Aging and Disease
www.aginganddisease.org
Volume 6, Number 3; 196-207, June 2015
http://dx.doi.org/10.14336/AD.2014.0623

Original article

Information Theoretical Analysis of Aging as a Risk Factor for Heart Disease

David Blokh[1] and Ilia Stambler[2]*

[1]C.D. Technologies Ltd., Israel
[2]Department of Science, Technology and Society, Bar Ilan University, Ramat Gan, Israel

[Received April 23, 2014; Revised June 19, 2014; Accepted June 23, 2014]

ABSTRACT: We estimate the weight of various risk factors in heart disease, and the particular weight of age as a risk factor, individually and combined with other factors. To establish the weights we use the information theoretical measure of normalized mutual information that permits determining both individual and combined correlation of diagnostic parameters with the disease status. The present information theoretical methodology takes into account the non-linear correlations between the diagnostic parameters, as well as their non-linear changes with age. Thus it may be better suited to analyze complex biological aging systems than statistical measures that only estimate linear relations. We show that individual parameters, including age, often show little correlation with heart disease. Yet in combination, the correlation improves dramatically. For diagnostic parameters specific for heart disease the increase in the correlative capacity thanks to the combination of diagnostic parameters, is less pronounced than for the less specific parameters. Age shows the highest influence on the presence of disease among the non-specific parameters and the combination of age with other diagnostic parameters substantially improves the correlation with the disease status. Hence age is considered as a primary "metamarker" of aging-related heart disease, whose addition can improve diagnostic capabilities. In the future, this methodology may contribute to the development of a system of biomarkers for the assessment of biological/physiological age, its influence on disease status, and its modifications by therapeutic interventions.

Key words: biomarkers of aging, biomarkers of disease, system aging, normalized mutual information, in silico assessment of anti-aging interventions

Non-communicable chronic diseases are the greatest cause of mortality in the world, yearly claiming more than 34.5 million lives worldwide (66% or 2/3 of global deaths, or nearly 100,000 deaths daily) [1]. Hence major efforts are directed toward their alleviation. Yet, a crucial point is often missing in these considerations, namely, the due emphasis on the fact that these diseases are age-related diseases, and their main risk factor is not necessarily related to environmental risks or life-style choices, but to the aging process itself!

The degenerative aging process lies at the basis of most processes of chronic pathogenesis. Thus some of the basic processes of aging include such processes as somatic mutations, cross-linkage, loss of viable stem cell populations, and impairment of the immune function which increases the susceptibility of the elderly to severe infectious and communicable diseases and reduces their responsiveness to vaccination, while at the same time aggravating the inflammatory damage to their tissues (the so-called "inflammaging") [2]. And these are precisely the main causative factors for the major non-communicable diseases, such as diabetes, neuro-degenerative diseases, cancer, and heart disease [3-6]. Moreover, the processes of aging exacerbate and reinforce the effects of other risk factors of non-communicable diseases.

The relation between non-communicable diseases and senescence is exacerbated by the fact that the world

***Correspondence should be addressed to:** Ilia Stambler, Department of Science, Technology and Society, Bar Ilan University, Ramat Gan 52900, Israel. Email: ilia.stambler@gmail.com

population is rapidly aging. Between 2000 and 2050, the proportion of the world's population over 60 years is expected to double from about 11% to 22%. The absolute number of people aged 60 years and over is expected to increase from 605 million to 2 billion over the same period [7]. Yet, this relation is often underappreciated. Thus even some of the most comprehensive analyses of risk factors for chronic diseases do not include age. At best, various risk factors are assessed for different ages [8]. Yet age itself generally is not quantitatively considered either as an independent or linked risk factor. However, the effects of age may outweigh the seemingly well established biomarkers, diagnostic parameters and risk factors, which often stop being predictive or show unexpected behavior in higher age [9,10].

There is an appreciation that the incidence of non-communicable diseases increases with age steeply, unlike the effects of other environmental and life-style factors whose influence may be considered steady [11]. Yet, the exact weight of age in relation to other risk factors remains uncertain. Hence, there is a need to be able to determine this weight in order to provide a fuller diagnostic and prognostic assessment for age-related diseases and design interventions that would be able to affect the entire array of risk factors, rather that some single, unrelated and purely symptomatic biological and physiological markers.

Such an ability would be especially valuable for heart disease, the main age-related disease and cause of death in the world [1]. As of 2010, it was estimated that the cardiovascular and circulatory diseases represented the largest proportion among all causes of mortality, causing about 15.6 million deaths, or nearly 30% of the total 52.8 million deaths globally, mainly due to ischemic heart disease (13.3%), closely followed by ischemic and hemorrhagic stroke (11.1%) [1]. Yet, it is also known that cardiovascular diseases, and ischemic heart disease in particular, can be highly susceptible to therapeutic and lifestyle interventions, capable of dramatically extending the health and longevity of the subjects [12]. Hence it is of primary importance to be able to assess the entire array of risk factors as well as the effects of therapeutic interventions on the risk factors, either individually or in combinations, including age. If age is the main risk factor, then it may well be that the primary target of the therapeutic and lifestyle intervention would be the aging process itself [13].

Here we apply the information theoretical measure of normalized mutual information (uncertainty coefficient) to determine precisely the weight of various risk factors in heart disease, and the particular weight of age as a risk factor, individually and combined with other factors, on the given sample and range of parameters. To do so, we introduce the concepts: “combined” or “general marker” (parameter) and “metamarker”. The “combined” or “general marker” (parameter) is a group of biomarkers whose combined influence on the disease under consideration, substantially exceeds the influence of every separate marker in the group. The “metamarker” is the biomarker (parameter) whose presence in the combined markers substantially increases the influence of those combined markers. The problem can be stated as follows: There is a group of parameters (markers) and several sets, related to the disease under consideration. As a result of the sets analysis, we need to obtain such a partition of the group of parameters, in which every cluster contains parameters “related in the same measure” with the disease under consideration. Then, using the obtained partition of the group of parameters, we need to find the combined markers and the metamarkers of the disease under consideration.

MATERIALS AND METHODS

Algorithm of partition of a set of parameters

Let there be K sets, and the data in each set are represented in the form of Table 1, where each subject of the set is described by n parameters. The set of n parameters needs to be partitioned into subsets, where each subset contains parameters correlated in the same measure with the disease.

The partition algorithm consists of 4 procedures:

1. Discretization of continuous parameters.
2. Construction of a table of the parameters’ influences on the disease (the table of the parameters’ values of normalized mutual information).
3. Construction of a table of the parameters’ ranks.
4. Partition of the parameters.

The detailed partition algorithm is as follows:

1. Discretization of continuous parameters.

This procedure transforms parameters having continuous values into parameters having discrete values. The discretization can be done taking into account the properties of the parameters and their biological and physiological meaning (above or below a certain salient physiological threshold) [14]. If the parameters do not have the corresponding properties or we are unaware of them, then we can use the formal rules of discretization [15].

2. Construction of the table of the parameters’ uncertainty coefficients (values of normalized mutual information)

Table 1. The presentation form of the sample dataset

	Parameter 1 (X_1)	**Parameter 2 (X_2)**	**…**	**Parameter n (X_n)**	**Disease**
Subject 1	x(1,1)	x(1,2)	...	x(1,n)	y(1)
Subject 2	x(2,1)	x(2,2)	...	x(2,n)	y(2)
...	...	...		...	...
Subject m	x(m,1)	x(m,2)	...	x(m,n)	y(m)

For the parameter x_i where 1<i<n for each set j where 1<j<k, we calculate the value of normalized mutual information c_{ij} [16, 17].

$$c_{ij} = \frac{I(x_i; y)}{H(y)} = \frac{H(x_i) + H(y) - H(x_i, y)}{H(y)}$$

where $H(x_i)$, $H(y)$, $H(x_i, y)$ are entropies of random values x_i , y , $x_i \times y$, respectively. We construct Table 2, n × k of the uncertainty coefficients $[c_{ij}]$.

Table 2. Values of Normalized Mutual Information (C)

	Datasets		
Parameters	**Hungary**	**Va Long Beach**	**Cleveland**
P1 age	0.01902	0.02389	0.03124
P3 cp	0.15316	0.03541	0.12281
P4 trestbps	0.01392	0.02105	0.00786
P5 chol	0.02352	0.01895	0.00564
P6 fbs	0.01749	0.00543	0.01104
P7 restecg	0.01699	0.02103	0.02505
P8 thalach	0.04772	0.00933	0.05168
P9 exang	0.14928	0.04466	0.07769
P10 oldpeak	0.15251	0.03931	0.07685

For the description of parameters, see "Materials and Methods. Case Materials"

Properties of the uncertainty coefficient c_{ij} [17, 18]:

1) $0 \le c_{ij} \le 1$;

2) $c_{ij} = 0$ if and only if x_i and y are mutually independent in the set j;

3) $c_{ij} = 1$ if and only if there exists a functional relationship between x_i and y in the set j.

Thus, values of the uncertainty coefficient (normalized mutual information) closer to zero indicate a smaller degree of correlation, while the coefficient values closer to 1 indicate a larger degree of correlation.

3. Construction of the table of the ranks of parameters.

For each column of the table c_{ij}, we rank its elements and assign rank 1 to the smallest element of the column. We obtain the table $n \times k$ of ranks $[r_{ij}]$, where each column of the matrix contains ranks from 1 to n.

We estimate the influence of the parameter x_i on the disease under consideration, compared to other parameters, by the sum of elements i of the row of the table $[r_{ij}]$.

In other words, the parameter the least correlated with the disease receives the rank 1, while the parameter the most correlated with the disease receives the greatest numerical rank. Then the ranks from different datasets for particular parameters are summated.

4. Partition of the parameters.

Consider Table $[r_{ij}]$ as the Friedman statistical model [19] and examine the row effect of this table. If the row effect exists, then for the clustering, we use the Newman-Keuls test of multiple comparisons [20]. In this way, we determine clusters or groups of parameters correlated with the disease in a relatively larger or smaller extent.

Estimation of the correlation of a combined marker with disease

In order to estimate the correlation between a combined marker and disease, we need to estimate the combined correlation of all the markers comprising the combined marker with the disease under consideration. For a combined marker comprised of two markers, this is done in the following way:

Let the combined marker Z be comprised of two discrete markers z_1 and z_2, while the marker z_1 assumes two values: 0 and 1, and the marker z_2 assumes three values: 0, 1 and 2. Then the correlation of the combined marker Z with the disease under consideration is estimated by the correlation of a "single marker" assuming 6 values in accordance to the values of the single markers z_1 and z_2 : (0,0) – 0, (0,1) – 1, (0,2) – 2, (1,0) –3, (1,1) – 4, (1,2) – 5. We proceed in the same way for combined markers comprised by more than two markers.

It is important to note, that for the analysis of combined markers, there is a need for datasets much larger than for the analysis of single markers. Therefore, for the analysis of combined markers, we used the "Cleveland dataset" consisting of 297 patients and the most thorough set of parameters (see the section below "Case Materials"). Such a sample size allows us to analyze triple combined parameters.

Case Materials

For the heart disease assessment, we used the Heart Disease Data Set from the University of California, Irvine (UCI) Machine Learning Repository [21]. The dataset includes healthy subjects and heart disease patients, aged 34-77. The entire dataset contains 76 attributes, but all published experiments refer to using a subset of 14 of the parameters. Here we also use the maximal set of 14 parameters (markers), or less when no data were available for a specific subset. The parameters (P) are: P1 - Age (years); P2 - Sex; P3 - Chest pain type (*cp*) ; P4 - Resting blood pressure (in mm Hg on admission to the hospital, *trestbps*); P5 - Serum cholesterol in mg/dl (*chol*); P6 - Fasting blood sugar (discretized above and below 120 mg/dl, *fbs*); P7 - Resting electrocardiographic results (discretized, *restecg*); P8 - Maximum heart rate achieved (*thalach*); P9 - Exercise induced angina (*exang*); P10 - ST depression induced by exercise relative to rest (*oldpeak*); P11 - The slope of the peak exercise ST segment (*slope*); P12 - Number of major vessels colored by fluoroscopy (discretized, *ca*); P13 - Thalium heart scan (normal, fixed defect, reversible defect, *thal*); P14 - The predicted attribute - diagnosis of heart disease (angiographic disease status, *num*) with value 0 for diameter narrowing < 50% and value 1 for diameter narrowing > 50%, in any major vessel.

The overall dataset contains 4 databases concerning heart disease diagnosis. The data were collected from the following four locations: 1. Cleveland Clinic Foundation (Cleveland data); 2. Hungarian Institute of Cardiology, Budapest (Hungarian data); 3. V.A. Medical Center, Long Beach, CA (Long Beach VA data), 4. University Hospital, Zurich, Switzerland (Switzerland data). Hence the data will be referred to according to the location.

This study focuses on the role of age as a potential risk factor for the development of heart disease. In order to compare the influence of the parameter "age" (P1) with the influences of other parameters on the disease, we

considered three of the datasets and selected the parameters – P3 (Chest pain type), P4 (Resting blood pressure), P5 (Serum cholesterol), P6 (Fasting blood sugar), P7 (Resting electrocardiographic results), P8 (Maximum heart rate), P9 (Exercise induced angina) and P10 (ST depression) – that were the most fully represented in the subsets. After removing from consideration the subjects with missing parameter values, we obtained the samples of the following sizes: Hu (Hungary) – 261 people, Va (VA Long Beach) – 133 people, Cl (Cleveland) – 297 people. Further, for the analysis of combined markers, including age, we focused on the largest and most complete Cleveland dataset, using all the 14 parameters represented.

RESULTS

Partition of the parameters

Now we consecutively perform the four procedures of the partition algorithm.

1. First, we perform the discretization of the continuous parameters. For the Hungarian, Long Beach and Cleveland datasets, the discretization of the age (P1) was: under 50, 50-59, 60 and over. The discretization thresholds for the other continuous parameters were selected as roughly corresponding to the common clinical diagnostic ranges. For the parameter P4 (Resting blood pressure), the discretization was: less than 140, equal and above 140; for the parameter P5 (serum cholesterol) – less than 200, equal and above 200 and less than 240, equal or over 240; for the parameter P8 (maximum heart rate) – below 160, equal and above 160; for the parameter P10 (ST depression) – less than 1, equal and above 1.

Other parameters were already discretized in the original dataset as follows: P2 (sex) - 1 = male; 0 = female; P3 (chest pain type, *cp*) - Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic; P6 (fasting blood sugar, *fbs*) > 120 mg/dl, 1 = true; 0 = false; P7 (resting electrocardiographic results, *restecg*) - Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria; P9 (exercise induced angina, *exang*) - 1 = yes; 0 = no; P11 (the slope of the peak exercise ST segment slope, *slope*) - Value 1: upsloping, Value 2: flat, Value 3: downsloping; P12 (number of major vessels colored by fluoroscopy, *ca*) - 0-3; P13 (Thalium heart scan, *thal*) - 3 = normal; 6 = fixed defect; 7 = reversible defect. Table 1 shows the general form of data presentation.

2. Then we compute the matrix c_{ij} of uncertainty coefficients. We obtain Table 2.

3. We rank entries of each column of the coefficients matrix c_{ij}. We obtain the matrix $[r_{ij}]$ shown in Table 3.

4. Finally, we consider Table 3 as the Friedman statistical model [19] and examine the row effect of this table.

Table 3. Parameter ranks

Parameters	Hungary	Va Long Beach	Cleveland	Sum of ranks
P1 age	4	6	5	15
P3 cp	9	7	9	25
P4 trestbps	1	5	2	8
P5 chol	5	3	1	9
P6 fbs	3	1	3	7
P7 restecg	2	4	4	10
P8 thalach	6	2	6	14
P9 exang	7	9	8	24
P10 oldpeak	8	8	7	23

Hypotheses:
H0: There is no row effect ("null hypothesis").
H1: The null hypothesis is invalid.
Critical range. The sample is "large", therefore the critical range is the upper 5%-range of the χ^2_8 distribution. Calculation of the χ^2 -criterion [19] gives $\chi^2 = 19.28$. The critical range is $\chi^2_8 > 15.51$. Since $19.28 > 15.51$, the null hypothesis with respect to Table 3 is rejected. Thus, according to the Friedman test, the row effect exists.

Hence, there is a difference between the rows under consideration.

For clustering (partition of parameters), we use the Newman-Keuls test for multiple comparisons [20]. For α_T =0.05 (α_T is the probability at least once to erroneously identify differences) we obtain the critical range for the comparison interval 2 equal 3.39 and $|R_j - R_{j+1}| > 3.39$ where Rj and Rj+1 are the elements of the column "Sum of ranks" in the j-th and (j+1)-th rows of Table 4 respectively (in other words, Rj and Rj+1 are the parameters represented by sums of ranks of the corresponding samples).

By the multiple comparisons, we construct the clustering (partition) shown in Table 4. The obtained clustering possesses the following properties: For two neighboring clusters of Table 4, the smallest element of one cluster and the greatest element of another cluster located nearby are significantly different (α_T =0.05); Elements belonging to the same cluster do not differ from each other (α_T =0.05). These properties allow us to categorize several groups of parameters according to their extent of correlation with disease.

Table 4. Parameters partition

No.	Clusters	Parameters	Sum of ranks
1	Cluster 1	P3 cp	25
2		P9 exang	24
3		P10 oldpeak	23
4	Cluster 2	P1 age	15
5		P8 thalach	14
6	Cluster 3	P7 restecg	10
7		P4 trestbps	9
8		P5 chol	8
9		P6 fbs	7

The parameters' influence on the disease

Table 4 shows the partition of parameters P1, P3, …, P10, according to the estimate of their correlation with the disease (Parameter P14). The parameters P3 (Chest pain type), P9 (Exercise induced angina) and P10 (ST depression, a salient marker of ischemia) show the strongest correlation with the disease. This is little surprising, as these parameters are directly indicative of a clinical, even severe state of heart disease. Yet, out of the parameters less directly connected with the disease, the influence of age (Parameter P1) is the strongest (the second cluster of Table 4). Right next to it is such a salient diagnostic marker as the maximum heart rate achieved (Parameter P8, Cluster 2). The influence of age on heart disease is much greater than the influence of the parameters comprising the third cluster: P4 (Resting blood pressure), P5 (Serum cholesterol), P6 (Fasting blood sugar) and P7 (Resting electrocardiographic results) that are routinely used for assessing heart condition. Thus the parameter "age" contains the largest amount of information about heart disease compared to all non-specific parameters (i.e. parameters not directly related to a severe clinical state). Hence it is justified to consider the parameter "age" as a "metamarker", i.e. a biomarker (parameter) whose presence substantially increases the diagnostic capacity of other markers. And hence, we will consider combined (multivalent) markers, containing the parameter "age".

Combined Markers

For the assessment of combined markers we use the Cleveland database, as the most complete one, containing 14 parameters (the 14th parameter is the predicted parameter – the presence of heart disease). Table 5 shows the correlation of single markers with the disease. Table 6 shows the influence on the disease from combined double markers, each containing age and another marker. Table 7 shows the influence of combined markers, each comprised of two commonly used clinical diagnostic parameters, on the disease. Finally, Table 8 shows the influence of combined markers, each containing 3 markers including age. As can be seen, the combined markers are much more informative regarding the disease status than single markers.

Table 5. Influence of single parameters

No.	Parameters	Values of normalized mutual information (C)
1	P13 thal	0.1316
2	P12 ca	0.13083
3	P3 cp	0.12281
4	P9 exang	0.07769
5	P10 oldpeak	0.07685
6	P11 slope	0.07549
7	P8 thalach	0.05168
8	P2 sex	0.03164
9	P1 age	0.03124
10	P7 restecg	0.02505
11	P6 fbs	0.01104
12	P4 trestbps	0.00786
13	P5 chol	0.00564

For the description of parameters, see "Materials and Methods. Case Materials"

As Table 5 shows, the influence of single parameters on the disease is rather low, with the normalized mutual information values ranging from 0.1316 for the heart defect scan to 0.00564 for cholesterol. The parameters directly associated with the clinical, even severe heart conditions provide better correlation (mutual information) values than the less specific parameters. Thus the single parameter results can be roughly divided into two groups. The parameters specific for heart disease have a higher correlation with the disease (C): P13 (heart defect scan), P12 (number of colored blood vessels), P3 (chest pain), P9 (exercise induced angina), P10 (ST depression), P11 (ST segment slope). The C values in this group range from 0.1316 for the heart defect scan to 0.07549 for the ST segment slope. The less specific parameters have lower correlation with the disease: P8 (maximum heart rate), P2 (sex), P1 (age), P7 (resting ECG), P6 (fasting blood sugar), P4 (resting blood pressure), P5 (Cholesterol). In this group, the mutual information values range from 0.05168 for the maximum heart rate to 0.00564 for cholesterol.

Yet, the combined consideration of several markers improves the correlation values dramatically. As shown in Table 6, when considering any diagnostic parameter in combination with age, the correlation value of the combined marker is increased. Still the combinations of age with specific markers of heart disease rank higher. However, the increase in correlative value is not extremely pronounced. For example, for the heart defect scan (P13) alone the correlation with the disease is 0.1316. For age (P1) alone, the correlative value is 0.03124. Yet, for their combination (P1+P13) it is 0.18501. Yet, for the less specific parameters, the combined consideration with age substantially increases the correlation value. Thus, for fasting blood sugar (P6) alone the correlation is 0.01104. Yet for its combination with age, the correlation value is already 0.05269. Nonetheless, despite the apparent large relative increase, the obtained absolute value is still rather low and would not allow to make any reliable diagnostic conclusions.

Table 6. Combined influence of age (P1) with other parameters

No.	Parameters	Values of normalized mutual information (C)
1	P1, P12 ca	0.18501
2	P1, P13 thal	0.17928
3	P1, P3 cp	0.17886
4	P1, P9 exang	0.11853
5	P1, P11 slope	0.11404
6	P1, P8 thalach	0.09757
7	P1, P10 oldpeak	0.09735
8	P1, P2 sex	0.09135
9	P1, P7 restecg	0.06918
10	P1, P5 chol	0.06582
11	P1, P6 fbs	0.05269
12	P1, P4 trestbps	0.03981

Also when considering other double diagnostic markers (apart from age), the correlation value is substantially increased as compared to each marker separately (Table 7). For the double diagnostic markers, the highest correlation value is obtained for the combination of P12 (the number of major vessels colored by fluoroscopy, a clear and specific indicator of ischemic insufficiency) and P13 (Thalium heart scan, clearly and specifically indicating a heart defect). The combined influence for P12 and P13 is 0.27686. Yet, for each of these specific parameters, the correlation value is also relatively high: 0.1316 for P13 (heart scan) which is the highest rank among the single parameters, and 0.13083 for P12 (number of vessels) – the second highest rank (Table 5). Their combination further substantially increases the correlation value. For the less salient, yet still fairly specific parameters for heart disease, both the individual and combined correlation values are somewhat lower. Thus, for the maximal heart rate alone (P8), the correlation with heart disease is 0.05168 and for the rest ECG (P7) it is 0.02505, whereas their combined influence is 0.07135. For the even less specific parameters, both

individual and combined influences are lower still. Thus, for the resting blood pressure (P4) alone the correlation is 0.00786, and for cholesterol (P5) it is 0.00564. Yet, for their combination (P4+P5), the correlation is already 0.02157. Apparently the relative increase in correlation values is larger for the less specific than for the more specific parameters. Nonetheless, even the combination of two of the non-specific parameters produces rather small absolute correlation values. Hence it is preferable to add additional parameters to the combinations, starting from triple combined parameters.

Table 7. Influence of several double combined parameters on the disease status

No.	Parameters	Values of normalized mutual information (C)
1	P12 ca, P13 thal	0.27686
2	P3 cp, P9 exang	0.1726
3	P10 oldpeak, P11 slope	0.1174
4	P7 *restecg*, P8 *thalac*h	0.07135
5	P4 *trestbps*, P8 *thalac*h	0.06878
6	P5 *chol*, P8 *thalac*h	0.06877
7	P6 *fbs*, P8 *thalac*h	0.06405
8	P5 *chol*, P7 *restecg*	0.05368
9	P2 *sex*, P5 *chol*	0.05067
10	P4 *trestbps*, P7 *restecg*	0.04209
11	P6 *fbs*, P7 *restecg*	0.04082
12	P4 *trestbps*, P6 *fbs*	0.02443
13	P5 *chol*, P6 *fbs*	0.02298
14	P4 *trestbps*, P5 *chol*	0.02157

When considering triple combined parameters, their correlation with the disease is much greater than for separate or double parameters (Table 8). Thus for example, as noted, the obtained results show that the Parameter P5 (cholesterol) by itself shows almost no correlation with the disease (the value of normalized mutual information is 0.00564). Age and sex separately are also very weakly linked with the disease, with the normalized mutual information of 0.03124 for age and 0.03164 for sex. Yet the triple combined marker, containing the age, sex and cholesterol, is strongly correlated with the disease, with the normalized mutual information of 0.15166. It should be noted that further addition of diagnostic parameters into large combined markers would require larger samples.

Statistical analysis

In order to facilitate the interpretation of the results as well as provide a comparison of the proposed method with the more commonly used ones, we performed statistical analysis. For the three continuous parameters considered in this study – P4 (Resting blood pressure), P5 (Serum cholesterol) and P8 (Maximum heart rate) – we performed an additional analysis of the correlation of the parameters, age and disease, using the ANOVA method. For each parameter we considered 6 groups: 3 age groups of healthy subjects (<50, 50-59, 60+) and 3 corresponding age groups of heart patients. Since the sizes of the groups are significantly different and for each parameter the group dispersions are also different, then instead of the standard parametric single-factor ANOVA, we used the non-parametric Kruskal-Wallis single-factor ANOVA.

For each parameter we consider the hypotheses:
H0: all the 6 groups are equally distributed.
H1: the null hypothesis is rejected.
Critical range. The samples are "large", therefore the critical range is the upper 5%-range of the χ^2_5 distribution. The critical range is $\chi^2_5 > 11.07$.

The parametric Kruskal-Wallis criterion H [22] equals: for the parameter P4 (blood pressure) H=32.69; for P5 (cholesterol) H=14.31; and for P8 (maximum heart rate) H=67.42. For the parameters P4, P5 and P8, H > 11.07 and in all the three cases the null hypothesis is rejected. That is to say, the distributions of parameter values in the 6 groups are different. Thus, the ANOVA results are not at a discrepancy with the proposed method. Yet, the proposed method allows to obtain information that is in principle unobtainable by ANOVA or other linear statistical methods (see the Discussion).

DISCUSSION

The importance of taking into consideration the patients' age in diagnosis and treatment cannot be overestimated. In accordance with the Antagonistic Pleiotropy theory of aging, entirely opposite diagnostic results can be obtained with the same biomarker, and opposite therapeutic results can be produced with the same treatment for younger and older individuals [23-25]. According to the Antagonistic Pleiotropy theory of aging, biological processes adaptive and increasing the organism's reproductive fitness early in life, can become maladaptive and lead to senescent deterioration late in life. Hence, the same process, and the same marker manifesting that process, can be a sign of health in young age, and a sign of disease in old age. For example, a rapid accumulation of calcium early in life can be beneficial for bone and muscle development, hence increased stamina. However, later in life, enhanced calcium deposition can contribute to atherosclerosis. Similarly, high levels of testosterone may give a good edge in sexual competition, yet later in life may contribute to prostate growth [26-28]. Hence, it appears very important to always consider biomarkers and risk factors

for diseases in their weighted relation with age. The proposed methodology provides, for the first time, the capability to precisely weigh age as a risk factor for heart disease, as a prime example of age-related disease, either individually or in combination with other biomarkers.

Table 8. **Influence of triple combined parameters**

No.	Parameters	Values of normalized mutual information (C)
1	P1 *age*, P12 *ca*, P13 *thal*	0.3753
2	P1 *age*, P3 *cp*, P12 *ca*	0.36317
3	P1 *age*, P11 *slope*, P12 *ca*	0.3503
4	P1 *age*, P2 *sex*, P12 *ca*	0.28037
5	P1 *age*, P5 *chol*, P12 *ca*	0.2764
6	P1 *age*, P7 *restecg*, P12 *ca*	0.27107
7	P1 *age*, P7 *restecg*, P13 *thal*	0.26664
8	P1 age, P2 *sex*, P3 *cp*	0.26048
9	P1 *age*, P3 *cp*, P5 *chol*	0.2516
10	P1 *age*, P5 *chol*, P11 *slope*	0.21026
11	P1 *age*, P2 *sex*, P11 *slope*	0.20168
12	P1 *age*, P2 *sex*, P9 *exang*	0.19274
13	P1 *age*, P5 *chol*, P9 *exang*	0.18908
14	P1 *age*, P8 *thalac*h, P11 *slope*	0.18687
15	P1 *age*, P2 *sex*, P5 *chol*	0.15166

Currently, the most popular methods for estimating heart disease risk factors are scoring (e.g. [29]) and the method of logical regression (e.g. [30]). The gist of the scoring method is that each risk factor is assigned a score and the risk estimate is equal to the sum of particular factors' risk scores found in the patient under investigation. In the logical regression method, the risk is estimated using a linear equation. Yet, the combined influence of factors cannot be estimated by an algebraic sum, insofar as for many factors their combined influence can greatly exceed the influence of each factor separately as well as their sum, that is to say, there can be a cumulative effect ("the whole is greater than the sum of parts"). Such cumulative effects were found in the present study. Also the linear equations of the logical regression method are insufficient to describe relations in a complex biological system. Unlike the former methods of scoring and regression, the proposed approach provides a methodologically adequate quantitative estimate for the combined influence of weighted risk factors, including age, which uncovers the cumulative and non-linear influences and which also allows to compare those influences.

The current analysis showed that individual parameters exhibit rather weak correlations with the disease, even for such routinely used diagnostic parameters as cholesterol. In contrast, in combination with other parameters, especially age, they often provide a good correlation (normalized mutual information). This is reasonable, as aging and age-related diseases are extremely complex multifactorial processes that can be neither treated by a single "magic bullet" nor described by a single "magic word". Thus it could be expected that a combination of diagnostic parameters would increase the descriptive/diagnostic value, as was indeed shown by our results. (By implication, a combination of treatments addressing a combination of biomarkers of aging and disease might increase therapeutic benefits, but a verification of this supposition is far beyond the scope of the present work.)

As our results show, the correlative value was commonly increased by combining virtually any parameters relevant to the disease process. For parameters

specifically indicative of heart disease, such as those clearly indicating an ischemic process (e.g. ST depression or number of blood vessels) or advanced dysfunction (such as chest pain or exercise-induced angina), both the individual and combined correlative values were high. For the less direct and less specific parameters, such as blood cholesterol and sugar, blood pressure and heart rate, both the individual and combined correlative values were lower relative to the specific parameters. Yet, the combination of the non-specific parameters commonly produced a greater relative increase in correlative capacity (rise in normalized mutual information) than the combination of specific parameters (Tables 5-7). This is also reasonable, as with the specific parameters there is very little ambiguity regarding every single parameter, hence their combination does not improve the diagnostic result dramatically. On the other hand, with the less specific parameters, each parameter is highly ambiguous, yet together they provide an indication of interactions determining the functional state of the interrelated system. It should be noted, however, that, in absolute terms, double non-specific parameters still had rather low correlations with the disease status. This finding indicates the necessity to consider a larger number of diagnostic parameters.

Time (or age) was shown to be a particularly valuable parameter. Among all the less specific individual parameters, age was the most informative regarding the disease status, providing the highest degree of correlation. Furthermore, seemingly non-specific, its addition dramatically increased the diagnostic correlations in every case and for every other parameter. These findings illustrate the crucial role of the aging process for the development of age-related pathology.

The combination of age with cholesterol represents a particularly interesting case, also due to the recent controversies surrounding the pathogenic role of cholesterol. Thus, regarding cholesterol, it has been increasingly shown that general cholesterol, the LDL or HDL forms, and the use of cholesterol-lowering medications are of limited predictive value for heart disease or for mortality from heart disease [31]. Often various forms of cholesterol can be associated with protection against heart disease and extended longevity [32]. Furthermore, cholesterol has been suggested to be more predictive for heart disease for young individuals, and less predictive for older individuals [9,10], and more predictive for men and less for women [33]. The present findings show almost no correlation of general cholesterol with the disease (Normalized mutual information C=0.00564), while it is much stronger correlated with the disease when combined with age (C=0.06582), and in the triple combination with sex, the correlation is even stronger (C=0.15166). These findings further emphasize the importance of considering metamarkers and combined markers, especially taking into account the age of the patient. Yet, it should also be emphasized that these findings mainly illustrate the validity of the proposed methodology and a proof of principle. The validity of the actual specific epidemiological results will yet require a thorough verification on a much larger and more representative sample.

The present methodology, using information theoretical measures of correlation (normalized mutual information) is well suited to consider the combined influence of various parameters on the disease status, as well as the contribution of age. It is well recognized that in biological systems, the relations between parameters are non-linear. Yet many (perhaps even most) studies a priori assume linear relations and normal distribution of parameters, and use linear measures of correlation, such as the correlation coefficient [34, 35]. In particular, the common use of the correlation coefficient in the assessment of disease risk factors a priori assumes a Gaussian distribution of parameters and linearity of correlations. The present use of the uncertainty coefficient (normalized mutual information) as the measure of correlation [17] does not assume any limitations on the distribution and correlation of parameters. Using the uncertainty coefficient, interesting results have been obtained in medicine [17], and in particular in oncology [36-38]. The Newman-Keuls method for multiple comparisons has been further successfully used for the analysis of biomedical data [39-41].

It is especially important to consider the parameters' non-linearity when referring to age and aging. It is well recognized that biomarkers of "physiological" or "biological age" are better correlated with "chronological age" for young and healthy adults. Yet, at older ages, the correlations break down, mainly due to non-linearity of diagnostic parameters [42]. Yet, many biomarkers assays still attempt to use linear measures that do not fully correspond to physiological reality [35]. The present methodology takes into account the non-linear correlations between the diagnostic parameters, as well as their non-linear changes with age. Thus the information-theoretical methodology is well equipped to demonstrate the crucial weight of age as a risk factor in the formation of heart disease.

To provide a comparison with the commonly used methods, we performed non-parametric Kruskal-Wallis single-factor ANOVA, as ANOVA is currently the most popular method of parameter analysis. We found that the results of the ANOVA method are not at a discrepancy with the information-theoretical method. Yet, the parametric ANOVA is suitable for the analysis of

parameters having Gaussian distribution, while the non-parametric ANOVA is suitable for parameters whose values can be represented as ranks. Unlike ANOVA, the information-theoretical method, using mutual information, is suitable for discrete (nominal) parameters and can analyze any type of parameters, after transformation of continuous and ranked parameters into discrete ones. The information-theoretical method thus provides a quantitative dimensionless estimate of the influence (correlation) between parameters and this allows the comparison of different parameters, obtained in different types of experimental models, which is not possible with the ANOVA method. However, in some cases, it may be beneficial to supplement the information-theoretical method with statistical methods, such as ANOVA, to facilitate the interpretation of results, for example to establish trends of decrease or increase of particular values with age. Yet, only using the information-theoretical measure it becomes possible to establish precisely the weight of age as a risk factor, alone and in combination with other factors.

The assertion of the importance of age (aging) as a risk factor for heart disease may have far reaching implications for diagnosis and treatment as it may motivate the physicians to apply greater discrimination for aged patients. This assertion may even have implications for research policy and public health policy. There is a growing realization that a promising and cost-effective strategy to combat severe non-communicable diseases is to give a greater focus of health research from attempting to address individual diseases and symptoms to addressing their underlying root cause and main risk factor – the degenerative process of aging [43]. Such an approach has already yielded in the past valuable strategies to combat non-communicable diseases. Historical examples include probiotic diets, cell therapy and adjuvant immunotherapy that were born from biological research of aging [44]. Further emphasis on treating, delaying or even reversing the seemingly "general" and "systemic" biological processes of aging may likely produce not just a general improvement of the functional state of the aged, but also further advances in the treatment of specific age-related non-communicable diseases, such as heart disease. The current work, for the first time quantitatively demonstrating the weight of age (aging) as a risk factor for heart disease, gives further support to this approach. It further emphasizes the need to intervene into the basic aging processes for developing effective therapies for age-related diseases.

Conflict of interest

We have no conflict of interest.

References

[1] Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet, 380:2095-2128.

[2] Chung HY, Cesari M, Anton S, Marzetti E, Giovannini S, Seo AY, et al. (2009). Molecular inflammation: Underpinnings of aging and age-related diseases. Ageing Res Rev, 8:18-30.

[3] Lunn JS1, Sakowski SA, Hur J, Feldman EL (2011). Stem cell technology for neurodegenerative diseases. Ann Neurol, 70:353-361.

[4] Baylis D, Bartlett DB, Patel HP, Roberts HC (2014). Understanding how we age: insights into inflammaging. Longev Healthspan, 3:6.

[5] Dai DF, Chiao YA, Marcinek DJ, Szeto HH, Rabinovitch PS (2014). Mitochondrial oxidative stress in aging and healthspan. Longev Healthspan, 3:6.

[6] Zieman S, Kass D (2004). Advanced glycation end product cross-linking: Pathophysiologic role and therapeutic target in cardiovascular disease. Congest Heart Fail, 10:144-149.

[7] World Health Organization. Ageing and Life Course. Accessed June 2014, Retrieved from: http://www.who.int/ageing/en/.

[8] Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010, Lancet, 380:2224-2260.

[9] Jacobs JM, Cohen A, Ein-Mor E, Stessman J (2013). Cholesterol, statins, and longevity from age 70 to 90 years. J Am Med Dir Assoc, 14:883-888.

[10] Petersen LK, Kaare Christensen K, Kragstrup J (2010). Lipid-lowering treatment to the end? A review of observational studies and RCTs on cholesterol and mortality in 80+-year olds. Age Ageing, 39:674-680.

[11] Rae MJ, Butler RN, Campisi J, de Grey ADNJ, Finch CE, Gough M, et al. (2010). The demographic and biomedical case for late-life interventions in aging. Science Transl Med, 2 (40), 40cm21.

[12] Meadows J, Danik JS, Albert MA (2007). Primary prevention of ischemic heart disease. In: Antman EM, editor. Cardiovascular Therapeutics: A Companion to Braunwald's Heart Disease. Third edition. Philadelphia PA: Saunders Elsevier, 178-220.

[13] Yu R, Navab K, Navab M (2010). Near term prospects for ameliorating cardiovascular aging. In: Fahy GM, West MD, Coles LS, Harris SB, editors. The Future of Aging: Pathways to Human Life Extension. New York: Springer, 279-306.

[14] Nicolis G, Prigogine I. Exploring Complexity. New York: W.H. Freeman, 1990.

[15] Glass GV, Stanley JC. Statistical Methods in Education and Psychology. New Jersey: Prentice-Hall, 1970.

[16] Renyi A (1959). On measures of dependence. Acta Math Acad Sci Hungar, 10:441-451.
[17] Zvarova J, Studeny M (1997). Information theoretical approach to constitution and reduction of medical data. Int J Med Inform 45:65-74.
[18] Cover TM, Thomas JA. Elements of Information Theory. Second edition. New York: Wiley-Interscience, 2006.
[19] Conover WJ. Practical Nonparametric Statistics. New York: Wiley-Interscience, 1999.
[20] Glantz SA. Primer of Biostatistics. Fourth edition. New York: McGraw-Hill, 1994.
[21] UCI Machine Learning Repository. Heart Disease Data Set. Creators: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.; University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.; University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.; V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. Donor: David W. Aha. Accessed June 2014, Retrieved from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.
[22] Zar JH. Biostatistical Analysis. Fifth edition. New Jersey: Prentice Hall, 2010.
[23] Carter AJ, Nguyen AQ (2011). Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. BMC Med Genet, 12:160.
[24] Le Couteur DG, Simpson SJ (2011). Adaptive senectitude: the prolongevity effects of aging. J Gerontol A Biol Sci Med Sci, 66:179-182.
[25] Goto M (2008). Inflammaging (inflammation + aging): A driving force for human aging based on an evolutionarily antagonistic pleiotropy theory? Biosci Trends, 2:218-30.
[26] Williams GC (1957). Pleiotropy, natural selection, and the evolution of senescence. Evolution, 11:398-411.
[27] Rose MR. Evolutionary Biology of Aging. Oxford: Oxford University Press, 1991.
[28] Kirkwood T. Time of Our Lives: The Science of Human Aging. Oxford: Oxford University Press, 1999.
[29] Pocock SJ, McCormack V, Gueyffier F, Boutitie F, Fagard RH, Boissel JP (2001). A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. Br Med J, 323:75-81.
[30] Wahman K, Nash MS, Lewis JE, Seiger A, Levi R (2010). Cardiovascular disease risk factors in persons with paraplegia: the Stockholm spinal cord injury study. J Rehabil Med, 42:272-278.
[31] Ray KK, Seshasai SRK, Erqou S, Sever P, Jukema JW, Ford I, Sattar N (2010). Statins and all-cause mortality in high-risk primary prevention: a meta-analysis of 11 randomized controlled trials involving 65,229 participants. Arch Intern Med, 170:1024-1031.
[32] Milman S, Atzmon G, Crandall J, Barzilai N (2013). Phenotypes and genotypes of high density lipoprotein cholesterol in exceptional longevity. Curr Vasc Pharmacol [Epub ahead of print].
[33] Baron AA, Baron SB (2007). High levels of HDL cholesterol do not predict protection from cardiovascular disease in women. Prev Cardiol, 10:125-127.
[34] Foulley JL, Quaas RL (1995). Heterogeneous variances in Gaussian linear mixed models. Genet Sel Evol, 27:211-228.
[35] Horvath S (2013). DNA methylation age of human tissues and cell types. Genome Biol, 14:R115.
[36] Blokh D, Stambler I, Afrimzon E, Shafran Y, Korech E, Sandbank J, et al. (2007). The information-theory analysis of Michaelis-Menten constants for detection of breast cancer. Cancer Detect Prev, 31:489-498.
[37] Blokh D, Zurgil N, Stambler I, Afrimzon E, Shafran Y, Korech E, et al. (2008). An information-theoretical model for breast cancer detection. Methods Inf Med, 47:322-327.
[38] Blokh D, Stambler I, Afrimzon E, Platkov M, Shafran Y, Korech E, et al. (2009). Comparative analysis of cell parameter groups for breast cancer detection. Comput Methods Programs Biomed, 94:239-249.
[39] Wu SS, Wang W, Annis DH (2008). On identification of the number of best treatments using the Newman-Keuls Test. Biometr J, 50:861-869.
[40] Blokh D and Stambler I (2014). Estimation of heterogeneity in diagnostic parameters of age-related diseases. Aging Disease, 5 [Epub ahead of print].
[41] Blokh D (2013). Information-theory analysis of cell characteristics in breast cancer patients. Int J Bioinform Biosci, 3:1-5.
[42] Krutko VN, Dontsov VI, Smirnova TM (2005). Theories, methods and algorithms for diagnosing aging. ISA RAS Proceed, 13:105-143 (Russian).
[43] Goldman DP, Cutler D, Rowe JW, Michaud PC, Sullivan J, Peneva D, Olshansky SJ (2013). Substantial health and economic returns from delayed aging may warrant a new focus for medical research. Health Aff, 10:1698-1705.
[44] Stambler I (2014). The unexpected outcomes of anti-aging, rejuvenation and life extension studies: an origin of modern therapies. Rejuvenation Res, 17 [Epub ahead of print].