Centroid of Age Neighborhoods: A New Approach to Estimate Biological Age

Syed Ashiqur Rahman and Donald A. Adjeroh, *Member, IEEE*

***Abstract*—Estimation of human biological age is an important and difficult challenge. Different biomarkers and numerous approaches have been studied for biological age prediction, each with its advantages and limitations. In this paper, we propose a new biological age estimation method, and investigate the performance of the new method. We introduce a centroid based approach, using the notion of age neighborhoods. Specifically, we develop a model, based on which we compute biological age using blood biomarkers, by considering the centroid or mediod of specially selected age neighborhoods. Experiments were performed on the National Health and Human Nutrition Examination Survey dataset with biomarkers (21 451 individuals). Compared with current popular methods for biological age prediction, our experiments show that the proposed age neighborhood model results in an improved performance in human biological age estimation.**

***Index Terms*—Age estimation, aging, bio-markers, biological age, age centroid, age medoid, all-cause mortality.**

I. Introduction

HUMAN age estimation is an important problem that has witnessed an increased attention, given its role in various daily activities, from health assessment, to social interaction, to security and identity profiling. Although age estimation has been practiced for centuries, accurate age estimation is known to be a difficult problem. Doing this automatically by a machine is an even more onerous task [1], [2]. The major challenge is that most of the measures used to characterize age, for instance, visual appearance, and biological/physiological markers vary significantly from person to person, even for people of the same chronological age.

Age has a deep connection with health and mortality [3]–[5]. Aging is a gradual process that results in increased health risk, and mortality over time. In general, a younger person is expected to have a better health condition and his/her mortality hazard should be low in comparison with a relatively older person. But two different people of the same age may have very different health conditions and mortality hazards. This brings up the debate on "chronological " versus "biological" age. Chronological age is typically what we know and is based on the date of birth. Chronological age estimation from face image [2] is most popular. However, biological age is based on the interesting, yet confounded, idea that a person's true age can be different from his/her chronological age. Biological age (sometimes called functional age [6]) lacks a precise definition, but it is often viewed as the true age of an individual in the gerontology and aging research community [7]. The common idea is that, biological age provides a better estimator of the true life expectancy of the individual than his or her chronological age. Quantification of biological age is a difficult challenge, since there is no well defined criteria. To estimate biological age, some age-dependent variables are used [8]–[10], and chronological age may or may not be a required attribute/variable depending on the application.

Klemera and Doubal's approach [7] is the most popular, and perhaps, the most effective biological age estimation method [3]–[5]. The biological age (BA) estimates are derived based on minimizing the distance between biomarker points and regression lines. Other approaches include multiple linear regression (MLR) [3], and combination of MLR with principal component analysis (PCA) features [3]. Levine [3] compared the performance of five BA estimation algorithms in terms of their ability to predict mortality. Klemera and Doubal's (KD) method was found to be the most reliable predictor for mortality. Overall, the performance of biological age (BA) in mortality prediction was significantly better than using chronological age (CA). Cho *et al.* [10] studied various BA estimation methods to examine the relation with work ability index (WAI). WAI is a measure that reflects present health condition rather than how it changes with age and their analysis showed that the KD method on PCA features produced the most reliable results. Mitnitski *et al.* [5] compared the performance of the frailty index (FI) with biomarker-based measures of BA. They employed the KD algorithm in predicting mortality. Belsky *et al.* [4] described biological age as a reflection of ongoing longitudinal change within a person. They estimated the BA for subjects at age 38 using the Klemera-Doubal equation with parameters estimated from the NHANES-III dataset. The study also tested the hypothesis that young adults with older biological age at age 38 years were actually aging faster than those with a younger biological age. They analyzed within-individual longitudinal change in 18 biomarkers from the Dunedin Study [11] across chronological ages 26 y, 32 y, 38 y to quantify each study member's personal rate of physiological deterioration. Cole *et al.* [12] studied the use of structural neuro-imaging such as MRI under a

Manuscript received December 15, 2018; revised June 14, 2019; accepted July 13, 2019. Date of publication July 24, 2019; date of current version April 6, 2020. This work was supported in part by the National Science Foundation under Grant 1636933, and the National Institute of General Medical Sciences of the National Institutes of Health under Grant 5U54GM104942-03. *(Corresponding author: Donald A. Adjeroh.)*
The authors are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26505 USA (e-mail: srahman2@mix.wvu.edu; don@csee.wvu.edu).
Digital Object Identifier 10.1109/JBHI.2019.2930938

Gaussian process regression framework to estimate biological age. The predicted age was identified as "brain-predicted age" or "brain age" for short. They combined DNA-methylation with brain age and showed that the combination improved mortality risk prediction. However, combining brain age with grey matter and cerebrospinal fluid volumes did not improve mortality risk prediction. Bobrov *et al.* [13] proposed a deep learning based model (called PhotoAgeClock) to estimate chronological age using images of eye corners.

Other methods that have been used to assess biological age or speed of aging includes handgrip strength [14], locomotor activity [15], [16], and deep learning on biomarker data [17]. There is no clear consensus on which method is best for BA estimation, nor on how best to quantify the BA itself. Although KD method is the most popular, it is limited to biomarker features. Thus, biological age estimation in humans remains a significant challenge and evaluating the estimated BA is still a difficult problem. In this work, we introduce a novel centroid based method to estimate biological age.

First, we propose a basic algorithm to estimate biological age using the centroid of selected age neighborhoods. Then we show a more refined algorithm that considers the distribution of the features. Similarly, we show that this approach can be used for medoid and inter quartile range. We then show three approaches to quantify the performance of the estimated biological age and compare with the state of the art biological age estimation methods. The paper is organized as follows: in Section II we describe the data set used and the proposed methodology. Section III shows the experimental results, and in Section IV we provide a discussion on the proposed approach and compare with other related methods. Section V draws some conclusions on the proposed work.

TABLE I
KEY BIOMARKER ATTRIBUTES FOR STUDY PARTICIPANTS IN THE NHANES DATASET, ALONG WITH THEIR CORRELATION WITH CHRONOLOGICAL AGE USING DIRECT MEASUREMENTS FOR BOTH PEARSON'S ρ, AND KENDALL'S τ

		Correlation with Age	
Biomarkers (N=21451)	**Average** $\pm$ **SD**	ρ	τ
C-reactive protein (mg/dL)	0.38 $\pm$ 0.78	0.09	0.13
Glycated hemoglobin $(\%)$	5.50 $\pm$ 0.91	0.33	0.35
Serum Albumin (ug/mL)	4.26 $\pm$ 0.37	-0.17	-0.15
Total Cholesterol (mg/dL)	197.77 $\pm$ 42.27	0.22	0.19
Serum Urea Nitrogen (mg/dL)	13.32 $\pm$ 5.55	0.44	0.31
Serum Alkaline Phosphatase (U/L)	74.75 $\pm$ 28.00	0.06	0.06
Systolic blood pressure $(mmHg)$	123.88 $\pm$ 20.02	0.54	0.38
Diastolic blood pressure $(mmHg)$	70.09 $\pm$ 13.05	0.07	0.11
Pulse $(60sec)$	72.12 $\pm$ 12.21	-0.14	-0.08
High density lipoprotein (mg/dL)	53.36 $\pm$ 15.91	0.05	0.02
Hemoglobin (g/dL)	14.26 $\pm$ 1.53	-0.07	-0.04
Lymphocyte percent $(\%)$	30.60 $\pm$ 8.71	-0.10	-0.07
White blood cell count (SI)	7.09 $\pm$ 2.42	-0.06	-0.04
Hematocrit $(\%)$	42.00 $\pm$ 4.39	-0.06	-0.03
Red blood cell count (SI)	4.68 $\pm$ 0.51	-0.17	-0.1
Platelet count $(\%SI)$	261.62 $\pm$ 67.85	-0.12	-0.08
Age $(years)$	46.37 $\pm$ 19.74		

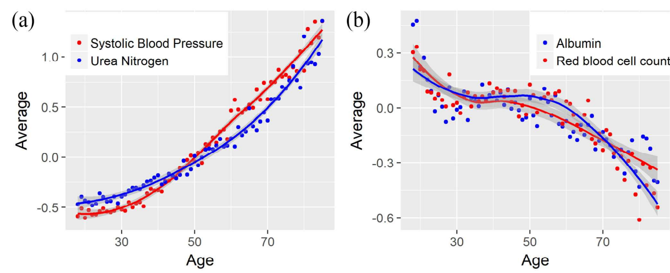


Fig. 1. Variation of biomarkers with age. Features plotted for average measurements for individuals grouped by age (in years).

II. METHODOLOGY

A. Dataset

We used biomarkers from the National Health and Human Nutrition Examination Surveys (NHANES) 1999–2010 [17]. NHANES employs a complex cluster design to sample members of the civilian USA population who are not institutionalized. NHANES uses stratified multistage probability to sample the data. Ethnicity included white, black, Mexican and others. For biomarkers of aging, we considered 16 of the biomarkers available in NHANES, namely, C-reactive protein, glycated hemoglobin, albumin, total cholesterol, urea nitrogen, alkaline phosphatase, systolic blood pressure, diastolic blood pressure, pulse, high density lipoprotein, hemoglobin, lymphocyte percent, while blood cell count, hematocrit, red blood cell count, platelet count. Subsets of these have been used in earlier work as key biomarkers of biological age [3], [4], [8]. To begin with, we had 62160 individuals from year 1999 to 2010 dataset. We merged the datasets of different years and then performed matching with the mortality follow-up data that was updated in 2015. Thus, we obtained 21451 individuals with 1664 deaths during the 4–16 years of follow-up (1999–2015). Table I shows some information on the key biomarkers used in this study.

B. Characteristics of the Dataset or Sample

Table I also shows the correlation between chronological age and the biomarkers. The table shows the correlation using direct measurements for both Pearson's ρ, and Kendall's τ. Age has higher correlation with some of the biomarker features (Systolic blood pressure, blood urea nitrogen, glycated hemoglobin) and low correlation with C-reactive protein, seram alkaline phosphatase, diastolic blood pressure, etc. Some biomarkers (e.g., pulse, red blood cell count and albumin) have negative correlation with age. Fig. 1(a) shows how two positively correlated biomarkers (systolic blood pressure ($\rho = 0.54$), blood urea nitrogen ($\rho = 0.44$) vary with age on average. Subjects in the NHANES dataset had ages in the range 18–85. Both mean systolic blood pressure and mean blood urea nitrogen increase consistently with age. Conversely, Fig. 1(b) shows how two negatively correlated biomarkers (albumin ($\rho = -0.17$), red blood cell count ($\rho = -0.17$), vary with age on average. Both mean albumin and red blood cell count decrease consistently with age. However, the variation of the decrease is not similar.

C. Symbols/Notations Used

Table II shows the notations used in this paper.

D. Centroid BA: New Approach to Estimating BA

We propose a new neighborhood-based method to predict biological age using biomarkers. Fig. 2 shows the general structure

TABLE II
NOTATIONS USED IN THE PAPER

T_R	training subjects	C_σ	trained standard deviation centroids
T_E	test subjects	C_{IQR}	trained inter-quartile range medoid
C_μ	trained centroids	BA_i	biological age of person P_i
C_{me}	trained medoids	C_μ^S	selected trained centroids
C	Centroid	ED	Euclidean distance vector
P_i	ith person	C_P	average age of selected N neighbors
N	# of neighbors	A_{C_k}	age corresponding to the $C_k{}^{th}$centroid
		μ_f^C	the f^{th} feature of the C^{th} centroid of C_μ
		σ_f^C	the f^{th} feature of the C^{th} centroid of C_σ
		T_{R_A}	training subjects with Age A

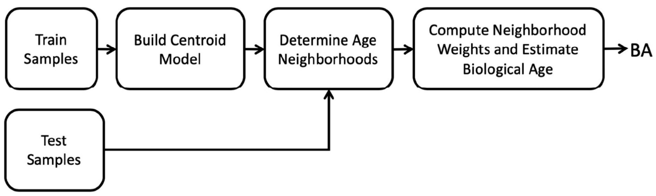


Fig. 2. Proposed framework for centroid-based biological age estimation.

Algorithm 1: Biological Age Estimation using Centroid of Age Neighborhoods.

Input: Training data T_R, test data T_E, # of neighbors (N)
Output: BA, the predicted biological age
for each age $A \in$ age range of training data T_R **do**
  $< C_{\mu_A}, C_{\sigma_A} > \leftarrow$ BUILDCENTROIDMODEL (T_{R_A})
end for
for each person $P_i \in T_E$ **do**
  $BA_i^1 \leftarrow$ COMPUTEBA1 (C_μ, P_i, N)
  $BA_i^2 \leftarrow$ COMPUTEBA2 (C_μ, C_σ, P_i, N)
end for

of the proposed framework. Below we describe each component of the framework.

Each subject is represented using the biomarker attributes, forming a multidimensional feature space. Each subject is viewed as a point in the multidimensional feature space defined by the individual biomarker attributes (the features). Subjects in our NHANES dataset had ages in the range 18–85. During training, we calculate the centroid for each age. First, we compute the centroid based on the training dataset. We divide the dataset in 68 groups based on the chronological age. We have 68 centroids (age range 18–85 inclusive). Then we calculate the mean and standard deviation of each feature for all the age groups. We denote these as C_μ and C_σ, respectively. At the testing stage, given a person P with an unknown biological age, we first determine the age neighborhoods for this individual P based on person P's biomarker attributes and the precomputed age centroids at training. Then we estimate the BA using the age neighborhoods. Algorithm 1 shows the pseudo-code of the proposed centroid-based approach to BA estimation. Clearly, how we determine the age neighborhoods is an important element in our proposed approach. Below, we describe two approaches to address this problem.

Approach 1: Using only C_μ.

For a given person P, we first calculate and record the Euclidean distances from all the 68 centroids. Now, based on

Algorithm 2: BA Estimation via Age Neighborhoods: Approach 1.

Input: train centroids (C_μ), test subject P_i, # of neighbors (N)
Output: BA_i, the predicted biological age
COMPUTEBA1(C_μ, P_i, N)
for each Centroid $C \in C_\mu$ **do**
  $ED_C \leftarrow ||P_i, C_{\mu C}||$
end for
Sort ED
Select N_P, the N neighbors from sorted ED
Compute $\bar{C_P}$
for each neighbor $j \in N_P$ **do**
  Compute the distances ($\Delta_j = |\bar{C_P} - C_j|$)
end for
for each neighbor $j \in N_P$ **do**
  Compute $a_j = 1 - \frac{\Delta_j}{\sum \Delta_j}$
  Compute weight $w_j = \frac{a_j}{\sum a_j}$
end for
$BA_i = \sum_{j=1}^{N} w_j * A_{C_j}$.

the sorted Euclidean distances, we select the required N number of neighbors added with two centroids (these centroids are used later on for removing outliers). Then we compute $\bar{C_P}$, mean age of the selected centroids, and record the distances Δ_j of each selected neighbor from the $\bar{C_P}$. We consider two weighting schemes in computing the biological age.

a) Simple Average: This is the simplest approach, where based on the selected centroids we calculate the mean age, $\bar{C_P}$. Based on the distance from the mean age, we remove two outliers that are farthest from the mean. Now we calculate the average age of the remaining centroids.

b) Weighted exponential squared distance: First, we calculate the mean age ($\bar{C_P}$). Now we compute the squared distances ($\Delta_i = (\bar{C_P} - C_i)^2$) of each neighbor from the mean; we now calculate weight $w_j = \exp^{-\Delta_j}$ and sum of weights $W = \sum_{j=1}^{N} w_j$. Finally, we calculate the centroid BA. $BA_i = \sum_{j=1}^{N} (\frac{w_j}{W} * A_{C_j})$. Algorithm 2 shows the pseudo-code summarizing the proposed approach.

Approach 2: Using Both C_μ and C_σ.

In the above basic approach, we used the centroid which is calculated based on the mean of the individual features. But the distributions of the features were ignored. To address this issue, in Approach 2, we incorporate the standard deviation (of each individual feature) along with their mean. Both C_μ and C_σ are calculated based on the training dataset. We have 68 centroids and 68 standard deviations for each individual features, one pair for each age range between 18 to 85. Given person P, we estimate the biological age based on these (feature centroid, dispersion) pairs. For a given person P, we first calculate and record the Euclidean distances from all the 68 trained centroids C_μ. Now, based on the sorted Euclidean distances, we select a subset of centroid C_μ and the standard deviation C_σ. Then we apply two different parameters (α and $\tau-$ see below) to determine the neighbors that will be considered to calculate

Algorithm 3: BA Estimation via Age Neighborhoods: Approach 2.

Input: mean training centroids (C_μ), standard deviation train centroids (C_σ), test dataset (T_E), # of neighbors (N), τ, α
Output: BA_i, the estimated biological age
COMPUTEBA2 $(C_\mu, C_\sigma, P_i, N, \alpha, \tau)$
$C_\mu^S \leftarrow \phi$
for each person $P_i \in T_E$ **do**
  for each Centroid $C \in C_\mu$ **do**
    count $\leftarrow$ # features k of $P_i \in (\mu_k^C \pm \alpha * \sigma_k^C)$
    if count $>= \tau$ **then**
      $C_\mu^S \leftarrow C_\mu^S \cup C$
    end if
  end for
  if $|C_\mu^S| >= N$ **then**
    $BA_i \leftarrow$COMPUTEBA1(C_μ^S, P_i, N)
  else
    $BA_i \leftarrow$COMPUTEBA1(C_μ, P_i, N)
  end if
end for

biological age. Based on the selected number of neighborhood centroids, the algorithm will use the new centroids that pass the thresholds. If not enough neighbors are found, the algorithm defaults to Approach 1. Algorithm 3 shows the pseudo-code for this improved approach.

Two key parameters in the proposed approach are α and τ. The first parameter α is a factor that we used to compute the lower range $(\mu - \alpha * \sigma)$ and higher range $(\mu + \alpha * \sigma)$ for evaluating similarity between corresponding features. The ranges are calculated for each feature individually. Essentially, α is used to restrict the allowed distance between the given feature for a subject say P, and the corresponding feature from the centroid for a given age category. With increasing values of α, more distance is allowed, and hence leading to a less stringent criteria. Conversely, when α is small, only centroids that are very close to P, on the given feature will be involved in computing the BA for P. The parameter τ is a threshold count that is used to determine how many similar features (% of the matches) are allowed in selecting a centroid. Here, even if one feature is found to be very close between P, and a given centroid, say C, this centroid may still not be used to estimate the BA for P unless some other features are similar between P and C, and the fraction of similar (or matching) features are above τ, the threshold on the number of feature matches. With higher values of τ, we have a more stringent criteria for selecting the thresholds.

Both α and τ can affect the performance of the proposed method. In our experiments, to find the best combination of α and τ in estimating the BA, we varied α in the range $\alpha = 0.25$, 0.5, 1.0, 1.5, 2.0 and τ in the range $\tau = 0.5$, 0.75, 0.9.

E. Medoid BA: Estimating BA Using Medoid of Age Neighborhoods

Similar to the mean of each age category (the centroids), we also considered median representation of each age cluster.

Algorithm 4: BA Estimation using Medoid of Age Neighborhoods.

Input: Training data T_R, test data T_E, # of neighbors (N)
Output: BA, the predicted biological age
for each age $A \in$ age range of training data T_R **do**
  $<C_{me_A}, C_{IQR_A}> \leftarrow$ BUILDMEDOIDMODEL (T_{R_A})
end for
for each person $P_i \in test$ **do**
  $BA_i^1 \leftarrow$ COMPUTEBA1 (C_{me}, P_i, N)
  $BA_i^2 \leftarrow$ COMPUTEBA2 $(C_{me}, C_{IQR}, P_i, N)$
end for

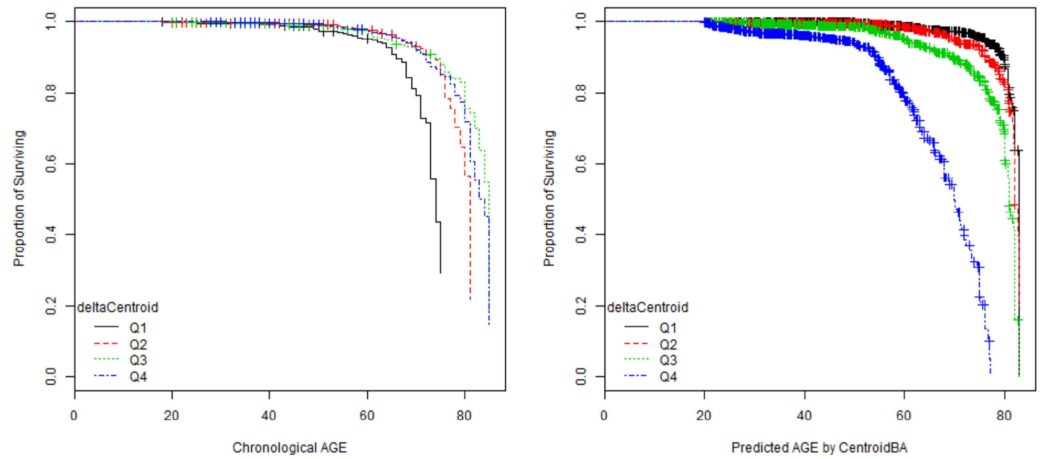
We call the estimated biological age (BA) based on median, as medoid-based BA. Algorithm 4 shows the pseudo-code of the proposed medoid-based biological age estimation. Similar to Algorithm 1, for every age (in the range 18–85) we calculate a train medoid of the attributes, where the medoid is a vector in the feature space that contains the respective median of each attribute at a given chronological age. First, we compute medoid based on the training dataset. We divide the dataset in 68 groups based on the chronological age. We have 68 medoids (age range 18–85 inclusive). Now we calculate the median and the inter quartile range (IQR) of each feature for each age group. We denote them as C_{me} and C_{IQR}, respectively. Following the centroid approach, we can now estimate a person's biological age based on the trained medoids and the recorded IQRs. The medoid approach is summarized in Algorithm 4. Algorithm 1 and Algorithm 4 are similar. The differences are, C_μ is replaced by C_{me}, and C_σ is replaced by C_{IQR}. So, for medoid BA, to compute COMPUTEBA1(), we then use C_{me} and similarly for COMPUTEBA2(), we use both C_{me} and C_{IQR}.

III. Results

For validation and comparison of the proposed BA algorithms, we have applied three statistical analysis methods, namely, Cox proportional hazard (Cox PH) model, Kaplan-Meier (KM) curves, and survival area under the curve (AUC) of receiver operating characteristic (ROC). We randomly partitioned the dataset into training set and test set, using 2/3 for training, the remaining 1/3 for testing. All statistical analyses were performed using the R Language, Ver. 3.3.5 (The R Foundation for Statistical Computing, Vienna, Austria). The following packages were used: survival, gtools, ggplot2, tidyverse, keras, e1071, matrixStats, SurvAUC.

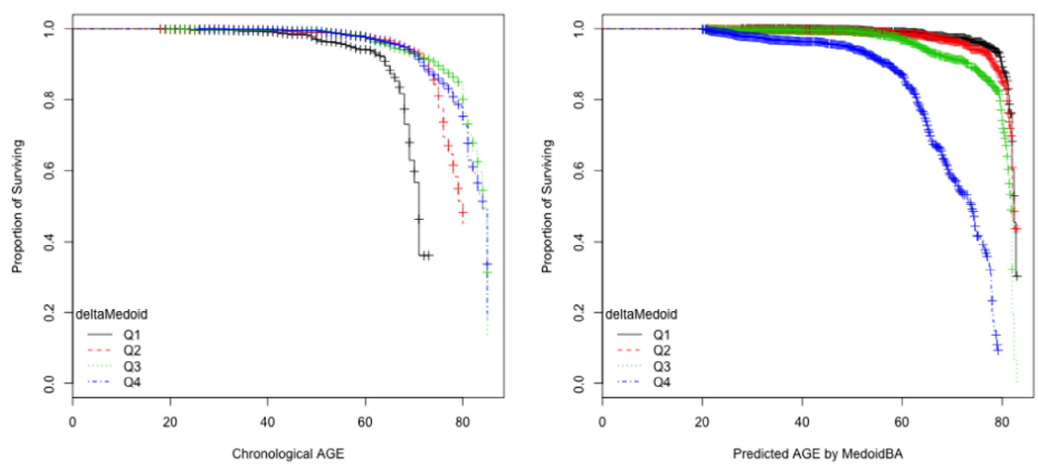
A. Cox PH Model

We used Cox proportional mortality hazard modeling [18] to quantify the association of the proposed centroid BA or medoid BA with all-cause mortality. Under the Cox model, the relationship between hazard and the covariates is described by considering the logarithm of the hazard as a linear function of the variables. The larger the hazard ratio the better the method. We have considered five BA estimation algorithms to calculate the hazard ratio (HR) [18]. First, we estimate BA using MLR method [3], KD method [7], DNN (Deep Neural Networks) [17], and



(a) Centroid with chronological age (b) Centroid with estimated biological age

Fig. 3. The Kaplan Meier curves for proposed centroid-based BA estimation algorithms. Results are reported for applying $\Delta = CA - BA$ on both (a) chronological age, and (b) estimated BA. Q1, Q2, etc. denote 1st quartile, 2nd quartile, etc.



(a) Medoid with chronological age (b) Medoid with estimated biological age

Fig. 4. The Kaplan Meier curves for proposed Medoid based approach. Results are reported for applying ($\Delta = CA - BA$) on both (a) chronological age, and (b) estimated BA. Q1, Q2, etc. denote 1st quartile, 2nd quartile, etc.

TABLE III
COX PH RESULTS FOR BIOLOGICAL AGE ESTIMATION METHODS

	Biological Age	
	HR	p-value
MLR	0.99	2.21E-24
KD	1.04	9.28E-41
DNN	**1.12**	1.34E-63
Centroid	**1.12**	9.16E-32
Medoid	**1.12**	4.02E-18

our proposed models. For DNN, we used the network reported in [17] by Putin *et al.* which had the best performance. Then we calculate $\Delta = CA - BA$ for each BA estimation algorithm. We then use the Δ quartiles to apply the Cox model. Table III shows the results for the biomarker features. We use the estimated BA as the parameter for the Cox model, and recorded the HR for each method (Centroid 1.12, Medoid 1.12, MLR 0.99, KD 1.04, and DNN 1.12). From the perspective of Cox PH model, we found that proposed centroid-based and medoid-based BA estimation methods had similar or slightly better performance than the other methods.

B. KM Curves and Log-Rank Test

To further study the performance of centroid BA, we analysed the Kaplan-Meier (KM) survival curves [19] obtained using the quartiles of delta ($\Delta = CA - BA$). Fig. 3 shows the KM plots using the estimated BA from the proposed Centroid approach. Fig. 4 shows the results for the proposed Medoid approach. In general, biological age performs well in distinguishing the proportion of survivors for each method. From the figures, it is apparent that when applied in survival model, using the estimated biological age from each method seemed to perform better than using chronological age. Among the methods, the

TABLE IV
LOG RANK RESULTS (χ^2-DISTANCE) FOR MORTALITY MODELING USING FOUR BIOLOGICAL AGE ESTIMATION METHODS

	Biomarker			
	Chronological Age		Estimated BA	
	Chi-Sq	p-value	Chi-Sq	p-value
MLR	74.24	5.55E-16	32.75	3.18E-16
KD	22.68	4.07E-05	157.59	3.21E-15
DNN	74.25	2.16E-16	439.83	1.30E-16
Centroid	81.51	2.38E-16	689.37	1.29E-17
Medoid	**115.15**	1.39E-16	**707.42**	1.12E-17

proposed centroid and medoid based approaches have similar survival curves. To further quantify the performance, we used the log-rank test to compare the survival distributions obtained using the different BA algorithms. The log-rank test can be used to compare different Kaplan-Meier curves to see if they are statistically equivalent. The output of the test is a χ^2-distance, and the p-value associated with the distance. Higher χ^2-distances and low p-values indicate a better separation between the curves, and hence a better performance in mortality modeling. The differences among the biological ages estimated by the four methods are more evident using quantitative measures, e.g., the χ^2-distance between their respective KM curves, as captured by the logrank test (Table IV). Proposed medoid based approach has the best overall results using either chronological age, or biological age.

R package "survival" is used for Cox PH model, KM plots, and log-rank test. The Surv() function is the primary function. The parameters for Surv() are as follows: Surv(time, status) $\sim$ variable. So, for time we can use either chronological or estimated biological age, status is the mortality status. For the variable we use $\Delta = CA - BA$. If chronological age is used as the time, we can also test the performance of estimated BA as a variable.

C. ROC of AUC

We have used the receiver operating characteristics (ROC) curves to examine the sensitivity and specificity of chronological age and the predicted biological ages in mortality modeling. We have applied estimators of cumulative and incident/dynamic area under curve (AUC) proposed by Song and Zhou [20]. These estimators are given by the areas under the time dependent ROC curves estimated by sensitivity and specificity. Fig. 5 shows the estimated ROC curves for the biological age prediction methods. Using the ROC curve on $\Delta = CA - BA$, the best performing biological age estimate was the proposed medoid BA (AUC = 0.66) followed by centroid (AUC = 0.61). MLR, KD, and DNN have AUC values of 0.60, 0.57, and 0.65, respectively. The results improved using the estimated BA rather than Δ. Observing the results on the biomarker features, the proposed medoid and centroid based BA estimation algorithms produced the best results with respect to the area under the curve of ROC. Cox model performance in terms of hazard ratio (HR) is similar in comparison to MLR, KD, or DNN approaches. Using the KM curves and log-rank test on the results from the proposed centroid-based and medoid-based approaches resulted in improved performances over the previous approaches. These results suggest that the centroid based model is a competitive BA predictor. The medoid-based model showed an improvement over all the other methods.

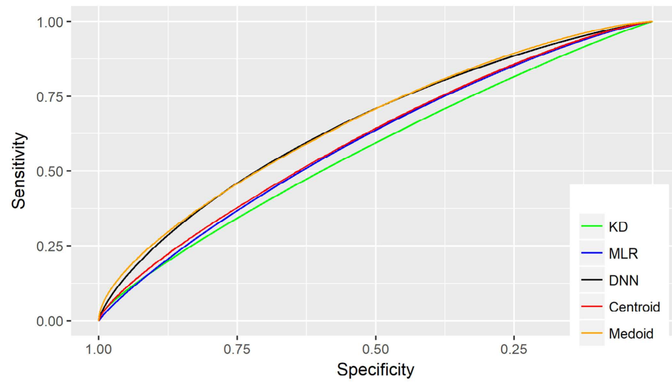


Fig. 5. ROC curves for MLR, KD, DNN, Centroid, and Medoid algorithms using biomarker features. Results are reported for applying $\Delta = CA - BA$ for the respective estimation approach.

TABLE V
RESULTS FOR VARYING α AND τ IN APPROACH2 USING CENTROID BA (ALGORITHM 3), FOR N = 5

	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
α	χ^2-dist	AUC	χ^2-dist	AUC	χ^2-dist	AUC
0.25	**805.27**	0.74	755.98	0.73	755.98	0.73
0.5	767.75	**0.75**	779.3	0.74	756.01	0.73
1	775.69	0.74	771.6	0.74	762.19	0.74
1.5	760.79	0.74	744.28	0.74	748.88	**0.75**
2	755.98	0.73	742.55	0.74	733.95	0.74

All the corresponding p-values are significant ($p \approx 0$) for log-rank test.

D. Results for Approach 2

Applying the proposed Approach 2 (Algorithm 3) using both the C_μ, C_σ improved the results. Table V shows the results for applying log-rank test, and AUC of ROC survival curves for variation of different values of the parameter α and τ. We varied parameter α ($\alpha = 0.25$, 0.5, 1, 1.5, 2) and τ ($\tau = 0.5$, 0.75, 0.9) to test the performance of the proposed approach. Parameter α is varied to test the impact of the range allowed, and parameter τ is varied to check the impact of percentage of feature matches. We have considered 15 possible variations of α and τ for these sets of values. As mentioned earlier, applying α and τ is a more robust approach. We notice that the above two mentioned criteria (χ^2-distance using the log-rank test, survival AUC) improved for the centroid method using Approach 2. Although for parameter $\alpha = 0.25$, and $\tau = 0.5$ the χ^2-distance is highest, but the range using $\alpha = 0.25$ will be too small. Thus we chose $\alpha = 0.5$, and $\tau = 0.5$ as our best combination for the centroid method. We observed that survival AUC improved from 0.61 (Table VI) to 0.75, χ^2 distance of the log-rank test increased from 689.37 to 767.75, and the hazard ratio of Cox PH model also improved from 1.120 (Table III) to 1.128 (data not shown).

TABLE VI
AREA UNDER THE CURVE (AUC) OF RECEIVER OPERATING CHARACTERISTICS (ROC) CURVES

	Delta	Estimated BA
MLR	0.60	0.64
KD	0.57	0.60
DNN	0.65	0.51
Centroid	0.61	0.64
Medoid	**0.66**	**0.66**

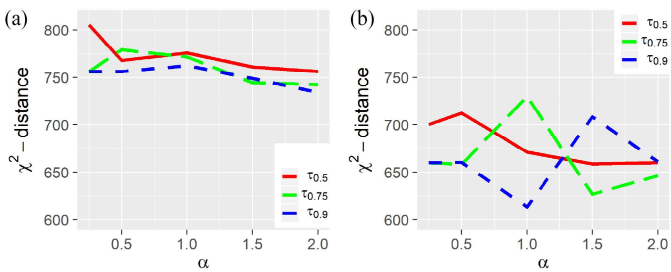


Fig. 6. Impact of parameter α and τ in Approach2 (a) Centroid BA, and (b) using Medoid BA.

TABLE VII
RESULTS FOR VARYING α AND τ IN APPROACH2 USING MEDOID BA (ALGORITHM 4), FOR N = 5

	$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.90$	
α	χ^2-dist	AUC	χ^2-dist	AUC	χ^2-dist	AUC
0.25	700.20	0.75	660.17	0.75	660.16	0.75
0.5	712.10	**0.76**	658.22	0.75	660.22	0.75
1	671.20	0.75	**729.65**	0.75	613.19	0.75
1.5	658.70	0.75	626.75	0.75	708.25	0.76
2	660.20	0.75	646.18	0.75	661.25	0.75

All the corresponding p-values are significant ($p \approx 0$) for log-rank test.

Fig. 6 shows the impact of α, and τ parameters on χ^2-distance of the log-rank test using Approach 2 for both the centroid and medoid methods. Similarly, Table VII shows the results of varying α and τ for medoid BA using Approach 2. We observe that survival AUC value improved from 0.66 (Table VI) to 0.75, χ^2 distance of the log-rank test increased from 707.42 to 712.10, and the hazard ratio of Cox PH model remained similar (1.12) (using $\alpha = 0.5$ and $\tau = 0.5$). For both the centroid approach and the medoid approach, our observation is that parameter setting ($\alpha = 0.5$, $\tau = 0.5$) usually leads to the best result, or close to the best. Thus we have used this setting in comparative analysis of the methods.

E. Comparative Results

To place the results of our proposed approaches in perspective, we have compared with current popular approaches to BA estimation using biomarker data, namely KD [7], MLR [3], and DNN [17]. We have shown the corresponding results with these methods while we discuss the results of our proposed approaches (see Tables III, IV, and VI).

Table VIII shows the comparative performance of existing popular methods and the proposed approaches. The methods are compared using three popular performance metrics, namely,

TABLE VIII
OVERALL COMPARATIVE RESULTS ON BIOLOGICAL AGE PREDICTION

	CoxPH		Log-rank		AUC
α	HR	p-value	χ^2-dist	p-value	
MLR [3]	0.99	2.21E-24	32.75	3.18E-16	0.60
KD [7]	1.04	9.28E-41	157.59	3.21E-15	0.57
DNN [16]	1.12	1.34E-63	439.83	1.30E-16	0.65
Centroid	1.12	9.16E-32	689.37	1.29E-17	0.61
Centroid2	**1.13**	7.18E-16	**767.75**	4.05E-18	0.75
Medoid	1.12	4.02E-18	707.42	1.12E-17	0.66
Medoid2	1.12	6.30E-16	712.1	4.81E-18	**0.76**

Hazard Ratio (HR) using the Cox proportional hazard model, χ^2-distance from the LogRank test, based on Kaplan-Meier (KM) curves, and the survival area under the curve (AUC). Using these three performance measures, the results show that our proposed Approach 2 using both Centroid2 (C_μ, C_σ) and Medoid2 (C_{me}, C_{IQR}) have the best overall results. Between the proposed centroid-based and mediod-based age neighborhood approaches, the centroid-based approach resulted in a superior performance over the mediod-based approach. The results for Approach 2 are reported for using $\alpha = 0.5$ and $\tau = 0.5$.

IV. DISCUSSION

A. Impact of Number of Neighbors

Two key steps in the proposed method are how to select the neighbors, and how to determine the number of neighbors to be involved. Algorithm 2 and Algorithm 3 have shown how the neighbors are selected, once we know the number of neighbors. Given the significance of neighborhoods in our approach, it becomes important to study how the neighborhood size can influence our results. To check the impact of neighborhood size on our proposed centroid of age neighborhoods for biological age estimation, we experimented with different values for N, the number of neighbors. We have considered 10 values of N, namely, N = 1, 2, 3, 5, 7, 10, 15, 18, 20, 22, 25, 30, 40, 50. We show the impact of N based on the KM plots and the corresponding log-rank test. We also studied changes in error density with increasing N.

Table IX shows the results of log-rank test using both estimated BA and delta ($\Delta = CA - BA$) for different number of neighbors. Based on the χ^2 distance, we observe that BA estimated using 10 neighbors has the maximum value followed by 5 neighbors. When using delta, best results were obtained with $N = 2$, followed by $N = 5$. Table IX also shows how mean absolute error (MAE) vary with increasing number of neighbors in our proposed centroid-based method. The MAE starts high ($N = 1$), then it reduces sharply at $N = 2$. Considering just a single neighbor leads to a significant error in the estimation. The sharp decrease at $N = 2$ is understandable because it averaged out the error introduced by the single neighbor. MAE is similar for $N = 2$ to $N = 20$. After that the MAE starts to increase almost linearly with the increasing N. $N = 20$ has the overall lowest MAE. The performance criteria for choosing N was to get lower MAE, and higher χ^2-distance using both estimated

TABLE IX
IMPACT OF NUMBER OF NEIGHBORS ON MEAN ABSOLUTE ERROR (MAE), LOG-RANK TESTS ON ESTIMATED BIOLOGICAL AGE, DELTA (CA-BA)

	MAE	Estimated BA		delta (CA - BA)	
N		χ^2-dist	p-val	χ^2-dist	p-val
1	26.82	**566.60**	0	**935.14**	0
2	11.51	399.8	0	723.7	0
3	11.47	376.3	0	653.8	0
5	11.35	491.2	0	703.9	0
7	11.17	480.9	0	664.6	0
10	10.92	506.1	0	562.1	0
15	10.80	494.5	0	408.01	0
18	10.81	497.5	0	334.4	0
20	**10.74**	476.5	0	320.4	0
22	10.83	536.4	0	326.3	0
25	10.91	568.5	0	212.1	0
30	11.20	494.6	0	154.6	0
40	12.36	419.6	0	108.4	0
50	14.02	394.1	0	84.0	0

TABLE X
RESULTS FOR DIFFERENT DISTANCE MEASURES

	Centroid			Medoid		
Distance	HR	χ^2-dist	AUC	HR	χ^2-dist	AUC
Euclidean	1.13	767.75	0.75	1.12	712.11	0.75
Jaccard	1.13	**874.61**	**0.76**	1.12	**860.06**	0.75
Manhattan	1.13	837.39	0.75	1.12	707.52	0.75

TABLE XI
STABILITY OF THE RESULTS

	Centroid			Medoid		
	CA	BA		CA	BA	
Fold	χ^2-dist	χ^2-dist	HR	χ^2-dist	χ^2-dist	HR
1	132.79	811.69	1.128	176.63	717.19	1.124
2	84.92	761.61	1.123	138.45	687.53	1.119
3	105.39	749.93	1.120	150.97	750.93	1.115
mean	107.70	774.41	1.124	155.35	718.550	1.119
std	24.02	32.81	0.004	19.46	31.719	0.005

BA and delta ($\Delta = CA - BA$). From these results, we selected $N = 5$ as the overall best number of neighbors.

B. Impact of Distance Measures

Table X shows the hazard ratio (Cox PH), χ^2-distance (Log-rank test), and survival area under the curve (AUC) using different distance measures in the proposed centroid-based neighborhood approach to BA estimation. We tested the performance of different approaches using 5 neighbors. We observe that, in Table X Jaccard distance has the largest χ^2-distances and AUC values applying delta (Δ = CA-BA) as a co-variate while the general trend is similar for all the distance measures. Although the Jaccard distance was slightly better than the Euclidean distance that we used to develop the model, these results show the generality of the proposed age neighborhood approaches.

C. Stability of Results

Table XI shows the χ^2-distance (Log-rank test) and hazard ratio (Cox PH) using 3-fold cross validation for both the proposed centroid-based and medoid-based approach to BA estimation.

TABLE XII
RESULTS REPORTED FOR LOG-RANK TEST AND AREA UNDER CURVE (AUC) OF RECEIVER OPERATING CHARACTERISTICS (ROC) CURVES FOR SINGLE MODEL GROUPED BY GENDER

	χ^2-dist		AUC	
	Female	Male	Female	Male
MLR	17.98	20.26	0.6	0.61
KD	64.54	81.05	0.57	0.58
DNN	191.09	265.74	0.65	0.66
Centroid	285.31	**421.56**	0.59	0.64
Centroid2	354.55	419.95	0.59	0.67
Medoid	269.38	386.36	0.67	0.68
Medoid2	294.82	363.75	**0.68**	**0.71**

We observe that, in Table XI the χ^2-distances and hazard ratios are similar to the results of simple train and test dataset. The relatively small values for the standard deviation demonstrate the stability of the results. We also performed 10-fold cross validation using the same data set. Similar to the above results, the standard deviation of the results was equally small, about 0.71% of the mean for centroid-BA, and 0.89% of the mean for medoid-BA approach.

D. Impact of Gender

Gender is expected to have some influence on the performance of an age estimation scheme [2]. We have grouped the results of the single model applied to all subjects into male and female subjects. In Table XII we show the results for log-rank (χ^2-distance) test, and AUC for each BA estimation algorithm for both female and male. For log-rank test, proposed centroid-based model has the highest χ^2-distance (Male = 421.56, Female = 354.55). However, the proposed medoid based method has the overall best result for AUC (Male = 0.71, Female = 0.68). We also observe that, in every case independent of the particular method applied, or the specific features used, better results were obtained for male subjects than for female subjects. These results are consistent with other existing work that show that age prediction and mortality analysis for female subjects is generally more difficult than for male subjects [2], [21], [22].

E. Computational Complexity

Of the proposed methods, Approach 2 requires more computation. This depends on the time used in COMPUTEBA2(). The running time of COMPUTEBA2() is the sum of running time of each statement of the pseudocode. The outer for loop (Line 2 to Line 14) has a running time of $O(|T_E|)$, T_E is the test dataset. The nested for loop (Line 3 to Line 8) looks up the two centroid matrices. The complexity is $(O(k) + O(k)$, where k is the number of features. Now depending on the condition of $|C_\mu^S| >= N$ method COMPUTEBA1() is called either using C_μ^S or C_μ. In Algorithm 2 for person P_i, we calculate the Euclidean distances from each centroid $C \in C_i$. This can be done in time of $O(|C| \times k)$, where k is the number of features and $|C|$ is the number of centroids. Then we perform quick sort on the calculated distances which is on average $O(|C| \log |C|)$, and essentially is also a constant. The for loop (Line 7 to Line 9) computes the distances in $O(N)$, where N is number of neighbors. Similarly, the for loop (Line 10 to Line 13) have time of $O(N)$. The overall complexity of Approach 2 is $O(|T_E| \times |C|)$, where T_E is the test dataset, and $|C|$ is the number of centroids. Since $|C|$ and N are essentially constants ($|C| = 68$, $N \leq |C| = 68$), and k is relatively small, the time required will thus depend linearly on the size of the dataset.

V. Conclusion

In this work, we studied age estimation using human blood biomarkers from the NHANES dataset. We presented a new centroid/medoid-based model to estimate biological age. We grouped individuals of same age to the same centroid. The proposed method utilized specially selected age neighborhood to perform biological age estimation. Although both centroid and medoid based approaches have similar performances, we observed that considering the three different methods for quantifying the performance of estimated BA as used in the paper (i.e., Cox PH, Log-rank from KM curves, and AUC) the centroid based approach (with Approach 2) is the overall best. Practical results demonstrate the significant improvement in BA estimation using the proposed methods when compared with existing approaches, such as KD [7], MLR [3], and DNN [17]. Although the performance of Cox Proportional Hazard model (Cox PH) provides similar hazard ratio for DNN and proposed methods, the methods differ in terms of log-rank test and survival area under the curve. In this work, we established that centroid based method can be used on blood biomarkers to estimate biological age. A potential future work, will be to study whether other modalities, for instance, human body measurements, or human locomotor activity, could be used for estimation of biological age, using the centroid-based or medoid-based age neighborhoods.

We mention one potential limitation of the proposed approach. As a learning-based approach, the results reported clearly depend on the specific dataset used. We have tested on a large dataset, and also performed cross-validation on the dataset. We have discussed the stability of the results produced with the proposed approaches. The results will not change much for a dataset with similar characteristics as the NHANES dataset. However, the NHANES dataset is based mainly on subjects from the US population, and thus the trained models based on NHANES may not generalize to people from a very different ethnic makeup. For instance, a dataset containing mainly people of East Asian origin may require us to train the model on the new dataset, before using it for prediction. But the general methodology remains the same. Thus, another potential future work would be to test the approach on biomarker datasets with possibly different characteristics from NHANES, example, data from populations with a different ethnic makeup when compared with the US population.

Acknowledgment

The authors declare no competing interest.

References

[1] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[2] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[3] M. E. Levine, "Modeling the rate of senescence: Can estimated biological age predict mortality more accurately than chronological age?" *J. Gerontology Ser. A, Biomed. Sci. Med. Sci.*, vol. 68, no. 6, pp. 667–674, 2012.

[4] D. W. Belsky *et al.*, "Quantification of biological aging in young adults," *Proc. Nat. Acad. Sci.*, vol. 112, no. 30, pp. E4104–E4110, 2015.

[5] A. Mitnitski, S. E. Howlett, and K. Rockwood, "Heterogeneity of human aging and its assessment," *J. Gerontology Ser. A, Biomed. Sci. Med. Sci.*, vol. 72, no. 7, pp. 877–884, 2016.

[6] B. J. Sharkey, "Functional vs chronologic age," *Med. Sci. Sports Exercise*, vol. 19, no. 2, pp. 174–178, 1987.

[7] P. Klemera and S. Doubal, "A new approach to the concept and computation of biological age," *Mechanisms Ageing Develop.*, vol. 127, no. 3, pp. 240–248, 2006.

[8] P. Sebastiani *et al.*, "Biomarker signatures of aging," *Aging Cell*, vol. 16, no. 2, pp. 329–338, 2017.

[9] S. H. Jackson, M. R. Weale, and R. A. Weale, "Biological age—What is it and can it be measured?" *Archives Gerontology Geriatrics*, vol. 36, no. 2, pp. 103–115, 2003.

[10] I. H. Cho, K. S. Park, and C. J. Lim, "An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI)," *Mechanisms Ageing Develop.*, vol. 131, no. 2, pp. 69–78, 2010.

[11] R. Poulton, T. E. Moffitt, and P. A. Silva, "The Dunedin multidisciplinary health and development study: Overview of the first 40 years, with an eye to the future," *Social Psychiatry Psychiatric Epidemiology*, vol. 50, no. 5, pp. 679–693, 2015.

[12] J. Cole *et al.*, "Brain age predicts mortality," *Mol. Psychiatry*, 23, pp. 1385–1392, 2018.

[13] E. Bobrov *et al.*, "PhotoAgeClock: Deep learning algorithms for development of non-invasive visual biomarkers of aging," *Aging (Albany NY)*, vol. 10, no. 11, pp. 3249–3259, 2018.

[14] W. C. Sanderson and S. Scherbov, "Measuring the speed of aging across population subgroups," *PLoS ONE*, vol. 9, no. 5, 2014, Art. no. e96289.

[15] T. V. Pyrkov *et al.*, "Extracting biological age from biomedical data via deep learning: Too much of a good thing?" *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 5210.

[16] S. A. Rahman and D. A. Adjeroh, "Deep learning using convolutional LSTM estimates biological age from physical activity," *Sci. Rep.*, vol. 9, no. 11425, 2019.

[17] E. Putin *et al.*, "Deep biomarkers of human aging: Application of deep neural networks to biomarker development," *Aging*, vol. 8, no. 5, pp. 1021–1033, 2016.

[18] D. R. Cox and D. Oakes, *Analysis of Survival Data*, vol. 21. Boca Raton, FL, USA: CRC Press, 1984.

[19] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Amer. Statist. Assoc.*, vol. 53, no. 282, pp. 457–481, 1958.

[20] X. Song and X.-H. Zhou, "A semiparametric approach for the covariate specific ROC curve with survival outcome," *Statist. Sinica*, vol. 18, pp. 947–965, 2008.

[21] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[22] S. A. Rahman and D. Adjeroh, "Surface-based body shape index and its relationship with all-cause mortality," *PLoS ONE*, vol. 10, no. 12, 2015, Art. no. e0144639.