Enhancing the Human Health Status Prediction: the ATHLOS Project

Panagiotis Anagnostou¹, Sotiris Tasoulis¹, Aristidis G. Vrahatis¹, Spiros Georgakopoulos¹, Matthew Prina^{2,3}, José Luis Ayuso-Mateos^{4,5,6}, Jerome Bickenbach^{7,8}, Ivet Bayes^{9,4}, Francisco Félix Caballero^{11,12}, Laia Egea-Cortés⁹, Esther García-Esquinas^{11,12}, Matilde Leonardi¹⁴, Sergei Scherbov^{20,22,23}, Abdonas Tamosiunas²⁴, Aleksander Galas²⁵, Josep Maria Haro^{9,4}, Albert Sanchez-Niubo^{9,4}, Vassilis Plagianakos¹, and Demosthenes Panagiotakos²⁶

¹Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece.

²Social Epidemiology Research Group. Health Service and Population Research Department, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.

³Global Health Institute, King's College London, London, UK.

⁴Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Madrid, Spain.

⁵Department of Psychiatry, Universidad Autónoma de Madrid, Madrid, Spain.

⁶Hospital Universitario de La Princesa, Instituto de Investigación Sanitaria Princesa (IIS Princesa), Madrid, Spain.

⁷Swiss Paraplegic Research, Guido A. Zäch Institute (GZI), Nottwil, Switzerland.

⁸Department of Health Sciences & Health Policy, University of Lucerne, Lucerne, Switzerland.

⁹Research, Innovation and Teaching Unit. Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain.

¹¹Department Preventive Medicine and Public Health, Universidad Autónoma de Madrid/Idipaz, Madrid, Spain.

¹²Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, CIBERESP, Madrid, Spain.

¹⁴Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy.

²⁰International Institute for Applied Systems Analysis, World Population

Program, Wittgenstein Centre for Demography and Global Human Capital,
Laxenburg, Austria.

²²Austrian Academy of Science, Vienna Institute of Demography, Vienna,
Austria.

²³Russian Presidential Academy of National Economy and Public
Administration (RANEPA), Moscow, Russian Federation.

²⁴Lithuanian University of Health Sciences, Kaunas, Lithuania.

²⁵Department of Epidemiology and Preventive Medicine, Jagiellonian
University, Krakow, Poland.

²⁶Department of Nutrition and Dietetics, School of Health Science and
Education, Harokopio University, Athens, Greece.

Corresponding Author: mr. Anagnostou Panagiotis. email: panagno@uth.gr
Address: Papasiopoulou 2-4, Lamia, Greece

Acknowledgement

This work is supported by the ATHLOS (Ageing Trajectories of Health: Longitudinal Opportunities and Synergies) project, funded by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement number 635316.

Declaration of interest

We confirm that the manuscript represents our own work, is original and has not been copyrighted, published, submitted, or accepted for publication elsewhere. We further confirm that we all have fully read the manuscript and give consent to be co-authors of the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

Preventive healthcare is a crucial pillar of health as it contributes to staying healthy and having immediate treatment when needed. Mining knowledge from longitudinal studies has the potential to significantly contribute to the improvement of preventive healthcare. Unfortunately, data originated from such studies are characterized by high complexity, huge volume and a plethora of missing values. Machine Learning, Data Mining and Data Imputation models are utilized as part of solving the aforementioned challenges, respectively. Towards this direction, we focus on the development of a complete methodology for the ATHLOS (Ageing Trajectories of Health: Longitudinal Opportunities and Synergies) Project - funded by the European Union's Horizon 2020 Research and Innovation Program, which aims to achieve a better interpretation of the impact of aging on health. The inherent complexity of the provided dataset lie in the fact that the project includes 15 independent European and international longitudinal studies of aging. In this work, we particularly focus on the HealthStatus (HS) score, an index that estimates the human status of health, aiming to examine the effect of various data imputation models to the prediction power of classification and regression models. Our results are promising, indicating the critical importance of data imputation in enhancing preventive medicine's crucial role.

Keywords– Imputation, Vtreat, Big data, Healthcare, Prediction

1 Introduction

We live in the “Big Data” era, which offers a great potential for revolutionizing all aspects of society, including the healthcare domain [1]. The recent advancements in medical and health domains generate, with an exponentially increasing rate, several heterogeneous data types, such as medical, demographic, molecular, cellular, and so on, offering great potential for personalized healthcare [2]. Digitization in medicine and healthcare has revolutionized how various biological processes and complex diseases are interpreted, since there is the possibility for analysis and knowledge extraction from large scale data [3]. In this direction, longitudinal studies, such as cohort studies and natural experiments, offer a platform to study more profound challenges in the biomedical and healthcare domain [4]. The predominant way to deal with the inherent data complexity is the Machine Learning frameworks [5], which allow the exploration, development and construction of algorithms that can learn from data and produce accurate predictions [6, 7].

Meanwhile, clinical and population health research usually includes questionnaires within longitudinal studies, creating a significant challenge regarding the missing values for several apparent reasons (non-response, loss tracking, etc.) [8]. The proper way to deal with this critical limitation is to impute the missing information without losing the imposed data structure [9]. The recent literature is plentiful with imputation methods for longitudinal studies [9–14], while it has been shown that such a pre-process step can lead to improvements in various Machine Learning tasks, such as the data classification [15]. Longitudinal studies often contain categorical information, which is restricted-transition variables. This information is usually incomplete; hence its imputation is a significant challenge [9]. The difficulty lies in the fact that there is little guidance on whether these restrictions need to be

accommodated when using imputation methods [16]. Nevertheless, most researches on this aspect focus on continuous data with fewer recommendations for categorical data [17].

A further essential challenge when dealing with longitudinal studies is the accuracy and validity of prediction models [18]. Towards this perspective, we study the prediction power in classification and regression tasks utilizing various imputation methodologies. Our aim is to enhance the predictability of machine learning tasks on a large-scale longitudinal study, characterized by high rate of missing values, and simultaneously mine new insights and knowledge for health status across the human age. For this purpose, we employ a harmonized dataset produced within the ATHLOS (Ageing Trajectories of Health: Longitudinal Opportunities and Synergies, <http://athlosproject.eu/>) Project (EU HORIZON2020-PHC-635316), which integrates 15 independent European and international longitudinal studies of aging.

The ATHLOS Project is based on improving a critical factor in the healthcare domain, extending the healthy lifespan. Its importance lies in the fact that in the last decades is observed a remarkable worldwide population aging due to the increase in life expectancy [19]. The healthcare demand increases along with the increasing number of elderly populations, which consists of a high cost every year for states [20]. Health status measurements are an immensely important indicator that helps prevent human health near or even in the distant future [21]. Under this perspective, the ATHLOS project aims to better study the impact of aging on health by developing a new single measure of health status. The project intends to identify patterns of healthy aging trajectories and their determinants, the critical points in time when changes in trajectories are produced, and propose timely clinical and public health interventions to optimize and promote healthy aging.

Furthermore, one of the unique features of ATHLOS is a metric of health, called Health-Status (HS), proposed in [22]. It is a novel index for human health condition calculation using an Item Response Theory (IRT) approach. This measure is based on specific characteristics that are part of the ATHLOS features list and has shown its reliability in several studies [23–38]. The challenge that now emerges is to discover meaningful relationships between the wide range of features found across various HS studies. Efficiently predicting HS through features that have been considered unrelated or not informative and thus excluded from its designing process would promote its expandability in further health-related studies while opening up new avenues for HS optimization.

This study focuses on strengthening the prediction power of the ATHLOS HealthStatus score in terms of classification and regression through various imputation methods. A large variety of both well-established and cutting-edge imputation and prediction methodologies were applied. More specifically, we applied five and six imputation and prediction methods, respectively by covering the broader range of imputation and prediction methods family. Our study shows that applying different imputation methods can greatly affect the various classification and regression schemes in terms of their accuracy and robustness. By discovering the most appropriate and methodological framework, we attempt to offer new insights and identify a better interpretation of the impact of aging on health status. The aforementioned outcomes can be considered promising, completing the puzzle of the strong outcomes that have emerged from previous studies of ATHLOS [23–38].

2 Methods & Experiments

Five imputation methods were applied (Mean [39], Linear Regression [40], Dual Imputation Model [41], Multiple Linear regression [42], and Vtreat method [43]) the harmonized ATHLOS dataset. Subsequently, the imputed matrices have been used as input to the six mentioned traditional and state-of-the-art regression and classification tools (Logistic/Linear Regression [44], k Nearest Neighbors [45], Random Forests [46], Extreme Gradient Boosting (XGBoost) [47], and two Deep Neural Network models). Details about the methodological pipeline, the brief analysis of all algorithms and the experimental setup are given in the following subsections.

2.1 Data preprocessing - Cleaning and Preparation

The ATHLOS project provides a harmonized dataset [23], built upon several longitudinal studies, originated from five continents. More specifically, it contains samples coming from more than 355,000 individuals who participated in 17 general population longitudinal studies in 38 countries. In this paper, we used 15 of these studies, namely the 10/66 Dementia Research Group Population-Based Cohort Study [24], the Australian Longitudinal Study of Aging (ALSA) [25], the Collaborative Research on Ageing in Europe (COURAGE) [26], the ELSA [27], the study on Cardiovascular Health, Nutrition and Frailty in Older Adults in Spain (ENRICA) [28], the Health, Alcohol and Psychosocial factors in Eastern Europe Study (HAPIEE) [29], the Health 2000/2011 Survey [30], the HRS [31], the JSTAR [32], the KLOSA [33], the MHAS ([34]), the SAGE [35], SHARE [36], the Irish Longitudinal Study of Ageing (TILDA) [37] and the the Longitudinal Aging Study in India (LASI) [38].

The 15 general population longitudinal studies utilised in this work consist of 990,000 samples in total, characterized by 184 variables, of which two are considered as response variables and the rest as independent variables. Response variables are the raw and the scaled HealthStatus scores for each individual. Regarding the independent variables (please see supplementary material, sheet S1), nine variables were removed including various indexes (sheet S2), 13 variables were removed including obviously depended variables that cannot be taken into account (sheet S3), and six variables were removed including information that cannot be considered within a prediction model (sheet S4).

It is worth mentioning that the 47 variables (sheet S5) that were originally used to calculate the HS score [22] have been excluded. These features create a statistical bias regarding the HS, which is the response variable in our analysis. Keep in mind, that these variables have been previously identified as more critical in defining the HS, as such we expect that using the rest for predicting it would be a challenging task, which however may reveal new insights. Finally, removing the samples for which the HS metric is missing, the resulting data matrix is constituted by 770,764 samples and 107 variables.

2.2 Missing Data Imputation

A brief analysis of all imputation tools used (Mean Imputation, Linear Regression, Dual Imputation Model, traditional Multiple Imputations, Vtreat) is provided. Mean Imputation (MI), Linear Regression (LR) belongs to the single imputation methods family having the

procedure to replace an unknown missing value by a single value. MI is a simple process based on the calculations of the mean value of the non-missing values for each response variable with a numerical data type. Missing values are filled with the respective mean value for each response variable. It is a relatively fast and straightforward approach, giving good results for relatively small numerical datasets. However, it exhibits various limitations, such as it does not consider the correlation among the response variables along with the data uncertainty in the imputations [48]. In the Linear Regression (LR) imputation, the missing values are replaced with predicted values that arise from applying a linear regression model in the non-missing data. The idea behind this process is the fact that the variables tend to be correlated. Hence, the predicting values should be affected by the observed data [40].

On the other hand, the Dual Imputation Model (DIM) and the Multiple Linear Regression Imputation (MLR) belong to the MI family where the missing values are replaced by n simulated values ($n > 1$) while its main assumption is that data are missing at random. The MI frameworks deal with the uncertainty about which values to impute, a major limitation of singular imputation methods. MLR imputation reduces this limitation by combining several different plausible imputed datasets. Initially, multiply imputed datasets are constructed and then standard statistical methods are applied to fit each one. Under the MI perspective, a recently proposed imputation method is the Dual Imputation Model (DIM) [41], which approximates an offset between the data distribution of the observed and missing data. This approximation is performed by iterating over the response and imputation models, assuming that the response as the imputation model has the same predictors.

Vtreat [43] is a recent imputation method that differs from similar traditional tools since it considers the uncertainty in the imputations. It includes a unique strategy with respect to the type of the response variable, creating additional variables along with the well-known dummy variables process. For a numerical type response variable, one more dummy variable is created to address the high cardinality categorical variables. Two dummy variables are created to cope with the novel (or rare) levels found in each categorical variable of the data set. For a categorical type response variable, one more dummy variable is created to address the high cardinality categorical variables. One dummy variable is created to cope with the novel (or rare) levels found in each categorical variable of the dataset. This inherent feature enhances the knowledge discovery of ultra-noisy data. Furthermore, vtreat is useful in the modeling process because it allows the model to re-estimate the response on systematically missing values. This property motivated us to focus our attention upon it since the nature of longitudinal studies imposes the emergence of systematically missing values at a high degree.

Regarding the experimental study and for the MI method, 141 dummy variables were created during the conversion of 66 categorical predictor variables. Mean values from each of the 34 remaining predictor variables values were calculated by replacing the respective missing values indexes. For the LR Imputation, 141 dummy variables were constructed following the same process for creating the dummy variables. A simple model was created for the remaining 34 predictor variables:

$$X_m = a_0 + \sum_{i=0} a_i X_i + \epsilon, \quad (1)$$

where X_m is the predictor variable to be imputed, X_i is the list of n complete predictor variables, a_i are the regression coefficients, a_0 is the intercept, and ϵ is the normally distributed

error.

Regarding the application of MLR (Multiple Linear Regression) imputation, we followed the multiple imputation main steps by using the Linear Regression model to perform the m complete data analyses. Initially it defines the posterior predictive density $p(Y_{mis}|X, R)$, where Y_{mis} indicates the missing parts of Y , X is the 34 predictor variables, and R is the $770,764 \times 33$ binary matrix indicating the elements of Y that are observed ($R_{ij} = 1$ if Y_{ij} is observed). Note that Y indicates the partially observed outcome variables in a sample of size $n = 770,764$. Subsequently, imputations are generated based on this density by concatenating $m = 5$ complete datasets. In the next step, a Linear Regression model was applied on each completed data matrix performing $m = 5$ complete data analyses, while an ensemble criterion exports the outcome for each point. Here, we employ the R implementation of the MICE (Multivariate Imputation by Chained Equations) [49] for the MLR model.

In the DIM (Dual Imputation Model), we integrated the multiple imputation along with the doubly robust weight frameworks to deal with misspecification included on ATHLOS data with missing at random values. Briefly, let the observed data be $\mathbf{V} = (\mathbf{Y}_{\text{obs}}, \mathbf{R}, \mathbf{X}, \mathbf{U})$, and $P(\boldsymbol{\theta}, \boldsymbol{\alpha} | \mathbf{V}) \propto P(\mathbf{V} | \boldsymbol{\theta}, \boldsymbol{\alpha})p(\boldsymbol{\theta}, \boldsymbol{\alpha})$, and $p(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is the prior density. The main equation of DIM which defines the posterior predictive distribution of missing data given the observed data, is:

$$P(\mathbf{Y}_{\text{mis}} | \mathbf{V}) = \iint P(\mathbf{Y}_{\text{mis}} | \mathbf{V}, \boldsymbol{\theta}, \boldsymbol{\alpha}) p(\boldsymbol{\theta}, \boldsymbol{\alpha} | \mathbf{V}) d\boldsymbol{\theta} d\boldsymbol{\alpha}.$$

Initially, we imputed the missing values by a random process taking into account the observed ATHLOS data. Then an iterative process was applied by (i) drawing a random value $\hat{\alpha}_t$ from its posterior distribution, (ii) calculating the propensity scores $\hat{\pi}_t$ given the drawn value $\hat{\alpha}_t$, (iii) adding an additional predictor into the imputation model, (iv) estimating the parameters of the imputation model for the observed values, (v) drawing a random value from their posterior distributions, and (vi) imputing the missing value adding a noise factor. We implemented this iterative process ten times, as suggested by [41].

The cutting-edge vtreat was applied on ATHLOS dataset following its previously described principles. We imputed the missing values through 458 dummy variables created for the continuous numerical response variable and 392 dummy variables created for the categorical response variable. The resulting matrices are given as input in the next step of the proposed framework, the significance pruning process. Each variable was evaluated based on its correlation with the HealthStatus (HS) [22] score (response variable) in order to isolate the most significant variables. More specifically, we applied a treatment plan by decoding and removing the noisy variables. For this purpose, we calculated their correlation with HS using the F-test, since it is a numeric variable and a X^2 -Test for the categorical target created with respect to the classification tasks.

The significance of each predictor variable was calculated using F-statistic on the single-variable regression model and X^2 -statistic on the logistic model between the response variable and each predictor. Subsequently, we pruned the variables with higher significance value than the threshold $1/n_vars$ in both cases, where n_vars indicates the resulting number of variables obtained from the matrix imputation step. As a result of the significance pruning process, for the numerical target, 51 noisy variables were removed. The rest 407 variables

were normalized using the linear regression model. Thus, for each predictor variable we have:

$$y = mx + b, \quad (2)$$

where x is a predictor variable, m is the linear regression coefficient, and b is the intercept. Then, scaling and centering takes place with the following procedure:

$$x' = mx - \text{mean}(mx). \quad (3)$$

For the categorical response variable, about 51 noisy variables were removed, resulting in 351 remaining variables. Subsequently, the predictor variables were normalized using the logistic regression model for each predictor variable:

$$\log_b \frac{p}{1-p} = mx + b, \quad (4)$$

where $p = P[y = TRUE]$ and y is the response variable, x is a predictor variable, m is the logistic regression coefficient and b is the intercept. Similarly, scaling and centering of each predictor variable took place with the following procedure:

$$x' = mx - \text{mean}(mx). \quad (5)$$

It is worth mentioning that all five methods utilize the dummy variable creation process for the categorical variables. In this approach, each categorical level is represented separately, including the values 0 and 1 for the participation or not in a specific level. The level that is not represented contains zeros on all dummy variables, acting as the reference level [50]. In our workflow, we deal with missing values of all categorical variables following the principles highlighted in [51]. The missing values are considered as regular categorical levels and are included in the dummy variable process. Concluding, five independent imputed datasets were generated and were used as input in the next section to examine the prediction power from various Machine Learning algorithms.

2.3 Prediction Methods

Six different classification and regression models, namely Logistic/Linear Regression, kNN, Random Forests, XGBoost, and two Deep Neural Network (DNN) models, have been applied on the five resulting datasets retrieved by applying the five imputation methods. Briefly, let $A \in R^{s \times d}$ be the ATHLOS dataset, where s is the number of samples, and d is the number of predictor variables. Let $H \in R^{s \times 1}$ denote the response variable, a vector with the previously described HealthStatus (HS) score. For the classification process, a three-class categorization was considered after consulting previously similar categorization [52]. More precisely, the three groups are defined based on HS scores as follows: scores lower than or equal to 30, scores higher than 30 and lower than or equal to 50, and scores higher than 50.

Linear and Multinomial Logistic regression was used for the regression and classification tasks, respectively. In linear regression, we used HS as the continuous dependent variable, while in logistic regression we used the aforementioned three-class categorization of HS. We applied kNN classification, searching for 5 nearest neighbors to assign each sample to one of

the three classes using the Euclidean distance. In kNN regression, we calculated the average of the HS values of the k-Nearest Neighbors search of a given test point. Let $H \in R^{s \times 1}$ denote the set of s training points that contain the HS values. The kNN estimator is defined as the mean function value of the nearest neighbors [53, 54]. The Random Forests (RF) regression operates in a similar way, with the main difference being that the final prediction output is the average of all trees' output. RF method builds N trees, using randomization to decrease the correlation between them and achieve high accuracy. In this respect, trees operate on a random subset of the data. Furthermore, for each tree node, a random subset of the variables is considered. In our analysis, we use 100 trees and the M_{try} variable was defined as $p/3$ and $\sqrt{p}$, where p are the number of variables. Extreme Gradient Boosting (XGBoost) is a novel classifier based on an ensemble of classification and regression trees [47]. Let the output of a tree be:

$$f(x) = wq(x_i), \quad (6)$$

where x is the input vector and wq is the score of the corresponding leaf q . The output of an ensemble of K trees will be:

$$y_i = \sum_{k=1}^K f_k(x_i). \quad (7)$$

The XGBoost algorithm tries to minimize the following objective function J at step t :

$$J(t) = \sum_{i=1}^n L(y_i \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{i=1}^t \Omega(f_i), \quad (8)$$

where the first term contains the train loss function L (e.g. mean squared error) between real class y and output $\hat{y}$ for the n samples, and the second term is the regularization term that controls the complexity of the model and helps to avoid overfitting.

We also applied two Deep Learning architectures using the well-known Backpropagation (BP) method (based on the Gradient Descent method) for the training phase. BP requires a vector of input patterns x and the corresponding target y . Given an x_i where i corresponds to an individual input sample, the model produces an output o_i . The main aim is to reduce the error in the training process through these proper weights that minimize the following cost function:

$$E(w) = \frac{1}{n} \sum_{p=1}^P \sum_{j=1}^{N_L} (o_{j,p}^L - y_{j,p})^2, \quad (9)$$

where P the number of patterns, $o_{j,p}^L$ the output of j neuron of layer L , N_L is the number of neurons in output layer, $y_{j,p}$ is the desirable target of j neuron of pattern p . The BP algorithm is utilized to minimize the cost function $E(w)$, as mentioned above. The aim of optimization algorithm is to find the minimized $w^* = (w_1^*, w_2^*, \dots, w_v^*) \in R^v$, such that:

$$w^* = \min_{w \in R^v} E(w). \quad (10)$$

We selected the Deep Learning strategy to examine its behavior in avoiding underfitting or overfitting on ATHLOS training step since the examined imputation methods varying in terms of the number of features. More specifically, the first DNN (DNN_1) consisted

	Linear Regression	kNN	Random Forests	XGBoost	DNN1	DNN2
R-squared						
LR	53.06 (0.61)	2.00 (0.29)	55.20 (0.86)	55.53 (1.44)	9.83 (0.42)	5.75 (0.36)
Mean	53.30 (0.81)	22.19 (0.72)	55.91 (0.53)	56.66 (0.64)	40.55 (0.56)	19.70 (0.69)
MLR	53.02 (1.04)	2.55 (0.30)	54.83 (1.28)	56.21 (1.24)	44.67 (0.67)	0.01 (0.00)
DIM	52.94 (0.83)	1.85 (0.36)	54.47 (0.79)	55.72 (0.86)	9.44 (0.37)	3.69 (0.27)
Vtreat	53.42 (0.63)	48.40 (1.10)	55.90 (0.70)	57.84 (0.78)	55.84 (0.63)	45.04 (86.00)
Root Mean Square Error						
LR	6.841 (0.048)	10.418 (0.110)	6.770 (0.104)	6.658 (0.080)	9.448 (0.134)	9.648 (0.137)
Mean	6.822 (0.037)	9.003 (0.062)	6.642 (0.037)	6.579 (0.041)	7.579 (0.089)	8.801 (0.095)
MLR	6.841 (0.078)	10.363 (0.095)	6.743 (0.054)	6.641 (0.044)	9.473 (0.126)	9.978 (0.114)
DIM	6.870 (0.070)	10.484 (0.101)	6.728 (0.046)	6.640 (0.057)	8.824 (0.115)	9.787 (0.127)
Vtreat	6.808 (0.067)	7.279 (0.078)	6.659 (0.054)	6.573 (0.048)	6.653 (0.090)	7.299 (0.097)

Table 1: Comparison of five imputation methods (Linear Regression (LR), Mean, Multiple Linear Regression (MLR), Dual Imputation Method (DIM), Vtreat) in regression tasks using 6 different regression techniques (Deep Neural Network (DNN) 1, DNN2, k-Nearest Neighbors (kNN), Linear Regression (LR), Random Forests (RF), XGBoost). The table contains the mean (standard error) values (%) of the R-squared measure, the mean (standard error) values of Root Mean Square Error (RMSE) as well, from 80 independent executions. The best value among imputation methods for each classifier is depicted in bold and the highest value of all imputation methods for all classifiers is depicted in bold italics.

of two hidden layers of 100 neurons and one output layer of one neuron, while the second network (DNN_2) has only one hidden layer with 100 neurons. In order to avoid the gradient vanishing problem, the ReLU activation function is utilized in hidden layers [55]. Since the response variable takes values in the range of (0,100) to produce comparable results, the linear activation functions are utilized in the output layer. Finally, the learning rate is set to 0.001.

3 Results & Discussion

All executions were validated through the Monte Carlo cross-validation technique [56], creating multiple random splits of the dataset into train and test by selecting 100,000 and 10,000 samples, respectively. We implement 80 independent executions assuring that each sample was used as a test sample at least once. The regression performance of each imputation method was calculated using the R-squared measure [57] and the Root Mean Squared Error (RMSE) [58] upon the test set. Also, Accuracy, F1-score, Sensitivity, and Specificity were evaluated to measure the classification performance. The results are summarized in the Figure 1 and Table 1.

In general, we observed that the Vtreat imputation method consistently improves both the classification and regression performance compared to the other methods. Its superiority

Figure 1: Each radar plot contains the visual representation of the classification results for each imputation method used in this paper. The methods are: Mean (Mean) imputation, Linear Regression (LR) imputation, Multi Linear Regression (MLR) imputation, Dual Imputation Model (DIM), and Vtreat imputation. The axes of the radar plots are metrics accuracy, precision, recall, sensitivity, and specificity. Finally, there is one radar plot for each of the classification models utilized. Namely, for the implementation of the Logistic Regression model, the kNN Classification model, the Random Forests model, the XGBoost model, and the two Deep Neural Network models (DNN1 and DNN2).

	LR	Mean	MLR	DIM	Vtreat
R-squared	6.07	2.89	5.93	4.9	43.5
RMSE	9.705	9.862	9.733	9.767	7.545

Table 2: Comparison of 5 imputation methods (Linear Regression (LR), Mean, Multi Linear Regression (MLR), Dual Imputation Method (DIM), Vtreat) using the Principal Components Regression technique. The table contains the (%) of the R-squared measure, the mean (standard error) values of Root Mean Square Error (RMSE) as well. The best value among imputation methods for each measure is depicted in bold.

appears in almost all evaluation measures and cases. Interestingly, we observe a significant improvement for the deep neural network and k-Nearest Neighbors concerning classification and regression. A minor difference is observed in logistic and linear regression for classification and regression performance, respectively. Additionally, we observe that the combination of XGBoost and Vtreat outperforms any other method/combination for both regression and classification tasks. Given that XGBoost is an efficient, cutting-edge classification technique, Vtreat reveals its true potential by enhancing its performance. Apparently, the better structure offered affects tree-based ensemble algorithms known to be appropriate for structured data. However, we also observe that Vtreat significantly improves the performance of both DNN models, known to perform better for unstructured data, such as images and text, further showing its potential in the data imputation task.

3.1 Visualizing the prediction power

Visualization is a central part of ML approaches, since it offers insights into the data structure. Most visualization approaches apply a dimensionality reduction technique to transform the multidimensional data into two or three dimensions, while preserving their relationships. Here, we focus on visualizing each imputation method’s predictive power through a two-dimensional representation. For this purpose, we employed the Principal Components Regression (PCR) technique to predict the Healthstatus score. Through this, we can investigate prediction power along with the corresponding two-dimensional representation produced by the respective Principal Component Analysis (PCA) method. Our motivation also lies in PCR’s advantages in both statistical and computational aspects, which reduce the variance estimation at the expense of other bias and offers a better model fitting respectively [59].

Briefly, let $A \in R^{s \times d}$ be the ATHLOS dataset, where s is the number of samples, and d is the number of predictor variables after applying the significant pruning step. Let $H \in R^{s \times 1}$ denote the vector of HealthStatus scores. Through PCR we initially perform PCA on the centered data matrix A and subsequently only consider the first two principal components, producing the matrix $A^{PC} \in R^{s \times 2}$. For the N number of principal components, the formula for the HS evaluation is calculated as:

$$HSEvaluation = a_0 + \sum_{i=1}^p a_i A_i^{PC}, \quad (11)$$

where a_0 is the intercept, a_i is the i^{th} regression coefficient and A_i^{PC} is the i^{th} principal

Figure 2: Scatter plots (left column) depict the first two principal components of PCA performed on the five imputed ATHLOS datasets using Linear Regression, Mean, Dual Imputation Model, and Vtreat imputation. Circular points with orange, yellow, and light blue colors illustrate the low, medium, and high HealthStatus scores. Above and right to each scatter plot, their data distribution is illustrated. Heatmap-Scatter plots (right column) depict the correlation of predicted and real HS score of the five imputation methods using the Principal Components Regression (PCR) technique. The red to green color graduation of boxes indicates the number of samples from low to high amounts, respectively. Above and right to each heatmap-scatter plot is illustrated the marginal distribution of the HS and the HS estimation as univariate histograms with a density curve on the vertical and horizontal axes of the scatter plot, respectively.

component resulted from PCA. The PCR technique is applied upon all five datasets produced by the application of each corresponding imputation method.

Initially, we observe that the low, medium and high values of HealthStatus score appear better separability for the Vtreat method. In addition, observing the class distribution with respect to the first Principal Component, a clear distinction between classes is only achieved for the Vtreat method, while the respective distributions arising from LR, Mean, and DIM imputation methods have quite similar behavior (see Figure 2 for details).

Utilizing the PCR technique to visualize regression efficiency (see Figure 2, right column) allow us to observe that the best performance is achieved by the Vtreat method, indicated by the higher correlation between predicted and real HS scores. Zooming into the respective heatmap-scatter, we can see the high number (green graduation in boxes) of samples across the dashed line, highlighting the optimal regression as reported in Table 2. We also observe all other methods perform poorly compared to Vtreat, while the mean imputation method behaves differently from the other two, with almost identical results.

3.2 Feature Importance for HealthStatus prediction

In this Section, our aim is to identify the factors that significantly affect HealthStatus prediction. Based on the indicated performance evaluation within the aforementioned experimental analysis, we captured this information by isolating the variable importance measure of the XGBoost algorithm when applied to the dataset imputed using the Vtreat methodology. The tree based nature of this algorithm allowed us to export the relative importance or contribution of each input variable in predicting the response.

Let T , be a single decision tree, based on [60]. We apply the following equation as an index of relevance for each of our features (predictor variables) X_ℓ :

$$\mathcal{I}_\ell^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = \ell),$$

where $J - 1$ indicates the internal nodes of the tree, t , indicates each one of the input variables $X_{v(t)}$ and $\hat{i}_t^2$ indicates the maximal estimated improvement which defines the particular variable in squared error risk over that for a constant fit over the entire region $\hat{i}_t^2$. Utilizing this equation through the XGBoost model [47], we obtain a list with the importance of each feature.

The outcomes illustrated in Figure 3 offer new insights for the HealthStatus evaluation. More specifically, the most important variables (left plot in Figure 3) regarding their effectiveness in the HealthStatus prediction were the features: "Respondent's self-rated/self-reported health" (srh), "Age at time of measure" (age), "History of arthritis, rheumatism or osteoarthritis" ("h_joint_disorders"), "Current depressive status" (depression), "Engage in vigorous exercise during the last two weeks" (vig_pa), "Highest score, to the nearest kg, of all hand grip measurements, regardless" (grip), "Psychological measure of anxiety symptoms" (anxiety_symp), "Weekly frequency of less vigorous exercise" (f_mod_pa), "Participant has paid employment" (employed), and "Highest level of formal education achieved" (education).

Figure 3: The horizontal bars illustrate the most (left) and the least (right) important variables regarding their effectiveness in the HealthStatus prediction by applying the XGBoost classification algorithm. The x-axis imprints the variable importance score while the y-axis includes the feature names defined by the ATHLOS project (see supplementary sheet S1).

On the other hand, the least important features (right plot in Figure 3) were the "Current smoker, any type of tobacco smoking" (current_smoking), "Ever experienced any natural disaster" (nat_dis), "Respondent is the participant, spouse/partner or other" (respondent), "Serum HDL Cholesterol" (hdl_chol), "Triglycerides" (triglycerides), "Experience of divorce/separation" (divorce), "Blood/serum glucose levels" (glucose), "Total Cholesterol" (total_chol), "Serum LDL Cholesterol" (ldl_chol) and "Is the relationship with the spouse close" (close_spouse).

Through the variable importance analysis we interestingly observe that there are factors that significantly affect HealthStatus prediction, while being unrelated to health, as well as factors that do not affect the HealthStatus but apparently should have, based on the recent literature. More specifically, the levels of depression were considered an important factor that affects the respondent's HealthStatus prediction, a condition that exists at various age levels and does not necessarily indicate aging. Factors such as Serum HDL Cholesterol, Triglycerides, Blood/serum glucose levels, and Total Cholesterol were considered minor to determine the accurate HealthStatus prediction. However, all these factors are quite relevant to the human status health [61,62].

Additionally, it is impressive how relevant is everyone's personal opinion about assessing their state of health. Our results indicated that it is the most important factor. We may consider encouraging the fact that most can personally assess their state of health, since this could affect prevention of adverse events and bad incidents, contributing to the Medical and Healthcare domain's effectiveness that follows the participatory and preventive factors, two of the four main axes that constitute the future of Medicine [63].

4 Concluding Remarks

In this paper, we emphasize on the development of a complete ML methodology for the ATH-LOS (Ageing Trajectories of Health: Longitudinal Opportunities and Synergies) Project, aiming in the accurate prediction of the HealthStatus (HS) score, an index that estimates the human status of health. We deal with the most critical aspect of the inherent complexity of the provided dataset, the high degree of missing values, by exploring extensively and discovering the most effective methodologies for missing value imputation. Our results indicated the effect of different imputation strategies on Machine Learning tasks exposing the capacity of accurate predictions when combined with cutting-edge tools. We additionally concluded that the highlighted Vtreat imputation method can be rendered as an imputation tool-guide for multiple unified, independent longitudinal studies. Finally, the variable importance analysis highlighted new factors that may affect the state of human health exposing the potential of further findings within similar large scale longitudinal studies. In our future research, we intend to further investigate the development of a unified model on the basis of a bidirectional ML scheme for data imputation and prediction.

References

- [1] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, G.-Z. Yang, Big data for health, *IEEE journal of biomedical and health informatics* 19 (4) (2015) 1193–1208.
- [2] M. Viceconti, P. Hunter, R. Hose, Big data, big knowledge: big data for personalized healthcare, *IEEE journal of biomedical and health informatics* 19 (4) (2015) 1209–1215.
- [3] L. Flitcroft, W. S. Chen, D. Meyer, The demographic representativeness and health outcomes of digital health station users: Longitudinal study, *Journal of Medical Internet Research* 22 (6) (2020) e14977.
- [4] J. Zhao, Q. Feng, P. Wu, R. A. Lupu, R. A. Wilke, Q. S. Wells, J. C. Denny, W.-Q. Wei, Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction, *Scientific reports* 9 (1) (2019) 1–10.
- [5] J. Wiens, E. S. Shenoy, Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology, *Clinical Infectious Diseases* 66 (1) (2018) 149–153.
- [6] L. Zhou, S. Pan, J. Wang, A. V. Vasilakos, Machine learning on big data: Opportunities and challenges, *Neurocomputing* 237 (2017) 350–361.
- [7] A. M. Tai, A. Albuquerque, N. E. Carmona, M. Subramanieapillai, D. S. Cha, M. Sheko, Y. Lee, R. Mansur, R. S. McIntyre, Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry, *Artificial intelligence in medicine* (2019) 101704.
- [8] P. Diggle, P. J. Diggle, P. Heagerty, K.-Y. Liang, P. J. Heagerty, S. Zeger, et al., *Analysis of longitudinal data*, Oxford University Press, 2002.

- [9] M. H. Huque, J. B. Carlin, J. A. Simpson, K. J. Lee, A comparison of multiple imputation methods for missing data in longitudinal studies, *BMC medical research methodology* 18 (1) (2018) 168.
- [10] Z. Zhang, Missing data imputation: focusing on single imputation, *Annals of translational medicine* 4 (1) (2016).
- [11] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, *Artificial intelligence in medicine* 50 (2) (2010) 105–115.
- [12] N. Nooraee, G. Molenberghs, J. Ormel, E. R. Van den Heuvel, Strategies for handling missing data in longitudinal studies with questionnaires, *Journal of Statistical Computation and Simulation* 88 (17) (2018) 3415–3436.
- [13] Y. Yamaguchi, M. Ueno, K. Maruo, M. Gosho, Multiple imputation for longitudinal data in the presence of heteroscedasticity between treatment groups, *Journal of Biopharmaceutical Statistics* 30 (1) (2020) 178–196.
- [14] M. Vilardell, M. Buxó, R. Clèries, J. M. Martínez, G. Garcia, A. Ameijide, R. Font, S. Civit, B. working group, et al., Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (mod-graprodep): an application to breast cancer survival, *Artificial Intelligence in Medicine* (2020) 101875.
- [15] I. Jordanov, N. Petrov, A. Petrozziello, Classifiers accuracy improvement based on missing data imputation, *Journal of Artificial Intelligence and Soft Computing Research* 8 (1) (2018) 31–48.
- [16] A. P. De Silva, M. Moreno-Betancur, A. M. De Livera, K. J. Lee, J. A. Simpson, Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study, *BMC medical research methodology* 19 (1) (2019) 1–14.
- [17] M. R. Stavseth, T. Clausen, J. Røislien, How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data, *SAGE open medicine* 7 (2019) 2050312118822912.
- [18] S. Licher, M. J. Leening, P. Yilmaz, F. J. Wolters, J. Heeringa, P. J. Bindels, A. D. N. Initiative, M. W. Vernooij, B. C. Stephan, E. W. Steyerberg, et al., Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies, *American Journal of Psychiatry* 176 (7) (2019) 543–551.
- [19] G. Passarino, F. De Rango, A. Montesanto, Human longevity: Genetics or lifestyle? it takes two to tango, *Immunity & Ageing* 13 (1) (2016) 12.
- [20] Y.-M. Liu, Population aging, technological innovation, and the growth of health expenditure: Evidence from patients with type 2 diabetes in taiwan, *Value in health regional issues* 21 (2020) 120–126.

- [21] D. Kelly, K. Curran, B. Caulfield, Automatic prediction of health status using smartphone-derived behavior profiles, *IEEE journal of biomedical and health informatics* 21 (6) (2017) 1750–1760.
- [22] F. F. Caballero, G. Soulis, W. Engchuan, A. Sánchez-Niubó, H. Arndt, J. L. Ayuso-Mateos, J. M. Haro, S. Chatterji, D. B. Panagiotakos, Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the athlos project, *Scientific reports* 7 (2017) 43955.
- [23] A. Sanchez-Niubo, L. Egea-Cortés, B. Olaya, F. F. Caballero, J. L. Ayuso-Mateos, M. Prina, M. Bobak, H. Arndt, B. Tobiasz-Adamczyk, A. Pająk, et al., Cohort profile: The ageing trajectories of health—longitudinal opportunities and synergies (athlos) project, *International journal of epidemiology* 48 (4) (2019) 1052–1053i.
- [24] A. M. Prina, D. Acosta, I. Acosta, M. Guerra, Y. Huang, A. Jotheeswaran, I. Z. Jimenez-Velazquez, Z. Liu, J. J. Llibre Rodriguez, A. Salas, et al., Cohort profile: the 10/66 study, *International journal of epidemiology* 46 (2) (2017) 406–406i.
- [25] M. A. Luszcz, L. C. Giles, K. J. Anstey, K. C. Browne-Yung, R. A. Walker, T. D. Windsor, Cohort profile: The australian longitudinal study of ageing (alsa), *International journal of epidemiology* 45 (4) (2016) 1054–1063.
- [26] M. Leonardi, S. Chatterji, S. Koskinen, J. L. Ayuso-Mateos, J. M. Haro, G. Frisoni, L. Frattura, A. Martinuzzi, B. Tobiasz-Adamczyk, M. Gmurek, et al., Determinants of health and disability in ageing population: the courage in europe project (collaborative research on ageing in europe), *Clinical psychology & psychotherapy* 21 (3) (2014) 193–198.
- [27] A. Steptoe, E. Breeze, J. Banks, J. Nazroo, Cohort profile: the english longitudinal study of ageing, *International journal of epidemiology* 42 (6) (2013) 1640–1648.
- [28] F. Rodríguez-Artalejo, A. Graciani, P. Guallar-Castillón, L. M. León-Muñoz, M. C. Zuluaga, E. López-García, J. L. Gutiérrez-Fisac, J. M. Taboada, M. T. Aguilera, E. Regidor, et al., Rationale and methods of the study on nutrition and cardiovascular risk in spain (enrica), *Revista Española de Cardiología (English Edition)* 64 (10) (2011) 876–882.
- [29] A. Peasey, M. Bobak, R. Kubinova, S. Malyutina, A. Pajak, A. Tamosiunas, H. Pikhart, A. Nicholson, M. Marmot, Determinants of cardiovascular disease and other non-communicable diseases in central and eastern europe: rationale and design of the hapiee study, *BMC public health* 6 (1) (2006) 255.
- [30] K. S., Health 2000 and 2011 surveys—thl biobank. national institute for health and welfare., <https://thl.fi/fi/web/thl-biobank/for-researchers/sample-collections/health-2000-and-2011-surveys>, [Online; accessed 18-July-2008] (2018).

- [31] A. Sonnega, J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. Phillips, D. R. Weir, Cohort profile: the health and retirement study (hrs), *International journal of epidemiology* 43 (2) (2014) 576–585.
- [32] H. Ichimura, S. Shimizutani, H. Hashimoto, Jstar first results 2009 report, Tech. rep., Research Institute of Economy, Trade and Industry (RIETI) (2009).
- [33] J. H. Park, S. Lim, J. Lim, K. Kim, M. Han, I. Y. Yoon, J. Kim, Y. Chang, C. B. Chang, H. J. Chin, et al., An overview of the korean longitudinal study on health and aging, *Psychiatry investigation* 4 (2) (2007) 84.
- [34] R. Wong, A. Michaels-Obregon, A. Palloni, Cohort profile: the mexican health and aging study (mhas), *International journal of epidemiology* 46 (2) (2017) e2–e2.
- [35] P. Kowal, S. Chatterji, N. Naidoo, R. Biritwum, W. Fan, R. Lopez Ridaura, T. Maximova, P. Arokiasamy, N. Phaswana-Mafuya, S. Williams, et al., Data resource profile: the world health organization study on global ageing and adult health (sage), *International journal of epidemiology* 41 (6) (2012) 1639–1649.
- [36] A. Börsch-Supan, M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, S. Zuber, Data resource profile: the survey of health, ageing and retirement in europe (share), *International journal of epidemiology* 42 (4) (2013) 992–1001.
- [37] B. J. Whelan, G. M. Savva, Design and methodology of the irish longitudinal study on ageing, *Journal of the American Geriatrics Society* 61 (2013) S265–S268.
- [38] P. Arokiasamy, D. Bloom, J. Lee, K. Feeney, M. Ozolins, Longitudinal aging study in india: Vision, design, implementation, and preliminary findings, in: *Aging in Asia: findings from new and emerging data initiatives*, National Academies Press (US), 2012.
- [39] H. M. Dodeen, Effectiveness of valid mean substitution in treating missing data in attitude assessment, *Assessment & Evaluation in Higher Education* 28 (5) (2003) 505–513.
- [40] I. Eekhout, H. C. de Vet, J. W. Twisk, J. P. Brand, M. R. de Boer, M. W. Heymans, Missing data in a multi-item instrument were best handled by multiple imputation at the item score level, *Journal of clinical epidemiology* 67 (3) (2014) 335–342.
- [41] S. Jolani, L. E. Frank, S. van Buuren, Dual imputation model for incomplete longitudinal data, *British Journal of Mathematical and Statistical Psychology* 67 (2) (2014) 197–212.
- [42] M. J. Azur, E. A. Stuart, C. Frangakis, P. J. Leaf, Multiple imputation by chained equations: what is it and how does it work?, *International journal of methods in psychiatric research* 20 (1) (2011) 40–49.
- [43] N. Zumel, J. Mount, vtreat: a data. frame processor for predictive modeling, arXiv preprint arXiv:1611.09477 (2016).

- [44] F. Y. Hsieh, D. A. Bloch, M. D. Larsen, A simple method of sample size calculation for linear and logistic regression, *Statistics in medicine* 17 (14) (1998) 1623–1634.
- [45] T. Hastie, R. Tibshirani, Discriminant adaptive nearest neighbor classification and regression, in: *Advances in Neural Information Processing Systems*, 1996, pp. 409–415.
- [46] L. Breiman, Random forests, *Machine learning* 45 (1) (2001) 5–32.
- [47] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [48] W.-C. Lin, C.-F. Tsai, Missing value imputation: a review and analysis of the literature (2006–2017), *Artificial Intelligence Review* 53 (2) (2020) 1487–1509.
- [49] S. v. Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in r, *Journal of statistical software* (2010) 1–68.
- [50] H. Alkharusi, Categorical variables in regression analysis: A comparison of dummy and effect coding, *International Journal of Education* 4 (2) (2012) 202.
- [51] N. Zumel, J. Mount, J. Porzak, *Practical data science with R*, Manning Shelter Island, NY, 2014.
- [52] D. Panaretos, E. Kolooverou, A. C. Dimopoulos, G.-M. Kouli, M. Vamvakari, G. Tzaveas, C. Pitsavos, D. B. Panagiotakos, A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the attica study, *British Journal of Nutrition* 120 (3) (2018) 326–334.
- [53] L. Devroye, The uniform convergence of nearest neighbor regression function estimators and their application in optimization, *IEEE Transactions on Information Theory* 24 (2) (1978) 142–151.
- [54] C. G. Atkeson, A. W. Moore, S. Schaal, Locally weighted learning, in: *Lazy learning*, Springer, 1997, pp. 11–73.
- [55] K. Eckle, J. Schmidt-Hieber, A comparison of deep networks with relu activation function and linear spline-type methods, *Neural Networks* 110 (2019) 232–242.
- [56] W. Dubitzky, M. Granzow, D. P. Berrar, *Fundamentals of data mining in genomics and proteomics*, Springer Science & Business Media, 2007.
- [57] A. C. Cameron, F. A. Windmeijer, An r-squared measure of goodness of fit for some common nonlinear regression models, *Journal of econometrics* 77 (2) (1997) 329–342.
- [58] T. Chai, R. R. Draxler, Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature, *Geoscientific model development* 7 (3) (2014) 1247–1250.

- [59] M. Slawski, et al., On principal components regression, random projections, and column subsampling, *Electronic Journal of Statistics* 12 (2) (2018) 3673–3712.
- [60] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and regression trees*, CRC press, 1984.
- [61] A. E. Morgan, M. T. Mc Auley, Cholesterol homeostasis: An in silico investigation into how aging disrupts its key hepatic regulatory mechanisms, *Biology* 9 (10) (2020) 314.
- [62] S. Jian, N. Su-Mei, C. Xue, Z. Jie, W. Xue-sen, Association and interaction between triglyceride–glucose index and obesity on risk of hypertension in middle-aged and elderly adults, *Clinical and Experimental Hypertension* 39 (8) (2017) 732–739.
- [63] L. Hood, P4 medicine and scientific wellness: catalyzing a revolution in 21st century medicine, *Molecular Frontiers Journal* 1 (02) (2017) 132–137.