American Journal of Bioinformatics

Original Research Paper

Applying Information Theory Analysis for the Solution of Biomedical Data Processing Problems

[1]David Blokh and [2]Ilia Stambler

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

Article history
Received 22-11-2014
Revised 25-11-2014
Accepted 05-01-2015

Corresponding author:
Ilia Stambler,
Department of Science, Technology and Society, Bar Ilan University,
Ramat Gan, Israel
Tel/Fax: 972-3-961-4296
Email: ilia.stambler@gmail.com

Abstract: The use of information-theoretical methods can be highly valuable for the solution of biomedical data processing problems. Some of the problems that can be solved by those methods include: The assessment of the influence of diagnostic parameters, biomarkers and risk factors, on the emergence of disease; the discretization of diagnostic parameters; the analysis of a combined influence of a group of parameters; the partition of a group of diagnostic parameters according to the amount of diagnostic information contained in those parameters; the analysis of the parameters' heterogeneity or variability and more. To illustrate the solution of those problems, we use a data base on diabetes patients. There are grounds to believe that an increasing application of information-theoretical methodologies in biomedical research will lead to significant practical dividends for diagnosis and therapy.

Keywords: Information Theory, Bioinformatics, Biomedical Data Processing

Introduction

Biomedical data refer to complex processes with non-linear, stochastic and non-analytic characteristics. These characteristics are determined by the very nature of the object under study: The complex human organism. Therefore, the processes under study cannot be adequately described by simple models. Yet, adequate modeling is crucial for the processing of biomedical data.

Some of the major problems arising during the processing of biomedical data can be summarized as follows:

- Evaluation of the influence (or correlation) of diagnostic parameters, biomarkers and risk factors on the actual emergence of disease-establishing causal relations (Lim *et al*., 2012)
- Optimal discretization of diagnostic parameters-finding physiologically meaningful thresholds (Nicolis and Prigogine, 1990)
- Evaluation of a joint influence of a group of parameters, crucial for multi-parametric biological systems (Lim *et al*., 2012)
- Partitioning of a group of parameters by the amount of information contained in these parameters and selection of a subgroup of parameters containing the greatest amount of information about all the parameters of the group under study, selecting the most meaningful and economical diagnostic parameters (Preckova *et al*., 2012; Molina-Pena and Alvarez, 2012)
- Evaluation of informational variety (heterogeneity or dissimilarity) of a group of parameters, potential measure of system adaptation and homeostasis (Radtke *et al*., 2009; Lipsitz and Goldberger, 1992)

Here are some brief characteristics of the methods commonly used for the solution of the problems listed above, as well as some of the drawbacks of those methods:

- For the problem of evaluation of influence (correlation), in most cases a correlation coefficient is used, i.e., the linearity of relations and gaussianity (normality) of distributions are assumed (Zar, 2010)
- For the discretization of parameters and risk factors (continuous or ranked), formal methods are used irrespective of the processes under study (Glass and Stanley, 1970)
- For the evaluation of joint influences of parameters, regression models are used, that is, linear forms with quantitative variables are postulated (Reynolds *et al.*, 2003)
- To determine the interconnection between variables (structural dependence) and to reveal common factors, using the analysis of principal components and factor

Science Publications

analysis, the continuity of parameters and the possibility of parameters' presentation as linear forms are assumed (Afifi and Azen, 1979)
- To evaluate the heterogeneity, variability or dispersion of parameters, gaussianity of distributions is assumed (Foulley and Quaas, 1995)

Thus, the methods commonly used for the solution of biomedical data processing problems *a priori* assume the fulfillment of one or several of the following hypotheses:

- Continuity of parameters
- Gaussian (normal) distributions of parameters
- Linear relations between parameters and the possibility of presenting parameters in linear forms

However, as a rule, in biomedical processes, discrete parameters are present side by side with continuous parameters; continuous parameters are not Gaussian; the interrelations between parameters are not linear and many parameters cannot be represented as linear forms. Therefore, the application of conventional methods for the solution of biomedical data processing problems often yields unsatisfactory results.

In the present paper, a unified approach, i.e., the information-theoretical analysis, is used for the solution of these problems. This approach makes it possible to examine models containing both continuous and discrete parameters, thus allowing for good inter-operability between diverse biomedical models. Moreover, it does not impose any constraints on the distribution of parameters, their interrelations and presentations. The suggested approach has been used for the solution of data processing problems in oncology (Blokh *et al.*, 2007; 2008; 2009; Blokh, 2013) and cardiology and biogerontology (Blokh and Stambler, 2014; 2015). The method developed in (Blokh *et al.*, 2007) is presented in the monograph (Gutierrez Diez *et al.*, 2012).

It is important to note that, at present, the information-theoretical analysis is the only theoretically substantiated method for evaluating both the influence of a single risk factor and a joint influence of several risk factors on the occurrence of a disease, as compared to various statistical methods. It is also rigorously and formally defined, as compared to the various concepts of "machine learning". To illustrate the solution of the relevant problems of biomedical data processing, a database on diabetes patients is used (see the Materials and Methods section).

Materials and Methods

Below we present the methodologies that are used to apply information-theory analysis for the solution of particular classes of biomedical data processing problems.

Data Presentation

The information-theoretical methods allow the researchers to work with any kind of parameters, both continuous and discrete. The parameters are presented in the form shown in Table 1.

Sample

In the illustrative present case, we used 8 parameters related to diabetes diagnosis. Though, any number of any clinically relevant parameters can be processed using the present methods. For the present diabetes assessment, the Pima Indians Diabetes Database of the Johns Hopkins University was analyzed (UCI, 2014). The representative sample comprised 130 diabetes patients and 262 healthy subjects. All the subjects were women at least 21 years old of Pima Indian origin, from Arizona, US. All the cases of diabetes that were included in the database were Type 2 diabetes (Baier and Hanson, 2004). The data set included 8 parameters: P1-Number of times pregnant; P2-Plasma glucose concentration at 2 hours in an oral glucose tolerance test; P3-Diastolic blood pressure (mm Hg); P4-Triceps skin fold thickness (mm); P5-2 h serum insulin (mu U/ml); P6-Body mass index (weight in kg/(height in m)^2); P7-Diabetes pedigree function; P8-Age (years).

Classes of Data Processing Problems

The following classes of biomedical data processing problems were considered:

Estimation of the Influence of a Parameter on the Development of Disease, Establishing Causal Relations

In the following, in order to measure the influence of one random value on another, we use the normalized mutual information (the uncertainty coefficient). Below is the definition of normalized mutual information (the uncertainty coefficient).

Let X be a discrete random value with a distribution function

X	x_1	x_2	....	x_n
Q	q_1	q_2	....	q_n

Entropy of random value X is:

$$H(X) = -\sum_{i=1}^{n} q_i \log q_i$$

For 2 discrete random values: *X*, *Y*, the uncertainty coefficient (the normalized mutual information) equals (Renyi, 1959; Zvarova and Studeny, 1997):

$$c = \frac{I(X;Y)}{H(Y)} = \frac{H(X) + H(Y) - H(X,Y)}{H(Y)}$$

where, $H(X)$, $H(Y)$, $H(X,Y)$ represent entropies of random values X, Y and XY, respectively.

The uncertainty coefficient has the following properties (Zvarova and Studeny, 1997; Cover and Thomas, 2006):

- $0 \le c \le 1$
- $c = 0$ if and only if X and Y are mutually independent (no correlation between the parameters)
- $c = 1$ if and only if there exists a functional relationship between X and Y (a complete correlation)

In other words, the values of the normalized mutual information closer to unity indicate a better correlation between the parameters. Yet, unlike the correlation coefficient, the use of normalized mutual information implies no *a priori* linearity of relations and can be used for non-linear relations. Further, unlike the correlation coefficient, for the uncertainty coefficient $c(X,Y) \neq c(Y,X)$. That is to say, the influence of one parameter on the second is not equal to the reverse influence of that second parameter on the first one. That makes it easier to address the possible reversed causality problem that can emerge when using the correlation coefficient which makes no such distinctions. Furthermore, the uncertainty coefficient is a dimensionless measure, which allows comparing parameters from different systems and models.

The Optimal Discretization of Parameters, Finding the Physiologically Relevant Thresholds

The questions of discretization (or boundary setting) arise when, alongside discrete parameters, there are also considered continuous parameters, or when the discretization of a parameter does not match the parameter's biological significance or the setting of the problem. An optimal discretization of parameter X relative to a discrete parameter Y is such a discretization in which the uncertainty coefficient $c(X,Y)$ assumes the maximal values. Below we will show the optimal discretization of the parameters under consideration.

Assessment of a Combined Influence of Parameters, Vital for Multi-Parametric Biological Systems

Let there be parameters related to Disease Y, designated as X_1, X_2,..., X_m, having the categories α_1, α_2,…,α_m respectively. Let us consider parameter $Z = X_1 \times X_2 \times \ldots \times X_m$, with $\alpha_1 * \alpha_2 * \ldots * \alpha_m$ categories. Then the combined influence of the parameters (biomarkers) X_1, X_2,..., X_m on the disease Y will be the influence of the parameter Z on the disease Y, that is, the value of the uncertainty coefficient $c(Z,Y)$. We will term the parameter Z: "*general/combined parameter*" or *"general/combined marker"*.

Let the combined marker Z be comprised of two discrete markers X_1 and X_2, while the marker X_1 assumes two values: 0 and 1 and the marker X_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 X_1 and X_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.

We should note a very important property of the combined influence of a group of parameters. Let, for example, $Z = X_1, X_2,..., X_m$, then $r(X_i,Y) \le r(Z,Y)$ for all $1 \le i \le m$. That is to say, the combined influence of a group of parameters is greater or equal than the influence of any parameters from this group (Gel'fand and Yaglom, 1957). This property emphasizes the adequacy of the metrics under consideration, insofar as the influence of several risk factors on the emergence of disease is always greater than the influence of some single factor among many.

Selecting a Sub-Group of Parameters, Containing the Largest Amount of Information Regarding all the Parameters in the Group Under Consideration, Selecting the Most Meaningful and Economical Diagnostic Parameters

First, we shall formulate the problem. Assume that the initial data on n objects are presented in the form of a $n \times m$ array $[a_{kj}]$ (in the form of Table 1), where each row k is an object described by m discrete parameters. It is needed to find a parameter or a subgroup of parameters containing the greatest amount of information about all m parameters.

The algorithm for selecting a subgroup of the most informative parameters from the entire group of parameters includes three procedures. A short description of each procedure is as follows. A more complete description of the application of information-theory analysis for the selection problem is presented elsewhere (Blokh, 2012).

1. Construction of the Uncertainty Coefficients Matrix: For i-th and j-th parameters $1 \le i,j \le m$, we calculate the uncertainty coefficient c_{ij} and construct $m \times m$ uncertainty coefficients matrix $[c_{ij}]$.

2. Construction of the Rank Matrix: For each column of the matrix $[c_{ij}]$, we rank its elements and assign rank 1 to the smallest element of the column. We obtain the matrix $m \times m$ of ranks $[r_{ij}]$, where each column of the matrix contains ranks from 1 to m.

We estimate the amount of information about all the m parameters contained in the i-*th* parameter by the sum of all the entries of the i-*th* row of the matrix $[r_{ij}]$.

Table 1. The presentation form of the sample dataset

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

3. Application of the Multiple Comparison Method: We apply the multiple comparison method to the sums of $[r_{ij}]$ matrix rows (Glantz, 2001). This gives a clustering of parameters that includes the desired subgroup of parameters containing the largest amount of information about all the other parameters in the group.

The most informative sub-group (the highest ranking cluster) can serve as a more economic diagnostic measure than the entire group, as it already contains the largest amount of information about all the other elements of the group. Moreover, this sub-group can serve to estimate causal connections and pathways among different elements of the entire group, as it shows the exact amount of information contained in the sub-group of elements regarding other elements of this group. The weights of relations between the parameters in the group or network can thus be estimated.

Estimation of the Information Variability (Heterogeneity) of a Group of Parameters, a Potential Measure of System Adaptation and Homeostasis

Let $M_1, M_2, \ldots M_k$ $1 \le k \le m$ be the clustering of the set of parameters M and $| M_i |$ the number of parameters in the set M_i and $|M| = m$. We shall estimate the heterogeneity (information variability) of the set of parameters, using the normalized Shannon entropy (Alter *et al.*, 2000):

$$S = -\frac{1}{\ln m}\sum_{i=1}^{k}\left(| M_i | / m\right)\ln\left(| M_i | / m\right)$$

Properties of the normalized Shannon entropy:

- $0 \le S \le 1$
- S = 0 if and only if the only cluster is the initial set (there is no heterogeneity)
- S = 1 if and only if every element of the initial set is a cluster (the maximal heterogeneity)

Simply put, zero value of the normalized Shannon entropy would indicate a complete identity of all the parameters (complete homogeneity) with reference to information content variability. In contrast, the value of 1 would indicate that any individual parameter has nothing in common with any other (complete heterogeneity), also with reference to information content variability. The greater variability may serve to indicate the adaptability of the system, its range of variation and ability to maintain function in response to a disturbance or insult. Notice that the use of normalized Shannon entropy, with values ranging from 0 to 1 (unlike regular Shannon entropy which can be unlimited), has the advantage of non-dimensionality. Hence it can be used for comparing any model systems with any number and kind of parameters. The same advantage of non-dimensionality is provided by the use of normalized mutual information over regular mutual information.

Results and Discussion

Estimation of the Influence of a Parameter on the Development of Disease

Table 2 shows the uncertainty coefficients for the influence of single parameters on the appearance of diabetes. Not surprisingly, the parameters P2 (Plasma Glucose) and P5 (Serum Insulin) exert the highest influence or correlation with diabetes (the values of the Uncertainty Coefficient of 0.17761 and 0.12918 respectively). Indeed, these parameters are in fact definitive of diabetes. Age (P8) is the third best correlated parameter (C = 0.11212), showing the crucial role of the aging process for the emergence of Type 2 Diabetes. Body Mass Index (P6), commonly understood to indicate the level of obesity, is also quite informative on the emergence of diabetes. On the other hand, the less specific parameters, not directly related to the disease mechanism, were shown to be less indicative, such as Triceps Skin Fold thickness (P4, C = 0.03736), Number of times pregnant (P1, C = 0.03601), Pedigree Function (P7, C = 0.02615) and Diastolic Blood Pressure (P3, C = 0.01978). Thus the use of information theory allowed to obtain clinically meaningful analysis.

Table 2. The influence of single parameters on the appearance of diabetes, as shown by the Normalized Mutual Information (the Uncertainty Coefficient-NMI/C)

Parameter	NMI (C)
P2-Plasma Glucose	0.17761
P5-Serum Insulin	0.12918
P8-Age (years)	0.11212
P6-Body Mass Index (BMI)	0.07120
P4-Triceps Skin Fold thickness	0.03736
P1-Number of Times Pregnant	0.03601
P7-Diabetes Pedigree Function	0.02615
P3-Diastolic Blood Pressure	0.01978

The Optimal Discretization of Parameters-Finding Physiologically Relevant Thresholds

Figure 1 and 2 show the borders of discretization and the uncertainty coefficients corresponding to the discretization thresholds. The median values for healthy and diseased subjects are also shown for each parameter. The numbers of categories for each parameter, either two or three categories, were determined by the clinical significance of the parameters. That is to say, in some cases it makes sense to speak of parameters "above and below the norm", where "the norm" is understood as some approximate dividing point, hence we consider two categories. In other cases it may be more appropriate to see the norm as an extended interval, while the values outside of that interval can be seen as abnormal, hence 3 categories. Or else, there may be a middle interval where the distinction between diseased and healthy subjects is difficult, but becomes clearer outside that interval. Here "abnormality" is understood simply as enhanced correlation with the disease, while "normality" as higher correlation with health, as shown by the maximal values of the Uncertainty Coefficient. Figure 1 presents "point boundaries" while Fig. 2 presents "interval boundaries".

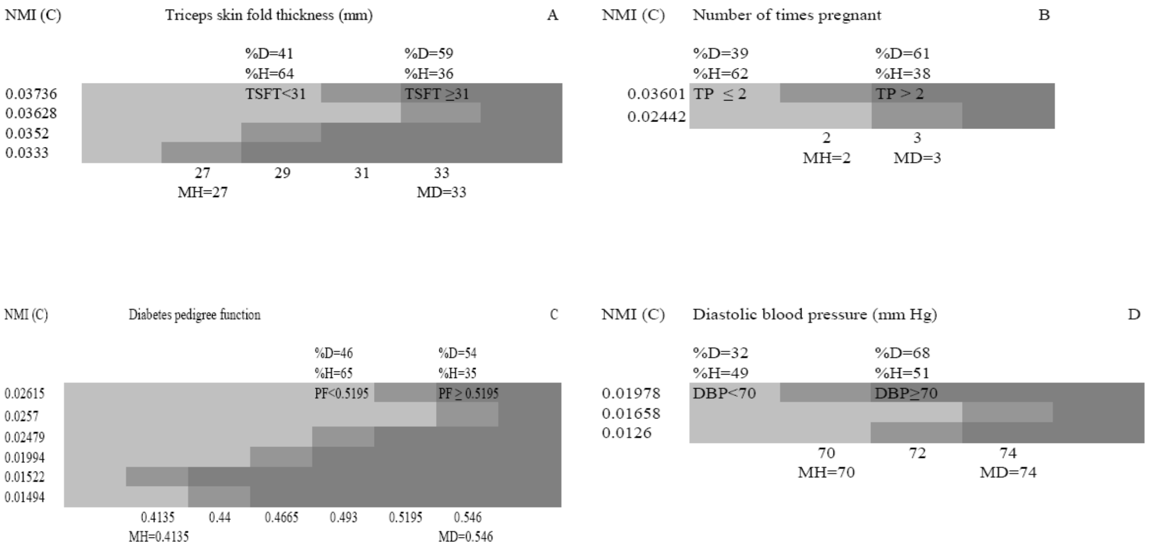

Fig. 1. "Point" boundaries (Diabetes). NMI-Normalized Mutual Information (Uncertainty Coefficient-C), %D-Proportion of Diseased, %H-Proportion of Healthy, MD-Median Diseased, MH-Median Healthy. Parameters: (A) P4-Triceps Skin Fold Thickness-mm (TSFT); (B) P1-Number of Times Pregnant (TP); (C) P7-Diabetes Pedigree Function (PF); (D) P3-Diastolic blood pressure (mm Hg)

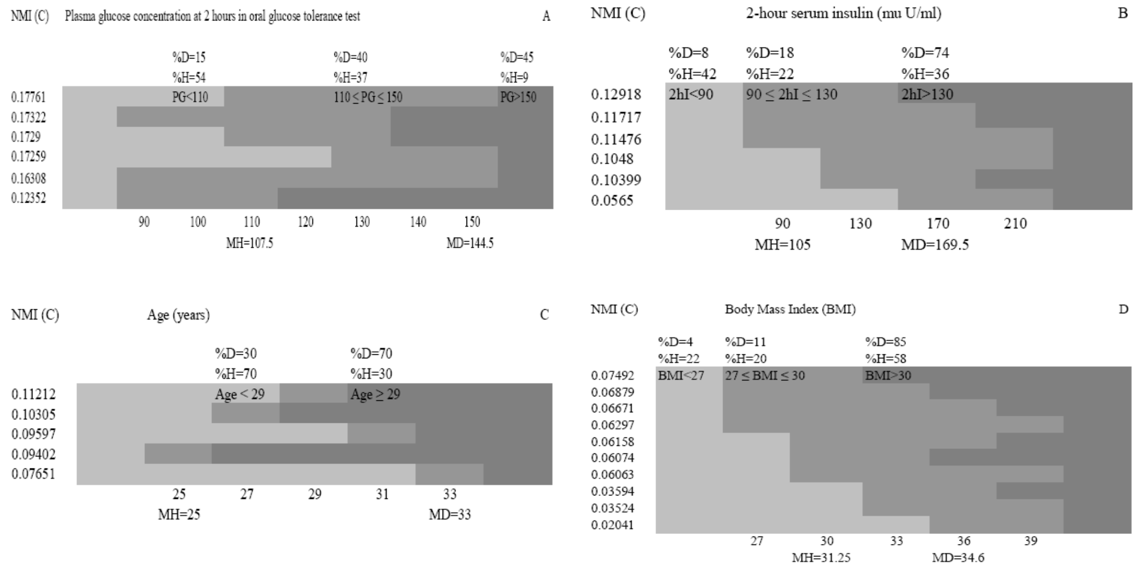

Fig. 2. "Interval" boundaries (Diabetes). NMI-Normalized Mutual Information (Uncertainty Coefficient-C), %D-Proportion of Diseased, %H-Proportion of Healthy, MD-Median Diseased, MH-Median Healthy. Parameters: (A) P2-Plasma Glucose Concentration at 2 h in oral glucose tolerance test; (B) P5-2 h serum insulin (mu U/ml); (C) P8-Age (years); (D) P6-Body Mass Index (BMI)

Consider the use of optimal discretization to establish "Point boundaries" using the example of Triceps Skin Fold Thickness (Fig. 1A). The entire range of the parameter values is divided into several steps with putative boundaries at the points 27, 29, 31 and 33 mm. The continuous values below and above the boundary are assigned discrete values 0 and 1 respectively. Then the normalized mutual information (C) is calculated for those discrete values in the relation to the absence or presence of disease (0 or 1). Among all the boundaries, the highest C values are found for the boundary at 31 mm (C = 0.037). Indeed, at this boundary, the distinction between the diseased and healthy subjects is most pronounced. Thus, of all the diseased subjects, 41% are found below the boundary and 59% above it. Among all the healthy subjects, 64% are found below the boundary and 36% above it. For other boundaries, with less values of Normalized Mutual Information, the distinctions are less clear.

Consider the use of optimal discretization to establish "Interval boundaries" using the example of Plasma Glucose Concentration at 2 h oral glucose tolerance test (Fig. 2A). The entire range of the parameter values is divided into different intervals, using a regular incremental step. Each time there are 3 intervals (lower, middle and upper) of different lengths and with different boundaries. We assign the discrete values 0, 1 and 2 to the lower, middle and upper interval, respectively. Then, for each set of intervals (boundaries) we calculate the normalized mutual information (C) in the relation to the absence or presence of disease (0 or 1). The highest normalized mutual information (C = 0.178) was when the lower interval was below 110, the middle interval was between and including 110 and 150 and the upper interval was above 150. With such interval boundaries, the distinction between the diseased and healthy subjects can be most reliably established. Thus, the greatest probability for health is below the glucose value of 110. Of all the healthy subjects, 54% are found in that interval and of all the diseased subjects only 15% are in this interval. On the other hand, the probability for disease is the greatest in the upper interval above 150, including 45% of the diseased subjects and 9% of the healthy subjects. For the middle interval ($110 \leq$ plasma glucose ≤ 150), the distinction is more ambiguous as the interval includes 40% of the diseased subjects and 37% of the healthy subjects. For other boundaries, with less values of Normalized Mutual Information, the distinctions are less clear. The boundaries for other parameters for the present sample were optimized in the same way and the boundaries can be optimized in the same way for any other diagnostic parameter for any disease.

Assessment of a Combined Influence of Parameters,Vital for Multi-Parametric Biological Systems

Tables 3-6 show the estimates for the influence of combined parameters containing two or three parameters. The last column of every table contains the sum of coefficients of parameters, included in the combined parameter. We shall state that a combined parameter has a cumulative effect, if the uncertainty coefficient (normalized mutual information) of the combined parameter is more than the sum of coefficients of parameters included in that combined parameter ("the whole is greater than the sum of parts").

Specifically Table 3 shows the combined influence of age (parameter P8) with other parameters on the appearance of diabetes. The addition of age strongly increases the informative value of the parameters, illustrating the importance of aging for the emergence of age-related diseases and the importance of considering the patients' age in diagnosis and treatment. Yet, also the combination of two other parameters increases the informative value as compared to individual parameters (Table 4) and in some cases, there is even a cumulative (synergistic or holistic) effect, when the combined influence of two parameters is greater than the sum of individual influences. For example, as Table 4 shows, there is a cumulative effect when combining the Number of Times Pregnant and Body Mass Index (BMI), or the Number of Times Pregnant with the Pedigree Function. This may indicate an important role the number of pregnancies may play in the rate of metabolic imbalance and aging. Also, cumulative effect is seen when combining the Pedigree Function with Diastolic Blood Pressure or the Pedigree Function with BMI. This may indicate the need to consider heredity not just as an independent risk factor but in combination with present environmental and life-style factors. In any case, the determination of cumulative effects is a unique capability of the information-theoretical approach and is impossible to achieve with the use of statistical methods. Yet, even without the emergence of cumulative effects, the combination of several diagnostic markers into a single, more informative marker, via the use of information theory, may provide valuable diagnostic and prognostic capabilities.

When combining 3 parameters, the informative value is increased even more, compared to individual or double parameters. Thus the combined consideration of age, together with two other diagnostic parameters, increases the informative value (Table 5). For the combination of Age, Glucose and BMI, the Normalized Mutual Information is 0.326, almost twice as much as for glucose alone, 3 times more than for age alone and almost 5 times more than for BMI alone. But also, as shown in Table 6, the combination of any other three parameters, increases the informative value, compared to individual or double parameters. Interestingly, here too the combination with the Number of Times Pregnant produces a cumulative effect.

Table 3. Combined influence of age (P8) with other parameters on the appearance of diabetes, as shown by Normalized Mutual Information (NMI/C). * Cumulative effect (the combined influence is greater than the sum of individual influences)

Parameters	NMI (C)	Sum
P8, P2-Plasma Glucose	0.24836	0.28973
P8, P5-Serum Insulin	0.21408	0.2413
P8, P6-Body Mass Index (BMI)	0.1747	0.18332
P8, P4-Triceps Skin Fold Thickness	0.1516*	0.14948*
P8, P7-Pedigree Function	0.1321	0.13827
P8, P1-No. of Times Pregnant	0.11766	0.14813
P8, P3-Diastolic Blood Pressure	0.1139	0.1319

Table 4. Combined influence of double diagnostic parameters on the appearance of diabetes, as shown by normalized mutual information (NMI/C). * Cumulative effect (the combined influence is greater than the sum of individual influences)

Parameters	NMI (C)	Sum
P2-Plasma Glucose, P6-Body Mass Index	0.2318	0.24881
P2-Plasma Glucose, P5-Serum Insulin	0.21344	0.30679
P2-Plasma Glucose, P4-Triceps Skin Fold Thickness	0.20571	0.21497
P2-Plasma Glucose, P7-Pedigree Function	0.19793	0.20376
P2-Plasma Glucose, P1-No. Times Pregnant	0.18986	0.21362
P2-Plasma Glucose, P3-Diast. Blood Pressure	0.18062	0.19739
P5-Serum Insulin, P6-Body Mass Index (BMI)	0.17524	0.20038
P5-Serum Insulin, P1-No. Times Pregnant	0.15446	0.16519
P1-No. Times Pregnant, P6-BMI	0.10995*	0.10721*
P6-Body Mass Index, P7-Pedigree Function	0.09836*	0.09735*
P3-Diast. Blood Pressure, P6-Body Mass Index	0.0863	0.09098
P4-Triceps Skin Fold, P6-Body Mass Index	0.08249	0.10856
P1-No. Times Pregnant, P7-Pedigree Function	0.06849*	0.06216*
P3-Diast. Blood Pressure, P4-Triceps Skin Fold	0.05207	0.05714
P1-No. Pregnancies, P3-Diast. Blood Pressure	0.0489	0.05579
P3-Diast. Blool Pressure, P7-Pedigree Function	0.04686*	0.04593*

Table 5. Combined influence of age (P8) with two other parameters on the appearance of diabetes, as shown by normalized mutual information (NMI/C). * Cumulative effect (the combined influence is greater than the sum of individual influences)

Parameters	NMI (C)	Sum
P8-Age, P2-Glucose, P6-BMI	0.32605	0.36093
P8-Age, P2-Glucose, P5-Insulin	0.31013	0.41891
P8-Age, P5-Insulin, P6-BMI	0.28608	0.31250
P8-Age, P2-Glucose, P4-Triceps Fold	0.28463	0.32709
P8-Age, P2-Glucose, P7-Pedigree	0.27352	0.31588
P8-Age, P1-No. Pregnancies, P2-Glucose	0.27159	0.32574
P8-Age, P2-Glucose, P3-Diast. Blood Pressure	0.26197	0.30951
P8-Age, P4-Triceps Fold, P5-Insulin	0.24493	0.27866
P8-Age, P5-Insulin, P7-Pedigree	0.23106	0.26745
P8-Age, P1-No. Pregnancies, P5-Insulin	0.22399	0.27731
P8-Age, P3-Diast. Blood Pressure, P5-Insulin	0.22362	0.26108
P8-Age, P6-BMI, P7-Pedigree	0.20905	0.20947
P8-Age, P1-No. Pregnancies, P6-BMI	0.19104	0.21933
P8-Age, P4-Triceps Fold, P6-BMI	0.18805	0.22068
P8-Age, P3-Diast. Blood Pressure, P6-BMI	0.17923	0.20310
P8-Age, P4-Triceps Fold, P7-Pedigree	0.17401	0.17563
P8-Age, P1-No. Pregnancies, P4-Triceps Fold	0.16071	0.18549
P8-Age, P3-Diast. Blood Pressure, P4-Triceps	0.15329	0.16926
P8-Age, P1-No. Pregnancies, P7-Pedigree	0.14443	0.17428
P8-Age, P3-Diast. Blood Pressure, P7-Pedigree	0.13576	0.15805
P8-Age, P1-No. Pregnancies, P3-Diast. Press.	0.12349	0.16791

Table 6. Combined influence of a selection of triple diagnostic parameters on the appearance of diabetes, as shown by Normalized Mutual Information (NMI/C). * Cumulative effect (the combined influence is greater than the sum of individual influences). Parameters: P1-Number of Times Pregnant, P2-Plasma Glucose, P3-Diastolic Blood Pressure, P4-Triceps Skin Fold thickness, P5-Serum Insulin, P6-Body Mass Index (BMI), P7-Diabetes Pedigree Function, P8-Age (years)

Parameters	NMI (C)	Sum
P3,P4,P5	0.16457	0.18632
P1,P4,P8	0.16071	0.18549
P1,P6,P7	0.15524*	0.13336*
P3,P4,P8	0.15329	0.16926
P1,P7,P8	0.14443	0.17428
P3,P7,P8	0.13576	0.15805
P1,P4,P6	0.12876	0.14457
P1,P3,P6	0.12741*	0.12699*
P1,P3,P8	0.12349	0.16791
P3,P6,P7	0.11453	0.11713
P4,P6,P7	0.11239	0.13471
P1,P4,P7	0.11137*	0.09952*

It is important to note that the present combined triple diagnostic markers could be reliably produced based on the current sample of several hundred subjects (130 diabetes patients and 262 healthy subjects). To produce combined diagnostic parameters composed of four individual parameters and more, there would be a need for much larger samples. Yet, for diagnostic purposes, any number of parameters, from even a limited sample, can be brought together into a diagnostic rule or diagnostic model (for example using a "decision tree" or "weighted Hamming Distance"). Such a rule would be based on all the individual values of Normalized Mutual Information of all the parameters involved in relation to the disease and can produce good diagnostic results. This capability was previously exemplified for the detection of breast cancer, building information-theoretical diagnostic models, combining a large set of parameters via the use of "decision trees" or "Weighted Hamming Distances" (Blokh *et al.*, 2007; 2008).

Selecting a Sub-Group of Parameters, Containing the Largest Amount of Information Regarding all the Parameters in the Group Under Consideration- or Selecting "the Most Informative Parameters"

We calculate the uncertainty coefficients c_{ij} $1 \leq i,j \leq 8$ for the parameters of the Database and construct the $[c_{ij}]$ matrix of uncertainty coefficients (Table 7).

We rank the $[c_{ij}]$ matrix (Table 7) columns and obtain the rank matrix $[r_{ij}]$ (Table 8). Here, the smallest uncertainty coefficient obtains the rank 1 and the largest has the rank 8.

We consider Table 8 as a Friedman statistical model (Conover, 1999) and examine the row effect of this table.

Hypotheses: H_0: There is no row effect ("null hypothesis"); H_1: The null hypothesis is invalid

Critical Range: The sample is "large"; therefore, the critical range is the upper 5%-range of χ_7^2 distribution.

Calculation of the χ^2-criterion (Glantz, 2001) gives $\chi^2 = 15$.

The critical range is $\chi_7^2 > 14.07$. Since 15>14.07, the null hypothesis with respect to Table 8 is rejected. Thus, according to the Friedman test, the row effect exists. Hence, there is a difference between the rows under consideration.

For multiple comparisons, we use the Newman-Keuls test (Glantz, 2001). We obtain $|R_j - R_{j+1}| > 5.544$, where R_j and R_{j+1} are the j-th and $(j+1)$-th elements of the "Sum of ranks" column of Table 8. Using the multiple comparisons method, we construct the parameter clustering shown in Table 9.

The obtained clustering has the following properties:

For two neighboring clusters of Table 9, the smallest element of one cluster and the greatest element of another cluster located nearby are significantly different (α_T = 0.05 where α_T is the probability at least once to erroneously identify differences);

Elements belonging to the same cluster do not differ from each other (α_T = 0.05).

The parameters of Cluster 1 contain the most information regarding all the 8 parameters (in the descending order of information content within the cluster). The parameters are: P2-Plasma glucose concentration at 2 h in an oral glucose tolerance test; P8-Age (years); P5-2 h serum insulin; P6-Body mass index. These results appear to fit medical intuition, as abnormal insulin and glucose concentration are the parameters definitive of diabetes; the body mass index appears as a major parameter affecting the body metabolism, while age (or aging) is a crucial contributor for the development of age-related diseases, including diabetes.

Table 7. The matrix [c_{ij}] of uncertainty coefficients for the database parameters

Parameter	P1	P2	P3	P4	P5	P6	P7	P8
P1	1.00000	0.02893	0.03360	0.00432	0.00797	0.00962	0.00002	0.20734
P2	0.04450	1.00000	0.05209	0.02757	0.17954	0.04579	0.01030	0.06914
P3	0.03335	0.03362	1.00000	0.02181	0.00460	0.02323	0.00005	0.07774
P4	0.00430	0.01783	0.02185	1.00000	0.02183	0.11504	0.00153	0.02098
P5	0.01202	0.17602	0.00698	0.03311	1.00000	0.04534	0.01814	0.04335
P6	0.01207	0.03736	0.02936	0.14514	0.03773	1.00000	0.00360	0.02441
P7	0.00002	0.00658	0.00005	0.00151	0.01182	0.00282	1.00000	0.00650
P8	0.20539	0.04452	0.07758	0.02090	0.02847	0.01927	0.00656	1.00000

Table 8. The rank matrix [r_{ij}] of the parameters uncertainty coefficients. The smallest uncertainty coefficient (normalized mutual information) has the rank 1 and the largest has the rank 8

Parameter	P1	P2	P3	P4	P5	P6	P7	P8	Sum of ranks
P1	8	3	5	2	2	2	1	7	30
P2	6	8	6	5	7	6	6	5	49
P3	5	4	8	4	1	4	2	6	34
P4	2	2	3	8	4	7	3	2	31
P5	3	7	2	6	8	5	7	4	42
P6	4	5	4	7	6	8	4	3	41
P7	1	1	1	1	3	1	8	1	17
P8	7	6	7	3	5	3	5	8	44

Table 9. Clustering of 8 parameters according to sums of uncertainty coefficients ranks. The highest ranking parameters clusters contain the largest amount of information

Clusters	Parameter name	Sum of ranks
Cluster 1	P2-Plasma Glucose	49
	P8-Age (years)	44
	P5-Serum Insulin	42
	P6-Body Mass Index	41
Cluster 2	P3-Diast. Blood Pressure	34
	P4-Triceps Fold Thickness	31
	P1-No. Times Pregnant	30
Cluster 3	P7-Diabetes Pedigree	17

Less information is contained in the parameters of Cluster 2 and 3 (in the descending order of information content). Those parameters are: P3-Diastolic blood pressure; P4-Triceps skin fold thickness; P1-Number of times pregnant; P7-Diabetes pedigree function. The less informative values of those parameters also appear as plausible, as they are less strongly related to the recognized specific clinical symptoms and mechanisms of diabetes.

We shall add to the 8 parameters the 9th parameter: the presence or absence of the disease and conduct the following procedures:

Calculate the uncertainty coefficients c_{ij} $1 \leq i,j \leq 9$ for the parameters of the Database and construct the [c_{ij}] matrix of uncertainty coefficients (Table 10).

Rank the [c_{ij}] matrix (Table 10) columns and obtain the rank matrix [r_{ij}] (Table 11).

Consider Table 11 as the Friedman statistical model (Conover, 1999) and examine the row effect of this table.

Hypotheses: H_0: There is no row effect (“null hypothesis”); H_1: The null hypothesis is invalid

Critical range: The sample is “large”; therefore, the critical range is the upper 5%-range of χ_8^2 distribution.

Calculation of the χ^2-criterion (Glantz, 2001) gives $\chi^2 = 20.97$.

The critical range is $\chi_8^2 > 15.51$. Since $20.97 > 15.51$, the null hypothesis with respect to Table 11 is rejected. Thus, according to the Friedman test, the row effect exists. Hence, there is a difference between the rows under consideration.

For multiple comparisons, we use the Newman-Keuls test (Glantz, 2001). We obtain $|R_j - R_{j+1}| > 5.88$, where R_j and R_{j+1} are the j-th and (j+1)-th elements of the “Sum of ranks” column of Table 11. Using the multiple comparisons method, we construct the parameter clustering shown in Table 12.

The obtained clustering has the following properties:

For two neighboring clusters of Table 12, the smallest element of one cluster and the greatest element of another cluster located nearby are significantly different ($\alpha_T = 0.05$).

Elements belonging to the same cluster do not differ from each other ($\alpha_T = 0.05$).

Table 10. The matrix $[c_{ij}]$ of uncertainty coefficients for the database parameters, including the presence of the disease itself as a parameter (diabetes-P9)

Parameter	P1	P2	P3	P4	P5	P6	P7	P8	P9
P1	1.00000	0.02893	0.03360	0.00432	0.00797	0.00962	0.00002	0.20734	0.03601
P2	0.04450	1.00000	0.05209	0.02757	0.17954	0.04579	0.01030	0.06914	0.17761
P3	0.03335	0.03362	1.00000	0.02181	0.00460	0.02323	0.00005	0.07774	0.01978
P4	0.00430	0.01783	0.02185	1.00000	0.02183	0.11504	0.00153	0.02098	0.03736
P5	0.01202	0.17602	0.00698	0.03311	1.00000	0.04534	0.01814	0.04335	0.12918
P6	0.01207	0.03736	0.02936	0.14514	0.03773	1.00000	0.00360	0.02441	0.07120
P7	0.00002	0.00658	0.00005	0.00151	0.01182	0.00282	1.00000	0.00650	0.02615
P8	0.20539	0.04452	0.07758	0.02090	0.02847	0.01927	0.00656	1.00000	0.11212
P9	0.03319	0.10641	0.01837	0.03462	0.07894	0.05229	0.02453	0.10431	1.00000

Table 11. The rank matrix $[r_{ij}]$ of the parameters uncertainty coefficients, including the presence of the disease as a parameter: P9. The smallest uncertainty coefficient (normalized mutual information) has the rank 1 and the largest has the rank 9

Parameter	P1	P2	P3	P4	P5	P6	P7	P8	P9	Sum of ranks
P1	9	3	6	2	2	2	1	8	3	36
P2	7	9	7	5	8	6	6	5	8	61
P3	6	4	9	4	1	4	2	6	1	37
P4	2	2	4	9	4	8	3	2	4	38
P5	3	8	2	6	9	5	7	4	7	51
P6	4	5	5	8	6	9	4	3	5	49
P7	1	1	1	1	3	1	9	1	2	20
P8	8	6	8	3	5	3	5	9	6	53
P9	5	7	3	7	7	7	8	7	9	60

Table 12. Clustering of 9 parameters according to sums of uncertainty coefficients ranks. The highest ranking parameters clusters contain the largest amount of information

Clusters	Parameter name	Sum of ranks
Cluster 1	P2-Plasma Glucose	61
	P9-Diabetes	60
Cluster 2	P8-Age (years)	53
	P5-Serum Insulin	51
	P6-Body Mass Index	49
Cluster 3	P4-Triceps Fold Thickness	38
	P3-Diast. Blood Presssure	37
	P1-No. Times Pregnant	36
Cluster 4	P7-Diabetes Pedigree	20

The parameters included in the highest ranking cluster in Table 12 (Cluster 1) contain the most information about the other parameters. Thus we find that the parameters P2 (glucose concentration) and P9 (presence and absence of the disease) contain the largest amount of information regarding all the other parameters in this selection. This is quite reasonable, as the parameter P9 is the very presence of diabetes, which includes all the parameters serving for its diagnosis. On the other hand, the glucose level is the most significant parameter in type 2 diabetes, which is clinically defined as impaired glucose utilization. Thus, using the information theoretical measures, we were able to rigorously and formally validate the significance of these parameters.

The ability to select the most informative parameters, either just by the value of normalized mutual information for individual parameters or their combinations (Examples 1 and 3 of this study), or by determining the amount of information each individual parameter or combination contains about all the other parameters in the group (the example of the current section), can have significant diagnostic utility. This ability would economize the diagnostic tasks, allowing the diagnostician to indicate the most meaningful parameters and perhaps discard the parameters whose information is already contained in the more informative ones. This capability could be especially useful in “OMICS” studies (genomics, metabolomics, proteomics, etc.) which may contain vast numbers of parameters, for a great part of which the diagnostic significance is uncertain. Determining the most informative parameters within those vast arrays may make those data more manageable and clinically applicable. Using this method, it may also be possible to enrich and elaborate biological mechanisms, pathways and networks, as it would allow the researchers to determine the amount of information one element in a pathway or network has about all the others and in this way determine the weights of mutual influences.

Estimation of the Information Variability (Heterogeneity) of a Group of Parameters-a Potential Measure of System Adaptation and Homeostasis

We shall estimate the heterogeneity (information variability of the parameters) for the clustering's shown in Table 9 and 12, using normalized Shannon entropy. For Table 9, the normalized Shannon entropy S = 0.468546 and for Table 12, S = 0.596563. That is to say, with the addition of the parameter "Diabetes" (P9) to the 8 diagnostic parameters, the heterogeneity of the obtained set of parameters increased. This may be due to several reasons. First of all, the parameter P2 (glucose concentration) contains the largest amount of information about the parameter P9 (diabetes). Hence, with the addition of the parameter "Diabetes" into the group, the parameter "Glucose" increased its information content about all the parameters to the largest extent, compared to the rest of the parameters and secondly, the parameter "Diabetes" itself contains a large amount of information about the other 8 parameters. As a result, the parameters "Glucose Concentration" (P2) and "Diabetes (P9) formed a new cluster, whose parameters contain the largest amount of information about the other parameters.

Yet another application of information theory to establish the heterogeneity (variability) of parameters related to diabetes was presented earlier, using the same database (Blokh and Stambler, 2014). Briefly, we used normalized Shannon entropy as the measure of heterogeneity or variability. The model employed 4 parameters: P2-plasma glucose concentration, P5-2 h serum insulin, P3-diastolic blood pressure and P6-body mass index. We determined the heterogeneity (the normalized Shannon entropy) for the entire set of those parameters for healthy individuals and diabetes patients of four age groups: 21-25, 26-29, 30-39 and 40-49 years old. The results are summarized in Table 13. Crucially, only young and healthy individuals exhibited high levels of entropy, indicating a high level of heterogeneity and variability, which can be interpreted as enhanced adaptability and homeostatic capacity. For older individuals (both healthy and diseased), same as for younger diseased individuals, the entropy values were diminished, indicating a greater homogeneity and a narrow range of change, which can be interpreted as a lessened ability for adaptation and homeostatic capacity. The common entropy change pattern also suggested a formal analogy between aging and aging-related disease, with reference to those particular parameters in the present sample.

Moreover, the loss of complexity, variability or heterogeneity, shown by the lower system entropy, has been suggested as a potentially powerful dynamic biomarker of disease and aging and as a potential metrics to test therapeutic interventions by measuring the therapy's ability to restore the entropy levels (Lipsitz and Goldberger, 1992; Li *et al.*, 2014). The latter studies focused on heart rate variability as a convenient surrogate measure of the system adaptive homeostasis. In the current example, we focused on several parameters connected with diabetes diagnosis. Yet, in fact, with the use of information theory, measuring a wide range of parameters, cross-sectionally and longitudinally, individually and in combinations, it may be possible to establish a truly comprehensive measure of youthful, healthy homeostasis and an evidence-based quantitative framework to assess therapeutic and anti-aging interventions by their effects on the homeostasis.

Table 13. Entropy (heterogeneity or variability) as a measure of age-related change and disease

Subjects (women, number)	Disease status	Age	Entropy
53	Healthy	21-25	0.527
22	Healthy	26-29	0.592
41	Healthy	30-39	0.583
18	Patients	40-49	0.000
24	Patients	21-25	0.000
18	Patients	26-29	0.000
34	Patients	30-39	0.000
26	Patients	40-49	0.000

Conclusion

The importance of using information theory in biomedical research has been recognized earlier and sometimes emphatically stressed. Thus, according to Paninski, "The mathematical theory of information transmission represents a pinnacle of statistical research: The ideas are at once beautiful and applicable to a remarkably wide variety of questions" (Paninski, 2003) and according to one of the pioneers of the use of information theory in biomedicine, Quastler (1958), "The basic concepts of information theory-measures of information, of noise, of constraint, of redundancy-establish the possibility of associating precise (although relative) measures with things like form, specificity, lawfulness, structure, degree of organization. … Closely related is the problem of destruction of orderliness. In biology, this is the problem of aging and decay" (Quastler, 1958). Nonetheless, the information-theoretical measures have not yet entered into routine use of biomedical researchers and practitioners.

Here we illustrate several applications of information theory for the solution of biomedical data processing problems that are commonly encountered in biomedical research and practice and for which information theory offers an adequate and often the only possible and grounded methodology. Those problems include: Evaluation of the influence (or correlation) of diagnostic parameters, biomarkers and risk factors, on the actual emergence of disease in order to assess causal relations; optimal discretization of diagnostic parameters in order to find physiologically meaningful thresholds and boundaries; evaluation of a joint influence of a group of parameters, necessary in order to describe complex multi-parametric biological systems; partitioning parameters by their information content and selecting subgroups of parameters containing the greatest amount of information about all the parameters of the group, which may be used to select the most meaningful and economical diagnostic parameters, as well as evaluate pathways and networks in biological systems; and finally evaluating information variety (heterogeneity) of a group of parameters, which may serve as a potential surrogate measure of system adaptation and homeostasis. There are grounds to hope that increased use of information-theoretical methods for the solution of such and similar problems will yield more enhanced and adequate diagnostic capabilities as well as more reliable, quantitative evidence-based treatments.

Acknowledgment

We thank the creators and contributors of the UCI Machine Learning Repository for making data freely and openly available for research.

Funding Information

This study was funded by the authors independently.

Author's Contributions

David Blokh provided bioinformatic information-theoretical analysis. Ilia Stambler provided biological interpretation.

Ethics

No ethical issues may arise following this work.

Conflict of Interest

We have no conflict of interest.

References

Afifi, A.A. and S.P. Azen, 1979. Statistical Analysis: A Computer Oriented Approach. Academic Press, New York, ISBN-10: 0120444607, pp: 442.

Alter, O., P.O. Brown and D. Botstein, 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Nat. Acad. Sci. USA, 97: 10101-10106. DOI: 10.1073/pnas.97.18.10101

Baier, L.J. and R.L. Hanson, 2004. Genetic studies of the etiology of type 2 diabetes in Pima Indians. Hunting for pieces to a complicated puzzle. Diabetes, 53: 1181-1186. DOI: 10.2337/diabetes.53.5.1181

Blokh, D., I. Stambler, E. Afrimzon, Y. Shafran and E. Korech *et al.*, 2007. The information-theory analysis of Michaelis-Menten constants for detection of breast cancer. Cancer Detect. Prev., 31: 489-498. DOI: 10.1016/j.cdp.2007.10.010

Blokh, D., N. Zurgil, I. Stambler, E. Afrimzon and Y. Shafran *et al.*, 2008. An information-theoretical model for breast cancer detection. Methods Inf. Med., 47: 322-327. DOI: 10.3414/ME0440

Blokh, D., I. Stambler, E. Afrimzon, M. Platkov and Y. Shafran *et al.*, 2009. Comparative analysis of cell parameter groups for breast cancer detection. Comput. Methods Programs Biomed., 94: 239-249. DOI: 10.1016/j.cmpb.2009.01.005

Blokh, D., 2013. Information-theory analysis of cell characteristics in breast cancer patients. Int. J. Bioinform. Biosci., 3: 1-5.

Blokh, D. and I. Stambler, 2014. Estimation of heterogeneity in diagnostic parameters of age-related diseases. Ag. Dis., 5: 218-225. DOI: 10.14336/AD.2014.0500218

Blokh, D. and I. Stambler, 2015. Information theoretical analysis of aging as a risk factor for heart disease. Ag. Dis.

Blokh, D., 2012. Clustering financial time series via information-theory analysis and rank statistics. J. Patt. Reco. Res., 7: 106-115. DOI:10.13176/11.396

Conover, W.J., 1999. Practical Nonparametric Statistics. 3rd Edn., Wiley-Interscience, New York, ISBN-10: 8126507756, pp: 584.

Cover, T.M. and J.A. Thomas, 2006. Elements of Information Theory. 2nd Ed., Wiley-Interscience, New York, ISBN-10: 0471241954, pp: 776.

Foulley, J.L. and R.L. Quaas, 1995. Heterogeneous variances in Gaussian linear mixed models. Genet. Sel. Evol., 27: 211-228. DOI: 10.1051/gse:19950302

Gel'fand, I.M. and A.M. Yaglom, 1957. Computation of the amount of information about a stochastic function contained in another such function. Uspekhi Mat. Nauk, 12: 3-52.

Glantz, S.A., 2001. Primer of Biostatistics. 5th Edn., McGraw-Hill, New York, ISBN-10: 0071379460, pp: 489.

Glass, G.V. and J.C. Stanley, 1970. Statistical Methods in Education and Psychology. Prentice-Hall, New Jersey, ISBN-10: 0138449287, pp: 596.

Gutierrez Diez, P.J., I.H. Russo and J. Russo, 2012. The Evolution of the Use of Mathematics in Cancer Research. 1st Edn., Springer, New York, ISBN-10: 1461423961, pp: 404.

Li, N., J. Cruz, C.S. Chien, S. Sojoudi and B. Recht *et al.*, 2014. Robust efficiency and actuator saturation explain healthy heart rate control and variability. Proc. Nat. Acad. Sci. USA, 111: E3476-85. DOI: 10.1073/pnas.1401883111

Lim, S.S., T. Vos, A.D. Flaxman, G. Danaei and K. Shibuya *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. DOI: 10.1016/S0140-6736(12)61766-8

Lipsitz, L.A. and A.L. Goldberger, 1992. Loss of 'complexity' and aging: Potential applications of fractals and chaos theory to senescence. JAMA, 267: 1806-1809. DOI: 10.1001/jama.1992.03480130122036

Molina-Pena, R. and M.M. Alvarez, 2012. A simple mathematical model based on the cancer stem cell hypothesis suggests kinetic commonalities in solid tumor growth. PLoS One, 7: e26233-e26233. DOI: 10.1371/journal.pone.0026233

Nicolis, G. and I. Prigogine, 1990. Exploring Complexity. 1st Edn., W.H. Freeman, New York, ISBN-10: 0716718596, pp: 328.

Paninski, L., 2003. Estimation of entropy and mutual information. Neur. Comp., 15: 1191-1253. DOI: 10.1162/089976603321780272

Preckova, P., J. Zvarova and K. Zvara, 2012. Measuring diversity in medical reports based on categorized attributes and international classification systems. BMC Med. Inform. Dec. Mak., 12: 31-31. DOI: 10.1186/1472-6947-12-31

Quastler, H., 1958. The Domain of Information Theory in Biology. In: Symposium on Information Theory in Biology, Yockey, H.P. (Ed.), Pergamon Press, New York, ISBN-10: 1245144200, pp: 187-196.

Radtke, M.A., K. Midthjell, T.I., Nilsen and V. Grill, 2009. Heterogeneity of patients with latent autoimmune diabetes in adults: Linkage to autoimmunity is apparent only in those with perceived need for insulin treatment: Results from the Nord-Trøndelag Health (HUNT) study. Diabetes Care, 32: 245-250. DOI: 10.2337/dc08-1468

Renyi, A., 1959. On measures of dependence. Acta Math. Acad. Sci. Hungar., 10: 441-451. DOI: 10.1007/BF02024507

Reynolds, S.R., J. Albrecht, R.L. Shapiro, D.F. Roses and M.N. Harris *et al.*, 2003. Changes in the presence of multiple markers of circulating melanoma cells correlate with clinical outcome in patients with melanoma. Clin. Cancer Res., 9: 1497-1502. PMID: 12684425

UCI, 2014. UCI Machine Learning Repository. Pima Indians Diabetes Data Set. Original owner: US National Institute of Diabetes and Digestive and Kidney Diseases. Donor of Database: The Johns Hopkins University.

Zar, J.H., 2010. Biostatistical Analysis. 5th Edn., Prentice Hall, New Jersey, ISBN-10: 0321656865, pp: 960.

Zvarova, J. and M. Studeny, 1997. Information theoretical approach to constitution and reduction of medical data. Int. J. Med. Inform., 45: 65-74. DOI: 10.1016/S1386-5056(97)00036-1