	ALL	UWH	OW	OB	ALL
Model-ALL	0.53	0.57	0.52	0.54	0.56	0.68	0.57	0.61
Model-ALL-Q	0.55	0.59	0.53	0.56	0.59	0.69	0.58	0.63
Model-UWH	0.52	-	-	-	0.56	-	-	-
Model-UWH-Q	0.55	-	-	-	0.59	-	-	-
Model-OW	-	0.55	-	-	-	0.65	-	-
Model-OW-Q	-	0.55	-	-	-	0.66	-	-
Model-OB	-	-	0.59	-	-	-	0.64	-
Model-OB-Q	-	-	0.63	-	-	-	0.68	-

Models using $\mathcal{J} = \{5, 8\}$								
	Training				Test			
Strata / Model	UWH	OW	OB	ALL	UWH	OW	OB	ALL
Model-ALL	0.44	0.50	0.47	0.47	0.48	0.61	0.51	0.54
Model-ALL-Q	0.38	0.44	0.43	0.41	0.43	0.56	0.46	0.49
Model-UWH	0.38	-	-	-	0.43	-	-	-
Model-UWH-Q	0.39	-	-	-	0.44	-	-	-
Model-OW	-	0.45	-	-	-	0.56	-	-
Model-OW-Q	-	0.51	-	-	-	0.60	-	-
Model-OB	-	-	0.46	-	-	-	0.50	-
Model-OB-Q	-	-	0.44	-	-	-	0.48	-

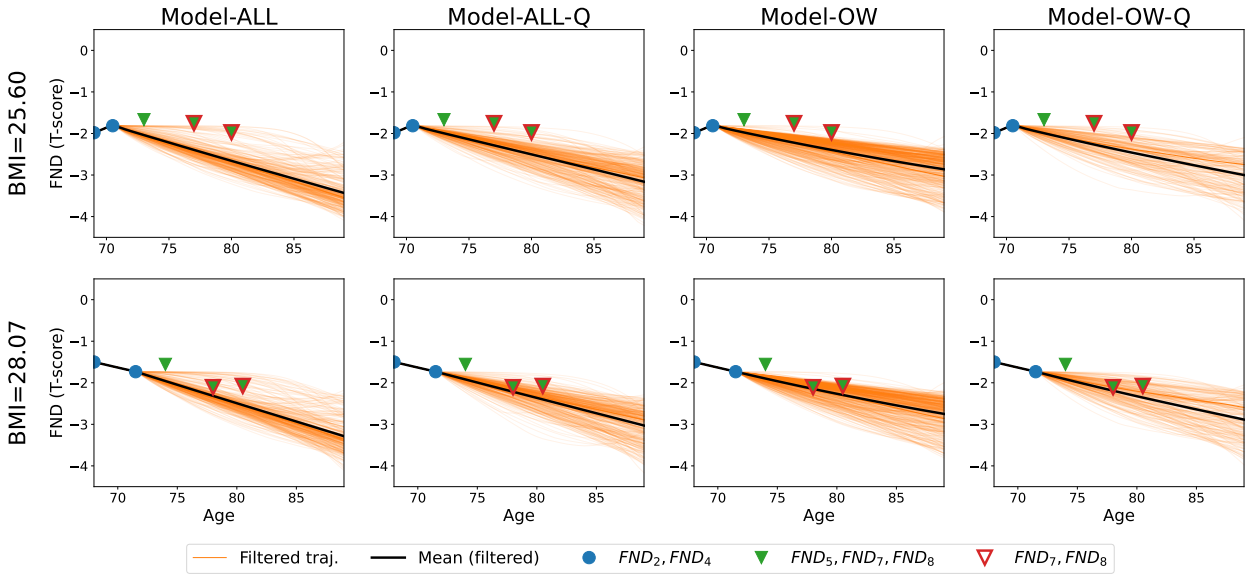


(a) Results for models using $\mathcal{J} = \{5\}$.

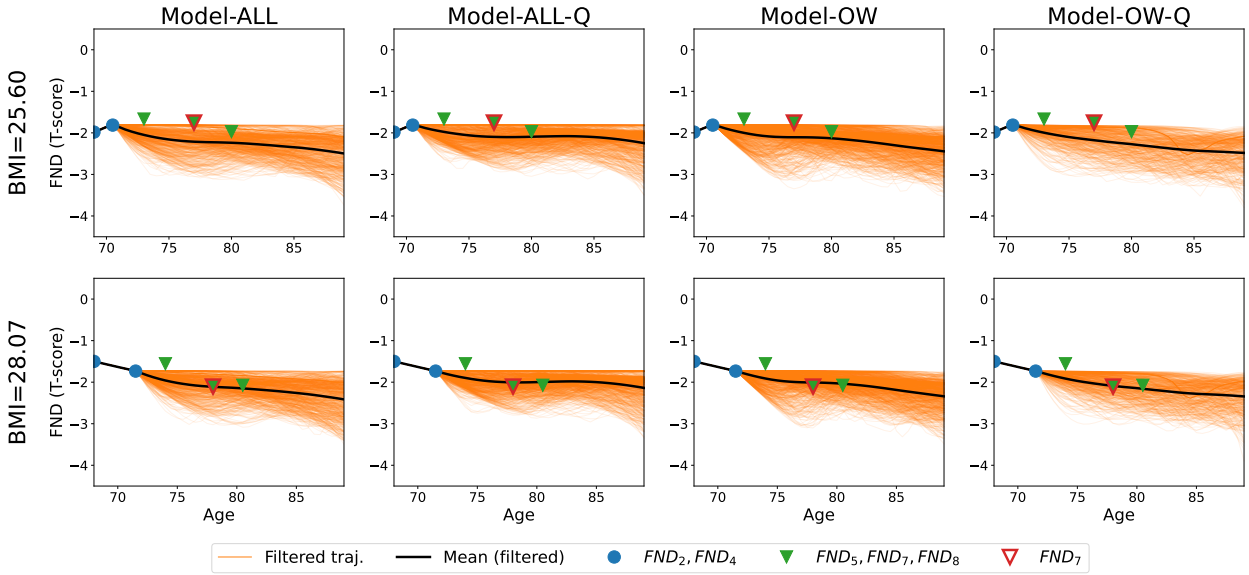


(b) Results for models using $\mathcal{J} = \{5, 8\}$.

Figure 2: Predictions with various models for two test participants from strata UWH. Each plot depicts the filtered trajectories (orange curves) out of 1000 original trajectories, the mean filtered personalized phenotype trajectory (black curve), and the phenotype measurements at times t_2, t_4, t_5, t_7, t_8 (dots and triangles). We depict the generator input using blue dots, the generator predictions using green triangles, and mark the prediction at Visit 7 (never shown to the DyViA-GAN) using red (unfilled) triangles. The BMI of the participant is listed on the left of each row.

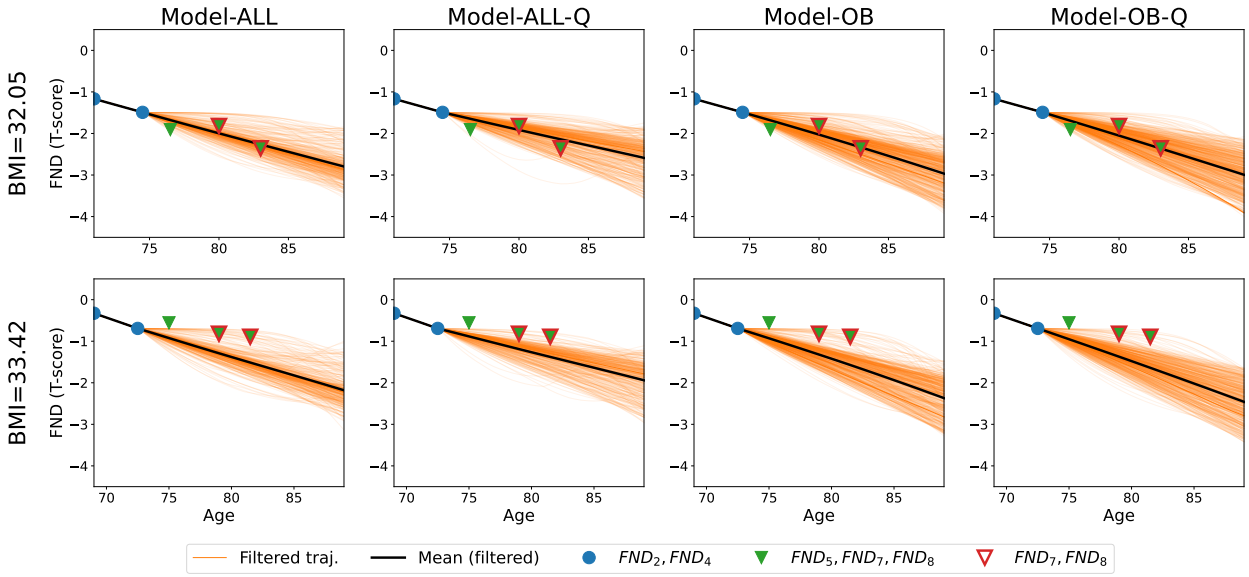


(a) Results for models using $\mathcal{J} = \{5\}$.

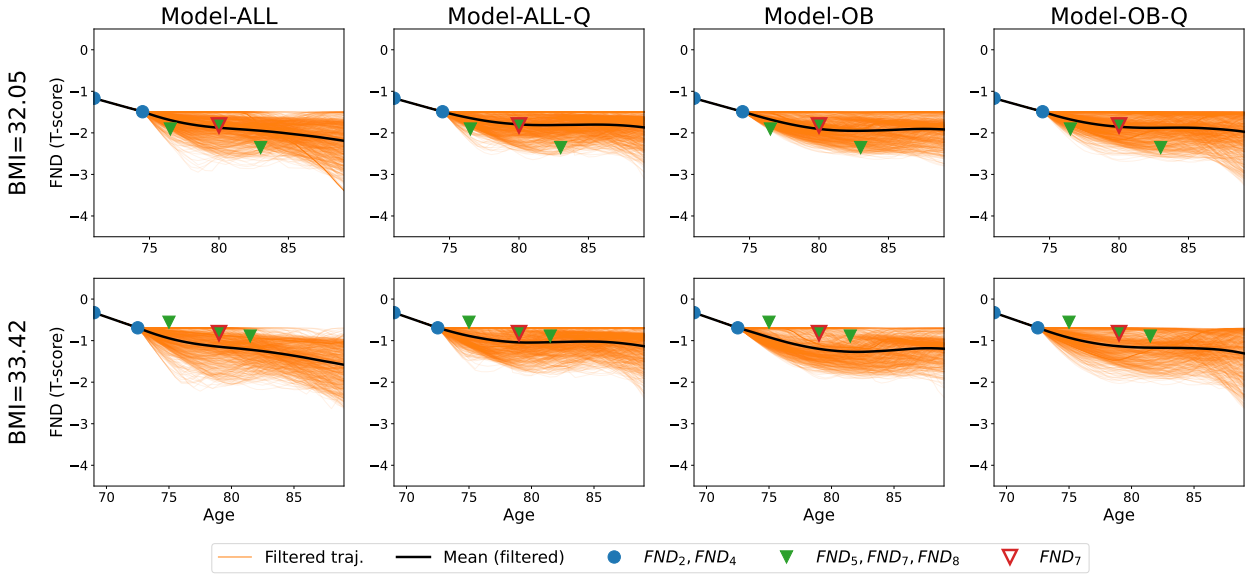


(b) Results for models using $\mathcal{J} = \{5, 8\}$.

Figure 3: Predictions with various models for two test participants from strata OW. Each plot depicts the filtered trajectories (orange curves) out of 1000 original trajectories, the mean filtered personalized phenotype trajectory (black curve), and the phenotype measurements at times t_2, t_4, t_5, t_7, t_8 (dots and triangles). We depict the generator input using blue dots, the generator predictions using green triangles, and mark the prediction at Visit 7 (never shown to the DyViA-GAN) using red (unfilled) triangles. The BMI of the participant is listed on the left of each row.



(a) Results for models using $\mathcal{J} = \{5\}$.



(b) Results for models using $\mathcal{J} = \{5, 8\}$.

Figure 4: Predictions with various models for two test participants from strata OB. Each plot depicts the filtered trajectories (orange curves) out of 1000 original trajectories, the mean filtered personalized phenotype trajectory (black curve), and the phenotype measurements at times t_2, t_4, t_5, t_7, t_8 (dots and triangles). We depict the generator input using blue dots, the generator predictions using green triangles, and mark the prediction at Visit 7 (never shown to the DyViA-GAN) using red (unfilled) triangles. The BMI of the participant is listed on the left of each row.

Table 5: Comparing the performance of Model-ALL and Model-ALL-Q with a spline-based regression model. The RMSE of the predictions at ages 74.5 and 81.5 years are shown.

Age \ Model	$\mathcal{J} = \{5\}$		$\mathcal{J} = \{5, 8\}$		
	Model-ALL	Model-ALL-Q	Model-ALL	Model-ALL-Q	Splines
74.5	0.26	0.26	0.27	0.26	0.32
81.5	0.70	0.72	0.57	0.51	0.67

a smooth historical trend but also a linear forecasting ability, which is hardly affected by the restricted parameter space [25]. The dataset to train the spline model is constructed by considering the maximal number of participants in the SOF dataset who have FND measurements from Visits 2, 4, 5 and 8 while also sharing the same ages t_2, t_4, t_5 and t_8 during these visits. This leads to 49 participants with (unscaled) ages 69.0, 72.5, 74.5 and 81.5 years. The spline model was trained on the first two ages and then use to predict the T-scores during the last two ages. We compare the RMSE of the spline predictions with the mean predictions at these two ages using the various DyViA-GAN models trained previously on unstratified data, i.e., Model-ALL and Model-ALL-Q. As show in Table 5, the DyViA-GAN models outperform the spline model when predicting the T-scores at 74.5 years. However, only the $\mathcal{J} = \{5, 8\}$ variants lead to lower errors than splines at age 81.5 years. Furthermore, including BMI as input improves the performance of this latter variant, which is also consistent with the observations from Table 4.

3.4.2 Linear mixed-effects models

We consider LMMs with fixed (population-level) effects modeled using a polynomial basis (in time) of dimension p and random (individual-level) effects modeled using a polynomial basis of dimension q. The random basis coefficients are modeled using a normal distribution $\mathcal{N}(0, \Sigma)$, and additional (individual-level) observation noise is modeled using $\mathcal{N}(0, \sigma^2 I)$. The coefficients of the fixed basis, covariance Σ for the random coefficients, and σ^2 for the noise model are determined by maximizing the log-likelihood on the (unstratified) training data. Once trained, the predictions for an individual are conditioned on their measurements at the first two time-points. We use the nomenclature LMM (p,q) to denote the LMM trained with basis dimension p and q.

We consider three variants of LMM based on the values of p and q. As shown in Table 6, the RMSE (of 0.49) on the test samples remains fairly robust across all variants, which perform better than Model-ALL and at par with Model-ALL-Q. We also compare the performance of the models on the six test participants considered in the previous sections. We observe from Figure 5 that all LMMs predict fairly linear (mean) trajectories. However, this is not ideal when the individual-level evolution of phenotypes is non-monotonic, as can be seen in Figure 5 for participants with BMI 25.60 and 32.06. In comparison, the DyViA-GAN predictions are far more expressive and capable of capturing non-monotone behavior. Furthermore, we have the flexibility of choosing a network architecture that ensures the predictions pass through (or close to) the given observations, which need not hold true for LMMs.

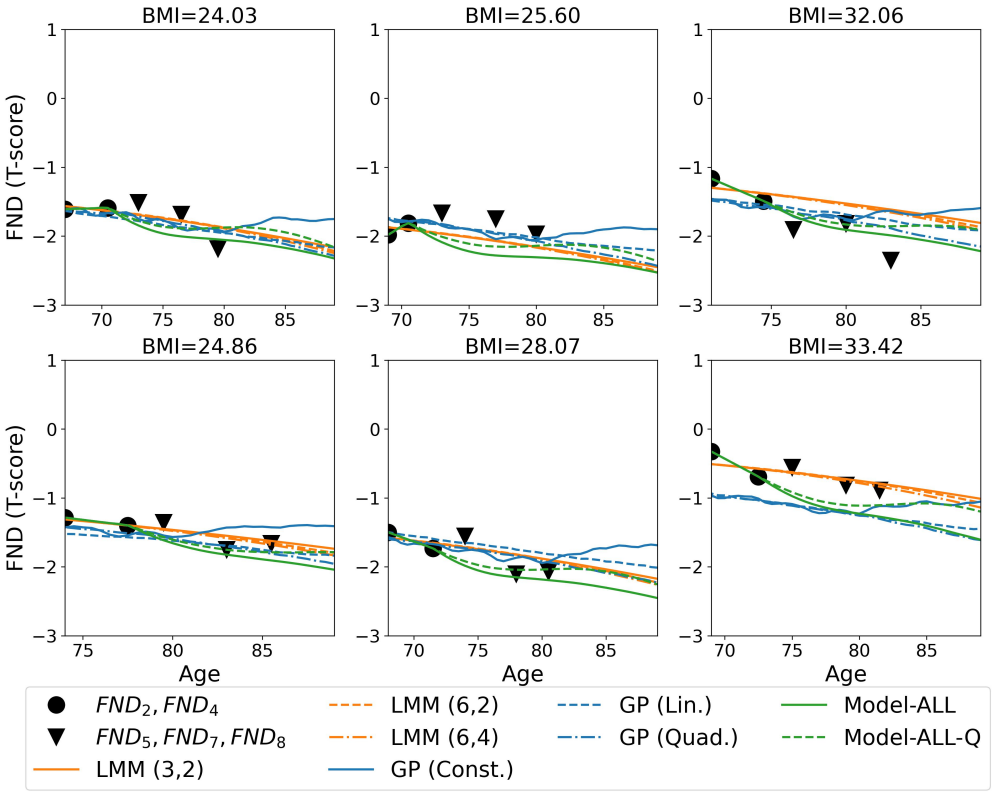


Figure 5: Comparing the mean predicted trajectories with LMMs, GPs, Model-ALL, and Model-ALL-Q, with $\mathcal{J} = \{5, 8\}$ for the later two approaches. The six test participants are the same as those considered in Figures 2, 3 and 4. Each plot depicts the phenotype measurements at the initial time-points t_2, t_4 (black dots) which are fed to the models as input, and future time-points t_5, t_7, t_8 (black triangles). The mean trajectories with the LMM models are shown in orange, with the GPs are shown in blue, and the DyViA-GAN are shown in green. The BMI of the participant is listed on the top of each plot.

Table 6: RMSE of the predictions of various models on the unstratified test data.

Model	RMSE
LMM (3,2)	0.49
LMM (6,2)	0.49
LMM (6,4)	0.49
GP Const.	0.62
GP Lin.	0.61
GP Quad.	0.62
Model-ALL	0.54
Model-ALL-Q	0.49

3.4.3 Gaussian process regression

We consider GP models with mixed effects. For individuals i and j whose FND are observed at times t and t', the mixed kernel is taken as $k((i,t),(j,t')) = k_{\text{pop}}(t,t') + \delta_{i,j} k_{\text{indiv}}(t,t')$, where k_{pop} models the population-level dynamics, k_{indiv} models the individual-level dynamics, and $\delta_{i,j} = 1$ only if $i = j$ (i.e., both measurements are for the same individual). In the present work, we take k_{pop} to be the Matérn kernel with parameter $\nu = 2.5$ and k_{indiv} to be a rougher Matérn kernel with $\nu = 0.5$. Furthermore, we consider the mean of GP to be either a constant mean, linear, or quadratic. The coefficients of the mean and kernel length-scales are determined by maximizing the log-likelihood on the (unstratified) training data. Like the LMMs and DyViA-GAN , the predictions with trained GP models are conditioned on the measurements at the first two time-points. We use the nomenclature GP Const., GP Lin., and GP Quad., based on how the mean is modeled.

We observe from the RMSE comparison in Table 6 as well as the predicted mean trajectories in Figure 5 that the GP models perform poorly when compared to LMMs and DyViA-GAN. Unlike the LMMs, the predictions with the GPs are not linear, and appear to possess similar expressivity to DyViA-GAN. Thus, it may be possible to fine-tune the GP, or choose a more sophisticated GP variant to improve its performance. This will be explored in future extensions of this work.

4 Discussion

While noting that the concept of *healthspan* is relatively new in geroscience research [5] and there is no universal agreement yet on its definition, the author of a perspective paper in *GeroScience* made two key observations: (ii) "health itself may be better considered as a continuous variable that changes in a dynamic way throughout life"; and (2) "[t]he health trajectory will be different in different individuals, but will generally trend downward with age".[28] However, understanding and prolonging individual-specific healthspan is challenging for many reasons, including not being timely aware of the trajectories of the different aging phenotypes. Advance knowledge of progression of such an outcome as osteopenia, the precursor of osteoporosis,

is crucial, as it can lead to an accelerated reduction in one's bone healthspan, and DyViA-GAN provides precisely that capability.

While there are several powerful approaches that one could use for the trajectory prediction problem, we used a GAN-based approach motivated by the following considerations:

- GANs allow for the generation of many possible outputs associated with a given input. In other words, it is a probabilistic method versus a deterministic one.
- Deep learning frameworks, such as GANs, allow the easy integration of different data modalities. For example, we can easily handle input comprising CT scans (i.e., images type data) and femoral neck density at different time points (i.e., longitudinal scalar data). Furthermore, we have the flexibility of having a different data type for the predictions. This level of flexibility is seldom granted by more traditional algorithms.
- Once trained, inference/predictions on new data is instantaneous. This is in fact an advantage over more sophisticated generative algorithms such as diffusion models [7].

The present study has, in fact, some distinct advantages. Using only very few initial observations as input, it demonstrated the capability of cGAN-type models to predict the personalized trajectory of a single phenotype – femoral neck BMD – that is particularly relevant to one's healthspan, and increasingly so as one continues to age. The generators of the proposed DyViA-GAN takes as input the (scaled) age variable t and the output generated is a *continuous* trajectory for the entire age-span, and not just for some discrete time points at which data is available during training/testing. The continuous representation also allows for easy calculation of the gradient (and higher derivatives) of each trajectory using automatic differentiation on neural nets.

Notably, there are several key findings from the results presented in section 3. First, the score-based metric indicates that the $\mathcal{J} = \{5\}$ DyViA-GAN models generate more spread-out trajectories. However, this might lead to the generation of outlier trajectories which can skewed statistical predictions. This was observed when using the RMSE metric which clearly showed that $\mathcal{J} = \{5, 8\}$ DyViA-GAN models lead to better predictions (in the mean), even when compared with a standard spline-based regression model. Second, the inclusion of additional risk factors, such as BMI, may lead to better personalized predictions in certain situations. Third, the $\mathcal{J} = \{5\}$ models lead to generation of simplistic trajectories as compared $\mathcal{J} = \{5, 8\}$, indicating the latter are more suitable to capture nonlinear dynamics. Fourth, filtering the predicted trajectories provides a useful strategy to eliminate outliers, leading to improved predictions for the unseen data from Visit 7. Fifth, the Model-ALL-Q variant of DyViA-GAN is at par with standard LMMs in terms of the RMSE metric while being more expressive when predicting individual-level non-monotone trajectories. The Model-ALL-Q and LMMs outperform the GP models.

According to the U.S. Census Bureau's 2017 National Population Projections, the year 2030 will mark an important demographic turning point in U.S. history when one in five Americans will be 65 years or older. Clearly, planning for and achieving individual healthspan goals will assume increasing importance in the society. Towards this, a variety of new sources of rich, high-resolution data such as genomics, epigenetics, wearable monitoring devices, etc., is making healthcare more preventive and personalized. Such data also allows for complex interplays among the factors of healthy aging – ranging from static to dynamic, and both internal and external – to inform the models for predicting a wider and more insightful set of outcomes. For

instance, a better understanding of the non-linear relationship between one's BMI and BMD, over the course of aging, could emerge [11]. Towards this end, the translational capacity of AI-based approaches to preventive and precision medicine including the present platform can be refined and tested in clinical trials.

We understand that there are certain limitations of our study. As future work, our approach could be extended to predict simultaneously the trajectories of multiple phenotypes, say, BMD values of multiple regions of the body or fusion thereof. Moreover, key insights are available from other markers of osteoporosis beyond BMI, which we chose for illustrative purposes and for its easy availability and relevance, e.g., [2]. In fact, DyViA-GAN has the flexibility to include, in addition to static baseline risk factors, also dynamic covariates such as a participant's physical activity level, chronic inflammation, bone biomarkers, epigenetic changes, etc. The effect of including such covariates in the trajectory prediction problem will be considered in future extensions. We also plan to train the models with longitudinal BMD data from heterogeneous populations, e.g., [22]. The presently generated trajectories are not dynamically updated, which could be revised in the future to allow for intermediate events, lifestyle alterations, and medical interventions. Further, we plan to extend the application to more sophisticated deep learning architectures to build the cGAN generator that accounts for the temporal nature of data (e.g., RNN, LSTM), alternative deep generative frameworks (e.g., [3, 13]), as well as other relevant longitudinal datasets on aging phenotypes and groups to train our generative models (e.g., [12]).

Acknowledgments

The authors thank the SOF Research Group for the publicly available data.

References

[1] Dataset: SOF Online. https://sofonline.ucsf.edu/, 2020.

[2] G. Barrera, D. Bunout, V. Gattás, M. P. de la Maza, L. Leiva, and S. Hirsch. A high body mass index protects against femoral neck osteoporosis in healthy elderly subjects. *Nutrition*, 20(9):769–771, 2004.

[3] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[4] D. W. Belsky, A. Caspi, D. L. Corcoran, K. Sugden, R. Poulton, L. Arseneault, A. Baccarelli, K. Chamarti, X. Gao, E. Hannon, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife*, 11: e73420, 2022.

[5] J. B. Burch, A. D. Augustine, L. A. Frieden, E. Hadley, T. K. Howcroft, R. Johnson, P. S. Khalsa, R. A. Kohanski, X. L. Li, F. Macchiarini, et al. Advances in geroscience: impact on healthspan and chronic disease. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 69(Suppl_1):S1–S3, 2014.

[6] V. M. Campello, T. Xia, X. Liu, P. Sanchez, C. Martin-Isla, S. E. Petersen, S. Segui, S. A. Tsaftaris, and K. Lekadir. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Frontiers in Cardiovascular Medicine*, 9:983091, 2022.

[7] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li. A survey on generative diffusion models. *IEEE transactions on knowledge and data engineering*, 36(7):2814–2830, 2024.

[8] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[9] E. Choi, M. T. Bahadori, J. Sun, J. A. Kulas, A. Schuetz, and W. F. Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Neural Information Processing Systems*, 2016. URL `https://api.semanticscholar.org/CorpusID:948039`.

[10] C. Cooper, G. Campion, and L. Melton III. Hip fractures in the elderly: a world-wide projection. *Osteoporosis international*, 2(6):285–289, 1992.

[11] S. Cummings, M. Nevitt, W. Browner, K. Stone, K. Fox, K. Ensrud, J. Cauley, D. Black, and T. Vogt. Risk factors for hip fracture in white women. study of osteoporotic fractures research group. *N Engl J Med.*, 332(12):767–773, 1995. doi: 10.1056/NEJM199503233321202.

[12] S. R. Cummings, P. M. Cawthon, K. E. Ensrud, J. A. Cauley, H. A. Fink, and E. S. Orwoll. BMD and risk of hip and nonvertebral fractures in older men: a prospective study and comparison with older women. *Journal of bone and mineral research*, 21(10):1550–1556, 2006.

[13] A. Dasgupta, A. M. da Cunha, A. Fardisi, M. Aminy, B. Binder, B. Shaddy, and A. A. Oberai. Unifying and extending diffusion models through pdes for solving inverse problems. *arXiv preprint arXiv:2504.07437*, 2025.

[14] L. DiPietro, G. Welch, D. Davis, J. Drane, and C. Macera. Body mass and risk of hip fracture among a national cohort of postmenopausal white women: a reanalysis. *Obes Res.*, 1(5):357–363, 1993. doi: 10.1002/j.1550-8528.1993.tb00013.x.

[15] S. Dove, M. Böhm, R. Freeman, S. Jellesmark, and D. J. Murrell. A user-friendly guide to using distance measures to compare time series in ecology. *Ecology and Evolution*, 13(10):e10520, 2023.

[16] V. Ecker, M. Früh, B. Yang, S. Gatidis, and T. Küstner. Deep regression for biological age estimation in multiple organs: Investigations on 40,000 subjects of the uk biobank. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2255–2259. IEEE, 2024.

[17] C. K. Fisher, A. M. Smith, and J. R. Walsh. Machine learning for comprehensive forecasting of alzheimer's disease progression. *Scientific reports*, 9(1):13622, 2019.

[18] G. C. Fox, G. von Laszewski, F. Wang, and S. Pyne. AICov: An integrative deep learning framework for covid-19 forecasting with population covariates. *Journal of Data Science*, 19(2):293–313, 2021. ISSN 1680-743X. doi: 10.6339/21-JDS1007.

[19] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.

[20] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso. Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 10335–10342. IEEE, 2021.

[21] D. Hanley and R. Josse. Prevention and management of osteoporosis: consensus statements from the scientific advisory board of the osteoporosis society of Canada. *CMAJ*, 55(7):921–923, 1996. doi: doi:10.1001/jama.285.6.785.

[22] D. He, H. Liu, W. Wei, Y. Zhao, Q. Cai, S. Shi, X. Chu, X. Qin, N. Zhang, P. Xu, et al. A longitudinal genome-wide association study of bone mineral density mean and variability in the uk biobank. *Osteoporosis International*, 34(11):1907–1916, 2023.

[23] Q. Hong, J. W. Siegel, Q. Tan, and J. Xu. On the activation function dependence of the spectral bias of neural networks. *arXiv preprint arXiv:2208.04924*, 2022. URL `https://arxiv.org/abs/2208.04924`.

[24] S. Horvath and K. Raj. Dna methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature reviews genetics*, 19(6):371–384, 2018.

[25] R. J. Hyndman, M. L. King, I. Pitrun, and B. Billah. Local linear forecasts using cubic smoothing splines. *Australian & New Zealand Journal of Statistics*, 47(1):87–99, 2005.

[26] E. Jung, M. Luna, and S. H. Park. Conditional gan with 3d discriminator for mri generation of alzheimer's disease progression. *Pattern Recognition*, 133:109061, 2023.

[27] M. Jung, V. K. Raghu, M. Reisert, H. Rieder, S. Rospleszcz, T. Pischon, T. Niendorf, H.-U. Kauczor, H. Völzke, R. Bülow, et al. Deep learning-based body composition analysis from whole-body magnetic resonance imaging to predict all-cause mortality in a large western population. *EBioMedicine*, 110, 2024.

[28] M. Kaeberlein. How healthy is the healthspan concept? *Geroscience*, 40(4):361–364, 2018.

[29] L. Kahwati, C. Kistler, G. Booth, N. Sathe, R. Gordon, E. Okah, R. Wines, and M. Viswanathan. Screening for osteoporosis to prevent fractures: A systematic evidence review for the US preventive services task force. *JAMA.*, 333(6):509–531, 2025. doi: 10.1001/jama.2024.21653.

[30] J. Kanis. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: Synopsis of a WHO report. *Osteoporosis International*, 4(6):368–381, 1994. doi: 10.1007/BF01622200.

[31] G. Li, L. Thabane, A. Papaioannou, G. Ioannidis, M. Levine, and J. Adachi. An overview of osteoporosis and frailty in the elderly. *BMC Musculoskelet Disord*, 18(1):1–5, 2017. doi: 10.1186/s12891-017-1403-x.

[32] B. Lim and M. van der Schaar. Disease-atlas: Navigating disease trajectories using deep learning. In *Machine Learning for Healthcare Conference*, pages 137–160. PMLR, 2018.

[33] A. C. Looker, E. S. Orwoll, C. C. Johnston JR, R. L. Lindsay, H. W. Wahner, W. L. Dunn, M. S. Calvo, T. B. Harris, and S. P. Heyse. Prevalence of low femoral bone density in older US adults from NHANES III. *Journal of Bone and Mineral Research*, 12(11):1761–1768, 1997.

[34] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[35] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. URL `https://arxiv.org/abs/1411.1784`.

[36] C. Moffatt and J. Leshin. Best practices in evolving privacy frameworks for patient age data: Census data study. *JMIR formative research*, 8:e47248, 2024.

[37] B. L. Moreland, J. K. Legha, K. E. Thomas, and E. R. Burns. Hip fracture-related emergency department visits, hospitalizations and deaths by mechanism of injury among adults aged 65 and older, United States 2019. *Journal of aging and health*, 35(5-6):345–355, 2023.

[38] W. K. Nicholson, M. Silverstein, J. B. Wong, D. Chelmow, T. R. Coker, E. M. Davis, C. R. Jaén, M. Krousel-Wood, S. Lee, L. Li, et al. Screening for osteoporosis to prevent fractures: US preventive services task force recommendation statement. *JAMA*, 333(6):498–508, 2025.

[39] NIH Consensus Development Panel on Osteoporosis Prevention, Diagnosis, and Therapy. Osteoporosis prevention, diagnosis, and therapy. *JAMA*, 285(6):785–795, 2001. doi: doi:10.1001/jama.285.6.785.

[40] A. Paganini-Hill, A. Chao, R. Ross, and B. Henderson. Exercise and other factors in the prevention of hip fracture: The leisure world study. *Epidemiology.*, 2(1):16–25, 1991. doi: 10.1097/00001648-199101000-00004.

[41] E. Putin, P. Mamoshina, A. Aliper, M. Korzinkin, A. Moskalev, A. Kolosov, A. Ostrovskiy, C. Cantor, J. Vijg, and A. Zhavoronkov. Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging (albany NY)*, 8(5):1021, 2016.

[42] S. Pyne, A. K. S. Vullikanti, and M. V. Marathe. Big data applications in health sciences and epidemiology. In *Handbook of statistics*, volume 33, pages 171–202. Elsevier, 2015.

[43] V. K. Raghu, J. Weiss, U. Hoffmann, H. J. Aerts, and M. T. Lu. Deep learning to estimate biological age from chest radiographs. *Cardiovascular Imaging*, 14(11):2226–2236, 2021.

[44] A. Rao and S. Pyne, editors. *Handbook of Statistics: Disease Modelling and Public Health*, volume 36. Elsevier, 2017.

[45] A. Rao and S. Pyne, editors. *Handbook of Statistics: Disease Modelling and Public Health*, volume 37. Elsevier, 2017.

[46] I. R. Reid. Fat and bone. *Arch Biochem Biophys*, 503(1):20–27, 2010. doi: 10.1016/j.abb.2010.06.027.

[47] J. F. Rodrigues-Jr, M. A. Gutierrez, G. Spadon, B. Brandoli, and S. Amer-Yahia. Lig-doctor: Efficient patient trajectory prediction using bidirectional minimal gated-recurrent networks. *Information Sciences*, 545:813–827, 2021.

[48] M. Shah, M. H. de A. Inácio, C. Lu, P.-R. Schiratti, S. L. Zheng, A. Clement, A. de Marvao, W. Bai, A. P. King, J. S. Ware, et al. Environmental and genetic predictors of human cardiovascular ageing. *Nature communications*, 14(1):4941, 2023.

[49] V. Shankar, E. Yousefi, A. Manashty, D. Blair, and D. Teegapuram. Clinical-GAN: trajectory forecasting of clinical events using transformer and generative adversarial networks. *Artificial Intelligence in Medicine*, 138:102507, 2023.

[50] S. Siami-Namini, N. Tavakoli, and A. S. Namin. A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1394–1401, 2018.

[51] T. Siggaard, R. Reguant, I. F. Jørgensen, A. D. Haue, M. Lademann, A. Aguayo-Orozco, J. X. Hjaltelin, A. B. Jensen, K. Banasik, and S. Brunak. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nature communications*, 11(1):4952, 2020.

[52] Y. E. Tian, V. Cropley, A. B. Maier, N. T. Lautenschlager, M. Breakspear, and A. Zalesky. Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature medicine*, 29(5):1221–1231, 2023.

[53] N. B. Watts, W. D. Leslie, A. J. Foldes, and P. D. Miller. 2013 international society for clinical densitometry position development conference: task force on normative databases. *Journal of Clinical Densitometry*, 16(4):472–481, 2013.

[54] D. Wilczok. Deep learning and generative artificial intelligence in aging research and healthy longevity medicine. *Aging (Albany NY)*, 17(1):251, 2025.

[55] N. C. Wright, A. C. Looker, K. G. Saag, J. R. Curtis, E. S. Delzell, S. Randall, and B. Dawson-Hughes. The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine. *Journal of bone and mineral research*, 29(11):2520–2526, 2014.

[56] L. Zhang, L. Cheng, H. Li, J. Gao, C. Yu, R. Domel, Y. Yang, S. Tang, and W. Liu. Hierarchical deep-learning neural networks: finite elements and beyond. *Computational Mechanics*, 67(1):207–230, Jan. 2021. ISSN 0178-7675. doi: 10.1007/s00466-020-01928-9. Publisher Copyright: © 2020, Springer-Verlag GmbH Germany, part of Springer Nature.

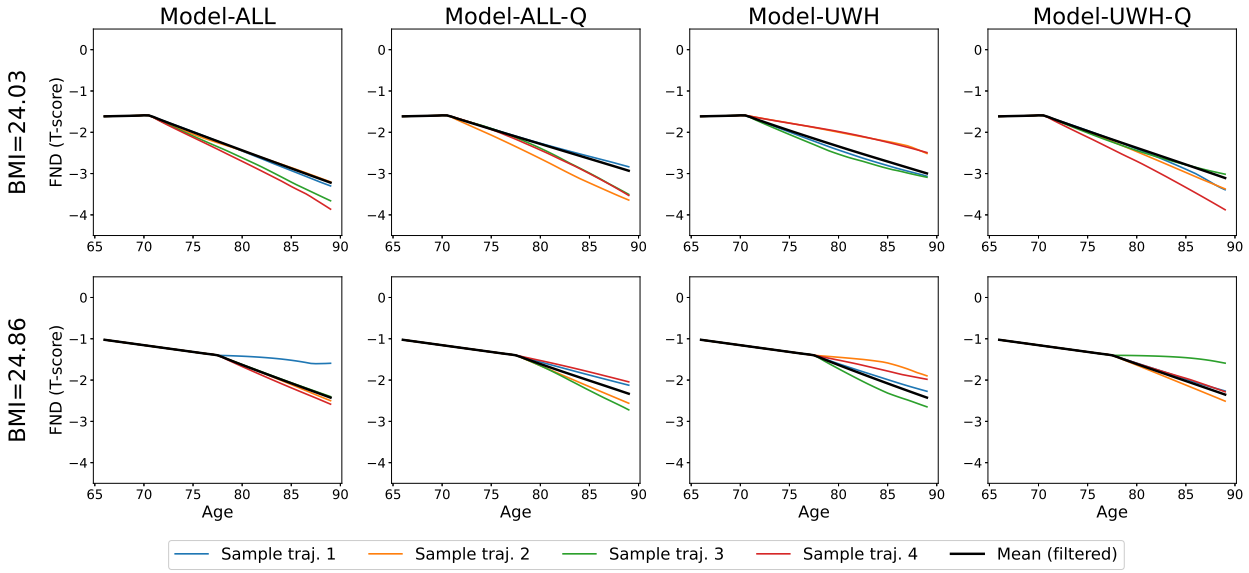
[57] X. Zhang, S. Pyne, and B. Kedem. Multivariate tail probabilities: predicting regional pertussis cases in Washington state. *Entropy*, 23(6):675, 2021.

[58] Y. Zhang, R. Wu, S. M. Dascalu, and F. C. Harris Jr. A novel extreme adaptive GRU for multivariate time series forecasting. *Scientific Reports*, 14(1):2991, 2024.

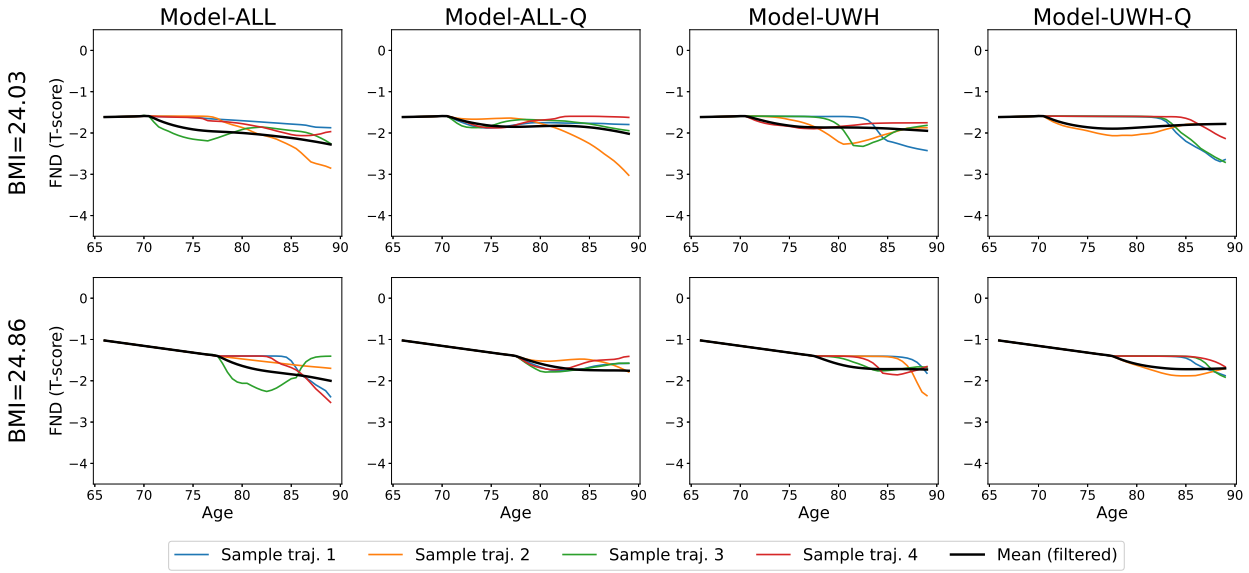
[59] A. Zhavoronkov, P. Mamoshina, Q. Vanhaelen, M. Scheibye-Knudsen, A. Moskalev, and A. Aliper. Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing research reviews*, 49:49–66, 2019.

Appendix A Additional results

As a companion to the test samples considered in Figure 2, Figure 3, Figure 4, we plot a few (filtered) sample trajectories predicted by DyViA-GAN for each of these test samples. These are shown in Figure A.1, Figure A.2 and Figure A.3 below.

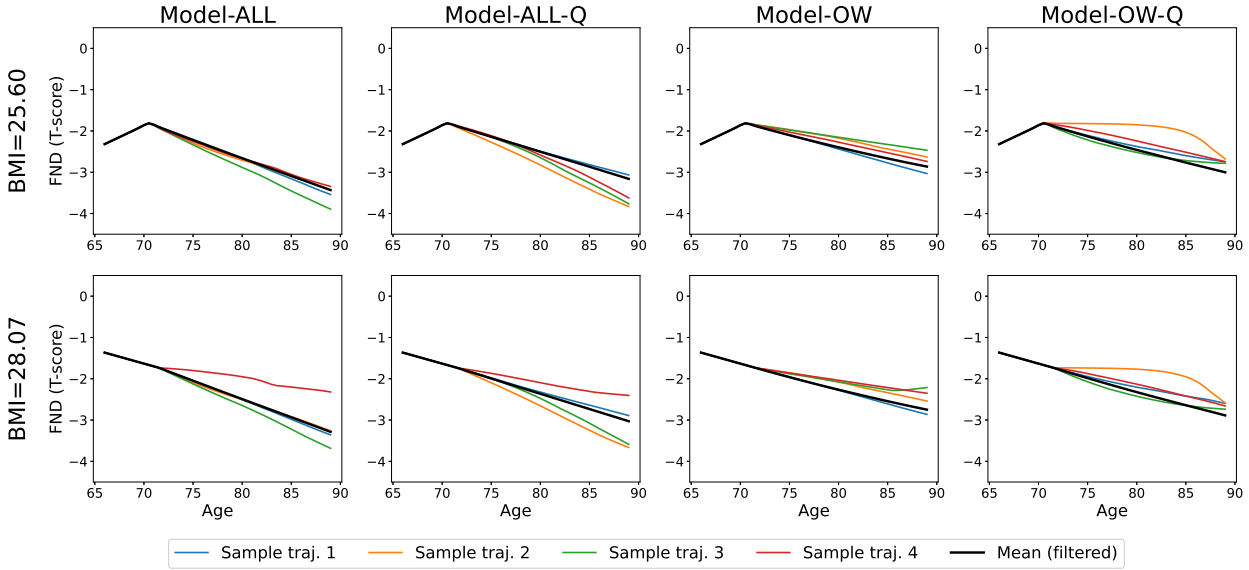


(a) Results for models using $\mathcal{J} = \{5\}$.

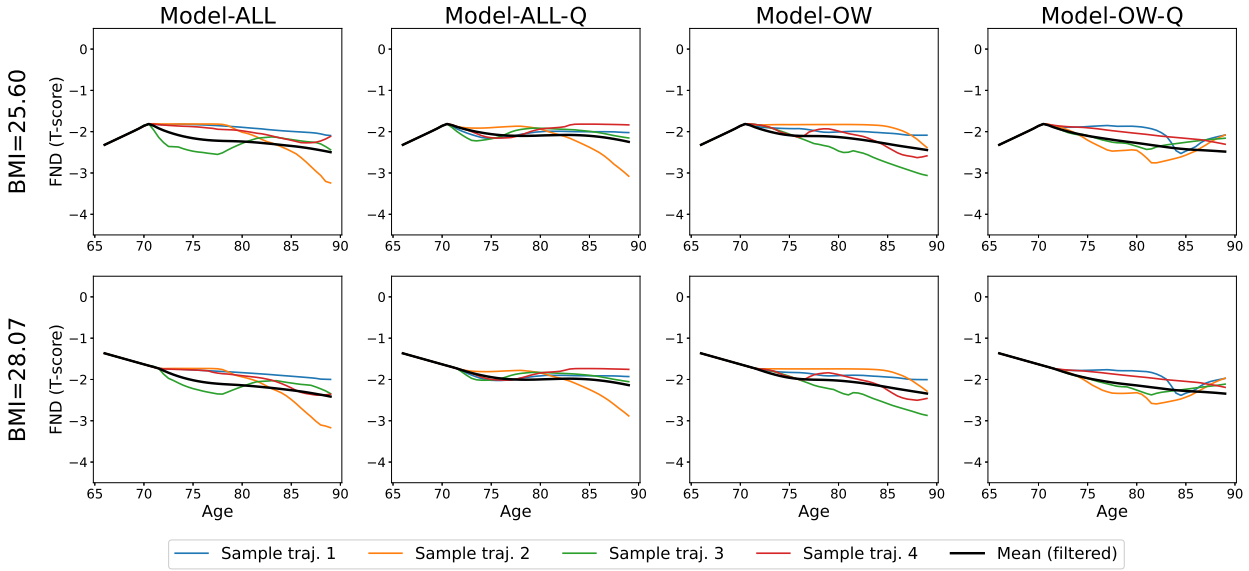


(b) Results for models using $\mathcal{J} = \{5, 8\}$.

Figure A.1: Predictions with models for two test participants from strata UWH. Each plot depicts the (filtered) mean personalized phenotype trajectory (black curve) and four (filtered) sample trajectories. The BMI of the participant is listed on the left of each row.

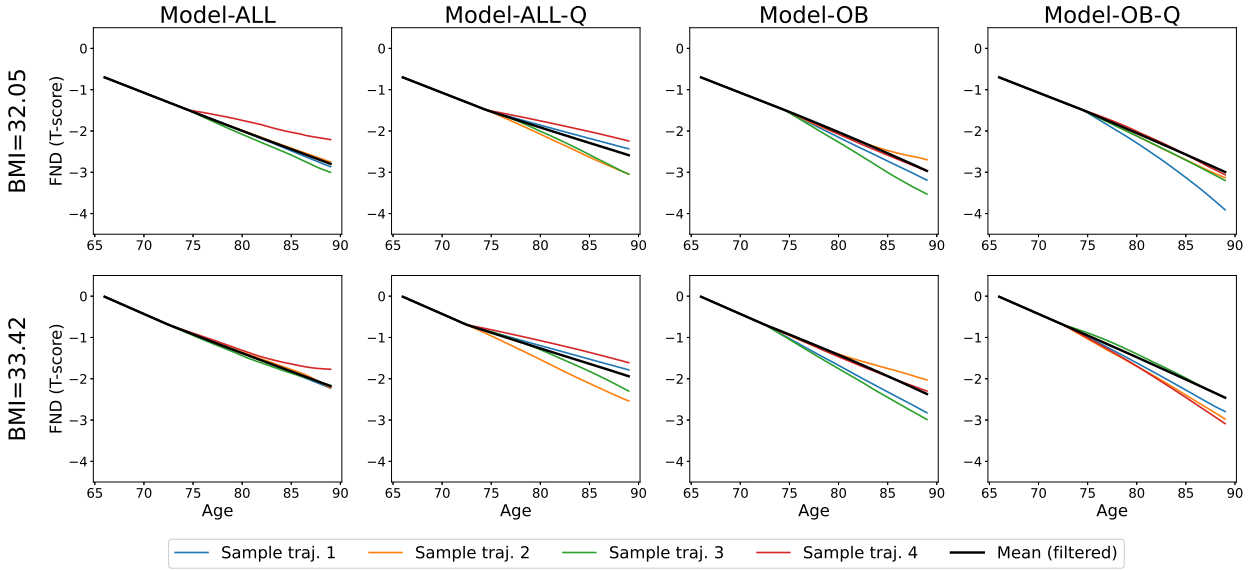


(a) Results for models using $\mathcal{J} = \{5\}$.

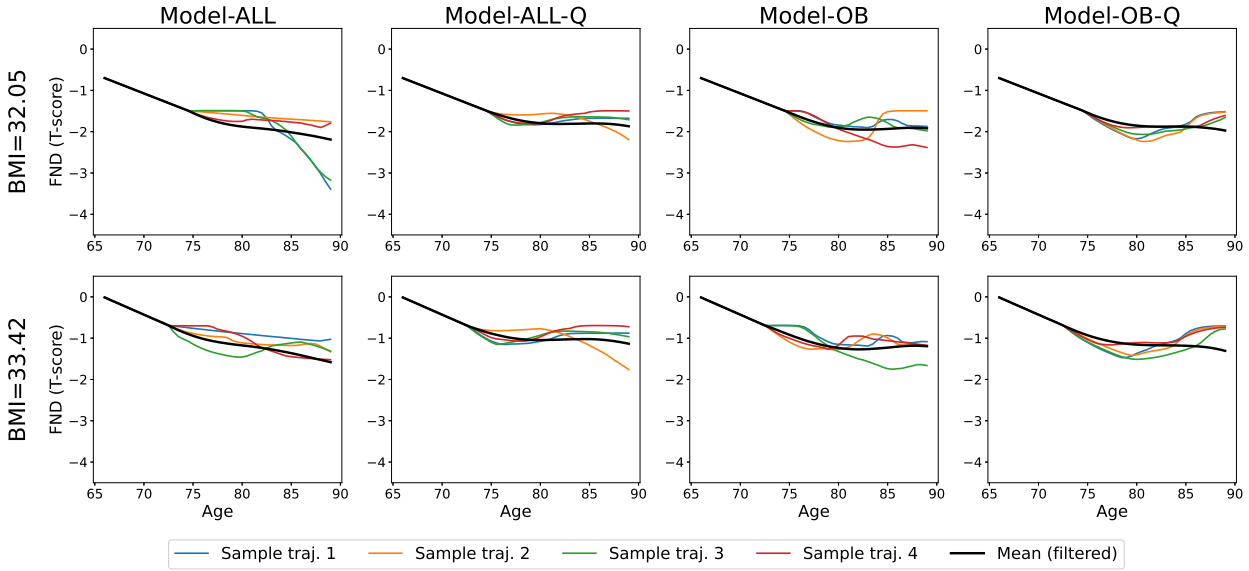


(b) Results for models using $\mathcal{J} = \{5, 8\}$.

Figure A.2: Predictions with models for two test participants from strata OW. Each plot depicts the (filtered) mean personalized phenotype trajectory (black curve) and four (filtered) sample trajectories. The BMI of the participant is listed on the left of each row.



(a) Results for models using $\mathcal{J} = \{5\}$.



(b) Results for models using $\mathcal{J} = \{5, 8\}$.

Figure A.3: Predictions with models for two test participants from strata OB. Each plot depicts the (filtered) mean personalized phenotype trajectory (black curve) and four (filtered) sample trajectories. The BMI of the participant is listed on the left of each row.

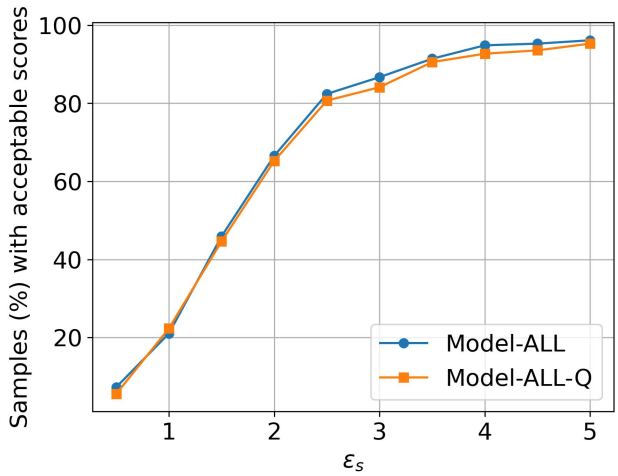


Figure B.1: Ablation study to understand the effect of ϵ_s in selecting samples with an acceptable score. The results are shown for Model-ALL (blue curve) and Model-ALL-Q (orange curve), with the y-axis indicating the percentage of test samples flagged as have a suitable score, i.e., $\mathcal{S}_M \leq \epsilon_s$.

Appendix B Ablation study: choosing ϵ_s and ϵ_f

To determine a suitable value of ϵ_s which is used to define the acceptable scores, we study how the percentage of test samples with an acceptable score, i.e., $\mathcal{S}_M \leq \epsilon_s$, changes as ϵ_s is increased. Figure Figure B.1 plots this change with Model-ALL and Model-ALL-Q (with $\mathcal{J} = \{5, 8\}$), both of which are trained on the same unstratified dataset. We observe a sharp transition near $\epsilon_s = 2.5$ at which point about 80% of the test sample predictions have acceptable scores. Beyond this threshold, there is a very gradual change in the influence of ϵ_s over the acceptable scores of test samples. Thus, we choose $\epsilon_s = 2.5$ in our experiments.

Next, we determine a suitable value for ϵ_f which is used to filter the samples trajectories on individuals based on DTW distance. In Figure Figure B.2, we plot the fraction of trajectories predicted (by Model-ALL-Q) that are retained for various values of ϵ_f. The six colored solid dotted lines show the fractions for the same six test participants considered in the previous section, while the black dashed dotted line shows the median fractions for each value of ϵ_f considered. We observe a sharp transition near $\epsilon_f = 2.0$, at which roughly 88% of trajectories are retained. The fraction increase more gradually beyond this value of ϵ_f. This motivates our choice for $\epsilon_f = 2.0$ in our experiments.

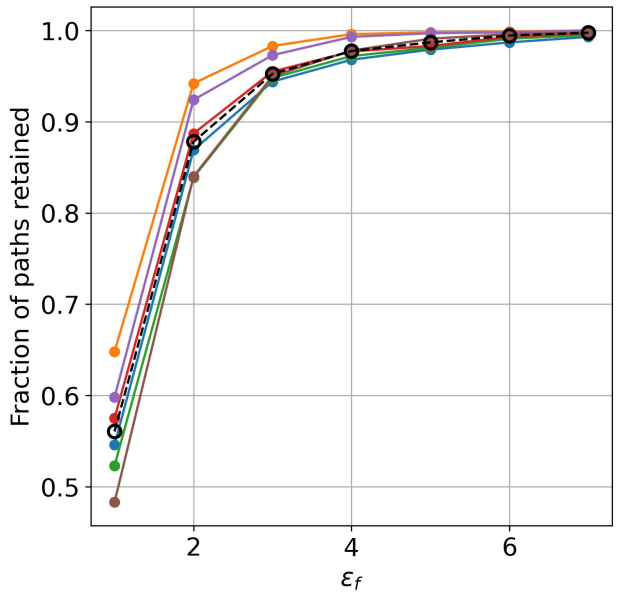


Figure B.2: Ablation study to understand the effect of ϵ_f in filtering trajectories based on the DTW distance for different individuals. The colored solid dotted lines show the fraction of predicted trajectories retained as ϵ_f is varied for six test participants for illustrative purposes. The black dashed dotted line shows the median fractions for each value of ϵ_f considered.