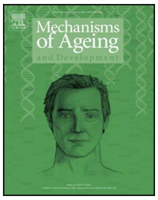
Mechanisms of Ageing and Development 151 (2015) 45–53

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

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



Review

Statistical strategies and stochastic predictive models for the MARK-AGE data

Enrico Giampieri[b,e,*], Daniel Remondini[b,e], Maria Giulia Bacalini[a,b], Paolo Garagnani[a,b], Chiara Pirazzini[a,b], Stella Lukas Yani[a,b,c], Cristina Giuliani[a,b], Giulia Menichetti[b,e], Isabella Zironi[b,e], Claudia Sala[b,e], Miriam Capri[a,b], Claudio Franceschi[a,b], Alexander Bürkle[d], Gastone Castellani[b,e]

[a] *Department of Experimental, Diagnostic and Specialty Medicine, Via S. Giacomo, 12 University of Bologna, Bologna, Italy*
[b] *Interdepartmental Center Galvani "CIG", Via Selmi, 3 University of Bologna, Bologna, Italy*
[c] *Institute for Biomedical Aging Research, University of Innsbruck, Austria*
[d] *Molecular Toxicology Group, Department of Biology, University of Konstanz, 78457 Konstanz, Germany*
[e] *Physics and Astronomy Department, Viale Berti Pichat 6/2, University of Bologna, Bologna, Italy*

ARTICLE INFO

Article history:
Received 2 December 2014
Received in revised form 26 May 2015
Accepted 9 July 2015
Available online 21 July 2015

Keywords:
Biomarkers
Biological age
Chronological age
Statistics models
MARK-AGE

ABSTRACT

MARK-AGE aims at the identification of biomarkers of human aging capable of discriminating between the chronological age and the effective functional status of the organism. To achieve this, given the structure of the collected data, a proper statistical analysis has to be performed, as the structure of the data are non trivial and the number of features under study is near to the number of subjects used, requiring special care to avoid overfitting. Here we described some of the possible strategies suitable for this analysis. We also include a description of the main techniques used, to explain and justify the selected strategies. Among other possibilities, we suggest to model and analyze the data with a three step strategy:

1. Introduction

MARK-AGE (European Study to Establish Biomarkers of Human Aging) aims at the identification of biomarkers of human aging capable of distinguishing between chronological and biological aging, as described thoroughly in this special issue, where the chronological age represent the amount of time from birth and biological age is linked to the underlying aging processes happening in the body.

For this prediction systemic and tissue related parameters are taken into account, not only regarding biological samples (blood, urine, buccal mucosa cells of volunteers), but also with anthropometric, health reported, cognitive, and functional assessments.

To achieve this objective a robust human model with clear-cut assumptions was conceptualized accordingly. Theoretically, the model is based on three different aging rates related to three different populations:

i.) a population representing the "normal" aging or randomly recruited age-stratified individuals from the general population (RASIG), covering the age range 35–74 years;

ii.) a population representing the successful or "decelerated" aging: subjects born from a long-living parent belonging to a family with long living sibling(s) already recruited in the framework of the GEHA -genetic of healthy aging- project (Skytthe et al., 2011). These individuals ("GEHA Offspring" or GO) were recruited together with their spouses or SGO ("Spouses of GEHA Offspring") that represent the best control to evaluate possible lifestyle effects, since sharing the same environmental for many years with their partner;

iii.) a population representing accelerated "segmental" aging, *i.e.* patients with progeroid syndromes (Cockaine, Werner, and Down syndromes), were recruited (see in this issue Capri et al., 2015).

[*] Corresponding author at: Interdepartmental Center Galvani "CIG", Via Selmi, 3 University of Bologna, Bologna, Italy.
E-mail address: enrico.giampieri@unibo.it (E. Giampieri).

http://dx.doi.org/10.1016/j.mad.2015.07.001
0047-6374/© 2015 Published by Elsevier Ireland Ltd.

Expected results are tightly related both to biological meaning and relative power of tracking aging-rate-related changes of each investigated parameter (among hundreds analyzed in MARK-AGE). Thus, the definition and the role of biomarkers as a panel of measurements that captures and quantifies features of the chronological versus biological aging are at the core of MARK-AGE project. The former and the latter are two faces of the same coin, *i.e*. the aging process, but only their combination can empower their predictive value for determining successful or unsuccessful aging. This is a critical step, still far to obtain a general consensus and validation being also closely connected to the fast production of new potential biomarkers with high throughput technology enhancement (Deelen et al., 2013).

Recent published data have highlighted the complexity of the genetics of aging (Capri et al., 2014) and the role of new classes of biomarkers for the detection of differences between chronological and biological aging, such as epigenetics changes, *N*-glycans and metabolites profiles from blood. The recent discovery of a subset of CpG sites that together form an aging clock in blood (Hannum et al., 2013; Weidner et al., 2014) and in a wide range of tissues (Horvath, 2013) has theoretically open the possibility to date the age of these tissues, predicted to be differently aged in the same individual (Cevenini et al., 2008). Further, the methylation levels at specific CpG sites of ELOVL2 and FHL2 genes showed the strongest correlation with age in a population constituted of about 500 donors from newborns until centenarians (Garagnani et al., 2012), suggesting that some mechanisms, like methylation at 5′Cytosine in the nuclear DNA, could better represent the chronological age of humans. Another important class of biological age-markers is represented by microRNAs (miRs) and in particular those miRs able to modulate the inflammatory response with aging or inflammmiRs (Olivieri et al., 2012, 2013). Concomitantly, *N*-glycans from serum blood have received during the last years the attention of many research groups. In particular, the increase of agalactosylated *N*-glycan structures during aging appears to be confirmed in many studies (Dall'Olio et al., 2013) and specific *N*-glycan structures are used for different predictive models (Vanhooren et al., 2007; Krištić et al., 2014). Lastly, metabonomics and lipidomics technologies have recently risen up many blood metabolites, such as phospho/sphingolipids (Collino et al., 2013; Montoliu et al., 2014), that could potentially be putative biological markers and modulators of healthy aging.

Currently, the challenge is the use of *ad hoc* advanced statistical models to elaborate properly the available huge amount of data. MARK-AGE project has faced this challenge exploiting the best fitting and modeling of data, combining both chronological and biological aging markers in the above described "human models".

The database resulting from MARK-AGE project includes both qualitative and quantitative data belonging to several categories:

- Clinical and social data: this category includes mainly qualitative/categorical/ordinal data, such as demographic information (family composition, marital status, education, occupation, housing conditions), lifestyle information (use of tobacco and alcohol, daily activities), health status information (present and past diseases, self-perceived health, number, and type of prescribed drugs) and cognitive/functional status (activities of daily living, Norton scale, STROOP test, 15-picture learning test, ZUNG depression scale).
- Anthropometric data: this category includes quantitative data relative to classical candidate markers of aging, such as waist and hip circumferences, blood pressure and heart rate at rest, lung capacity, near vision, five-times chair standing, and handgrip strength.
- Molecular biomarkers: this category includes a wide range of both qualitative and quantitative data. Qualitative measurements result from the analysis of *APOE* genotype should be managed using statistical tools specific for genetic data. The vast majority of molecular measurements are expressed as quantitative data, both on an infinite scale (for example albumin levels, which are expressed as g/l) or on a finite scale (for example methylation levels, which are expressed as continuous numbers ranging from 0 to 1).

2. Statistical methods

Here we discuss some statistical methods alongside with their potential and limitation for datasets like MARK-AGE. These methods cover all the required steps of the analysis, from data preparation, through feature selection, modeling, biological age assessment, to prediction, and divergence from chronological age.

Selecting the most important passages of the analysis is fundamental for the choice of the proper analysis strategy. Dealing with this kind of problems involves a long sequence of analysis and according to this chain structure, the robustness of the final results is upperly bound by the robustness of the frailest component. Even a spotless analysis can be severely distorted by a single, non accurate step. The steps of our analysis will be the following:

I) Variable pre-selection, to remove non appropriate variables from the analysis.
II) Feature extraction from the raw variables, to obtain more biologically relevant information.
III) Selection of the appropriate method for the analysis, with a proper standardization of the variables to make them follow the assumptions of the method, where possible.
IV) Feature selection, to discriminate the most relevant derived features for the analysis.
V) Parameter estimation, to adapt the model to the data.
VI) Model selection, to select the most appropriate model among all the proposed one (sets of features and parameters); this step will include priors biological knowledge to help in the discrimination of good models from non meaningful ones.
VII) Prediction robustness estimation, to verify how much the proposed model is able to generalize to the population, verifying the biological hypothesis underlying the project.

2.1. Classical test limits

All statistical tests relies on specific hypotheses to be performed. These hypotheses represents the assumptions that the test needs to operate. Most classical tests for example rely on the assumption that the data can be described (or at least approximated) with a Normal distribution. These test are can not be applied on data that do not respect this hypothesis, as the results would be unreliable. These hypotheses, on the other side, allows to include more information in the test, increasing its power. A subset of the classical tests try to use a smaller set of hypotheses, typically removing the request of adherence to a specific distribution in favor of just considering the ranking of the value in the population. These tests are usually referred as non-parametric, as most distributions are described by specific "parameters" (like mean and variance), so the tests that assume specific distribution implicitly are testing the values of these parameters.

It is now known that several biological variables do not respect the requirement for the classical tests, such as the normality (or the possibility to normalize it with an appropriate transformation), even when common transformation (location-scale, logarithm, square root) are used. When the proper distribution of the data is unknown, or can't be described with numerical (ideally continuous or real values), we are forced to use a non parametric test for our hypothesis.

In the MARK-AGE database we have several parameters that could not conform to the requirement for the parametric testing: the numerical continuous data, measurements mainly based on biochemical methods, or those arising from continuous data, such as blood pressure, body weight, etc. can show significant deviations from the Gaussian distribution and are not immediately ascribable to any known distribution.

This problem is common in biomedical data analysis as most biological data, such as cell's volume, proteins molecular weights, gene lengths, and other biochemical variables, are not distributed in a Normal way (they are not following a Normal distribution, or equivalently a Gaussian distribution). This deviation from Normal distribution was historically defined as an exception, while nowadays is considered very common (Zhang and Popp, 1994; Koch, 1966; Bahr et al., 1987; Russo et al., 2012, 2011). A simple intuition of this behavior can be obtained by observing that the majority of these measurements are strictly positive and with an average close to zero, and thus can not be described as a Gaussian distribution as this would give a definite probability also to negative values. Metabolites, for example, usually have a distribution closer to an Exponential distribution, with a strong peak on the value of zero, and being a measure of concentration can of course be only positive.

2.2. Non parametric correlations

The most widely known type of correlation is the Pearson's product-moment correlation coefficient, usually referred as the *Pearson's r*. This value measure the linear dependence of the parameters, and is comprised between 1 and −1.

This quantity has the limit of losing any information about non-linear relationships between the variables, and requires the data to be numeral. To circumvent these problems a common solution is to use a non-parametric correlations method. These methods usually require only information about the ordering of the data, working both for numerical and ordinal variables. Working only on the ranking of the data, these methods allow to include non-linear (but still monotonic) effects in account.

The two most common non-parametric tests are the Kendall's Tau and the Spearman's Rho. The former (endall's Tau) generated the couples of observed ranks of each variables, then confront all the couples to see if they are concordant (both the value of the couple are greater or smaller than the other couple); the statistic of the test is the expected number of concordant couple when there is no relationship between the variables. The latter (Spearman's Rho) is defined as the Pearson's correlation between the ranking of the two variables.

The Kendall's Tau is usually more conservative, so less sensible to errors in the data and less biased, but has less statistical power and takes more time to compute in presence of large datasets like MARKAGE, as the time it requires for the computation grows with the square of the number of observations. These two analyses are not equivalent, and the most appropriate one should be chosen depending on the goal of the analysis (Xu et al., 2013).

2.3. Robust regression for truncated data

With the term "robust regression" we refer to a large family of methods for estimating the linear relationship between two or more variables. The regression methods are in general performed by searching the combination of parameters that minimize the sum of all the prediction errors. The classical linear regression use as an estimate of the prediction error the square of the difference between the predicted value and the observed one. This measure correspond to hypothesize that all the errors have a Normal distribution. Evaluating the errors with this method can be very fragile, as any anomaly of the distribution of the errors can heavily affect the regression results. This can be due, for example, to outliers or uneven differences that may happen when the dependent variables has hard limit like 0 for the concentration of molecules. To circumvent this problem several methods have been proposed during the years. Most of them modify the underlying error function in a way that allows to be less affected by outliers. Some robust regression use a Student's *T* distribution, other a truncated exponential, and so on. This corresponds to different hypothesis on the underlying distribution of the estimation error. Most of the observed variables in the database does not conform to the Gaussian hypothesis; hence; using safe assumptions on the error distribution is the best approach without removing outliers from the dataset, a practice that is frowned upon as it is completely arbitrary and can lead to unpredicted biases in the model Fig. 1.

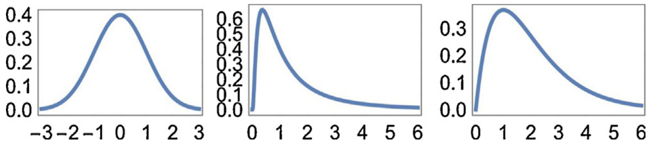


Fig. 1. Comparison between parametric distributions, on the left a Gaussian standardized distribution with zero mean and unitary standard deviation and two non Gaussian distributions. In the center is a LogNormal distribution with location parameter equal to zero and scale parameter equal to 1. On the right a Gamma distribution with shape parameter equal to 2 and scale parameter equal to 1.

The regression needs to consider also an important feature of the data: when the regression is done with the age as a function of a set of regressors, the dependent variable (the age) does not respect the assumptions of the ordinary regressions methods, that is to be randomly sampled on the basis of the independent variable. Having a precise limit on the value of the dependent variable transform this problem in one where there are missing not at random (MNAR) data. This can lead to a severe loss of power of the predictor, as the missing data are the one in the extreme position of the spectrum, the most important for the prediction. This would also lead to strong biases in the estimation, as the underlying data are biased. The effects of this bias can be seen in Fig. 2, where a toy model is used to show the severity of the distortion. Methods to deal with the missing data are well known (Schafer and Graham, 2002; Ibrahim et al., 2005), and include adjusted maximum likelihood, multiple imputation, and full Bayesian methods. All of these methods are based on a modelization of the mechanism that can lead to the missing data, and can significantly decrease the error in the regression (Mason et al., 2012).

2.4. Multiplex network approaches

Since, in the MARK-AGE project, we are interested in multivariate analyses that keep into account the relations between variables (*e.g.* biochemical, behavioral) and the relations between samples (*e.g.* gender, age group, health history, lifestyle, nationality), a network-based approach can be fully exploited for such purposes. Very recently the concept of network, namely a set of elements (nodes) that share specific relationships between each other (links), has been extended to multiplex networks (Boccaletti et al., 2014; Menichetti et al., 2014a,b; Castellani et al., 2014). In a multiplex network the same nodes are represented in each layer, corresponding to different types of relations or interactions. A possible analysis involves a multiplex in which the nodes are the measured parameters (*e.g.* biochemicals), and the links in each layer are defined by similarity distances, evaluated through correlations or similar

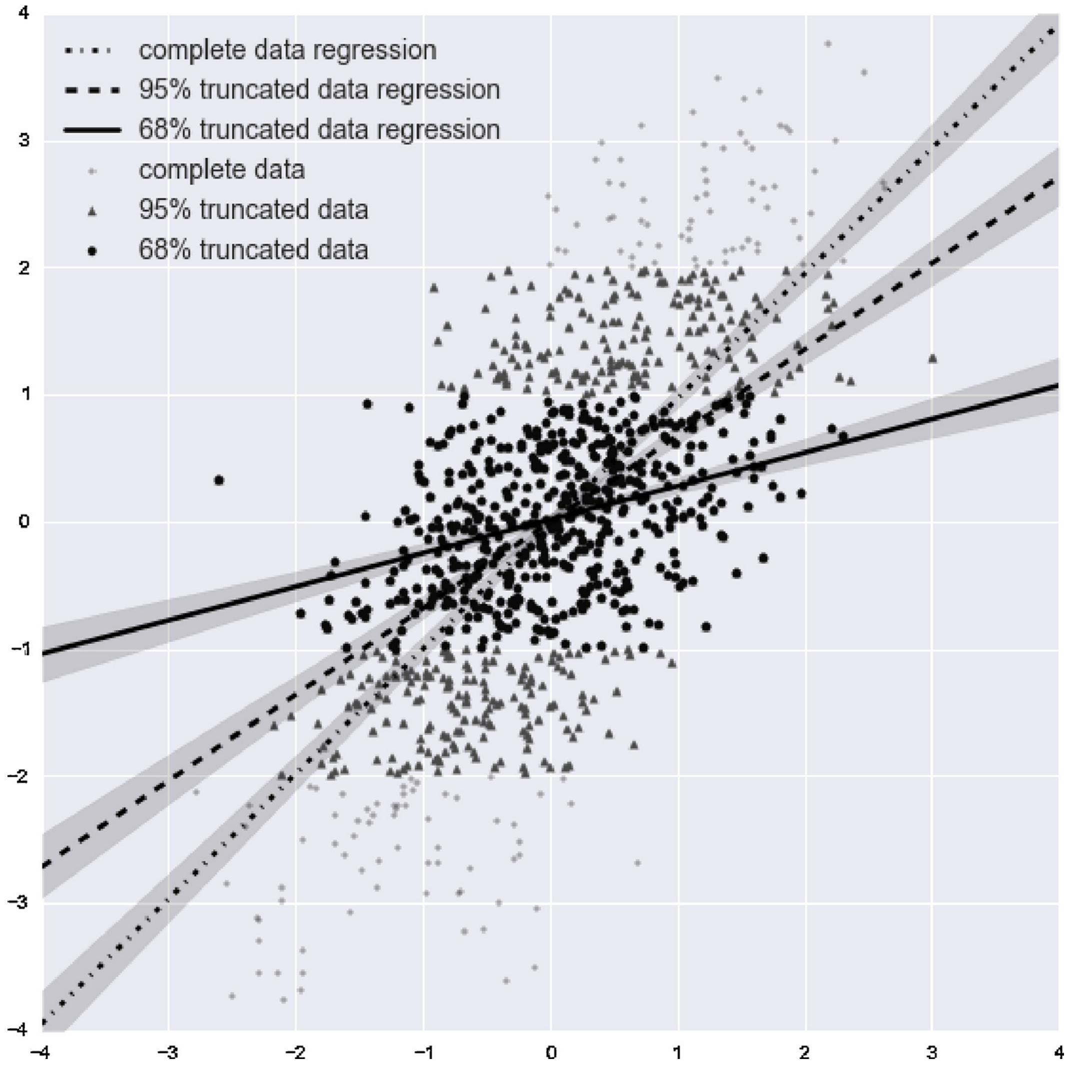


Fig. 2. The effect of the truncation of data on the predicted variable, depending on a different amount of truncated datas. This figure depict a trivial model, where both variables are normally distributed, with the dependent one equal to the first one with the addition of white noise: x is distributed as $N(0, 1)$ and y as $x + N(0, 1)$. It is clear how the effect of the truncation is evident on the regression even when only the extreme 5% is excluded. When 32% of the data from the extreme is removed (one standard deviation from the mean), the regression line that should be close to the bisecting line is almost flat. While the true regression has a slope of 1, the estimated slope from the truncated data is 0.27. Even using 95% of the data, the regression line is 0.67.

measurements (Menichetti et al., 2014a,b). Each layer could be associated with a particular population subgroup, *e.g.* GO, SGO, RASIG, etc. (see Fig. 3) In this way we can characterize the parameters and the relations that are common to all the groups in which we divide the dataset, by means of specific multiplex observables. An example is shown in (Menichetti et al., 2014a,b) in which we have characterized the relation between node connectivity degree (the sum of the link for a node) and its strength (the sum of the weights for a node) for the set of links shared between more than one layer, or specific for each layer.

This multiplex approach allows also to define more statistical observables useful to evaluate the quality of our models. One of the possible measurements is the Network Entropy (Menichetti et al., 2015), that is related to the selectivity the observed characteristics in the real dataset, *i.e.* it estimates the number of structures (networks or multiplex networks) compatible with some given real features.

These network approaches can be used in the context of our analysis strategy to include biological relevant information on the relationships between variables and to quantify the meaningfulness of a set of variables Fig. 4.

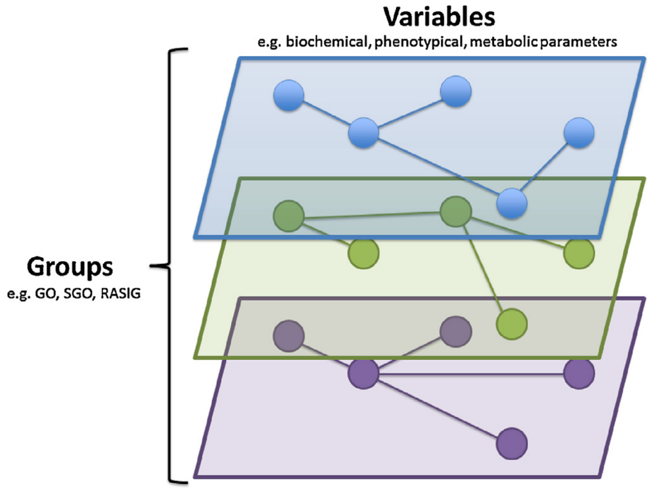


Fig. 3. A Multiplex is a set of networks or layers with common nodes (the observed variables). Each layer corresponds to a given population subgroup. These links are generated by correlation and causal relationships. The relationship between the layers can be used to improve the estimation of the relationship between the nodes.

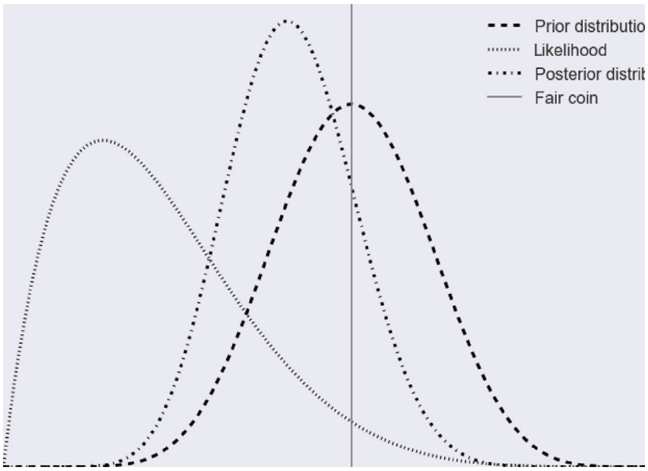


Fig. 4. Bayesian investigation of the fairness of a coin. The investigator start with a mild prior hypothesis over the frequency of the head, as being balanced, but with a margin of uncertainty. The observations are 2 heads and 7 tails, that correspond to the depicted likelihood. The final knowledge about the coin is represented by the posterior distribution, given by the normalized product of the prior and the likelihood. This posterior distribution show that after these observations the confidence about the fairness of the coin is reduced, but it's still a plausible hypothesis.

2.5. Training, testing and validation

Statistical analysis can be used to reach two major, distinct, objectives: explanation and prediction (Shmueli, 2010). Most scientific research focus on explanatory analysis, that tries to explain which is the causal relationship among variables. In this case our interest is on the prediction of the biological age of a subject, and thus the statistical approach should be different from the standard for explanatory analysis.

Prediction analysis does not concern so much on causation relationships as much on reliable correlations (Domingos, 2012). This is easily quantified for simply linear models by the adjusted r-squared statistics, that conveys the predicted r-squared that would be observed on a new sample from the same process. It can be considered a measure of the true correlation after removing the effects of the overfitting.

This kind of analytical estimates are not possible on general non-linear models, or non parametric one. The best solution so far to this problem is to split the data in two subsets (usually unequal in size), fit the statistical model on a set (called train subset) and verifying how better the model perform on new, unobserved data (called the test subset). This procedure is usually repeated several times with different splitting of the data, to make the estimation more robust and not dependent on a specific split.

The division of the data in train and test can be done in several ways. One of the most used is the k-fold cross validation, where the data are partitioned in k different subset, of which one is used as test and the other as train, repeating the procedure until all the subsets has been used. An extreme version of this procedure is the Leave One Out method, where the test correspond to a single observation and the train to all the others. This kind of approach get the most out of the data, but can easily become unfeasible as the number of observation increase.

A similar approach can be used in the context of model selection, and is referred as train-test-validation split. With model selection we mean choosing one model among several others; an example of this may be the selection of the regressors to use in the linear model, or the use of parametric or nonparametric models. For the sake of the explanation, let the problem be the selection of the variables to be used as regressors in a linear model. We want to use the train test to fit each possible model, and then use the test set to choose the best performing model. The resulting estimate of the prediction power is biased upward, as we chose the best performing model among the set of tested ones. To avoid this bias, we perform a second estimation of the expected error using the third set, called validation set, that is separate both from the test and the train one. A typical division of the data would be 60% for train, 20% for validation, and 20% for test.

Using these strategies we can evaluate the expected error for any kind of model, independently from how complicated it is. These methods include a form of automatic Occam Razor, as the greater the number of parameters included, the bigger the overfitting of the model and thus greater the error on the test set.

A proper procedure of model selection should also consider the clinical practice, as all these parameters requires exams that are expensive in terms of time and money. Two variables with the same statistical properties can be very different in terms of cost of measurements, and thus the selection should be informed of the relative cost of the two. Often variables are measured in batches, and thus removing one variable from the analysis would not save any cost and it is worth including even if the improvement to the model is small.

2.6. Feature selection, feature extraction and the curse of dimensionality

In a biological analysis, it is often assumed that higher number of features (the variables used) lead to a better model for the data. This idea is wrong, especially when the included features are not directly linked to the problem under study.

The first, and by far the most important problem, is known with the name of curse of dimensionality. This problem is due to the number of examples needed to sufficiently sample the possible cases describable using a set of variables. If we consider only two variables, each of which can only assume two value, we have four possible combinations. If we assume that we need more or less ten subjects for each case for the statistical methods to properly work, we would need around forty subjects. If we are using six different variable, we may have around 64 combination and thus 640 subject for minimal coverage. With ten variables we would need around 10,000 subjects to maintain the same power as before.

This, and other effects, make dangerous increasing indiscriminately the number of features included, as we will not have enough subject to discriminate easily which variable are relevant and which are not. This is even worst when few features are similar between each other. In this case the algorithm may have an hard time discriminating between them, and reach an undetermined state. In this kind of situation is more proficuous to summarize all these similar and related features in a single one. This would at the same time reduce the number of features used and improve the behavior of most predictors. This kind of procedure is referred as feature extraction, and is a key procedure in high dimensional data analysis. The term feature extraction refers to the process of generation of new features from the existing one, that are then usually replaced by the new ones. One of the easiest and most common kind of feature extraction is the principal component analysis (usually referred by its acronym as PCA), a non-supervisioned method that generate new features by grouping features that have strong correlation between them. This method is best suited to work with features that have a similar scaling, as the variance of the individual feature becomes a weighting factor in the process.

This makes the PCA appropriate to refine raw measurements of a common concept, for example methylation levels (where several probes are used for each gene) and body fat composition (where most analysis have region with overlaps).

2.7. Bayesian estimation

The Bayesian statistic is an alternative approach to the whole standard set of analysis commonly done (Kruschke, 2010). The whole approach is based on a single way of considering the analysis that is implemented differently depending on the case. Most of the standard statistics (referred as frequentist statistic) can be seen as a special case of this more general approach.

The main ideas of the Bayesian statistics are easy to grasps, and allows to answer several question in a direct and intuitive way. Firstly, the analyst create a model describing how the data can be generated. This model contains a probabilistic description of the phenomenon in terms of certain parameters. A model is composed by two parts: the likelihood function, that describes how likely is to have an observation value given the value of the parameters, and the priori distributions that express the plausible value of the unknown parameters. The Bayesian analysis update the plausibility of the parameters after the observations. All the classical tests can be rewritten in terms of this kind of parameter estimation.

If one want to study the plausibility of nefarious development of an illness under a new treatment compared to a placebo, one would estimate the frequency of nefarious development in each of the two cases and to check that these frequencies do in fact differ.

Firstly, one need to describe the plausible values of frequencies before the observations. This knowledge is represented as a probability distribution over the possible values that the frequency can have. In the Bayesian statistics probability is not considered an objective entity, but rather an expression of our knowledge about the process. This distribution, representing the previous knowledge, is called the prior distribution of the parameter. The choice of the priors is one of the most delicate parts of the Bayesian modeling, and is good practice to test the model with different priors, ranging from very wide distribution (or a completely flat one) that does not express any preference for any possible value of the parameter, to more committed ones that can include previous observations, accepted knowledge in the field or the evaluation of experts.

The likelihood function describe the plausibility of observing a certain number of development in the population under study given a certain frequency, using the Binomial distribution. This represents how the researcher think that the data are generated. The choice of the likelihood function represent the model, and determine the parameters under consideration. Different models can imply very different likelihood functions, with very different statistical properties. It is then always recommended to not limit the analysis to a single model, but take different ones into consideration.

The final results of a Bayesian analysis is the posterior distribution of the parameters, that encodes all the available information about it. In the real practice is necessary take a decision based not on the distribution, but rather on the best possible estimate of the parameter based on the information. To do this is necessary to synthesize the information from the posteriori distribution in a single point estimate. This can be done by choosing a risk function that describe the penalty incurring in choosing a value of the parameter when a different value is the true one. The final estimation is the value that minimize the expected risk, weighted over the plausibility of the other values as indicated by the posteriori distribution. Different risk function will generate different estimates.

2.8. Prediction using expert mixtures

A statistical model used to predict the outcome of an experiment or an intervention is often referred as an Expert System. Each possible model is a partial representation of the reality, that include certain elements in the prediction neglecting others, and use different approximation of the phenomenon. Linear regression models hypothesize that all effects are linear and independent, neural networks approximate the response as the superposition of binary signals and so on.

To improve the performances of the predictions a common strategy is to improve the model used, including more and more details in it, reducing the approximations used trying to get closer to the truth. This approach is useful with very simple models, but have a very low payoff for more complicated ones, especially when considering complicated effects like human aging.

Research have shown (Masoudnia and Ebrahimpour, 2014) that a more effective approach is to pool several simple models together, weighting the prediction based on their accuracy. Combining several of these independent models in a higher level one, the performance boost can be significative. In the case of biological age prediction, each model returns an estimation of the real biological age. Each one of this prediction will be imprecise, but if the model is good, should also be unbiased and independent from the others. A better predictor can be generated by taking the weighted average of the estimations.

3. Proposed statistical approach

Given the volume of data and the complexity of the research question, several different predictive modeling approaches will have to be used on MARK-AGE data.

To increase the replication power of the analysis, this should be focused on robust statistical methods, like nonparametric ones, to avoid that outliers, due to exceptional situation or human error, influence too much the final results (see Sections 2.1 and 2.2). The results of these feature selections should be tested with a train-test protocol to assess the robustness and replicability of the method (see Section 2.5). In particular, the tested signatures should be expected to be able to separate young and senile people with an high degree of accuracy to be meaningful.

The steps necessary for a proper modeling of the MARK-AGE data are thus the following:

1. cleaning of the dataset, selecting the relevant features and generating new biologically relevant one;
2. divide the dataset in train, test, and validation sets, balancing for gender, country, and other covariates;
3. divide the features in group of interest: clinical features, biochemical, genetic, and epigenetic. These groups will be the basis for the multi-expert evaluation of the biological age;
4. for each one of these sets generating all the models that use these features (up to two or three features) to predict the chronological age. These models should use prior biological knowledge where available through Bayesian modeling, and should be evaluated with models that can correct for the truncation effect of the dependent variable;
5. Combine all of these models in a multiplex network of the different population groups (RASIG, GO, SGO, etc.) and use the information between these layers to choose a reliable model (only the information common in most layers should be considered real and useful);
6. Generate higher level model for each feature set, and combine these in a general, composable model, robust to data collection issues.

3.1. Data correction, feature extraction and dimensionality reduction

The database should first checked for any irregularity of data, both for compliance to the standard range of values and for differences between nationality, study groups, and gender. Due to

Statistical softwares

Statistical analysis can be seen on two different levels: high level decision about the analysis workflow, and number crunching of the data. Various software has been created to lift the burden of the number crunching as much as possible from the analyst, and automatizing the most common higher level procedures. Different systems have different approaches to this problem, but we can roughly distinguish them into two broad categories: low level, programming suites, and high-level automated interfaces.

The authors sustains that data analyst should focus more on programming environments rather than higher level interfaces. This will increase the marginal cost of a new analysis, but in the long term it will force the analyst to keep a precise and reproducible trace of the analysis done. It would also help other analyst to check for the correctness of the operations performed, and ease the long distance collaborations by code sharing.

The most common programming environments are:

- **R** (www.r-project.org)
- **Python** (www.python.org)
- **SAS** (www.sas.com)

There are other suitable environments, such as MATLAB and Mathematica, but they are less common choices among data analysts.

Two main distinctions can be done between these suites.

Firstly, Both R and Python are free and open source, while SAS is a commercial product. This means that SAS is not free to use, but gives more support to paying users and is more standardized. R and Python, on the other way, rely more on their community for support, both of which are very active and available to help, but there is no guarantee that there will be someone with the competence to solve the user's problem. On the upside they allow a more rich and fast development of solution, taking advantage of the concurrent nature of the development from multiple sources. These libraries are often implemented by the original authors of the methods, and are quickly available, but managing and installing these libraries are left to the user (even if improvements are made each year to simplify the library managements).

Secondly, SAS and R are Domain Specific Languages for data analysis, while Python is a general programming language. This makes R and SAS slightly more comfortable for interactive analysis, with a richer set of dedicated models. Python, on the contrary, has a less rich set of bleeding edge models, but it eases the incorporation of the analysis in more complicated pipelines, such as data management and mangling, output production and so on.

the strong differences found in gender during aging, all the analysis should be performed separately between the two groups. This will avoid to incur in the Simpson's paradox, where an imbalance between a covariate variable can lead to wrong, even paradoxical results.

The first step is data cleaning: clinical studies can have a relatively high percentage of missing data in some relevant biological and anthropometric variables, and most analysis techniques have problem dealing with missing values. Removing all the patients with missing values is impossible as, given the amount of parameters, it will mean to drop the whole database. One solution is to make a selection of a subset of variables and remove all the patients that are missing values in those variables. Multiple signature selection based on different variable subsets should be then tried. These analysis should be performed on both subsets of high coverage samples and extended samples sets with data imputation based on external informations.

Where meaningful, new derived variables not directly encoded in the database should be generated, like factorized expression of multiple methylation sites with high correlation *via* dimensionality reduction methods (see Section 2.6), and combination of anthropometric values *via* medical-driven reasoning, like the average strength of the dominant hand. This procedure increases the robustness of the estimation of correlation coefficients and avoid discrimination problem due to highly correlated variables. It also increases the interpretability of the results of the signature selection.

Several related variables, for example the expression of neighbors methylation probes, should be grouped with a hierarchical clustering and reduced with a Principal Component Analysis to a single value (see Section 2.6). This is supposed to represent more biologically relevant values, removing spurious correlation due to variability. This approach should be verified with a train test validation and is expected to yield a significant improvement in robustness and reliability of the measure.

Several relevant anthropometric and biochemical quantities should be analyzed with robust linear regression and logistic regression to assess their interdependence and choosing which variables should be used as covariate in the following analysis.

3.2. Dataset split in train-test and validation

As noted in Section 2.5, dividing the dataset allows to have a more solid estimation of the predictive ability of the model under analysis. The selection should be done in a random way, but it should maintain a balance between the most relevant variables of the dataset, like age, smoking abit, and country. This will ensure that the test will not be altered by unbalancing between these groups.

The various population groups (RASIG, GO, SGO, etc.) should be kept completely separate: a good predictor model is expected to perform properly on all these groups, but with a relevant bias toward lower ages in the GO group and toward higher ages in the progeric-like groups. This is a key feature of the models: the base biological assumption is that part of the discrepancy between the predicted and observed age is due to the underlying biological age, but without a test population it would be impossible to discriminate this from a bad predictive model.

3.3. Feature grouping

There are two main reasons for feature grouping: differential aging and clinical availability. Firstly, it is suspected that aging is not a single process, but rather a multi-spectrum process that can develop with different speed in different part of the body. A good predictor should not try to predict all of them at once, as it would mean losing statistical power, mixing effects potentially independent one from the other, averaging them in a null effect.

Secondly, different measurements require different set of exams, not always available or convenient. Having a single model that use all the possible information available in the best case scenario where all the exams results are available may not be useful in practice.

3.4. Age signature

To generate the feature signature for the prediction of the biological age single, couple, and triple correlations among all the variables should be performed, ranking them on the basis of robust, non-parametric regression methods (see Section 2.3), then the validity should be checked with a train-test setup, including in the regression and the test the appropriate considerations for the truncated data (see Section 2.3). On each subset a full sub-matrix should be generated, with only the used regressors, to avoid imputation and the removal of subjects if not absolutely necessary. Each signature should be tested both for the regression ability and for the capability of discrimination between young and elderly subjects in the control group. All these correlations should separated by gender and the country of origin should be used as covariate.

Having obtained a complete set of small signatures, a network based method can be used to generate a bigger signature. The informations contained in the relationships network can be used to weight the performances of each subset, and to combine them into biologically relevant supersets.

Different signatures based on biochemical and anthropometric variables should be generated, and on a superset of the these. The predicted biological ages obtained with these variable sets should be confronted to understand the amount of agreement between these approaches. These predictions should also be tested using the GEHA Offspring (GO) group, with their spouses (SGO), and Progeric groups, as under the biological hypothesis of the project these groups should behave differently in a reliable way if the predicted age is representative of the underlying biological age. This biological comparison allows to remove several spurious correlations from the prediction, especially once compared with subjects from similar environment.

A reliable method to confront the predictors of different population is with a reasonable set of initial assumptions. These can be given by direct estimation and update using the Bayesian methods (See section 2.7) and by using the information present in other databases through multiplex methods (see Section 2.4).

3.5. Multiplex model selection

The final model for each feature set has to be generated from a set of well performing submodels. A good tradeoff between exploring the model space and time requirement is the ensemble of all the couples of features (with the appropriate covariates included). This allows us to generate a network between all the features, where the features are the nodes and the quality of the prediction of the models are the links. Generating different networks for the population groups allows to generate a multiplex.

We can use this multiplex to select the most relevant features set by weighting two parameters. Firstly, one link can be considered relevant if it is present in most layers with the same value, excluding high predictive results due only to random outcomes. Secondly, a subset of features can be considered relevant when the clustering of these features is higher than average. This means that each couple of features in the selected group is relevant by itself; this selection allows to limit the amount of correlated features that other methods could risk to include in the selection. This selected group should then be checked against the validation group of RASIGs, to assess its real predictive power.

3.6. Prediction with mixed models

Having different predictors that comes from different ranges of exams will allow to combine these prediction as an expert set. The coherence between their prediction will allow a better prediction, that would be more robust not only to the intrinsic distortion of each method, but also to the lack of information due to the impossibility (due to cost or health risk) of performing certain tests (see Section 2.8).

These analysis should also test for nonlinear behavior in the trends, that can arise from survivors biases, where the less healthy subjects get removed from the viable study subjects due to health issues. This can generate a reduction in the sensitivity of the method in older ages, that should be kept in consideration during the model selection and final testing.

4. Conclusions

MARK-AGE is an ambitious projects, and to reach the proposed goals a huge ensemble of data has been collected. These data have a complex and heterogeneous structure, being the composition of clinical, social, anthropometric, and biochemical data. These data cannot be described by any of the standard distribution; in the case of ordinal data, it is often improper to consider an analytical distribution at all. Given the intrinsic difficulties with human data collection, both the database and the future observation on the subjects will contains some non-observed values as well as wrong values derived from errors in the data entry. The chosen strategy should be designed to include a proper statistical approach that is robust enough to this kind of errors in the data. This leads to a preference for robust and nonparametric methods, that are more conservative when facing non ordinary data structures. This robustness is crucial when dealing with high dimensional data, as the sampling of the variable space will be uneven and the variability overwhelming.

Being the objective of the MARK-AGE project the prediction for the individual of its aging status, a great deal of attention should be given to the prediction capability and robustness of the prediction, employing techniques apt to reduce as possible the prediction error.

It is also worth noting that the MARK-AGE predictor hopefully will find application in several fields of the public health management. When the policy maker will be called to take a decision on the base of the prediction of the MARK-AGE model, it will have to take into consideration the effects and the risk associated with this estimation of the biological age. It is thus necessary for this model to be as honest as possible and give not a single point estimation of the biological age, but rather a whole posterior distribution of the plausible values of the biological age. This will allow the users of this model to perform an adequate decision based on a risk evaluation that encompass the whole prediction, and not only a part of it.

To realize a prediction as unbiased and informed as possible, it is also necessary for the analytic strategy to allow the previous biological knowledge as an explicit information, both as prior in the parameter estimation, plausible expected distribution for the missing values and biologically informed model selection.

The analysis strategy that we propose should be able to cope with all these challenges, and outperform simpler, more naive approaches, in term of predictive accuracy and robustness of the results. Dividing the model in several sub-models will allow us to improve the prediction robustness to missing data, but also to allow to use this model in situation where only partial data are available. The division between submodels it is also important in the clinical practice, where the cost-benefit ratio of performing a test can be crucial in the model selection. Our goal should be not only to develop a precise model, but one that can and will be used in the clinical practice.

Acknowledgements

We wish to thank the European Commission for financial support through the FP7 large-scale integrating project “European Study to Establish Biomarkers of Human aging” (MARK-AGE; grant agreement no.: 200880)

References

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C.I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., Zanin, M., 2014. The structure and dynamics of multilayer networks. Phys. Rep vol. 544 (1 (July)), 1–122.

Capri, M., Santoro, A., Garagnani, P., Bacalini, M.G., Pirazzini, C., Olivieri, F., Procopio, A., Salvioli, S., Franceschi, C., 2014. Genes of human longevity: an endless quest? Curr. Vasc. Pharmacol. (in press).

Capri, M., Moreno-Villanueva, M., Zoli, M., Scurti, M., Pini, E., Cevenini, E., Borelli, V., Schön, C., Fiegl, S., de Craene, A.J., Hervonnen, A., Bernhardtd, J., Sikora, E., Gonos, E., Toussaint, O., Grubeck-Loebenstein, B., Bürkle, A., Franceschi, C., 2015. MARK-AGE population: from the human model to new insights. Mech. Aging Dev., April (2) http://dx.doi.org/10.1016/j.mad.2015.03.010, pii: S0047-6374(15)00035-4.

Castellani, G., Remondini, D., Intrator, N., 2014. Systems biology and brain activity in neuronal pathways by smart device and advanced signal processing. Front. Genet. 5 (253).

Cevenini, E., Invidia, L., Lescai, F., Salvioli, S., Tieri, P., Castellani, G., Franceschi, C., 2008. Human models of aging and longevity. Expert. Opin. Biol. Ther. 8 (9), 1393–1405.

Collino, S., Montoliu, I., Martin, F.P., Scherer, M., Mari, D., Salvioli, S., Bucci, L., Ostan, R., Monti, D., Biagi, E., Brigidi, P., Franceschi, C., Rezzi, S., 2013. Metabolic signatures of extreme longevity in northern Italian centenarians reveal a complex remodeling of lipids, amino acids, and gut microbiota metabolism. PLoS One 8 (3), e56564.

Dall'Olio, F., Vanhooren, V., Chen, C.C., Slagboom, P.E., Wuhrer, M., Franceschi, C., 2013. *N*-glycomic biomarkers of biological aging and longevity: a link with inflammaging. Ageing Res. Rev. 12 (2), 685–968.

Deelen, J., Beekman, M., Capri, M., Franceschi, C., Slagboom, P.E., 2013. Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. Bioessays 35 (4), 386–396.

Domingos, P., 2012. A Few Useful Things to Know about Machine Learning. Commun. ACM Volume 55 (10 (October)), 78–87.

Garagnani, P., Bacalini, M.G., Pirazzini, C., Gori, D., Giuliani, C., Mari, D., Di Blasio, A.M., Gentilini, D., Vitale, G., Collino, S., Rezzi, S., Castellani, G., Capri, M., Salvioli, S., Franceschi, C, 2012. Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell 11 (6), 1132–1134.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., Zhang, K., 2013. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell 49, 359–367.

Horvath, S., 2013. DNA methylation age of human tissues and cell types. Genome Biol. 14 (10), R115.

Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R., Herring, A.H., 2005. Missing-data methods for generalized linear models: a comparative review. J. Am. Stat. Assoc. 100 (469), 332–346.

Koch, A.L., 1966. The logarithm in biology. 1. Mechanisms generating the log-normal distribution exactly. J. Theor. Biol. 12 (2 (November)), 276–290.

Krištić, J., Vučković, F., Menni, C., Klarić, L., Keser, T., Beceheli, I., Pučić-Baković, M., Novokmet, M., Mangino, M., Thaqi, K., Rudan, P., Novokmet, N., Sarac, J., Missoni, S., Kolčić, I., Polašek, O., Rudan, I., Campbell, H., Hayward, C., Aulchenko, Y., Valdes, A., Wilson, J.F., Gornik, O., Primorac, D., Zoldoš, V., Spector, T., Lauc, G., 2014. Glycans are a novel biomarker of chronological and biological ages. J. Gerontol. A Biol. Sci. Med. Sci. 69 (7), 779–789.

Kruschke J.K., Doing bayesian data analysis: a Tutorial with R and BUGS (10 November 2010).

Mason, Alexina, Richardson, Sylvia, Plewis, Ian, Best, Nicky, 2012. Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods. J. Off. Stat. Vol. 28 (2), 279–302.

Masoudnia, S., Ebrahimpour, R., 2014. Mixture of experts: a literature survey. Artificial Intelligence Review Vol. 42 (2 (August)), 275–293.

Menichetti, G., Remondini, D., Panzarasa, P., Mondragón, R.J., Bianconi, G., 2014a. Weighted multiplex networks. PLoS One vol. 9 (6 (June)), e97857.

Menichetti, G., Remondini, D., Bianconi, G., 2014b. Correlations between weights and overlap in ensembles of weighted multiplex networks. Phys. Rev. E vol. 90 (6 (December)), 062817.

Menichetti, G., Bianconi, G., Castellani, G., Giampieri, E., Remondini, D., 2015. Multiscale characterization of aging and cancer progression by a novel Network Entropy measure. Mol. BioSyst., http://dx.doi.org/10.1039/C5MB00143A

Montoliu, I., Scherer, M., Beguelin, F., DaSilva, L., Mari, D., Salvioli, S., Martin, F.P., Capri, M., Bucci, L., Ostan, R., Garagnani, P., Monti, D., Biagi, E., Brigidi, P., Kussmann, M., Rezzi, S., Franceschi, C., Collino, S., 2014. Serum profiling of healthy aging identifies phospho- and sphingolipid species as markers of human longevity. Aging (Albany NY) 6 (1), 9–25.

Olivieri, F., Spazzafumo, L., Santini, G., Lazzarini, R., Albertini, M.C., Rippo, M.R., Galeazzi, R., Abbatecola, A.M., Marcheselli, F., Monti, D., Ostan, R., Cevenini, E., Antonicelli, R., Franceschi, C., Procopio, A.D., 2012. Age-related differences in the expression of circulating microRNAs: miR-21 as a new circulating marker of inflammaging. Mech. Aging Dev. 133 (11–12), 675–685.

Olivieri, F., Rippo, M.R., Monsurrò, V., Salvioli, S., Capri, M., Procopio, A.D., Franceschi, C., 2013. MicroRNAs linking inflamm-aging, cellular senescence and cancer. Aging Res. Rev. 12 (4), 1056–1068.

Russo, D., Bombardi, C., Castellani, G., Chiocchetti, R., 2011. Characterization of spinal ganglion neurons in horse (Equus caballus). A morphometric, neurochemical and tracing study. Neuroscience 176 (10 (March)), 53–71.

Russo, D., Castellani, G., Chiocchetti, R., 2012. Expression of high-molecular-mass neurofilament protein in horse (Equus caballus) spinal ganglion neurons. Microsc. Res. Tech. 75 (5 (May)), 626–637.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychol. Methods 7, 147–177.

Shmueli, G., 2010. To Explain or to Predict? Stat. Sci. 25 (3), 289–310, http://dx.doi.org/10.1214/10-STS330 http://projecteuclid.org/euclid.ss/1294167961

Skytthe, A., Valensin, S., Jeune, B., Cevenini, E., Balard, F., Beekman, M., Bezrukov, V., Blanche, H., Bolund, L., Broczek, K., Carru, C., Christensen, K., Christiansen, L., Collerton, J.C., Cotichini, R., de Craen, A.J., Dato, S., Davies, K., De Benedictis, G., Deiana, L., Flachsbart, F., Gampe, J., Gilbault, C., Gonos, E.S., Haimes, E., Hervonen, A., Hurme, M.A., Janiszewska, D., Jylhä, M., Kirkwood, T.B., Kristensen, P., Laiho, P., Leon, A., Marchisio, A., Masciulli, R., Nebel, A., Passarino, G., Pelicci, G., Peltonen, L., Perola, M., Poulain, M., Rea, I.M., Remacle, J., Robine, J.M., Schreiber, S., Scurti, M., Sevini, F., Sikora, E., Skouteri, A., Slagboom, P.E., Spazzafumo, L., Stazi, M.A., Toccaceli, V., Toussaint, O., Törnwall, O., Vaupel, J.W., Voutetakis, K., Franceschi, C., 2011. GEHA consortium. Design, recruitment, logistics, and data management of the GEHA (Genetics of Healthy aging) project. Exp. Gerontol. 46 (11), 934–945.

Vanhooren, V., Desmyter, L., Liu, X.E., Cardelli, M., Franceschi, C., Federico, A., Libert, C., Laroy, W., Dewaele, S., Contreras, R., Chen, C., 2007. *N*-glycomic changes in serum proteins during human aging. Rejuvenation Res. 10 (4), 521–531a.

Weidner, C.I., Lin, Q., Koch, C.M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D.O., Jöckel, K.-H., Erbel, R., Mühleisen, T.W., Zenke, M., Brümmendorf, T.H., Wagner, W., 2014. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 15, R24.

Xu, W., Hou, Y., Hung, Y.S., Zou, Y., 2013. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. Signal Process. 93, 261–276 http://dx.doi.org/10.1016/j.sigpro.2012.08.005

Zhang, C.L., Popp, F.A., 1994. Log-normal distribution of physiological parameters and the coherence of biological systems. Med. Hypotheses 43 (1 (July)), 11–16.