Linear regression

Prediction of Biological Age

Anastasiia Buslova

June 12, 2017

Abstract

The biological age (BA) equation is a prediction model that utilizes an algorithm to combine various biological markers of ageing. This model has been used to assess the ageing process in a more precise way and may predict possible diseases better as compared with the chronological age (CA). Here study focus on predicting BA as deviation from healthy state of aging and treats CA as known and unkown value. The multiple linear regression (MLR), principal component analysis (PCA) and Klemera and Doubal method (KDM) are applied on data set from Research Laboratories, Russia. Thesis summarize the up-to-date knowledge about the BA formula construction and discuss the influential factors, so as to give an overview of BA estimate by PCA, MLR, KDM and group miltiple linear regression, choices of test items, and selection of ageing biomarkers. It is also discussed the advantages and disadvantages of every method with reference to the construction mechanism, accuracy, and practicability of several common methods in the construction of the BA formula.

1 Introduction

There is no final definition of aging since this process is influenced by a number of different factors and conditions. In general terms, aging can be defined as a structural decline which results in developing illnesses and increases the risk of mortality [1]. Despite the fact that nearly all species are affected by the aging process, there is not consensual agreement of what factors cause age-related decline as the ageing rate is interconnected with various factors [2]. It is of great importance whether one is exposed to certain environment or has genetic predisposition as well as the presence of smoking and other damaging habits [3].

Consequently, chronological age (CA) is not the best reliable indicator of the body breakdown, although it gives some understanding of the rate of aging. For better measurement of aging a concept of biological age (BA) was introduced as an indicator of body's condition, reflecting the degree of age related frailty. Despite the correlation between chronological and biological ages, individuals of the same chronological age can have absolutely different health conditions. As the reserve capacity has dropped to a critically low mark the signs of frailty are visible, which results in complications after even slight disturbance. BA is a hidden state of the body. That makes it hard to understand which factors are crucial in foreseeing the case.

Frailty is important while both searching for a health care plan and identifying those in need of medical attention. Moreover, prediction of BA could give understanding whether individual's medication, way of life and environment help prolong life or not.

There are two general approaches to BA prediction. Common observations of elderly people suggest a natural connection between getting older and frailty of some kind such as falls, less mobility, less independence, hospitalization, disability, and death. A reasonable conclusion is to measure BA as deviation from young state of the body. Other method consists in an idea of comparison rate of aging. Although a middle-aged person may not be as healthy as a young adult, he or she still can be relatively healthy in contrast with a sick individual. Therefore, BA is a measurement of healthy aging. Depending on the goal of research, studies are choosing different methods of measuring BA. There is no agreement about the approaches [4][5] as calculating BA has been a relatively recent activity [6].

Alex Comfort was the first to suggest an idea to measure age-related changes [7]. Since there are a lot of factors involved in the aging process one of the main interests of many researches is to find out which biomarkers are responsible for aging. A biomarker is defined as a quantitative indicator of some biological condition as well as state such as a blood test. Some major work has been done to identify the biomarkers of aging so that it could benefit the research of the processes of growing old in humans and animals [8]; yet little success has been achieved [9]. It has been suggested that it is unlikely to point a single marker responsible for the aging process as the process in extremely complex especially in humans.

Despite the fact that there have been several publications on the measurement of BA there is little agreement about the approach to BA measurement and whether such measures are valid and reliable. Over the years, a number of varying mathematical algorithms have been suggested, such as multiple linear regression (MLR) [8][11][12], principal component analysis (PCA) [13][14], and more recently, an approach proposed by Klemera and Doubal [3][15][17].

The problem is that it is impossible to measure the intrinsic value of BA, which makes the acceptance of calculated estimates difficult. However, the common criteria should be used to evaluate the reliability and validity of BA measurements. For instance, BA calculations should have realistic measurements within the limits of a recorded life span. BA estimations should be able to find and pinpoint the at risk individuals before they become ill. The peculiarity of modern methods used to single out the at risk individuals is based on indexes of disease, weaknesses and deficiencies.

In the end, BA should meet the criteria introduced for biomarkers of aging which means that a biomarker unlike the chronological age should be a better indicator of different age-associated biological and functional consequences [10].

Using these criteria, the focus of this study is to compare BA measurements, estimated using various methods that have been proposed in the literature (multiple linear regression, PCA, Klemera and Doubal method), and also propose a new way of estimation of BA with the goal of determining their validity and usefulness in predicting aging rate, within a large representative human sample.

2 Study population

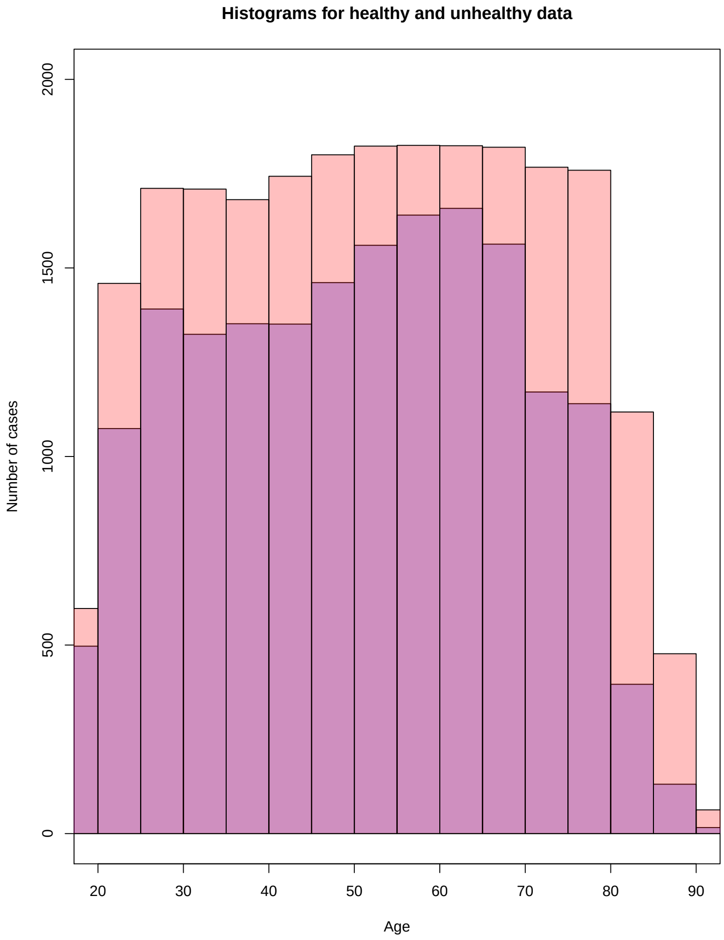

Figure 1: Visualisation of data set for female population. Unhealthy data (red) contains more observation than healthy data (blue). There is small disproportional represented in data depending on age group; there are more observation for middle-aged and elderly people.

The study population included anonymous data records of patients between 2014 and 2016 from Invitro Research Laboratories, Russia. Mostly all blood tests were taken as a prescription of a therapist, consequently, data set is highly corrupted by unhealthy measurements. Data set consists of patient sex, age and anonymous ID, test name and result, date of tests and diagnoses.

This study was limited to adults aged 20-90 years, in order to ensure that participants were old enough to be experiencing detectable age-related changes in biomarkers, yet not too old as to represent a select group with above-average health and longevity. Of the 1.409.867 data set subjects, aged 20-90 years, final analytic sample included 1.387.186 observations (or 109.666 participants), 75% of them are female. The final sample consists from 80% unhealthy observations. As a result, that the data was obtained from Research Laboratory only about 20% of observations from the final sample are representing ages from 20 to 40.

Overall the data has measurements of 85 blood test, but many of them are not related to aging or has few results.

3 Theory

3.1 Multiple linear regression

One of the most common method for prediction of BA is linear regression. General MLR approach predicts BA such as

$$\mathrm{BA}_i = \mathrm{PredictedCA}_i = a_0 + \sum_{j=1}^{n} a_j x_{ij} \tag{1}$$

Here BA is assumed to be equal to the predicted CA of an individual and is based upon the relationship of true (measured) CA and several biomarkers (n). The misinterpretation of BA at the regression edge is determined by mathematical factors and MLR also neglects the discontinuity of the aging rate during a life span [16].

3.2 Multiple linear regression groups method

In order to reduce bias between CA and BA, the following approach is suggested: a model was built only on results of healthy people. Thus reduces impact of outliers - "unhealthy test results" - and takes into account a general natural observation that unhealthy people are most likely to have diseases and have a greater mortality rate. A resulting model measures the healthy aging.

One of the challenges of predicting BA by MLR models is to get a reasonable confidence interval for an estimated age. Evaluated BA should be in a rather narrow age range, or it became impossible to verify a legitimacy of prediction. Many researches introduced an age prediction model done by MLR [3][17], but nearly in every case it has a wide confidence interval starting from 15-20 up to 100 years. Under such conditions estimation of BA looks unreliable as nearly every age is inside that confidence interval.

In order to obtain a better estimation of BA the following method was suggested. Several MLR models were built for different age groups (20-30, 31-40, 41-50, 51-60, 61-70 and 71-90). This approach gives better understanding what blood markers are relevant to predicting of various ages and how they differ during the lifetime. Thus models have good prediction power and generally estimated BA varies inside rather narrow interval of ±4 years as later will be shown.

Although the MLR groups models have clear benefits in comparison to a simple MLR model calculated by formula 1 this method has a major drawback. In order to estimate BA it is necessary to conclude which group model should be used for prediction. Generally, there are two approaches to this matter. One is to consider CA as known value, the other is do prediction without CA. At the last case there is a duplicity in estimation as the prediction suggests that CA is the same as BA. Most recent studies suggest that CA is known and include it in the final model. This principe gives noticeably better results and it holds for both PCA and Klemera and Doubal method (KDM) [15][16]. This study check two approaches with known and unknown CA.

For initial rough estimation of CA the Mahalanobis distance was used. In order to better represent an individual's degree of aging, a model has to consider relation between biomarkers measurements as all of them are simultaneous play part in BA. The general idea is to measure how far a given vector deviates from the other vector. The distance gives understanding of similarity between them and takes into account the correlations of the data set.

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})S^{-1}(\vec{x} - \vec{\mu})} \tag{2}$$

In general case $\vec{x}$ is observation vector, $\vec{\mu}$ is mean vector, and S is covariance matrix. In our case $\vec{x}$ is blood measurements of person with unknown CA and μ is a matrix of vectors matching mean for different ages.

One of the main challenges of prediction BA consist in verification of estimation value. An estimation of BA relies on CA which gives bias and BA intrinsic value which adds uncertainty into validation. Suggested approach provides a BA estimated by the model built on healthy people. In other words, the BA represents a measurement of the healthy aging. Consequently, residuals between the BA and CA quantifies variation from normal state of aging. Resulting calculations of Mahalanobis distance to every group will provide a rough suggestion of a BA. Thus allows to choose between appropriate MLR age group's models and estimate of BA with minimal variance. This method takes into consideration a) complexity of relations between biomarkers b) change in prediction power of different blood markers with age and c) drawbacks of a simple MLR model.

3.3 Principal component analysis

Then PCA is applied to reduce the dimension of the variables. Indices with eigenvalues greater than 1.0 are determined as principal components, and the greater one is called the first principal component. Other variables load onto the first principal component to explain the variation of BA. Two kinds of PCA are done, with CA and without CA. PCA with CA determines the relationship between CA and principal components. While PCA without CA is used to show whether the relationships will be held without the influence of CA. Hence the biomarkers of ageing are selected.

PCA was applied on the data with a full range of biomarker's measurements. As the result, reduced data was reconstructed only upon relevant principal components and new estimation was made by MLR 1. The formula for reconstructed data is

$$\text{data} = \sum_{i=1}^{k} \left(\text{scores of blood marker}_i * \text{loadings of blood markers}_i^{\mathrm{T}} + \text{blood markers mean}_i\right) \tag{3}$$

where k is a chosen number of principal components. Because the BA formula is underestimated of the means of BA for the upper end of regression and overestimated for the lower end, researchers correct the estimation using the following method:

$$\text{corrected BA} = \text{BA} + z \tag{4}$$

$$z = (y_i - y)(1 - b)$$

where y_i is the individual's CA, y is the average of CA and b stands for the coefficient of linear regression between CA and BA [16].

3.4 Klemera and Doubal method

This method is based on estimation of BA estimates as minimizing the distance between m regression lines and m biomarker points, within an m dimensional space of all biomarkers. In their article, the authors used computer-generated simulations to validate the method they propose. They defined BA as equal to CA, plus some random variable, R_{BA}, with a mean of zero and a variance s_{BA}^2. Klemera and Doubal presented two alternative methods for calculating the optimum estimates of BA, equations 5 and 6, in which the later method utilizes CA in the final equation and using simulations, was shown to be superior:

$$\text{BA}_{\text{E}} = \frac{\sum_{j=1}^{m} \left(x_j - q_j\right) \frac{k_j}{s_j^2}}{\sum_{j=1}^{m} \left(\frac{k_j}{s_j^2}\right)^2} \tag{5}$$

$$\text{BA}_{\text{EC}} = \frac{\sum_{j=1}^{m} \left(x_j - q_j\right) \frac{k_j}{s_j^2} + \frac{\text{CA}}{s_{\text{BA}}^2}}{\sum_{j=1}^{m} \left(\frac{k_j}{s_j^2}\right)^2 + \frac{1}{s_{\text{BA}}^2}} \tag{6}$$

where k is the slope of the linear function, intercept q, and x biomarkers measurement. In order to produce an estimate for BA, using equation 6, s_j^2 and s_{BA}^2 have to be calculated. The value, s_j , represents the root mean squared error of a biomarker regressed on BA. However, given that BA is not measurable, root mean squared errors from the regressions between each biomarker and CA, rather than BA, were used, as suggested by Cho and colleagues [18]. Finally, in order to calculate s_{BA}^2 the following two equations were used:

$$r_{\text{char}} = \frac{\sum_{j=1}^{m} \frac{r_j^2}{\sqrt{1 - r_j^2}}}{\sum_{j=1}^{m} \frac{r_j}{\sqrt{1 - r_j^2}}}$$

$$s_{\text{BA}}^2 = \frac{\sum_{j=1}^{n} \left((\text{BA}_{\text{E}i} - \text{CA}_i) - \sum_{i=1}^{n}(\text{BA}_{\text{E}i} - \text{CA}_i)/n\right)^2}{n} - \frac{1 - r_{\text{char}}^2}{r_{\text{char}}^2} \times \frac{\left(\text{CA}_{\text{max}} - \text{CA}_{\text{min}}\right)^2}{12m}$$

The value r_j^2 used to calculate the characteristic correlation coefficient, refers to the variance explained by regression CA on m biomarkers. Finally, in accordance with the assumption made by Klemera and Doubal, s_{BA}^2 was transformed so that s_{BA} maintained the same mean but was now linearly increasing with age, with a difference of five between subjects at $\mathrm{CA}_{\mathrm{min}}$ and $\mathrm{CA}_{\mathrm{max}}$.

4 Preparation of data

As described above the MLR models are conducted on healthy people results. From initial models there have been excluded the participants whose biomarkers test are out of normal boundaries known from medical laboratories [19].

After such procedure and taking away irrelevant blood test the data set shrank into 141.618 separate measurements for female and 46.266 for male. Unfortunately, there are a lot of missing values in the data as participants were checked only by prescribed tests, not full-range of various blood test markers.

These leaves a question of missing data. Firstly, from the data there were excluded biomarkers with less than 20% of measurements, so it has left 15 markers out of 54: C-reactive protein (R), Alkaline phosphatase, Alanine aminotransferase (R), Aspartate aminotransferase (R), Gamma-glutamyl transferase (R), Bilirubin, Creatinine, Cholesterol, Glucose, Protein total, Triglycerides, Urea, Uric acid, Thyroid-stimulating hormone (R) and Free thyroxine (R). Although markers were downgraded more than a half, there have been kept the most important biomarkers known to have a relation to age. Secondly, it allowed to have patients only with no more than 5 missing measurements.

After this procedure data set dropped to 3.326 and 220 results for different ages for female and male respectively. Due to huge disproportional in data for male this study do not include results for male population. Among healthy data 75% was used as a training set, 25% as a test set. Whole unhealthy data was left for testing.

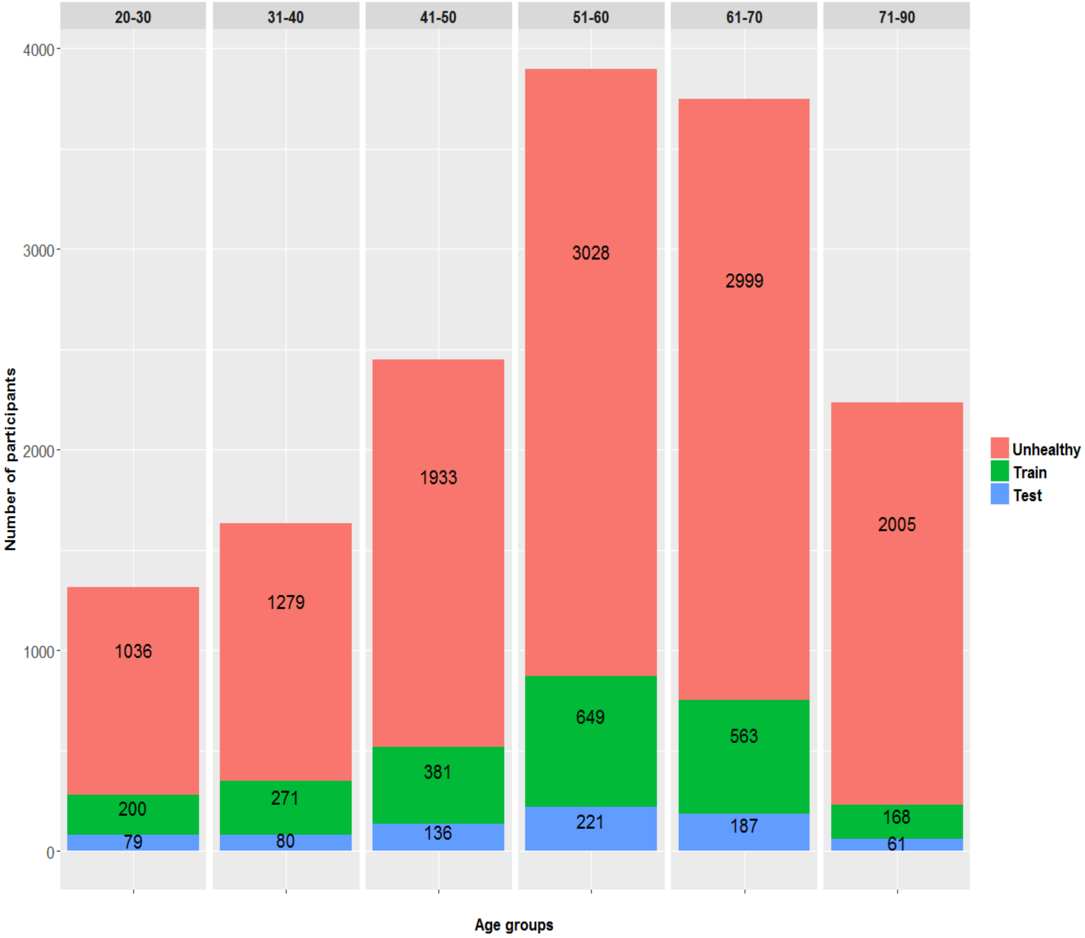

Figure 2: Number of participants for each age group.

Finally, the missing blood test results were filled by the mean of two closest age neighbors' measurements. Also it was tried out filling the data with three closest neighbors as well as predict not available results by density of the matching blood test, but both of these methods were close to mean of two neighbors due to the nature of data. For this reason the closest two neighbors was found the best way of estimation of NA results. For reconstruction of unhealthy data there were used 3 closest neighbours as its results can vary in wide range.

5 Results

5.1 Multiple linear regression

Several MLR models were constructed with different clustering of biomarkers. The best MLR model was chosen on the basis of smallest AIC.

$$\begin{aligned}\mathrm{BA_E} = -36.08 + 1.21 * \text{C-reactive protein} + 0.2 * \text{Alkaline phosphatase}-\\ - 0.23 * \text{Alanine aminotransferase} + 0.63 * \text{Aspartate aminotransferase}+\\ + 0.09 * \text{Gamma-glutamyl transferase} + 5.23 * \text{Cholesterol} + 8.74 * \text{Glucose}-\\ - 0.4 * \text{Protein total} + 3.55 * \text{Urea} + 0.46 * \text{Thyroid-stimulating hormon}\end{aligned} \tag{7}$$

Results of estimated BA for the best MLR model are represented in a following table 1:

	Train data	Test data	Unhealthy data
(-2,2) in%	18.46	18.03	4.76
(-4,4) in%	37.37	38.7	9.46
(-10,10) in%	77.51	74.04	24.64
(min,max)	[23.23,84.03]	[25.4,78.6]	[17.42, 665.07]
mean of residuals	6.75	7.01	26.48

Table 1: Results of BA prediction by MLR method. Table shows accuracy of prediction by presenting in percentage how many estimations are different from CA and in what range: ± 2, ± 4 and ± 10 years. Also it demonstrates mean of residuals as well as a maximum and a minimum of estimated BA. Overall, method almost completely fails in prediction of unhealthy data. The best results are marked in green, the worst in red.

Prediction of MLR model barely holds for healthy people but absolutely fails in estimation for unhealthy population and it is out of reasonable boundaries of the life span. Therefore the model cannot be used as a good predictor of either BA or CA; however the model 7 could be used as first rough estimation of the age for healthy population.

5.2 Multiple linear regression group models

In order to improve MLR model 7, 6 group models were built on train data and following models were obtained.

$$\begin{aligned} \mathrm{BA}_{\mathrm{E}_{20-30}} = {} & 18.78 + 0.7 * \text{C-reactive protein} - 0.03 * \text{Alkaline phosphatase} - 0.1 * \text{Bilirubin} + 0.98 * \text{Cholesterol}+ \\ & + 0.9 * \text{Glucose} + 1.38 * \text{Triglycerides} - 0.48 * \text{Thyroid-stimulating hormone} \end{aligned} \tag{8}$$

$$\begin{aligned} \mathrm{BA}_{\mathrm{E}_{31-40}} = {} & 31.48 + 0.48 * \text{C-reactive protein} + 0.02 * \text{Alkaline phosphatase} + 0.05 * \text{Alanine aminotransferase}+ \\ & + 0.11 * \text{Bilirubin} + 0.96 * \text{Cholesterol} - 0.29 * \text{Free thyroxine} \end{aligned} \tag{9}$$

$$\begin{aligned} \mathrm{BA}_{\mathrm{E}_{41-50}} = {} & 32.05 + 0.35 * \text{C-reactive protein} + 0.05 * \text{Alkaline phosphatase} + 0.07 * \text{Alanine aminotransferase}+ \\ & + 0.09 * \text{Bilirubin} + 0.48 * \text{Cholesterol} + 0.99 * \text{Glucose} + 0.92 * \text{Triglycerides} \end{aligned} \tag{10}$$

$$\begin{aligned} \mathrm{BA}_{\mathrm{E}_{51-60}} = {} & 43.82 + 0.02 * \text{Alkaline phosphatase} + 0.08 * \text{Aspartate aminotransferase} - 0.09 * \text{Bilirubin}+ \\ & + 0.57 * \text{Cholesterol} + 0.55 * \text{Glucose} + 0.38 * \text{Urea} + 0.11 * \text{Free thyroxine} \end{aligned} \tag{11}$$

$$\begin{aligned} \mathrm{BA}_{\mathrm{E}_{61-70}} = {} & 61.72 + 0.05 * \text{Aspartate aminotransferase} - 0.06 * \text{Gamma-glutamyl transferase}- \\ & - 0.37 * \text{Triglycerides} + 0.28 * \text{Thyroid-stimulating hormone} + 0.22 * \text{Free thyroxine} \end{aligned} \tag{12}$$

$$\begin{aligned} \mathrm{BA}_{\mathrm{E}_{71-90}} = {} & 80.34 - 0.16 * \text{Aspartate aminotransferase} - 0.1 * \text{Gamma-glutamyl transferase} + 0.15 * \text{Bilirubin}- \\ & - 0.53 * \text{Cholesterol} - 0.69 * \text{Thyroid-stimulating hormone} + 0.25 * \text{Free thyroxine} \end{aligned} \tag{13}$$

The table 2 shows that GMLR ia much better in predicting age than a simple MLR model. Almost every measure for different groups is inside an [-4,4] interval for healthy people which gives a good comparison power BA with CA. The main goal is to understand individual's pace of aging and estimation on unhealthy people gives clear result of difference between CA and BA. Moreover, collected results look promising as they are inside the life span interval and reflect prediction changes between healthy and unhealthy data. The worst prediction result was obtained on the most unhealthy population: participants from 71 to 90 years old from unhealthy data which has to be expected. Nevertheless, the results are calculated under assumption of known CA which is not always the case.

Group model		Train data	Test data	Unhealthy data
20-30	[-2,2] in %	53	40.51	35.71
	(-4,4) in %	90	86.08	60.91
	(-10,10) in %	100	100	90.15
	(min,max)	[22.02, 29.33]	[23.81, 29.83]	[-174.78, 87.19]
	mean of residuals	2.08	2.5	5.01
31-40	[-2,2] in %	49.45	53.75	34.4
	(-4,4) in %	91.14	92.5	66.85
	(-10,10) in %	100	100	94.76
	(min,max)	[32.95, 38.89]	[32.33, 37.96]	[15.11, 107.26]
	mean of residuals	2.13	2.66	3.77
41-50	[-2,2] in %	55.64	48.53	32.64
	(-4,4) in %	94.49	85.29	57.27
	(-10,10) in %	100	100	91.57
	(min,max)	[43.12, 48.91]	[43.78, 49.94]	[42.36, 116.87]
	mean of residuals	1.9	2.31	4.4
51-60	[-2,2] in %	46.53	53.85	43.43
	(-4,4) in %	93.53	90.05	76.72
	(-10,10) in %	100	100	99.17
	(min,max)	[53.01, 57.41]	[53.65, 57.41]	[44.17,79.54]
	mean of residuals	2.11	2.01	2.76
61-70	[-2,2] in %	50.44	41.71	41.95
	(-4,4) in %	92.72	87.17	75.99
	(-10,10) in %	100	100	97.7
	(min,max)	[63.57, 67.26]	[63.66, 66.72]	[18.67,92.1]
	mean of residuals	2.05	2.32	2.95
71-90	[-2,2] in %	54.17	42.62	28.63
	(-4,4) in %	86.31	80.33	52.97
	(-10,10) in %	99.4	98.36	84.24
	(min,max)	[72.77, 79.28]	[73.46, 79.12]	[-59.82,89.03]
	mean of residuals	2.16	2.76	6.26

Table 2: Results of BA prediction by group MLR method with assumption of known CA. Overall, the method shows a good prediction power CA on healthy population therefore reasonable estimation of BA on both healthy and unhealthy data sets. The best results are marked in green, the worst in red.

To get results with unknown CA it is necessary to get a rough estimation of age which could provide information what model(8 - 13) has to be used for estimation purposes. In order to get it Mahalanobis distance was calculated for different ages by the formula 2. The table 3 summarizes the results. Mahalanobis distance was showed to be worse in a rough prediction of age than a simple MLR model for healthy people, but much better estimation for unhealthy population. Overall this approach outperformed a simple MLR model 7.

	Train data	Test data	Unhealthy data
[-2,2] in %	12.47	11.06	11.31
(-4,4) in %	23.06	24.04	22.56
(-10,10) in %	51.96	52.4	51.18
(-15,15) in %	70.73	68.87	68.28
(min,max)	[20.79,78.93]	[22.28,78.35]	[20.45,78.66]
mean of residuals	11.71	12.24	12.75

Table 3: A prediction of age calculated by Mahalanobis distance could be used only as rough estimation. It showed approximately same prediction power on both data sets. The best results are marked in green, the worst in red.

5.3 Principal component analysis

For estimation there have been chosen 6 principal components with cumulative explained variance 40% and eigenvalues > 1 (1656 for the first PC). Loading of PCA showed that Alkaline phosphatase, Gamma-glutamyl transferase, Alanine aminotransferase, Aspartate aminotransferase, Creatinine and Protein total are the most important biomarkers for age prediction. The table 4 shows result for the estimation, where prediction without CA is calculated by formula 3 and with CA by formula 4.

		Train data	Test data	Unhealthy data
Without CA	[-2,2] in %	13.1	15.38	11.83
	[-4,4] in %	27.67	28	23.71
	(-10,10) in %	63.79	63.34	53.11
	(-15,15) in %	81.36	78	71.01
	(min,max)	[31.07, 87.27]	[32.18, 79.6]	[27.63, 154.11]
	mean of res	9.06	9.34	11.83
Knowing CA	[-2,2] in %	24.66	24.64	33.5
	[-4,4] in %	45.95	45.91	60.69
	(-10,10) in %	87.81	85.46	94.08
	(-15,15) in %	97.11	96.88	97.98
	(min,max)	[11.68, 97.11]	[11.62, 87.06]	[6, 154.16]
	mean of res	5.25	5.3	4.1

Table 4: Estimation of BA calculated by PCA with 6 chosen blood markers showed good prediction power with known CA and average results for prediction without CA. The best results are marked in green, the worst in red.

Overall, estimation of BA by PCA has a good prediction power regardless that covariants were reduced to only 6 out of 15 blood markers. Also it could be noticed that PCA gauges unhealthy population with high accuracy. Although some of the predictions are out of plausible boundaries of human's life even taking into account that sickness could impact estimation in wide spectrum. If PCA was done by using the variable set consisting of the six biomarkers and CA, the loadings of CA showed high correlation with the components which proves connection between BA and CA.

5.4 Klemera and Doubal method model

KDM shown to be a more reliable predictor of mortality than MLR or PCA [15] as this approach look at prediction in multidimensional space and connects every biomarker into prediction, although it relies on complicated calculations. Below table shows results for KDM method where estimation without knowing CA is calculated by the formula 5 and with CA by the formula 6.

		Train data	Test data	Unhealthy data
Without CA	[-2,2] in %	14.55	14.66	8.33
	[-4,4] in %	28.35	27.88	16.12
	(-10,10) in %	65.84	64.54	39.1
	(-15,15) in %	83.04	79.09	56.58
	(min,max)	[2.85, 97.54]	[2.4, 92.67]	[-17.02, 429.46]
	mean of residuals	8.6	9.03	17.22
Knowing CA	[-2,2] in %	35.65	34.25	9.93
	[-4,4] in %	64.84	64.3	19.23
	(-10,10) in %	97.27	96.88	46.57
	(-15,15) in %	99.92	100	66.14
	(min,max)	[13.99, 85.24]	[11.98, 83.5]	[-9.41, 361.46]
	mean of residuals	3.52	3.64	14.32

Table 5: Estimation of BA calculated by KDM with 15 blood markers demonstrate better results for given CA and overall showed plausible results for both populations. The best results are marked in green, the worst in red.

KDM showed best results for not-knowing CA case, but was outperformed by group MLR models for knowing CA. KDM is superior to Mahalanobis distance calculation as well as PCA method on healthy people. One of disadvantages of this approach is that KDM does not choose relevant biomarkers and they have to be chosen before estimation. For testing reasons other two estimations were applied: KDM with 6 biomarkers selected by PCA and KDM with biomarkers selected by a simple MLR. Full-scale KDM mighty outperformed KDM on 6 covariats, and showed a slightly better result with MLR chosen markers.

	KDM + PCA		**KDM + MLR**	
	Test data	Unhealthy data	Test data	Unhealthy data
(-2,2) in %	8.41	5.08	12.26	7.13
(-4,4) in %	15.99	10.11	25.61	14.71
(-10, 10) in %	38.46	24.63	60.94	36.39
(-15, 15) in %	52.64	36.59	77.41	52.54
mean of residuals	16.71	29.01	9.57	18.77

Table 6: KDM with different set of biomarkers without CA. Results indicates that KDM calculates better results with 15 biomarkers. The best results are marked in green, the worst in red.

5.5 Combining methods

In order to understand impact of additional bias added by calculation a rough prediction of age for Group MLR models (GMLR) several estimations were make by different approaches. All of these calculation was done without an assumption of given CA. Summary of the result for 51-60 age group is shown in the following table 7:

	GMLR + Mahalanobis distance		**GMLR + KDM**		**GMLR + PCA**	
	Test data	Unhealthy data	Test data	Unhealthy data	Test data	Unhealthy data
(-2,2) in %	11.06	11.31	28.87	13.8	20.74	13.16
(-4,4) in %	24.04	22.56	44.37	28.21	34.78	26.52
(-10, 10) in %	52.4	51.18	77.46	62.78	68.23	59.22
(-15, 15) in %	68.87	68.28	94.37	78.44	81.27	75.36
mean of residuals	9.8	10.41	6.02	9.21	8.22	9.97

Table 7: Combing methods on age group 51-60 shows that the best results without known CA is calculated by doing a rough prediction by KDM and then applying GMLR method. The best results are marked in green, the worst in red.

The estimated BA obtained by calculation of an initial prediction of age by KDM then applying GMLR was found most plausible for both tested data sets. Combination of Mahalanobis distance and PCA with GLMR showed approximately the same results on unhealthy population but second method outperformed Mahalanobis approach on healthy population. Results for other age groups slightly differ from presented table but overall combination of KDM and group method showed better prediction power than other covered approaches.

Following table 8 summarize all methods without known CA for all age groups:

	Test data	Unhealthy data
GMLR + Mah.	13.54	13.5
GMLR + KDM	8.5	10.9
GMLR + PCA	10.79	13.8
PCA	9.34	11.83
KDM	9.03	17.22

Table 8: Residuals of estimated BA without CA by several methods. A rough estimation of age by KDM combined with final prediction of BA by GMLR approach showed the best results (green).

6 Selection of Biomarkers

For estimation BA chosen set of biomarkers plays important role. Group method allows to evaluate how different biomarkers varied over time and what blood test has most strong impact on prediction.

	20-30	31-40	41-50	51-60	61-70	71-90	whole	PCA
C-reactive protein	***	**	**				***	
Alkaline phosphatase	.	.	***	**			***	***
Alanine aminotransferase		.	**				***	***
Aspartate aminotransferase				*	.	*	***	***
Gamma-glutamyl transferase					**	*	*	***
Bilirubin	.	*	.	*		.		
Creatinine								***
Cholesterol	**	***	**	***		.	***	
Glucose	.		**	.			***	
Protein total							***	***
Triglycerides	.		**		.			
Urea				**			***	
Uric acid								
Thyroid-stimulating hormone	.				.	*	.	
Free thyroxine		**		.	***	*		

Table 9: Significance of biomarkers for different age group for MLR methods. Three age stage matching biomarkers patterns are marked pink, blue and green.

From the table of importance of each covariant it could be concluded that significance of biomarkers changes over time. C-reactive protein is highly relevant for prediction in 20-50 group models, but for elderly this biomarker is not included in final model. Likewise, cholesterol has impact on prediction power in 20-50 groups, but losses importance of estimation for 61-90 groups. To the contrary, aspartate aminotransferase gains relevancy over age as well as gamma-glutamyl transferase.

Also it could be noticed that overall there are three major patterns in prediction matching three stages of aging: 20-40, 41-60 and 61-70 clusters (marked with colors in the above table). Especially, significant biomarkers for elderly people are approximately the same, there is a change only in the impact on prediction of a chosen covariant. The table shows clear shift in relevance for prediction of BA biomarkers in various age group, hence the whole model fails in prediction.

There is similarity between detached biomarkers of PCA and MLR models. The final models both included alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, gamma-glutamyl transferase and protein total as highly relevant for estimation covariants. It is interesting to point out that KDM applied only on PCA chosen biomarkers showed worse result that the KDM with all markers (mean of residuals increased from 8.99 to 15.72 on train data set without knowing CA). Whereas KDM on markers chosen through simple MLR gave approximately the same results, mean of residuals increased just slightly from 8.99 to 9.11 on train data set without knowing CA.

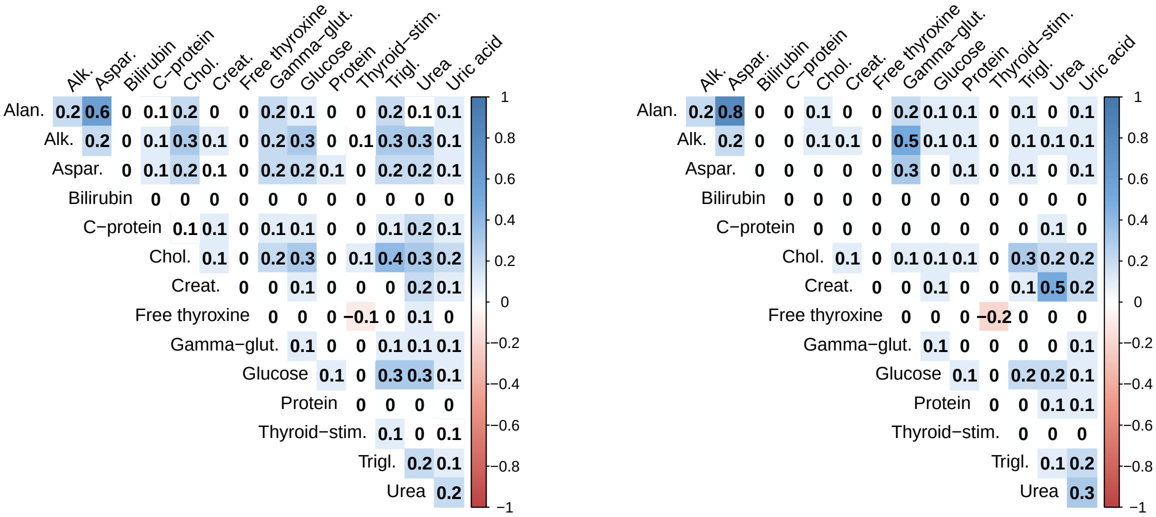

Figure 3: Correlation between biomarkers on healthy (left) and unhealthy (right) data sets

Although correlation slightly differs for healthy and unhealthy population overall correlation matrix shows the same patterns in dependence between biomarkers. The strongest correlation could be noticed between alanine aminotransferase and aspartate aminotransferase (0.6 and 0.8 for healthy and unhealthy sets respectively). For unhealthy data there are stronger correlations than for healthy although some of them are decreasing as in a case with alkaline phosphatase and glucose (correlation dropped from 0.3 to 0.1 in unhealthy population). Generally, many of biomarkers have not got strong correlation or do not have it at all (as bilirubin or thyroid-stimulating hormone) which gave justification to use all of them in prediction for KDM.

7 Discussion

Overall, the best results have been shown with known CA for whole models investigated in this study: KDM, PCA, MLR and group models. The group MLR algorithm performed the best on healthy data and gave the smallest residuals and almost every estimation was in ± 4 interval. Also this age prediction satisfy all initial goals for BA: estimated age for healthy people is close to CA, it is within the life span and represents healthy aging. Thus gives a reason to conclude that estimated age by MLR group model is indeed BA. Therefore, the estimated BA shows deviation from average healthy aging. Results collected on unhealthy population support this statement. BA of unhealthy people deviates from healthy group with noticeable difference inside plausible interval. And the highest deviation between BA and CA is contained in most elderly group. Also PCA method showed a good prediction power with known CA on unhealthy population, but it has inferior estimation on healthy people which could suggest that PCA rather gauges CA than BA. Whereas results of KDM is close to group MLR: method gave rather good predictions for healthy people but an estimation for unhealthy population decreased in comparison which can be interpreted as estimation of BA. Nevertheless, some of results obtained by KDM look unrealistic in a sense of the life span.

The best method of estimation BA without correction for known CA is KDM, but this approach still shows unreliable results as percentage of BA close to CA in healthy population is rather small. Prediction by PCA is close to KDM, however PCA tends to predict CA as it was discussed before so KDM should be considered as best predictor of BA without known CA. Mahalanobis distance showed most reasonable estimation within the life span, but overall was outperformed by PCA and KDM.

Combination of KDM and MLR group method produce the best result of calculating BA without known CA. First, KDM gave a rough prediction of age, and then MLR group method estimates final BA. Although inial evaluation of age add additional residuals to the final result it still outperforms predictions that were done by a single method.

The method of selecting variables is not conclusive as the data set has a lot of missing observation. KDM had shown approximately the same results on full-range of biomarkers and only by 10 included through MLR. Klemera and Doubal suggest that all the biomarkers included in the algorithm be functionally uncorrelated and that factor analysis or PCA or other approach could used to reach this goal [15].

Although the research shows how useful the KDM and MLR group algorithms were to estimate BA, their potential is yet to be realized to the full. The developments in technology and expanding knowledge and observations of the aging process were crucial in identifying and estimating even more sophisticated age-associated biomarkers such as telomere length and oxidative damage measures [20]. Moreover, among the oldest, an amount of physical performance and blood cell count measures proved to be helpful biomarkers of aging [21].

Age-related degradation has a complex of structure and systematic risk of escalation of chronic illnesses over the life span [1]. Based on some evolutionary theories of aging the process of living involves exposure to damaging factors and environment therefore there are intrinsic protective mechanisms to deal with the hazards. Consequently, the damage is accumulated to some degree at different levels over the lifetime which results in the decrease in fertility.

BA estimates can be used to track the curve of damage for a period of time with the help of measures of an individual's level of damage accumulation in the long run. As a result, valid BA estimates can be instrumental in understanding and studying some questions connected with the biology of aging. For example, changes in BA should reflect the changes in the aging rate due to genes or environmental factors which can change the distribution of energy responsible for maintenance and repair or the extent of the damaging causes [22],[23].

Although there are number of useful physiological measures in existence, biomarkers used in BA calculations were limited only to 15 blood biomarkers. Secondly, data set has a lot of not available values which has to be reconstructed and which consequently added a bias to further estimations. Also male population was underrepresented in data and had to be excluded from research. Thirdly, about 60% of the data was centered in ages from 50 up to 90 for healthy people and 65% for unhealthy. Finally, data was collected from participants from South part of Russia and is limited in race and variety of other factors such as environmental soundings, eating habits, or genetic characteristics. For this reason, more work is needed to identify a population that would be the most appropriate from which to generate an equation for BA that represents human aging in general.

MLR group models showed the best result as it captures complexity of aging over time, and paired with Klemera and Doubal method gives the most trustful estimation of biological age as measure of diversity from healthy aging. The development and validation of BA construct is valuable given its impact on our theoretical understanding of the aging process and may facilitate future development of preventative interventions with implications for health and longevity.

References

[1] Yin D, Chen K. The essential mechanisms of aging: Irreparable damage accumulation of biochemical side-reactions. Exp Gerontol. 2005;40:455-465.

[2] Hayflick L. Biological aging is no longer an unsolved problem. Ann N Y Acad Sci. 2007;1100:1Ű13. doi:10.1196/annals.1395.001

[3] Morgan E. Levine Modeling the Rate of Senescence: Can Estimated Biological Age Predict Mortality More Accurately Than Chronological Age? Journals of Gerontology: Biological Sciences Cite journal as: J Gerontol A Biol Sci Med Sci. 2013 June;68(6):667-674

[4] Mitnitski A, Rockwood K (2014) Biological age revisited. J Gerontol A Biol Sci Med Sci 69(3):295-296. 48.

[5] Levine ME (2013) Response to Dr. Mitnitski's and Dr. Rockwood's Letter to the Editor: Biological age revisited. J Gerontol A Biol Sci Med Sci 69A(3):297-298

[6] Jackson SHD, Weale MR, Weale RA (2003) Biological age - What is it and can it be measured? Arch Gerontol Geriatr 36(2):103-115.

[7] Comfort A. Test-battery to measure ageing-rate in man. Lancet.1969;2:1411-1415.

[8] Hollingsworth JW, Hashizume A, Jablon S. Correlations between tests of aging in Hiroshima subjects - an attempt to define "physiologic age". Yale J Biol Med. 1965;38:11-26.

[9] Sprott RL. Biomarkers of aging and disease: introduction and definitions. Exp Gerontol. 2010;45:2Ű4.

[10] Seplaki CL, Goldman N, Glei D, Weinstein M. A comparative analysis of measurement approaches for physiological dysregulation in an older population. Exp Gerontol. 2005;40:438-449.

[11] Krll J, Saxtrup O. On the use of regression analysis for the estimation of human biological age. Biogerontology. 2000;1:363-368.

[12] Takeda H, Inada H, Inoue M, Yoshikawa H, Abe H. Evaluation of biological age and physical age by multiple regression analysis. Med Inform (Lond). 1982;7:221-227.

[13] Nakamura E, Miyao K. A method for identifying biomarkers of aging and constructing an index of biological age in humans. J Gerontol A Biol Sci Med Sci. 2007;62:1096-1105.

[14] MacDonald SW, Dixon RA, Cohen AL, Hazlitt JE. Biological age and 12-year cognitive change in older adults: findings from the Victoria Longitudinal Study. Gerontology. 2004;50:64-81.

[15] Klemera P, Doubal S. A new approach to the concept and computation of biological age. Mech Ageing Dev. 2006;127:240-248.

[16] Nakamura E, Miyao K, Ozeki T. Assessment of biological age by principal component analysis. Mech Ageing Dev. 1988;46:1-18.

[17] Linpei Jia, Weiguang Zhang, Rufu Jia, Hongliang Zhang, Xiangmei Chen. Construction Formula of Biological Age Using the Principal Component Analysis. BioMed Research International, Volume 2016, Article ID 4697017

[18] Cho IH, Park KS, Lim CJ. An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI). Mech Ageing Dev. 2010;131:69-78.

[19] Kamihnikov V.S., Normal range in laboratory medicine. Formulary, 2014, ISBN 978-5-00030-047-3

[20] Johnson TE. Recent results: biomarkers of aging. Exp Gerontol. 2006;41:1243-1246.

[21] Martin-Ruiz C, Jagger C, Kingston A, et al. Assessment of a large panel of candidate biomarkers of ageing in the Newcastle 85+ study. Mech Ageing Dev. 2011;132:496-502.

[22] Kenyon C. A conserved regulatory system for aging. Cell. 2001;105:165-168.

[23] Verbeke P, Fonager J, Clark BF, Rattan SI. Heat shock response and ageing: mechanisms and applications. Cell Biol Int. 2001;25:845-857