Selection of Genetic and Phenotypic Features Associated with Inflammatory Status of Patients on Dialysis Using Relaxed Linear Separability Method

Leon Bobrowski[1,2]*, Tomasz Łukaszuk[2], Bengt Lindholm[3], Peter Stenvinkel[3], Olof Heimburger[3], Jonas Axelsson[3], Peter Bárány[3], Juan Jesus Carrero[3,4], Abdul Rashid Qureshi[3], Karin Luttropp[3,4], Malgorzata Debowska[1,3], Louise Nordfors[3,4], Martin Schalling[3,4], Jacek Waniewski[1]

1 Institute of Biocybernetics and Biomedical Engineering, Warsaw, Poland, **2** Bialystok University of Technology, Bialystok, Poland, **3** Baxter Novum and Renal Medicine, Karolinska Institutet, Stockholm, Sweden, **4** Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

Abstract

Identification of risk factors in patients with a particular disease can be analyzed in clinical data sets by using feature selection procedures of pattern recognition and data mining methods. The applicability of the relaxed linear separability (*RLS*) method of feature subset selection was checked for high-dimensional and mixed type (genetic and phenotypic) clinical data of patients with end-stage renal disease. The *RLS* method allowed for substantial reduction of the dimensionality through omitting redundant features while maintaining the linear separability of data sets of patients with high and low levels of an inflammatory biomarker. The synergy between genetic and phenotypic features in differentiation between these two subgroups was demonstrated.

Citation: Bobrowski L, Łukaszuk T, Lindholm B, Stenvinkel P, Heimburger O, et al. (2014) Selection of Genetic and Phenotypic Features Associated with Inflammatory Status of Patients on Dialysis Using Relaxed Linear Separability Method. PLoS ONE 9(1): e86630. doi:10.1371/journal.pone.0086630

Editor: Yi-Hsiang Hsu, Harvard Medical School, United States of America

Received June 6, 2012; **Accepted** December 16, 2013; **Published** January 28, 2014

Funding: This work was supported by Polish National Center for Research and Development [N R13 0014 04]; Institute of Biocybernetics and Biomedical Engineering PAS [4.2/St/2012]; Bialystok University of Technology [S/WI/2/2013]; GENECURE [grant LSHM-CT-2006-037697]; Swedish Research Council (Dnr 521-2012-2721). Baxter Novum is the result of a grant from Baxter Healthcare Corporation to Karolinska Institutet. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have the following interests. Baxter Novum - a university division within the Karolinska Institutet - is the result of a grant from Baxter Healthcare Corporation to Karolinska Institutet. Bengt Lindholm is employed by Baxter Healthcare Corporation and serves as well as head of Baxter Novum at the Karolinska Institutet. Jacek Waniewski, Juan Jesus Carrero, Abdul Rashid Qureshi and Malgorzata Debowska are or have been affiliated to Baxter Novum. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: l.bobrowski@pb.edu.pl

Introduction

Statistical models for analysis of risk factors for a disease or clinical complications, a main focus of medical research, require that the number of patients is larger than the number of variables (factors) to ensure that the statistical significance of the results can be appropriately established. In practice, most studies assess only the influence of each variable separately rather than the combined importance of a set of variables; the former oversimplistic but yet prevailing approach ignores the possibility of interactions between variables or between groups of variables [1]. The obvious need of developing new statistical tools that take into account the extensive interactions between the very large numbers of variables determining biological processes and hence clinical outcomes is increasingly emphasized in modern medical and bioinformatics research.

Medical data sets collected today often have a large number of variables for a relatively low number of patients. This may happen for genetic data sets, where the number of variables (genetic variability, as single nucleotide polymorphism, or gene expression data) can be thousand times greater than the number of patients. Statistical methods are not fully justified in this situation [1]. In such a case, data mining methods can be used instead of, or in addition, to statistical methods [2]. The methods of feature subset selection developed in the scope of data mining play an increasingly important role in the exploratory analysis of multidimensional data sets.

Feature selection methods are used to reduce feature space dimensionality by neglecting features (factors, measurements) that are irrelevant or redundant for the considered problem. Feature selection is a basic step in the complex processes of pattern recognition, data mining and decision making [3,4]. Interesting examples of applications of feature selection procedures can be found, among others, in bioinformatics [5]. A survey of noteworthy methods of feature selection in the field of pattern recognition is provided in [6].

The feature subset resulting from feature selection procedure should allow building a model on the basis of available learning data sets that can be applied for new problems. In the context of designing such prognostic models, the feature subset selection procedures are expected to produce high prediction accuracy.

We apply here the *relaxed linear separability* (*RLS*) method of feature selection for the analysis of data on clinical and genetic factors related to *inflammation*. These data were obtained from the so called *malnutrition, inflammation and atherosclerosis* (MIA) cohort of incident dialysis patients with end-stage renal disease [7] in whom extensive and detailed phenotyping and genotyping have been performed [8,9]. The cohort was split into two groups: inflamed patients (as defined by blood levels of C-reactive protein, *CRP*, above median) and non-inflamed patients (as defined by a *CRP* below median). Then, genetic and phenotypic (anthropometric, clinical, biochemical) risk factors that may be associated with the plasma *CRP* levels were identified by exploring the linear separability of the high and low *CRP* patient groups. Particular attention was paid in this work to study the complementary role of genetic and phenotypic feature subsets in differentiation between inflamed and non-inflamed patients.

Four benchmarking feature selection algorithms were selected for the comparisons with *RLS* method on the given clinical data set: 1) *ReliefF*, based on feature ranking procedure proposed by Kononenko [10] as an extension of the *Relief* algorithm [11], 2) *Correlation-based Feature Subset Selection - Sequential Forward* algorithm (*CFS-SF*) [12], 3) *Multiple Support Vector Machine Recursive Feature Elimination* (*mSVM-RFE*) [13] and 4) *Minimum Redundancy Maximum Relevance* (*MRMR*) algorithm [14]. The *CPL* method and four other frequently used classification methods (*RF* (*Random Forests*) [15], *KNN* (*K - Nearest Neighbors*, with K = 5) [3], *SVM* (*Support Vector Machines*) [16], *NBC* (*Naive Bayes Classifier*) [3]) were applied for classification of patients on the basis of the selected features.

Methods

Relaxed Linear Separability Method

A detailed description of the *relaxed linear separability* (*RLS*) method as applied in the present study is provided in Appendix S1 together with all the definitions. A brief summary of the method is presented below.

The *RLS* method of feature subset selection is linked to the basic concept of linear separability. The linear separability means possibility of two learning sets separationby a hyperplane [17,18]. The linear separability notion originated from the perceptron model linked to the beginning of neural networks [19]. Detection and evaluation of linear separability can be carried out efficiently by minimizing the *perceptron criterion function* [3]. This function belongs to the more general class of the *convex and piecewise-linear (CPL)* criterion functions [20].

The *perceptron* criterion function was modified by adding a regularization component for the purpose of the feature subset selection task [20]. The regularization component has similar structure to those used in the *Lasso regression* [21]. The main difference between the *Lasso* and the *RLS* methods is in the types of the basic criterion functions. The basic criterion function used in the *Lasso* method is that of the *least squared* method, whereas the perceptron criterion function and the modified criterion function are used in the *RLS* method. This difference effects the computational techniques used to minimize the criterion functions. The modified criterion function, similarly to the perceptron criterion function, is convex and piecewise-linear (*CPL*). The basis exchange algorithms allow the identification of the minimum of each of these *CPL* criterion functions [22]. The basis exchange algorithms are similar to linear programming and allow to find the optimal solution efficiently even in the case of large, high dimensional learning sets.

The (*RLS*) method of feature subset selection is based on minimization of the modified perceptron criterion function and allows for successive reduction of unnecessary features while preserving the linear separability of the learning sets by increasing the cost parameter in the modified criterion function. The stop criterion for discarding the unnecessary features was based on the cross-validation error (CVE) rate (defined as the average fraction of wrongly classified elements) estimated by the leave-one-out method.

The evaluation of the *RLS* approach was previously carried out with good results both when applied on simulated high dimensional and numerous data sets as well as on benchmarking genetic data sets [18]. For example, the *RLS* method were used for processing the *Breast cancer* data set [23]. The number of features (genes) in this set is equal to 24481. The *RLS* method allowed to select from this set the optimal subset of 12 genes and such linear combination of these genes (*linear key*), which allows to correctly distinguish with 100% accuracy two leaning sets composed of 46 cancer and 51 non-cancer patients.

Alternative Methods for Feature Selection and Classification

The *RLS* method of feature subset selection involves generation of the sequence of the reduced feature subspaces F_k (see Appendix S1, equation 7). The sequence is generated in the deterministic manner through a gradual increase of the cost level λ in the minimized criterion function $\Psi_\lambda(\mathbf{w},\theta)$ (see Appendix S1, equation 5). In order to determine the best (final) subspace $F_{k'}$ in the sequence an evaluation of the quality of individual subspaces F_k is needed. Traditionally, the quality of the feature subspaces F_k is evaluated through the quality evaluation of the classifiers built in this subspace. Statistical methods for evaluation and comparison of classifiers can be found in [24]. This section presents a few other previous methods of feature selection and classification that were applied for the analysis of the MIA data sets, for comparison of the results, see Results.

Four benchmarking feature selection algorithms were chosen for an experimental comparison with the *RLS* method. One of the selected algorithms, *ReliefF*, is based on feature ranking procedure proposed by Kononenko [10] as an extension of the *Relief* algorithm [11]. The *ReliefF* searches for the nearest objects from different classes and weighs features according to how well they differentiate these objects. The second one is a subset search algorithm denoted as *CFS-SF* (*Correlation-based Feature Subset Selection - Sequential Forward*) [12]. The *CFS-SF* algorithm is based on a correlation measure which evaluates the goodness of a given feature subset by assessing the predictive ability of each feature in the subset and a low degree of correlation between features in the subset. These two feature selection algorithms are considered as "the state of the art" tools for feature selection [4]. The third algorithm, *mSVM-RFE*, is a relatively new idea. It is an extension of the *SVM-RFE* algorithm (*Support Vector Machine Recursive Feature Elimination*). The *SVM-RFE* is an iterative procedure that works backward from an initial set of features. At each round it fits a simple linear *SVM*, ranks the features based on their weights in the *SVM* solution, and eliminates the feature with the lowest weight [25]. Multiple *SVM-RFE* (*mSVM-RFE*) extends this idea by using resampling techniques at each iteration to stabilize the feature rankings [13]. The fourth algorithm *MRMR* (*Minimum Redundancy - Maximum Relevance*) [14] is also a relatively new idea. It bases on feature ranking procedure with special ranking criterion. The position of single feature in the list depends both on its correlation with class and dissimilarity to each feature above it in the ranking.

To compare feature selection algorithms and to evaluate the selected feature subspaces, four frequently used classification methods, beside the *CPL* method, were applied:

$$\begin{array}{l}1.\ \text{RF (Random Forests) [15]}\\ 2.\ \text{KNN (K}-\text{Nearest Neighbors, with K}=5\text{) [3]}\\ 3.\ \text{SVM (Support Vector Machines) [16]}\\ 4.\ \text{NBC (Naive Bayes Classifier) [3]}\\ 5.\ \text{CPL (Convexand Piecewise}-\text{Linear criterion functions) [20]}\end{array} \qquad (1)$$

The four first classifiers (1) were designed by using *Weka*'s implementation [26]. The *Weka*'s implementation of *ReliefF* and *CFS-SF* was used also for the feature selection and cross validation evaluation of designed classifiers. The *R* implementation of *mSVM-RFE* was used (*SVM-RFE* package) [27]. The results of *MRMR* was obtained with the help of the code provided by its author [28]. The *CPL* classifiers based on the search for optimal separating hyperplane $H(\mathbf{w}^*,\theta^*)$ (see Appendix S1, equation 1) through minimization of the *CPL* criterion functions $\Phi(\mathbf{w}[n],\theta)$ (see Appendix S1, equation 4) was applied using our own implementation. Our own implementation was also used for the *RLS* method of feature selection [18].

Clinical Data Sets

Two learning sets G^+ and G^- were selected from a cohort of patients with chronic kidney disease, the *MIA* cohort [7]. The set G^+ contained $m^+=112$ patients O_j with a high *CRP* levels (above the median value) and the set G^- contained $m^-=113$ patients O_j with a low plasma *CRP* levels (below the median value). Each patient O_j from the learning sets G^+ and G^- was characterized by numerical results x_i ($x_i \in R$) of 57 anthropometric or biochemical measurements and by 79 sites of genetic polymorphism (single nucleotide polymorphisms (*SNPs*) or deletions/insertions). The 79 polymorphisms were selected from 45 different candidate genes each harboring one to four of these variations. Each site of the genetic polymorphism was characterized by (usually three) binary features x_i ($x_i \in \{0,1\}$), $i=1,2,3$, that described three possible genotypes at this site (for example A/A, C/C, A/C). The value one ($x_i=1$) of the binary feature x_i represented the appearance of a particular genotype at the polymorphic site. Thus, each patient O_j was represented by the n-dimensional feature vector $\mathbf{x}_j=[x_{j1},...,x_{jn}]^T$, where $n=228$ is the total number of features and $j \in \{1,...,225\}$ represents the order number (*index*) of a patient O_j in the cohort of 225 patients. The number of genetic features, $n=228$, is lower than the expected value of $237=3\times 79$ because several genes appeared in the studied population as only one or two genotype forms, i.e., the polymorphism in these genes was not found or was reduced - such cases were coded with less than three binary features. There was also one gene with three alleles and it was coded with five binary features.

These cohort and feature sets were selected from a larger data set and included only those patients for whom at least 85% of features were available and those features that were measured for at least 65% of the patients. In the selected cohort there were still missing data; therefore, for each missing datum, its value for the nearest neighbor in the respective learning set (G^+ or G^-) was assigned. The phenotypic and genetic features were considered separately in the procedure of allocating the missing data. In the case of a missing phenotypic feature value, the nearest neighbour was the patient that had the most similar phenotype, whereas for a missing genetic feature value, the nearest neighbour was the patient that had the most similar genotype. The *ce.impute* procedure of *dprep* package of the *R* programming language was used for the substitution of missing values.

During exploration of this database, the computations were performed in feature subspaces F_k ($F_k \subset F$) divided in two learning sets G_k^+ and G_k^-. The vectors $\mathbf{x}_j$ from the set G_k^+ described patients O_j with high plasma *CRP* levels in the feature subspaces F_k. Similarly, the vectors $\mathbf{x}_j$ from the set G_k^- described the patients with low plasma *CRP* levels.

Three basic feature spaces F_k were distinguished as follows:

$$\begin{array}{ll} I.\ \ F_I & -\textit{phenotypic space}\\ & \quad (n_I=57 \textit{ standardized features } x_i\ (x_i \in R^1))\\ II.\ \ F_{II} & -\textit{genetic space}\\ & \quad (n_{II}=228 \textit{ binary features } x_i\ (x_i \in \{0,1\}))\\ III.\ F_{III} & -\textit{phenotypic and genetic space}\\ & \quad (n_{III}=285 \textit{ standardized or binary features } x_i)\end{array} \qquad (2)$$

The *RLS* procedure of feature selection was carried out in each of the basic feature spaces (2) separately.

Results

The apparent error rate $AE=e_a(\mathbf{w}_k^*,\theta_k^*)$ (see Appendix S1, equation 9) and the crossvalidation error rate $CVE=e_{CVE}$ (see Appendix S1, equation 10) of the optimal linear classifier $LC_k(\mathbf{w}_k^*,\theta_k^*)$ (see Appendix S1, equation 8) as a function of the dimension k of feature subspaces F_k in the sequence (see Appendix S1, equation 7) of the feature spaces F_I, F_{II} and F_{III}, definition (2), are presented in Figures 1–3.

The apparent error rate (*AE*) and the cross-validation error (*CVE*) in feature subspaces F_k of the *phenotypic* space F_I are shown in Figure 1. The lowest value of (*CVE*) equal to 13,8% appeared in the feature subspace $F_{k'}$ of the dimension $k=21$. The features that define this subspace $F_{k'}$ are presented in Table 1. The features listed in Table 1 were ordered according to the absolute values $|w_i^*|$ (*factors*) of the components of the optimal weight vector $\mathbf{w}_k^*=[w_{k1}^*,...,w_{kn}^*]^T$.

The features listed in Table 1 was identified as the one subset F_k of the feature subspace F_I. This subset was not composed from the best single features x_i. It includes the features that are correlated to *CRP* plasma levels as well as those that are not. Most of the *phenotypic features* listed in Table 1 are in fact expected by medical experts to be related to inflammation but their relative importance is less clear.

Whereas the list of phenotypic features in general appears to be biologically plausible, the ranking of the strength of the association as expressed by the value of the factor coefficient w_{ki}^* provides novel and potentially important insights into the links between the investigated features and the biomarker selected to represent inflammation, i.e. *CRP*. Thus, some of the identified phenotypic features in Table 1 (i.e., serum fibrinogen, (low) plasma iron, serum ferritin, serum interleukin-6, and white blood cells count) are well established *biomarkers of inflammation*, whereas others are linked to *cardiovascular disease* (plasma troponin T and systolic blood pressure) which is in turn linked to inflammation [29]. However, the negative value for the factor coefficient for systolic blood pressure is an intriguing finding which might reflect that a low blood pressure could be associated with cardiac dysfunction and heart failure, conditions which are known to be associated with inflammation [30]. Other phenotypic features in Table 1 (height, serum creatinine, plasma insulin, plasma calcium, bone mineral density, hand grip strength, S-triiodothyronine T3, plasma uric

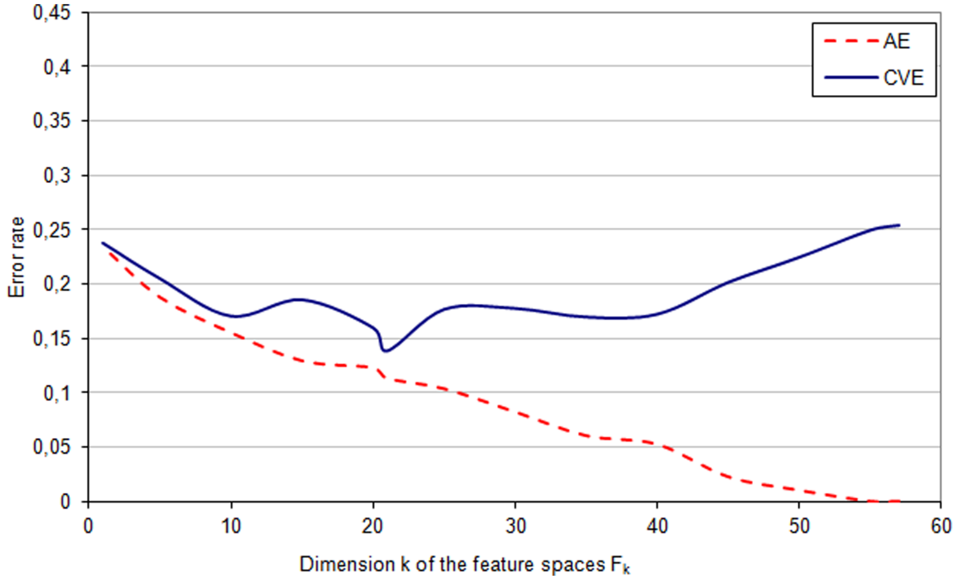

Figure 1. AE and CVE - *phenotypic* space. The apparent error rate (*AE*) and the cross-validation error (*CVE*) in different feature subspaces F_k of the *phenotypic* space F_I.
doi:10.1371/journal.pone.0086630.g001

acid, plasma fetuin, truncal fat mass, body mass index, glycated hemoglobin) are linked to *nutrition* (height, serum creatinine, bone mineral density, hand grip strength, truncal fat mass and body mass index). It is well established that an abnormal nutritional status with protein-energy wasting in this patient population is strongly linked to inflammation [31]. Several features were linked to *hormonal status* or *metabolism* (plasma insulin, plasma calcium, S-triiodothyronine T3, plasma uric acid, plasma fetuin, glycated hemoglobin); in general, relations between these features and inflammation have been described previously, but the relation with plasma calcium is not expected. Finally, high age and smoking are factors which are associated with inflammation.

Feature selection from the *genetic* space F_{II} is illustrated in Figure 2. The learning sets G^+ and G^- of the space F_{II} are linearly separable, i.e., the apparent error AE is equal to zero. Moreover, the linear separability was preserved during feature reduction from $k=228$ to $k=55$. In contrast, the lowest value of the average cross-validation error rate $CVE \approx 16{,}9\%$ appeared for $k=81$. It should be stressed, that the cross-validation procedure does not separate fully those feature subspaces that are linearly separable (Figure 2).

The process of feature selection from the combined *phenotypic* and *genetic* space F_{III} yielded interesting results shown in Figure 3. The linear separability in the combined space F_{III} was found in a

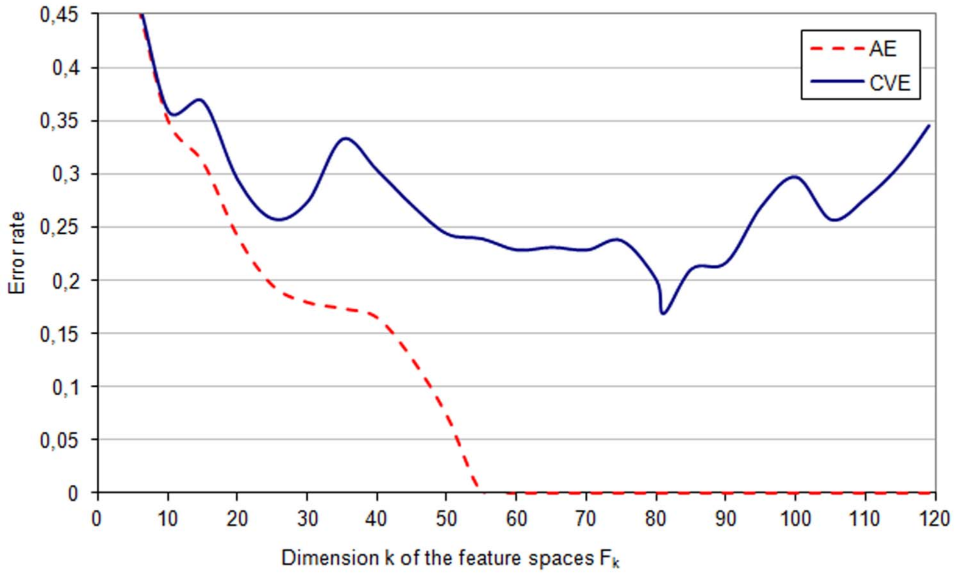

Figure 2. AE and CVE - *genetic* space. The apparent error rate (*AE*) and the cross-validation error (*CVE*) in different feature subspaces F_k of the *genetic* space F_{II}.
doi:10.1371/journal.pone.0086630.g002

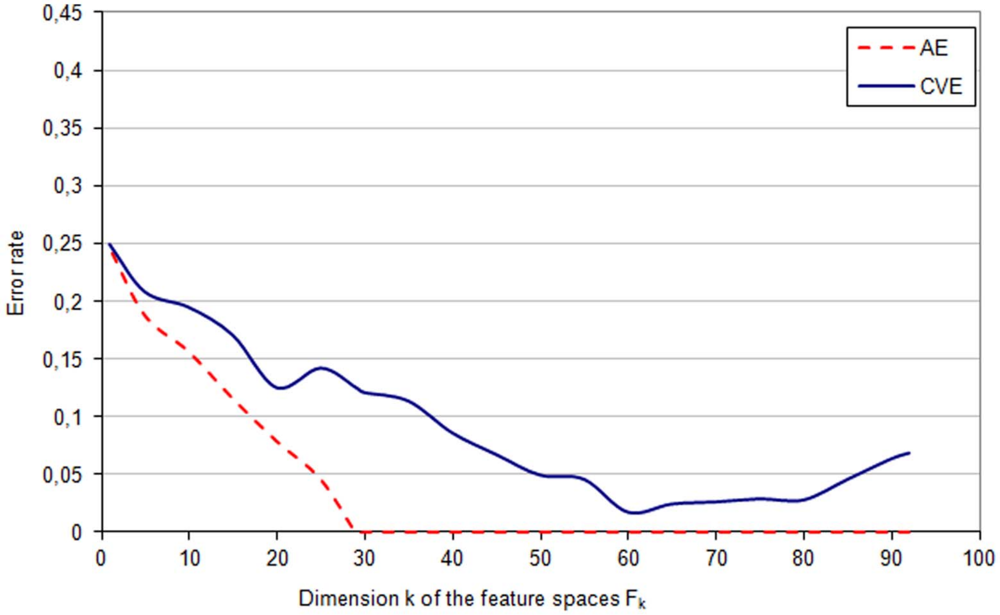

Figure 3. AE and CVE - *phenotypic* and *genetic* space. The apparent error rate (*AE*) and the cross-validation error (*CVE*) in different feature subspaces F_k of the *phenotypic* and *genetic* space F_{III}.
doi:10.1371/journal.pone.0086630.g003

Table 1. Features that define the optimal phenotypic subspace F_k characterized by the lowest cross-validation error (*CVE*), their factor coefficients w^*_{ki} in the minimal value of the criterion function $\Psi_\lambda(\mathbf{w},\theta)$ (see Appendix S1, equation 5) and their correlation coefficients with *CRP* plasma concentrations.

Feature	Factor	Pearson's correlation	*p*-value
Serum fibrinogen	1,478	0,483	0,000
Plasma iron	1,066	−0,389	0,000
Serum ferritin	1,023	0,238	0,000
Height	0,841	0,098	0,141
Serum interleukin-6	0,806	0,396	0,000
Serum creatinine	−0,778	−0,070	0,298
White blood cells count	0,758	0,351	0,000
Smoking	0,754	0,106	0,114
Plasma insulin	−0,740	0,017	0,796
Plasma calcium	−0,657	−0,085	0,201
Bone mineral density	−0,493	−0,084	0,212
Plasma Troponin T	0,493	0,225	0,001
Systolic blood pressure	−0,433	−0,039	0,559
Handgrip strength	0,404	−0,064	0,336
S-triiodothyronine T3	−0,393	−0,219	0,001
Plasma uric acid	0,301	0,093	0,165
Age	0,289	0,323	0,000
Plasma fetuin	−0,278	−0,120	0,071
Truncal fat mass	0,237	0,225	0,001
Body mass index	0,237	0,075	0,264
Glycated hemoglobin	−0,153	−0,088	0,189

doi:10.1371/journal.pone.0086630.t001

large range of subspace dimensions from $k=285$ till $k=29$. The minimal feature subspace F_k with the linear separability of the learning sets for $k=29$ is composed from both *phenotypic* (i.e., clinical, anthropometric and laboratory) features and *genotypes*. The minimal value of the average cross-validation error rate was low: $CVE=1{,}8\%$. This minimum value appeared at the dimension $k=60$ inside the linear separability zone. The optimal feature subspace F_k with $k=60$ was composed from 29 phenotypic features and 31 genotypes.

The minimal cross validation error rate in the *phenotypic* space F_I was $CVE=25{,}8\%$ (Figure 1), and the *genetic* space F_{II} it was $CVE=22{,}7\%$ (Figure 2). Combining the *phenotypic* and *genetic* factors (features) resulted in a marked reduction of the CVE error rate to 1,8%. These results indicate that the *phenotypic* and *genetic* factors are not independent and play complementary roles in describing the inflammatory status of the patients in the MIA cohort.

The *confusion matrices* $T_k(\mathbf{w}^*_k,\theta^*_k)$ with the mean values obtained by the *leave-one-out* procedure for the *phenotypic* and *genetic* features are presented in Table 2 for a few selected feature subspaces. The lowest error was found for the subspace with dimension $k=60$ in agreement with the *RLS* method of feature selection.

The optimal parameters $\mathbf{w}^*_k$ and θ^*_k may be used to define the linear (affine) transformation of the feature vectors $\mathbf{x}$ ($\mathbf{x}\in F^*_k$) on the one dimensional space R^1:

$$y=(\mathbf{w}^*_k)^T\mathbf{x}-\theta^*_k \qquad (3)$$

The above transformation described by equation (3) was applied in designing the *scatter diagram* (*diagnostic map*) showed in Figure 4. The horizontal axis (called *phenotypic fraction*) was obtained by transformation (3) applied for 29 *phenotypic* features that constitute the optimal feature subspaces F^*_{60} of the *phenotypic* and *genetic* space F_{III}. Similarly, the vertical axis (called *genetic fraction*) of the diagram

Table 2. The *confusion matrices* $T_k(\mathbf{w}_k^*,\theta_k^*)$ (see Appendix S1, equation 11), for the combined *phenotypic* and *genetic* subspaces F_{III} with dimensionalities $k=285$, 92, 60, and 25.

$k=285$	G_k^+	G_k^-
ω^+	89	23
ω^-	24	89
$k=92$	G_k^+	G_k^-
ω^+	104	8
ω^-	8	105
$k=60$	G_k^+	G_k^-
ω^+	110	2
ω^-	2	111
$k=25$	G_k^+	G_k^-
ω^+	95	17
ω^-	20	93

doi:10.1371/journal.pone.0086630.t002

was obtained by transformation (3) applied for 31 *genetic* features x_i which constitute the optimal feature subspaces F_{60}^*.

The *diagnostic map* showed in Figure 4 can be used for diagnosis support. A new patient represented by a feature vector $\mathbf{x}$ ($\mathbf{x}\in F_k^*$) can be situated on the *diagnostic map* as the point y determined by equation (3). If most of the K nearest neighbors y_j of the point y (3) on the map belong to the set G^+ of the high *CRP* patients, then we infer that the new patient is inflamed. If most of the K nearest neighbors y_j of the point y (3) on the map belong to the set G^- of the low *CRP* patients, then we infer that the new patient is not inflamed. Similar schemes of decision support are called the *K-nearest neighbours* (*KNN*) in the *pattern recognition* or as the *Case Based Reasoning* (*CBR*) scheme [18].

The transformation of the multidimensional feature vectors $\mathbf{x}_j$ ($j=1,\ldots,m$) from the learning sets G_k^+ and G_k^- and the feature vector $\mathbf{x}$ of a currently diagnosed patient on a two-dimensional diagnostic map are aimed at obtaining a *similarity measure* $s(\mathbf{x},\mathbf{x}_j)$ [20]. The measure $s(\mathbf{x},\mathbf{x}_j)$ allows for the determination of the similarity between the vector $\mathbf{x}$, representing a newly diagnosed patient, and the *m precedents* (*cases*, *verified examples*) from the learning sets (clinical database). Such scheme of the decision support based on the *diagnostic maps* has been used successfully in the medical diagnosis support system *Hepar* [32].

The performance of *RLS* selection method and *CPL* classifier applied in our study was compared to other selection methods and classifiers (see Section "Alternative methods for feature selection and classification") using the error rate (fraction of misclassified objects from the test set), *CVE*, evaluated in the *cross-validation* (*leave-one-out*) procedure [3]. The results are presented in Tables 3–5. The methods *CFS-FS* and *mSVM-RFE* alongside with *RLS* select an optimal subset of features and their prediction power can be assessed using different classifiers. In contrast, *ReliefF* and *MRMR* methods are ranking procedures and od not provide any intrinsic criteria for selection of any optimal subset of features. Such criterion need to be chosen separately. For the purpose of comparison of all these methods, the optimal sets of features for *ReliefF* and *MRMR* were determined for each classifier separately as those with minimal *CVE* for the applied classifier. Thus, the optimal set (and number) of features for these two methods can vary with the choice of classifier (see Tables 3–5).

All the applied methods of feature selection were able to reduce the initial number of features (Tables 3–5). The highest reduction was obtained by *CFS-FS* method, which substantially outperformed in this respect four other methods. The features selected by *RLS* method provided however the lowest average cross validation error *CVE* for all three feature spaces. Especially low errors of $1-2\%$ (with standard deviation of 10%) obtained for *RLS* method in the combined phenotypic and genotypic feature space (Table 5) demonstrate its good efficiency. The number of features was reduced in this case five times. For the space of genetic features, only *RLS* selection method combined with *CPL* classifier was able to obtain the low average error around 10%, much lower that values of around 30% or higher obtained by other selection methods and classifiers (Table 4). In the case of phenotypic features, the five selection methods had a similar performance, but *RLS* method yielded slightly lower errors than the four other methods (Table 3). *MRMR* provided in all three feature spaces lower error values than other methods alternative to *RLS*, especially for *SVM* and *CPL* classifiers; however, the optimal sets of features defined according to the minimal *CVE* value for *MRMR* depended on the selected classifier and this reasult would need further attention and investigation of the scope of these different optimal sets. It is also worth to notice that by allowing for higher errors (similar to those obtained for *CFS-FS* method), one can easily reduce further the number of features selected by *RLS* method as it can be seen in Figures 1–3. Among classifiers, *SVM* and/or *CPL* yielded the lowest errors when combined with *RLS* or *CFS-FS* selection methods. *ReliefF* method worked also well with *RF* and *KNN* classifiers. The errors related to the application of *mSVM-RFE* were similar to those related to *ReliefF* and *CFS-FS* methods (Tables 3 and 4).

The overlap between the features selected by different methods was not high. For example, among the 15 features selected by *CFS-FS* method from the combined phenotypic and genetic features (Table 5), three were shared among all three methods and seven with only one of the two other methods; five features were specific for the *CFS-FS* method. However, the problem of overlapping between features cannot be easily interpreted because many features are more or less correlated and different methods may select different features fromthose that are mutually correlated. Therefore, an additional analysis would be necessary to investigated this problem; however, this is outside the scope of this study.

Among the four applied feature selection methods, *CFS-FS* was the fastest (computation time of the order of 1 sec). *ReliefF* and *MRMR* (together with the selection of optimal set) needed between a few and a few tens of minutes (depending on the applied classifier). The computation time of the *RLS* method was of the order of tens of minutes. The *mSVM-RFE* method had the computation time of about 20 hours. It should be stressed that the relatively long computation time of the *RLS*, *mSVM-RFE*, *ReliefF* and *MRMR* methods was caused mainly by repeated computation in the framework of the cross-validation procedure used by these methods.

Discussion and Conclusions

Feature selection is an integral - but often implicit - component in statistical analyses. An explicit systematic feature selection process is of value for identifying features that are important for prediction, and for analysis on how these features are related, and furthermore it provides a framework for selecting a subset of relevant features for use in model construction. The most common approach for feature selection in clinical and epidemiological research is based so far on evaluation of the impact of single

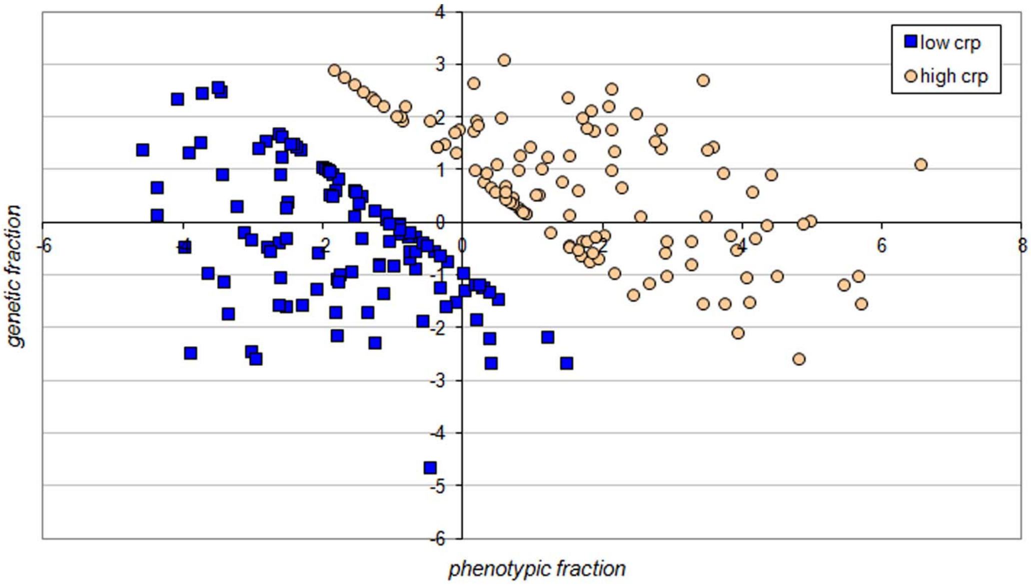

Figure 4. The diagnostic map. Linear separation of the high *CRP* from the low *CRP* patients for the cohort of incident dialysis patients in the optimal feature subspace F_{60}^{*} of the phenotypic and genetic space F_{III}.
doi:10.1371/journal.pone.0086630.g004

features [4]. In this approach, the resulting feature subsets are composed of such features (factors) which have the strongest individual influence on the analyzed outcome (in this case inflammation). Such approach is related to the assumption about the independence of the factors. However, in a complex system, such as the living organism, these factors are more often related than not related. The role of particular factors in a living organism depends among others on (time-dependent) environmental factors and internal conditions, and on (permanent) genetic factors. An advantage of the relaxed linear separability (*RLS*) method is that it may identify directly and efficiently a subset of related features that influences the outcome and that it assesses the *combined* effect of these features as prognostic factors. This characteristic of the approach presented here is clearly visible in the dataset of phenotypic features with minimal cross validation error rate, Table 1: this set contains also features that individually do not

Table 3. The cross validation error *CVE* (mean ± SD) for different classifiers in the *phenotypic* space F_I and their subspaces obtained by using five features selection methods (*RLS*, *ReliefF*, *CFS-FS*, *mSVM-RFE*, *MRMR*) and five classifiers (*RF*, *KNN*, *SVM*, *NBC*, *CPL*), see Section ''Alternative methods for feature selection and classification''.

Feature selection method	Number of features	Classifier				
		RF	KNN	SVM	NBC	CPL
No selection	57	0,231	0,329	0,258	0,302	0,258
		±0,422	±0,470	±0,437	±0,459	±0,437
ReliefF	*	0,173	0,240	0,160	0,240	0,156
		±0,379	±0,428	±0,367	±0,428	±0,362
		(25)	(28)	(26)	(3)	(26)
CFS-FS	15	0,218	0,196	0,178	0,267	0,191
		±0,413	±0,397	±0,382	±0,442	±0,393
mSVM-RFE	26	0,200	0,338	0,151	0,231	0,178
		±0,400	±0,473	±0,358	±0,422	±0,382
MRMR	*	0,182	0,182	0,169	0,240	0,173
		±0,387	±0,387	±0,375	±0,427	±0,379
		(30)	(12)	(11)	(21)	(8)
RLS	21	0,191	0,311	0,156	0,280	0,138
		±0,393	±0,463	±0,362	±0,449	±0,345

*ReliefF and MRMR are ranking procedures. The optimal sets of features for these two methods were determined for each classifier separately; the number of features (shown in parentheses) corresponds to the size of the subset of features characterized by the smallest cross validation error for the specific classifier.
doi:10.1371/journal.pone.0086630.t003

Table 4. The cross validation error *CVE* (mean ± SD) for different classifiers in the *genetic* space F_{II} and their subspaces obtained by using five features selection methods (*RLS*, *ReliefF*, *CFS-FS*, *mSVM-RFE*, *MRMR*) and five classifiers (*RF*, *KNN*, *SVM*, *NBC*, *CPL*), see Section ''Alternative methods for feature selection and classification''.

Feature selection method	Number of features	Classifier				
		RF	**KNN**	**SVM**	**NBC**	**CPL**
No selection	228	0,502	0,436	0,444	0,493	0,462
		±0,500	±0,496	±0,497	±0,500	±0,499
ReliefF	*	0,338	0,293	0,347	0,369	0,369
		±0,473	±0,455	±0,476	±0,483	±0,483
		(22)	(76)	(82)	(26)	(39)
CFS-FS	3	0,458	0,427	0,427	0,422	0,427
		±0,498	±0,495	±0,495	±0,494	±0,495
mSVM-RFE	140	0,48	0,342	0,356	0,458	0,378
		±0,500	±0,474	±0,478	±0,498	±0,485
MRMR	*	0,347	0,333	0,280	0,280	0,276
		±0,476	±0,471	±0,449	±0,449	±0,447
		(21)	(70)	(38)	(21)	(25)
RLS	81	0,489	0,418	0,338	0,418	0,169
		±0,500	±0,483	±0,473	±0,493	±0,375

*ReliefF and MRMR are ranking procedures. The optimal sets of features for these two methods were determined for each classifier separately; the number of features (shown in parentheses) corresponds to the size of the subset of features characterized by the smallest cross validation error for the specific classifier.

doi:10.1371/journal.pone.0086630.t004

correlate to the level of *CRP* in plasma, the clinical biomarker used here for discrimination of inflamed and non-inflamed patients.

The *RLS* method of feature selection is based on the minimization of the criterion function $\Psi_\lambda(\mathbf{w},\theta)$ (see Appendix S1, equation 5) for selected values of the cost level λ and repeated minimizations of the perceptron criterion function $\Phi(\mathbf{w},\theta)$ (see Appendix S1, equation 4) in consecutive reduced feature subspaces F_k (see Appendix S1, equation 7). The *CPL* criterion function $\Psi_\lambda(\mathbf{w},\theta)$ can be defined for different values of the cost level λ ($\lambda > 0$) in the same feature space F. Successive increasing of the

Table 5. The cross validation error *CVE* (mean ± SD) for different classifiers in the *phenotypic* and *genetic*space F_{III} and their subspaces obtained by using five features selection methods (*RLS*, *ReliefF*, *CFS-FS*, *mSVM-RFE*, *MRMR*) and five classifiers (*RF*, *KNN*, *SVM*, *NBC*, *CPL*), see Section ''Alternative methods for feature selection and classification''.

Feature selection method	Number of features	Classifier				
		RF	**KNN**	**SVM**	**NBC**	**CPL**
No selection	285	0,293	0,382	0,218	0,293	0,209
		±0,455	±0,486	±0,413	±0,455	±0,407
ReliefF	*	0,191	0,240	0,187	0,200	0,213
		±0,393	±0,427	±0,390	±0,400	±0,410
		(80)	(2)	(54)	(16)	(61)
CFS-FS	15	0,218	0,196	0,178	0,267	0,191
		±0,413	±0,397	±0,382	±0,442	±0,393
mSVM-RFE	153	0,262	0,382	0,156	0,302	0,182
		±0,440	±0,486	±0,362	±0,459	±0,386
MRMR	*	0,160	0,267	0,129	0,213	0,156
		±0,367	±0,442	±0,335	±0,410	±0,362
		(25)	(1)	(44)	(27)	(39)
RLS	60	0,231	0,378	0,018	0,258	0,018
		±0,422	±0,485	±0,132	±0,437	±0,132

*ReliefF and MRMR are ranking procedures. The optimal sets of features for these two methods were determined for each classifier separately; the number of features (shown in parentheses) corresponds to the size of the subset of features characterized by the smallest cross validation error for the specific classifier.

doi:10.1371/journal.pone.0086630.t005

parameter λ in the function $\Psi_\lambda(\mathbf{w},\theta)$ allows to reduce increasing number of features and, as the result, the obtain the descended sequence of feature subspaces F_k. A feasibility of feature subspaces F_k can be evaluated on the basis of the cross validation experiment with the optimal linear classifier $LC(\mathbf{w}^*,\theta^*)$ (see Appendix S1, equation 8). The parameters $\mathbf{w}^*$ and θ^* of the optimal classifier are defined on the basis of repeated minimizations of the perceptron criterion function $\Phi_k(\mathbf{w},\theta)$ on elements $\mathbf{x}_j$ of the learning sets G_k^+ and G_k^- in subspace F_k.

The application of this method for identifying genetic and phenotypic (anthropometric, clinical and biochemical) risk factors that are associated with inflammation was implemented using a clinical database of patients with chronic kidney disease. A few important properties of the computation results obtained from this cohort can be pointed out. The results show, among others, the scale of the bias of the apparent error (*AE*) estimator (see Appendix S1, equation 9). The bias is illustrated as the difference between the *CVE* curve and the *AE* curve (Figures 1–3). The optimal feature subspace F_k^* characterized by the lowest *CVE* error rate e_{CVE} (see Appendix S1, equation 10) cannot be identified on the basis of the apparent error *AE* curve because of this bias. The minimum of the *CVE* rate is clear and narrow for the analysis of genetic data (Figure 2), whereas it is less marked for phenotypic and phenotypic-genetic data sets (Figures 1 and 3) with *CVE* curves fluctuating for a wide range of feature numbers. These two cases may need an analysis of not only the feature space with minimal *CVE* but also the feature spaces with similar, albeit slightly higher *CVE* values. It is also interesting to observe that the lowest values of *CVE* occur for feature subspaces with zero apparent error rate, if genetic and phenotypic-genetic feature spaces are analyzed (Figures 2 and 3), whereas for phenotype feature space the minimum is within the range of subspaces with non-zero apparent error rate (Figure 1).

Working with large medical data bases one meets often the problem of missing data, which was encountered also in our database. The patients with too many features missing and features that are measured for too low number of patients must be excluded. However, with sufficiently many data one can restore missing values by hypothetical values, and in our study this was done by the value of the nearest neighbour, separately for the phenotypic and genotypic features. Another practical problem is the overfitting of the data that happens when many features are studied for a relatively low number of patients, and this problem occurs also in our database: the two sets of patients with different inflammatory status can be linearly separated as indicated by zero apparent error for all features in the case of genetic, phenotypic and combined sets of features (Figures 1–3). Therefore, to provide a more reliable method for identifying the most predictive subset of features, the cross validation error was applied together with the *leave-one-out* procedure. These two problems preclude actually any statistical *proof* of the studied associations between features in our patient populations and the study should be considered rather as an example of *exploratory analysis* for associations that should be further investigated. We hope that our approach can supplement the current methods for analyses of such complex data which are difficult to collect, and, at the same time, represent unique and medically promising sets of data.

An important characteristic of feature selection methods is the predictive power of the selected feature set, as assessed in the present study by cross validation error (Tables 3–5). The *RLS* method combined with *CPL* classifier was of similar effectiveness as some other methods if applied for phenotypic features represented mostly by continuous variables (Table 3), considerably better than all other methods if applied for genetic features represented by discrete (zero - one) variables, Table 4, and much better than all other methods if applied for combined phenotypic and genetic features represented by mixed type mathematical variables (Table 5). Therefore, the *RLS/CPL* approach may be considered as a viable and promising tool for analysis of the extent by which the genetic pool, and, especially the combination of genetic variability and phenotypic characteristics of the patient, may associate with selected features in patient populations.

The computational time for our method depends on two factors: 1) the number of cases and features, and 2) the repetition of calculations for the cross-validation method. The actual computing time for personal computer implementations was in the order of tens of minutes, and was longer than for some alternative methods (see Results), but all the computational times were reasonably short for the current research purpose. However, the computation time may be a limitation of the *RLS* method if applied in the future for data bases with huge amount of data and many patients, or both, and the parallelization of the code or the application of main frame computers may be necessary. Our results suggest that the considerably lower prediction errors obtained for our approach compared to those yielded by faster methods, especially for combined genetic and phenotypic data, make such extensions of the code worthwhile.

The comparison between the optimal feature subspaces F_k^* of the three feature spaces (*phenotypic*, *genetic*, *combined*) showed that the combined *phenotypic* and *genetic* subspace can provide a very low *CVE* error rate of 2% (Figure 3 and Table 5). Such a low error rate opens the possibility for effective computer support of medical diagnosis on the basis of optimal linear combination of selected phenotypic and genetic features. Moreover, an individualization ofdiagnosis and/or therapy can also be considered on the basis of our methods, as, for example, the application of the diagnostic map (Figure 4). Nevertheless, the results of the current study should be considered as hypothesis generating and need to be confirmed in separate evaluations, if possible in another larger group of patients.

Supporting Information

Appendix S1 Mathematical foundations of the *RLS* method of feature selection.
(PDF)

Author Contributions

Conceived and designed the experiments: LB TŁ BL PS OH JA PB JJC ARQ KL LN MS JW. Performed the experiments: LB TŁ PS OH PB ARQ KL LN. Analyzed the data: LB TŁ JA MD JW. Contributed reagents/materials/analysis tools: LB TŁ BL PS OH JA PB JJC ARQ KL LN MS JW. Wrote the paper: LB TŁ BL PS JA JJC MD LN JW.

References

1. Johnson RA, Wichern DW (1991) Applied Multivariate Statistical Analysis. New York: Prentiice-Hall Inc., Englewood Cliffs.
2. Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press.
3. Duda OR, Heart PE, Stork DG (2001) Pattern Classification. John Wiley & Sons.
4. Liu H, Motoda H (2008) Computational methods of feature selection. Chapman & Hall/CRC.
5. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46: 389–422.
6. Ding S, Zhu H, Jia W, Su C (2012) A survey on feature extraction for pattern recognition. Artificial Intelligence Review 37: 169–180.

7. Stenvinkel P, Heimburger O, Paultre F, Diczfalusy U, Wang T, et al. (1999) Strong association between malnutrition, inammation, and atherosclerosis in chronic renal failure. Kidney International 55: 1899–1911.
8. Luttropp K, Lindholm B, Carrero JJ, Glorieux G, Schepers E, et al. (2009) Genetics/genomics in chronic kidney disease - towards personalized medicine? Semin Dial 22: 417–422.
9. Pecoits-Filho R, Nordfors L, Lindholm B, M HC, Schalling M, et al. (2003) Genetic approaches in the clinical investigation of complex disorders: Malnutrition, infammation, and atherosclerosis (mia) as a prototype. Kidney International 63: 162–167.
10. Kononenko I (1994) Estimating attributes: analysis and extensions of Relief. Machine Learning, ECML-94 : 171182.
11. Kira K, Rendell LA (1992) A practical approach to feature selection. Proceedings of the ninth international workshop on machine learning : 249–256.
12. Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. Proceedings of the 17th International Conference on Machine Learning : 359–366.
13. Duan KB, Rajapakse JC, Wang H, Azuaje F (2005) Multiple svm-rfe for gene selection in cancer classification with expression data. IEEE Trans Nanobiosci : 228–234.
14. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3: 185–205.
15. Breiman L (2001) Random Forests. Machine Learning 45: 5–32.
16. Vapnik VN (1988) Statistical Learning Theory. New York: J. Wiley.
17. Bobrowski L, Lukaszuk T (2009) Feature Selection Based on Relaxed Linear Separability. Biocybernetics and Biomedical Engineering 29: 43–59.
18. Bobrowski L, Lukaszuk T (2011) Relaxed Linear Separability (RLS) Approach to Feature (Gene) Subset Selection. In: Xia X, editor, Selected Works in Bioinformat- ics, InTech. Availabl: http://www.intechopen.com/articles/show/title/relaxed-linear- separability-rls-approach-to-feature-gene-subset-selection.
19. Minsky ML, Papert SA (1987) Perceptrons - Expanded Edition. An Introduction to Computational Geometry. MIT Press.
20. Bobrowski L (2005) Data mining based on convex and piecewise linear (CPL) criterion functions (in Polish). Bialystok: Bialystok University of Technology.
21. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.
22. Bobrowski L (1991) Design of piecewise linear classifiers from formal neurons by some basis exchange technique. Pattern Recognition 24: 863–870.
23. van0020't Veer L J, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.
24. Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7: 1–30.
25. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine learning 46: 389–422.
26. Witten I, Frank E (2000) Data Mining - Pracitcal Machine Learning Tools and Techniques with JAVA Implementations. Morgan Kaufmann Publisher.
27. (multiple) Support Vector Machine Recursive Feature Elimination - mSVM-RFE. Available: http://www.colbyimaging.com/wiki/statistics/msvm-rfe. Accessed 2013 Sep 1.
28. mRMR (minimum Redundancy Maximum Relevance Feature Selection). Available: http://penglab.janelia.org/proj/mRMR/. Accessed 2013 Dec 1.
29. Stenvinkel P, Carrero JJ, Axelsson J, Lindholm B, Heimburger O, et al. (2008) Emerging biomarkers for evaluating cardiovascular risk in the chronic kidney disease patient: how do new pieces fit into the uremic puzzle? Clinical Journal of the American Society of Nephrology 3: 505–521.
30. Ahmad T, Fiuzat M, Felker GM, O'Connor C (2012) Novel biomarkers in chronic heart failure. Nature Reviews Cardiology 9: 347–359.
31. Fouque D, Kalantar-Zadeh K, Kopple J, Cano N, Chauveau P, et al. (2008) A proposed nomenclature and diagnostic criteria for protein-energy wasting in acute and chronic kidney disease. Kidney International 73: 391–398.
32. Bobrowski L, Wasyluk H (2001) Diagnosis Supporting Rules of the Hepar Systems. In: MEDINFO 2001. IOS Press, Studies in Health Technology and Informatics, 1309–1313.