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

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

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

August 22, 2025

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 total hip Bone Mineral Density (BMD) of a healthy individual as the phenotype of interest, and baseline individual Body Mass Index (BMI) as the 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 98 years, for eight different combinations both with and without involving the covariate. Our results clearly demonstrate the potential of generative deep learning frameworks in healthspan research.

1 Introduction

Forecasting the trajectories of diseases has been a cornerstone of the field of epidemiology involving numerous approaches ranging from the classical such as SIR and related models to more recent ones

[†]These authors contributed equally to this paper.
*Corresponding author, Email: spyne@healthanalytics.net

based on time series and predictive analytics, genomics, etc. [28]. 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 [30]. Since the COVID-19 pandemic, novel approaches have emerged for outbreak forecasting including AI-driven deep learning of spatio-temporal patterns, prediction by real and synthetic data fusion, e.g., [36, 10]. 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.

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. [5] 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 just one 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.[24] Thus, osteoporosis acts as a key determinant of one's 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.[26] 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).[18] As individuals age, BMD decreases and, consequently, osteoporosis becomes more prevalent in older adults. [13, 16, 19] While several factors may be involved in the etiology of osteoporosis, historically, the body mass index (BMI) has been linked to bone health as a protective

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

factor.[12, 16] 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.[27] Their findings were later corroborated by other studies.[6, 8]

The longest running cohort study in the U.S. on this topic is the NIH-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 While initially the study enrolled white women, beginning in 1997, the study 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.[6] 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. ([25],[17]) 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, hip 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 [11, 33, 37]. 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.[3, 4, 31] Some have used GANs to predict such codes for multiple subsequent visits but not for tracking any actual phenotype of an individual.[32] Other deep learning applications have focused on specific health conditions or diseases, e.g., cystic fibrosis [20], Alzheimer's disease [9]. Measures of DNA methylation biomarkers that estimate epigenetic aging serve a somewhat distinct purpose.[15] 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.[34]

[†]https://research.kpchr.org/News/CHR-Stories/Landmark-Study-of-Osteoporotic-Fractures-Draws-to-a-Close-After-31-Years

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 [22]. cGANs have been successfully designed to impute corrupted images in dynamic contrast enhanced computed tomography [29] and probabilistic brain segmentation from magnetic resonance images [23].

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 total hip BMD of a healthy individual as the phenotype of interest and baseline individual BMI as the covariate. We used DyViA-GAN to generate 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 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 work with the data for 9,704 (primarily Caucasian women) participants in the SOF dataset, which were recorded over multiple clinic visits at four sites. 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 consider participant data from Visit 2 (1989-90), Visit 4 (1992-94), Visit 5 (1995-96) and Visit 8 (2002-04), and consider the total hip BMD (THD) measured from scans using Hologic QDR 1000 workstations. We use 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. These adjusted measurements are available in the SOF dataset. In total, we have 2,257 participant records which are further stratified based on the 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. Moving forward, we use the following notations to denote the variables serving as input or predicted outputs of our deep learning model: THD_i is the THD value of a participant at visit $i \in 2, 4, 5, 8$, t_i is the (scaled) age of the participant during the i-th visit, and $Q \in \mathbb{R}^q$ is a vector of q baseline parameters or risk factors of the participant. For the present study, we chose the BMI (scaled by 100) to represent Q, with $q = 1$. We note that the age of the participants is scaled according to

$$t \to \frac{t - t_{\min}}{t_{\max} - t_{\min}}, \quad t_{\min} = 66,\ t_{\max} = 98, \tag{1}$$

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 (1) ensures that $t \in [0, 1]$. 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.

Table 1: Description of data 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. THD	Max. THD	Mean THD (± SD)
Training	UWH	900	0.25	1.28	0.71 (± 0.12)
	OW	800	0.30	1.66	0.77 (± 0.12)
	OB	400	0.36	1.21	0.83 (± 0.13)
Test	UWH	66	0.39	1.03	0.72 (± 0.12)
	OW	75	0.43	1.31	0.76 (± 0.12)
	OB	16	0.57	1.12	0.81 (± 0.14)

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 $\mathcal{D}$. The generator is fed as input the vector $Y = (\text{THD}_2, \text{THD}_4, t_2, t_4, Q) \in \mathbb{R}^{4+q}$, and a variable $t \in [0, 1]$ denoting the (scaled) age at which we want to predict the value of THD given Y. We denote the predicted value as $\widehat{\text{THD}}(t, Y, Z)$. In addition, we feed as input a random latent variable $Z \in \mathbb{R}^{N_z}$ which is sampled from an uncorrelated standard Gaussian distribution. Thus, the generator is mapping of the form $\mathcal{G} : (t, Y, Z) \in \mathbb{R}^{4+q+N_z} \times [0, 1] \times \mathbb{R}^{N_z} \mapsto \mathbb{R}$.

The discriminator $\mathcal{D}$ is a scalar-valued network that is fed as input either a "true" data tuple (X, Y, t_5, t_8) where $X = (\text{THD}_5, \text{THD}_8) \in \mathbb{R}^2$ or a "fake" tuple $(\widehat{X}, Y, t_5, t_8)$ where $\widehat{X} = \big(\widehat{\text{THD}}(t_5, Y, Z), \widehat{\text{THD}}(t_8, Y, Z)\big) \in \mathbb{R}^2$ is generated using $\mathcal{G}$. Thus, the discriminator is a mapping of

the form $\mathcal{D} : (X, Y, t_5, t_8) \in \mathbb{R}^2 \times \mathbb{R}^{4+q} \times [0,1]^2 \mapsto \mathbb{R}$. 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_5, t_8), i.e., $P_{X(\mathsf{Y,t_5,t_8})}$"\$.

The two network are trained in an adversarial fashion by solving the following optimization problem associated with a least-squares GAN [21]

$$\begin{aligned} \mathcal{D}^* &= \arg\min_{\mathcal{D}} \mathop{\mathbb{E}}_{P_{X,Y,t_5,t_8}} \left[\left(\mathcal{D}(X, Y, t_5, t_8) - 1 \right)^2 \right] + \mathop{\mathbb{E}}_{P_{\widehat{X},Y,t_5,t_8}} \left[\left(\mathcal{D}(\widehat{X}, Y, t_5, t_8) \right)^2 \right] \\ \mathcal{G}^* &= \arg\min_{\mathcal{G}} \mathop{\mathbb{E}}_{P_{\widehat{X},Y,t_5,t_8}} \left[\left(\mathcal{D}(\widehat{X}, Y, t_5, t_8) - 1 \right)^2 \right] \end{aligned} \tag{2}$$

where P_{X,Y,t_5,t_8} is the joint probability distribution of the true data, while $P_{\widehat{X},Y,t_5,t_8}$ 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 THD 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 $\mathcal{G}^*(t, Y, Z^{(k)})$ for $1 \leq k \leq K$. By allowing t to vary in $[0,1]$, we can generate an ensemble of plausible trajectories of the temporal values of THD for an individual. We can then compute the point-wise empirical mean and variance of the predicted trajectories (given a Y) for any $t \in [0,1]$ as

$$\overline{\mathcal{G}}^*(t, Y, Z) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{G}^*(t, Y, Z^{(k)}), \quad \mathrm{Var}[\mathcal{G}^*](t, Y, Z) = \frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{G}^*(t, Y, Z^{(k)}) - \overline{\mathcal{G}}^*(t, Y, Z) \right)^2 \tag{3}$$

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

2.3 Network architectures

The generator $\mathcal{G}$ is taken to be a fully-connected network with input dimension $4 + N_z + 1$ (or $4 + N_z + q$ if covariates Q are used) with a scalar-valued output. We used four hidden layers of width 50 each with the hat activation function [14, 35]. A softplus output function is used to ensure the predicted outputs are non-negative. Furthermore, to ensure that the generator's prediction is constrained to be close to the given initial measurements at t_2 and t_4, the output of the generator is transformed as follows

$$\mathcal{G}(t, Y, Z) \to \mathcal{G}(t, Y, Z) \times (1 - \phi(t;\ t_2) - \phi(t;\ t_4)) + \phi(t;\ t_2) \times \mathrm{THD}_2 + \phi(t;\ t_2) \times \mathrm{THD}_4 \tag{4}$$

where $\phi(t;\ \tilde{t}) = \exp(-(t - \tilde{t})^2/(2\sigma))$ is a smooth function that takes takes the value 1 at $t = \tilde{t}$ and exponentially decays to 0 away from $\tilde{t}$, with the rate of decay controlled by σ. Thus, with the transformation Equation 4, the output of the generator is forced to be close to THD_2 and THD_4

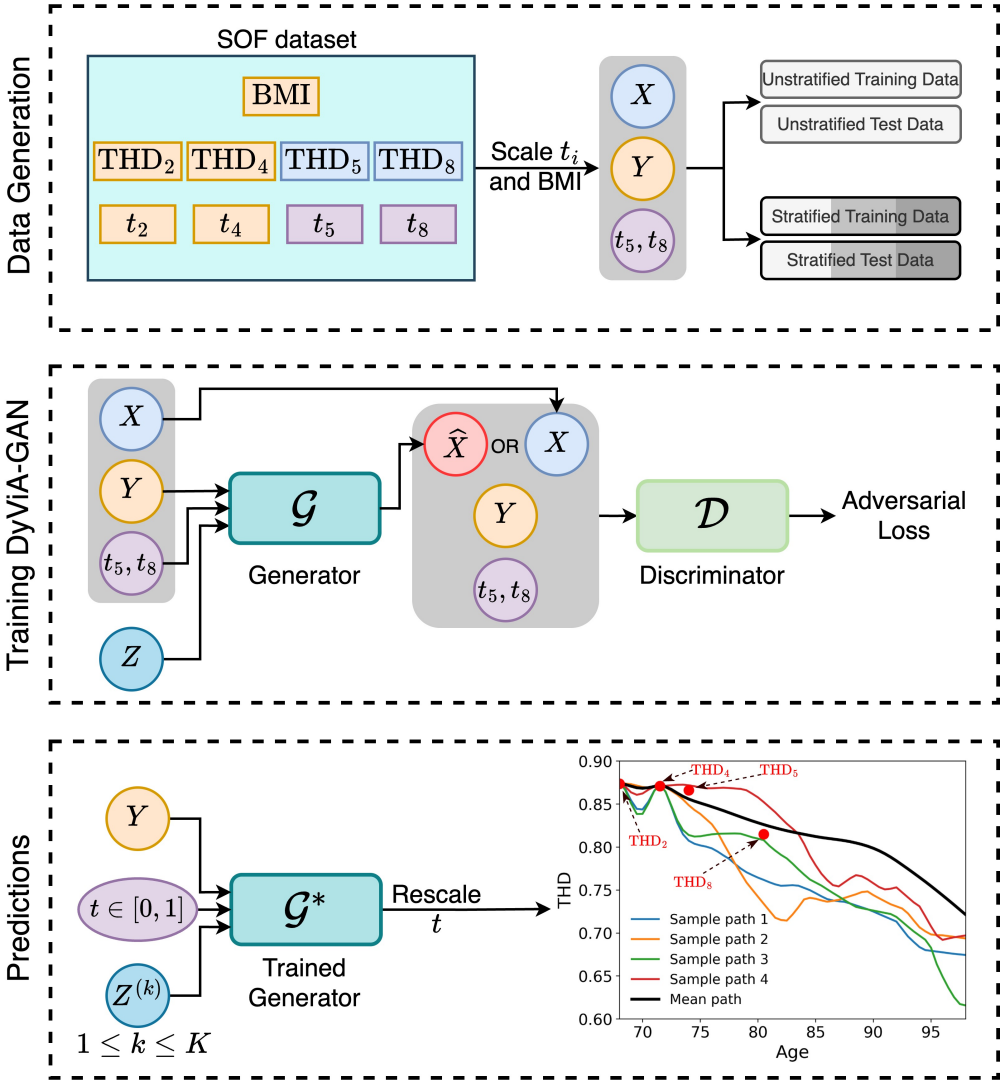

Figure 1: Schematic depicting the construction of the datasets, the inputs and outputs of $\mathcal{G}$ and $\mathcal{D}$ when training the DyViA-GAN, and how the trained generator $\mathcal{G}^*$ is used to predict plausible phenotype trajectories. **Top Panel:** From the SOF dataset, the THD and ages at during visits 2,4,5,8, along with the BMI measured at visit 1, are extracted for all participants without gaps in the data. After scaling the ages and BMI, unstratified as well as stratified training/test data tuples of the form (X, Y, t_5, t_8) are created. **Middle Panel:** During training phase, Y, Z and the scaled ages t_5, t_8 are fed to $\mathcal{G}$ to obtain the prediction vector $\widehat{X}$ at t_5, t_8. The discriminator $\mathcal{D}$ is fed either a true tuple (X, Y, t_5, t_8) or a fake tuple $(\widehat{X}, Y, t_5, t_8)$. Both network are trained simultaneously according to Equation 2. **Bottom Panel:** The trained generator $\mathcal{G}^*$ is fed the conditional data Y for a participant, any $t \in [0, 1]$, and K samples of the latent variable. The output is an ensemble of K personalized trajectories (4 sample trajectories are shown) and the mean trajectory (black curve), which can be compared to the known phenotype measurements (red dots). We observe that all predicted trajectories pass through the true data at t_2, t_4 (due to the Equation 4) and at least one trajectory passes close to the data at future time t_5 (and similarly t_8).

when $t = t_2$ and $t = t_4$, respectively. In this work, we choose $\sigma = 10^{-3}$.

The discriminator $\mathcal{D}$ is a also taken to be fully-connected network with input dimension 8 (or $8 + q$ if covariates Q are used) 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.

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: 90, OW: 80, OB: 40
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.

3 Results

Using the trained generator $\mathcal{G}^*$, we quantify the performance of the various models by considering the score $\mathcal{S}$ for each training/test tuple (X, Y, t_5, t_8), which is defined as follows using empirical

Table 3: Percentage (%) of samples with acceptable scores ($\mathcal{S} \leq 1.5$) for 8 different DyViA-GAN models. 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.

	Training				Test			
Model \ Strata	UWH	OW	OB	ALL	UWH	OW	OB	ALL
Model-ALL	91.44	88.75	84.75	89.14	89.39	89.33	81.25	87.82
Model-ALL-Q	82.00	78.00	67.75	77.76	86.36	82.67	68.75	81.60
Model-UWH	68.44	-	-	-	68.18	-	-	-
Model-UWH-Q	77.78	-	-	-	72.73	-	-	-
Model-OW	-	83.75	-	-	-	81.33	-	-
Model-OW-Q	-	75.88	-	-	-	82.67	-	-
Model-OB	-	-	74.25	-	-	-	75.00	-
Model-OB-Q	-	-	77.75	-	-	-	75.00	-

statistics (see Equation 3)

$$\mathcal{S}(X, Y, t_5, t_8) = \max\left\{ \frac{\mathsf{THD}_5 - \overline{\mathcal{G}}^*(t_5, Y, Z)}{\mathrm{Var}[\mathcal{G}^*](t_5, Y, Z)} \ , \ \frac{\mathsf{THD}_8 - \overline{\mathcal{G}}^*(t_8, Y, Z)}{\mathrm{Var}[\mathcal{G}^*](t_8, Y, Z)} \right\} \tag{5}$$

with the empirical statistics computed using $K = 1000$ random samples of Z. The score evaluates the normalized distance of the true values of the phenotype at $t = t_5, t_8$ from the mean predicted values. Thus, predictions with lower scores are preferable. We say that the *score is acceptable* for a sample if $\mathcal{S} \leq 1.5$. With this notion, we evaluate the percentage of training and test samples with acceptable scores, which are listed in Table 3. 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 scores of the other models trained on stratified data are only evaluated on the corresponding strata. We make the following observations based on the values listed in Table 3:

- Models trained on the unstratified data perform better than models trained on stratified data. Model-ALL and Model-ALL-Q have significantly larger percentages of training and test samples with acceptable scores ($\mathcal{S} \leq 1.5$) when compared to the other models. This is not very surprising as we have significantly fewer samples when working with stratified training data, especially for the OB strata (see Table 1).

- Feeding the covariate Q (i.e., the BMI) as a generator input deteriorates the performance of the models trained on unstratified data (compare the first two data rows of Table 3). The percentage of training samples with acceptable scores drops from 89% to 78%, while the percentage of test samples with acceptable scores drops from 88% to 82%.

- Feeding the BMI as a generator input appears to improve the performance of the models trained on stratified data, especially for the UWH and OB strata.

- Although not presented here, we observed that in general the predictions at t_8 (as compared to t_5) have larger contributions towards samples having a score $\mathcal{S} > 1.5$, i.e., having unacceptable scores.

We also demonstrate the qualitative performance of the trained models on six test participants chosen to cover the entire spectra of BMI values in the dataset. We show the mean predicted personalized trajectory (with $K = 1000$) using models trained on unstratified data in Figure 2, and using models trained on stratified data in Figure 3. For each test sample, we also plot the true measurements $\text{THD}_2, \text{THD}_4, \text{THD}_5, \text{THD}_8$ (red dots), and the region with acceptable scores (gray shading). In addition, we have access to the adjusted phenotype value at Visit 7 (1999) for these six test participants, which is depict in the plots (blue triangle). Note that the generator and discriminators never see the phenotype value for this visit during the training phase of DyViA-GAN. The Appendix contains additional plots depicted a few sample trajectories for each of these test samples. We make the following observations from the plots in Figure 2 and Figure 3:

- The mean trajectories pass through the data at t_2, t_4 with almost zero width of the acceptable regions at these time points. This demonstrates the effect of the constraint Equation 4 applied on the generator output, indicating the high certainty of the trajectory at the initial measurements.

- With all models, the new unseen data at future time point t_7 (blue triangle) falls within the region of acceptable score for all six test participants. In other words, the mean trajectory value is close enough to this data point.

- For models trained on unstratified data (Figure 2), acceptable score region in gray does a better job at enveloping the true phenotype measurements at future time points t_5, t_7 (red dots) when BMI is not used as a covariate. This is most prominent for test participant 5 in row 3 (with BMI = 37.76) where the THD_8 measurement lies significantly outside the gray region when BMI is used as a covariate.

- The predictions with models trained on stratified data (Figure 3) appear to have narrower regions of acceptable scores about the mean predicted path, when compared to the regions

for models trained on unstratified data (see Figure 2). This indicates that the predicted paths have a narrower spread, which also explains why a larger percentage of samples have unacceptable scores with the stratified models (see Table 3).

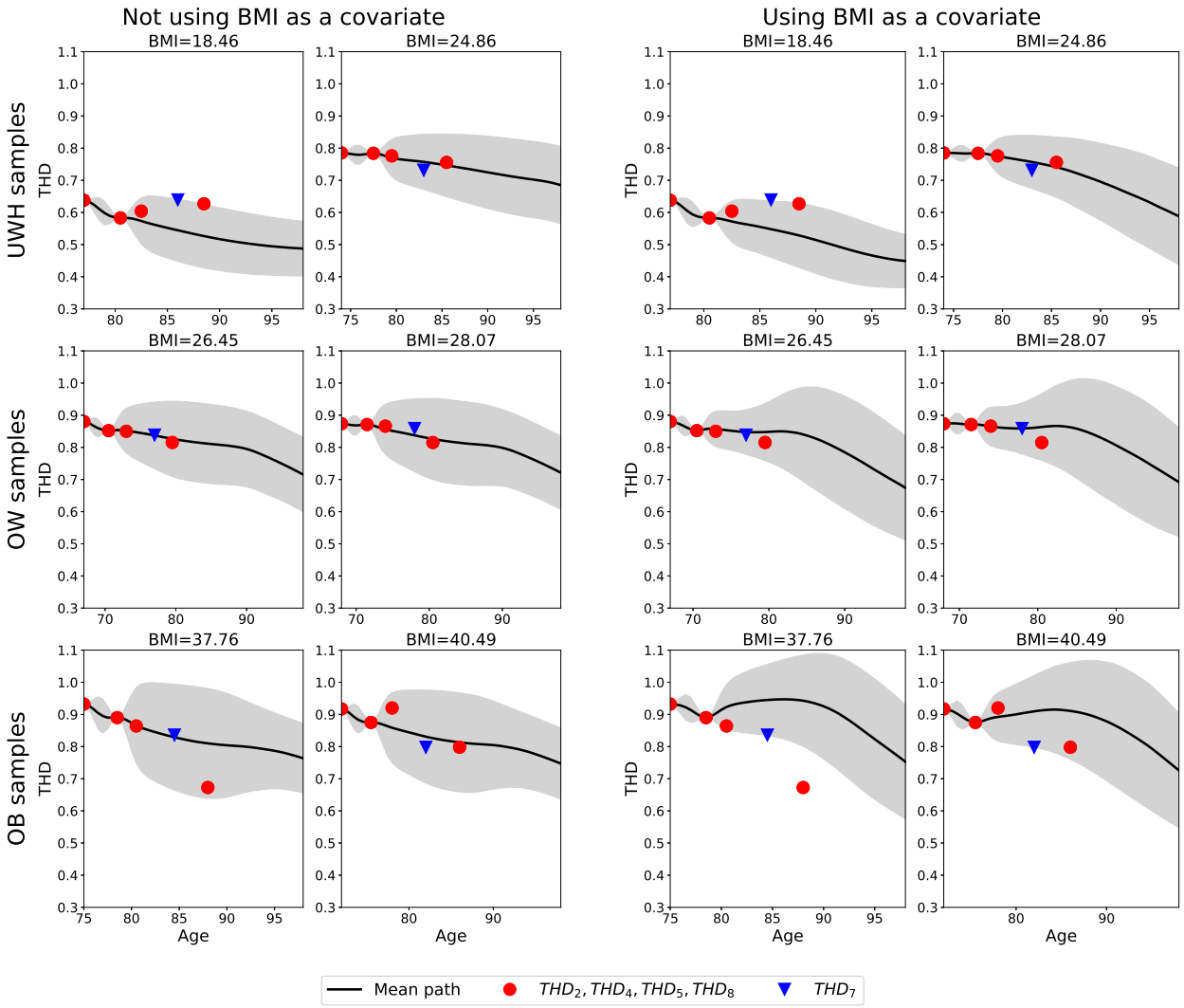

Figure 2: Predictions with Model-All (left panel) and Model-ALL-Q (right panel) for two test participants each from strata UWH (row 1), OW (row 2) and OB (row 3). Each plot depicts the mean personalized phenotype trajectory (black curve), the region of acceptable scores (gray shading), the phenotype measurements at times t_2, t_4, t_5, t_8) (red dots), and an additional measurement at t_7 (blue triangle) which was never seen during the training phase of DyViA-GAN. The BMI of the participant is listed on the top of each plot.

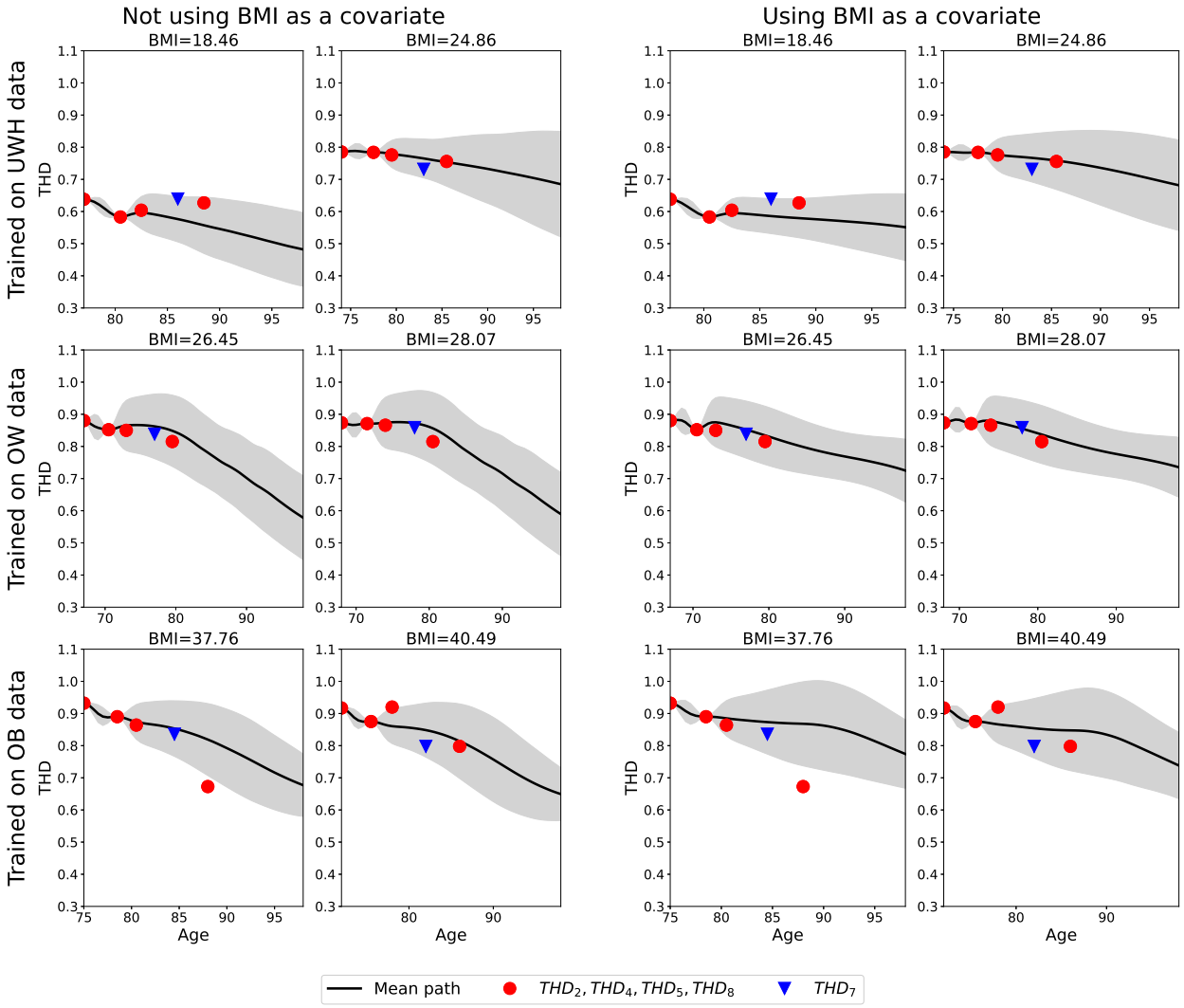

Figure 3: Predictions with Model-⟨DT⟩(left panel) and Model-⟨DT⟩-Q (right panel) for two test participants each from strata UWH for ⟨DT⟩=UWH (row 1), OW for ⟨DT⟩=OW (row 2) and OB for ⟨DT⟩=OB (row 3). Each plot depicts the mean personalized phenotype trajectory (black curve), the region of acceptable scores (gray shading), the phenotype measurements at times t_2, t_4, t_5, t_8) (red dots), and an additional measurement at t_7 (blue triangle) which was never seen during the training phase of DyViA-GAN. The BMI of the participant is listed on the top of each plot.

4 Discussion

The present study has 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 – total hip 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.

Notably, there are several key findings from the results presented in section 3. First, the quality of the model predictions measured in terms of the score $\mathcal{S} \leq 1.5$, i.e., the proximity of the personalized mean trajectory to true phenotype measurements, are better if more samples are considered to train the model. Second, the inclusion of additional risk factors, such as BMI, may lead to better personalized predictions in situations where training data is stratified. However, such an inclusion may have a negative effect on the performance on test data if unstratified training data is used. Third, the mean trajectory with all eight versions of the DyViA-GAN considered in this study are close (in terms of the score function) to the true measurement of the phenotype from visit 7, which was never shown to the model during training.

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.[6]

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 other regions of the body. Moreover, 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 or chronic inflammation or epigenetic changes. The fact that the presently generated trajectories are not dynamically updated could be revised in the future to allow for intermediate events, lifestyle alterations, and medical interventions. Further, we plan to design phenotype trajectory predictors using other deep generative frameworks, such as conditional diffusion models [2, 7] which have been shown to lead to better predictions compared to cGANs in certain Bayesian inference applications.

Acknowledgments

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

References

[1] Dataset: Study of osteoporotic fractures (SOF). URL `https://agingresearchbiobank.nia.nih.gov/studies/sof/details`.

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

[3] 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.

[4] 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`.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] D. Guh, W. Zhang, N. Bansback, Z. Amarsi, C. Birmingham, and A. Anis. The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis. *BMC Public Health*, 9, 2009. doi: 10.1186/1471-2458-9-88. URL `10.1186/1471-2458-9-88`.

[13] 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.

[14] 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`.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] S. Moazami, D. Ray, D. Pelletier, and A. A. Oberai. Probabilistic brain extracation in MR images via conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 2023.

[24] L. J. T. K. Moreland, B.L. 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*, 35:345–355, 2022.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] R. Raad, D. Ray, B. Varghese, D. Hwang, I. Gill, V. Duddalwar, and A. A. Oberai. Conditional generative learning for medical image imputation. *Scientific Reports*, 14(1):171, 2024.

[30] R. C. Rao AS, Pyne S, editor. *Handbook of Statistics: Disease Modelling and Public Health (Parts A and B)*, volume 36. Elsevier, 2017.

[31] 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.

[32] 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.

[33] 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. Ieee, 2018.

[34] 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.

[35] 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.

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

[37] 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.

A Additional results

As a companion to the six test samples considered in Figure 2 and Figure 3 of the main text, we plot a few sample trajectories predicted by DyViA-GAN for each of these test samples. These are shown in Figure 4 and Figure 5 below.

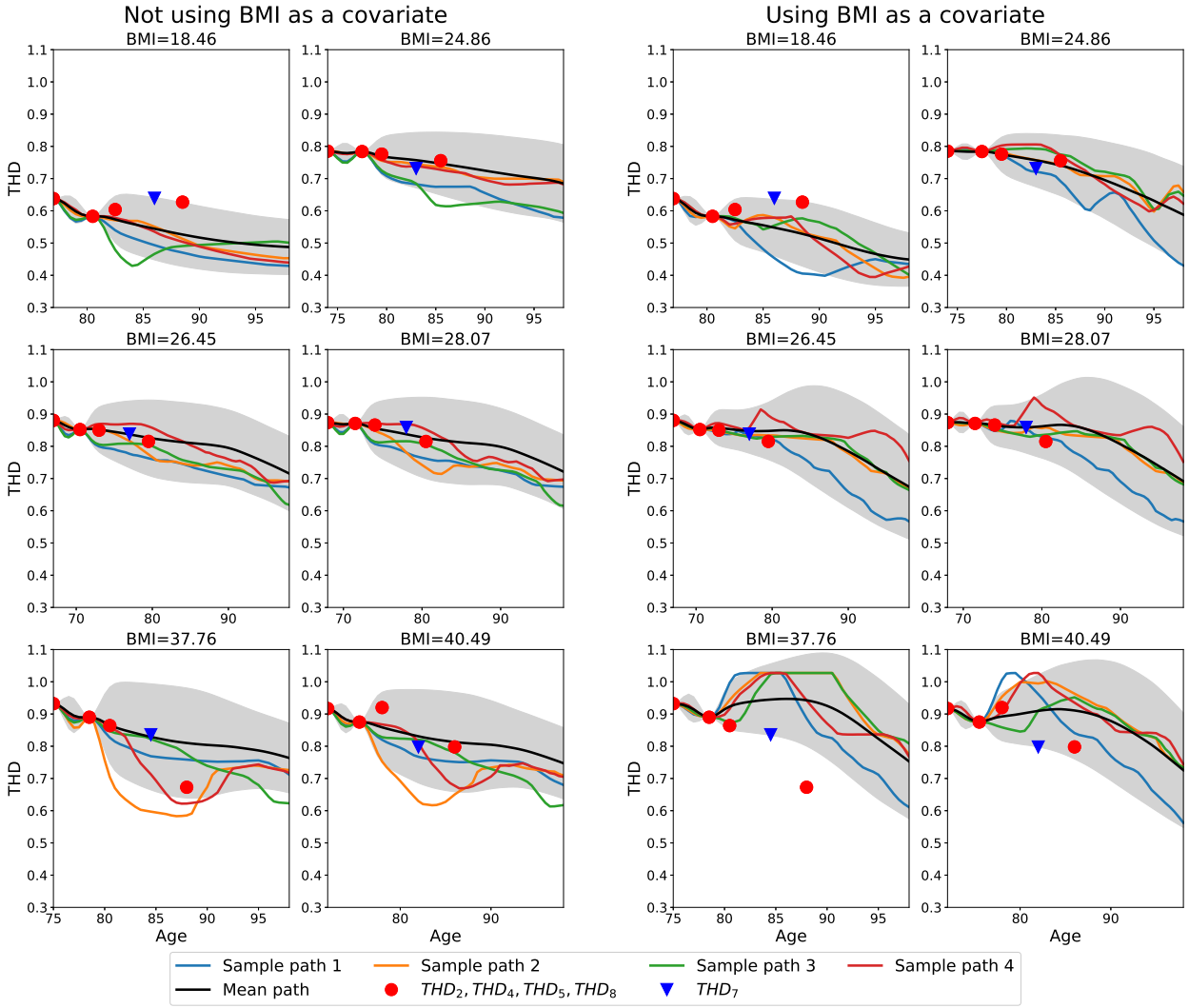

Figure 4: Predictions with Model-All (left panel) and Model-ALL-Q (right panel) for two test participants each from strata UWH (row 1), OW (row 2) and OB (row 3). Each plot depicts the mean personalized phenotype trajectory (black curve), four sample trajectories, the region of acceptable scores (gray shading), the phenotype measurements at times t_2, t_4, t_5, t_8) (red dots), and an additional measurement at t_7 (blue triangle) which was never seen during the training phase of DyViA-GAN. The BMI of the participant is listed on the top of each plot.

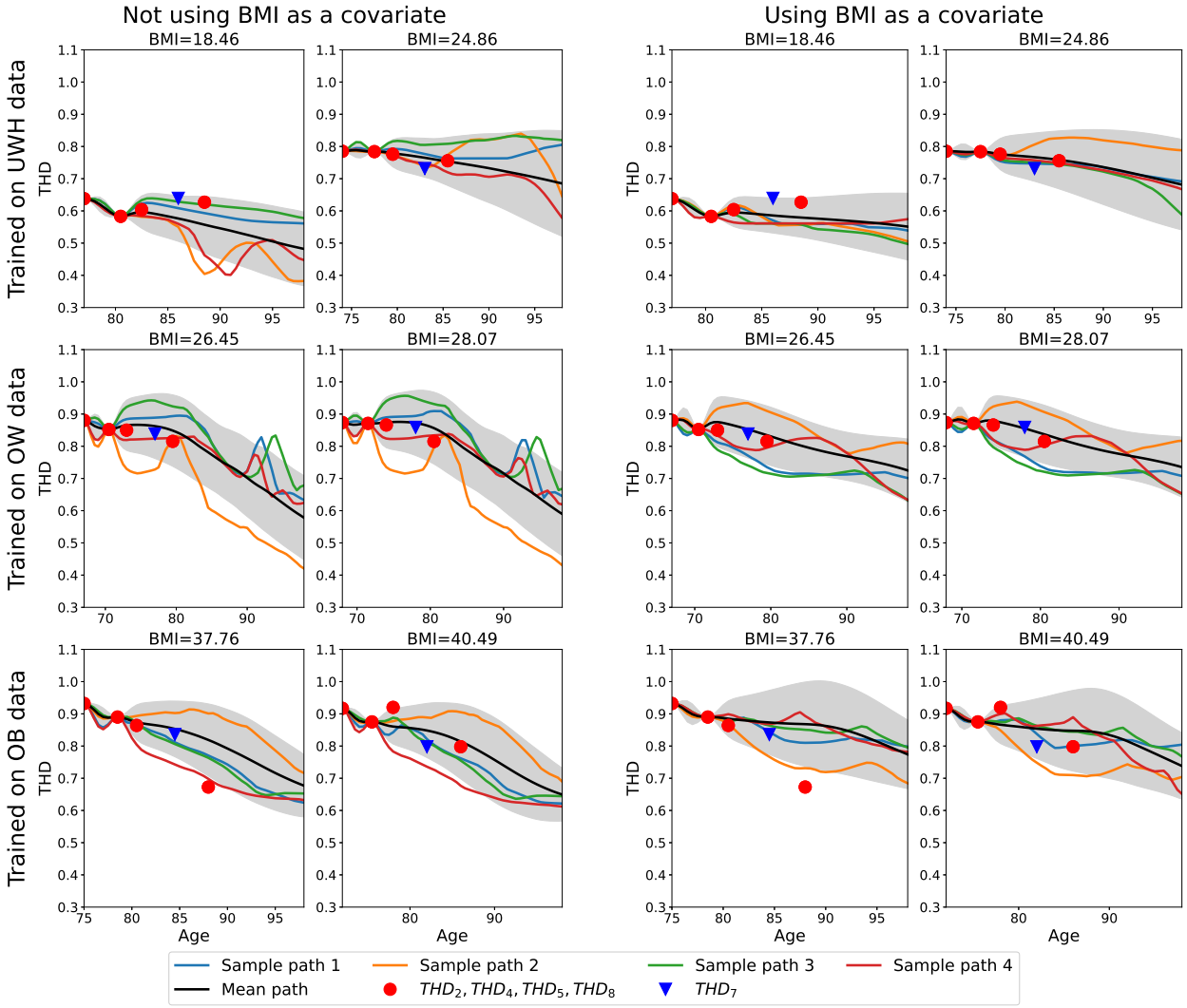

Figure 5: Predictions with Model-⟨DT⟩(left panel) and Model-⟨DT⟩-Q (right panel) for two test participants each from strata UWH for ⟨DT⟩=UWH (row 1), OW for ⟨DT⟩=OW (row 2) and OB for ⟨DT⟩=OB (row 3). Each plot depicts the mean personalized phenotype trajectory (black curve), four sample trajectories, the region of acceptable scores (gray shading), the phenotype measurements at times t_2, t_4, t_5, t_8) (red dots), and an additional measurement at t_7 (blue triangle) which was never seen during the training phase of DyViA-GAN. The BMI of the participant is listed on the top of each plot.