Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age

Zichen Wang[a], Li Li[b], Benjamin S. Glicksberg[b], Ariel Israel[c], Joel T. Dudley[b], Avi Ma'ayan[a,*]

[a] Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA
[b] Department of Genetics and Genomic Sciences, Institute of Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, New York, NY 10029, USA
[c] Department of Family Medicine, Clalit Health Services, Jerusalem 90258, Israel

ARTICLE INFO

ABSTRACT

Determining the discrepancy between chronological and physiological age of patients is central to preventative and personalized care. Electronic medical records (EMR) provide rich information about the patient physiological state, but it is unclear whether such information can be predictive of chronological age. Here we present a deep learning model that uses vital signs and lab tests contained within the EMR of Mount Sinai Health System (MSHS) to predict chronological age. The model is trained on 377,686 EMR from patients of ages 18–85 years old. The discrepancy between the predicted and real chronological age is then used as a proxy to estimate physiological age. Overall, the model can predict the chronological age of patients with a standard deviation error of ~7 years. The ages of the youngest and oldest patients were more accurately predicted, while patients of ages ranging between 40 and 60 years were the least accurately predicted. Patients with the largest discrepancy between their physiological and chronological age were further inspected. The patients predicted to be significantly older than their chronological age have higher systolic blood pressure, higher cholesterol, damaged liver, and anemia. In contrast, patients predicted to be younger than their chronological age have lower blood pressure and shorter stature among other indicators; both groups display lower weight than the population average. Using information from ~10,000 patients from the entire cohort who have been also profiled with SNP arrays, genome-wide association study (GWAS) uncovers several novel genetic variants associated with aging. In particular, significant variants were mapped to genes known to be associated with inflammation, hypertension, lipid metabolism, height, and increased lifespan in mice. Several genes with missense mutations were identified as novel candidate aging genes. In conclusion, we demonstrate how EMR data can be used to assess overall health via a scale that is based on deviation from the patient's predicted chronological age.

1. Introduction

While there is strong correlation between physiological age and chronological age, it is well established that these can be different in individuals [1]. Physiological age is a useful metric because it reflects the general health status of a patient; however, estimating physiological age is not trivial [2]. Multiple efforts have been made to estimate physiological age by combining physiological indicators such as physical activity level, mental health, nutrition, and medications [1–8]. The aging process is characterized by a variety of physiological and molecular changes at the cell and tissue levels [9]. At the tissue level, anemia, osteoporosis, and decrease in muscle fiber are induced due to the exhaustion of the somatic stem cell pool. At the cellular level, among other mechanisms, aging induces the loss of proteostasis, mitochondrial dysfunction, and altered intercellular communication that induces inflammation. Genetic variations and changes in gene expression have been studied in the context of aging. Genome-wide association studies (GWAS) and gene knockout studies have uncovered genetic variants and genes that are associated with extended lifespan in humans and model organisms, such as yeast, fly, worm, and mice. Several manually curated resources list genes associated with aging; for example, GenAge [10] and LongevityMap [11]. In addition, gene expression analysis of human tissues across individuals of different age, generated by the GTEx project [12], was utilized for identifying age-associated genes in different tissues [13]. Another study examined the correlation between the transcription level of genes, and chronological age, to determine a

---

"transcriptomic age" [14]. Similarly, it was demonstrated that DNA methylation is highly predictive of chronological age [15].

At the same time, data from electronic medical records (EMR), including disease diagnoses and biomedical trait values, can be used for characterizing the human phenotype. However, associating human phenotypes from EMR data is challenging because of data incompleteness and data inaccuracy [16]. To overcome some of these challenges, the Electronic MEdical Records and GEnomics (eMERGE) network [17] has developed and validated many EMR-phenotyping algorithms stored in the Phenotype Knowledgebase (PheKB) [18]. Many of these algorithms are rule-based. These algorithms operate on diverse types of EMR data, including diagnoses, procedures, medications, vital signs, laboratory tests, as well as clinical notes processed by natural language processing (NLP) algorithms [19]. Recently, a growing number of statistical methods, employing supervised and unsupervised machine learning approaches, have been proposed for various EMR-phenotyping tasks such as identifying cohorts with certain diseases and predicting patient survival [20]. Computationally, traditional linear statistical models do not scale well to very large datasets with missing values such as EMR data which contain billions of data points. With advent algorithms that employ deep learning, through non-linear transformations, it is now possible to better handle data incompleteness, inaccuracy, and scalability to learn from EMR. For instance, DeepPatient uses denoising Autoencoder (dAE) to perform unsupervised feature extraction from EMR [21]. Most relevant to the work presented here is a study that implemented a Least-Angle Regression (LARS) model to predict the age of adolescents using 39 biomarkers from a blood test [5]. The model identified alkaline phosphatase and creatinine levels as predictive biomarkers for both male and female age; hematocrit and mean-cell-volume levels were found to be predictive markers for males, whereas total serum globulin was predictive for females.

Here we present a predictive model of chronological age utilizing deep learning applied to the Mount Sinai Health System (MSHS) EMR data. Physiological measurements, including vital signs and lab tests from the EMR, were used as features to train the model. The predictive model performs well to predict the age of most patients, but for some patients the predictor makes a significantly large error. These outliers define two unique cohorts: those patients that display significant discrepancy between their predicted age and their chronological age, either predicted to be much younger than their chronological age, or much older. By investigating the unique features of these two cohorts, we confirmed known, and discovered novel, clinical characteristics and medicines associated with aging. Furthermore, we performed GWAS analysis on these cohorts comparing them to the rest of the population to identify genetic variants that correlate with both accelerated and delayed aging phenotypes.

2. Data and methods

2.1. EMR data preprocessing

The MSHS-EMR data used for this project contain over 4 million de-identified patient records spanning from 1980 to March 2015, covering 85 unique vital signs and 2968 unique lab test variables. Age in years was calculated for all patients based on year of birth. Rare vital signs and lab test variables that appeared in less than 5 EMR were removed, resulting in 30 vital signs and 1891 lab tests for subsequent analyses. We only included patients with at least 10 visits and excluded patients with no vital signs or lab tests, resulting in 456,748 patients. Further filtering retaining only adults, ages 18–100 years, resulted in a final set of 385,918 individual patients. The median was used to aggregate vital signs and lab tests from multiple visits. Chronological age at the time of the latest clinical visit was used for age prediction for each patient. Missing data across all feature set variables were first filled with the population average, and then z-score transformed.

2.2. Predictive model for aging

We implemented an artificial neural network (ANN) regression model that implements a feed-forward neural network architecture consisting of three hidden layers, each containing 1000 sigmoid activation units. There was a 40% dropout probability between each layer. During training of the ANN, weights were initialized using the method developed by Glorot and Bengio [22]. The model was trained for 20 epochs with mini-batch size of 128 using Adam as an optimizer [23]. Training samples were shuffled before each epoch. 10-fold cross-validation (CV) was applied to train and evaluate the performance of the model. Mean absolute error (MAE) of the discrepancy between the predicted age and the real age in years was used to evaluate performance. We implemented the ANN model using Keras [24] with Theano [25].

2.3. Identifying cohorts with significant discrepancy between predicted and chronological age

To identify patients with significantly older or younger predicted age compared with their chronological age, we first filtered out patients older than 85, leading to 377,686 patients with ages between 18 and 85. Then, Z-test was performed to quantify the error between the predicted age and chronological age, with the null hypothesis that the error is 0. Patients that are significantly older (group O) or younger (group Y) than their chorological age are called if the error is significantly smaller or larger than 0 with a p-value < 0.05.

2.4. EMR-wide association analyses

To perform the EMR-wide association analyses, rare features from the patients' diagnoses, includingprocedures, medicines, lab tests, and vital signs, characterizing fewer than 500 patients were excluded. Patients were associated with diagnoses or medicines from their EMR regardless of the time of encounter. To standardize nomenclature, we used the International Classification of Disease (ICD)-9 code as the diagnostic and procedure representation, and generic names for medicines. The Fisher exact test was performed for evaluating each discrete feature, including diagnosis, procedure and medicine, to test whether those are enriched within the younger/older groups versus the entire patient population. For continuous variables, such as those from lab tests and vital signs, the Wilcoxon rank sum test was applied. Multiple hypothesis testing correction was performed with the Benjamini-Hochberg procedure [26].

2.5. Genome-wide association study (GWAS)

The Charles Bronfman Institute of Personalized Medicine Mount Sinai BioMe biobank is comprised of over 32,000 patients who also have records in the MSHS-EMR. For this study, we included a subset of 9897 individuals. These patients have been genotyped using the Illumina Human Omni Express Exome Bead-8 BeadChip v1.1, which has 866,864 markers per sample. The analysis included first performing phasing of the genotype data using SHAPEIT2 (r644) [27] and imputing using IMPUTE2 v2.3 [28] with chunk size of 5 Mb. For the imputation reference panel, the 1000 Genomes Project Phase 3 data was used. Next, quality control (QC) steps were applied to filter the genetic data for subsequent analyses. Variants that had low imputation scores were removed, specifically any sites with INFO scores < 0.3 from IMPUTE2. Then filtered variants that were rare, specifically minor allele frequency (MAF) < 0.01 were removed. To determine the effect of genotype on the aging phenotype of interest, an additive, linear mixed model GWAS was performed with genotype set as either 0, 1, or 2 referring to the number of variant alleles using the Efficient Mixed-Model Association eXpedited software (EMMAX; beta version, March 7, 2010) [29]. Due to high levels of cryptic relatedness within the population, a kinship

matrix was incorporated into the mixed model. Specifically, the procedures outlined by Kang et al. [29] were followed to calculate an n x n matrix of pairwise genetic relatedness. The entries in the matrix were corrected for age and sex. Due to the high racial diversity of the cohort, principal component analysis (PCA) was utilized for dimensionality reduction of the genetic data to obtain a metric for genetic ancestry to include as covariates in the model. Several QC steps were performed on the genotype data before generating principal components (PC): rare variants (MAF < 0.01), thinned linkage disequilibrium (r2 of 0.3), as well as human leukocyte antigen and lactase genetic regions were filtered out. The data was further cleaned for site and individual level missingness (threshold of 95%). PCA was performed with these cleaned data by using the EIGENSOFT smartpca tool [30,31], training the space on the individuals with European ancestry from Utah (CEU); Yoruba in Ibidan, Nigeria (YRI); and Han Chinese in Beijing, China (CHB), which were used as European, African, and East Asian ancestry reference panels from the 1000 Genomes Project, respectively.

2.6. Enrichment analyses of SNPs

SNPs with suggestive significant associations (p-value $< 5e^{-6}$) were annotated to genes using the R package VariantAnnotation [32], with intergenic SNPs excluded. The direction of effect for the SNPs that were enriched in either population was determined by the sign of the BETA values. Enrichment analysis was performed with Enrichr [33,34]. MetaXcan [35] was also used to analyze the GWAS results using the transcriptome prediction models generated from the GTEx resource [12].

3. Results

3.1. Descriptive statistics of physiological variables and their relation to chronological age

The MSHS patient cohort analyzed for this study contains ~60% women and ~40% men, with 41% Caucasians, 16.9% African Americans, 5.4% Hispanics, 3.9% Asians, and 19.5% unclassified (Table 1, first column). We first performed correlation analysis to identify vital signs and lab tests that are correlated with chronological

Table 1
Summary statistics for demographics of the cohorts in this study. N: Normal; Y: Young; O: Old.

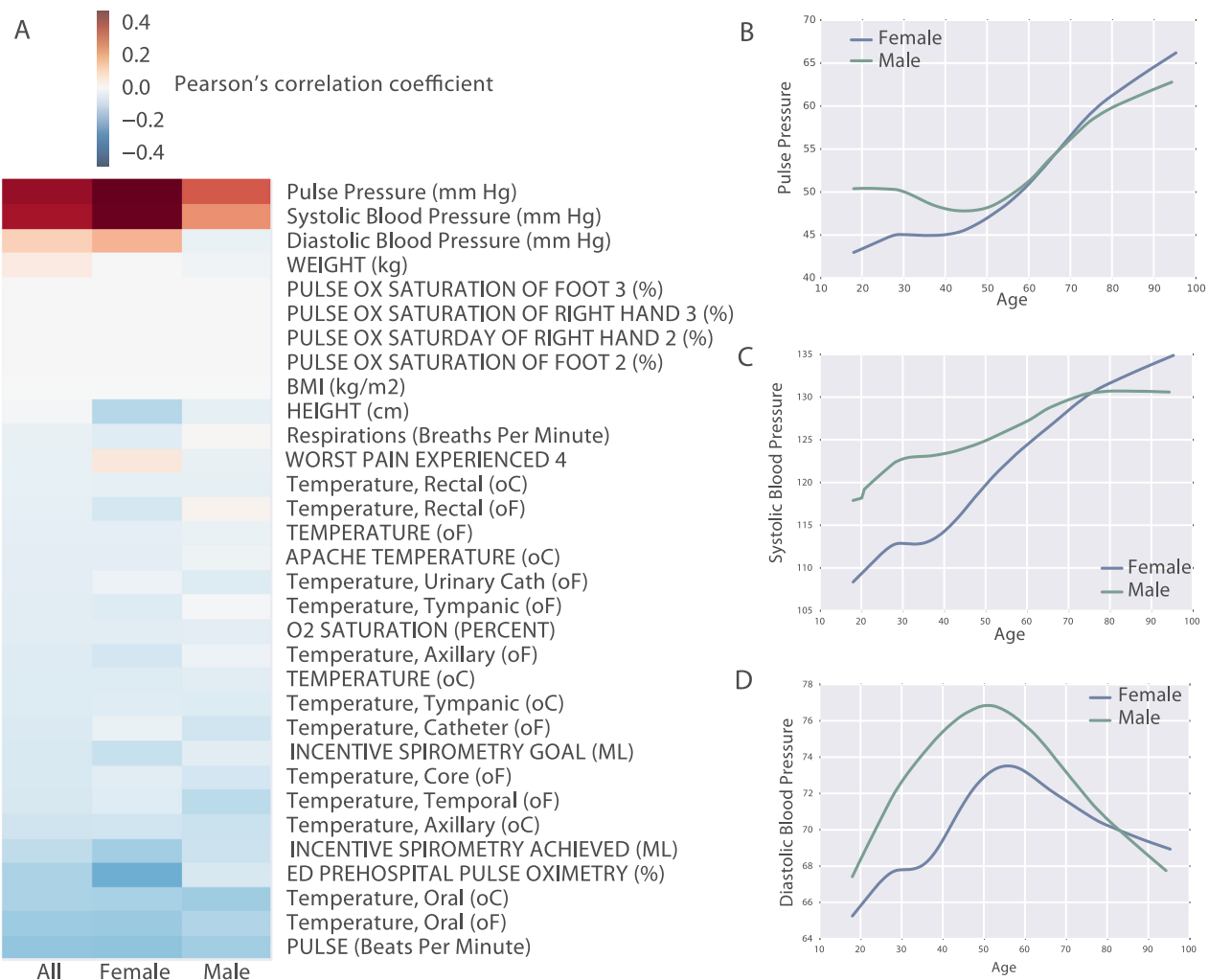| Variable	Cohorts			
	All	N	Y	O
Total	385,918	363,235	10,817	11,866
Chronological age				
25% quartile	35.9	36.4	60.0	24.3
Median	52.4	52.5	69.0	29.9
75% quartile	66.3	66.2	78.2	37.5
Error (chronological - predicted age)				
25% quartile	−5.7	−5.2	21.9	−29.2
Median	0.0	0.1	24.5	−24.7
75% quartile	5.8	5.4	29.1	−22.0
Gender				
Male	40.4%	40.0%	39.6%	53.1%
Female	59.6%	60.0%	60.4%	46.9%
Race				
Caucasian (White)	41.0%	40.9%	48.2%	38.5%
African American (Black)	16.9%	17.0%	13.7%	17.2%
Asian	3.9%	3.9%	3.1%	3.9%
Other	19.2%	19.4%	15.0%	19.1%
Hispanic/Latino	5.4%	5.4%	4.7%	6.3%
Native American	0.2%	0.2%	0.3%	0.2%
Pacific Islander	0.1%	0.1%	0.1%	0.1%

age across the entire cohort. Such correlation analysis was applied to the entire cohort as well as to gender-specific groups. Among all the vital signs, pulse pressure, which is the difference between systolic blood pressure and diastolic blood pressure, and systolic blood pressure, show the strongest positive correlation with chronological age for both genders (Fig. 1A). Interestingly, we also observed positive correlation between diastolic blood pressure in males but not in females (Fig. 1A). In order to more accurately interpret those correlations, we plotted the trends of blood pressure over chronological age across the entire cohort using the locally weighted scatterplot smoothing (LOWESS) algorithm [36]. We find that both pulse pressure and systolic blood pressure almost monotonically increase with age in both genders, whereas diastolic blood pressure appears to increase with age, similarly to systolic blood pressure, but begins to decrease around age 50 (Fig. 1B–D). Blood pressure generally increases with age due to increased arterial wall stiffening. However, the decrease in diastolic blood pressure around age 50 may be explained by the effect of blood pressure medications. It is well known that systolic blood pressure is less affected by antihypertensive medications. The correlation analysis also identifies lab tests that are highly positively correlated with age, these include urea nitrogen, glucose, hemoglobin A1C, PROTIME/INR; whereas lab tests that are most negatively correlated with age are glomerular filtration rate (GFR) estimate, albumin, total protein, red blood cell count and hematocrit (Fig. 2A). Urea nitrogen and GFR estimate are both measures of renal function (Fig. 2B–C). Glucose and hemoglobin A1C (Fig. 2D–E) are both markers of glucose intolerance and diabetes mellitus, and their positive correlation with age suggests gradual increase in glucose intolerance and higher risk to develop diabetes with age. PROTIME/INR are tests measuring blood clotting ability; their increase with age (Fig. 2F) is likely due to the use of anticoagulation medications to treat diseases such as stroke and atrial fibrillation. Hematocrit and red blood cell count both reflect the level of red blood cells in the blood; their decrease with age (Fig. 2G) is likely due decrease in oxygen demand in males as they age. Interestingly, in females, hematocrit is much lower, and it decreases until around age 40, most probably because of blood dilution during pregnancy. Albumin, the main protein component of the blood, and total protein, decrease in levels with age (Fig. 2H–I) is likely due to a decrease in nutritional status in older patients.

To obtain a global view of the correlations between physiological measurements and chronological age, we performed unsupervised hierarchical clustering on the LOWESS smoothened trends of the most common physiological measurements across all patients. This clustering method uncovers different modes of trends that cluster different physiological measurements with similar trends (Fig. 3). Specifically, we observe a cluster of nutritional status markers, which includes: total protein, albumin, calcium, total iron binding capacity (TIBC), and platelets. These nutritional markers generally decrease with age. Another cluster of variables has a bell shape curve that generally increase with age before 50–60, and then decrease with age after that inflection point. This cluster contains variables that are related to body mass index (BMI), including weight, height, diastolic blood pressure, cholesterol, low density/high density lipoprotein (LDL/HDL) ratio, triglycerides, and glutamyl transferanse (gamma GT). Higher values for these measurements are known to increase risk for cardiovascular disease. We speculate that their decrease after 50–60 could be the result of decreased BMI, which is due to survival bias which stems from premature mortality of obese patients [37]. There is also a cluster of variables that monotonically increase with age. These include: mean corpuscular volume, PROTIME/INR, urea nitrogen, creatinine, pulse pressure, ferritin, glucose, hemoglobin A1C, and systolic blood pressure. Variables in this cluster represent diverse mechanisms such as decreased renal function, attributed to urea nitrogen, and creatinine; increased glucose intolerance, attributed to increase in glucose and hemoglobin A1C levels; and increased inflammation, which is reflected in an increase in ferritin. We also examined the trends of all

**Fig. 2.** Exploration of lab tests and chronological age. (A) Pearson's correlation coefficients between individual lab tests and chronological age across all, female, and male patients in the three columns. Lab tests are sorted by descending order by the correlation coefficients with all patients. (B–I). The smoothened trends of representative lab tests across chronological age from female (blue) and male (green) patients. The age range of patients used to produce these figures is 18–100 years. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the largest error between their real chronological age and the age predicted by the model. Of these, 11,712 were placed in a group of predicted age significantly older than their real age (group O), and 10,611 were placed in a group of predicted age significantly younger than their real age (group Y) (Fig. S6). To rule out the possibility that patients from those two groups are merely due to poor prediction performance of the regression model, we performed the Hotelling's T-squared test, a multivariate generalization of Welch's *t*-test, to compare whether the physiological measurements for patients in group O and group Y are different from those patients without significant differences between their predicted age and their chronological age (group N) (Fig. S7). As expected, we found that members of group O are similar to older members of group N, and members of group Y are similar to younger members of N (Fig. S7). Interestingly, group O has significantly fewer women (p-value = $6.9e^{-175}$), whereas members of group Y have significantly more Caucasians (p-value = $6.4e^{-52}$) and fewer African Americans (p-value = $2.0e^{-19}$) (Table 1). It was previously suggested that women age slower than men [38], whereas some of these other

differences might be explained by the demographics associated with economic disparity within the Mount Sinai's patient cohort. The majority cohort population of the MSHS data that was analyzed originates from three different New York City neighborhoods with sharp demographic differences. The cohort contains patients from the Upper East Side which has a majority of Caucasian population, East Harlem which has a majority of Hispanic population, and Central Harlem with a majority of African American population. These unique diverse populations may confound the results.

3.4. Characterization of groups O and Y using EMR-wide association study

Next, we aimed to characterize the O and Y groups by comparing them with group N. We performed Wilcoxon rank sum tests to compare vital signs (Table S1) and lab tests (Table S2) between the different groups of patients. In Table S1 we only included frequent vital signs that are profiled in at least 10% of the patients. We also performed Fisher exact tests to identify diagnosis codes and procedures (Table S3),
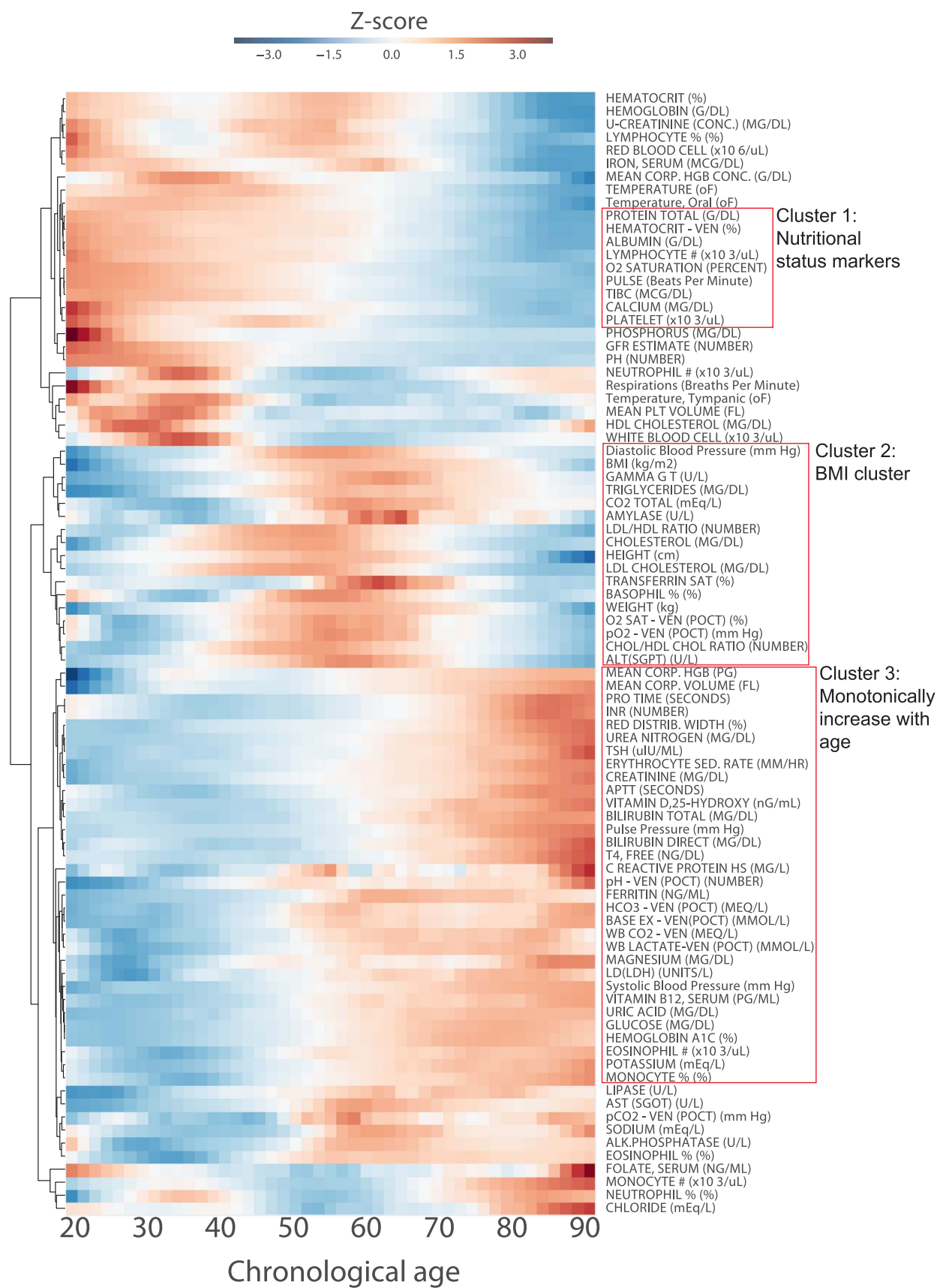
Fig. 3. Hierarchical clustering of physiological measurements across chronological age. The trends of physiological measurements across chronological age for all the 385,918 adult patients were derived from applying the locally weighted scatterplot smoothing (LOWESS) algorithm. Prior to performing hierarchical clustering, the trends for all measurements were z-score transformed. Hierarchical clustering was performed using Euclidean distance and average as the linkage function. Physiological measurements with similar trends across age are clustered together. The age range of patients used to produce the figure is 18–100 years.
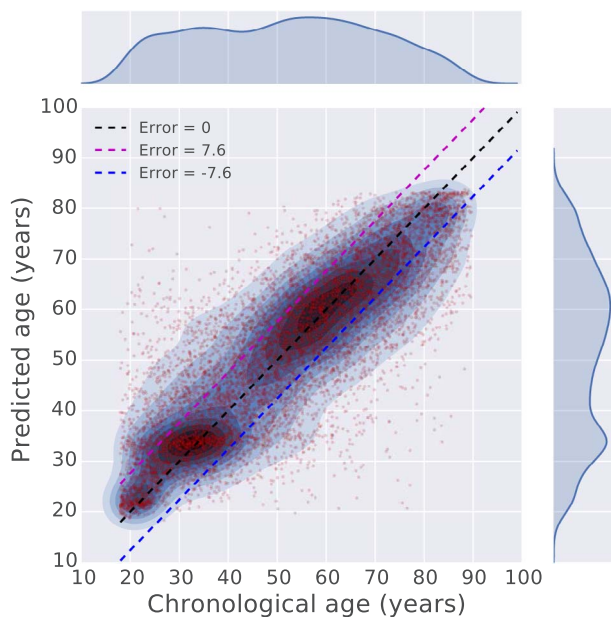
Fig. 4. Comparison between chronological age and predicted age. The chronological age and predicted age of 10,000 randomly sampled patients are plotted against each other in a scatter plot. To show the densities of the data points, kernel density estimation (KDE) is applied to overlay contours on top of the scatter plot. The figure also visualizes the outliers where there is a large discrepancy between the predicted and chronological age. The age range of patients used to produce the figure is 18–100 years.

as well as prescribed medicines (Table S4) that are enriched in groups O and Y compared to N.

The significant findings with interpretations for patients in group O are tabulated below:

- **Increased prevalence of hypertension and cardiovascular disorders:** patients of both genders in this group have significantly higher systolic blood pressure and pulse pressure. Consistently, they are disproportionally diagnosed with hypertension. The enrichment of diagnoses such as internal hemorrhoids, lumbago, which is a painful condition affecting the lower spine, and sciatica, which is pain along the lower back projected to the legs, also implies a sedentary lifestyle, which is a known risk factor for hypertension. The observed high levels of PROTIME/INR/APTT lab tests are consistent with the prescription of the anticoagulant drug alteplase, as well as the diagnosis with hyperlipidemia and carotid artery stenosis, which indicate that these patients are likely prescribed preventive treatment for cardiovascular diseases such as myocardial infarction and atrial fibrillation. The top enriched prescribed drugs also include cholesterol lowering drugs including ezetimibe and pravastatin. Interestingly, the levels of cholesterol and triglycerides in group O are lower compared to group N, this suggests that these are effects of cholesterol lowering medicines. Another possible explanation is that cholesterol levels are driven by BMI (Fig. 3), and patients in group O have significantly lower BMI compared to group N. Patients in group O are also more likely to be prescribed sildenafil, vasodilator and tadalafil, all of which are indicated for erectile dysfunction and high blood pressure in the lungs.
- **Increased chronic inflammation:** the decrease of lymphocytes and proliferation of inflammatory cells including neutrophils and monocytes are signs of age-related chronic inflammation [39]. The higher level of ferritin also suggests increased inflammatory processes. The chronic inflammation possibly leads to the enrichment of chronic obstructive pulmonary disease (COPD) including emphysema chronic obstructive asthma and obstructive chronic bronchitis, as well as enriched prescription of treatment with steroids such as fluticasone/salmeterol and hydrocortisone.

- **Possibility of chronic anemia:** patients in group O also have significantly lower blood test readouts including lower red blood cells, platelets, hemoglobin, hematocrit and serum iron, all of which suggest chronic anemia. High levels of Vitamin B12 and ferritin are also observed in this group. This is seemingly contradictory to the indicators of anemia, but could be a related physiological mechanism to compensate for the loss of red blood cells.
- **Poor nutritional status:** multiple lines of evidence from physiological measurements and prescribed medicines also suggest that patients in group O suffer from decreased nutritional status. Many of the physiological variables in the nutritional status clusters identified (Fig. 3), such as BMI, albumin, calcium, total iron binding capacity (TIBC) and hemoglobin, have significantly lower values compared to patients in group N of the same chronological age range. The lack of nutrition probably leads to the enrichment of dietary supplements prescribed such as calcium carbonate/Vitamin D3 and multivitamins.
- **Decreased kidney function:** the significant differences in many lab tests related to kidney function, including increased urea nitrogen, creatinine, and uric acid, as well as decreased GFR estimate, indicate that patients in the O group have decreased kidney function. Consistent with these observations, we also found that patients in this group are significantly enriched for kidney transplant procedure, as well as prescription of immunosuppressants usually used before and after organ transplantation such as mycophenolate mofetil, methylprednisolone and mercaptopurine.
- **Potential liver damage:** the increased levels of multiple liver enzymes, including gamma glutamyl transferase, ALT (SGPT), ALK phosphatase, and LD (LDH) lactate dehydrogenase suggest liver damage.

In general, patients in group Y have opposite physiological patterns compared to group O for many of the physiological measurements, but there are few exceptions. The significant findings with interpretations for patients in group Y are tabulated below:

- **Low risk for hypertension and hyperlipidemia:** patients in group Y have significantly lower blood pressure, BMI, body weight and lipids levels including cholesterol and triglycerides, indicating they have lower risk for developing hypertension and hyperlipidemia.
- **Healthy kidney and liver functions:** lab tests that shows significant differences also indicate healthy kidney function, such as lower levels of creatinine, urea nitrogen, uric acid and higher level of GFR estimate. The healthy liver indicators are lower level of liver enzymes in the blood including gamma glutamyl transferase and ALT (SGPT). Interestingly, indicators of glucose tolerance, including glucose and hemoglobin A1C, also have significantly lower levels compared to group N, suggesting lower risk for developing diabetes.
- **Healthier nutritional status:** as oppose to the observations from group O, patients in group Y demonstrate higher levels of positive nutritional status indicators such as higher levels of albumin, red blood cells, hematocrit, hemoglobin, platelets and calcium.
- **Higher risk for venereal diseases:** patients in group Y are enriched for screening for venereal diseases; female patients are enriched for irregular menstrual cycle, abortion, and unwanted pregnancy. Patients in this group are also enriched for HPV vaccine and birth control pills for females, suggesting members in group Y may be more sexually active and as a result are at high risk for venereal diseases.

Other enriched diagnoses and medicines for group Y include acute pharyngitis, fluconazole, a medication that is used to treat fungal infection, and furosemide, a drug to treat edema and swelling. Interestingly, tobacco use is also enriched in group Y. This should be viewed with caution because it is well established that tobacco use has a positive correlation with mortality [40]. It is worth noting that the

model does not predict mortality, therefore, although more tobacco users are disproportionally found within group Y, it does not negate the fact that tobacco use is associated with higher mortality.

In summary, both groups have significantly lower BMI, body weight, and lipid levels. Despite their young chronological age, the physiological profiles of patients in group O resembles profiles of patients that are above 70 (Fig. 3), with only few exceptions such as high hemoglobin A1C. On the contrary, the physiological profiles of members of group Y are more similar to patients younger than 50.

3.5. GWAS identified genes and pathways enriched in groups O and Y

Next, we performed GWAS analysis comparing the O and Y groups to the general population. Of the patients whom we analyzed, 10,123 have genomic profiling data available from the Charles Bronfman Institute of Personalized Medicine Mount Sinai Bio*Me* BioBank. From those, 146 patients belong to the O group, and 197 to the Y group. Using a linear mixed model to control for cryptic relatedness, population structure, age, and gender, we annotated those SNPs that were identified as significant (Figs. S8 and S9). The significance cutoff was set to a p-value of $5e^{-6}$. To interpret the functions of those enriched SNPs, we conducted the following analyses. First, we examined coding variants that cause missense mutations. We found five genes in the O group and two genes in the Y group that harbor significantly enriched missense mutation SNPs (Table 2). Next, we queried gene sets created from significant SNPs, and the genes within their proximity that were found in intronic and promoter regions. These gene sets were subjected to enrichment analysis with Enrichr [33,34] (Table S5–6). Finally, using the software MetaXcan [35] to summarize the GWAS results from the variant level to the gene level using eQTLs generated from GTEx [12], we identified significantly up- and down-regulated genes predicted to be aberrantly expressed in different tissues (Table S7–10).

Below we summarize the findings and their interpretations for all the variants identified to be enriched in group O:

- **Altered inflammation-related genes and pathways:** out of the five genes found to harbor missense mutations in their coding regions, NFAM1 and IL11RA are both implicated in inflammatory-related pathways: NFAM1 is a known regulator of IL-13 and TNF-alpha, and it is involved in regulating B-cell development [41]; whereas IL11RA is a cytokine receptor and a member of the Jak/STAT signaling pathway [42]. Moreover, inflammatory pathways such as PI3K-Akt signaling, AGE-RAGE signaling, and inflammatory mediator regulation of TRP channels are also enriched among the set of genes found in close proximity to the enriched SNPs (Table S5). This observation is consistent with the clinical indicators that suggest that patients in group O are more likely to suffer from chronic-inflammation. It is possible that the identified genetic variants in the inflammation-related genes make the patients more prone to premature age-related chronic inflammation.
- **Anemia:** Another missense mutation that was found to be significantly enriched is in the coding region of ZFPM1. ZFPM1 is known to be involved in platelet activation and anemia [43]. Our analysis predicts that this SNP in ZFPM1 may impair the normal

function of ZFPM1 and this abnormality makes patients harboring this variant to have increased risk to develop anemia.

- **Hypertension, cardiovascular disease and obesity:** out of the five significant SNPs causing missense mutations, one is located in the coding region of PLEKHS1, which might be linked to obesity [44]. Furthermore, the enrichment analysis of SNPs in gene proximity also suggest that there is significant overlap between these genes and genes previously associated with myocardial infarction, blood pressure, and high cholesterol level from GWAS studies collected in dbGAP [45] (Table S5). This observation is also consistent with the clinical indicators that have over-representation for risk for hypertension and cardiovascular disease, indicating that the genetic variants could partially explain the increased risk.
- **Decrease in unfolded protein response (UPR):** from the gene expression level association analysis, we found downregulation of genes in adipose tissue, and enrichment for UPR based on Reactome [46] (Table S10). The UPR response is one of the hallmarks of the aging process [9].

The variants found to be enriched in group Y and their interpretation are listed below:

- **Variants protective against hypertension and obesity:** a SNP enriched in group Y is found to cause a missense mutation in the gene OSBPL9. OSBPL9 is associated with lipid transport and, as a result, it is known to be involved in regulating body weight [47]. Moreover, the gene set created from SNPs in proximity to genes is also enriched for genes known to be associated with blood pressure and BMI based on a gene set library created from dbGAP (Table S6). Together, and with the observation that patients in this group have lower blood pressure and BMI, these enriched genetic variants could have a protective role against hypertension and obesity.
- **Suppressed inflammatory response:** genes in proximity to significant SNPs were found to be enrich for proteoglycans and inflammatory pathways (Table S6). Furthermore, we found that inflammation-related Toll-like receptor signaling is also downregulated adipose tissue by the SNPs eQTL analysis (Table S8).
- **Regulation of lifespan-related genes:** from the MetaScan analysis, we also found two genes, SURF1 and PPARG, predicted to be downregulated in adipose tissue by the SNPs enriched for group Y (Table S8). These two genes have been shown to extend lifespan in knockout mice. From the list of 189 genes that are predicted to be expressed at a lower level in adipose tissue (Table S9), only 63 have been knocked out in mice, and 3.17% of them have shown to increase lifespan. This is a much higher percent compared to the 0.26% for the 27 genes that are marked to extend lifespan in MGI out of 10,493 with any listed phenotype. Hence, the genes from the list of 189 genes without knockout mouse phenotypes are excellent candidates for future knockout experiments, and if successful, these can be potential drug targets for therapeutic intervention for aging. Among these are GPR156, MRGPRX2, CNGA1 and KCND3 which are targets of interest for the NIH Illuminating the Druggable Genome Project (https://commonfund.nih.gov/idg).

Table 2
Significantly enriched SNPs causing missense mutations. BETA: fitted coefficient; SE: standard error for BETA; P: p-value.

Position	Gene ID	Strand	Gene Symbol	SNP	BETA	SE	P	Comparison
chr22:42781198	150372	−	NFAM1	rs151280792:42781198:C:G	0.0321	0.0062	1.97E − 07	O vs N
chr9:127213445	347088	+	ADGRD2	rs76928670:127213445:G:A	0.0358	0.0070	3.72E − 07	O vs N
chr9:34655629	3590	+	IL11RA	rs76429508:34655629:C:T	0.0402	0.0079	4.20E − 07	O vs N
chr16:88600792	161882	+	ZFPM1	rs75878389:88600792:C:T	0.0197	0.0041	1.74E − 06	O vs N
chr10:115535561	79949	+	PLEKHS1	rs34024791:115535561:G:A	0.0220	0.0047	2.22E − 06	O vs N
chr1:52231560	114883	+	OSBPL9	rs61739207:52231560:C:T	0.0523	0.0100	1.59E − 07	Y vs N
chr1:247614606	127623	−	OR2B11	rs74583618:247614606:G:A	0.0403	0.0082	9.26E − 07	Y vs N

4. Discussion

Here we have demonstrated how EMR data can be used to assess the overall health of patients by a scale that is based on the deviation from the patient's predicted chronological age; this measure can be considered the patient's physiological age. We found that a deep neural network regression model is both scalable and accurate in predicting age from thousands of features and hundreds of thousands EMR. However, interpreting deep neural network models is rather challenging. A recent study proposed an approach to interpret deep neural network models by annotating hidden units with conceivable semantic concepts by feeding the model sets of labeled training samples [48]. Similarly, our deep neural network model can potentially be interpreted by aligning the hidden units from the last hidden layer with original individual features or combinations of features. By examining the weights of the last hidden layer, one can potentially interpret how the prediction of age is made from the final hidden units. Regardless, as it is, the deep learning model that we have constructed can be implemented to inform both physicians and patients about the status of the patient physiological age as a general indicator of overall health which can be traced over time.

Our model identified cohorts of patients with significant deviation between their predicted age and their chorological age. EMR-wide association study applied to those cohorts identified physiological measurements that are consistent with existing clinical knowledge about parameters that indicate accelerated or delayed aging. Such analysis also identified medicines that may affect the aging process. Since there are many potential confounding factors that are not included in the model, conclusions should be drawn carefully. For example, we used the median to aggregate physiological measurements of patients across multiple clinical visits, which removes intra-patient variation. The temporal dynamics of physiological aging at the individual level is not captured by our model and could be an interesting follow up study. Future studies that include temporal analysis could track the progression of patients on a timeline of their predicted aging. Such analyses can potentially uncover causality; for example, identifying environmental factors that influence physiological aging. Environmental factors are currently completely missing from our model. Temporal analysis can also identify the aging rate for each patient and warn patients that display accelerated aging. One caveat of the present analysis is that the patient data used to create the model is mostly from individuals who needed medical assistance. Hence, this fact alone likely confound the study because the general healthy population is not sampled. While the GWAS analysis performed has identified novel mutations in genes, and potential regulatory regions near genes that are likely associated with the aging process, the sample size we have utilized is relatively small and racially heterogeneous. Further studies with larger cohorts that employ deep sequencing, and/or deeper SNP profiling are expected to become available soon. Reapplying a similar analysis on a larger cohort would confirm or refine the reported observations. Regardless of these shortcomings, our success in predicting age from vital signs and lab tests can inform physicians and their patients to assist in promoting healthier lifestyle and early detection of human disease.

Author contributions

ZW designed, implemented, and benchmarked the predictive deep learning model; performed the EMR-wise association analyses, interpreted part of the GWAS results, and created all figures. LL and JTD provided access to a cleaned and de-identified version of the MSHS-EMR. AI interpreted the physiological measurements and advised in regards to clinical interpretations. BSG, LL and JTD performed the GWAS analysis. AM and ZW wrote the manuscript with contributions from other co-authors. AM conceived, directed and managed the study.

Conflict of interest

Acknowledgements

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.11.003.

References

[1] G.T. Baker Iii, R.L. Sprott, Biomarkers of aging, Exp. Gerontol. 23 (4–5) (1988) 223–239.
[2] R.L. Sprott, Biomarkers of aging and disease: introduction and definitions, Exp. Gerontol. 45 (1) (2010) 2–4.
[3] A. Bürkle, M. Moreno-Villanueva, J. Bernhard, M. Blasco, G. Zondag, J.H.J. Hoeijmakers, O. Toussaint, B. Grubeck-Loebenstein, E. Mocchegiani, S. Collino, et al., MARK-AGE biomarkers of ageing, Mech. Ageing Dev. 151 (2015) 2–12.
[4] J. Lara, R. Cooper, J. Nissan, A.T. Ginty, K.-T. Khaw, I.J. Deary, J.M. Lord, D. Kuh, J.C. Mathers, A proposed panel of biomarkers of healthy ageing, BMC Med. 13 (1) (2015) 222.
[5] D.P. Chen, A.A. Morgan, A.J. Butte, Validating pathophysiological models of aging using clinical electronic medical records, J. Biomed. Inform. 43 (3) (2010) 358–364.
[6] A. Simm, N. Nass, B. Bartling, B. Hofmann, R.-E. Silber, A. Navarrete Santos, Potential biomarkers of ageing, Biol. Chem. 389 (3) (2008) 257–265.
[7] T.E. Johnson, Recent results: biomarkers of aging, Exp. Gerontol. 41 (12) (2006) 1243–1246.
[8] E. Lixie, J. Edgeworth, L. Shamir, Comprehensive analysis of large sets of age-related physiological indicators reveals rapid aging around the age of 55 years, Gerontology 61 (6) (2015) 526–533.
[9] C. López-Otín, M.A. Blasco, L. Partridge, M. Serrano, G. Kroemer, The hallmarks of aging, Cell 153 (6) (2013) 1194–1217.
[10] R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. Costa, V.E. Fraifeld, J.P. de Magalhães, Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing, Nucleic Acids Res. 41 (D1) (2012) D1027–D1033.
[11] A. Budovsky, T. Craig, J. Wang, R. Tacutu, A. Csordas, J. Lourenço, V.E. Fraifeld, J.P. de Magalhães, LongevityMap: a database of human genetic variants associated with longevity, Trends Genet. 29 (10) (2013) 559–560.
[12] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al., The genotype-tissue expression (GTEx) project, Nat. Genet. 45 (6) (2013) 580–585.
[13] J. Yang, T. Huang, F. Petralia, Q. Long, B. Zhang, C. Argmann, Y. Zhao, C.V. Mobbs, E.E. Schadt, J. Zhu, et al., Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases, Sci. Rep. 5 (2015) 15145.
[14] M.J. Peters, R. Joehanes, L.C. Pilling, C. Schurmann, K.N. Conneely, J. Powell, E. Reinmaa, G.L. Sutphin, A. Zhernakova, K. Schramm, et al., The transcriptional landscape of age in human peripheral blood, Nat. Commun. 6 (2015) 8570.
[15] S. Horvath, DNA methylation age of human tissues and cell types, Genome Biol. 14 (10) (2013) 3156.
[16] G. Hripcsak, D.J. Albers, Next-generation phenotyping of electronic health records, J. Am. Med. Inf. Assoc. 20 (1) (2013) 117–121.
[17] O. Gottesman, H. Kuivaniemi, G. Tromp, W.A. Faucett, R. Li, T.A. Manolio, S.C. Sanderson, J. Kannry, R. Zinberg, M.A. Basford, et al., The electronic medical records and genomics (eMERGE) network: past, present, and future, Genet. Med. 15 (10) (2013) 761–771.
[18] J.C. Kirby, P. Speltz, L.V. Rasmussen, M. Basford, O. Gottesman, P.L. Peissig, J.A. Pacheco, G. Tromp, J. Pathak, D.S. Carrell, et al., PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inf. Assoc. 23 (6) (2016) 1046–1052.
[19] H. Mo, W.K. Thompson, L.V. Rasmussen, J.A. Pacheco, G. Jiang, R. Kiefer, Q. Zhu, J. Xu, E. Montague, D.S. Carrell, et al., Desiderata for computable representations of electronic health records-driven phenotype algorithms, J. Am. Med. Inf. Assoc. 22 (6) (2015) 1220–1230.
[20] C. Shivade, P. Raghavan, E. Fosler-Lussier, P.J. Embi, N. Elhadad, S.B. Johnson, A.M. Lai, A review of approaches to identifying patient phenotype cohorts using

electronic health records, J. Am. Med. Inf. Assoc. 21 (2) (2014) 221–230.

[21] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (2016) 26094.

[22] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Aistats, 2010, pp. 249–256.

[23] D. Kingma, J. Ba, Adam: a method for stochastic optimization, 2014. arXiv preprint arXiv:14126980.

[24] F. Chollet, Keras, 2015. URL <http://keras io>.

[25] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: a CPU and GPU math compiler in Python, in: Proc 9th Python in Science Conf, 2010, pp. 1–7.

[26] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J.R. Stat. Soc. 57 (1995).

[27] O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes, Nat. Meth. 9 (2) (2012) 179–181.

[28] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLOS Genet. 5 (6) (2009) e1000529.

[29] H.M. Kang, J.H. Sul, S.K. Service, N.A. Zaitlen, S.-Y. Kong, N.B. Freimer, C. Sabatti, E. Eskin, Variance component model to account for sample structure in genome-wide association studies, Nat. Genet. 42 (4) (2010) 348–354.

[30] J. Galinsky Kevin, G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, J. Patterson Nick, L. Price Alkes, Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia, Am. J. Hum. Genet. 98 (3) (2016) 456–472.

[31] J. Galinsky Kevin, P.-R. Loh, S. Mallick, J. Patterson Nick, L. Price Alkes, Population structure of UK Biobank and ancient eurasians reveals adaptation at genes influencing blood pressure, Am. J. Hum. Genet. 99 (5) (2016) 1130–1139.

[32] V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, M. Morgan, VariantAnnotation: a bioconductor package for exploration and annotation of genetic variants, Bioinformatics 30 (14) (2014) 2076–2078.

[33] M.V. Kuleshov, M.R. Jones, A.D. Rouillard, N.F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S.L. Jenkins, K.M. Jagodnik, A. Lachmann, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucleic Acids Res. (2016).

[34] E. Chen, C. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (1) (2013) 128.

[35] A. Barbeira, K.P. Shah, J.M. Torres, H.E. Wheeler, E.S. Torstenson, T. Edwards, T. Garcia, G.I. Bell, D. Nicolae, N.J. Cox, MetaXcan: summary statistics based gene-

level association method infers accurate PrediXcan results, bioRxiv (2016) 045260.

[36] W.S. Cleveland, Robust locally weighted regression and smoothing scatterplots, J. Am. Stat. Assoc. 74 (368) (1979) 829–836.

[37] J.E. Manson, W.C. Willett, M.J. Stampfer, G.A. Colditz, D.J. Hunter, S.E. Hankinson, C.H. Hennekens, F.E. Speizer, Body weight and mortality among women, N. Engl. J. Med. 333 (11) (1995) 677–685.

[38] M.V. Blagosklonny, Why men age faster but reproduce longer than women: mTOR and evolutionary perspectives, Aging 2 (5) (2010) 265–273.

[39] A. Ahmad, S. Banerjee, Z. Wang, D. Kong, A.P.N. Majumdar, F.H. Sarkar, Aging and inflammation: etiological culprits of cancer, Curr. Aging Sci. 2 (3) (2009) 174–186.

[40] D.S. Timberlake, D. Nikitin, N.J. Johnson, S.F. Altekruse, A longitudinal study of smokeless tobacco use and mortality in the United States, Int. J. Cancer 141 (2) (2017) 264–270.

[41] M. Ohtsuka, H. Arase, A. Takeuchi, S. Yamasaki, R. Shiina, T. Suenaga, D. Sakurai, T. Yokosuka, N. Arase, M. Iwashima, NFAM1, an immunoreceptor tyrosine-based activation motif-bearing molecule that regulates B cell development and signaling, Proc. Natl. Acad. Sci. USA 101 (21) (2004) 8126–8131.

[42] F. Van Leuven, L. Stas, C. Hilliker, Y. Miyake, P. Bilinski, A. Gossler, Molecular cloning and characterization of the human interleukin-11 receptor α-chain gene, IL11RA, located on chromosome 9p13, Genomics 31 (1) (1996) 65–70.

[43] H.-Y. Yang, S.H. Kim, S.-H. Kim, D.-J. Kim, S.-U. Kim, D.-Y. Yu, Y.I. Yeom, D.-S. Lee, Y.-J. Kim, B.-J. Park, The suppression of zfpm-1 accelerates the erythropoietic differentiation of human CD34+ cells, Biochem. Biophys. Res. Commun. 353 (4) (2007) 978–984.

[44] J. Kotoh, D. Sasaki, K. Matsumoto, A. Maeda, Plekhs1 and Prdx3 are candidate genes responsible for mild hyperglycemia associated with obesity in a new animal model of F344-fa-nidd6 rat, J. Vet. Med. Sci. 78 (11) (2016) 1683.

[45] M.D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, et al., The NCBI dbGaP database of genotypes and phenotypes, Nat. Genet. 39 (10) (2007) 1181–1186.

[46] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, et al., Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Res. 39(Suppl. 1) (2011) D691–D697.

[47] M. Ngo, N.D. Ridgway, Oxysterol binding protein–related protein 9 (ORP9) is a cholesterol transfer protein that regulates Golgi structure and function, Mol. Biol. Cell 20 (5) (2009) 1388–1399.

[48] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: quantifying interpretability of deep visual representations, 2017. arXiv preprint arXiv:170405796.