

SCIENTIFIC REPORTS



OPEN

Extracting biological age from biomedical data via deep learning: too much of a good thing?

Timothy V. Pyrkov¹, Konstantin Slipensky², Mikhail Barg², Alexey Kondrashin², Boris Zhurov¹, Alexander Zenin¹, Mikhail Pyatnitskiy¹, Leonid Menshikov¹, Sergei Markov² & Peter O. Fedichev^{1,3}

Age-related physiological changes in humans are linearly associated with age. Naturally, linear combinations of physiological measures trained to estimate chronological age have recently emerged as a practical way to quantify aging in the form of biological age. In this work, we used one-week long physical activity records from a 2003–2006 National Health and Nutrition Examination Survey (NHANES) to compare three increasingly accurate biological age models: the unsupervised Principal Components Analysis (PCA) score, a multivariate linear regression, and a state-of-the-art deep convolutional neural network (CNN). We found that the supervised approaches produce better chronological age estimations at the expense of a loss of the association between the aging acceleration and all-cause mortality. Consequently, we turned to the NHANES death register directly and introduced a novel way to train parametric proportional hazards models suitable for out-of-the-box implementation with any modern machine learning software. As a demonstration, we produced a separate deep CNN for mortality risks prediction that outperformed any of the biological age or a simple linear proportional hazards model. Altogether, our findings demonstrate the emerging potential of combined wearable sensors and deep learning technologies for applications involving continuous health risk monitoring and real-time feedback to patients and care providers.

Many physiological parameters demonstrate profound correlations with age. This observation has led to the growing popularity of “biological clocks”, designed as linear predictors of chronological age from, e.g., DNA methylation¹, gene expression², plasma proteome³ profiles, etc. A broader acceptance of the biomarkers of age, however, will depend on a better understanding of the observed correlations to the incidence of specific diseases, improved transferability of the models across populations and reduction of costs of the studies. We note, that large-scale biochemical or genomic profiling is not impossible, but is still logistically difficult and expensive. Instead, the recent introduction of low-power and compact sensors, based on micro-electromechanical systems (MEMS) has led to a new breed of the wearable and affordable devices providing unparalleled opportunities for the collecting and cloud-storing personal digitized activity records in a fully standardized and controlled way. This tracking is already done without interfering with the daily routines of hundreds of millions of people all over the world. Moreover, the analysis of human locomotion has already led to widely accepted health recommendations, such as the “10,000 steps per day” minimum activity advisory⁴, which is now a basic health recommendation in many wellness applications. We have recently shown that a simple set of hand-engineered features, representing statistical properties of one-week long human physical activity time series can be used to produce digital biomarkers of aging and frailty⁵.

Deep learning is a powerful tool in pattern recognition and has demonstrated outstanding performance in visual object identification, speech recognition, and other fields requiring hierarchical analysis of input data. Recent promising examples in the field of biomedical signal analysis include convolutional neural networks (CNNs) trained to process electrocardiograms showing cardiologist-level performance in detection of arrhythmia⁶, biomarkers of age from clinical blood biochemistry^{7,8} or electronic medical records⁹, and mortality prediction¹⁰. Inspired by these examples, we explored deep learning architectures for Health Risks Assessment (HRA)

¹Gero LLC, Novokuznetskaya street 24/2, Moscow, 119017, Russia. ²ActiveBusinessCollection LLC (Sberbank group), Moscow, 117312, Russia. ³Moscow Institute of Physics and Technology, 141700, Institutskii per. 9, Dolgoprudny, Moscow Region, Russia. Correspondence and requests for materials should be addressed to T.V.P. (email: tim.pyrkov@gero.com) or P.O.F. (email: peter.fedichev@gero.com)

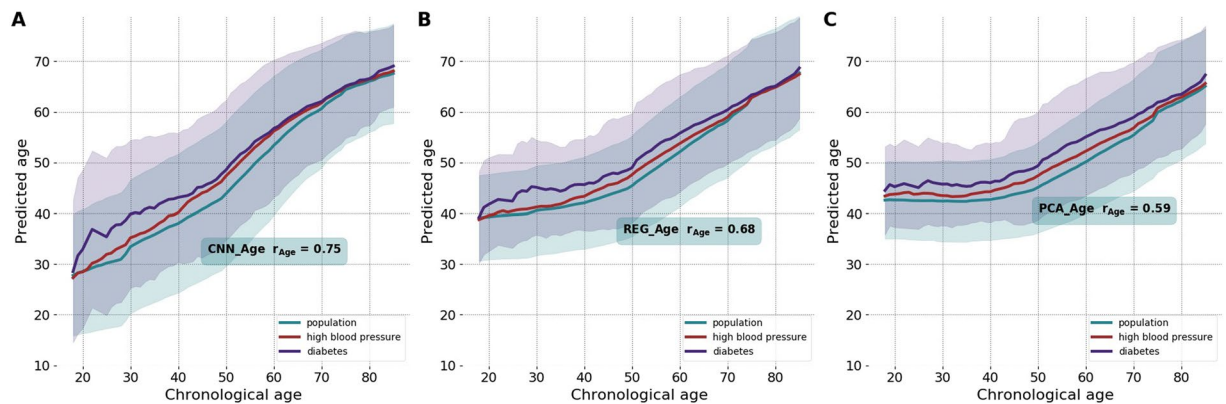


Figure 1. Accuracy of age-predicting models. The biological age estimation according to deep convolutional neural network CNN_Age (A), the multivariate regression REG_Age (B), and the unsupervised PCA_Age (C) models. The solid lines and the transparent \pm standard deviation bands are color coded as green, blue, and red, representing the whole population, the patients diagnosed with diabetes by a doctor, and individuals with self-reported high blood pressure, respectively. All the calculations were produced using the NHANES 2003–2006 cohort wearable accelerometers data, comprising one-week long activity tracks (1 min^{-1} sampling rate).

applications, involving human physical activity streams from wearable devices. First, we trained a series of deep CNNs to predict the chronological age of NHANES participants and obtained a substantial improvement over a multivariate linear regression fit from the same signal. By design, every such supervised biological age estimation aims at minimizing the difference between the predicted and the actual chronological age of the same individual. This quantity is referred to as aging acceleration and is associated with prevalence of disease^{11–13} and mortality^{14–16}. Therefore, we found, as expected, that any improvement yielding a decrease in the age prediction error rate often translates into a reduced ability to differentiate among disease states or mortality risks.

Thus we conclude that the chronological age of an individual may not be the best target for biological age model development. More sensible ways to quantify aging should involve the inclusion of metrics characterizing life expectancy, disease burden or frailty. In this study, we focused on mortality as the ultimate health status variable and proposed a novel technique to train a proportional hazards model in a form, suitable for an out-of-the-box implementation with modern deep-learning toolkits. We used the death register and raw one-week long time series representing physical activity records of NHANES participants and obtained the best results with a deep CNN architecture, simultaneously and automatically producing the set of features engineered from the raw physical activity time series and the most accurate non-linear representation of the risks function. Altogether, our findings demonstrate the emerging potential of a combination of wearable sensors and deep learning technologies for future HRA applications involving continuous health risk evaluation and real-time feedback to patients and care providers.

Results

Deep learning chronological age from physical activity records. The “biological age” is a quantitative measure of aging - and thus an expected lifespan - based on biological data. State-of-the-art approaches for biological age evaluation take advantage of the strong association of physiological variables with age and thus rely on linear (see, e.g., DNA methylation age^{1,17}) or non-linear regressions^{7,8,10} to estimate the chronological age of an individual directly from the biological data. Following these examples, we started by building a deep convolutional neural network (CNN) trained to predict the chronological age of the same NHANES participants from the raw one-week long physical activity records. We used CNN to automatically extract relevant features from time series (see, e.g.⁶) and to unravel the apparent non-linear dependencies between the locomotor activity variables and age. We trained the CNN_Age model with four convolution and two dense layers (see Materials and Methods for the details of the CNN architecture). The model achieved Pearson’s $r = 0.75$ between the estimated and the chronological age of the study participants, see Fig. 1A).

To highlight the superior performance of the CNN_Age, we also characterized a regularized multivariate regression, trained to predict chronological age from a linear combination of hand-crafted features, representing statistical properties of the physical activities time series, borrowed from⁵. The result was the REG_Age model, which is similar by design to the most commonly used biological age metrics, such as, e.g., DNA methylation clock, a regularized linear regression of DNA methylation features to the chronological age^{1,17}. REG_Age worked fairly well, but was inferior in accuracy to CNN_Age (Pearson’s $r = 0.68$, see Fig. 1B).

In our recent study⁵ we observed that Principal Component (PC) analysis reveals that most of the variance in the multi-dimensional parameter space spanned by all NHANES participant representations is associated with chronological age. Therefore, we proposed the first principal component score as the unsupervised definition of biological age, PCA_Age (Pearson’s correlation coefficient of the first PC score and chronological age $r = 0.59$, see Fig. 1C). PCA_Age does not change much at first, but increases approximately linearly with age thereafter, roughly, the age of 40, and exhibits an excellent correlation with the negative logarithm of average daily activity⁵.

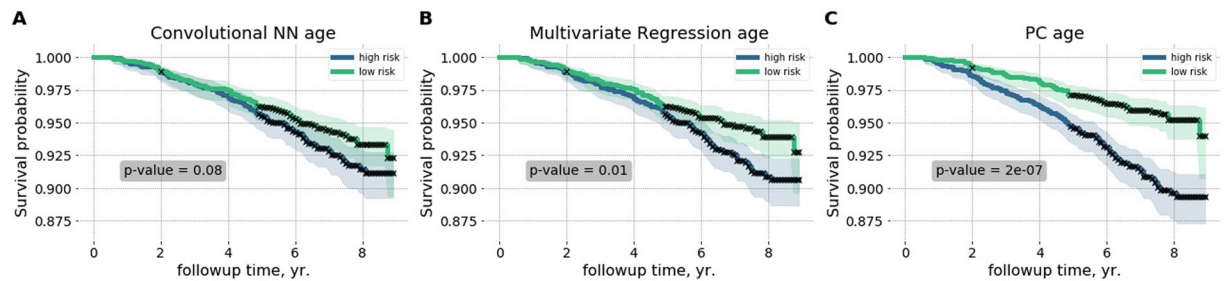


Figure 2. Performance of the biological age models to distinguish between the longer- and the shorter-living individuals illustrated with Kaplan-Meier survival curves. Each participant was classified into the “high-” and the “low-” risks groups according to the deep convolutional neural network CNN_Age (A), the multivariate regression REG_Age (B), and the unsupervised PCA_Age (C) models. The p-values characterize the survival curves separation significance.

Covariate	Hazard Ratio, 95% CI, p-value
age	1.090, CI[1.082, 1.099] ($p = 2e-109$)
gender	1.503, CI[1.281, 1.765] (2.4 yr., $p = 6e-07$)
diabetes	1.690, CI[1.401, 2.038] (2.2 yr., $p = 4e-08$)
smoking	2.300, CI[1.866, 2.836] (3.8 yr., $p = 6e-15$)
high blood pressure	1.198, CI[1.018, 1.409] (1.0 yr., $p = 3e-02$)
log(activity)	2.378, CI[2.051, 2.756] (4.9 yr., $p = 1e-30$)

Table 1. Cox-Gompertz parametric proportional hazards model parameters.

CNN_Age outperformed any other model in terms of the chronological age prediction accuracy. REG_Age and PCA_Age used the same set of physical activity derived features. Therefore, the CNN performed well and extracted the relevant set of age-associated properties of the raw time series in a fully automated way.

Improved chronological age estimation accuracy undermines the biological age association with diseases and lifespan. Patients diagnosed with diabetes or hypertension appear to be biologically older than their healthy peers, see Fig. 1A–C. To find out if the biological age difference translates into a lifespan change, we used NHANES death register representing survival data for 7837 participants (3750 male, 4087 female, aged 18–85, follow-up time up to 9 years, 701 participants died, in total). Figure 2A–C show a summary of Kaplan-Meier survival curves for NHANES participants, stratified into the high- and the low- risk groups according to the difference between the estimated biological age of an individual and the averaged estimated age of gender- and age-matched peers. The procedure is similar in spirit but formally different from and produces better results than the group separation according to the sign of aging acceleration. Using the predicted biological age as the reference naturally offsets the apparently non-linear relation between the biological and chronological age.

The unsupervised PCA_Age model fared exceptionally well and produced significantly different survival curves ($p = 2 \times 10^{-7}$, see Fig. 2C). The increasing accuracy of chronological age estimation by each of the two supervised models, however, came at a price of a significant drop in the ability to distinguish the longer- and the shorter-living individuals. The multivariate regression REG_Age and the deep CNN_Age models failed to rank the individuals according to mortality risks ($p = 0.08$ and $p = 0.01$, respectively, see Fig. 2A,B).

The apparent progressive loss of biologically relevant information by increasingly accurate models, involving a regression to chronological age, highlights the biological significance of aging acceleration, a quantity closely related to the chronological age determination error from the physiological data.

Biological age is a non-trivial factor in health risk assessment. Health Risk Assessment (HRA) is a systematic approach to collecting information from individuals that identifies risk factors, provides individualized feedback, and links the person with health-promoting interventions, see, e.g.¹⁸. Biological age acceleration is associated with major diseases, and hence we asked ourselves if any of the biological age models could provide any useful additional information and improve health risks evaluation accuracy over standard HRA questionnaires.

HRA approaches involve capturing demographic (such as age and gender), lifestyle (including exercise, smoking, alcohol intake, diet), and physiological (e.g., weight, height, blood pressure, cholesterol) characteristics¹⁸. To model a HRA we turned to all-cause mortality, as an ultimate variable characterizing an organism with respect to overall health and aging. We built a Cox proportional hazards model including the most important demographic, lifestyle and physiological factors, see Table 1 for the summary of the model parameters. We then tested the statistical significance of the biological age estimators in association with all-cause mortality using a Cox proportional hazards model with the corresponding aging acceleration and the health risk factors taken as covariates, see the top of Table 2 for the summary of the results. The biological age scores were corrected by the health risks factors in advance to avoid possible fitting instabilities due to potential collinearity between the covariates. In this way, we

HRA Covariates	Age, Gender, Smoking, Diabetes, Hypertension	Age, Gender, Smoking, Diabetes, Hypertension, Activity
Model	Hazard Ratio, 95% CI, p-value	Hazard Ratio, 95% CI, p-value
PCA Age	1.45, CI[1.35, 1.55] (4.4 yr., $p < 10^{-10}$)	1.07, CI[0.99, 1.15] (0.8 yr., $p = 0.08$)
REG Age	1.31, CI[1.21, 1.42] (3.2 yr., $p < 10^{-10}$)	1.01, CI[0.93, 1.08] (0.0 yr., $p = 1.0$)
CNN Age	1.05, CI[0.96, 1.15] (0.6 yr., $p = 0.3$)	0.94, CI[0.86, 1.03] (-0.7 yr., $p = 0.2$) (wrong sign)
Hazard Weighted regr	1.27, CI[1.21, 1.32] (2.8 yr., $p < 10^{-10}$)	1.08, CI[1.01, 1.16] (0.9 yr., $p = 0.03$)
Hazard CNN	1.53, CI[1.42, 1.65] (5.0 yr., $p < 10^{-10}$)	1.15, CI[1.07, 1.24] (1.6 yr., $p = 3 \times 10^{-4}$)

Table 2. Association of the biological age- and all-cause mortality -predicting models with all-cause mortality. Two class of models were evaluated: one including the HRA parameters: age, gender, diabetes, smoking and hypertension as covariates (the left column) and the other including additionally the negative logarithm of the average daily physical activity (the biological age proxy, the right column). The corresponding 95% hazard ratio intervals are given along with the significance p-value and the effect levels, expressed in years of life lost (see the text for details).

explicitly tested the biological age residual for the ability to explain the risk function variance that is not already accounted for by the standard HRA factors.

As in every statistical test so far, the unsupervised PCA_Age model prediction produced the largest proportional hazards effect, $HR = 1.45$ ($p < 10^{-10}$). The supervised REG_Age and CNN_Age models achieved better age predicting accuracy at the expense of biological significance loss, see the upper-left column in Table 2. The multivariate regression REG_Age provided a smaller yet significant effect, $HR = 1.31$ ($p < 10^{-10}$). The most accurate chronological age estimator CNN_Age, however, failed to produce any statistically significant contribution ($p = 0.3$). Therefore, at least in our model scenario, two of the biological age estimators, PCA_Age and REG_Age, produced biological age measures that substantially improve patient health risk estimation beyond any standard procedure involving HRA metrics.

On the other hand, the negative logarithm of the average daily physical activity is a good biological age proxy⁵ and provides a highly significant contribution to the hazard function ($p < 10^{-10}$), see Table 1. We checked explicitly and found that none of the biological age models estimations produced any proportional hazards effect, once the physical activity level is included as an independent HRA variable, see the upper-right column of Table 2, $p > 0.05$.

Deep learning mortality model improves health risks assessment performance. Next, we explored whether we could build a useful health risks prediction model involving signatures of aging or diseases beyond biological age. Proportional hazards models (PHM) in Cox-¹⁹ or Cox-Gompertz-²⁰ variants are trusted tools for such an analysis and are readily available as standard software, including a recent deep learning architecture implementation²¹. We employed CNN to automate extraction of the most relevant features from the raw time series. To train a CNN to predict mortality, we would need a PHM likelihood as the network cost function in a form, suitable for an efficient minimization with back-propagation. Naturally, we expected that the contribution to the all-cause mortality (or hazard ratio) of features, associated with the locomotor activity to be small on top of the already significant effects of age, gender, and the other HRA factors. Therefore, instead of implementing a full non-linear likelihood minimization, we performed a linear perturbation theory expansion of the Cox-Gompertz likelihood. The result was a simplified linear model, closely related to regression to Martingale residuals, or the unexplained risks variance of the model involving the standard HRA factors. More importantly, the cost function was formally equivalent to a linear regression with sample weights and therefore can be implemented out-of-the-box, with essentially any modern deep learning software (see Materials and Methods).

We used the proposed linearized version of the Cox-Gompertz proportional hazards model to train a deep CNN. The network received raw physical activity streams and performed considerably better than our reference, a linear hazard predictor built using the same perturbation theory expansion but with the help of hand-crafted features from⁵ and already used here as descriptors by REG_Age and REG_Age models. The deep CNN hazard rate model outperformed all the biological age scores above, see the bottom of Table 2. Both the linear and the CNN hazard rate predictors produced a significant association with the all-cause mortality after a correction by age, gender, diabetes, smoking and high blood pressure ($p < 10^{-10}$, see the bottom-left column of Table 2). The correction by the negative logarithm of the average daily physical activity made the hazard prediction of the linear hazard model statistically irrelevant. The CNN hazard rate model predictor remained significant, $HR = 1.15$, ($p = 0.0003$). Assuming proportional hazards model this contribution accounts for, approximately, 1.6 years of life gained or lost at the standard deviation level. The most significant association of the CNN hazard rate model residual after detrending by the physical activity level and the major HRA factors with the NHANES Questionnaire and Laboratory data variables was the self-reported “general health condition” (labels “excellent/very good” vs “fair/poor”).

Discussion

We report a systematic investigation of biological aging acceleration in relation to disease states and all-cause mortality in a large-scale human study. In particular, we used the NHANES physical activity records along with the medical meta-data and the death register to produce a series of biological age models. The most popular supervised learning examples, such as a multivariate regression and a deep CNN, exploit the apparent linear

dependence of physiological changes with age and hence were trained to produce a chronological age estimate from individual locomotor activity time series. Both cases involved a minimization of the biological age acceleration, the difference between the “physiological” age, estimated by the model and the chronological age of an individual. Alternatively, we used an unsupervised technique taking advantage of the intrinsic low-dimensionality of the aging trajectories, closely related to criticality^{5,22} of the biological state variables kinetics, and thus presenting a natural biomarker of age from Principal Component Analysis (see, e.g.²³ for a more detailed discussion of PCA in the context of biological age determination).

To date, different forms of regularized multivariate linear regressions of biologically relevant variables to chronological age are the most popular technique behind the recently proposed biological age signatures, such as, e.g., biological clocks using IgG glycosylation²⁴, blood biochemical parameters²⁵, gut microbiota composition²⁶, and cerebrospinal fluid proteome²⁷. The “epigenetic clock” based on DNA methylation (DNAm) levels^{1,17} appears to be the most accurate and the most extensively studied measure of aging. The biological age acceleration, measured by the DNAm clock, explains all-cause mortality in later life better than chronological age¹⁴, is elevated in people with HIV, Down syndrome^{11,12}, obesity¹³, but is not correlated with smoking²⁸. The same biomarker of age is lower for super-centennial’s offspring¹⁵, and predicts mortality in a longitudinal twins study¹⁶.

By design, such a supervised methodology involves a form of minimization of the biological age determination error, i.e., an attempt of minimizing the difference between the predicted “physiological” age and the chronological age of the individual. This is, by definition, the biological age acceleration, a presumably biologically relevant variable. Therefore, we expected and provided an evidence for here that a systematic improvement of the chronological age determination error minimization leads to an immediate degradation of the biological age acceleration significance in any test, involving health or risk of death. Our calculations show that the loss of biological information can be aggravated if even more powerful machine learning tools, such as deep learning architectures, are employed to unravel complex and possibly nonlinear relations between the features in the data and produce even more “accurate” models. The more accurate chronological age estimation from biological samples could nevertheless find applications in forensic studies, see, e.g.²⁹.

Diabetes is one of the most significant health risk factors affecting lifespan by shortening life expectancy up to 8 years according to some studies³⁰. However, the most popular DNAm clock did not label patients with diabetes diagnosed by a doctor as “biologically older” in at least one study³¹. In our studies, the patients diagnosed with diabetes and hypertension appear to be significantly older biological-age wise according to the unsupervised PCA_Age model⁵. The multivariate regression REG_Age is similar in spirit to DNAm age and produced a less significant separation between the groups of patients with the diseases and healthy controls. It remains to be seen, however, which of the interpretations would be supported with future versions of the DNAm clock, or, if our findings are specific to the source of the signal derived from human physical activity time series properties, known to be associated with both of the diseases.

Physical activity measurements recorded by wearable devices are, in theory, an ideal data source for building fully automated HRA systems for continuous health risk monitoring and real-time feedback to patients and care providers. The unsupervised PCA_Age and a linear multivariate regression REG_Age yielded valuable biological age models producing a biological aging acceleration estimate predictive of all-cause mortality, even after detrending by the HRA variables. The more accurate models, such as REG_Age and the deep CNN_Age predictors explained a lesser degree of the death risk variation, respectively. We also observed, as expected, that an explicit addition in a HRA routine of a variable in high correlation with the biological age, such as the negative logarithm of the average daily activity, makes any of biological age predictors statistically irrelevant.

The idea of identifying patterns in biological signals in association with phenotypic differences is not new³². However, deep CNNs bring the idea to an entirely new level and are particularly useful for automation of the most relevant features engineering from the data in relation to human activity recognition^{33,34}, specific diseases or risks factors¹⁰. To demonstrate CNN capabilities for all-cause mortality evaluation from physical activity records, we introduced a novel way to train a proportional hazards model effectively. We started by observing that the fundamental HRA factors such as age, gender, and major disease status allow for the production of an excellent survival model. Therefore, the expected contribution of any combination of the locomotor activity derived features, after detrending by the HRA variables, would be expected to be small. Under the circumstances, the full mortality can be obtained by iterations, with the zero-order approximation model being the Cox-Gompertz mortality model, embracing the HRA descriptors as independent covariates. The first subsequent perturbation theory correction is then equivalent to a regression with sample weights, depending on the zero-order model parameters (a generalization of the method for a more generic non-parametric mortality model is also possible and will be reported elsewhere). The statistical power of the method is limited by the total number of death events, note the explicit dependence in Eq. (3) in Materials and Methods. In that view, ubiquitous deployment of wearable sensors promises unparalleled opportunities to achieve large population-wide coverage and thus to make possible the identification of additional smaller health risk effects at significance levels in the future.

The unsupervised PCA_Age yields an important insight on the dynamics of the physiological state in association with age. Biological age turns out to be related to an order-parameter, associated with the organism development and as such, undergoes a random walk on top of the systematic aging drift^{5,22}. The variance of the biological age distribution increases with age, which is a sign of the increasing heterogeneity of the human population. The effect is a challenge to supervised methods, such as regressions to biological age and even proportional hazards models. We envision future improvements of mortality prediction models by taking into account the diffusion of the biological variables into the likelihood function directly. Given the success of the unsupervised biological age model in our study, we further expect a development of unsupervised deep learning architectures, such as deep auto-encoders³⁵ for aging research.

Any practical applications of HRA tools based on human physical activity analysis will depend on whether the developed models could perform well across different sources of signals. In our earlier work⁵, we adopted

the aggregated descriptors, the transition matrix elements, and a simple form of quantile normalization (see, e.g.³⁶) to demonstrate that the models trained using the NHANES data could be used to predict health risks in UK Biobank³⁷ cohort. CNNs are more sophisticated tools capable of capturing far more intricate dependencies of the inputs, sometimes at the expense of the model sensitivity to small variations in the data representations inevitable due to population differences⁸ as well as to differences in sensors hardware and study protocols. The CNN hazard model in this work is provided as a proof-of-concept example, is trained using NHANES locomotor activity records and, most probably, is not ready for immediate applications to any other human physical activity data source without a prior adaptation.

Life and health insurance programs have begun to provide discounts to their users based on physical activity monitored by fitness wristbands³⁸. We report that a deep CNN can be used to further refine the risks models by inclusion of an apparently biological age-independent risk factor, producing a significant effect on lifespan. We believe that the result highlights the power and practical utility of semi-analytic approaches, combining aging theory with the most powerful modern machine learning tools. This synthesis will eventually produce even better health risks models for HRA, to mitigate longevity risks in insurance, help in pension planning, and contribute to upcoming clinical trials and future deployment of anti-aging therapies.

Methods

Data preparation, quantification of locomotor activity. Locomotor activity records and questionnaire/laboratory data from the National Health and Nutrition Examination Survey (NHANES) 2003–2004 and 2005–2006 cohorts were downloaded from [www.cdc.gov/nchs/nhanes/index.htm]. NHANES provides locomotor activity in the form of 7-day long continuous tracks of “activity counts” sampled at 1 min^{-1} frequency and recorded by a physical activity monitor (ActiGraph AM-7164 single-axis piezoelectric accelerometer) worn on the hip. Of 14,631 study participants (7176 in the 2003–2004 cohort and 7455 in the 2005–2006 cohort), we filtered out samples with abnormally low (average activity count <50) or high (>5000) physical activity. We also excluded participants aged 85 and older since the https://www.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm#RIDAGEYRNHANES age data field is top coded at 85 years of age and we desired precise age information for our study. The mortality data for NHANES participants is obtained from the National Center for Health Statistics <https://www.cdc.gov/nchs/data-linkage/mortality-public.htmpublic> resources (4017 in the 2003–2004 cohort and 3985 in the 2005–2006 cohort). We excluded days with less than 200 minutes corresponding to activity states >0 . Only participants with 4 or more days that passed this additional filter were retained, yielding a total of 7454 samples (age, years: 35 ± 23 , range 6–84; women: 51%). Quantification of locomotor activity of the subject was carried out in two ways: by hand-engineered features for all linear analysis and by CNN (see corresponding sections below).

Statistical description time series representing physical activity. To calculate a statistical descriptor of each participant’s locomotor activity we followed the prescription from⁵. We first converted activity counts into discrete states with bin edges b_k , $k = 1 \dots K$. Activity level states $1 \dots K - 1$ were then defined as half-open intervals $b_k \leq a < b_{k+1}$, state 0 as $a < b_1$ and state K as $a \geq b_K$, where a is the activity count value. In this study, we defined $K = 8$ activity states with bin edges $b_k = e^k - 1$, $k = 1 \dots 7$. Thus, each sample was converted into a track of activity states and a transition matrix (TM) was then calculated for each participant (see below). To ensure that our analysis dealt only with days on which a participant actually performed some physical activity, we applied an additional filter. Transition matrices (TM) T_{ij} , $i = 1 \dots 8$, $j = 1 \dots 8$ were calculated as a set of transition rates from each state j to each other state i (the diagonal elements correspond to the probability of remaining in the same activity state). We flattened 8×8 TM of each sample into a 64-dimensional descriptor vector and converted the flattened descriptor to log-scale to ensure approximately normal distribution for elements of the locomotor descriptor (a useful property for the stability of the linear models that we applied in PCA and Survival analysis). All near-zero elements ($<10^{-3}$, which corresponds to less than 10 transitions during a week) were imputed by the value of 10^{-3} before log-scaling.

Age predicting models. We compare three age-estimating models. CNN_Age is a convolutional neural network (CNN) model trained to predict the chronological age of an individual based on input raw activity counts track, see Fig. 3 (more details of CNN architecture are described in the relevant section below). The learning curves for CNN_Age model are shown in Fig. 4A,B. Since Mean Squared Error (MSE) increased after epoch 500–800, we stopped training at epoch 500. The root mean squared error (RMSE) of the resulting biological age model was 14.0 years.

The remaining models, REG_Age and PCA_Age, estimate age in response to the vector of hand-crafted features representing statistical properties of the physical activity time series and borrowed from⁵. We trained the models using the activity records of 7186 NHANES participants aged 18–85. REG_Age is a L_2 -regularized multivariate regression trained with 5-fold cross-validation using RidgeCV package from scikit-learn python library. The regularization path exhibits a very shallow minimum corresponding to the RMSE of 14.5 years (the details of the calculation are available on request).

PCA_Age model is based on principal component (PC) analysis of the same set of features with the help of SVD factorization from numpy python library. We retained the first principal component score as the biological age metric and reported the PC score scaled to the chronological age by a univariate linear regression. Given the large dataset size of several thousand samples compared to the descriptor size of as little as 64 transition matrix components, we did not expect an overfit and hence did not use any cross-validation here. The RMSE of the biological age model was 16.1 years.

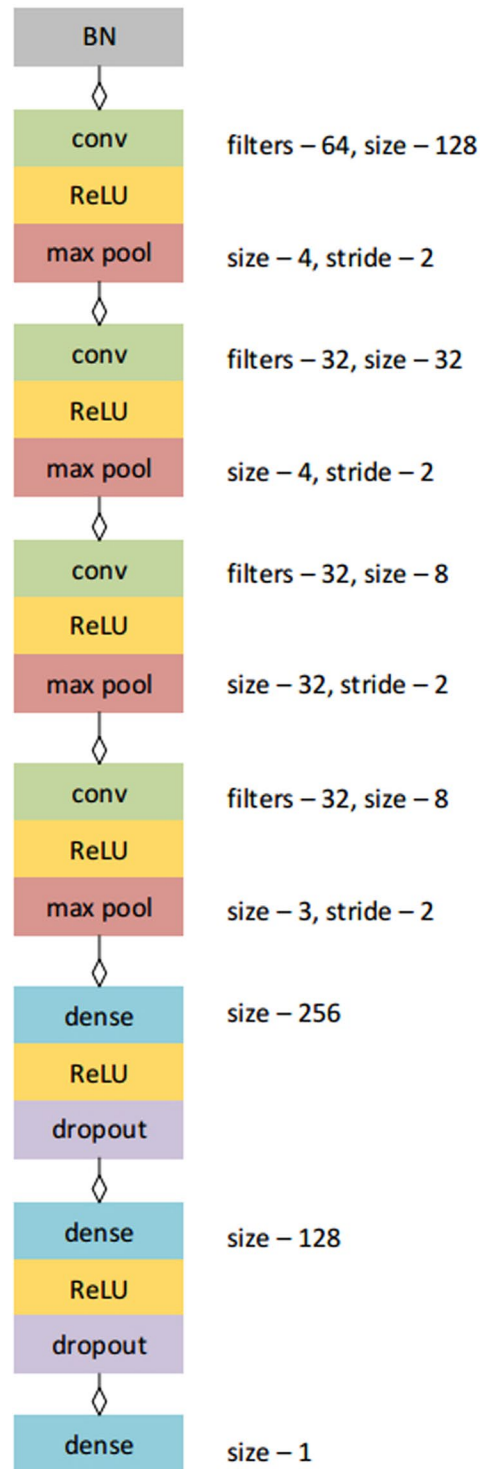


Figure 3. Architecture of convolutional neural network. The same network architecture was used to train both models for predicting age and for predicting mortality rates.

Cox-Gompertz proportional hazards model. According to the Gompertz law³⁹, the mortality rate in human populations increases exponentially starting at the age of about 40⁴⁰. Therefore an accurate estimation of a hazard rate, or mortality, for each participant can be obtained with the help of a parametric Cox-Gompertz proportional hazards model adapted from²⁰. The model predicts the mortality rate in the form $M(t^n, x^n) = M_0 \exp(\Gamma t^n) \exp(\beta, x^n)$, where t^n is the age of a participant n , x^n is a vector of independent predictor variables (covariates), such as the participant's gender and any set of physical activity related descriptors. The variables Γ and M_0 stand for the Gompertz exponent (inversely related to the mortality rate doubling time) and the initial mortality rate, respectively. If the covariates x^n do not explicitly depend on age, the unknown

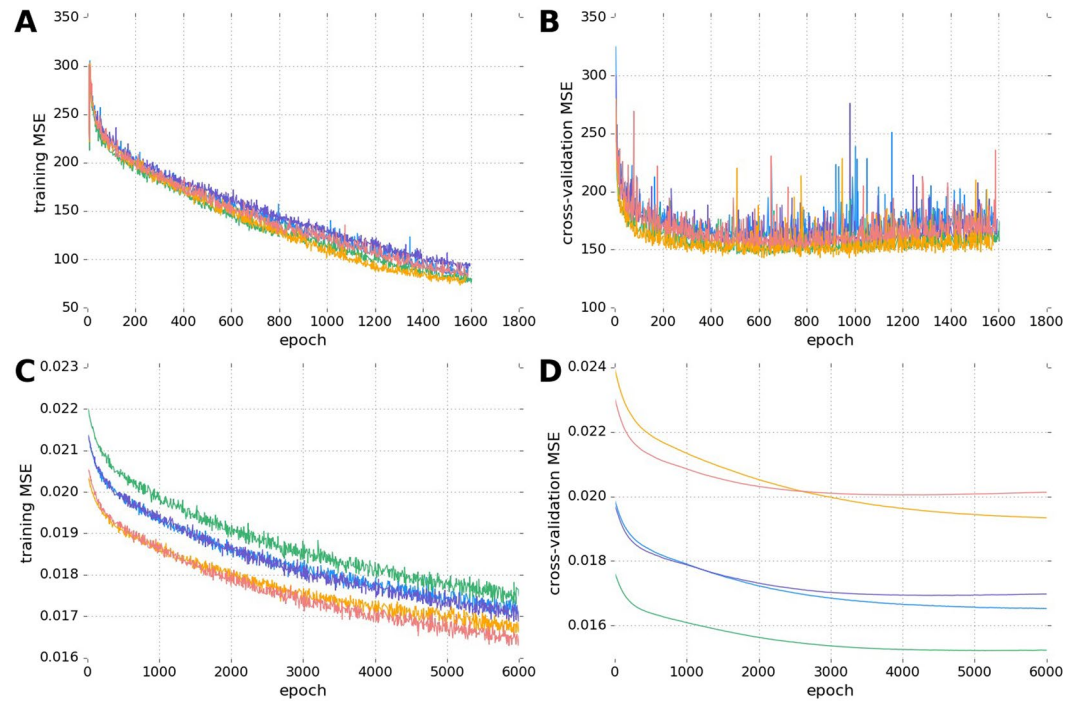


Figure 4. Learning curves for the CNN age and hazard predictor models. Learning curves for age-predicting model CNN_Age are shown for each of 5 folds for the training (A) and test (B) subsets. Cross-validation learning curves (B) show a minimum of MSE at epochs 500–800, while further training results in model overfitting. Panels (C,D) show the learning curves for the mortality risk-predicting CNN model. Here some of the cross-validation learning curves show a minimum of MSE at epoch 3000. Learning curves are color-coded according to the fold number in the 5-fold training/cross-validation setup.

parameters along with the vector of quantities β are to be fitted using the experimental data by minimizing the following negative log-likelihood function:

$$-\log LH = \sum_n \frac{M_0}{\Gamma} \exp(\beta x^n) f_n - \sum_n \delta^n (\log M_0 + (\beta x^n) + \Gamma t_2^n) \tag{1}$$

where the summation occurs over all the study participants, $f_n = \exp(\Gamma t_1^n) - \exp(\Gamma t_2^n)$, and t_1^n is the age of a participant when locomotor activity measurements were carried out. The second time variable, t_2^n , is the age at death, if the individual died during the follow-up time or the age of the last observation if the individual was alive, respectively. The indicator $\delta^n = 1$ if an individual n is dead at t_2^n and $\delta^n = 0$ otherwise. The log-likelihood can be further regularized by adding a proper term depending on a norm of the vector β (L_2 -regularization in our study). The optimization with respect to the scalar variables Γ , M_0 and the vector β can be performed using any convenient technique, including Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm or a batched stochastic gradient descent (SGD).

Perturbation theory expansion in a Cox-Gompertz model. We start assuming there's a well-defined minimum defined by a set of the best fit variables Γ , M_0 and $\bar{\beta}$. Let us now see how the model can be improved if we are allowed to add a set of additional variables ξ to improve the hazard rate model. Once the new parameters are plugged in the likelihood function (1), then, the position of the minimum of the likelihood function (1) also changes. If the corrections to the model parameters $\delta\Gamma$, δM_0 , $\delta\beta$ are sufficiently small, the solution of the minimization problem can be obtained by iterations. In typically small cohorts of human subjects, the simultaneous determination of the Gompertzian variables $\delta\Gamma$ and δM_0 is a poorly defined mathematical problem⁴¹. Therefore the variation of either one of the parameters can be arbitrary set to zero. We choose to fix Γ and without further derivation provide the final expression for the proportional hazards effect:

$$\beta_\xi = C_\rho^{-1} \sum_n \xi^n (\delta^n - N_d \rho^n). \tag{2}$$

Here $\rho^n = f_n \exp(\bar{\beta} x^n) / \sum_n f_n \exp(\bar{\beta} x^n)$, $N_d = \sum_n \delta^n$ is the total number of the death events in the dataset, and $\langle (\dots) \rangle_\rho \equiv \sum_n \rho^n (\dots)_n$ stands for averaging with the sample weights ρ^n . By definition the sample-weighted variance $C_\rho = N_d \langle \delta \xi^T \delta \xi \rangle_\rho = N_d (\langle \xi^T \xi \rangle_\rho - \langle \xi^T \rangle_\rho \langle \xi \rangle_\rho)$. The effect estimation error can also be obtained analytically,

$$\sigma_{\beta_{\xi}}^2 = \frac{1}{N_d \langle \delta \xi^2 \rangle_{\rho}}, \quad (3)$$

and depends explicitly on the number of death events, rather than on the total number of individuals, in the study.

Up to the appearance of sample-weighted average in the definition of the covariance matrix, Eq. (2) is a regression of independent variables ξ to the martingale residual of the Cox-Gompertz model. The latter is the difference between the actual survival, represented by the indicator variable δ^n , and the model mortality, integrated over the duration of the follow-up time for the same individual, $\int_{t_1}^{t_2} dtM(t, x^n) = N_d \rho^n$. What is more important, the solution can now be obtained by minimization of an extremely simple cost function

$$L = \sum_n \rho^n (\beta_{\xi} \xi^n - R^n)^2. \quad (4)$$

This is nothing else but a sample-weighted regression of ξ^n against the properly selected target function $R^n = \delta^n / (\rho^n N_d) - 1$. This form of the likelihood optimization can be easily done with any modern software library and hence is far more convenient for machine learning applications than the original Cox-Gompertz likelihood (1).

Significance of locomotor hazards and biological age in all-cause mortality evaluation. We first tested the age and hazard predicting models for the significance of their association with all-cause mortality with the help of a Cox proportional hazards model. Every test included the model prediction for each NHANES participant along with the age (data field RIDAGEMN), gender (RIAGENDR), diabetes (DIQ010), smoking (SMQ040), and hypertension (high blood pressure) (BPQ020). Also, the same test was carried out after including two additional covariates in the form of the average number of activity counts per day and the logarithm of this number, calculated from the physical activity recording tracks. Each covariate except for age were subsequently linearly detrended by all the other covariates and then standardized to zero mean and unit variance. After these preprocessing procedures, the significance for association was tested using Cox proportional hazards model yielding the effect, its 95% confidence intervals (CI) and p-value. The effect was further transformed into the corresponding difference in life expectancy by dividing the effect by the Gompertz exponential coefficient 0.085 per year.

The Cox model parameters and the significance p-values are summarized in Table 2 and were obtained using *survival*^{42,43} package implemented in R⁴⁴. Alternatively, we obtained the same p-values as probabilities of the effect deviations from zero using the analytical expressions for the effect and the effect determination error given by Eqs (2) and (3).

Convolutional neural networks for biological age and risks of death predictions. We trained a convolutional neural network to convert each physical activity record, a vector of 10080 values, representing subsequent activity counts per minute records (7-day-long recording sampled at 1 min^{-1} frequency), into a single value of estimated age or hazard rate. The network architecture was similar for both tasks and is shown in Fig. 3. We identified an optimal architecture for this deep learning model empirically, after exploring a range of combinations of layer depth and size. Before feeding data into convolutional network, we apply Batch Normalization⁴⁵. The network consisted of four convolutional layers with ReLU activation each followed by Max Pooling. The results were imported into two fully connected layers with ReLU activation. We applied the dropout⁴⁶ to regularize the dense layers. We used the RmsProp⁴⁷ optimizer with learning rate per sample set to 10^{-7} , minimal learning rate 10^{-17} and momentum 0.9. Dropout rate is set to 0.5 after first 9 epochs, and no dropout was used during these 9 epochs. Finally, output of the last layer is fed into a single linear neuron (a linear regression to the network target) producing the resulting value of age prediction.

The convolutional layers and output dense layer weights were initialized with Glorot uniform initializer⁴⁸. We used a gaussian initializer for the two internal dense layers weights. For CNN_Age model we scaled the weights of the output dense layer by the factor 85 to match the range of the possible output values.

To produce a deep CNN for mortality prediction using the log-likelihood (4), we started by building the simplest “zero-order” Cox-Gompertz PHM including the participant’s age and gender as independent covariate. In spite of the very small number of death cases in the data, the Gompertz exponent $\Gamma = 0.088$ per year was compatible with the currently accepted mortality rate $\Gamma = 0.085$ per year, corresponding to doubling time of 8 years⁴⁹. The initial mortality rate was $M_0 = 3 \times 10^{-5} \text{ yrs}^{-1}$, and therefore the model predicted 90.02 years of life expectancy, on average (gender contributed 5.2 years of the life expectancy difference).

A CNN trained to predict the linearized proportional hazards model target R^n from the raw physical activity tracks generated the best locomotor activity features ξ (output of the last hidden CNN layer) and identified the projection vector β_{ξ} (weights of last linear layer). The values R^n were inverse normal transformed and standardized to zero mean and unit variance with custom R⁴⁴ script. Figure 4C,D show learning curves for the training and cross-validation sets. Since some of the cross-validation curves show increase of MSE approximately after epoch 3000, we stopped training at epoch 3000. We tested the CNN output for significance by feeding it to a Cox model along with age, gender, and major risk factors: smoking, diabetes, high blood pressure and, optionally, average daily activity.

Data availability. All data needed to evaluate the conclusions in the paper are present in the paper. Additional data related to this paper may be requested from the authors.

References

- Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
- Peters, M. J. *et al.* The transcriptional landscape of age in human peripheral blood. *Nat Commun* **6**, 8570 (2015).
- Enroth, S., Enroth, S. B., Johansson, A. & Gyllenstein, U. Protein profiling reveals consequences of lifestyle choices on predicted biological aging. *Sci Rep* **5**, 17282, <https://doi.org/10.1038/srep17282> (2015).
- Choi, B. C., Pak, A. W. & Choi, J. C. Daily step goal of 10,000 steps: a literature review. *Clin. & Investig. Medicine* **30**, 146–151 (2007).
- Pyrkov, T. V. *et al.* Quantitative characterization of biological age and frailty based on locomotor activity records. *bioRxiv* 186569 (2017).
- Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C. & Ng, A. Y. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836* (2017).
- Putin, E. *et al.* Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging (Albany NY)* **8**, 1021 (2016).
- Cohen, A. A., Morissette-Thomas, V., Ferrucci, L. & Fried, L. P. Deep biomarkers of aging are population-dependent. *Aging (Albany NY)* **8**, 2253 (2016).
- Wang, Z. *et al.* Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *J. Biomed. Informatics* (2017).
- Oakden-Rayner, L. *et al.* Precision radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Sci. Reports* **7**, 1648 (2017).
- Horvath, S. & Levine, A. J. HIV-1 Infection Accelerates Age According to the Epigenetic Clock. *J. Infect. Dis.* **212**, 1563–73, <https://doi.org/10.1093/infdis/jiv277> (2015).
- Horvath, S. *et al.* Accelerated epigenetic aging in Down syndrome. *Aging Cell* **14**, 491–5, <https://doi.org/10.1111/accel.12325> (2015).
- Horvath, S. *et al.* Obesity accelerates epigenetic aging of human liver. *Proc. Natl. Acad. Sci. USA* **111**, 15538–15543 (2014).
- Marioni, R. E. *et al.* Dna methylation age of blood predicts all-cause mortality in later life. *Genome biology* **16**, 25 (2015).
- Horvath, S. *et al.* Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. *Aging (Albany NY)* **7**, 1159–70, <https://doi.org/10.18632/aging.100861> (2015).
- Christiansen, L. *et al.* DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell* **15**, 149–54, <https://doi.org/10.1111/accel.12421> (2016).
- Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
- Stellman, J. M. *Encyclopaedia of occupational health and safety* (International Labour Organization, 1998).
- Cox, D. R. Regression models and life-tables. In *Breakthroughs in statistics*, 527–541 (Springer, 1992).
- Efron, B. The efficiency of cox's likelihood function for censored data. *J. Am. statistical Assoc.* **72**, 557–565 (1977).
- Katzman, J. *et al.* Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931* (2016).
- Podolskiy, D. *et al.* Critical dynamics of gene networks is a mechanism behind ageing and gompertz law. *arXiv preprint arXiv:1502.04307* (2015).
- Levine, M. E. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *Journals Gerontol. Ser. A: Biomed. Sci. Med. Sci.* **68**, 667–674 (2012).
- Kristic, J. *et al.* Glycans are a novel biomarker of chronological and biological ages. *J. Gerontol. A Biol. Sci. Med. Sci.* **69**, 779–89, <https://doi.org/10.1093/gerona/glt190> (2014).
- Levine, M. E. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *J. Gerontol. A Biol. Sci. Med. Sci.* **68**, 667–674 (2013).
- Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* **16**, 90 (2016).
- Baird, G. S. *et al.* Age-dependent changes in the cerebrospinal fluid proteome by slow off-rate modified aptamer array. *Am. J. Pathol.* **180**, 446–56, <https://doi.org/10.1016/j.ajpath.2011.10.024> (2012).
- Gao, X. *et al.* Tobacco smoking and smoking-related dna methylation are associated with the development of frailty among older adults. *Epigenetics* (2016).
- Vidaki, A. *et al.* Dna methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci. Int. Genet.* **28**, 225–236 (2017).
- Franco, O. H., Steyerberg, E. W., Hu, F. B., Mackenbach, J. & Nusselder, W. Associations of diabetes mellitus with total life expectancy and life expectancy with and without cardiovascular disease. *Arch. internal medicine* **167**, 1145–1151 (2007).
- Horvath, S. *et al.* An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome biology* **17**, 171 (2016).
- Brown, A. E., Yemini, E. L., Grundy, L. J., Jucikas, T. & Schafer, W. R. A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion. *Proc. Natl. Acad. Sci.* **110**, 791–796 (2013).
- Ordóñez, F. J. & Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **16**, 115 (2016).
- Guan, Y. & Ploetz, T. Ensembles of deep lstm learners for activity recognition using wearables. *arXiv preprint arXiv:1703.09370* (2017).
- Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **313**, 504–507 (2006).
- Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinforma.* **19**, 185–193 (2003).
- Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779 (2015).
- Tedesco, S., Barton, J. & O'Flynn, B. A review of activity trackers for senior citizens: Research perspectives, commercial landscape and the role of the insurance industry. *Sensors* **17**, 1277 (2017).
- Gompertz, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philos. transactions Royal Soc. Lond.* **115**, 513–583 (1825).
- Olshansky, S. The law of mortality revisited: interspecies comparisons of mortality. *J. comparative pathology* **142**, S4–S9 (2010).
- Tarkhov, A. E., Menshikov, L. I. & Fedichev, P. O. Strehler-mildvan correlation is a degenerate manifold of gompertz fit. *J. theoretical biology* **416**, 180–189 (2017).
- Therneau, T. M. *A Package for Survival Analysis in S*, <https://CRAN.R-project.org/package=survival>, Version 2.38 (2015).
- Therneau, Terry M. & Grambsch, Patricia M. *Modeling Survival Data: Extending the Cox Model*. (Springer, New York, 2000).
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (2017).
- Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456 (2015).
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).
- Tieleman, T. & Hinton, G. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. Tech. Rep., Technical report, 31 (2012).
- Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256 (2010).
- Levy, G. L. B. *The Biostatistics of Aging: From Gompertzian Mortality to an Index of Aging-relatedness* (John Wiley & Sons, 2014).

Acknowledgements

The authors thank their colleagues from Gero K. Avchaciov, A. Tarkhov, D. Shishov, and G. Getmantsev for extensive help with handling the data and proof-reading of the work. The work has been funded by Gero LLC.

Author Contributions

T.V.P., S.M., L.I.M., and P.O.F. designed the computational experiments and analyzed the results. K.S., M.B., A.K. and S.M. performed calculations using convolutional neural networks. T.V.P., B.Z., A.Z., M.P. performed all other calculations, theoretical modeling and statistical analysis of all the results. All authors reviewed the manuscript.

Additional Information

Competing Interests: P.O. Fedichev is a shareholder of Gero LLC. T. V. Pyrkov, B. Zhurov, A. Zenin, M. Pyatnitskiy, L. Menshikov, and P.O. Fedichev are employees of Gero LLC. A patent application submitted by Gero LLC on the described methods and tools for evaluating health non-invasively is pending.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018