Determining the Probability of Heart Disease using Data Mining Methods

Kseniia Bazilevych[0000-0001-5332-9545], Ievgen Meniailov[0000-0002-9440-8378], Kirill Fedulov[0000-0001-9619-0299], Sergey Goranina[0000-0001-8988-3935], Dmytro Chumachenko[0000-0003-2623-3294] and Pavlo Pyrohov[0000-0002-6100-4406]

National Aerospace University "Kharkiv Aviation Institute", Chkalova str., 17, Kharkiv, 61070, Ukraine
k.bazilevych@khai.edu, j.menyailov@khai.edu, fedulov.kirill172@gmail.com, sgoranin@gmail.com, dichumachenko@gmail.com

Abstract. The article suggests methods for estimating the parameters of logistic regression for different conditions. In the case of a single polytomic input variable with a minimum number of categories - a method for assessing chances and probabilities. In this case, the quality of classification can be evaluated separately for each input variable: the assessment does not depend on the connectedness of the input variables, which allows not to check the correlation and preliminary selection of significant variables. For several variables, it is proposed to use a Bayesian classifier, which, if there is no correlation between the attributes, assigns specific individuals of the population to a certain class for health reasons. If there is a correlation of factor attributes and complex dependencies between input variables, it is proposed to use the maximum likelihood estimation. As a result of the analysis, a ready-made mathematical apparatus will be obtained, which makes it possible in practice to obtain the values of the the probabilities of diseases under various initial data..

Keywords: Classification, Probability Assessment, Logistic Regression, Bayesian Classifier, Odds Assessment Method, Maximum Likelihood Method

1 Introduction

Diagnostic methods [1-2] in medicine play a crucial role. The accuracy of the diagnosis and the speed with which it can be made depends on many factors: the condition of the patient, the available data on the symptoms and signs of the disease, the results of laboratory tests, but most importantly, the qualifications of the doctor himself [3-6]. An accurately diagnosed diagnosis as soon as possible allows increasing the chance of curing the patient [7]. Based on all these considerations, it is natural to try to determine the conditions under which the diagnosis can be made as quickly and accurately as possible.

For many centuries, doctors have been trying to solve this problem with varying degrees of success. However, in recent years, thanks to the use of modern methods of

2019 IDDM Workshops.

treatment and diagnostics based on the latest achievements of science and technology, the chances of obtaining successful results have increased significantly. Therefore, it is important to find the exact methods [8-9] for description, research, evaluation and monitoring of the diagnosis process, which makes the task of determining the likelihood of disease based on existing data on the patient's condition relevant.

If the study is associated with a large number of interdependent factors that exhibit significant natural variability, then for a sufficiently effective description of the complex pattern of their influence, there is only one way - using the appropriate statistical method [10]. If there is a need to determine the probability of falling into one of two classes of the disease, one of the simplest and most effective methods is the binary classifier. The quality of the classification can be evaluated for each input variable separately. If the number of factors or the number of data categories is very large, it is necessary to use the computing power of the computer [11] so that the desired results can be obtained in a fairly short time, which will reduce the likelihood of errors in the diagnosis, and will also make it as quick and efficient as possible.

Thus, the aim of the study is to determine the likelihood of a patient's disease [12-13] with specified diagnostic characteristics based on Data Mining methods, which will improve the accuracy of diagnosis.

The health status of each individual is influenced by a number of factors, such as: age, gender, illness, place of residence, temperature, blood condition, etc. [14]. The objective of the study is to identify and analyze methods that allow us to assess the likelihood of illness [15] of a patient with specified diagnostic characteristics.

This task is referred to the classification tasks "with the teacher", during which the test system is trained using the "stimulus-reaction" examples. It is required to find dependency that shows which patients belong to the "Healthy" class and which patients belong to the "Sick" class. For such a task, it is rational to use logistic regression, which is widely used to find the probabilities of an event with given characteristics [16].

2 Estimation of logistic regression parameters based on the method of assessing chances and probabilities

Consider a sample of patients based on data from the source [17]. For each patient, health information is known. The explanatory variable in this case is the result of an electrocardiogram (ECG) at rest. This variable is polytomic. Each patient can belong to the three classes "Normal", "Hyp" and "Abnormal" according to ECG results. Two events are also considered: the patient is sick (y = 1) and healthy (y = 0).

It is necessary to evaluate the parameters of the logistic equation for this problem and determine the probability with which it will belong to the "Sick" class, i.e. evaluate his state of health. Probability that the output variable $y=1$ for the given value of the explanatory variable x will be $P(y=1|x)=\rho(x)$, and the probability that $y=0$ at a given value x will be equal to $P(y=0|x)=1-\rho(x)$.

The conditional average for logistic regression in this case is determined as in formula (1):

$$\rho(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{1}$$

where $g(x) = \beta_0 + \beta_1 C_1 + \beta_2 C_2$; C_1, C_2 is variables for quantizing values in three intervals; X is explanatory variable; $\beta_0, \beta_1, \beta_2$ is desired parameters; $c(x)$ is event probability.

The function is defined on an infinite interval and takes values in a range [0, 1]. Required to find the best estimates of parameters $\beta_0, \beta_1, \beta_2$. We will organize the information about patients based on the data [17] in the form of a Table 1.

Table 1. ECG Patient Data

Outcome	*Normal*	*Hyp*	*Abnormal*	*Total*
y = 0	96	68	1	165
y = 1	56	79	3	138
Total	152	147	4	303

In the Table 1 in the line with the "Normal" class, the quantization variables will be equal to: $C_1 = C_2 = 0$. In line with class "Hyp" $C_1 = 1, C_2 = 0$. In the line with the class "Abnormal" $C_1 = C_2 = 1$.

The chances of being a patient with a sick heart for all categories of conditions of the electrocardiogram are estimated by the formulas (2-4):

$$Ch_{y=1,C_1} = \frac{56}{96} \approx 0.58 \tag{2}$$

$$Ch_{y=1,C_2} = \frac{79}{68} \approx 1.16 \tag{3}$$

$$Ch_{y=1,C_3} = 3 \tag{4}$$

The odds ratio for the "Hyp" categories to the "Normal" category is estimated by the formula (5):

$$OR\left(\frac{C_2}{C_1}\right) = \frac{Ch_{y=1,C_2}}{Ch_{y=1,C_1}} = 2.01 \tag{5}$$

The odds ratio for the "Abnormal" to the "Normal" categories is estimated by the formula (6):

$$OR\left(\frac{C_3}{C_1}\right) = \frac{Ch_{y=1,C_3}}{Ch_{y=1,C_1}} = 5.17 \tag{6}$$

The experimental probability of a disease for the “Normal” category can be found (7) by dividing the number of positive outcomes by the total number of outcomes:

$$c_{exp} = 56 / 152 = 0.37 \tag{7}$$

From here the coefficient β_0 can be found as (8)

$$\beta_0 = \ln\left(\frac{c_{exp}}{1 - c_{exp}}\right) = -0.532 \tag{8}$$

For the “Hyp” category, the experimental probability of the disease can be estimated by the formula (9):

$$c_{exp} = 79 / 147 = 0.54 \tag{9}$$

From here, the coefficient β_1 can be found as (10):

$$\beta_1 = \ln\left(\frac{c_{exp}}{1 - c_{exp}}\right) - \beta_0 = 0.692 \tag{10}$$

For the category "Abnormal" the experimental probability of the disease can be estimated by the formula (11):

$$c_{exp} = 3 / 4 = 0.75 \tag{11}$$

From here the coefficient β_2 can be found as (12)

$$\beta_2 = \ln\left(\frac{c_{exp}}{1 - c_{exp}}\right) - \beta_0 - \beta_1 = 0.939 \tag{12}$$

The probability that the output variable y will be equal to one (that is, the patient will be ill) for the category "Normal" is calculated by the formula (13):

$$P(y = 1 | x) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} = \frac{e^{-0.532}}{1 + e^{-0.532}} \approx 0.37 \tag{13}$$

The probability that the output variable y = 1 for the “Hyp” category is calculated by the formula (14):

$$P(y = 1 | x) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} = \frac{e^{-0.532 + 0.692}}{1 + e^{-0.532 + 0.692}} \approx 0.54 \tag{14}$$

The probability that the output variable y = 1 for the "Abnormal" category is calculated by the formula (15):

$$P(y=1|x)=\frac{e^{\beta_0+\beta_1+\beta_2}}{1+e^{\beta_0+\beta_1+\beta_2}}=\frac{e^{-0.532+0.692+0.939}}{1+e^{-0.532+0.692+0.939}}\approx 0.75 \quad (15)$$

It can be concluded that if the result of the ECG is "Abnormal", then the probability of the disease is highest, if "Hyp", then less, and the probability of being healthy is highest if the result is "Normal".

3 Estimating the likelihood of a disease using a Bayesian classifier

Consider a sample of 30 patients with input variables defined in the nominal scale (Table 2) based on data from the source [17]. For analysis, we use the following signs: age (in years), blood sugar, patient gender, ECG result. According to the Table 2, the pair correlation coefficients were calculated, the values of which are in the interval [-0.303; 0.078], which indicates a low correlation between the input variables.

Table 2. Patient data for selected characteristics

№	*Age*	*Blood sugar < 120*	*Gender*	*ECG result*	*Disease state*
1	60-69	No	Man	Hyp	Yes
2	40-49	No	Man	Normal	No
3	60-69	No	Man	Hyp	No
4	60-69	No	Man	Normal	Yes
5	50-59	Yes	Man	Hyp	Yes
6	50-59	Yes	Woman	Normal	No
7	50-59	No	Man	Normal	Yes
8	50-59	No	Man	Normal	Yes
9	60-69	Yes	Man	Normal	No
10	60-69	No	Man	Hyp	Yes
11	60-69	No	Man	Hyp	Yes
12	30-39	No	Man	Normal	No
13	40-49	No	Woman	Normal	No
14	50-59	No	Man	Normal	No
15	60-69	No	Woman	Normal	Yes
16	50-59	No	Woman	Hyp	No
17	50-59	No	Man	Normal	No
18	50-59	No	Woman	Normal	No
19	50-59	Yes	Man	Hyp	Yes

20	40-49	No	Man	No	No
21	50-59	No	Woman	No	No
22	50-59	No	Woman	Normal	No
23	60-69	No	Woman	Normal	No
24	40-49	No	Man	Normal	No
25	40-49	No	Man	Hyp	Yes
26	60-69	No	Woman	Normal	No
27	60-69	Yes	Man	Hyp	Yes
28	60-69	No	Man	Normal	Yes
29	50-59	No	Man	Normal	No
30	40-49	No	Man	Hyp	No

Thus, in this case, you can use the Bayesian classifier, the application of which for this case is considered in detail in [18].

We denote by C1 the class “Sick” for whom the state of the disease is present (the value of the resulting variable is “yes”). Through C2, we can designate the class of patients “Healthy”, which have no signs of illness (the value of the resulting variable is “no”). The use of the Bayesian classifier does not make it possible to obtain the form of a statistical dependence based on the training sample, however, it makes it possible to determine the probability that a patient with given characteristics will fall into one or another class. For example, we define that a patient aged 50 to 59 years, with blood sugar less than 120 units, a man and with the result of ECG “Hyp” will fall into the class “Sick”.

It is necessary to maximize the product of probabilities $P(X \mid C_k)P(C_k)$ for $k = 2$, because there are only two classes in this problem. The prior probability of the appearance of class C1 is calculated by the formula (16):

$$P(C_1) = \frac{12}{30} = 0.4 \tag{16}$$

The prior probability of the appearance of a class C2 is calculated by the formula (17):

$$P(C_2) = \frac{18}{30} = 0.6 \tag{17}$$

There are 30 observed examples, 18 of them are “Healthy”, 12 are “Sick”. Conditional probabilities for determining $P(X \mid C_k)$ calculated in Table 3. Calculate the generalized probabilities $P(X \mid C_k)$ for events of the formula (18-19):

$$P(X \mid C_1) = 0.33 \bullet 0.25 \bullet 0.92 \bullet 0.58 = 0.044 \tag{18}$$

$$P(X \mid C_2) = 0.44 \bullet 0.11 \bullet 0.55 \bullet 0.17 = 0.0045 \tag{19}$$

Than probabilities $P(X \mid C_k)P(C_k)$ will be respectively equal (20-21):

$$P(X \mid C_1)P(C_1) = 0.044 \bullet 0.6 = 0.0264 \tag{20}$$

$$P(X \mid C_2)P(C_2) = 0.0045 \bullet 0.4 = 0.0018 \tag{21}$$

Table 3. Conditional Probabilities for Patient Data

Probability description	*Estimation*
P(Age 50–59|C_2)	8/18=0.44
P(Age 50–59|C_1)	4/12=0.33
P(Blood Sugar<120|C_2)	2/18=0.11
P(Blood Sugar<120|C_1)	3/12=0.25
P(Man|C_2)	10/18=0.55
P(Man|C_1)	11/12=0.92
P(Hyp|C2)	3/18=0.17
P(Hyp|C_1)	7/12=0.58

The class whose probability is greater is selected, i.e. the patient in question belongs to the class “Sick”.

The normalization of probabilities is as follows of formula (22-23):

$$P'(X \mid C_1)P(C_1) = \frac{0.0264}{0.0264 + 0.0018} = 0.94 \tag{22}$$

$$P'(X \mid C_2)P(C_2) = \frac{0.0018}{0.0018 + 0.0264} = 0.06 \tag{23}$$

Thus, a patient with the described characteristics will be sick with a probability of 0.94 (will fall into the “Sick” class), and with a probability of 0.06 will be healthy (will fall into the “Healthy” class).

4 Estimation of logistic regression parameters based on the maximum likelihood estimation

Consider a sample of 303 patients with input characteristics shown in Table 4 based on data from the source [17]. The resulting trait is measured in a dichotomous scale, and factor traits in metric and other types of scales. It is necessary to determine the likelihood of a patient's disease with this many characteristics.

Since the maximum likelihood estimation (MLE) is quite resource-intensive, we will use the software from IBM – SPSS Statistics for the demonstration. This software allows us not only to find the parameters of logistic regression, but also to evaluate the parameters of the model and probability, and also analyze the quality of the model.

Table 4. Patient sampling data

age	metric scale	29…77
sex	nominal scale	Male/Female
chest pain type	nominal scale	Asymptomatic, Abnormal Angina, Angina, No-Tang
blood pressure	metric scale	94…200
cholesterol	metric scale	126…564
fasting blood sugar <120	nominal scale	true/false
resting ecg	nominal scale	Normal/Hyp
maximum heart rate	metric scale	71…202
angina	nominal scale	true/false
peak	metric scale	0…6,2
slope	nominal scale	Flat, Down, Up
#colored vessels	metric scale	0,1,2,3
thal	nominal scale	Normal, Rev, Fix
class*	nominal scale	Sick, Healthy

**Row "Class" is necessary for analysis of simulation results performing.*

The most significant results are visible in the tables below. In the Table 5 presents the quality factors of the model.

Table 5. Model Summary Table

Step	*-2 Log probability*	*R-squared Cox & Snell*	*R-squared Nagelkerke*
1	348.461	0.210	0.281

Criterion -2 Log probability corresponds to the correspondence between the models and the source data. The smaller this indicator, the more adequate the model.

R-squared Cox & Snell and R-squared of the Nagelkerke are stably statistically consistent, which are used in the logit. The value of an equal object is achievable. In the second sign, this drawback is eliminated. These criteria shows the share of all factor characteristics. More detailed information can be taken from the source [17]. In the table 6 presents the values of the Chi-square test.

Table 6. Universal criterion for model coefficients

	Step	*Chi-squared*	*Degrees of freedom*	*Relevance*
1	Step	71,586	3	,000
	Block	71,586	3	,000
	Model	71,586	3	,000

Table 7-8 presents the Hosmer-Lemeshov criterion. In our case, part of the variance is 0.7%. This indicates a high degree of consistency in the model.

The Hosmer-Lemeshov criterion - shows an assessment of the agreement between the frequencies in the sample and the model [19-20]. It shows whether there is "garbage", which leads to a decrease in the quality of the model in the model.

Table 7. Hosmer-Lemeshov criteria

Step	*Chi-squared*	*Degrees of freedom*	*Relevance*
1	9.800	2	0.007

Table 8. Conjugation table for checking Hosmer-Lemeshov consent

		Illness = yes		Illness = no		
		Observed	Expected	Observed	Expected	Overall
Step 1	1	23	22,839	0	0,161	23
	2	27	29,294	3	0,706	30
	3	29	28,613	1	1,387	30
	4	27	26,486	3	3,514	30
	5	27	23,015	3	6,985	30
	6	16	17,586	14	12,414	30
	7	10	11,150	20	18,850	30
	8	5	4,997	24	24,003	29
	9	0	1,847	30	28,153	30
	10	1	0,495	40	40,505	41

Table 9 shows percentages representing different levels of classification of the model. Quite high indicators were obtained, i.e. 92.2% of cases were classified correctly.

Table 9. Table classification

	Observed		Predicted		
			Class		Correctness
			Healthy	Sick	
Step 1	Class	Healthy	147	18	93.1
		Sick	23	115	91.3
	Total percentage				92.2

Table 10 shows the parameters of the logistic regression equation.

Table 10. Parameters of logistic regression equation

Influencing variable	*Regression equation cofficients β*	*Stand ard error*	*Wald statistics*	*Signifi -cance level*
A – sex(1)	-1,464	0,490	8,932	0,003
B - chestpaintype			30,864	0,000
B1 - chestpaintype(1)	2,286	0,444	26,474	0,000
B2 - chestpaintype(2)	0,971	0,590	2,705	0,100
B3 - chestpaintype(3)	0,170	0,652	0,068	0,794
C - angina(1)	-0,763	0,380	4,044	0,044
D - slope			24,588	0,000
D1 - slope(1)	1,724	0,700	6,068	0,014
D2 - slope(2)	2,018	0,415	23,700	0,000
E - @#coloredvessels			36,481	0,000
E1 - @#coloredvessels(1)	-1,763	0,505	12,204	0,000
E2 - @#coloredvessels(2)	0,495	0,547	0,818	0,366
E3 - @#coloredvessels(3)	1,498	0,793	3,566	0,059
F - thal			14,056	0,001
F1 - thal(1)	-1,492	0,733	4,137	0,042
F2 - thal(2)	-1,452	0,411	12,472	0,000

The remaining variables were excluded from the formula due to data redundancy.

Based on this table, you can determine the most significant factors by which you can get the smallest errors with a high probability. The general form of the regression equation for the patient will have the form similar to formula (24):

$$\begin{aligned} g(x) = &-1,464 \bullet \mathrm{A} + 2,286 \bullet \mathrm{B1} + 0,971 \bullet \mathrm{B2} + \\ &+0,170 \bullet \mathrm{B3} - 0,763 \bullet \mathrm{C} + 1,724 \bullet \mathrm{D1} + 2,018 \bullet \mathrm{D2} - \\ &-1,763 \bullet \mathrm{E1} + 0,495 \bullet \mathrm{E2} + 1,498 \bullet \mathrm{E3} - 1,492 \bullet \mathrm{F1} - 1,452 \bullet \mathrm{F2} \end{aligned} \quad (24)$$

Then, for a male patient with a second type of chest pain that did not have a sore throat, with a bias of the first type, with vessels of the third type, as well as a thal of the first type, it will be true (25):

$$g(x) = -1,464 + 0,971 + 1,724 + 1,498 - 1,492 = 1,237 \quad (25)$$

Then the probability that such a patient will be healthy is calculated by the formula (26):

$$\rho(x) = e^{g(x)} / \left(1 + e^{g(x)}\right) \approx 0,77 \quad (26)$$

Moreover, as can be seen from the Table 10, according to Wald's statistics, the most significant are the following factors: chestpaintype (value 30.864), slope (value 24.588), coloredvessels (value 36.481).

Wald test – a statistical test used to check the restrictions on the parameters of statistical models estimated on the basis of sample data. It is the most appropriate of the three basic constraint checking tests such as the likelihood ratio test and the Lagrange multiplier test. The test is asymptotic, that is, a sufficiently large sample size is required for the reliability of the conclusions. The confidence interval (CI) of the test is also a closed form. The higher the statistics, the better.

The significance of the factors is confirmed using the appropriate level of significance. It is defined as the p-level, which is calculated during the test. The lower this level, the better.

Based on the data in Table 4, the probabilities of getting into the Healthy group were calculated for all data.

In the “Expected” and “Group” columns, you can see the probabilities of getting into the “Healthy” or “Sick” group.

The simulation results (Table 8) show high accuracy of the classification results in comparison with the classes previously known for the experimental sample (Table 4).

Based on the results, we can say that the model adequately describes this population.

5 Conclusions

In this work we identified, analyzed and implemented methods that allow us to assess the likelihood of a patient's disease with specified diagnostic characteristics.

It is shown in which cases it is advisable to use certain methods to determine the probability and estimate the parameters of models. These models are not static. Calculation of parameters can be carried out every time when the amount of data about patients changes, and the use of SPSS software tools will allow calculations to be made quite quickly. The data obtained will allow a more accurate assessment of the state of health in the face of constantly changing diagnostic parameters.

References

1. Baldi, P., Brunak, S.: Bioinformatics: The Machine Learning Approach (2nd ed.). MIT Press, 400 p. (2001).
2. Meniailov I., et. al.: Using the K-means Method for Diagnosing Cancer Stage Using the Pandas Library. In CEUR Workshop Proceedings, vol. 2386, pp. 107-116 (2019).
3. Chumachenko, D.: On Intelligent Multiagent Approach to Viral Hepatitis B Epidemic Processes Simulation, in Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018, pp. 415-419 (2018).

4. Chumachenko, D., Chumachenko, K., Yakovlev, S.: Intelligent simulation of network worm propagation using the code red as an example. In Telecommunications and Radio Engineering, vol. 78, iss. 5, pp. 443-463. (2019).
5. Polyvianna, Yu., Chumachenko, D., Chumachenko T.: Computer Aided System of Time Series Analysis Methods for Forecasting the Epidemics Outbreaks, 2019 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), pp. 1-4 (2019).
6. Chumachenko, D., Chumachenko, T.: Intelligent Agent-Based Simulation of HIV Epidemic Process. In Advances in Intelligent Systems and Computing, vol. 1020, pp. 175-188 (2019).
7. Berry, M.W.: Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 244 p. (2003).
8. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations, In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 281-297 (1967).
9. Deshpande, M., Kuramochi, M., Karypis, G.: Automated approaches for classifying structures. In Proc. 2002 Workshop on Data Mining in Bioinformatics (BIOKDD'02), Canada, 11 – 18 (2002).
10. Frakes, W., Baeza-Yates, R.: Information Retrieval: Data Structures and Algorithms (1992).
11. Bazilevych, K. et al.: Stochastic modelling of cash flow for personal insurance fund using the cloud data storage. In: International Journal of Computing, Vol. 17, Iss. 3, pp. 153-162 (2018).
12. Cancer control: early detection. WHO Guide for effective programmes. Geneva: World Health Organization; 2007, http://apps.who.int/iris/bitstream/10665/43743/1/9241547338_eng.pdf, last accessed 2019/10/28
13. Rubin, G, et. al.: The expanding role of primary care in cancer control Lancet Oncol, pp. 31 – 72 (2015).
14. Chumachenko, D., et. al.: Intelligent Expert System of Knowledge Examination of Medical Staff Regarding Infections Associated with the Provision of Medical Care, in CEUR Workshop Proceedings, vol. 2386, pp. 321-330 (2019).
15. Bowers, N.L., et. al.: Actuarial mathematic, Illinois, USA by Society Of Actuaries, 621 p. (1997).
16. Norman, T. J.: The mathematical approach to biology and medicine norman, Wiley, 296 p. (1967).
17. Sample Dataset, https://github.com/DorianDrain/Excel-Data-Sets/tree/master, last accessed 2019/10/28.
18. Cox, D.R., Snell, E.J.: Analysis of Binary Data, Chapman and Hall, CRC, 240 p. (1989).
19. Hosmer-Lemeshow Test, http://www.real-statistics.com/logistic-regression/hosmer-lemeshow-test/, last accessed 2019/10/28.
20. Bartlett, J.: The Hosmer-Lemeshow goodness of fit test for logistic regression. In The Stats Geek (2014).