bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Quantitative Characterization of Biological Age and Frailty Based on Locomotor Activity Records

Timothy V. Pyrkov[1], Evgeny Getmantsev[1], Boris Zhurov[1], Konstantin Avchaciov[1], Mikhail Pyatnitskiy[1], Leonid Menshikov[1], Kristina Khodova[1], Andrei V. Gudkov[2], Peter O. Fedichev[1,3]*

[1] *Gero LLC, 24/2 Novokuznetskaya street, Moscow 119017, Russia*
[2] *Roswell Park Cancer Institute, Buffalo, NY 14263, USA and*
[3] *Moscow Institute of Physics and Technology, 141700, Institutskii per. 9, Dolgoprudny, Moscow Region, Russian Federation*

Since chronological age is not a complete and accurate indicator of organism aging, the concept of biological age has emerged as a well-accepted way to quantify the aging process in humans and laboratory animals. In this study, we performed a systematic statistical evaluation of the relationships between locomotor activity and biological age, mortality risk, and frailty using human physical activity records from the 2003-2006 National Health and Nutrition Examination Survey (NHANES) and UK BioBank (UKBB) databases. These records are from subjects ranging from 5 to 85 years old and include 7-day long continuous tracks of activity provided by wearable monitors as well as data for a comprehensive set of clinical parameters, lifestyle information and death records, thus enabling quantitative assessment of frailty and mortality. We proposed a statistical description of the locomotor activity tracks and transformed the provided time series into vectors representing individual physiological states for each participant. Using this data, we performed an unsupervised multivariate analysis and observed development and aging as a continuous trajectory consisting of distinct phases, each corresponding to subsequent human life stages. Therefore, we suggest the distance measured along this trajectory as a definition of the biological age. Consistent with the Gompertz law, mortality, estimated with the help of a proportional hazard model, was found to be an exponential function of biological age as quantified herein. However, we observed that the significant contribution of clinical frailty to mortality risk can be independent of biological age. We used the biological age and mortality models to show that some lifestyle variables, such as smoking, produce a reversible increase in all-cause mortality without a significant effect on biological age. In contrast, medical conditions, such as type 2 diabetes mellitus (T2DM) or hypertension, are associated with significant aging acceleration and a corresponding increase in mortality as well. The results of this work demonstrate that significant information relevant to aging can be extracted from human locomotor activity data and highlight the opportunity provided by explosive deployment of wearable sensors to use such information to encourage lifestyle modifications and clinical development of therapeutic interventions against the aging process.

I. INTRODUCTION

Accurate and non-invasive quantification of the aging process is essential for successful translation of basic research in the field of aging into possible future clinical practice. Most studies of aging in model organisms, such as *C. elegans*, *S. cerevisiae*, and *D. melanogaster*, involve direct measurements of lifespan to characterize pro- or anti-aging effects of gene variants, nutrition conditions or experimental therapies. In longer-lived animals, such as mammals, and especially in humans, analysis of longevity itself is generally not practical since it would require long and exceptionally well-controlled experiments. Many age-related physiological changes are similar in mice and humans and hence can be described by a universal frailty index, which was recently proposed as a promising pre-clinical indicator to quantify aging [1]. Other useful metrics of aging include health span, maximum lifespan, and biological age [2], along with less commonly employed properties, rooted in aging theory phenomenology, such as Strehler and Mildvans concept of vitality [3]. It remains to be seen, however, if and how any of these measures of aging are related to each other in human populations and whether the same relations hold and hence can be reliably examined in animal models.

Aging is a continuous phenotypic change. Therefore, to answer questions regarding the age-dependencies of physiological variables, it is appropriate to apply the language and tools of dynamical systems theory to relate changes in physiological parameters to organism-level properties, such as mortality and lifespan. However, investigations of this kind are complicated by the high intrinsic dimensionality of high-throughput biological data, overlaid by batch effects and prohibitive costs of large-scale human studies. This situation is improving due to the recent explosive deployment of web-connected wearable devices that provide personal digitized health records, including measurements of locomotion, heart rate, skin temperature, etc. It is projected that $400M$ such devices will be in use worldwide by 2020 [4]. This provides an unprecedented opportunity to monitor physiological changes in large human populations [5] and improve our understanding of how such changes are related to health and lifespan [6] and hence present a rich and yet an untapped source

* tim.pyrkov@gero.com, peter.fedichev@gero.com

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of large scale human data for aging research.

In this study, we performed a systematic statistical evaluation of the relationships between locomotor activity and biological age, mortality risk, and frailty using human physical activity records from the 2003 – 2006 National Health and Nutrition Examination Survey (NHANES) and UK Bio Bank (UKBB) databases. These large databases contain data for 7-day long continuous tracks of activity provided by wearable monitors as well as health and lifestyle information, death records, and data for a comprehensive set of clinical parameters sufficient to compute a frailty index for every participant. We proposed a statistical description of 7-day long locomotor activity tracks and performed an unsupervised principal components analysis of the study participants physical activity representations. This revealed human development as a continuous process following a trajectory in the physical activity parameters space. According to the Gompertz law [7], the mortality rate in human populations increases exponentially starting at age 40. The dynamics of the physiological variables along the corresponding trajectory segment is approximately a linear function of age. Therefore, we proposed the distance measured along the path as the natural definition of biological age or bioage.

Although bioage was found to account for most of the variation in mortality observed in our study, there was a notable contribution to mortality that was independent of biological age and associated with a high frailty index score. Statistically, we observed an increasing fraction of such participants starting from approximately the age of the middle-age to old-age transition point on the aging trajectory. We used models of biological age and mortality to show that some lifestyle choices, such as smoking, produce a reversible increase in all-causes mortality without an appreciable effect on biological age. On the other hand, some medical conditions, such as type 2 diabetes mellitus (T2DM) and hypertension, were associated with significant aging acceleration (elevated biological age in the chronological age-matched cohorts).

Overall, the results of this work provide an indication of the power of the vast amounts of activity-related and physiological data that can be easily obtained via simple wearable monitoring devices. In particular, with respect to aging, our statistical analysis of locomotor activity data from such devices has revealed a new physical activity-based descriptor of biological age (bioage) and suggests that mortality risk is determined by both bioage-related and bioage-independent components, which may be differentially amenable to modification by lifestyle changes or therapeutics.

II. RESULTS

A. Quantification of Human Locomotor Activity

For this study, we used two large-scale repositories of wearable accelerometer track records made available by the $2003-2006$ National Health and Nutrition Examination Survey (NHANES, 12053 subjects, age range $5-85$ years old) and UK Biobank (UKBB, 95609 subjects, age range $45-75$ years old). For both NHANES and UKBB, a 7-day long continuous activity track was collected for each subject, as well as data for a comprehensive set of clinical variables, and a record of death within up to ten years following activity monitoring. Human physical activity is usually collected in the form of time series of direct sensor readouts, such a 3D accelerations, sampled at a specified frequency. Instead, NHANES provides sequences of transformed quantities such as the number of steps or activity counts per minute. Figure 1A shows plots of two representative 2-day long activity count tracks from a younger (age 43) individual and an older (age 65) individual, which we selected by the same level of total activity. Nevertheless, their patterns of activity were qualitatively different. Transitions between the states corresponding to different levels of physical activity appeared to be random. Therefore, instead of trying to approximate the precise shape of the activity time series, we chose to apply a probabilistic model, specifically Markov chain approximation, a simple yet powerful tool from stochastic processes theory apparatus (see [8] for a review of applications, including stochastic modelling of biological systems).

Statistical description of the participants' activity was based on the concept that any future state of a Markov chain is completely determined by its current state and the probabilities of transitions between different states. Therefore, we divided physical activity measurements ranges into eight discrete bins representing activity states (numbered from 1 to 8 and corresponding to increasing activity levels, see histograms to the left of the activity tracks in Figure 1A) and counted the transitions between every consecutive pair of states along the track. The number of transitions from state j to i was then normalized to the number of times that state j was encountered along the activity record. This yielded the kinetic transition rate, i.e. the probability of a stochastic "jump" from state j to state i per unit time. We then combined these transition rates into the transition matrix (TM) elements (shown as bins in heatmaps to the right of the activity tracks in Figure 1A), which is thus a complete description of the underlying Markov chain model (see Materials and Methods and Figure 1A additional details).

On a more technical level, the TM element values can be related to the organism's responses to external perturbations on physiological time scales. To make this connection, we checked explicitly that the TM elements satisfied a detailed balance condition [9] and hence the TM eigenvalues represent inverse equilibration times. Using

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

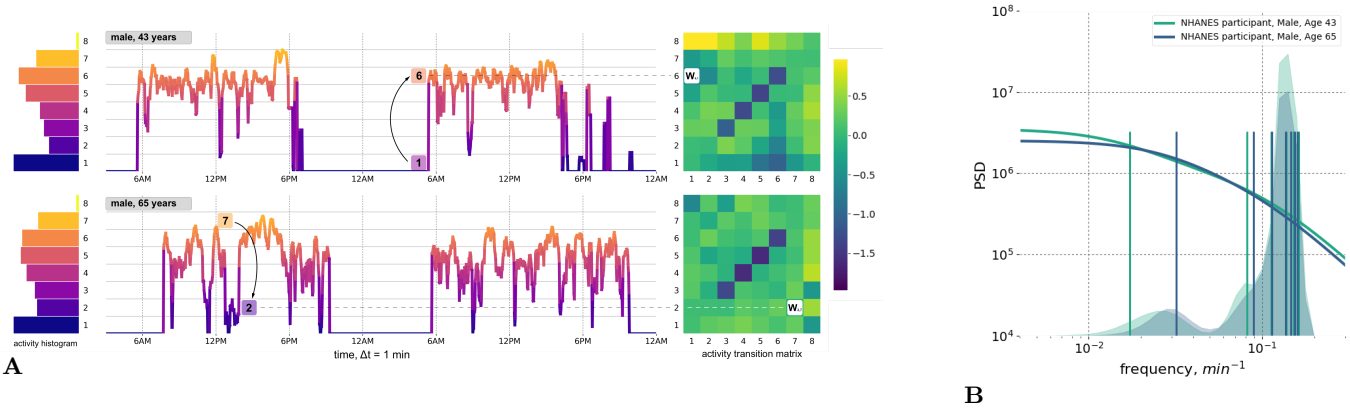

FIG. 1: **Quantitative Description of Human Locomotor Activity Tracks**
A. Representative 2-day long locomotor activity tracks from a younger individual (age 43) and an older (age 65) individual in the NHANES 2003 − 2006 cohort. The sampling rate was 1 min^{-1}. Activity levels were divided into 8 bins as indicated in the histograms to the left of the tracks. Both individuals demonstrated approximately the same levels of average physical activity. To the right of each activity track is the corresponding matrix of transition rates, each element W_{ij} describing the probabilities of random jumps from each discrete activity state j to every other possible state i per unit time. The kinetic rates are presented in the form of the transition matrices (TM), the color scale represents the differences relative to the population mean values, element-wise in logarithmic scale.
B. Power Spectral Densities (PSD) of simulated physical activity tracks computed for the same participants as in Figure 1A (age 43 subject, teal; age 65 subject, blue). PSDs were reconstructed using the corresponding TM elements as the Markov chain kinetic coefficients (see Appendix A for details on calculations). The color-coded ticks along the horizontal axis represent the positions of the TM eigenfrequencies. For comparison, histograms in the background of the graph show distributions of TM eigenfrequencies (the density of states) computed for two cohorts of study participants of age ranges 35 − 45 (light green) and 65 − 75 (light blue) years old.

the relation between the autocorrelation function of the time series and the Markov chain TM from Appendix A, we reconstructed a Power Spectrum Density (PSD) display (Figure 1B) of the physical activity signals for the same two study participants shown in Figure 1A. Figure 1B also shows the discrete sets of TM eigenvalues (the TM spectra) for the same individuals. The crossover frequency on the PSD plots coincides with the lowest TM eigenvalue, corresponding to a time scale in the range of tens of minutes. The time scale corresponding to the eigenvalue is considerably longer than any period associated with body motion and, therefore, should reflect the organism's physiological state. The observed decrease of the limiting time scale signifies a reduction of temporal correlations of physical activity, commonly observed in human [10] and animal [11] studies of aging and age-associated neurological and mental disorders, including Alzheimer's disease [12], depression [13], and bipolar disorder [14].

A transition matrix is a conceptually simple and physically intuitive way to illustrate the aggregate characteristics of physical activity-time series. TM elements are kinetic transition rates and the spectral properties of TM are directly related to the organism's responses at physiological time scales. Therefore, TM are useful descriptors containing information for a set of parameters representing human physiological state vectors in a form that can be used alone (as in this study) or along with other physiological metrics in human health-related applications.

B. Manifestations of aging in Locomotor Activity

According to our assumptions described above, TM provide an aggregate representation of each study participant's activity state during the time of monitoring. To reveal the intrinsic structure of the physical activity data for the entire NHANES study population, we used Principal Components Analysis (PCA, [15]). The results of this analysis are shown in Figure 2A, with the position of each point defined by the average PC score for a cohort of age-matched men or women separated by one year of age. The plot of PC_1 vs. PC_2 vs. PC_3 in Figure 2A clearly shows that the activity state vector evolves over the course of human lifespan. The aging trajectory is continuous and yet is visually broken into distinct phases that are easily recognizable as life stages (chronological age ranges). Following a previously established age classification [16], we divided the trajectory into four segments covering childhood and adolescence (younger than 16 years old), followed by early adulthood (16 − 35 y.o.), middle (35 − 65 y.o.) and older ages (older than 65 y.o.). Each life stage has a characteristic range of average PC values that is essentially the same for men and women. These results suggest a universal character of changes associated with each period in development

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

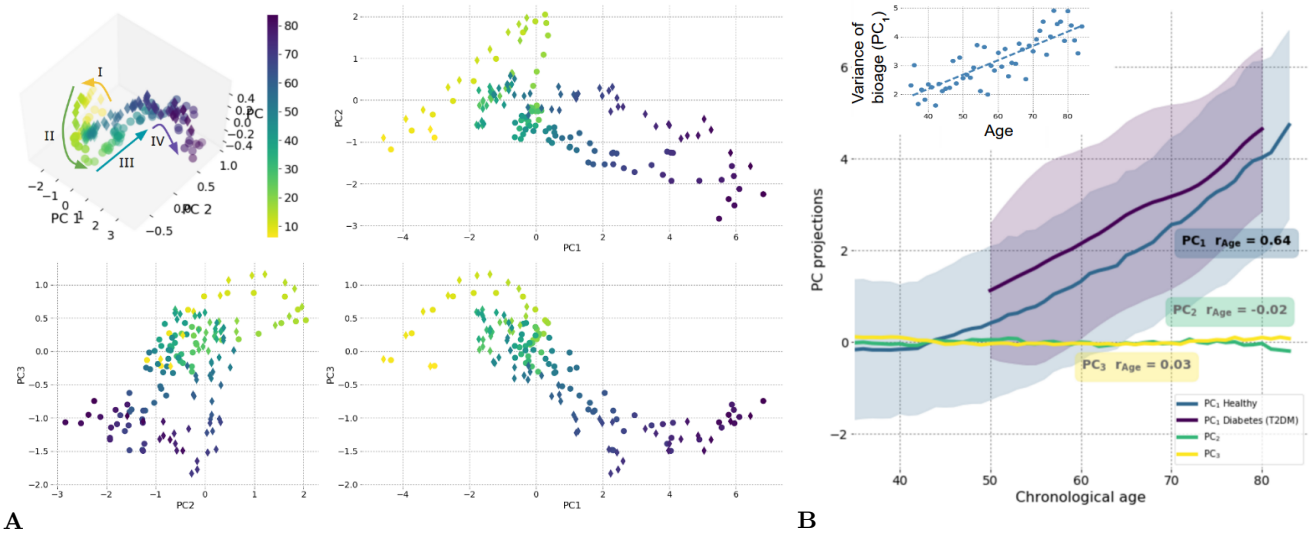

FIG. 2: **Principle Component Analysis of Physical Activity Descriptors**
A. Principal Component Analysis (PCA) of the descriptors representing one-week long physical activity tracks of NHANES participants (top left). The projections PC_2 vs. PC_1 (top right), PC_3 vs. PC_1 (bottom right) and PC_3 vs. PC_2 (bottom left) are also shown. To simplify the presentation, we averaged the PC scores in cohorts separated by one year of age (color scale ranging from yellow (< 10 y.o.) to purple (> 80 y.o.) for men (diamonds) and women (circles) separately. The Roman numerals I-IV and corresponding colored arrows illustrate the dynamics of PC scores over subsequent stages of human development: I) age < 16; II) age $16 - 35$; III) age $35 - 65$; IV) age > 65.
B. Relationship between PCA values and chronological age. PC_1, PC_2, and PC_3 values obtained for NHANES $2003 - 2006$ 40+ cohort reveals the evolution along the direction of the largest variance in data and significantly associated with age, named PC_1 (Pearson's coefficient of correlation with age $r = 0.64$). None of the other PCs show correlations greater than $r \sim 0.2$. PC_1 data is shown separately for healthy subjects and type 2 diabetes mellitus (T2DM) patients. The width of the PC_1 distribution in both healthy and T2DM populations is illustrated by shading (blue and purple, respectively) corresponding to a range of one standard deviation in each age cohort. The inset panel shows the variance in biological age (PC_1) in the age- and sex-matched cohorts as a function of chronological age.

and aging.

According to the Gompertz law, mortality risk in human populations increases exponentially in mid-life, starting from the age of about 40 [7, 17]. The corresponding age range boundary between the development trajectory fragments II and III is visible in Figure 2A and can be identified with the early-adulthood to middle-age transition. The subsequent cross-over between middle-age and older-age occurs in the vicinity of the average human lifespan. To better focus our study on aging, we limited all further analysis to participants older than 40 years old. In this restricted dataset, aging was observed to manifest itself as the evolution of the participants' physiological state along the PC_1 direction, which by definition is the direction of the most variance in the data. PC_1 scores in this group of participants were strongly associated with chronological age (Pearson's correlation coefficient $r = 0.64$, see Figure 2B). PCA is an unsupervised learning technique and, therefore, can be used to infer functional dependence of physiological state variables on age without a prior hypothesis. Therefore, we propose the first PC score, PC_1, as a natural definition of biological age, or bioage, a quantitative measure of the aging process in the most relevant age range. This argument is supported by our observation that no other PCs show any significant association with age (Figure 2B). A more careful examination of Figure 2A reveals that variation of PC_2 and PC_3 scores is associated with the transition between early adulthood and middle age stages.

Biological age defined as PC_1 was found to increase approximately linearly with chronological age for NHANES participants over 40 years old (Figure 2B). The variance in biological age (PC_1 distribution width squared in age- and sex-matched cohorts) in this population increased linearly with age (inset in Figure 2B). The latter result is a hallmark of diffusion, suggesting that biological age not only drifts in time but also undergoes a random walk under the influence of stochastic forces (see Discussion for more details).

Any appreciable difference in the rate (change in biological age per unit of time) or direction of aging among study participants would lead to faster growth of the bioage distribution width as a function of chronological age and, therefore, cannot be supported by the data. To highlight the apparent stability of the aging trajectory, we assessed the effect of a common age-related disease,

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

type 2 diabetes mellitus (T2DM), on biological age as defined by PC_1. The average and range (standard deviation) of biological age in age-matched cohorts of healthy subjects and T2DM patients are compared in Figure 2B. Generally, the T2DM patients appeared to be older, biological age-wise, than their chronological age-matched peers. The difference between the healthy and T2DM groups did not significantly change with age, which suggests that the disease does not modify the rate of aging.

C. Biological Age and the Gompertz law of Mortality

To examine how biological age (again, defined as PC_1 from the analysis of locomotor activity data described above) relates to lifespan and mortality, we first confirmed that the mortality rate indicated by the NHANES death register is an exponential function of chronological age as predicted by the Gompertz law. Next, since the state vector dynamics in the physical activity parameters space was approximately linear with age, we were able to naturally model the exponentially increasing individual risks of death by using a log-linear model, such as a variant of the Cox proportional hazards model [19]. We used the parametric Cox-Gompertz proportional hazard model in the form of maximal likelihood minimization adapted from [20], since this model allowed us to obtain explicit mortality predictions rather than implicit proportional hazards values, which are uncertain up to an unknown baseline hazard function. To produce a robust model, we used a batch mode stochastic gradient descent with cross-validation (see Materials and Methods section V C for details). The final model yielded the Gompertz exponent $\Gamma \approx 0.08 \pm 0.01\ y^{-1}$, which is very close to the commonly accepted value $\Gamma \approx 0.085 y^{-1}$ corresponding to the mortality rate doubling time of 8 years [21]. Mortality risk was determined in this way for every NHANES study participant and found to increase exponentially as a function of biological age (Figure 3A; determination coefficient of the corresponding log-linear model $R^2 = 0.81$). This further supports our identification of the PC_1 score as a quantitative measure of aging.

Variation in biological age explains most of the variation in mortality observed in the NHANES study (Figure 3A). However, there was also a significant spread in the predicted log-mortality values in biological age-matched cohorts, amounting to as much as 25% of the total difference in the population. Given the value of the Gompertz exponent of the mortality model, the observed differences correspond to up to ≈ 10 years of remaining lifespan. To understand the nature of the unexplained difference in mortality, we hypothesized that the log-mortality predictions into a sum of two parts, one of which is associated with (proportional to) biological age and one of which is statistically independent of biological age.

Next, we investigated the relation between biological age, the risk of death, and another commonly employed descriptor of aging: a clinical Frailty Index (FI). FI is a composite measure of the health of an individual calculated as the proportion of health deficits present in an individual out of the total number of age-related health variables considered [22]. Using a FI variant adapted for use with the NHANES meta-data from [18], we computed FI values for every study participant and categorized them into one of the three FI ranges: "non-frail", "vulnerable", and "frail/most frail", as defined in the same source. Figure 3B shows boxplots of the sex- and the chronological age-adjusted bioage distributions in cohorts of NHANES participants, split according to their FI category. These data revealed a statistically significant correspondence between aging acceleration, i.e. higher values for biological age after adjustment for sex and chronological age, and FI. The increases in aging acceleration between the non-frail and vulnerable groups and between the vulnerable and frail/most frail groups were both significant.

For comparison, in Figure 3C, we provide box-plots of the sex-adjusted distributions of the log-mortality component that is biological age-independent for the same three cohorts of NHANES participants. This shows that bioage-independent mortality is similar between the non-frail and vulnerable groups, see Figure 3C. At the same time, the bioage-independent component is much significantly higher in the highest FI range (frail/most frail). Therefore, except in cases with high clinical FI values, biological age serves as an indicator of frailty as well as mortality (see above). In this study FI increases gradually first and then reaches its maximum value, at approximately the age corresponding to the transition between middle- and old- age (development stages III and IV in Figure 2A), marking the end of health span (we will focus on this in Discussion). These observations suggest that the bioage-independent contribution to mortality can be attributed to most frail phenotype.

D. Association of unhealthy lifestyles and medical conditions with increased locomotor activity-based biological age and mortality risk

In the next stage of our study, we evaluated how biological age and mortality (biological age-dependent and -independent components) are influenced by known hazardous behaviours/lifestyle choices or medical conditions. As shown in Figures 4A and 4B, we constructed volcano plots to illustrate statistical associations between the predicted log-mortality components for NHANES participants and the NHANES 2003-2006 Questionnaire, Demographics and Examination lifestyles and medical conditions labels.

Figure 4A and its legend lists the lifestyles and medical conditions that showed statistically significant associations with biological aging acceleration, defined as the

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

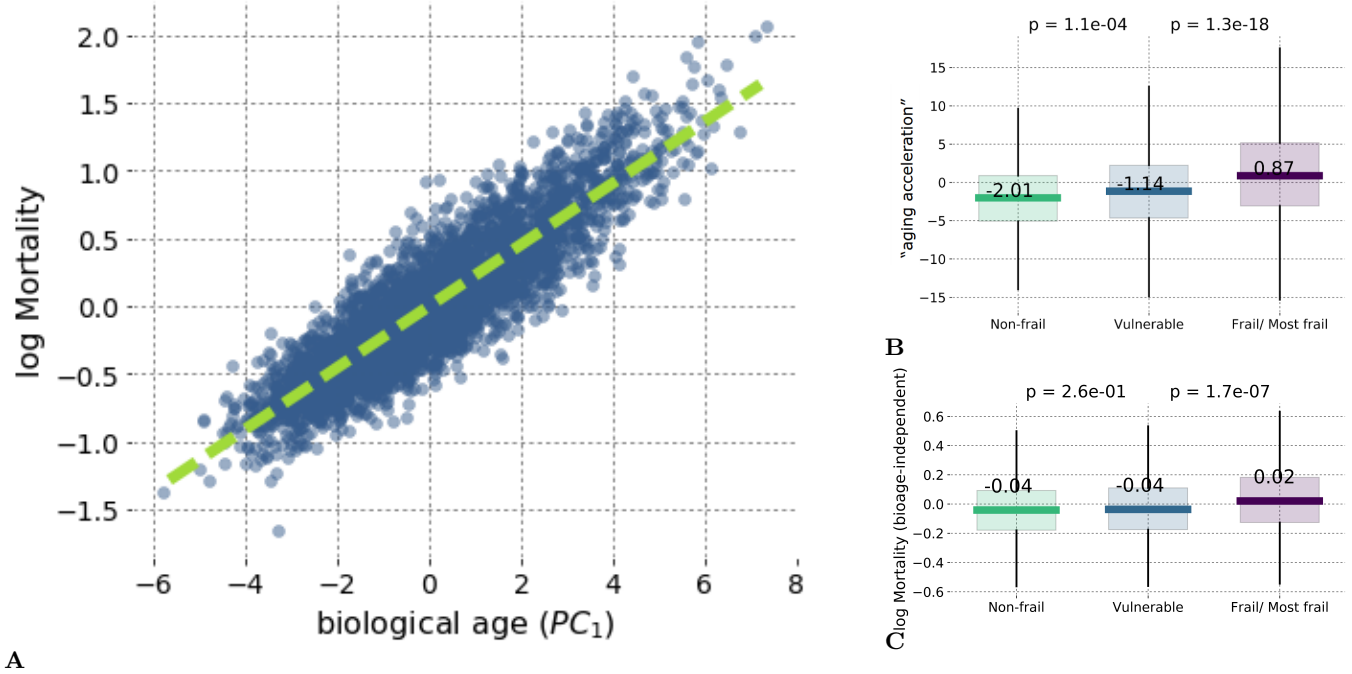

FIG. 3: **Relationships Between Biological Age, Mortality and Frailty**
A The logarithm of mortality (log-mortality) predicted for NHANES participants increases linearly as a function of biological age (PC_1). Biological age explains most, but not all, of the observed variance in mortality (coefficient of determination, $R^2 = 0.81$);
B Box-plots of biological age distributions calculated for sex- and age-matched cohorts of NHANES participants representing frailty index categories from [18]. The horizontal line within each box indicates the mean value and the shaded box indicates variance (one standard deviation). P-values for pairwise comparisons between FI categories by KS-test are shown above each graph;
C Box-plots (as in **B**) of the predicted biological age-independent log-mortality component for age and sex-matched cohorts of NHANES participants representing frailty index categories from [18].

elevated bioage in a group after adjustment for sex and chronological age (see e.g. [23]). The identified associations included high blood pressure ($A1$), C-reactive protein level ($B1$), clinical parameters associated with diabetes [the diagnosis itself ($E1$), taking insulin ($E2$), elevated levels of glycated haemoglobin ($G1$) and blood serum glucose ($J2$, $J5$)], along with self-reported weight ($K1$-$K3$) and nutrition status information ($F1$-$F2$). The association of clinical parameters related to diabetes with aging acceleration is consistent with the generally higher locomotor activity-based biological age (PC_1 value) of NHANES participants diagnosed with T2DM compared to healthy controls across age cohorts (Figure 2B).

In a similar manner, the volcano plot shown in Figure 4B shows lifestyles and medical conditions that demonstrated statistically significant associations with the bioage-independent log-mortality component. The most striking association here was with parameters related to smoking ($F1$-$F3$), including elevated blood levels of cotinine ($B1$), the predominant metabolite of nicotine, which is used as a biomarker for exposure to tobacco smoke. Associations with general ($C1$) and physical health ($C2$) conditions, along with mental health-related parameters ($C3$, $D1$) are in agreement with our association of bioage-independent hazard with frailty (see above).

Comparison of the predicted log-mortality component distributions between smokers and non-smokers, after adjustment for age and sex, showed little effect on biological age (i.e., no strong evidence of aging acceleration associated with smoking, see Figure 4C). In contrast, the difference in the bioage-independent component of log-mortality between smokers and non-smokers was highly significant (Figure 4D). In contrast, the difference in the bioage-independent component of log-mortality between smokers and non-smokers was highly significant (Figure 4D). Study participants who smoked in the past but quit smoking, demonstrate an significant improvement in the bioage-independent component of log-mortality compared to current smokers (Figure 4D) although there was no difference in bioage between the groups after adjustment for sex and age (Figure 4C). From a dynamic point of few, this is not very surprising, since the bioage-independent mortality component originates from deviations of the physiological state from the development trajectory due to lifestyle choices and diseases and, if

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

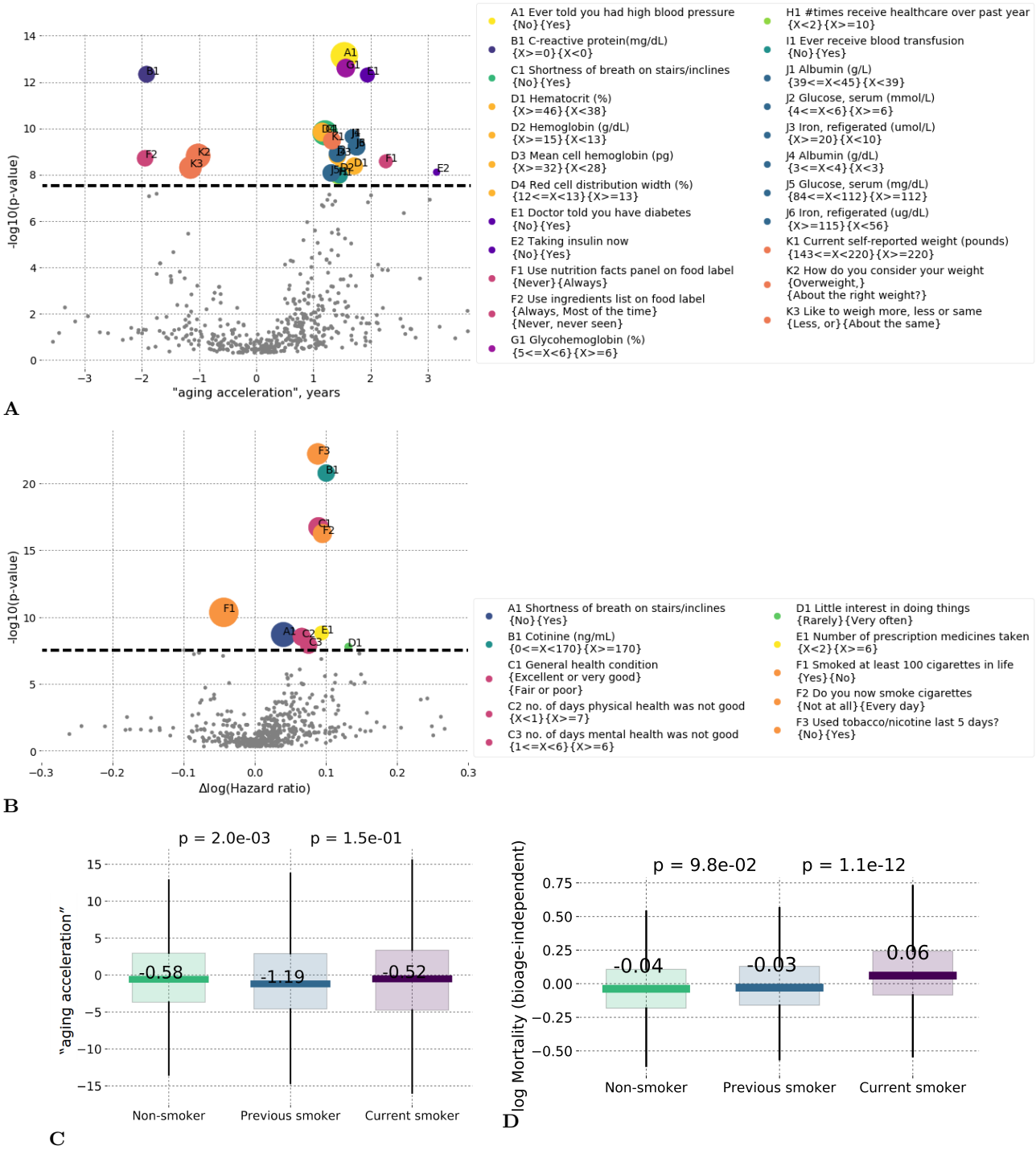

FIG. 4: **Association of lifestyle variables and health conditions with biological age-related and independent components of mortality**

A, **B** Lifestyle variables and health conditions are ranked in volcano plots according to the differences in log-mortality (horizontal axis) between groups of NHANES 2003-2006 study participants and the statistical significance (p-value, vertical axis) for bioage (**A**) and bioage-independent mortality component (**B**). The circle sizes represent the relative number of participants in the compared groups. Participants were divided into groups based on lifestyle habits (e.g., smoking, sleep quality, diet) or health conditions (e.g., blood pressure, general health status, diabetes, respiratory health, mental health, etc.). Lifestyle and health conditions are color-coded according to categories of NHANES 2003-2006 Laboratory and Questionnaire Data above the significance level 0.001 after the Bonferroni correction $n = 10^4$ (total number of considered metadata fields rounded up to power of 10).

C, **D** Boxplots show the distributions of bioge (**C**) and bioage-independent mortality component (**D**) calculated using locomotor activity data from the NHANES 2003-2006 study for current smokers, previous smokers and the rest of the population ("non-smokers"). Only subject in the age range 40-70 y.o. were included in the analysis. P-values for pairwise comparisons between FI categories by KS-test are shown above each graph.

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

small, should be reversible. This finding is also in qualitative agreement with an estimated longevity dividend of as much as 5 years upon quitting smoking [24].

The impacts of smoking on biological age and mortality were verified using an independent dataset of locomotor activity records from UK Biobank (UKBB, 95,605 locomotor activity samples after exclusion of poor quality records). We computed TM descriptors using time series of activity counts per minute for UKBB participants and applied the mortality model without any additional pre-training. As seen for the NHANES study, there was no significant difference in biological age means or distributions after correcting for sex and age (i.e., no aging acceleration) between current smokers, previous smokers and non-smokers, compare the UKBB and the NHANES analysis results in Figures 6A and 4C. Also similar to the NHANES findings, there was a statistically significant difference in the biological age-independent component of mortality between current smokers and either of the other two groups (participants with no smoking history and those who quit smoking in the past). The log-mortality ratio between previous smokers and current smokers was very similar to that observed for the NHANES, see Figure 6B.

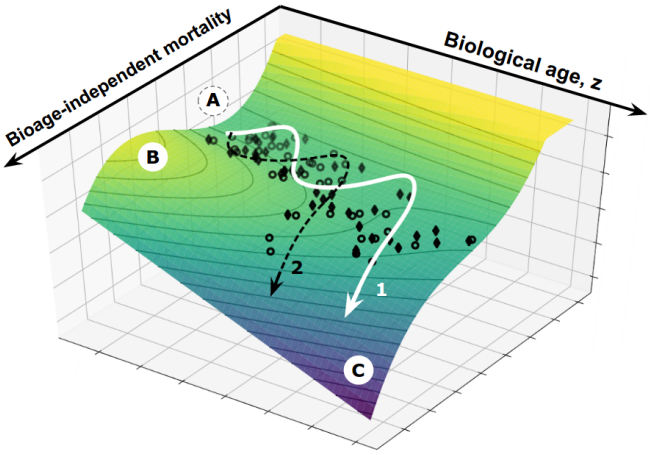

FIG. 5: **Graphic Explanation of Aging Dynamics and Mortality Model**
The trajectory of aging is shown superimposed on the potential energy landscape (vertical axis), which provides a schematic visualization of the constraints provided by the underlying regulatory network. Each dot represents the physical activity state vectors of an age- and sex-matched cohort of NHANES participants (men, diamonds; women, circles). Cohorts were separated by one year of age. The axes in the horizontal plane are (i) biological age, and (ii) biological age-independent mortality. The stability basin A is separated from the unstable region C by the potential energy barrier B. Linear coupling between the modes leads to a decrease in barrier height and thus to an exponential increase in the probability of crossing the barrier and subsequent organism death (see trajectory examples 1 and 2) as a function of biological age (see explanations in the Discussion).

III. DISCUSSION

We performed a systematic identification of biomarkers of age and frailty using an extensive human physical activity records collection. The phenotypic changes associated with the development and aging have different dynamics depending on the life stage. We identify the age range 40+ as the most relevant to the studies of aging in humans in relation to the Gompertz mortality law.

Biological age is a phenomenological organism-level property that is linearly associated with age and serves as a key indicator of aging and all-cause mortality. The linear association of any physiologically relevant variables with age is a hallmark of aging studies in human subjects and, therefore, can be used to construct useful "biological clocks". Examples of this include DNA methylation [25, 26], IgG glycosylation [27], blood biochemical parameters [28], gut microbiota composition [29], and cerebrospinal fluid proteome [30]. To date, the "epigenetic clock" based on DNA methylation (DNAm) levels [25, 26] appears to be the most accurate measure of aging, showing remarkably high correlation with chronological age. The DNAm clock predicts all-cause mortality in later life better than chronological age [31], is elevated in groups of individuals with HIV, Down syndrome [32, 33], obesity [34], but is not correlated with smoking status [35].

We find that the biological age signature computed from locomotor activity was found to be elevated in cohorts of NHANES participants diagnosed with T2DM or those characterized by hypertension, increased levels of C-reactive protein, excess weight, or elevated frailty index. An earlier analysis of aging acceleration using the epigenetic clock with data from the Woman's Health Initiative study did not identify significant associations with these indications [23]. We were able to obtain the same results with the help of the same approach in a similar sized age- and gender- matched subpopulation of NHANES participants. A more careful examination, however, reveals that the multivariate test from [23] was underpowered due to intrinsic correlations between the participants characteristics. We observed that, e.g., removal of the diabetes-associated features such as blood glucose and insulin levels from the model leads to a dramatic increase of the significance of the association between the disease and aging acceleration.

The tightly coordinated change of the physiological variables with age is a direct consequence of intrinsic low-dimensionality of the organism state dynamics, a hallmark feature of criticality. This observation should not be surprising since slow evolution of physiological variables associated with major "biological programs", such as morphogenesis [36] and aging [37], exhibit characteristic properties of critical dynamics, such as critical slowing down, rising variance, strong correlations between

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

key variables, and non-Gaussianity in the distribution of fluctuations [38]. We find that in humans between approximately age 40 and the average lifespan, the evolution of physiological parameters with age is dominated by the dynamics of a single mode. In [37] we suggested that the stochastic drift of the collective variable associated with the critical mode is the driving force behind the characteristic increase in mortality with age. We observed that the corresponding mode vector coincides with the singular eigenvector of the data covariance matrix and can, therefore, be reliably identified with the help of PCA in an unsupervised way (i.e., without prior assumptions regarding the functional dependence of the biologically-relevant variables on age). Since the first PC score (PC_1) grows approximately linearly with age, the stochastic broadening of the PC_1 variation is small, and the PC scores with lesser variance (PC_2, PC_3) are practically independent of age, we propose PC_1 as a natural measure of biological age (Figure 2B). The association between the aging mode variable and biological age was confirmed the finding that most of the mortality variation in our study can be explained by biological age alone (Figure 3A).

Biological age thus emerges as an organism-level property associated with the dynamical properties of the underlying regulatory network. Therefore, we expect multiple phenotypic changes associated with biological age on various levels of an organisms organization to occur in a coordinated fashion. This suggests that different biological clocks based on any convenient phenotypic feature, even features as different as DNA methylation and locomotor activity, should yield very similar biological age predictions for the same subject.

In our study, we observed that variance of biological age in sex- and age-matched cohorts increases linearly with age. This is a signature of a random walk, suggesting that the biological age, as an organism-level property, does not only increase as a function of age, but also undergoes Brownian motion under the influence of stochastic forces. Therefore, we suggest that popular regression models of biological age can be further refined by explicitly including the effects of stochastic broadening.

The exponential form of mortality as a function of biological age shown in Figure 3A is typical for situations where the lifespan of a system is limited by the decay of a metastable state (see, e.g., [39]). This is schematically illustrated in Figure 5, where the small fluctuations of the physiological state variables that are independent of biological age are reversible, which means that the mode variables describing the state vector deviations from the aging trajectory are dynamically stable. In contrast, the large amplitude fluctuations are most certainly not reversible and manifest themselves as diseases. The dots in Figure 5 represent the data for NHANES participants, averaged over the sex- and age-matched cohorts and projected onto a subspace spanned by the direction associated with biological age and bioage-independent mortality (axes in the horizontal plane).

We superimposed the experimental points on a schematic representation of the effective potential energy surface (the vertical axis) set by the underlying regulatory network constraints. Around the early adulthood to mid-life transition, the organism state vector starts in a potential energy basin A separated from the dynamically unstable regions C by sufficiently high potential energy barriers B. The dynamics of the state variables are critical, which means that there is no or almost no curvature in the potential energy in the direction associated with biological age. With the natural assumption that the mode coupling is weak, the barrier heights depend on biological age linearly, and hence the probability of barrier crossing increases exponentially with biological age. Once the (presumably lowest) barrier is crossed, the dynamic stability along the corresponding mode vector will be lost (see example trajectories 1 and 2 in Figure 5, which differ by the age at which barrier crossing occurs). The resilience understood as the ability to regain the homeostatic state disappears, and the deviations of the physiological parameters develop beyond control. The situation manifests itself as development of extreme frailty (Figure 3C) and, eventually, certain death of the organism. On a population level, the loss of stability happens approximately at the age range corresponding to the average lifespan. The point on the aging trajectory corresponds to the middle-age to older-age transition and signifies the end of health span.

The physical picture behind the presented scenario reveals the dynamic origins of the postulates underlying the Strehler-Mildvan theory of aging [3], in which a linear decrease in vitality leads to an exponential increase in mortality with age. We note that the aging at criticality conjecture is then the necessary mechanistic link behind the phenomenological vitality concept and the dynamics of the physiological state variables. We relate the biological age with Strehler-Mildvan vitality deficit. The stochastic drift along the aging direction is then naturally the driving force behind the vitality attrition in the Strehler-Mildvan theory. The exponential Gompertz mortality increase with age appears to be a consequence of a gradual loss of the dynamic stability in the directions independent of biological age.

Since evolution of the physiological state vector is not reducible to aging drift alone, biological age constitutes an essential, but not the only contribution to human mortality. Our work indicates that the bioage-independent component of the risk of death is significant and associated with extreme frailty; this is consistent with the conclusions of [40] where a frailty index showed superior performance compared to DNAm age in mortality predictions. Also, in an epigenome-wide association study [41], the reported DNAm signature of all-cause mortality, was found to comprise a component independent of the "epigenetic clock". We show that the physiological state vector fluctuations that are independent of bioage and the associated component of mortality risk are signatures of an organisms responses to stresses, diseases,

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and hazardous behaviors (e.g., smoking, Figure 4B). Our findings related to the impact of smoking agree with results obtained earlier in [35] where frailty index demonstrated a significant correlation with methylation sites associated with smoking. We also established that variations in the bioage-independent component of mortality can be induced by smoking early in life, but reversed if the individual quits smoking. Therefore, the biological age-dependent and -independent components of mortality are independent factors contributing to human lifespan determination.

In conclusion, this report demonstrates a way to quantify human physical activity time series and extract locomotor activity-based signatures of aging acceleration and increased mortality risk in association with diseases and hazardous lifestyles. A systematic study of aging and frailty in a large NHANES dataset revealed the dynamical origin of biological age and its relation to the characteristic increase in mortality with age, the Gompertz mortality law.

On a practical level, the results of this study lead us to propose using the hazard function, a property associated with all-causes mortality, as an ultimate measure of the health or wellness of an individual. This can then be decomposed into the biological age-associated component (which is determined by the accumulated effects of the individuals life history), and the potentially modifiable biological age-independent component. Both mortality components can be quantified from a single one-week long physical activity track collected by a consumer-grade wearable device.

Our findings highlight an opportunity for deployment of fully automated wellness intelligence systems capable of ambiently processing tracker information and providing dynamic feedback to the general public for improved engagement in health-promoting lifestyle modifications, disease interception, and clinical development of therapeutic interventions against the aging process.

IV. ACKNOWLEDGEMENT

This study was conducted using the UK Biobank Resource, application number 21988. We would like to thank G. Ivashkevich, I. Molodtsov, A. Tarkhov, V. Kogan from Gero LLC for extensive help in conducting the research.

P.O. Fedichev is a shareholder of Gero. T.V. Pyrkov, E. Getmantsev, B. Zhurov, K. Avchaciov, M. Pyatnitskiy, L. Menshikov, K. Khodova, and P.O. Fedichev are employees of Gero LLC. An international PCT patent application submitted by Gero LLC on the described methods and tools for non-invasively evaluating health is pending.

V. MATERIALS AND METHODS

A. NHANES dataset

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 $1min^{-1}$ 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 NHANES age data field is top coded at 85 years of age and we desired precise age information for our study.

To calculate a statistical descriptor of each participant's locomotor activity, we first converted activity counts into discrete states with bin edges b_k, $k = 1..K$. Activity level states $1...K-1$ 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 8 activity states with bin edges $b_k = e^k - 1, k = 1...7$. 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. 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 11839 samples (age, years: 35 ± 23, range $6-84$; women: 51%). For PCA and Survival analysis, the only samples used were those for participants aged 40 and older with known follow-up on survival/mortality outcome (age, years: 59 ± 12, range $40-84$; women: 50%). Once PCA loading vectors were identified, we plotted all NHANES samples' scores in Figure 2A, including those for which survival/mortality data were not available.

Transition matrices (TM) $T_{ij}, i = 1...8, j = 1...8$ 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). TM elements were calculated as $T_{ij} = N(j \to i)/N(j)$, where $N(j)$ is the number of minutes corresponding to state j and $N(j \to i)$ is the number of times the state j was immediately followed by state i (in the consecutive minute along the sample record). We next converted the TM from a discrete point map to continuous notation: $W_{ij} = T_{ij} - I$, where I is the identity matrix. W_{ij} is the proper TM for which the apparatus of the Markov chains can be used. We used this property to calculate Power

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spectral Densities (PSD) and eigenfrequencies (shown in Figure 1B) based on the assumption that the Markov chain model can be an approximation of observed activity records. We flattened 8×8 TM of each sample into a 64-dimensional descriptor vector for Principal Component Analysis (PCA) and Survival analysis. Additionally, we 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.

B. UKBB dataset

We accessed data from UK Biobank (UKBB) under the approved research project 21988 (formerly 9086). At the time the present study was conducted (2015-2017), locomotor activity data were available for 103710 UKBB participants. Physical activity was measured using Axivity AX3 tri-axial accelerometers worn on the wrist for 7 consecutive days. The data were recorded in the low-level format as continuous tracks of 3D acceleration values sampled at 100Hz. Some tracks indicated that hardware errors occurred during the monitoring period. Participants with more than 10 such hardware errors in their track were excluded from our analysis, leaving 102914 participants. To make it possible to apply the PCA and Survival analysis models established using NHANES data to the UKBB data, we downsampled the original UKBB tracks to $1min^{-1}$ (as used in NHANES). For this purpose, individual acceleration records were split into 1-minute slices, and for each slice, the natural logarithm of the sum over the power spectral density (PSD) of the signal within that slice was calculated. Each of these PSDs was calculated from the absolute values of acceleration using the Welch method with 512 points Hann window function and 50% window overlap.

The downsampled UKBB tracks represent the level of physical activity per minute but are quantitatively different from the NHANES activity counts. We used a quantile normalization procedure to re-scale the UKBB values to the range of discrete activity states of NHANES. We selected NHANES participants in the age range 45-75 and dropped 1/6 of participants with the lowest and highest average activities. The combined tracks from the remaining 2398 NHANES participants were used to calculate the occupancy fractions $p_k = N(k)/N$ for each NHANES activity state (here $N(k)$ is the number of times the state k was seen and N is the total number of minutes in all tracks). Then we randomly selected 5000 UKBB participants from the same age range and similarly dropped 1/6 of participants with the lowest and highest average activities; this resulted in selection of 3288 UKBB participants. Using the combined UKBB tracks from selected participants, we found UKBB bin edges b'_k such, that the occupancy fractions for the corresponding activity states, were equal to the occupancy fractions in NHANES. Note that such quantile normalization automatically accounts for shift, linear and monotonic non-linear scaling of values, and so the resulting UKBB activity states are roughly equivalent to the ones from NHANES. Once bin edges for UKBB were obtained, the downsampled UKBB tracks were processed exactly as described above for NHANES. TMs and corresponding descriptors were obtained for 95609 UKBB participants (age, years: average 62, median 63, range 43-79; women: 56%).

C. Survival analysis

We estimated hazard rate (i.e. mortality rate) for each participant using parametric Cox-Gompertz proportional hazard model in the form of maximal likelihood minimization adapted from [20]. The model estimates mortality rate in response to values of each sample explanatory covariates, the locomotor activity descriptors in our study. In a general form, Cox proportional hazards are uncertain up to an arbitrary baseline hazard function. Cox-Gompertz model, in contrast, explicitly accounts for Gompertz exponential increase of mortality with age and provides an estimation of Gompertz parameters for the studied dataset. We used NHANES survival and mortality follow-up data to train the model. The corresponding parametric log-likelihood was maximized using theano python package:

$$\log L\left(M_0, \Gamma, \beta\right) = \frac{1}{N}\sum_{i=1}^{N} \frac{M_0}{\Gamma} e^{(\beta, x_i)} \left(1 - e^{\Gamma \Delta t_i}\right) + \frac{1}{N}\sum_{i=1}^{N} \delta_i \left(\log M_0 + (\beta, x_i) + \Gamma \Delta t_i\right) - \lambda \beta^2$$

where M_0 and Γ are the initial mortality rate and mortality doubling time constant of Gompertz law, respectively; Δt_i is the time between sampling of locomotor activity and censored or death event (in years), x_i is the set of explanatory covariates (locomotor descriptors and gender label to account for gender-related differences in mortality) of i-th participant and β is the set of weighting coefficients corresponding to locomotor descriptors. To train the model we used the subset of participants for whom mortality/survival outcome was known. Covariates x_i were normalized to zero mean and unit variance across the said subset. Regularization parameter λ is introduced to account for overfitting. We screened λ in log-space range $10^{-5} - 10^{5}$ for stability of trained weighting coefficients β and selected $\lambda = 0.1$. We did not add chronological age explicitly to the model because our intention was to let the model attribute the age-related increase in mortality completely to locomotor biological age. The bioage-related and -independent components of

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

log Mortality were obtained by linear detrending of log Mortality on bioage.

D. Volcano plot

NHANES 2003-2006 study participants in the age range 40-70 y.o. were divided into groups for each relevant entry from NHANES Questionnaire and Laboratory data. For entries containing continuous data, participants were divided into three groups by percentiles: 0 to 13, 14 to 86 and 87 to 100. For entries with categorical data, participants were grouped by their categorical label. In addition, we allocated two marginal groups corresponding to the first two and the last two labels, if an entry had more than 3 categories. The means of log-mortality were calculated for each group and differences of means for all possible combinations of group pairs were evaluated for statistical significance ($p < 0.001$) by the Mann-Whitney U test. For each pair of groups, we assigned the group containing more participants as the control group, while the other group was assigned as hazardous. Therefore, the positive sign for $\Delta\log(\text{mortality})$ corresponds to the increased mortality for the hazardous group. The difference of $\Delta\log(\text{mortality})$ for the bioage-related component was referred to as aging acceleration, and p-values were transformed using the $-\log 10$ function. The significance level was adjusted for multiple testing with Bonferroni correction with the factor of $n = 10^4$ (total number of observed pairs rounded up to a power of 10).

All analyses were conducted using a set of in-house developed scripts in Python [`www.python.org`] and R [`www.r-project.org`].

[1] M. P. Antoch, M. Wrobel, K. K. Kuropatwinski, I. Gitlin, K. I. Leonova, I. Toshkov, A. S. Gleiberman, A. D. Hutson, O. B. Chernova, and A. V. Gudkov, Aging (Albany NY) **9**, 615 (2017).

[2] S. M. Jazwinski and S. Kim, Front Genet **8**, 64 (2017).

[3] B. L. Strehler and A. S. Mildvan, Science See Saiensu **132** (1960).

[4] P. Lamkin, "Wearable tech market to be worth $34 billion by 2020," `www.forbes.com/sites/paullamkin/2016/02/17/wearable-tech-market-to-be-worth-34-billion-by-2020/` (2016), accessed: 2017-08-14.

[5] X. Li, J. Dunn, D. Salins, G. Zhou, W. Zhou, S. M. S.-F. Rose, D. Perelman, E. Colbert, R. Runge, S. Rego, *et al.*, PLoS biology **15**, e2001402 (2017).

[6] B. M. Nes, C. R. Gutvik, C. J. Lavie, J. Nauman, and U. Wisløff, The American journal of medicine **130**, 328 (2017).

[7] B. Gompertz, Philosophical transactions of the Royal Society of London **115**, 513 (1825).

[8] D. J. Wilkinson, Nature Reviews Genetics **10**, 122 (2009).

[9] L. D. Landau, E. Lifshitz, and L. Pitaevskii, "Statistical physics, part i," (1980).

[10] T. Nakamura, T. Takumi, A. Takano, N. Aoyagi, K. Yoshiuchi, Z. R. Struzik, and Y. Yamamoto, PLoS One **3**, e2050 (2008).

[11] C. Gu, C. P. Coomans, K. Hu, F. A. Scheer, H. E. Stanley, and J. H. Meijer, Proceedings of the National Academy of Sciences **112**, 2320 (2015).

[12] K. Hu, E. J. Van Someren, S. A. Shea, and F. A. Scheer, Proceedings of the National Academy of Sciences **106**, 2490 (2009).

[13] T. Nakamura, K. Kiyono, K. Yoshiuchi, R. Nakahara, Z. R. Struzik, and Y. Yamamoto, Physical review letters **99**, 138103 (2007).

[14] P. Indic, P. Salvatore, C. Maggini, S. Ghidini, G. Ferraro, R. J. Baldessarini, and G. Murray, PloS one **6**, e20650 (2011).

[15] M. Ringnér, Nature biotechnology **26**, 303 (2008).

[16] R. S. Feldman, *Development across the life span* (Prentice Hall, 2003).

[17] L. A. Gavrilov and N. . S. Gavrilova, *The biology of life span: a quantitative approach.* (New York New York/Chur Switzerland Harwood Academic 1991., 1991).

[18] J. Blodgett, O. Theou, S. Kirkland, P. Andreou, and K. Rockwood, Archives of Gerontology and Geriatrics **60**, 464 (2015).

[19] D. R. Cox, in *Breakthroughs in statistics* (Springer, 1992) pp. 527–541.

[20] B. Efron, Journal of the American statistical Association **72**, 557 (1977).

[21] L. B. Levy G, *The Biostatistics of Aging: From Gompertzian Mortality to an Index of Aging-relatedness* (John Wiley & Sons, 2014).

[22] A. Mitnitski, J. Collerton, C. Martin-Ruiz, C. Jagger, T. von Zglinicki, K. Rockwood, and T. B. Kirkwood, BMC Med **13**, 161 (2015).

[23] S. Horvath, M. Gurven, M. E. Levine, B. C. Trumble, H. Kaplan, H. Allayee, B. R. Ritz, B. Chen, A. T. Lu, T. M. Rickabaugh, B. D. Jamieson, D. Sun, S. Li, W. Chen, L. Quintana-Murci, M. Fagny, M. S. Kobor, P. S. Tsao, A. P. Reiner, K. L. Edlefsen, D. Absher, and T. L. Assimes, Genome Biol. **17**, 171 (2016).

[24] D. H. Taylor Jr, V. Hasselblad, S. J. Henley, M. J. Thun, and F. A. Sloan, American journal of public health **92**, 990 (2002).

[25] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J. B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang, Mol. Cell **49**, 359 (2013).

[26] S. Horvath, Genome Biol. **14**, R115 (2013).

[27] J. Kriti, F. Vukovi, C. Menni, L. Klari, T. Keser, I. Beceheli, M. Pui-Bakovi, M. Novokmet, M. Mangino, K. Thaqi, P. Rudan, N. Novokmet, J. Sarac, S. Missoni, I. Koli, O. Polaek, I. Rudan, H. Campbell, C. Hayward, Y. Aulchenko, A. Valdes, J. F. Wilson, O. Gornik, D. Primorac, V. Zoldo, T. Spector, and G. Lauc, J. Gerontol. A Biol. Sci. Med. Sci. **69**, 779 (2014).

[28] M. E. Levine, J. Gerontol. A Biol. Sci. Med. Sci. **68**, 667 (2013).

[29] T. Odamaki, K. Kato, H. Sugahara, N. Hashikura, S. Takahashi, J. Z. Xiao, F. Abe, and R. Osawa, BMC

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Microbiol. **16**, 90 (2016).
[30] G. S. Baird, S. K. Nelson, T. R. Keeney, A. Stewart, S. Williams, S. Kraemer, E. R. Peskind, and T. J. Montine, Am. J. Pathol. **180**, 446 (2012).
[31] R. E. Marioni, S. Shah, A. F. McRae, B. H. Chen, E. Colicino, S. E. Harris, J. Gibson, A. K. Henders, P. Redmond, S. R. Cox, *et al.*, Genome biology **16**, 1 (2015).
[32] S. Horvath and A. J. Levine, J. Infect. Dis. **212**, 1563 (2015).
[33] S. Horvath, P. Garagnani, M. G. Bacalini, C. Pirazzini, S. Salvioli, D. Gentilini, A. M. D. Blasio, C. Giuliani, S. Tung, H. V. Vinters, and C. Franceschi, Aging Cell **14**, 491 (2015).
[34] S. Horvath, W. Erhart, M. Brosch, O. Ammerpohl, W. von Schönfels, M. Ahrens, N. Heits, J. T. Bell, P.-C. Tsai, T. D. Spector, *et al.*, Proceedings of the National Academy of Sciences **111**, 15538 (2014).
[35] X. Gao, Y. Zhang, K.-U. Saum, B. Schöttker, L. P. Breitling, and H. Brenner, Epigenetics (2016).
[36] D. Krotov, J. O. Dubuis, T. Gregor, and W. Bialek, Proceedings of the National Academy of Sciences **111**, 3683 (2014).
[37] D. Podolskiy, I. Molodtcov, A. Zenin, V. Kogan, L. Menshikov, V. Gladyshev, R. J. S. Reis, and P. Fedichev, arXiv preprint arXiv:1502.04307 (2015).
[38] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, Nature **461**, 53 (2009).
[39] L. Landau and E. Lifshitz, Course of Theoretical Physics (1981).
[40] S. Kim, L. Myers, J. Wyckoff, K. E. Cherry, and S. M. Jazwinski, GeroScience **39**, 83 (2017).
[41] Y. Zhang, R. Wilson, J. Heiss, L. P. Breitling, K.-U. Saum, B. Schöttker, B. Holleczek, M. Waldenberger, A. Peters, and H. Brenner, Nature Communications **8**, 14617 (2017).

Appendix A: Transition matrix theory.

Under the Markov chain model, the evolution of the probability $P_i(t)$ to find the system at state i for the system with N discrete states is governed by the master equation which in the linear mode can be written as

$$\frac{dP_i(t)}{dt} = \sum_{i=1}^{N} \left(k_{ij} P_j(t) - k_{ji} P_i(t)\right), \tag{A1}$$

where $k_{ij} \geq 0$ is the rate of transition from state j to state i. By introducing the transition matrix (TM) according to

$$W_{ij} = k_{ij} - \delta_{ij} \sum_{e=1}^{N} k_{ei}, \tag{A2}$$

we can rewrite Eq. A1 as

$$\dot{P}_i(t) = \sum_{j=1}^{N} W_{ij} P_j(t). \tag{A3}$$

Note from Eq. A2 and definition of k_{ij} we have $W_{ij} \geq 0$ for $i \neq j$, $W_{ii} \leq 0$ and

$$\sum_{i=1}^{N} W_{ij} = 0, \tag{A4}$$

from which it follows that the probabilty norm is preserved $\frac{d}{dt}\left(\sum P_i\right) = 0$, as it should be.

In the following analysis we will assume that the TM W is irreducible and has distict eignevalues. The reasoning for such asumptions will be provied later. Under this assumptions W can be diagonalized

$$W_{ij} = \sum_{k=1}^{N} \lambda_k A_{kj} B_{ki}, \tag{A5}$$

where A_k and B_k are left ($\sum_i W_{ij} A_{ki} = \lambda_k A_{kj}$) and right ($\sum_j W_{ij} B_{kj} = \lambda_k B_{ki}$) eigenvectors corresponding to eigenvalue λ_k. Note that the systems of left and right eigenvectors are the inverse for each other:

$$\sum_{k=1}^{N} A_{ki} B_{kj} = \delta_{ij} \text{ and } \sum_{i=1}^{N} A_{ki} B_{mi} = \delta_{km}. \tag{A6}$$

To solve Eq. A3 we introduce

$$Q_k(t) = \sum_{j=1}^{N} A_{kj} P_j(t), \tag{A7}$$

and using Eqs. A5 and A6 rewrite Eq. A3 as

$$\dot{Q}_k(t) = \lambda_k Q_k(t),$$

for which the solution is

$$Q_k(t) = Q_k(0)\, e^{\lambda_k t},$$

from which using Eqs. A6 and A7 we get

$$P_i(t) = \sum_{k=1}^{N} \sum_{j=1}^{N} A_{kj} B_{ki} e^{\lambda_k t} P_j^0 = \sum_{j=1}^{N} G_{ij}(t)\, P_j^0, \tag{A8}$$

$$G_{ij}(t) = \sum_{k=1}^{N} A_{kj} B_{ki} e^{\lambda_k t}, \tag{A9}$$

where $G_{ij}(t)$ is the probability $P(i,t|j,0)$ to find the system at state i at time t if the system originally was at state j at time 0 and P^0 is the initial distribution.

The assumption that W has distinct eignavalues together with Eq. A4 imply that W has exactly one zero eigenvalue. Since the order of eigenvalues is arbitrary, we can state that

$$\begin{cases} \lambda_1 = 0, \\ \operatorname{Re} \lambda_i < 0, \quad 1 < i \leq N, \end{cases} \tag{A10}$$

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where the later unequality follows from W being a TM.

Indeed, W is real-valued, therefore for any eigenvalue λ_i and corresponding left eigenvector A_i we have

$$\lambda_i A_{ij} = \sum_{k=1}^{N} W_{kj} A_{ik},$$

and

$$\lambda_i^* A_{ij}^* = \sum_{k=1}^{N} W_{kj} A_{ik}^*,$$

where * is complex conjugate. After multiplying the first equation by A_{ij}^*, the second by A_{ij} and summing we get

$$\operatorname{Re}\lambda_i \cdot |A_{ij}|^2 = \sum_{k\neq j} W_{kj}\left(A_{ik}A_{ij}^* + A_{ik}^* A_{ij}\right) + 2W_{jj}\left|A_{ij}\right|^2.$$

Representing all A_{ij} in exponential form $A_{ij} = \rho_{ij}\exp(\iota\phi_{ij})$, dividing by $|A_{ij}|^2$ and replacing W_{jj} using Eq. A4 we get

$$\operatorname{Re}\lambda_i = 2\sum_{k\neq j} W_{kj}\left[\frac{\rho_{ik}}{\rho_{ij}}\cos(\phi_{ik}-\phi_{ij}) - 1\right]. \quad \text{(A11)}$$

This equation holds for all i and j. For a given i let us choose a particular j such that $\rho_{ik} \leq \rho_{ij}$. Since all ρ_{ik} and W_{kj} for $k \neq j$ are non-negative by definition, the Eq. A11 becomes $\operatorname{Re}\lambda_i \leq 0$.

According to Eq. A3, any equlibrium state is the right eigenvector corresponding to the zero eigenvalue. Since W has only one such eigenvector (up to scaling), we have a unique equlibrium distribution given by

$$P_i^{eq} = B_{1i} / \sum_{j=1}^{N} B_{1j} \quad \text{(A12)}$$

The eigensystem has several interesting properties. From Eqs. A9 and A10 we get $G_{ij}(+\infty) = A_{1j}B_{1i}$ and the distribution at $t = +\infty$ is

$$P_i^{\infty} = B_{1i}\sum_{j=1}^{N} A_{1j}P_j^0. \quad \text{(A13)}$$

For any initial distribution P^0 the corresponding P^∞ is an equlibrium state:

$$\sum_{j=1}^{N} G_{ij}(t)\, P_j^{\infty} = \sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{m=1}^{N} A_{kj}B_{ki}e^{\lambda_k t}B_{1j}A_{1m}P_m^0 =$$
$$= \sum_{m=1}^{N} B_{1i}A_{1m}P_m^0 = P_i^{\infty},$$

and since equlibrium is unique $P_i^\infty = P_i^{eq}$ for any P^0. From this and Eq. A13 we have

$$A_{1i} = const = 1/\sum_{j=1}^{N} B_{1j}. \quad \text{(A14)}$$

Using Eq. A4, for the right eigenvectors B_k we get

$$\lambda_k \sum_{i=1}^{N} B_{ki} = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}B_{kj} = 0,$$

and therefor

$$\sum_{i=1}^{N} B_{ki} = 0, \text{ for } \lambda_k \neq 0. \quad \text{(A15)}$$

Let as consider a discrete real-valued stochastic process $x(t)$ having value x_i when system happens to be in state i. According to the Wiener–Khinchin theorem, the power spectral density $S_x(\omega)$ for the $x(t)$ is the Fourier transform of the autocorrelator

$$R_{xx}(\tau) = \lim_{T\to+\infty}\frac{1}{T}\int_0^T \mathrm{E}\left[x(t+\tau)\,x(t)\right]dt. \quad \text{(A16)}$$

Using the fact that $R_{xx}(\tau)$ is an even real-valued function we obtain

$$S(\omega) = 2\int_0^{+\infty} R_{xx}(\tau)\cos(\omega\tau)\,d\tau. \quad \text{(A17)}$$

Here we follow the common physical convention that the total power of the signal is given by $\int_{-\infty}^{+\infty} S_x(\omega)\frac{d\omega}{2\pi}$.

Expanding the Eq. A16 we get

$$R_{xx}(\tau) = \lim_{T\to+\infty}\frac{1}{T}\int_0^T \sum_{i=1}^{N}\sum_{j=1}^{N} x_i P(i,t+\tau|j,t)\,x_j P(j,t)\,dt, \quad \text{(A18)}$$

where $P(i,t+\tau|j,t)$ is the probaility to find the system at state j at time $t+\tau$ if the system was at state j at time t and $P(j,t)$ is the probability to find the system at state j at time t, with the evolution of the system starting from some state P^0. From the definitions we have $P(i,t+\tau|j,t) = G_{ij}(\tau)$ for $\tau \geq 0$ and $P(j,t) = P_j(t)$. Using this and Eqs. A8 and A9 rewrite Eq.A18 as

$$R_{xx}(\tau) = \lim_{T\to+\infty}\frac{1}{T}\int_0^T \sum_{i,j,k,m,n} x_i A_{kj}B_{ki}e^{\lambda_k\tau}x_j A_{mn}B_{mj}e^{\lambda_m t}P_n^0 dt,$$

where $\tau > 0$ and the summation is done for each index from 1 to N. By rearranging and using Eq. A10 we get

$$R_{xx}(\tau) = \sum_{i,j,k,n} x_i A_{kj}B_{ki}e^{\lambda_k\tau}x_j A_{1n}B_{1j}P_n^0,$$

from which using Eq. A12 we finally obtain

$$R_{xx}(\tau) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} x_i x_j A_{kj}B_{ki}P_j^{eq}e^{\lambda_k\tau}, \text{ for } \tau \geq 0. \quad \text{(A19)}$$

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Note that $R_{xx}(\tau)$ is not dependent on the initial distribution P^0, as it is expected for the system with equlibrium state. The integration of Eq. A19 using Eq. A17 is straightforward, and we get

$$S_x(\omega) = -2\sum_{k=2}^{N}\left\{\frac{\lambda_k}{\lambda_k^2+\omega^2}\sum_{i=1}^{N}x_iA_{ki}P_i^{eq}\sum_{j=1}^{N}x_jB_{kj}\right\}. \tag{A20}$$

The Eq. A20 is valid for any irreducible diagonalizable TM W. In particular, some λ_k may be complex. However, for the real-valued matrix W complex eigenvalues and corresponding eigenvectors always comes in complex conjugate pairs, which, together with Eq.A10, imply that $S_x(\omega)$ is always real positive, as any PSD should be.

Due to time symmentry of the fundamental physical laws, for the systems in thermodynamic equilibrium the detailed balance assumption is hold:

$$W_{ij}P_j^{eq} = W_{ji}P_i^{eq}. \tag{A21}$$

Biological organisms as a whole are not systems in thermodynamic equlibrium and the description of the motion using Markov chain model is only a rough approximation, so there are no *a priori* reasons to assume the detailed balance. However, experimetally the correlation between $W_{ij}P_j^{eq}$ and $W_{ji}P_i^{eq}$ is good, so it is interesting to see how $S_x(\omega)$ looks under detailed balance assumption.

First we introduce a derived matrix

$$\tilde{w}_{ij} = W_{ij}\sqrt{P_j^{eq}/P_i^{eq}}. \tag{A22}$$

With Eq. A21 hold, $\tilde{w}$ is symmetric and therefore can be eigendecomposed into

$$\tilde{w}_{ij} = \sum_{k=1}^{N}\lambda_k\mu_{ki}\mu_{kj}, \tag{A23}$$

where all eignvalues λ_k are real and eigenvectors μ_k are orthonormal:

$$\sum_{k=1}^{N}\mu_{ki}\mu_{kj} = \delta_{ij} \text{ and } \sum_{i=1}^{N}\mu_{ki}\mu_{mi} = \delta_{km}. \tag{A24}$$

From Eqs. A22 and A23 we get

$$W_{ij} = \sqrt{\frac{P_i^{eq}}{P_j^{eq}}}\sum_{k=1}^{N}\lambda_k\mu_{ki}\mu_{kj} = \sum_{k=1}^{N}\lambda_k\tilde{A}_{kj}\tilde{B}_{ki}, \tag{A25}$$

where

$$\tilde{A}_{ki} = \mu_{ki}/\sqrt{P_i^{eq}} \text{ and } \tilde{B}_{ki} = \mu_{ki}\cdot\sqrt{P_i^{eq}}, \tag{A26}$$

from which using Eq. A24 follows

$$\sum_{k=1}^{N}\tilde{A}_{ki}\tilde{B}_{kj} = \delta_{ij} \text{ and } \sum_{i=1}^{N}\tilde{A}_{ki}\tilde{B}_{mi} = \delta_{km},$$

which imply that Eq. A25 is an eigendecomposition of W as in Eq. A5, so we can use Eq. A20, which becomes

$$S_x(\omega) = -2\sum_{k=2}^{N}\frac{\lambda_k}{\lambda_k^2+\omega^2}\left(\sum_{i=1}^{N}x_i\tilde{B}_{ki}\right)^2. \tag{A27}$$

Here we used $\tilde{A}_{ki} = \tilde{B}_{ki}/P_i^{eq}$, obtained from Eq. A26, to express $S_x(\omega)$ via right eigenventors $\tilde{B}_k$ alone. Each of the right eigenvectors B_k is defined up to a multiplication constant, however the scaling is fixed for $\tilde{B}_k$: from Eqs. A24 and A26 we have

$$\sum_{i=1}^{N}\frac{\tilde{B}_{ki}\tilde{B}_{kj}}{\sqrt{P_i^{eq}P_j^{eq}}} = \delta_{ij}, \tag{A28}$$

from which we can find a proper scaling for an arbitrary right eigenvector B_k:

$$\tilde{B}_{ki} = B_{ki}\left(\sum_{j=1}^{N}\frac{(B_{kj})^2}{P_j^{eq}}\right)^{-\frac{1}{2}}. \tag{A29}$$

The $S_x(\omega)$ can be calculated under detailed balance assumption as follows: calculate the right eigensystem for W, scale the found eigenvectors B_k using Eq. rA29 and finally calculate $S_x(\omega)$ using Eq. A27. The same procedure can be applied when the detailed balance assumption holds only approximately, as long as we drop the imaginary part of the found eigenvalues and right eigenvectors. Note that even when all eigenvalues are real, the Eq. A27 is not equivalent to Eq. A20 without the detailed balance assumption. In particular, scaling according to Eq. A29 is not enought for Eq. A28 to hold, which is required for Eq. A27 to be precise.

Appendix B: Appendix figures

bioRxiv preprint doi: https://doi.org/10.1101/186569; this version posted September 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

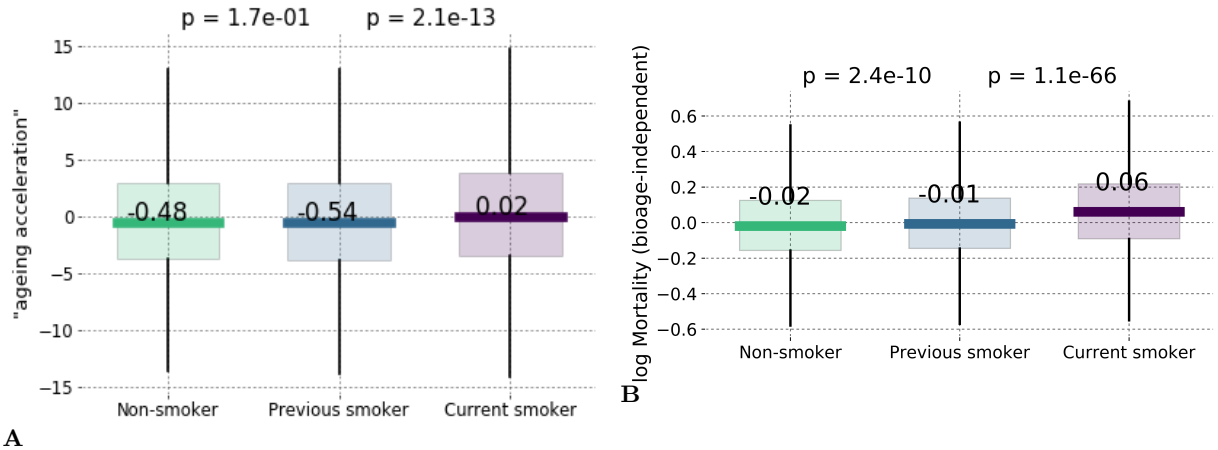

FIG. 6: **A, B.** Boxplots show distribution of hazard ratio calculated using locomotor activity data from UK Biobank study for previous and current smoker groups along the bioage-related and bioage-detrended hazard components in age range 40-70 y.o.