ISSN 0002-306X. Proc. of the RA NAS and NPUA Ser. of tech. sc. 2023. V. LXXVI, N2

UDC 004.832

COMPUTER SCIENCE AND INFORMATICS

DOI: 10.53297/0002306X-2023.v76.2-210

L.K. ANDREASYAN

A METHOD FOR ASSESSING THE RISK OF DEVELOPMENT OF DISEASE COMPLICATIONS WITH METABOLIC SYNDROME BASED ON THE CUMULATIVE DISTRIBUTION FUNCTION

It is relevant to assess the risks of related diseases based on the electronic data of patients in order to prevent the progress of further complications of metabolic syndrome (MS). This article describes the Archimedean copula-based method of assessing the risks of developing related diseases based on diagnosis and risk factors associated with complications and readmissions.

Keywords: metabolic syndrome, etiological model, logistic regression, Archimedean copula functions, cumulative distribution functions.

Introduction. According to medical guidelines, the primary prevention of diseases is the prevention of the development of a new disease, secondary prevention is the prevention of complications arising from an already started disease, and tertiary prevention is the reduction of the impact of an ongoing disease [1]. Disease diagnosis by doctors are traditionally made based on the results of diagnostic tests and medical research at the same time based on their own knowledge and experience, and as a rule, no assessment of the development of complications is made.

In order to prevent complications, different types of systems and methods have been created that model the risk processes to evaluate the progression of disease complications and their impact. Among the well-known systems are APACHE II (Acute Physiology and Chronic Health conditions score), SOFA (Sequential Organ Failure Assessment), Multiple Organ Dysfunction Score (MODS) expert-based systems, and disease progression (dynamics) assessment systems based on complication development modeling and probability methods [2-4]. The main disadvantages of the above-mentioned systems are the problems associated with the processing of incomplete medical data and the high approximation of the modeling result indicators. For middle-aged patients, 5- or 10-year disease risks are often calculated using the Weibull model [5, 6], but if the risk function takes the form of a more complex curve, the Weibull distribution leads to implausible predictions [3]. Charlson (17 categories) and Elixhauser (30 categories) evaluation

indices are among the well-known comorbidity risk assessment methods [7]. Although the Elixhauser index showed a better risk assessment result than the Charlson index, they are also approximations, as these indices were originally designed to assess the survival risks, not to predict the risks of developing a particular disease [7, 8].

For risk assessment, the correlation analysis method is used to reveal the existence of a linear relationship between random variables - the Pearson method is most often used to calculate correlation coefficients. However, if there is a strong non-linear relationship between two variables, the correlation analysis method may underestimate the true strength of the relationship [9]. In the presented method, multivariate functions of the cumulative distribution of connectivity were used to identify non-linear relationships between random variables and to determine the strength of the relationship for risk assessment [10].

The purpose of the article is to investigate the nonlinear relationships between the risk factors contributing to the development of metabolic syndrome complications and readmissions based on electronic patient data, and to develop a risk assessment method for predicting such complications using machine learning supervised classification logistic regression algorithm and Archimedean cumulative multivariate distribution functions. In this regard, the following research and development should be done:

1. To separate diseases related to MS according to the international system of classification of diseases ICD-09-CM/ICD-10-CM (International Classification of Diseases, Ninth or Tenth Revision, Clinical Modification) [11, 12].

2. To select the risk factors that contribute to the development of complications of MS diseases based on current medical guidelines [13, 14].

3. To evaluate the risks of the development of comorbidity complications using Archimedean cumulative multivariate distribution functions.

Disease complications, risk factors, and the etiological model. According to medical guidelines, there are several types of complications, the main ones being:

a) Comorbidity - a complication caused by the presence of diseases related to the given disease.

b) Multimorbidity - a complication caused by the presence of different types of diseases.

The main difference between the aforementioned complications is that in the case of comorbidity disease, the primary disease is distinguished from the related diseases, which is the primary focus of attention from the point of view of treatment. In the case of multimorbidity disease, the accompanying diseases are not superior to each other. According to medical research, risk factors can be internal,

pathos-physiological (genetic), external, environmental, behavioral, and social. These factors are further divided into modifiable and invariable risk factors. Modifiable risk factors are factors that can be modified by certain interventions, such as behavioral or lifestyle changes, medical procedures, or pharmaceutical treatments. Invariable risk factors include genetics, e.g. family history of cardiovascular disease or diabetes. To predict the progression of disease complications, it is necessary to take into account the influence of both modifiable and invariable risk factors.

The features of the etiological model of the development of disease complications were compared for the risk assessment and complication development model. According to Valderas and others [15], the etiological model of the development of diseases can basically be represented by several schemas (Fig. 1). In the etiological model, diseases are considered to be directly related when one causes the other, e.g. diabetes can cause cataracts, or when diseases are caused by a direct link between risks, e.g. heavy smoking and drinking causes cirrhosis of the liver or lung cancer. Another possibility is that there is no direct relationship between the risks in which case each may cause the same disease, e.g. smoking and age both may contribute to coronary heart disease, lung cancer, or type 2 diabetes mellitus (T2DM) (Fig. 1).

In the etiological model, the cause-and-effect relationships of the occurrence of diseases is presented. This is necessary to consider but not sufficient for risk assessment.

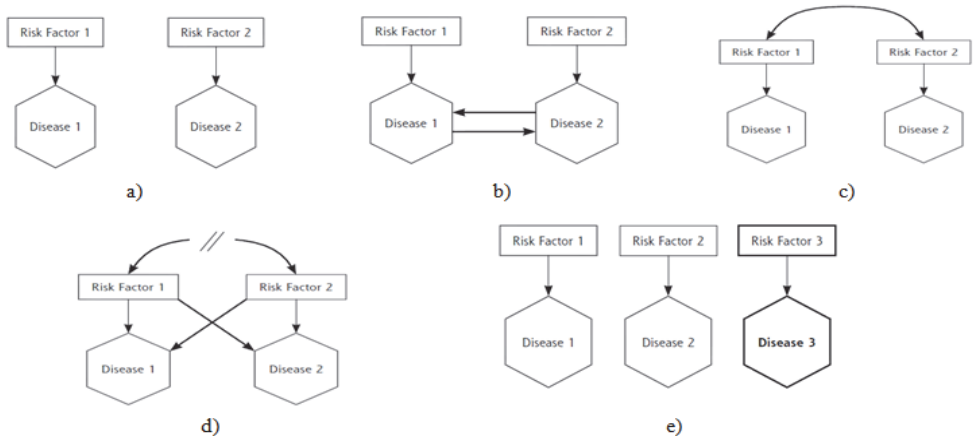

Fig.1. The etiological model of disease development. a - diseases caused independently of each other, b - directly related diseases, c - diseases caused by a direct connection between risks, d - independent risks can cause the same diseases, e - the presence of diseases is explained by the presence of another disease

ICD codes and MS description. ICDs are disease classification code schemes that describe diseases and their causes. These codes include two types of information: the name of the disease and the variants of the diseases caused by them, for example, in the codes E11.311, E11.319 and E11.341 in the ICD-10-CM version, E11 refers to T2DM disease, and 311, 319 and 341 codes are the different manifestations of retinopathy caused by that disease [11].

Metabolic syndrome is the combination of two or more metabolic disorders, increased insulin levels in the blood, presence of high-density lipoprotein cholesterol, thrombogenesis tendencies, arterial hypertension, and obesity [13, 14]. MS increases the risk of various cardiovascular diseases and T2DM. Although there is a lot of data on MS diseases based on which diagnoses are made, the existing forms of treatment often do not stop the development of disease complications as the manifestations of these diseases are heterogeneous from the point of view of clinical presentation. Therefore, it is relevant to assess the risks of developing complications in order to improve MS disease treatments.

A copula functions. According to Sklar's theorem, copulas are functions that combine multivariate distribution functions with their univariate marginal distribution functions. The copula-based approach makes it possible to separate the marginal distributions and estimate the dependence between them, if they are known [16]. The advantages of the copula method are:

- the ability to model non-linear dependence;
- the marginal distribution can be any univariate distribution;
- if the marginal distributions are fixed and the number of parameters to be estimated is the same, choosing the copula with the smallest information criterion is equivalent to choosing the copula with the largest log-likelihood value;
- the mathematical structure of the copula method is such that it allows one to easily add additional marginal distributions.

The use of copula makes it possible to transform random variables through their cumulative distributions into uniformly distributed variables, to characterize the dependence of a set of random variables, regardless of the marginal distributions:

$$C(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d), \quad (1)$$

where $U_i \sim U(0, 1)$ $i = 1, \ldots, d$ random variables with uniform distributions are obtained by applying the integral transformation of the probability to each of the marginals $F_1(x), \ldots, F_d(x)$, so that $U_1 = F_1(X), \ldots, U_d = F_d(X)$. Given $X_1, \ldots, X_d$ for random variables with joint probability distribution H and marginal distribution functions $F_i(x)$ $(i = 1, \ldots, d)$ the copula function C is defined by the following formula:

$$C(u_1, \ldots, u_d) = H[F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)] \quad (2)$$

of the copula function C and $F_1(x)$, …, $F_d(x)$ for arbitrary distribution functions, the function H is defined as:

$$H(X_1, \ldots, X_d) = C[F_1(x), \ldots, F_d(x)]. \quad (3)$$

Copula functions belonging to the class of Archimedean family are presented by the following formula:

$$C(u_1, \ldots, u_d; \theta) = \psi^{[-1]}(\psi(u_1; \theta) + \cdots + \psi(u_d; \theta); \theta), \quad (4)$$

where ψ is the generator function and $\psi^{[-1]}$ is its inverse, $\theta \in \Theta$ is the set of parameters (Fig. 2).

Archimedean family		$\varphi(t)$
AMH	$u_1 u_2 \{1-\theta(1-u_1)(1-u_2)\}^{-1}$	$\log\left\{\frac{1-\theta(1-t)}{t}\right\}$
Clayton	$(u_1^{-\theta}+u_2^{-\theta}-1)^{-1/\theta}$	$\theta^{-1}(t^{-\theta}-1)$
Frank	$-\theta^{-1}\log\left\{1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{(e^{-\theta}-1)}\right\}$	$-\log\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$
Gumbel	$\exp\left[-\{(-\log u_1)^\theta+(-\log u_2)^\theta\}^{1/\theta}\right]$	$\{-\log(t)\}^\theta$
Joe	$1-\{(\bar{u}_1)^\theta+(\bar{u}_2)^\theta-(\bar{u}_1\bar{u}_2)^\theta\}^{1/\theta}$	$-\log\{1-(1-t)^\theta\}$

Fig.2. Archimedean copula functions

For Archimedean copula functions, the strength of dependence through the parameter θ in the interval [-1, 1] is determined using the Kendall's τ correlation coefficient by the following formula:

$$\tau = 1+4\int \psi(t)/\psi'(t)\, dt. \quad (5)$$

The closer the value of the coefficient is to 1 (or −1), the greater is the positive (negative) dependence, and on the contrary, a value close to 0 means a weak dependence, equal to 0, no dependence.

The Archimedean copula functions are symmetric or asymmetric, depending on the direction of the tail dependence force. The Frank copula is a symmetric function and is used to present the weak dependences between the tails. Clayton and Gumbel are asymmetric functions. The former is used to present the negative, and the latter the positive tail dependences [10, 16].

Copula-based method description. The copula-based method includes 3 stages:

1. Creation of individual trajectory vectors based on electronic data and the grouping by main disease.

2. Development of dependency models between diseases to evaluate marginal distributions using logistic regression.

3. Based on the obtained results, the calculation of the marginal and joint distributions and the assessment of the development risks of the complications of the MS disease using the Frank Archimedean copula function.

Stage 1 – the formation of groups by the main disease. The individual trajectory-vector of the disease is compiled on the basis of diagnoses with ICD codes, related diseases, and the time interval of their diagnosis. The health trajectory can be represented as a graph where the nodes (vertices) represent the disease and the edges between two nodes indicate that the two diseases occurred at time t. If P is the set of m patients: $P = \{p_1, p_2, ..., p_m\}$, then for each P_i patient at time t we can construct a graph G_i with (V, E) nodes, where V is the vertices, and E is the edges connecting 2 vertices (Fig. 3), so that:

$$V = \{v \in D, 1 \le i \le n, 1 \le |V| \le k\}, \quad (6)$$

$$E = \{(v_1, v_2) \mid v_1 \in d_i, v_2 \in d_j, 1 \le i \le j, v_1 \ne v_2, \text{if } i = j\}. \quad (7)$$

In (6), n is the overall number of possible diseases, k is the number of comorbidities present at time t, 1 means the patient has no comorbidities at time t, and the edges between two nodes (7) indicate that those diseases have occurred. Each V vertex (node) of the G_i graph is the diagnoses with ICD codes from the set of diseases $D = \{d_1, d_2, ..., d_n\}$, to which related diseases are connected at the moment t. The first diagnosis is considered the main disease. Then, based on trajectory vectors with similar histories the graphs are grouped according to the main disease (Fig. 3).

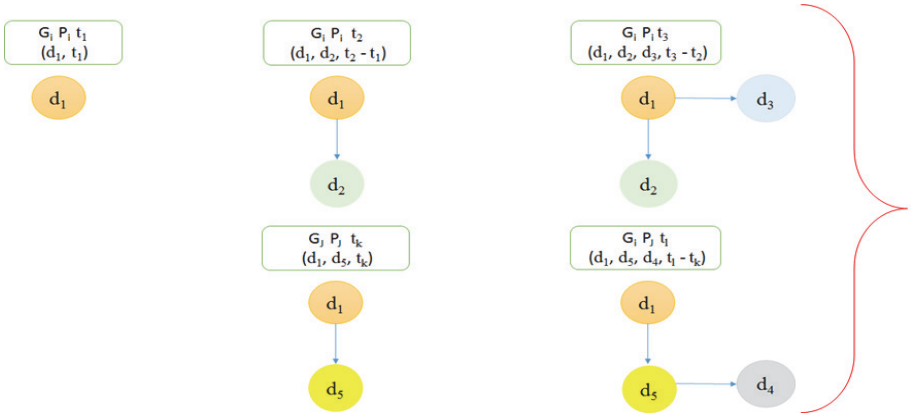

Fig.3. The process of generating and grouping the G_i and G_j graphs by establishing a relationship between related diseases with the main disease d_1 based on patient data P_i and P_j

Stage 2 – logistic regression modeling. Multivariate logistic regression models of the development of each disease were constructed, one for each disease to estimate its marginal distribution [17,18]. In logistic regression, we have a logistic function (also called the sigmoid) that predicts two values (Y = 0 or Y = 1) [19]. A logistic function curve shows a probability, for example, whether or not the risks of disease complications are present. We want to model P(Y = 1) in terms of a set of predictor variables $x_1,\ldots, x_n$, then the multivariate logistic regression equation on the probability scale may be written:

$$P(Y = 1) = \exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n) / 1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n), \quad (8)$$

where the β_i coefficients are the model's learned weights, and β_0 is the bias.

The marginal distribution describes the probability of acquiring a disease regardless of the presence of another disease. Then, the disease-free groups were separated, for which the logistic regression model was built by combining the risk factors. The MLE (Maximum Likelihood Estimation) method was used to estimate the β_i coefficients of the logistic regression:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-xT\beta i}}, \quad (9)$$

where input data $x_i \in X$ represents the feature vector for the i^{th} samples and the output is denoted by Y. The logistic regression model expresses the relationship between $y_i \in Y$ and x_i in terms of the conditional probability P(Y = 1|X) of risks of developing disease complications.

Stage 3 – using the copula function. In order to assess the risks of related diseases, it is necessary to find the joint distributions of random variables, which are all the possible distributions of the patient's disease manifestation. We assign 1 if the patient acquires a disease in a certain period of time, and 0 otherwise. Then, the joint distributions of the probability of acquiring two diseases in a certain time interval will be P(0,0), P(0,1), P(1,0), P(1,1). Marginal probability is the sum of all possible values of joint distributions. Using this logic, the joint distributions of the probability of acquiring two or more diseases can be calculated. Based on the definition of the copula function (2) and the formula for the joint density distributions for the variables $x_1, \ldots, x_n$ which is expressed as follows:

$$f(x_1, \ldots, x_n) = f(x_n)\, f(x_{n-1}|x_n)\, f(x_{n-2}|x_{n-1}, x_2) \ldots f(x_1|x_2,\ldots,x_n) \quad (10)$$

in the case of n = 3, the function will look as follows:

$$f(x_1|x_2, x_3) = c_{13|2}\,(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))\, f(x_1|x_2), \quad (11)$$

$$f(x_1|x_2, x_3) = c_{13|2}\,(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))\, c_{12}\,(F_1(x_1), F_2(x_2))\, f_1(x_1). \quad (12)$$

The symmetric Frank Archimedean function will be used as a copula in (12), and the numerical maximum likelihood estimation (MLE) method will be used to determine its parameters:

$$L(x,y,\Theta) = \Sigma \log c(x_i,y_i,\Theta), \text{ where } i \in [0,n]. \quad (13)$$

Healthcare data. The research is based on online available diabetic and heart disease datasets from UCI [20, 21]. The diabetic disease dataset is collected from 1999 to 2008. It includes 55 features. The heart disease dataset includes 76 features. The chosen variables are:

- Gender (1-male, 0-female); weight (BMI), age.
- The results of laboratory tests: total cholesterol (TCL), HAlc, Blood Pressure indicators, the amount of glucose in the blood.
- Number of diagnosis and diagnosis in ICD-9-CM format (presence of related diseases -1 for each existing disease and 0 otherwise).
- Harmful habits: smoking.

The following data preparations were completed:

- Insignificant missing data and outliers were ignored, and the rest were replaced by the average value or median.
- Duplicate data were merged.
- Patients with 3 or more related diseases were not included for secondary prevention.
- The data were normalized by the min-max normalization technique (Table 1), minimum and maximum value from data is fetched and each value is replaced according to the following formula:

$$\frac{x - \min(x)}{\max(x) - \min(x)}. \quad (14)$$

Table 1

The data AGE after normalization

AGE	min.max
44	0.421
41	0.342
65	0.974

Demographic, behavioral, biomedical factors were considered as risk factors.

Results. According to the presented three-stage method, the groups of patients with T2DM (type 2 diabetes mellitus), HD (hypertension disease) and CHF (congestive heart failure) as the main disease were distinguished in our dataset. Then, a pair-by-pair analysis of the dependence between these diseases was performed. Given

the risk factors, logistic regression models were constructed to calculate marginal distributions to estimate the risks of developing these related diseases in groups that do not have them (Fig. 4).

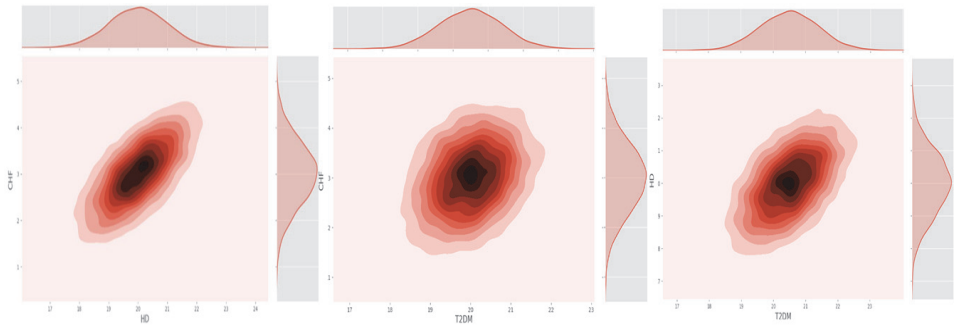

Fig.4. The joint and marginal distributions between CHF- HD, CHF-T2DM, HD-T2DM disease by sequence

Table 2

Disease dependence

Diseases	T2DM	HD	CHF
T2DM	1	1.57	1.33
HD	1.57	1	**1.92**
CHF	1.33	**1.92**	1

Based on the output of those models, the results were combined using the Frank copula function to capture their codependency. The results of the parameter Θ were calculated. The higher the positive value, the greater the occurrence of two diseases at the same time. The results of the Θ parameter, which can be seen in Table 2, are positive and show that the greatest dependence was the higher the positive value, the greater the occurrence of two recorded between HD and CHF diseases (Fig. 5).

Fig.5. The bivariate Frank copula for values of θ

Receiver Operating Characteristic (ROC) curves are plotted for each model (Fig. 6). The Area Under the Curve (AUC) (0.875, 0.863, 0.852) shows that the models discriminated very well.

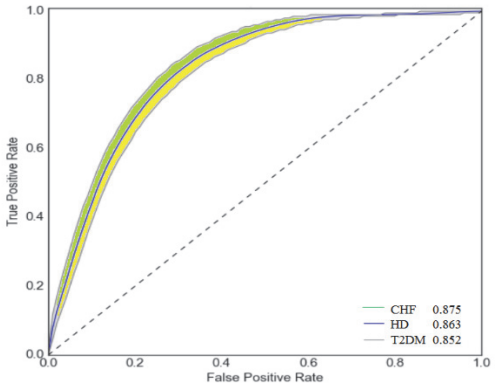

Fig.6. ROC curves and AUC for each model

Conclusion. The copula-based method evaluates the risks of developing complications of MS using Archimedean multivariate distribution functions. Because the copula function is independent of marginal distributions, it captures the dependence between random variables by connecting the marginal distributions with each other, and the "shape" of the resulting tail indicates the degree of dependence. By combining the marginal distributions with the joint distributions using the Frank copula function, the risks of developing the diseases with the highest degree of complexity were estimated for the disease-free groups. The presented method will be integrated in a multi-functional medical decision-making support system to automate risk assessment. The goal is to enable doctors to obtain data on the risks of developing related diseases.

REFERENCES

1. **Available**: https://www.iwh.on.ca/what-researchers-mean-by/primary-secondary-and-tertiary-prevention.
2. **Subbe C., Kruger M., Rutherford P., Gemmel I.** Validation of a modified early warning score in medical admission. - 2001. - Vol. 94, no. 10. - P. 521-526.
3. **Knaus W., Draper E., Wagner D., Zimmerman J.** APACHE II: as everity of disease classification system // Critical care medicine. - 1985. - 13(10). - P. 818-829.
4. Predictive risk algorithms in a population setting: an overview / **D. Manuel, L. Rosella, D. Hennessy, et al** // J Epidemiol Community Healt. - 2012. -66.-P. 859-865.
5. Rule Extraction from Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes / **L. Han, S. Luo, et al** // IEEE, J-BHI. - 2015. - 19(2). - P. 728-734.

6. A machine learning-based framework to identify type 2 diabetes through electronic health records / **T. Zheng, et al** // Elsevier, I JMI. - 2016.
7. **Available**: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3818341/
8. **Available**:http://mchp-appserv.cpe.umanitoba.ca/concept/Elixhauser%20Comorbidities%20-%20Coding%20Algorithms%20for%20ICD-9-CM%20and%20ICD-10.pdf#View=Fit.
9. **Rebekić A., Lonević Z., Petrović S., Marić S.** Pearson's or Spearman's Correlation Coefficient - Which One to Use? // Poljoprivreda (Osijek). -2015. - 21(2). - P. 47-54.
10. **Available**: https://github.com/sdv-dev/Copulas
11. **Available**:https://www.cdc.gov/nchs/icd/icd9cm.htm
12. **Available**:https://www.icd10data.com/ICD10CM/Codes/E00-E89/E08-E13/E11-
13. American Diabetes Association. **Available**: https://www.diabetes.org/
14. American Association of Clinical Endocrinologists and the American College of Endocrinology. **Available**: https://care.diabetesjournals.org/content/32/suppl_2/S151
15. Defining Comorbidity: Implications for Understanding Health and Health Services / **J. Valderas, et al** // The Annals of Family Medicine. - 2009.
16. **Nelson R.** An Introduction to Copulas.- Springer-Verlag, New York, Inc. - 1999.
17. **Agresti A., Kateri M**. Foundations of Statistics for Data Scientists: With R and Python. Python-Web-Appendix of Foundations of Statistics for Data Scientists // Chapman and Hall / CRC. - 2022. - P. 15-21.
18. **Genest C., Nikoloulopoulos A.K., Rivest L-P., Fortin M.** Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas // Brazilian J Probab Stat. - 2013. - 27(3). - P. 265-284.
19. **Available**:https://www.cantab.net/users/filimon/cursoFCDEF/will/logistic_reg.pdf.
20. **Available**:https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008.
21. **Available**: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

National Polytechnic University of Armenia. The material is received on 27.10.2022.

Լ.Կ. ԱՆԴՐԵԱՍՅԱՆ

ՄԵՏԱԲՈԼԻԿ ՀԱՄԱԽՏԱՆԻՇՈՎ ՀԻՎԱՆԴՈՒԹՅՈՒՆՆԵՐԻ ԲԱՐԴՈՒԹՅՈՒՆՆԵՐԻ ԶԱՐԳԱՑՄԱՆ ՌԻՍԿԻ ԿՈՒՄՈՒԼՅԱՏԻՎ ԲԱՇԽՄԱՆ ՖՈՒՆԿՑԻԱՅԻ ՀԵՆՔՈՎ ԳՆԱՀԱՏՄԱՆ ՄԵԹՈԴԸ

Արդիական է կանխել մետաբոլիկ համախտանիշով (ՄՀ) հիվանդությունների հետագա բարդությունների զարգացումը՝ գնահատելով հիվանդների էլեկտրոնային տվյալների հիման վրա հարակից հիվանդությունների ռիսկերը: Նկարագրվում է ՄՀ ունեցող հիվանդների ախտորոշումների և ռիսկի գործոնների հիման վրա բարդությունների ու հետհոսպիտալացման հետ կապված հարակից հիվանդությունների զարգացման ռիսկերի գնահատման Արքիմեդյան կոպուլահենքով մեթոդը:

Առանցքային բառեր. մետաբոլիկ համախտանիշ, էթիոլոգիական մոդել, լոգիստիկ ռեգրեսիա, Արքիմեդյան կոպուլա ֆունկցիաներ, կումուլյատիվ բաշխման ֆունկցիա:

Л.К. АНДРЕАСЯН

МЕТОД ОЦЕНКИ РИСКА РАЗВИТИЯ ОСЛОЖНЕНИЙ ЗАБОЛЕВАНИЙ С МЕТАБОЛИЧЕСКИМ СИНДРОМОМ НА ОСНОВЕ КУМУЛЯТИВНОЙ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Показана актуальность оценки рисков сопутствующих заболеваний на основании электронных данных пациентов с целью предупреждения прогрессирования дальнейших осложнений метаболического синдрома (МС). Описан метод оценки рисков развития сопутствующих заболеваний на основании диагнозов и факторов риска, связанных с осложнениями и повторными госпитализациями у больных МС на основе Архимедовы копулы.

Ключевые слова: метаболический синдром, этиологическая модель, логистическая регрессия, функции Архимедовы копулы, кумулятивная функция распределения.