Weighted clustering and ROC analysis in assessment of the quality of life in patients with chronic heart failure

Aleksander Owczarek[1], **Bożena Szyguła-Jurkiewicz**[2], **Michał Cogiel**[3], **Damian Grzechca**[4]

[1] Division of Statistics, Medical University of Silesia, Poland
[2] III Department of Cardiology, Medical University of Silesia, Poland
[3] Student Scientific Group, Medical University of Silesia, Poland
[4] Department of Electronics, Silesian Technical University in Gliwice, Poland

Abstract. Chronic heart failure is increasingly prevalent in population and has a significant impact on the length as well as the quality of patients' life. In Polish population there are no norms for the SF-36 test to assess the health related quality of life (HRQoL). Weighted *k*-means algorithm has been used to divide the population into 2 groups with better and worse quality of life and then cut-off points have been calculated based on the ROC curves analysis. Vitality has been the best discriminating factor. Poor quality of life was related with higher risk of depression development, MACE (major adverse cardiac event) occurrence and worse clinical parameters.

Introduction

Chronic heart failure (CHF) is associated with high mortality and morbidity regardless of the development in pharmacological treatment [1]. Several risk factors for major cardiac adverse events (MACE: sudden death, hospitalization due to exacerbation of CHF) have been already identified: elevated creatinine plasma level, age, female gender, NYHA class, left and right ventricular function, the peak exercise oxygen reuptake test and brain natriuretic peptide plasma level [2–4]. Depression is another very important risk factor (as it turned out recently). It is highly prevalent in this group of patients and may bias patients' reports of their Health Related Quality of Life (HRQoL) [5]. Such patients are not only more likely to develop depression, but once depressed, they are more likely to experience deteriorating heart disease, need repeated procedures, or die due to MACE.

The HRQoL, which was found to be very important in the assessment of CHF progression and treatment results, may be evaluated with The Short Form (36) Health Survey [6]. Results obtained in such a survey of patient's

ISBN 978–83–7431–336–0 ISSN 0860-150X

health are compared to cut-off points (different for various populations) and patient's well-being and its changes during the treatment are estimated. However, there are no cut-off points marked in the Polish population. These which are in the test instruction refer to American population [6]. In view of great social, cultural and economical differences between Polish and American population, referring the SF-36 results to these norms would not reflect the reality. This makes it difficult to evaluate outcomes yielded with the SF-36 test.

The purpose of the present study was to: 1) assess cut-off points for scales of the SF-36 test, 2) find which scales of the test best discriminate groups according to the HRQoL, 3) compare group with better and worst HRQoL regarding clinical parameters, depression and MACE occurrence.

Material and methods

Material

One hundred and ninety three consecutive patients with chronic systolic heart failure were included in the prospective study. Detailed (medical) description of patient treatment as well as measured clinical parameters one may find in our previous paper [7]. *Inclusion criteria*: 1. symptoms of systolic heart failure for at least 2 years; 2. increased LV end-diastolic diameter (LVEDD > 57 mm) and reduced LV ejection fraction (LVEF $< 45\%$) shown by the ECG; 3. 5-year or longer history of hypertension before the onset of heart failure symptoms (documented at least 2 episodes of systolic blood pressure ≥ 140 mmHg and/or diastolic blood pressure ≥ 90 mmHg); 4. lack of significant ($> 30\%$) narrowing in coronary arteries indicated by the coronary angiogram. *Exclusion criteria*: 1. confirmed coronary artery disease and/or history of myocardial infarction; 2. acquired or congenital valve disease leading to impairment of myocardial function excluding functional mitral and/or tricuspid regurgitation; 3. connective tissue disease and/or neoplasm; 4. infection; 5. endocrine diseases, i.e. diabetes mellitus, hyper- or hypothyroidism, Cushing disease; 6. advanced liver or kidney disease.

HRQoL was measured with the SF-36 test. The SF-36 encloses eight scaled scores, which are sums of the questions (36) in corresponding section. Each scale is directly transformed into a 0–100 scale on the assumption that each question carries equal weight. These eight parts are respectively: Physical Functioning *PF*, Role Physical *RP*, Bodily Pain *BP*, General Health *GH*, Vitality *V*, Social Functioning *SF*, Emotional Role Functioning *RE* and Mental Health *MH*. First four coefficients are related to physical functioning and four consecutive ones to mental health. The presence of depression was diagnosed according to patient's history, clinical observation, the Beck Depression Inventory [8] and the Hamilton rating scale for depression [9]. If depression was suspected patient were consulted by a psychiatrist. The clinical observation of patients began on admission to hospital and lasted for 36 months.

Methods

Weighted k-means clustering: The algorithm proposed in [10] has been used to find a partition of a dataset X, with $M = 193$ records and $N = 8$ features corresponding to the SF-36 test scales, into $k = 2$ clusters. It is a modification of classical k-means clustering, however to identify the importance of different features, a weight is assigned to each feature in the distance calculation. Formally, the minimization of the following objective function is being done:

$$Q(U,Z,W) = \sum_{l=1}^{k} \sum_{i=1}^{M} \sum_{j=1}^{N} u_{i,l} w_j^{\beta} d(x_{i,j}, z_{l,j}) \tag{1}$$

where:

U – an $M \times k$ partition matrix, $u_{i,l} \in \{0,1\}$,
$Z = \{Z_1, Z_2, \ldots, Z_k\}$ – a set of k vectors representing the k-clusters centers,
$W = [w_1, w_2, \ldots, w_N]$ – a set of weights,
$d(x_{i,j}, z_{l,j})$ – a distance or dissimilarity measure between object i-th and the center of l-th cluster on the j-th feature; in paper we used: $d(x_{i,j}, z_{l,j}) = (x_{i,j} - z_{l,j})^2$,
$\beta > 1$ – a fuzziness parameter.

W-k-means clustering algorithm:

A. Random generation of initial set of weights W^0, $\sum_{j=1}^{N} w_j = 1$ and partitioning matrix U^0; set $t = 0$.

B. In original paper authors suggested to choose randomly an initial set of Z. However, to improve final results, we used a classical k-means algorithm. Calculation of Q.

C. Update of matrix U: $u_{i,j}^{(t+1)} = 1$ if

$$\underset{1 \le t \le k}{\forall} \sum_{j=1}^{N} w_j^{\beta} d(x_{i,j}, z_{l,j}) \le \sum_{j=1}^{N} w_j^{\beta} d(x_{i,j}, z_{t,j}) \tag{2}$$

otherwise for $t \neq l$ $u_{i,j}^{(t+1)} = 0$.

D. Update of matrix Z:

$$\underset{1\le l\le k}{\forall}\ \underset{1\le j\le N}{\forall} z_{l,j}^{(t+1)} = \left(\sum_{i=1}^{M} u_{i,l}x_{i,j}\right)\cdot\left(\sum_{i=1}^{M} u_{i,l}\right)^{-1} \tag{3}$$

E. Update of matrix W: $w_j^{(t+1)} = 0$ if $D_j = 0$, otherwise

$$w_j^{(t+1)} = \left(\sum_{s=1}^{h} [D_j/D_s]^{\frac{1}{\beta-1}}\right)^{-1} \tag{4}$$

where: h is the number of features with $D_j \neq 0$ and

$$D_j = \sum_{l=1}^{k}\sum_{i=1}^{M} u_{i,j} d(x_{i,j}, z_{l,j}) \tag{5}$$

Then Q recalculation. If $Q^{(t+1)} = Q^{(t)}$ then stop, else go to step C.

In order to refine the optimal β coefficient we used the Bouldin-Davies (DB) index (6) and mean Cohen's effect size (ES) – for all eight scales.

$$DB = \frac{1}{2}\sum_{i=1,i\neq j}^{n} \max\left(\frac{\sigma_1+\sigma_2}{d(z_1,z_2)}\right) \tag{6}$$

where: σ_1 – the average distance of all patterns in the i-th cluster to their cluster center z_i, $d(z_1, z_2)$ – distance of cluster centers z_1 and z_2. Small values of DB correspond to clusters that are compact, and whose centers are far away from each other [11]. ES was also used as a measure of the strength of the particular SF-36 test scales in the relationship between two groups yielded by the w-k-means algorithm. Cohen's ES is defined as the difference between two means divided by a standard deviation for the data [12]. The larger ES, the bigger size of the effect (higher relevance of analyzed factor). Coefficient values are important in interpreting the data, as it is possible to determine, not only the statistical significance but clinically relevant changes (or differences) in the quality of life [13–14].

In order to illustrate the quality of obtained distribution of the population into two groups with different HRQoL we used the Principal Component Analysis. To specify cut-off points for each of the scales ROC curves were plotted and typical performance measures for the confusion matrix were calculated (including Matthews correlation coefficient MCC [15] and normalized mutual information NMI [16]). Relative risk with confidence intervals was calculated to assess the risk of development (or having) depression and MACE. Kaplan-Meier survival curves for both groups were computed and compared with the log-rank test. For clinical data we used: the χ^2 test for categorical data, for interval data (including the SF-36 scales) with normal distribution or after normalization with the Box-Cox transformation, the t-Student test, otherwise the U Mann-Whitney test. Variables distribution was evaluated with the Shapiro-Wilk test. Homogeneity of variances was assessed by the Levene test.

Results

Weighted k-means clustering results

We used the DB-index and mean effect size values to select the β parameter in order to get optimal clustering results. Based on results presented in the [Fig. 1] we chose $\beta = 2.4$.

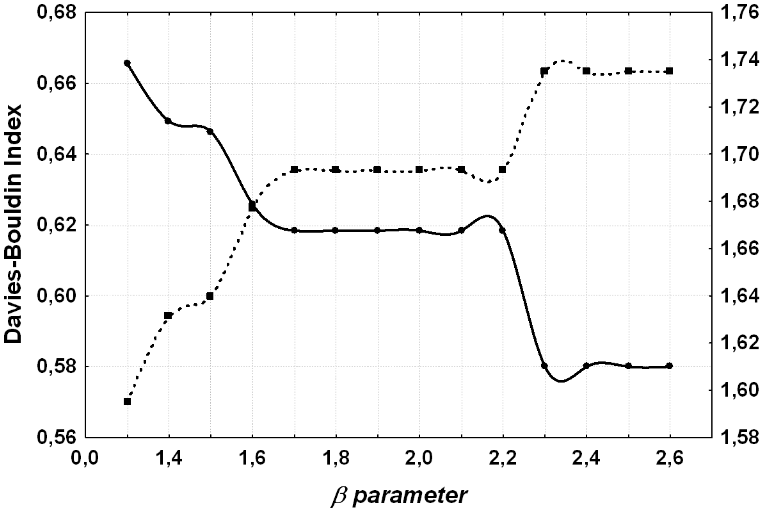

Fig. 1. Values of DB-Index (*solid line*) and mean Effect Size (*dashed line*) according to β parameter in weighted-k-means clustering algorithm

[Fig. 2] and [Tab. 1] present respectively the PCA results and weights yielded by the clustering algorithm as well as ES values for each scale of the SF-36 test. As it can be seen, the obtained groups are well separated from each other. The most important weights proved to be Vitality and the worst one the Role Emotional. Taking into account ES, the best factor was the same, however the worst one was the Physical Functioning.

[Fig. 3] shows comparison of the SF-36 scales between both groups. Statistically significant differences between all eight scales were found ($p < 0.001$).

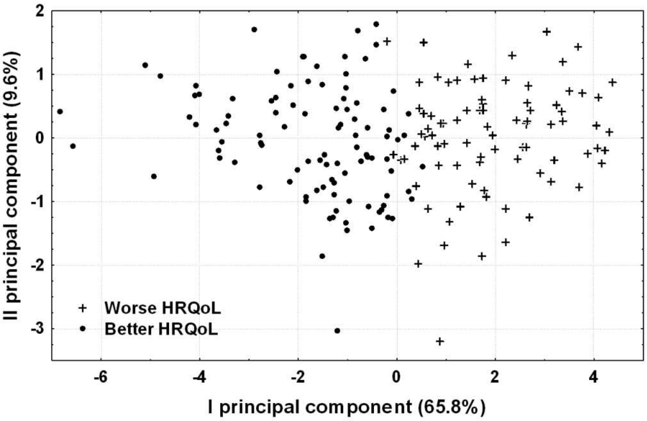

Fig. 2. The PCA projection for the SF-36 test scales

Tab. 1. Weights yielded by the W-k-mean algorithm and corresponding mean Effect Sizes values

SF-36 scale	PF	RP	BP	GH
Weight	0.1028	0.1147	0.1036	0.1526
Cohen's ES	1.2946	1.6044	1.7259	1.9981
SF-36 scale	V	SF	RE	MH
Weight	0.1815	0.1403	0.0820	0.1225
Cohen's ES	2.7457	2.1923	1.5417	1.7762

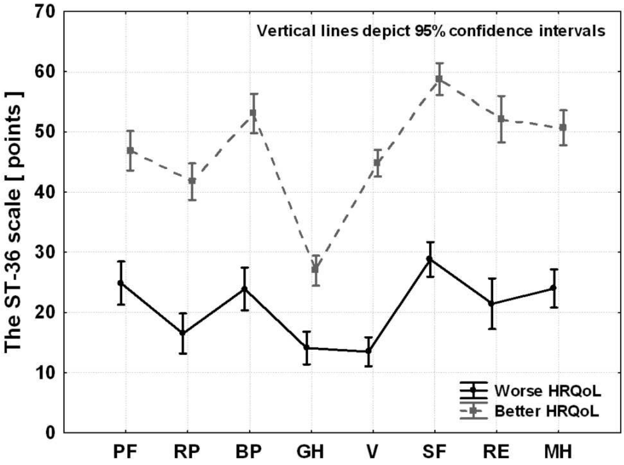

Fig. 3. Comparison of the SF-36 scales between both groups

ROC curves

[Tab. 2] presents results based on ROC curves. For all scales cut-off points were computed. Sensitivity and specificity are one approach to quantify the diagnostic ability of the test. This coefficients measure respectively the proportion of actual positives and the proportion of negatives which are correctly identified. A test with a high specificity has a low statistical significance (α), while a test with a high sensitivity has a low statistical power ($1 - \beta$). In clinical practice, however, the test result is all that is known, so we want to know how good the test is at predicting the disease.

Tab. 2. Results of the ROC analysis

	PF	RP	BP	GH	V	SF	RE	MH	All
	> 40	> 31	> 32	> 15	> 25	> 50	> 33	> 40	> 265
AUC	0.816	0.879	0.888	0.763	0.965	0.942	0.862	0.881	0.994
Se	0.705	0.648	0.933	0.867	0.924	0.924	0.781	0.905	0.981
Sp	0.795	0.932	0.727	0.580	0.932	0.852	0.818	0.761	0.955
ACC	0.746	0.777	0.839	0.420	0.927	0.891	0.798	0.839	0.969
PPV	0.804	0.919	0.803	0.736	0.942	0.882	0.837	0.819	0.963
NPV	0.693	0.689	0.901	0.711	0.911	0.904	0.758	0.870	0.977
FPR	0.205	0.068	0.273	0.420	0.068	0.148	0.182	0.239	0.045
FNR	0.295	0.352	0.067	0.133	0.076	0.076	0.219	0.095	0.019
MCC	0.499	0.594	0.682	0.470	0.854	0.781	0.597	0.678	0.937
NMI	0.883	0.897	0.909	0.880	0.946	0.928	0.896	0.908	0.971

Thus, we have also calculated other parameters, especially positive predictive value and normalized mutual information. Positive predictive value is the proportion of patients with positive test results who are properly diagnosed. It is a key measure of the diagnostic method as it reflects the probability that a positive test corresponds to the underlying condition being tested for, in our case, better HRQoL.

The problem with PPV is that its value depends also on the prevalence of the disease, which of course may vary. In order to deal with this problem it should only be used if the ratio of the number of patients in the disease group and the number of patients in the healthy control group is equivalent to the prevalence of the diseases in the studied population. However, our study on HRQoL in patients with systolic heart failure in Poland is unique and there is no information about the worst quality of life prevalence in CHF population. This led us to normalized mutual information which is interpreted as an amount by which the model reduces our uncertainty about the true state. As it can be seen, the best discriminating value has Vitality (highest area under curve, best accuracy and positive predictive value, Matthews correlation coefficient and mutual information) and then Social Functioning and Mental Health.

Statistical analysis

For the sum of all scales we obtained high value of NMI. Patient with CHF who has more than 265 points is very likely to have good HRQoL. Relative risk (RR) of depression development for a person with less than 265 points is 3.29 (95% CI: 2.04–5.32; $p < 0.0001$). RR for MACE occurrence is 1.83 (95% CI: 1.31–2.55; $p < 0.001$). The Number Needed to Treat which is the number of patients who need to be treated in order to prevent MACE outcome is 3.79, so we need to treat 4 patients to avoid one adverse cardiac event.

[Fig. 4] shows Kaplan-Meier survival curves in both yielded by the w-k-mean algorithm groups. Patients with the worst quality of life statistically significantly more often and earlier underwent adverse cardiac events than the other group. In [Tab. 3] we enclosed the comparison of relevant for the HRQoL clinical parameters between both groups.

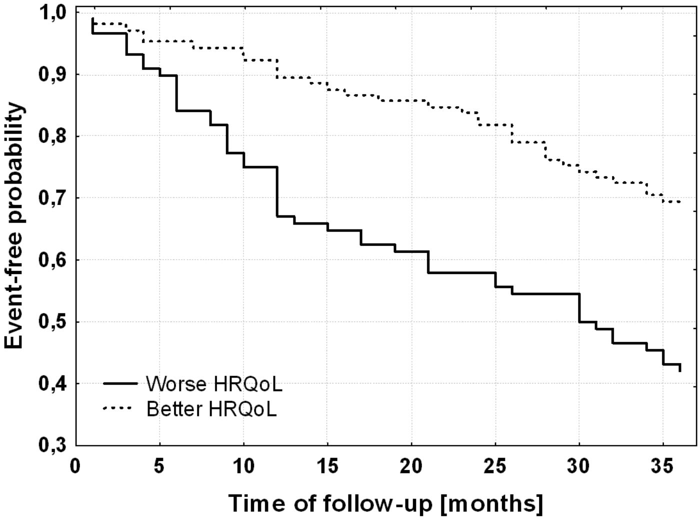

Fig. 4. Kaplan-Meier curves of MACE-free probability in group with better and worst HRQoL ($p_{\text{log-rank}} < 0.001$)

Tab. 3. Comparison of Clinical parameters Between group with better and worse HRQoL

Parameter	+ HRQoL	– HRQoL	p
Death	5 (4.76%)	31 (35.23%)	< 0.001
MACE	33 (31.43%)	51 (57.95%)	< 0.001
PAP>19 [mmHg]	49 (46.67%)	56 (64.64%)	< 0.05
RAP>5 [mmHg]	47 (44.76%)	54 (61.36%)	< 0.05
hs-CRP [mg/l]	1.64/4.85	2.80/4.22	< 0.01
NT-pro BNP [pg/ml]	548/972	1669/3055	< 0.001
Bilirubin [Bmol/l]	16.25/10.22	18.45/12.10	0.0964
Long QT	28 (26.67%)	39 (44.32%)	< 0.05
LVEDD [mm]	65.8±8.3	68.1±7.9	< 0.05
LVESD [mm]	51.0±9.5	53.5±9.6	< 0.05
IVRT [s]	71.0/40.0	60.0/30.0	< 0.01
TAPSE [mm]	21.0/5.0	19.0/10.0	< 0.001
E/A	1.4/1.2	1.8/2.7	< 0.01

Mean±STD or Me/IQR (Interquartile range)

As it can be seen, patients with the worst quality of life have statistically significant elevated pulmonary and right arterial pressure as well as bilirubin, high sensitive C-reactive protein (CRP) and brain natriuretic peptide NT-pro BNP plasma level. They also have almost twice often long QT syndrome (in electrocardiogram). In echocardiography, these patients asserted higher left ventricular (LV) end diastolic and systolic diameter, isovolumetric relaxation time, the E/A ratio of transmitral flow and lower tricuspid annular plane systolic excursion (TAPSE).

Conclusions

1. Application of the weighted k-means algorithm yields two well separated in the SF-36 scales dimension groups with poor and good HRQoL. Weights obtained in the clustering process correspond generally with clinically relevant differences measured with the Cohen's ES and evaluation measurements of ROC curves. The obtained groups differ significantly in all eight scales of the SF-36 test.

2. For all scales cut-off points were calculated and Vitality proved to be the best discriminating SF-36 scale. Vitality corresponds with patient's well-being and "life energy". So, there is less than 6% chance that the patient with more than 25 points has poor HRQoL. 92% of the patients with real good quality of life will be correctly identified by this scale (nevertheless all SF-36 scales should be taken into consideration in assessment of patient's HRQoL).

3. Taking into account the sum of all SF-36 scales, there is less than 4% chance that the patient with more than 265 points has worse HRQoL. 99.4% of the patients who actually have good quality of life will be correctly identified with this test.

4. It is undeniable that poor Health Related Quality of Life (less than 265 points in all scales of the SF-36 test) is associated with higher risk of hospitalization and death occurrence as well as with depression development. Therefore, it is **very important** to perform screening tests for quality of life (and further for depression) in all patients with chronic heart failure, because effective treatment of depression may improve their long-time prognosis.

5. Quality of life is strongly associated with the worst clinical parameters [Tab. 3]. Depression might promote an inflammatory response (represented by CRP and NT-pro BNP) by activating the immune response. Alternatively, the effects of depression on inflammation might be due to its links with psychological stress. Worse echocardiography parameters and higher values of arterial pressure are related with heart remodeling in chronic heart failure. On the other hand, heart insufficiency handicaps patient's physical and social activity.

List of abbreviations

ACC	Accuracy
AUC	Area Under ROC Curve
FNR	False Negative Rate
FPR	False Positive Rate
MCC	Matthews correlation coefficient
NMI	Normalized Mutual Information
NPV	Negative Predictive Value
PPV	Positive Predictive Value
ROC	Receiver Operating Curve
Se/Sp	Sensitivity/Specificity
STD	Standard Deviation
WKM	Weighted K-Means Algorithm

REFERENCES

[1] Jessup M., Brozena S. C., Heart failure, N Engl J Med, 348 (20), pp. 2007–2018, May 2003.

[2] Muntwyler J., Abetel G., Gruner C., Follath F., One-year mortality among unselected outpatients with heart failure, Eur Heart J, 23 (23), pp. 1861–1866, December 2002.

[3] Cohn J. N., Johnson G. R., Shabetai R., et al., Ejection fraction, peak exercise oxygen consumption, cardiothoracic ratio, ventricular arrhythmias, and plasma norepinephrine as determinants of prognosis in heart failure. The V-HeFT VA Cooperative Studies Group, Circulation, 87 (6), pp. 5–16, Juni 1993.

[4] de Groote P., Dagorn J., Soudan B., et al., B-type natriuretic peptide and peak exercise oxygen consumption provide independent information for risk stratification in patients with stable congestive heart failure, J Am Coll Cardiol, 43 (9), pp. 1587–1589, May 2004.

[5] Faller H., Störk S., Schuler M., et al., Depression and disease severity as predictors of health-related quality of life in patients with chronic heart failure – a structural equation modeling approach, J Card Fail, 15 (4), pp. 286–292, May 2009.

[6] Ware J. E., Kosinski M., Dewey J. E., How to score Version 2 of the SF-36® Health Survey (Standard and Acute Forms), Medical Outcomes Trust and QualityMetric, Incorporated 2002.

[7] Szyguła-Jurkiewicz B., Owczarek A., Duszańska A., et al., Long-term prognosis and risk factors for cardiac adverse events in patient with chronic systolic heart failure due to hypertension, PAMW, 118 (5), pp. 280–287, 2008.

[8] Beck A. T., Ward C. H., Mendelson M., et al., An inventory for measuring depression, Arch Gen Psychiatry, 4, pp. 561–571, 1961.

[9] Hamilton M., A rating scale for depression, J Neurol Neurosurg Psych, 23, pp. 56–62, 1961.

[10] Huang Z., Ng M. K., Rong H., Li Z., Automated variable weighting In k-means type clustering, IEEE PAMI, 27 (5), pp. 657–668, 2005.

[11] Halkidi M., Batistakis Y., Vazigriannis M., On clustering validation techniques, J Intell Inf Syst, 17 (2), pp. 107–145, 2001.

[12] Kazis L. E., Anderson J. J., Meenan R. F., Effect sizes for interpreting changes in health status, Med Care, 21, pp. 178–189, 1989.

[13] Osoba D., King M., Meaningful differences In: Fayers P., Hays R. (eds): Assessing quality of life in clinical trials. II Ed., Medical Press, pp. 243–259, 2005.

[14] Sprangers M. A., Moinpour C. M., Moynihan T. J., et al., Assessing meaningful change in quality of life over time: a user's guide for clinicians, Mayo Clin Proc, 77, pp. 561–571, 2002.

[15] Baldi P., Soren B., Chauvin Y., et al., Assessing the accuracy of prediction algorithm for classification, an overview, Bioinfo Rev, 16 (5), pp. 412–424, 2000.

[16] Bush W. S., Edwards T. L., Dudek S. M., et al., Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction, BMC Bioinformatics, 9 (238), pp. 1–17, 2008.