Information Sciences 586 (2022) 628–643

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Risk score-embedded deep learning for biological age estimation: Development and validation

Suhyeon Kim [a], Hangyeol Kim [b], Eun-Sol Lee [c], Chiehyeon Lim [d], Junghye Lee [d,*]

[a] Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[b] Department of Business Analytics, Graduate School of Interdisciplinary Management, UNIST, Ulsan 44919, Republic of Korea
[c] Bio-Age Medical Research Institute, Bio-Age Inc., Seoul 06170, Republic of Korea
[d] Department of Industrial Engineering & Graduate School of Artificial Intelligence, UNIST, Ulsan 44919, Republic of Korea

ARTICLE INFO

Article history:
Received 13 October 2020
Received in revised form 6 October 2021
Accepted 5 December 2021
Available online 17 December 2021

Keywords:
Biological age
Health index
Risk score
Deep learning
Autoencoder
Index validation

ABSTRACT

The health index measures a person's overall health status which provides useful information for people to manage their health, so developing a precise and relevant health index is urgent. Currently, many researchers have studied the biological age (BA) estimation, one of the beneficial health indices, by applying machine learning and deep learning techniques to health data. However, most of them have focused on the chronological age prediction or basic latent feature extraction methods. In this paper, we present a new algorithm to estimate BA, called Risk Score-Embedded Autoencoder-based BA (RSAE-BA). RSAE-BA can provide an accurate health index by using deep representation learning with an individual's health risk. We first proposed a notion of risk score (RS) calculation to monitor a person's health risk. Then we extracted latent features by using an autoencoder embedding the RS, and used them to generate BA. To evaluate RSAE-BA, we presented a new BA validation method using the RS, which is applicable to both unlabeled and labeled data. We compared the results of RSAE-BA with existing methods, and demonstrated the accuracy of RSAE-BA and its applicability to predict disease incidence. We believe that RSAE-BA will be a useful alternative method to measure a person's health.

1. Introduction

Aging is a gradual decline of the physical and mental capacity of an organism over time; the process is correlated with functional impairment, increasing susceptibility to disease, and risk of death [26,32,38]. Advances in medical technology have increased the average life expectancy of humans, and this trend has increased the importance of personal health care and prevention of diseases caused by aging. To estimate an individual's health and aging status, chronological age (CA) is used as a general indicator [21,29]. However, CA is recorded by a simple time flow, so it is considered as a low-confident indicator that cannot accurately evaluate the functional and structural capacity of the body or aging status [2,8]. Hence, estimation of 'age' requires the development of a new comprehensive health index that is superior to CA in ability to predict an individual's overall health status, aging degree, and even risk of disease [26,29].

* Corresponding author.
E-mail addresses: suhyeonkim@unist.ac.kr (S. Kim), hangyeol0225@unist.ac.kr (H. Kim), ilikerefill@daewoong.co.kr (E.-S. Lee), chlim@unist.ac.kr (C. Lim), junghyelee@unist.ac.kr (J. Lee).

https://doi.org/10.1016/j.ins.2021.12.015
0020-0255/© 2021 The Author(s). Published by Elsevier Inc.

Biological age (BA) is a representative numerical index that can represent a person's health and aging [48,12]. When people are subjected to a health check-up, the use of BA can simplify identification of their condition by summarizing electronic medical record (EMR) data. The development of BA algorithms begins with the assumption that physical age is different from CA. With advances in machine learning, several machine-learning methods have been applied to BA estimation, and meaningful BA indices have been developed using various data such as gene expression and EMR [3,16,18,19,24,28,29]. BA has also been estimated by incorporating with the age-dependent variables, mentioned as biomarkers that are associated with health status in existing health indicators [10].

Three general types of methods have been used to estimate BA: (1) regression for CA, (2) simulation learning, and (3) latent feature extraction. First, regression for CA-based BAs have been estimated commonly using multiple linear regression (MLR) or regularized linear regression (e.g. lasso regression and elastic net). They exploit the correlation between CA and biomarkers [17,33]. Several deep learning methods have also been used to estimate BA by using a large amount of data [1]. For example, Cole et al. [11] proposed a deep learning-based predictive modeling approach for CA using convolutional neural networks (CNN) where they estimated a brain-predicted age using raw brain MRI data. Pyrkov et al. [45] and Rahman and Adjeroh [46] utilized the physical activity data and presented CNN-based time-series approaches. These methods attempt to minimize the mean squared error between the estimated BA and CA (i.e., target value is CA). Next, as a simulation learning-based model, Klemera and Doubal's method (KDM) was proposed to minimize the distance between the regression lines and biomarker points in a multi-dimensional space of all biomarkers [30]; this method has been used in many studies that require BA estimation and has predicted mortality better than other methods [38,39,25,50]. Lastly, many researchers have estimated BAs by summarizing factors of principal component analysis (PCA) [3,40,42,41], which is widely used to convert high-dimensional data to low-dimensional data based on the orthogonal transformation [27]. Such literature has been focused on extracting latent features to represent the characteristics of original data well.

However, each existing BA estimation method has a demerit. Linear regression-based methods tend to distort BA to CA; they do not account for the discontinuity of the aging rate over an individual's lifetime. Previous deep learning-based methods (i.e., nonlinear regression) also focused on predicting CA as a target value (i.e., supervised learning); this may deviate from the main goal of BA estimation, which is to measure the age as a state of health, not as CA. KDM is a simulation-based method that relies on many assumptions and conditions for experimental settings and thus its accuracy cannot be guaranteed when data does not conform to the assumptions. Finally, in the case of the PCA-based techniques, the direction of extracting factors (i.e., maximizing the variance of data for BA estimation) may be ineffective to represent BA.

Furthermore, most existing methods do not consider the risk level or the direction of each variable, whether a bad level of the variable is less than the normal level, or greater than it. Especially, machine learning and deep learning-based BA estimates generated by using lots of variables are difficult to represent the health risks of all variables, because the criteria (e.g., normal criteria for clinical variables) and directions for the risk that they represent are ambiguous or different from each other; it means that there is a lack of numerical uniformity for risks of the variables. BA should serve as an indicator of an individual's health, so an approach to estimating BA should include individual health-risk information. In addition, an index must be validated before it can be adopted for widespread usage in the real world, but methods to validate the BA estimates have not been standardized [38], because the index has no specific target value. Conventional methods to estimate BA have been generally validated by identifying the correlation between BA and CA or by using the mortality or disease incidence of cohort data as a target value of BA [9,47]. However, BA is an indicator of the body's biological state, and therefore cannot be easily derived from CA, mortality, or disease incidence [20]. Besides, validating the effect and direct correlation of BA estimates on mortality is of questionable value [26], and existing methods that consider mortality or disease incidence cannot be used unless the data are from a cohort study with follow-up examinations.

In this paper, we present a new BA-estimation algorithm that uses representation learning with people's health risk. We first calculate the cumulative Gaussian probability function for each risk factor, and define this function as the risk score (RS). Then we use an RS-embedded autoencoder (RSAE) to extract latent features by modifying a loss function to consider the original data and RS. Finally, we use the latent features to generate a BA from this RSAE (RSAE-BA). We also present a new BA validation method that can sort the individuals into high-risk and low-risk groups, then use the RS to examine the health status of each group. Finally, we apply the algorithm to several real-world health datasets to demonstrate its validity and usefulness.

Our Contributions. Our study proposes methods that use the RS to estimate and validate BA. The main contributions of are as follows:

- We introduce a new method for health risk measurement, which represents the risk of clinical variables. We quantify the risk of each clinical variable into a score (i.e., RS) by considering its type of risk direction and distribution based on prior knowledge. Further, we use the RS in the BA estimation model construction, which allows our BA to include health risk information, and we also apply the RS to evaluation of BA from the model.
- We present RSAE, which is an unsupervised deep learning algorithm to estimate BA. Our undated autoencoder structure composed of a modified loss function to include the RS can generate embeddings that contain the health risk information and information from original health data specific to each person while capturing their non-linear patterns. Also, we propose a novel BA validation method using the RS that represents one's risk levels, which can be implemented to both unlabeled and labeled data. It can resolve limitations of existing methods that generally require cohort data labeled with mortality or disease incidence in evaluation.

- RSAE-BA enables people to diagnose their health status and provides precautionary information related to health care. This allows one's health to be quantitatively evaluated and can lead to hyper-personalized health management.

The structure of this paper is as follows. Section 2 describes the methodology for the new BA estimation and validation, which complements the limitations of existing methods. Section 3 shows the experimental results and comparison with the existing methods and Section 4 discusses the results and findings. Finally, Section 5 concludes with a brief summary.

2. Methodology

In this section, we describe a new method to estimate BA; the method consists of two parts: RSAE-BA estimation and BA validation.

2.1. RSAE-BA estimation

RSAE-BA estimation consists of four steps: RS calculation, data transformation, RSAE modeling, and BA calculation. Fig. 1 presents its overall process and detailed information on each step is described below.

2.1.1. RS calculation

Health conditions can be usually distinguished discretely into a normal or abnormal status (i.e., $\{0, 1\}$). However, individual conditions may differ in the same health condition and can be represented in a continuous form [31]. In this study, we introduce the RS to represent an individual's health risk as a continuous value. The assumption of RS is that the distribution of each variable on normal health status follows Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ under the central limit theorem [6] where the mean μ is determined using the normal criterion for a medical examination (i.e., medical prior knowledge) and the variance σ^2 is estimated from data samples empirically. The variable $\boldsymbol{x}_j \in \mathbb{R}^n, j = 1, \ldots, m$ in the EMR data $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ with n samples and m variables has its own independent Gaussian distribution ($\boldsymbol{x}_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$). Then for the RS calculation we consider three types of variables, which represent different directions of the health abnormality (Fig. 2): poor health is indicated by the large value (Type 1), by the small value (Type 2); by either the large or small value (Type 3).

The RS is calculated for $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ by the cumulative probability in the probability density function under the premise of $\mathcal{N}(\mu, \sigma^2)$. It is denoted as $\boldsymbol{P} \in [0, 1]^{n\times m}$, where $\boldsymbol{p}_i \in [0, 1]^m$ is the i-th sample of $\boldsymbol{P}$, and $\boldsymbol{p}_j \in [0, 1]^n$ is the j-th feature of $\boldsymbol{P}$. For $\boldsymbol{x}_j \in$Type 1 or $\boldsymbol{x}_j \in$Type 2, $\boldsymbol{p}_j$ of the one-sided Gaussian distribution is calculated in opposite directions. For $\boldsymbol{x}_j \in$Type 3, $\boldsymbol{p}_j$ is calculated under the two-sided Gaussian distribution since the health risk increases if the value of $\boldsymbol{x}_j$ is either larger or smaller than μ_j. Besides, the distribution used in the RS calculation can be easily extended to other distributions such as binomial or multinomial distribution according to data type (e.g., categorical data).

2.1.2. Data transformation

If the data and RS have different directions as in the three types, the information can be biased. Thus, to use both the original data and RS in RSAE modeling, the direction of the original data must be aligned beforehand with the direction of RS; this process (Algorithm 1) transforms original data $\boldsymbol{X}$ to new data $\boldsymbol{X}'$. First, we standardize the data in a feature-wise manner. For

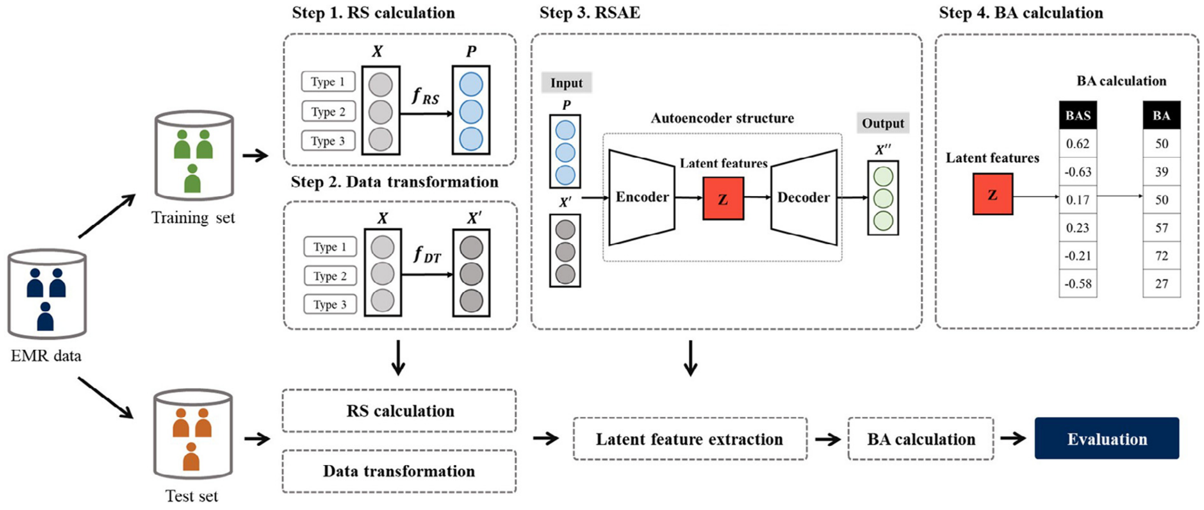
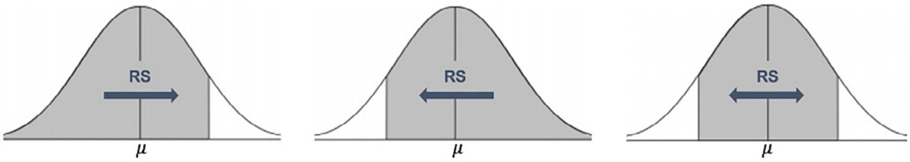


Fig. 1. Overall process of RSAE-BA estimation algorithm.



(a) Type 1
: poor health is indicated
by large value

(b) Type 2
: poor health is indicated
by small value

(c) Type 3
: poor health is indicated
by large or small value

Fig. 2. Three types of methods for RS calculation.

Type 1, we keep the value of variable as it is. For Type 2, we multiply by -1 to change the direction, and for Type 3, we convert the value to an absolute value. Finally, we use a sigmoid function to establish bounds of 0 and 1 on $\boldsymbol{x}'_j$ for all $j = 1, \ldots, m$, similar to the RS (i.e., $\boldsymbol{x}'_j \in [0,1]^n, \forall j = 1, \ldots, m$) where the calculated $\boldsymbol{x}'_j$ is the variable of the transformed data $\boldsymbol{X}'$; it can increase the learning stability of RSAE by adjusting the value range of input data. Thus, these data transformations avoid accumulating errors in the loss calculation of RSAE modeling, so the RSAE model can be trained effectively.

Algorithm 1: Data transformation

```
Input: Raw data $\boldsymbol{X}$
Output: Transformed input data $\boldsymbol{X}'$
function TRANS($\boldsymbol{X}$)
    $\boldsymbol{x}_j, j = 1, \ldots, m$ are the features of $\boldsymbol{X}$
    $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
    $\sigma_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}$
    $\mathbf{1}_n \in \mathbb{R}^n$ is an $n$-dimensional vector where all elements are 1
    for $j = 1$ to $m$ do
        $\boldsymbol{x}_j = (\boldsymbol{x}_j - \mu_j \mathbf{1}_n) \oslash (\sigma_j \mathbf{1}_n)$        ▹ $\oslash$ operator denotes element-wise division
        if $\boldsymbol{x}_j \in$ Type 2 then
            $\boldsymbol{x}'_j = -\boldsymbol{x}_j$
        else if $\boldsymbol{x}_j \in$ Type 3 then
            $\boldsymbol{x}'_j = |\boldsymbol{x}_j|$
        else
            $\boldsymbol{x}'_j = \boldsymbol{x}_j$
        end if
        $\boldsymbol{x}'_j = s(\boldsymbol{x}'_j)$        ▹ sigmoid function: $s(a) = \frac{1}{1+e^{-a}}$
    end for
end function
$\boldsymbol{X}' = $ TRANS($\boldsymbol{X}$)
```

2.1.3. RSAE modeling

An autoencoder, a neural network consisting of an encoder network and a decoder network, has been successfully used for feature extraction about various types of data in many applications [49,43]. Parameters of the autoencoder are optimized by minimizing the restoration error that accumulates during the process of projecting the input data to lower dimensions, then restoring the input data from the projected data [4]. The autoencoder learns a representation of data by capturing both the non-linear characteristics of various variables and their dependency in an unsupervised learning fashion [13]. In this study, we modify the vanilla autoencoder to extract latent features that are more beneficial in BA estimation than those of the vanilla one (Algorithm 2 and Fig. 3).

Let $\boldsymbol{H}_{Enc}^{(0)}$ be the input layer and $\boldsymbol{H}_{Enc}^{(k)} \in \mathbb{R}^{n \times D^{(k)}}$ be the k^{th} hidden layer for $\forall k \in 1, \ldots, K$ where K is the number of hidden layers and $D^{(k)}$ is the dimensionality of $\boldsymbol{H}_{Enc}^{(k)}$ encoded from $\boldsymbol{X}'$. $\boldsymbol{Z} \in \mathbb{R}^{n \times D^{(K)}}$ is the latent representation after K layers of the encoder of RSAE:

$$\boldsymbol{H}_{Enc}^{(0)} = \boldsymbol{X}', \tag{1}$$

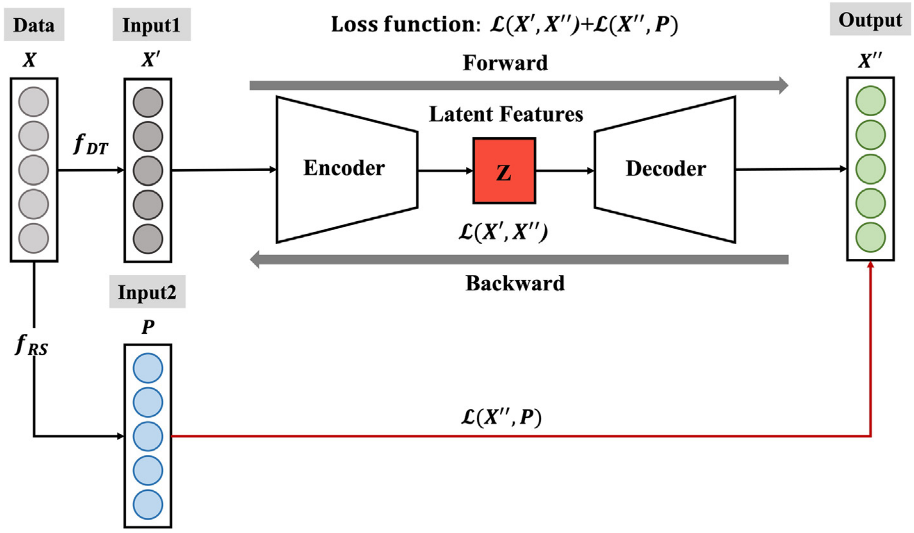


Fig. 3. Structure of RSAE.

$$\boldsymbol{H}^{(k)}_{Enc} = f_{ReLU}(\boldsymbol{H}^{(k-1)}_{Enc}\boldsymbol{W}^{(k)}_{Enc} + \boldsymbol{1}_n \cdot \boldsymbol{b}^{(k)T}_{Enc}), \tag{2}$$

$$\boldsymbol{Z} = \boldsymbol{H}^{(K)}_{Enc}, \tag{3}$$

where $\boldsymbol{W}^{(k)}_{Enc} \in \mathbb{R}^{D^{(k-1)} \times D^{(k)}}$ and $\boldsymbol{b}^{(k)}_{Enc} \in \mathbb{R}^{D^{(k)}}$ present respectively the weights and biases for the encoder of RSAE. $f_{ReLU}(a) = max(a, 0)$ is a rectified linear unit (ReLU) activation function for RSAE modeling. We use $\boldsymbol{H}^{(0)}_{Dec}$ to represent the input layer of the decoder of RSAE, which is equal to $\boldsymbol{Z}$. Then, we define

$$\boldsymbol{H}^{(k)}_{Dec} = f_{ReLU}(\boldsymbol{H}^{(k-1)}_{Dec}\boldsymbol{W}^{(k)}_{Dec} + \boldsymbol{1}_n \cdot \boldsymbol{b}^{(k)T}_{Dec}) \in \mathbb{R}^{n \times D^{(K-k)}}, \tag{4}$$

as the k^{th} hidden layer, where $\boldsymbol{W}^{(k)}_{Dec} \in \mathbb{R}^{D^{(K-k+1)} \times D^{(K-k)}}$ and $\boldsymbol{b}^{(k)}_{Dec} \in \mathbb{R}^{D^{(K-k)}}$ are the weights and biases for decoder of RSAE. We also define

$$\boldsymbol{X}'' = \boldsymbol{H}^{(K)}_{Dec}, \tag{5}$$

as the reconstructed data decoded from $\boldsymbol{Z}$. The key to this algorithm is to use the RS information as well as original data (after transformation); the modified loss function is defined as

$$\mathcal{L}_{RSAE} = \mathcal{L}(\boldsymbol{X}', \boldsymbol{X}'') + \mathcal{L}(\boldsymbol{X}'', \boldsymbol{P}), \tag{6}$$

where $\mathcal{L}(\boldsymbol{X}', \boldsymbol{X}'')$ is the reconstruction error of $\boldsymbol{X}'$ and $\mathcal{L}(\boldsymbol{X}'', \boldsymbol{P})$ is the difference error between $\boldsymbol{X}''$ and $\boldsymbol{P}$. $\mathcal{L}(x, y)$ is Kullback–Leibler (KL) divergence loss between x and y [34]. Unlike the vanilla autoencoder, these equations enable extraction of the $D^{(K)}$-dimensional RSAE latent features $\boldsymbol{z}_d \in \mathbb{R}^n$ $(d = 1, \ldots, D^{(K)})$ of $\boldsymbol{Z}$, which contain the risk information of the RS as well as capture the non-linearity of both the RS and original data.

Algorithm 2: RSAE training

```
Input: Transformed input data $\boldsymbol{X}'$, risk score $\boldsymbol{P}$
Output: Latent feature $\boldsymbol{Z}$
procedure RSAE($\boldsymbol{X}', \boldsymbol{P}, D^{(k)}, K, \eta, \epsilon, max\_iter$)
    $\boldsymbol{\theta} = \{\boldsymbol{W}^{(k)}_{Enc}, \boldsymbol{b}^{(k)}_{Enc}, \boldsymbol{W}^{(k)}_{Dec}, \boldsymbol{b}^{(k)}_{Dec}\}$ is the set of the parameters of autoencoder
    $\boldsymbol{H}^{(k)}_{Enc} \in \mathbb{R}^{n \times D^{(k)}}$ is the $k^{th}$ hidden layer of encoder for $\forall k \in 1, \ldots, K$
    $\boldsymbol{H}^{(k)}_{Dec} \in \mathbb{R}^{n \times D^{(K-k)}}$ is the $k^{th}$ hidden layer of decoder for $\forall k \in 1, \ldots, K$
    $D^{(k)}$ is the dimensionality of $\boldsymbol{H}^{(k)}_{Enc}$ (i.e., the number of hidden units)
    $K$ is the number of hidden layers
    $\eta$ is the learning rate
    $\epsilon$ is the convergence criterion
```

a (*continued*)

Algorithm 2: RSAE training

```
  $max\_iter$ is the maximul number of iteration
  **Initialize**: $\theta^{(0)}$, $\mathcal{L}^{(0)}$
  $iter \leftarrow 0$
  **repeat**
    $iter \leftarrow iter + 1$
    $\boldsymbol{H}^{(0)}_{Enc} \leftarrow \boldsymbol{X}'$
    **for** $k \in 1, \ldots, K$ **do**
      $\boldsymbol{H}^{(k)}_{Enc} \leftarrow f_{ReLU}(\boldsymbol{H}^{(k-1)}_{Enc}\boldsymbol{W}^{(k)}_{Enc} + \boldsymbol{1}_n \cdot \boldsymbol{b}^{(k)T}_{Enc})$
    **end for**
    $\boldsymbol{Z} = \boldsymbol{H}^{(K)}_{Enc}$
    $\boldsymbol{H}^{(0)}_{Dec} \leftarrow \boldsymbol{Z}$
    **for** $k \in 1, \ldots, K$ **do**
      $\boldsymbol{H}^{(k)}_{Dec} \leftarrow f_{ReLU}(\boldsymbol{H}^{(k-1)}_{Dec}\boldsymbol{W}^{(k)}_{Dec} + \boldsymbol{1}_n \cdot \boldsymbol{b}^{(k)T}_{Dec})$
    **end for**
    $\boldsymbol{X}'' = \boldsymbol{H}^{(K)}_{Dec}$
    $\mathcal{L} = \mathcal{L}(\boldsymbol{X}', \boldsymbol{X}'') + \mathcal{L}(\boldsymbol{X}'', \boldsymbol{P})$
    $\boldsymbol{g}$ = compute the gradients of the loss with respect to $\boldsymbol{\theta}$
    $\boldsymbol{\theta}^{(iter+1)} \leftarrow \boldsymbol{\theta}^{(iter)} - \eta \times \boldsymbol{g}^{(iter)}$
  **until** $|\mathcal{L}^{(iter+1)} - \mathcal{L}^{(iter)}| < \epsilon$ or $iter < max\_iter$
  **return** $\boldsymbol{\theta}$
**end procedure**
```

2.1.4. BA calculation

To estimate BA, the BA score (BAS) should be calculated by using the extracted $\boldsymbol{z}_d$ as

$$BAS = \frac{1}{D^{(K)}} \sum_{d=1}^{D^{(K)}} (\boldsymbol{z}_d - z^{\mu}_d \boldsymbol{1}_n) \oslash (z^{\sigma}_d \boldsymbol{1}_n), \tag{7}$$

where $z^{\mu}_d = \frac{1}{n}\sum_{i=1}^{n} z_{id}$ and $z^{\sigma}_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_{id} - z^{\mu}_d)^2}, i = 1, \ldots, n$. The n-dimensional vector BAS is an inadequate value for a general age scale, so we convert the BAS to the BA to match the range of possible age values. We transform individual BAS to BA in terms of years by using the T-scale idea, which is a technique to transform from a standard score to a T-score [14]. Finally, RSAE-BA can be generated.

2.2. BA validation

To assess the feasibility of BA estimates, we introduce a new validation method that uses the RS, and that can be used in unlabeled data beyond cohort data that are labeled with mortality or disease incidence. We first sort all individuals into two groups: Group 1 that has BA $<$ CA and Group 2 that has BA $>$ CA. Then we define a user-defined parameter q and define a low-risk (LR) group as the q% of samples in Group 1 that have the largest CA - BA, and a high-risk (HR) group as the q% of

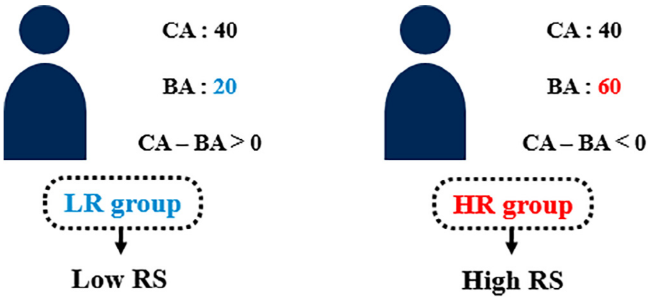


Fig. 4. Example of LR group and HR group for BA validation.

samples in Group 2 that have the largest BA - CA (Fig. 4). By exploiting the characteristic of RS, we can construct two hypotheses: (1) The RS is lower for participants in the LR group than in the HR group, and (2) well-estimated BA would give a good distinction of the RS difference between LR and HR groups. Finally, we determine whether the BA estimates satisfy these hypotheses.

3. Results

3.1. Data description and preprocessing

To verify the effectiveness of RSAE-BA and to show the diverse results regarding both the RS and disease incidence based on our BA validation method, experiments were conducted on three real-world datasets that had distinct characteristics, one is unlabeled data and the others are labeled data with disease incidence (Table 1). The dataset collected by the Korea National Health and Nutrition Examination Survey (KNHANES) is a result of an annual survey from 1988 to 2017 to identify the health and nutritional status of Koreans [35]. A sample-cohort database, collected from the Korean National Health Insurance Corporation (NHIC) from 2002 to 2010, includes health check-up and other information such as medical treatments of 2% of the population stratified by age, sex, and income level. From NHIC database, we used two cohorts: one (NHIC–HTN) to study determinants of hypertension (HTN) occurrence in the two and half years after the check-up [37], and the other (NHIC-T2DM) is about type-2 diabetes mellitus (T2DM); both were collected in the same way. To be specific, there is the main difference among these datasets. KNHANES is a cross-sectional study based on a one-time examination, which cannot provide disease incidence labels for prediction. On the contrary, NHIC–HTN and NHIC-T2DM are longitudinal studies (i.e., cohort studies) based on follow-up examinations, which include the incidence labels of certain diseases to predict future disease onset.

The normal criteria and types of variables for RS calculation were obtained for each dataset (see Tables A.1 and A.2). We set the gender-specific values for the normal criteria of each variable since they can be different by male and female. Furthermore, KNHANES dataset has many missing values, so we used 21 common clinical variables in all years after excluding variables that had $> 40\%$ missing values. Missing information can lead to biased estimates, so the missing data must be imputed to preserve as much information as possible [15]. To impute the missing values, we used fast unified random forests for survival, regression, and classification (RF-SRC) [23], which presented the best result in the preliminary experiments (Appendix B). For NHIC–HTN and NHIC-T2DM, we selected 10 clinical variables that represent diverse physiological systems that are commonly available in various examinations. NHIC–HTN and NHIC-T2DM have few missing values, so we excluded all samples that had any missing values.

3.2. Experimental settings and benchmark methods

The RSAE-BA estimation algorithm in this study involves the RS calculation, data transformation, RSAE training, and BA calculation processes. For each of three datasets, we first calculated the RS of data samples based on the Gaussian distribution whose mean and variance were defined by the normal criteria and empirical sample variance of each variable, respectively. Next, we implemented the data transformation method on each dataset, and then we conducted the RSAE training. We used several hyperparameters for RSAE training, including the number of hidden layers, hidden layer units, learning rate, and epoch numbers, which had been tuned appropriately (see Table C.1). After the RSAE training, we used the latent features extracted from the RSAE model to calculate a BAS. Then, we transformed the BAS to BA by applying the *T*-scale method. To avoid the problem of too large or too small BAS leading to an abnormal BA (e.g., age < 0), we bounded BAS in the interval $[-2, 2]$ and then calculated stable BA estimates. To validate the result of the proposed algorithm, we compared RSAE-BA with the most widely-used methods in BA estimation, which are previously described: MLR, PCA, and KDM, where the BA estimates are denoted as MLR-BA, PCA-BA, and KDM-BA, respectively. We additionally estimated BA using non-negative matrix factorization (NMF) [36] (denoted as NMF-BA), which is a naïve extension of PCA, to increase the diversity of comparison results. The accuracy of the BA estimation models was evaluated according to the averaged RS for the unlabeled data, and according to the disease-incidence ratio for the labeled data. We used 10-fold cross-validation to measure the accuracy of the five BA models for performance comparison. All experiments were implemented on Pytorch 1.4.0 in Python 3.6 with an i7-8700 CPU, 32 GB of RAM, and a single 16-GB NVidia RTX 2080Ti GPU.

Table 1
Descriptions of three datasets for BA estimation.

Dataset	Source	# of variables	# of samples	Imbalance ratio
KNHANES	KNHANES	21	85,460	-
NHIC–HTN	NHIC	10	149,967	7%
NHIC-T2DM	NHIC	10	145,718	11%

Notes: Imbalance ratio is the ratio of the number of samples in the majority class to the number of samples in the minority class. That is, it is the ratio of disease-free incidence to disease incidence in the datasets in this study.

3.3. Validation of BA estimates based on RS

To test the hypotheses of BA validation method using the RS, we compared the RS averages of the LR and HR groups, as produced by the five BA estimation models. We implemented data categorization based on the age groups to show diverse BA validation results regarding people of their age. The LR and HR groups consist of the botton 5% and top 5% respectively of samples from the age groups '30s', '40s', '50s', and '60s'. Average RS of LR and HR groups were taken for age-stratified subsamples (Table 2). Compared to other BA algorithms, RSAE-BA gave the lowest RS average in all age groups for the LR group, but the highest RS average for the HR group. We also used the *t*-test with three significance levels ($\alpha = 0.05, 0.01, 0.0001$) to determine significance of differences in average RS between LR and HR groups (Table 3). RSAE-BA obtained statistically significant difference between the two groups for all age groups ($p < 0.0001$). We gradually increased q from 1 to 10 to check the accuracy of models at different sampling ratios (Fig. 5). In all three datasets, RSAE-BA was more accurate than the other baseline methods at all sampling ratios.

3.4. Validation of BA estimates based on disease incidence

To examine the relationship between RSAE-BA and disease incidence, our model and four baseline models were evaluated in terms of HTN and T2DM incidence ratio in the NHIC–HTN and NHIC-T2DM datasets. In the same way as RS-based validation, BAs were estimated based on the five models, and then LR and HR groups were established according to the difference between CA and BA and vice versa. Then we tested two hypotheses related to disease incidence: (1) participants in the HR group would have higher disease incidence than the LR group, and (2) well-estimated BA would differentiate the disease incidence between LR and HR groups effectively.

We checked the disease incidence in two groups, then used the *t*-test to statistically quantify the difference in average disease incidence between two groups. RSAE-BA showed the largest difference of incidence between LR and HR groups in all age groups in both datasets (Table 4). Especially in RSAE-BA, the difference in the disease incidence is $>$ 20% for HTN in 60s age groups and $>$ 25% for T2DM in the 50s and 60s age groups; this result means that people in the HR group have a much higher potential risk of hypertension and diabetes than those in the LR group. Results of the disease-onset rate were compared at different q from 1 to 10% for the five models (Figs. 6 and 7). RSAE-BA outperforms other BA models in all age groups and at all sampling ratios.

Table 2
Average and standard deviation (SD) of RS of LR and HR groups by age when q=5 for five BA estimates.

Dataset	Risk group	Model	Age			
			30s	40s	50s	60s
KNHANES	LR	**RSAE-BA**	**0.1698** ± **0.0042**	**0.1990** ± **0.0020**	**0.2637** ± **0.0055**	**0.3033** ± **0.0030**
		KDM-BA	0.1887 ± 0.0013	0.2235 ± 0.0005	0.2952 ± 0.0017	0.3224 ± 0.0030
		MLR-BA	0.2980 ± 0.0001	0.3299 ± 0.0003	0.4009 ± 0.0012	0.4176 ± 0.0074
		NMF-BA	0.2005 ± 0.0035	0.2341 ± 0.0006	0.2940 ± 0.0043	0.3431 ± 0.0001
		PCA-BA	0.2379 ± 0.0010	0.2689 ± 0.0004	0.3456 ± 0.0038	0.3925 ± 0.0007
	HR	**RSAE-BA**	**0.6745** ± **0.0035**	**0.7059** ± **0.0043**	**0.7114** ± **0.0039**	**0.7057** ± **0.0051**
		KDM-BA	0.6592 ± 0.0028	0.6880 ± 0.0014	0.6844 ± 0.0023	0.6732 ± 0.0027
		MLR-BA	0.5377 ± 0.0069	0.5873 ± 0.0001	0.6047 ± 0.0032	0.5975 ± 0.0037
		NMF-BA	0.6453 ± 0.0026	0.6805 ± 0.0025	0.6773 ± 0.0022	0.6696 ± 0.0006
		PCA-BA	0.6169 ± 0.0001	0.6413 ± 0.0005	0.6403 ± 0.0011	0.6250 ± 0.0012
NHIC–HTN	LR	**RSAE-BA**	**0.3682** ± **0.0037**	**0.3749** ± **0.0042**	**0.3818** ± **0.0041**	**0.3680** ± **0.0068**
		KDM-BA	0.4986 ± 0.0023	0.5141 ± 0.0015	0.5298 ± 0.0031	0.5195 ± 0.0080
		MLR-BA	0.6323 ± 0.0011	0.6139 ± 0.0004	0.6190 ± 0.0038	0.5812 ± 0.0011
		NMF-BA	0.3876 ± 0.0005	0.3965 ± 0.0002	0.4004 ± 0.0007	0.3950 ± 0.0042
		PCA-BA	0.4484 ± 0.0003	0.4653 ± 0.0002	0.4748 ± 0.0073	0.4503 ± 0.0043
	HR	**RSAE-BA**	**0.8615** ± **0.0054**	**0.8531** ± **0.0081**	**0.8434** ± **0.0065**	**0.8276** ± **0.0112**
		KDM-BA	0.7634 ± 0.0019	0.7480 ± 0.0038	0.7171 ± 0.0034	0.6873 ± 0.0124
		MLR-BA	0.6536 ± 0.0008	0.6780 ± 0.0008	0.6848 ± 0.0053	0.6516 ± 0.0047
		NMF-BA	0.8483 ± 0.0016	0.8376 ± 0.0001	0.8277 ± 0.0041	0.8048 ± 0.0014
		PCA-BA	0.8340 ± 0.0001	0.8218 ± 0.0028	0.8041 ± 0.0024	0.7783 ± 0.0075
NHIC-T2DM	LR	**RSAE-BA**	**0.3801** ± **0.0025**	**0.3920** ± **0.0026**	**0.4003** ± **0.0038**	**0.3915** ± **0.0031**
		KDM-BA	0.5457 ± 0.0052	0.5515 ± 0.0017	0.5591 ± 0.0001	0.5453 ± 0.0065
		MLR-BA	0.6527 ± 0.0067	0.6274 ± 0.0038	0.6273 ± 0.0022	0.5909 ± 0.0116
		NMF-BA	0.3839 ± 0.0011	0.3954 ± 0.0010	0.4039 ± 0.0029	0.3996 ± 0.0006
		PCA-BA	0.4351 ± 0.0001	0.4518 ± 0.0004	0.4656 ± 0.0031	0.4544 ± 0.0049
	HR	**RSAE-BA**	**0.8676** ± **0.0026**	**0.8543** ± **0.0027**	**0.8490** ± **0.0037**	**0.8288** ± **0.0043**
		KDM-BA	0.7068 ± 0.0044	0.7083 ± 0.0003	0.6990 ± 0.0025	0.6711 ± 0.0010
		MLR-BA	0.6314 ± 0.0040	0.6564 ± 0.0017	0.6728 ± 0.0011	0.6574 ± 0.0044
		NMF-BA	0.8548 ± 0.0015	0.8405 ± 0.0019	0.8360 ± 0.0048	0.8147 ± 0.0006
		PCA-BA	0.8354 ± 0.0021	0.8178 ± 0.0009	0.8079 ± 0.0021	0.7922 ± 0.0002

Table 3
Averaged and SD of RS difference between LR and HR groups by age when $q = 5$ for five BA estimates.

Dataset	Model	Age			
		30s	40s	50s	60s
KNHANES	**RSAE-BA**	**0.5046 ± 0.0069*****,*	**0.5069 ± 0.0054*****,*	**0.4477 ± 0.008*****,*	**0.4024 ± 0.0074*****,*
	KDM-BA	0.4705 ± 0.0041***	0.4644 ± 0.0009***	0.3892 ± 0.0006***	0.3508 ± 0.0056***
	MLR-BA	0.2397 ± 0.0070***	0.2573 ± 0.0003***	0.2038 ± 0.0020**	0.1799 ± 0.0037**
	NMF-BA	0.4448 ± 0.0061***	0.4464 ± 0.0019***	0.3833 ± 0.0022***	0.3265 ± 0.0007***
	PCA-BA	0.3790 ± 0.0010***	0.3725 ± 0.0001***	0.2948 ± 0.0028**	0.2325 ± 0.0005**
NHIC–HTN	**RSAE-BA**	**0.4933 ± 0.0086*****,*	**0.4783 ± 0.0121*****,*	**0.4616 ± 0.0101*****,*	**0.4596 ± 0.0127*****,*
	KDM-BA	0.2648 ± 0.0004**	0.2338 ± 0.0052**	0.1873 ± 0.0065*	0.1677 ± 0.0044*
	MLR-BA	0.0213 ± 0.0004	0.0641 ± 0.0004	0.0658 ± 0.0015	0.0704 ± 0.0058
	NMF-BA	0.4606 ± 0.0020	0.4411 ± 0.0002	0.4272 ± 0.0048	0.4098 ± 0.0028
	PCA-BA	0.3856 ± 0.0002***	0.3564 ± 0.0030***	0.3293 ± 0.0097***	0.3279 ± 0.0032**
NHIC-T2DM	**RSAE-BA**	**0.4874 ± 0.0042*****,*	**0.4624 ± 0.0041*****,*	**0.4487 ± 0.0057*****,*	**0.4373 ± 0.0054*****,*
	KDM-BA	0.1610 ± 0.0009	0.1568 ± 0.0014	0.1399 ± 0.0025	0.1258 ± 0.0075
	MLR-BA	−0.0213 ± 0.0027	0.0289 ± 0.0022	0.0455 ± 0.0011	0.0664 ± 0.0072
	NMF-BA	0.4709 ± 0.0027***	0.4451 ± 0.0029***	0.4321 ± 0.0076***	0.4151 ± 0.0012***
	PCA-BA	0.4003 ± 0.0022***	0.3660 ± 0.0013***	0.3423 ± 0.0011***	0.3378 ± 0.0051**

Notes:*** $p < 0.0001$. ** $p < 0.01$. * $p < 0.05$.; p is denoted by p-value for t-test. The first group of asterisks indicates statistical significance for difference between the LR and HR groups; the second group of asterisks refers to difference between BA models.

4. Discussion

This study provides several methodological advances.

1. This study introduced the RS which can represent the degree of risk of a specific clinical variable quantitatively by considering the data distribution. The RS increases the detail provided by health risk information. We also constructed a new autoencoder model structure that can embed the RS in the model learning. Since the RS calculated from the Gaussian distribution has non-linear characteristics, RSAE can extract latent features that represent both original data and the potential health risk, better than the vanilla autoencoder can.
2. In general, an ‘index’ is an expression of the state of an object by reference to a specific value or level in the data for a given purpose (e.g., body mass index is calculated from height and weight). However, these indices cannot easily represent the features of various variables and their inter-relationship simultaneously; this condition demonstrates the need for an index that uses machine learning, especially deep learning, to efficiently and effectively represent the interrelations and characteristics among a large number of variables. Our BA model uses a deep autoencoder that can extract the latent representations of original variables from a large amount of unlabeled data. It can learn non-linear characteristics of the variables automatically, whereas linear models such as PCA cannot accomplish this task.
3. Indices do not have an actual target (label), so they cannot be validated by direct comparison with one. Therefore, indices have mainly been validated by consulting experts in the field; this approach is time consuming and expensive. For index validation, we presented a novel method that can identify risk levels by dividing the samples into LR and HR groups and evaluating the RS in each group.
4. We demonstrated the feasibility of RSAE-BA by comparing the existing BA estimation methods in terms of the RS and disease incidence. RSAE-BA also showed higher accuracy of prediction of disease incidence than the other methods (Table 4, Fig. 6, and Fig. 7); this success means that the RS is also highly associated with the health conditions with certain diseases. Also, the validation method using the RS enables us to evaluate the estimated BA regardless of the types of data (i.e., unlabeled or labeled data). It can significantly improve the usability of BA estimates by solving the limitations of existing studies related to BA estimation and validation.

Our BA model has several managerial implications in the healthcare industry. The use of a weight scale that can simply measure body compositions has been popularized, so this trend to provide a more comprehensive and supplementary index including BA rather than the body indices itself through its own software or connection with a healthcare service application (app). This study proposed a new BA estimation method that is more accurate than the existing algorithms, and that enables such healthcare products or apps to utilize accurate health index. People can easily and conveniently compare their BA and CA to understand their health status at a glance. Private companies related to BA estimation systems such as ‘Inner-age’ [22] and ‘Bio-age’ [7] have begun to provide BA management services by using body measurement and blood tests to measure overall BA. Beyond the function as an overall health index, this study provides individuals with the RS information to intuitively identify their risk level group (HR or LR group) and to identify clinical variables that have high potential risk. Thus, RSAE-BA can raise the awareness of the need for health management, and lead healthcare companies to build personalized healthcare services.

Nonetheless, the proposed method has some limitations. First, we assumed that variables are independent of each other and that each has a Gaussian distribution. The RS calculated using these assumptions may be questionable when information

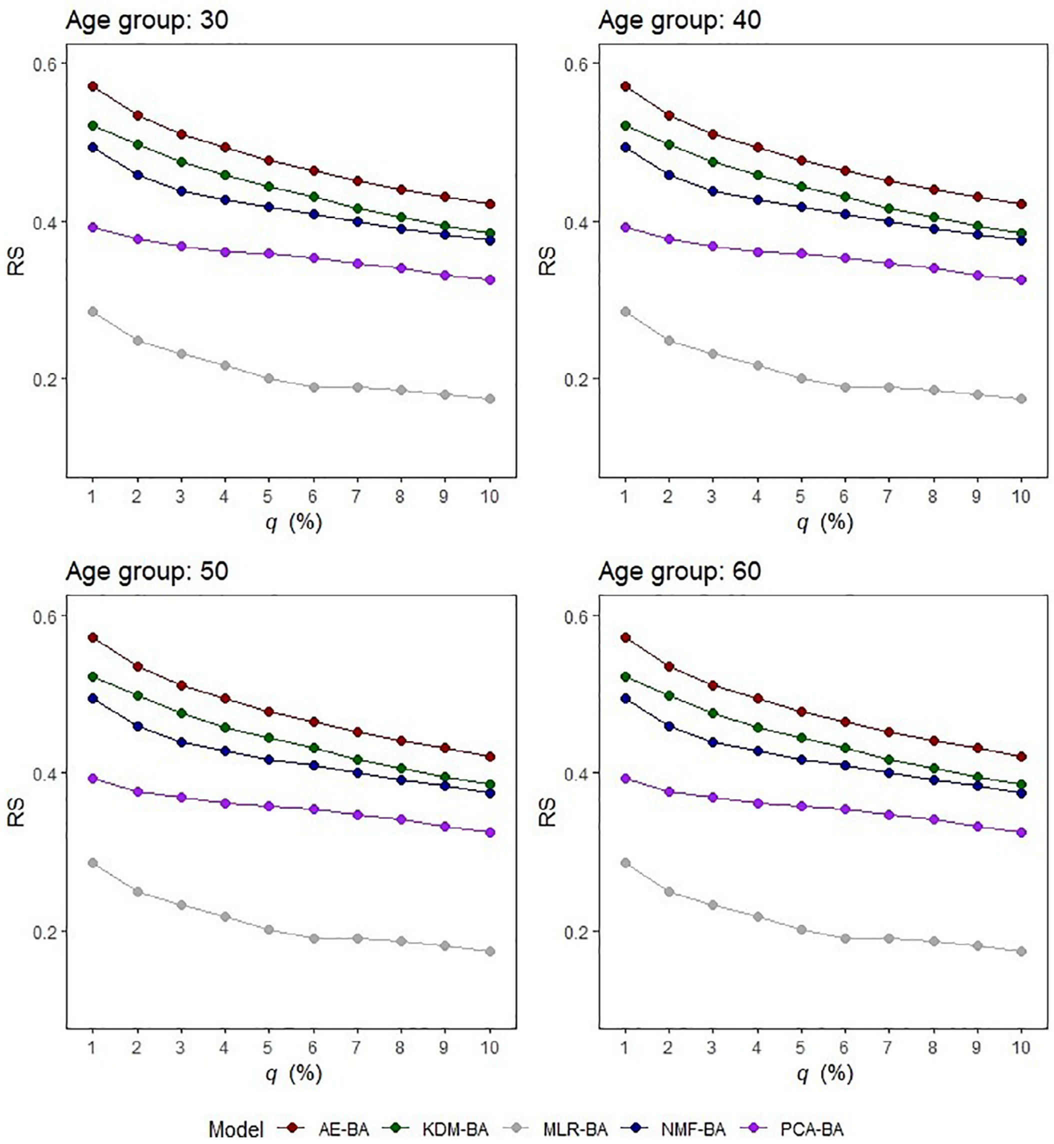


Fig. 5. RS comparison results of five BA estimates in terms of RS difference between LR and HR groups by age and sampling ratio.

Table 4
Disease incidence difference between LR and HR groups by age when q=5 for five BA estimates.

Disease	Model	Age			
		30s	40s	50s	60s
HTN	**RSAE-BA**	**0.1943** ± **0.0094*****,*	**0.1903** ± **0.0154*****,*	**0.1989** ± **0.0247*****,*	**0.2084** ± **0.0252*****,*
	KDM-BA	0.1034 ± 0.0069***	0.1038 ± 0.0203***	0.0825 ± 0.0002***	0.0132 ± 0.0137
	MLR-BA	−0.0047 ± 0.0096	0.0296 ± 0.0100	0.0271 ± 0.0012	0.0014 ± 0.0111
	NMF-BA	0.1593 ± 0.0077***	0.1527 ± 0.0032***	0.1379 ± 0.0016***	0.1254 ± 0.0097**
	PCA-BA	0.0983 ± 0.0008***	0.0460 ± 0.0027*	0.0330 ± 0.0050	0.0279 ± 0.0163
T2DM	**RSAE-BA**	**0.1728** ± **0.0099*****,*	**0.2347** ± **0.0119*****,*	**0.2535** ± **0.0251*****,*	**0.2568** ± **0.0192*****,*
	KDM-BA	0.0975 ± 0.0184***	0.1703 ± 0.0053***	0.2017 ± 0.0330***	0.1481 ± 0.0382***
	MLR-BA	−0.0141 ± 0.0081	0.0930 ± 0.0168***	0.1282 ± 0.0269***	0.0979 ± 0.0652
	NMF-BA	0.1559 ± 0.0155***	0.1956 ± 0.0079***	0.1976 ± 0.0252***	0.1944 ± 0.0185***
	PCA-BA	0.1204 ± 0.0014***	0.1781 ± 0.0026***	0.1902 ± 0.0076***	0.2398 ± 0.0159***

Notes:*** $p < 0.0001$. ** $p < 0.01$. * $p < 0.05$.; p is denoted by p-value for t-test. The first group of asterisks indicates statistical significance for difference between the LR and HR groups; the second group of asterisks refers to difference between BA models.

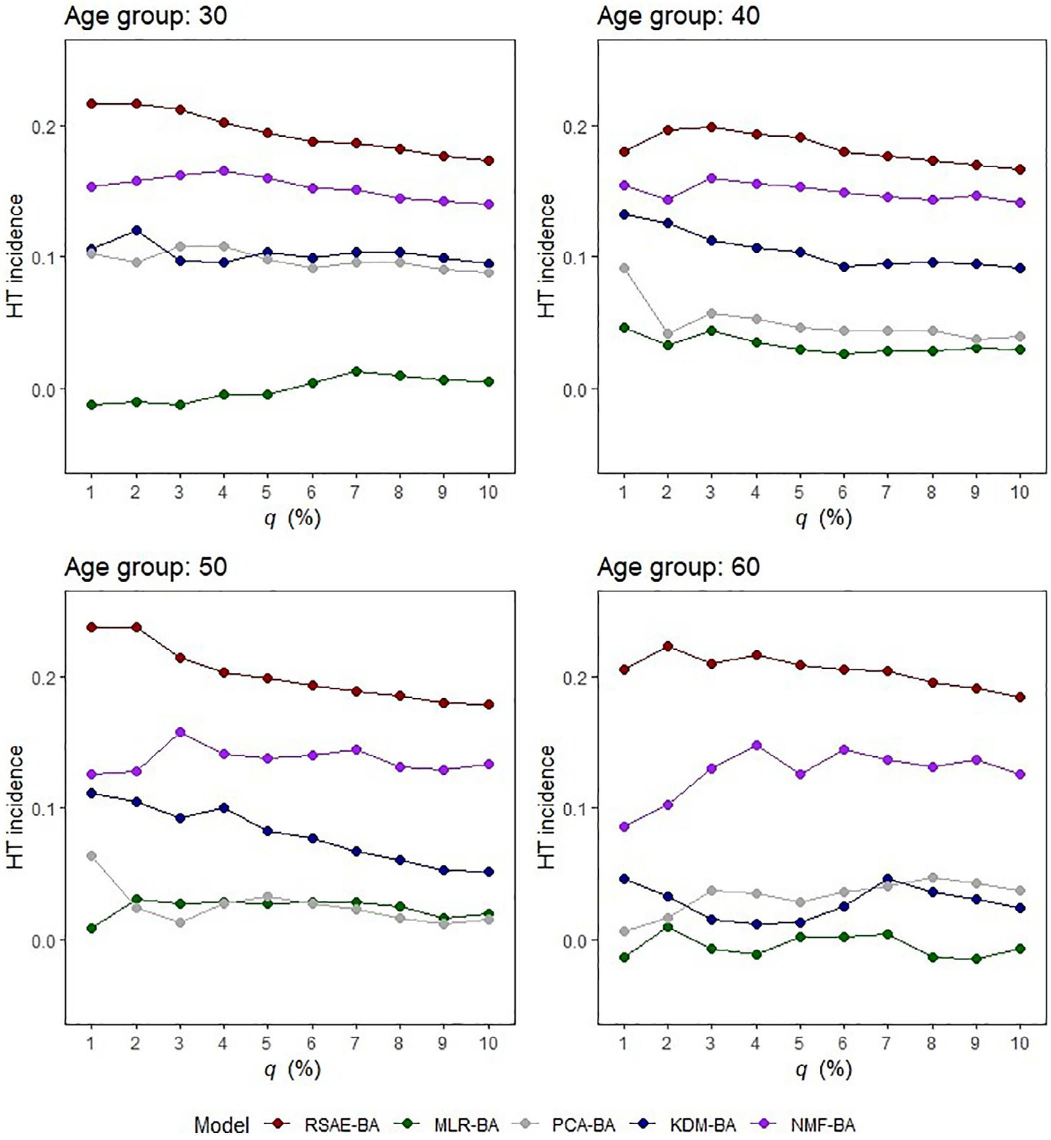


Fig. 6. HTN incidence difference between LR and HR groups by age and sampling ratio for five BA estimates.

about the types and normal criteria of the variables is insufficient. This problem should be studied using different empirical distributions in the RS calculation. In addition, although the experimental results from deep-representation learning are accurate, the relationships between the original variables and latent features extracted from the RSAE are not easily identified due to its intrinsic 'black-box' nature [5,44].

Several future research directions could improve our study. First, it can be extended to BAs that are specific to body functions such as cardiovascular age, immune age, and kidney age. Further, the concept of disease-specific BA (e.g., diabetes age) could be refined by supervised learning using a slight modification of our proposed method. Second, advances in healthcare are yielding new health-related data about such topics as DNA methylation, physical activity, and metabolites, so RSAE-BA is applicable to such data. Third, the RS-embedded method that uses the modified loss function can be applied to other manifold learning methods such as PCA and NMF, and a comparison of their results with RSAE-BA might yield interesting insights. Fourth, the RS was defined as the relative level of a person's health risk to others in this study, but we can also consider the risk level of the same person's health data in a different period as the absolute RS (i.e., the personalized RS). Thus, the RSAE-BA model is able to be extended by reflecting both the relative and absolute RS of each person's health status simultaneously into the RSAE-BA estimation. Lastly, the approach used in this study can be used to develop diverse indices in other fields beyond healthcare, such as finance, insurance, and transportation safety.

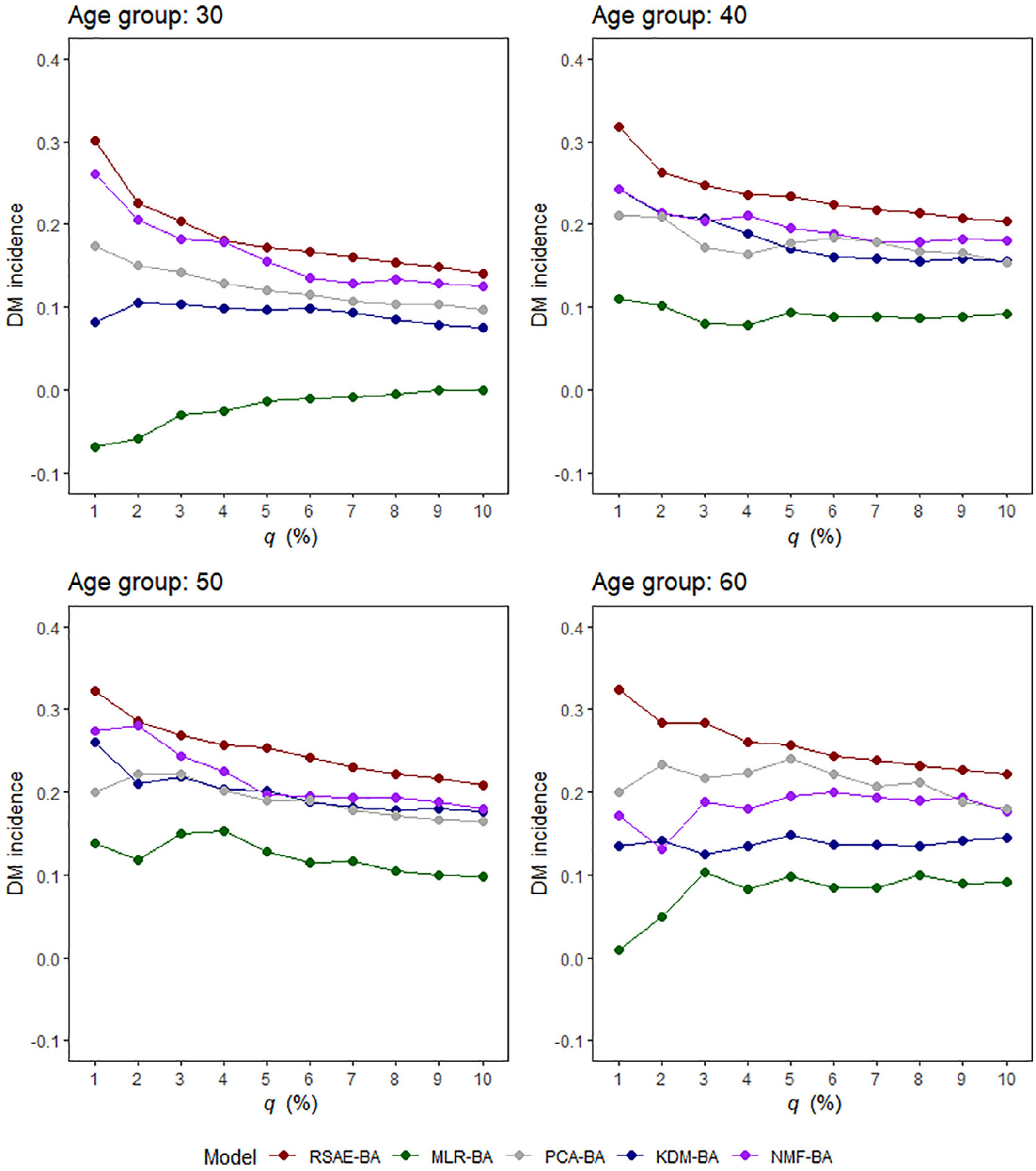


Fig. 7. T2DM incidence difference between LR and HR groups by age and sampling ratio for five BA estimates.

5. Conclusion

This study presented a new BA estimation method called RSAE-BA, which uses an autoencoder to achieve deep representation learning. The key idea behind this is the concept of RS, which can measure people's health risk level into BA estimation and validation. This technique enables extraction of latent features that aggregate information of both original data and the RS by embedding the RS into autoencoder learning. Further, we proposed BA-validation method that uses the RS by defining high-risk and low-risk groups based on BAs and comparing their average RSs. We conducted experiments on three datasets with or without labels to demonstrate the usefulness of the proposed method. It provides significant improvement in accuracy of the RS and disease incidence compared with existing methods. We believe that our approach can provide an accurate health index and lead to improvements in personal health management.

CRediT authorship contribution statement

Suhyeon Kim: Conceptualization, Methodology, Software, Writing - original draft. **Hangyeol Kim:** Conceptualization, Methodology, Software. **Eun-Sol Lee:** Data curation, Investigation. **Chiehyeon Lim:** Supervision, Writing - review & editing. **Junghye Lee:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by the National Research Foundation (NRF) grant funded by the Korea Government (MSIT) under Grant No. 2020R1C1C1011063 and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the MIST (No. 2020–0-01336, Artificial Intelligence graduate school support (UNIST)). This work was also supported by Korea Environment Industry & Technology Institute (KEITI) through (Digital Infrastructure Building Project for Monitoring, Surveying and Evaluating the Environmental Health), funded by Korea Ministry of Environment (MOE) (No. 2021003330001). Also, this work was supported by the Establishment of Demonstration Infrastructure for Regulation-Free Special Zones funded by the Ministry of SMEs and Startups (MSS, Korea) (No. P0016191).

Appendix A. Baseline characteristics

Table A.1
Baseline characteristics and clinical criteria of KNHANES.

Variable	Mean (SD)	Criteria for normality		Type	Missing ratio (%)
		Male	Female		
CA	49.22 (16.61)	-	-	-	0.0
BMI (kg/m^2)	23.66 (3.39)	21.75	21.75	3	0.0
WC (cm)	81.37 (9.92)	-	-	1	0.2
SBP (mmHg)	119.59 (21.25)	105	105	1	0.9
DBP (mmHg)	75.6 (13.07)	70	70	1	0.9
GLU0 (mg/dL)	98.97 (24.32)	90	90	1	4.9
TC (mg/dL)	188.8 (36.38)	150	150	1	4.8
HMG (g/dL)	13.93 (1.61)	14.75	13.75	3	5.0
SGOT (U/L)	23.25 (14.28)	20	20	1	4.7
SGPT (U/L)	22.26 (18.55)	22.5	18.5	1	4.8
TG (mg/dL)	133.78 (103.45)	100	100	1	4.9
HbA1c (%)	8.56 (16.03)	4.35	4.35	1	36.1
PLS	17.73 (2.32)	20	20	3	5.8
UPH	5.76 (0.83)	5.2	4.05	1	14.4
USG	1.02 (0.01)	1.019	1.019	3	14.4
HCT (%)	41.76 (4.37)	45	39	3	5.0
RBC (10^6/uL)	4.57 (0.47)	4.85	4.2	3	5.0
WBC (10^3/uL)	6.2 (6.53)	7	7	3	12.6
HDL (mg/dL)	49.25 (11.98)	55	55	2	4.8
LDL (mg/dL)	114.03 (31.81)	100	100	1	6.7
BUN (mg/dL)	14.43 (4.49)	13	13	3	4.7
CREA (mg/dL)	0.87 (0.26)	1	1	1	4.7

Notes: CA = chronological age; BMI = body mass index; WC = waistline; SBP = systolic blood pressure; DBP = diastolic blood pressure; GLU0 = fasting glucose; TC = total cholesterol; HMG = hemoglobin level; SGOT = serum glutamate oxaloacetate transaminase; SGTP = serum glutamic pyruvic transaminase; TG = triglycerides; HbA1c = hemoglobin A1c; PLS = pulse; UPH = urine acid; USG = urine specific gravity; HCT = hematocrit; RBC = red blood cells; WBC = white blood cells; HDL = high-density lipoprotein cholesterol; LDL = low-density lipoprotein cholesterol; BUN = blood urea nitrogen; CREA = creatinine

Table A.2
Baseline characteristics and clinical criteria of NHIC–HTN and NHIC-T2DM.

Variable	Mean (SD)		Criteria for normality		Type
	NHIC–HTN (N = 81,775)	NHIC-T2DM (N = 84,130)	Male	Female	
CA	41.46 (10.95)	45.15 (12.72)	-	-	-
Disease	10,814	15,665	-	-	-
BMI (kg/m^2)	23.81 (2.91)	24.06 (2.96)	21.75	21.75	3
WC (cm)	82.68 (7.53)	83.61 (7.56)	-	-	1
SBP (mmHg)	119.98 (10.35)	124.39 (14)	105	105	1
DBP (mmHg)	75.20 (7.34)	78.13 (9.67)	70	70	1
GLU0 (mg/dL)	95.22 (22.2)	92.78 (11.85)	90	90	1
TC (mg/dL)	191.57 (35.12)	193.24 (35)	150	150	1
HMG (g/dL)	15 (1.12)	14.96 (1.15)	14.75	13	3
SGOT (U/L)	25.92 (12.11)	26.84 (16.14)	-	-	1
SGPT (U/L)	28.58 (19.21)	29.34 (22.52)	22.5	18.5	1
r-GTP (U/L)	41.47 (35.41)	46.4 (52.35)	40.5	24	1

Notes: CA = chronological age; BMI = body mass index; WC = waistline; SBP = systolic blood pressure; DBP = diastolic blood pressure; GLU0 = fasting glucose; TC = total cholesterol; HMG = hemoglobin level; SGOT = serum glutamate oxaloacetate transaminase; SGTP = serum glutamic pyruvic transaminase; r-GTP = gamma glutamyl transpeptidase.

Appendix B. Missing value imputation

Missing values in the KNHANES dataset were imputed by comparing four representative data imputation methods, Harrell miscellaneous (Hmisc), missForest, RF-SRC, and multivariate imputation by chained equations (MICE). To select the most accurate imputation method, we artificially generated missing values and compared the estimated value with the corresponding real values. A normalized root mean squared error (NRMSE) was used as a metric to assess the accuracy of imputation. The computational speeds of the models were also compared, to assess their abilities to process large datasets. Of the four imputation methods for KNHANES data, RF-SRC showed the highest accuracy (e.g., the lowest NRMSE) and the smallest computation time (Table B.1); i.e., RF-SRC is an accurate and time-efficient method to impute missing values.

Appendix C. Hyperparameters

We used a grid search to select the optimal value for hyperparameters which minimizes our loss in the training set. For the RSAE-BA model, the search space of the number of hidden layers was {2, 3, 4} for three datasets. In case of the number of hidden layer units, it was explored at the range of [3,20] for the KNHANES data and [3,9] for the NHIC datasets, respectively. In addition, experiments with several learning rates of {0.05, 0.01, 0.001}, convergence criterion of 10^{-5}, and the number of epochs at every 50 interval with the range of [200, 500] were conducted. The determined optimal hyperparameters for three datasets are listed in Table C.1.

Appendix D. BA statistics

Table B.1
Comparison for missing value imputation methods in terms of NRMSE and computational costs.

	Hmisc	missForest	**RF-SRC**	MICE
NRMSE	0.855	0.303	**0.188**	0.399
Time cost (N = 1000)	3 min	30 min	**1 s**	1 min
Time cost (N = 10000)	12 min	1 h	**3.3 s**	7 min

Table C.1
Hyperparameters used in experiments for each dataset.

Hyperparameter	Value		
	KNHANES	NHIC–HTN	NHIC-T2DM
Hidden layers	4	2	2
Hidden layer units	18–14–10–6	7–4	7–4
Learning rate	0.01	0.01	0.01
Convergence criterion	10^{-5}	10^{-5}	10^{-5}
Epoch	300	350	350

Table D.1
Average and SD of difference between BA and CA by age group based on CA when $q = 5$.

	Age group with CA			
	30s	40s	50s	60s
|BA-CA|	5.95 (3.95)	5.54 (3.98)	4.85 (3.73)	4.33 (3.41)

References

[1] S. Ashiqur Rahman, P. Giacobbi, L. Pyles, C. Mullett, G. Doretto, D.A. Adjeroh, Deep learning for biological age estimation, Briefings in Bioinformatics..

[2] C.-Y. Bae, Y.G. Kang, Y.-S. Suh, J.H. Han, S.-S. Kim, K.W. Shim, A model for estimating body shape biological age based on clinical parameters associated with body composition, Clinical interventions in aging 8 (2013) 11.

[3] X. Bai, L. Han, Q. Liu, H. Shan, H. Lin, X. Sun, X.-M. Chen, Evaluation of biological aging process–a population-based study of healthy people in China, Gerontology 56 (2) (2010) 129–140.

[4] D.H. Ballard, Modular Learning in Neural Networks., in: AAAI, 279–284, 1987..

[5] R. Bhadani, AutoEncoder for Interpolation, arXiv preprint arXiv:2101.00853..

[6] P. Billingsley, Probability and measure, John Wiley & Sons, 2008.

[7] Bioage, Bioage, URL:http://www.bio-age.co.kr/, 2002..

[8] G.A. Borkan, A.H. Norris, Assessment of biological age using a profile of physical parameters, Journal of Gerontology 35 (2) (1980) 177–184.

[9] C.J. Bulpitt, R.L. Antikainen, H.L. Markowe, M.J. Shipley, Mortality according to a prior assessment of biological age, Current aging science 2 (3) (2009) 193–199.

[10] I.H. Cho, K.S. Park, C.J. Lim, An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI), Mechanisms of ageing and development 131 (2) (2010) 69–78.

[11] J.H. Cole, R.P. Poudel, D. Tsagkrasoulis, M.W. Caan, C. Steves, T.D. Spector, G. Montana, Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker, NeuroImage 163 (2017) 115–124.

[12] A. Comfort, Test-battery to measure ageing-rate in man, The Lancet 294 (7635) (1969) 1411–1415.

[13] L. Deng, Three classes of deep learning architectures and their applications: a tutorial survey, APSIPA transactions on signal and information processing 57 (2012) 58.

[14] T. Dubina, A.Y. Mints, E. Zhuk, Biological age and its estimation. III. Introduction of a correction to the multiple regression model of biological age and assessment of biological age in cross-sectional and longitudinal studies, Experimental gerontology 19 (2) (1984) 133–143.

[15] J.M. Engels, P. Diehr, Imputation of missing longitudinal data: a comparison of methods, Journal of clinical epidemiology 56 (10) (2003) 968–976.

[16] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, et al, Genome-wide methylation profiles reveal quantitative views of human aging rates, Molecular cell 49 (2) (2013) 359–367.

[17] J.W. Hollingsworth, A. Hashizume, S. Jablon, Correlations between tests of aging in Hiroshima subjects–an attempt to define physiologic age., The Yale journal of biology and medicine 38 (1) (1965) 11.

[18] A.C. Holly, D. Melzer, L.C. Pilling, W. Henley, D.G. Hernandez, A.B. Singleton, S. Bandinelli, J.M. Guralnik, L. Ferrucci, L.W. Harries, Towards a gene expression biomarker set for human biological age, Aging cell 12 (2) (2013) 324–326.

[19] S. Horvath, DNA methylation age of human tissues and cell types, Genome biology 14 (10) (2013) 3156.

[20] D.K. Ingram, Key questions in developing biomarkers of aging, Experimental gerontology 23 (4–5) (1988) 429–434.

[21] D.K. Ingram, E. Nakamura, D. Smucny, G.S. Roth, M.A. Lane, Strategy for identifying biomarkers of aging in long-lived species, Experimental Gerontology 36 (7) (2001) 1025–1034.

[22] InnerAge, InnerAge by Insidetracker, URL:https://www.insidetracker.com/innerage/, 2009..

[23] H. Ishwaran, U.B. Kogalur, E.H. Blackstone, M.S. Lauer, et al, Random survival forests, The annals of applied statistics 2 (3) (2008) 841–860.

[24] H. Jee, B.H. Jeon, Y.H. Kim, H.-K. Kim, J. Choe, J. Park, Y. Jin, Development and application of biological age prediction models with physical fitness and physiological components in Korean adults, Gerontology 58 (4) (2012) 344–353.

[25] H. Jee, J. Park, Selection of an optimal set of biomarkers and comparative analyses of biological age estimation models in Korean females, Archives of gerontology and geriatrics 70 (2017) 84–91.

[26] L. Jia, W. Zhang, X. Chen, Common methods of biological age estimation, Clinical interventions in aging 12 (2017) 759.

[27] I. Jolliffe, Principal component analysis, Springer, 2011.

[28] Y.G. Kang, E. Suh, H. Chun, S.-H. Kim, D.K. Kim, C.-Y. Bae, Models for estimating the metabolic syndrome biological age as the new index for evaluation and management of metabolic syndrome, Clinical interventions in aging 12 (2017) 253.

[29] Y.G. Kang, E. Suh, J.-W. Lee, D.W. Kim, K.H. Cho, C.-Y. Bae, Biological age as a health index for mortality and major age-related disease incidence in Koreans: National health Insurance service–health screening 11-year follow-up study, Clinical interventions in aging 13 (2018) 429.

[30] P. Klemera, S. Doubal, A new approach to the concept and computation of biological age, Mechanisms of ageing and development 127 (3) (2006) 240–248.

[31] T. Knaus, R. Nuscheler, Incomplete risk adjustment and adverse selection in the German public health insurance system..

[32] J. Krištić, F. Vučković, C. Menni, L. Klarić, T. Keser, I. Beceheli, M. Pučić-Baković, M. Novokmet, M. Mangino, K. Thaqi, et al, Glycans are a novel biomarker of chronological and biological ages, Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences 69 (7) (2013) 779–789.

[33] J. Krøll, O. Saxtrup, On the use of regression analysis for the estimation of human biological age, Biogerontology 1 (4) (2000) 363–368.

[34] S. Kullback, Information theory and statistics, Courier Corporation (1997).

[35] S. Kweon, Y. Kim, M.-J. Jang, Y. Kim, K. Kim, S. Choi, C. Chun, Y.-H. Khang, K. Oh, Data resource profile: the Korea national health and nutrition examination survey (KNHANES), International journal of epidemiology 43 (1) (2014) 69–77.

[36] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[37] J. Lee, W. Lee, I.-S. Park, H.-S. Kim, H. Lee, C.-H. Jun, Risk assessment for hypertension and hypertension complications incidences using a Bayesian network, IIE Transactions on Healthcare Systems Engineering 6 (4) (2016) 246–259.

[38] M.E. Levine, Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age?, Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences 68 (6) (2012) 667–674

[39] M.E. Levine, E.M. Crimmins, A comparison of methods for assessing mortality risk, American Journal of Human Biology 26 (6) (2014) 768–776.

[40] S. MacDonald, R. Dixon, A. Cohen, J. Hazlitt, Biologi-cal age and 12-year cognitive change in older adults, Gerontology 50 (2004) 64–81.

[41] E. Nakamura, K. Miyao, Further evaluation of the basic nature of the human biological aging process based on a factor analysis of age-related physiological variables, The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 58 (3) (2003) B196–B204.

[42] E. Nakamura, K. Miyao, T. Ozeki, Assessment of biological age by principal component analysis, Mechanisms of ageing and development 46 (1–3) (1988) 1–18.

[43] K. Pawar, V.Z. Attar, Assessment of autoencoder architectures for data representation, in: Deep Learning: Concepts and Architectures, Springer, 101–132, 2020..

[44] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M.P. Reyes, M.-L. Shyu, S.-C. Chen, S. Iyengar, A survey on deep learning: Algorithms, techniques, and applications, ACM Computing Surveys (CSUR) 51 (5) (2018) 1–36.

[45] T.V. Pyrkov, K. Slipensky, M. Barg, A. Kondrashin, B. Zhurov, A. Zenin, M. Pyatnitskiy, L. Menshikov, S. Markov, P.O. Fedichev, Extracting biological age from biomedical data via deep learning: too much of a good thing?, Scientific reports 8 (1) (2018) 1–11

[46] S.A. Rahman, D.A. Adjeroh, Deep learning using convolutional LSTM estimates biological age from physical activity, Scientific reports 9 (1) (2019) 1–15.

[47] M. Uttley, M.H. Crawford, Efficacy of a composite biological age score to predict ten-year survival among Kansas and Nebraska Mennonites..

[48] J. Yoo, Y. Kim, E.R. Cho, S.H. Jee, Biological age as a useful index to predict seventeen-year survival and mortality in Koreans, BMC geriatrics 17 (1) (2017) 7.

[49] F.-N. Yuan, L. Zhang, J. Shi, X. Xia, G. Li, Theories and applications of auto-encoder neural networks: A literature survey, Chinese Journal of Computers 42 (01) (2019) 203–230.

[50] X. Zhong, Y. Lu, Q. Gao, M.S.Z. Nyunt, T. Fulop, C.P. Monterola, J.C. Tong, A. Larbi, T.P. Ng, Estimating biological age in the Singapore longitudinal aging study, The Journals of Gerontology: Series A..