The Use of Data Mining Techniques for Analysis of Menstrual Cycle Parameters and Prognosis of Migraine Symptoms in Reproductive Age Women

Larysa Malanchuk
Department of Obstetrics and Gynaecology No.1
I. Horbachevsky Ternopil National Medical University
Ternopil, Ukraine
malanchuk@tdmu.edu.ua

Mariia Riabokon
Faculty of Medicine
I. Horbachevsky Ternopil National Medical University
Ternopil, Ukraine
ryabokon_maol@tdmu.edu.ua

Artem Malanchuk
Faculty of Medicine
I. Horbachevsky Ternopil National Medical University
Ternopil, Ukraine
malanchuk_arse@tdmu.edu.ua

Serhiy Malanchuk
Department of Pharmacology and Clinical Pharmacology
I. Horbachevsky Ternopil National Medical University
Ternopil, Ukraine
malanchuk_s@tdmu.edu.ua

Svitlana Riabokon
Department of Primary Health Care and General Practice of Family Medicine
I. Horbachevsky Ternopil National Medical University
Ternopil, Ukraine
ryabokon@tdmu.edu.ua

Olha Kovalchuk
Department of Applied Mathematics
Ternopil National Economic University
Ternopil, UKRAINE
olhakov@gmail.com

Abstract: **The article represents the results of using the data mining algorithms to determine probability of occurrence migraine symptoms related to changes of menstrual cycle parameters in women of reproductive age. The analysis was carried out on the basis of logistic regression models, developed by three different methods. Relations between menstrual amount changes, regularity cycle and onset of migraine symptoms were determined. Sensitivity and specificity of developed model was counted. Its quality assessment was appraised by means of ROC (Receiver Operating Characteristics) analysis.**

Keywords: menstrual cycle parameters, symptoms of migraine, women of reproductive age, data mining, logistic regression model, ROC-analysis.

I. Introduction

Menstrual cycle (MC) is a complex of cyclic changes in a woman's body implemented on the level of target organs and is under the supervision of extrareproductive (hypothalamus, anterior pituitary, thyroid, adrenal gland) and reproductive (ovaries) hormone-producting organs. Actually a vivid external sign of MC is menstruation, in the assessment matters is its regularity, duration, volume of blood loss considering occurrence of blood clots, change of state of health during premenstrual period and during menstruation, occurrence of pathological symptoms in target to steroids organs (including the brain) [1]. In addition, it is assessment duration of menstrual cycle, occurrence intermenstrual blood excretion and premenstrual symptoms [2].

According to present-day concepts a monthly physiological process of endometrium regeneration considered from an immunological point as inflammation, to respond fall of estradiol and progesterone level in premenstrual period. Degranulation of the mast cells and other immunocompetent cells in different organs, where local inflammation occurs in response to genetic or acquired vulnerabilities, play a major role in this process [3]. Particularly it is accomplished at the brain level.

Headache, which has been disturbed for the last 3 months or more, has following characteristics: duration more than 4 hours, unilateral (one-sided) localization, throbbing character, medium or high intensity, nausea and vomiting or intolerance of light or sounds, it is treated as a symptom of migraines [4].

Absence or insignificance of local and system signs of inflammation associated with menstruation, probably subject to a time limit and amount of menstruation, also to adequate recovery histological and functional characteristics of involved tissues [5].

So, the purpose of our study was evaluation of menstrual function parameters deviation (changes) and manifestation identify their possible association with the risk of appearance of symptoms of migraine in women of reproductive age using the data mining models.

II. Research Methodology

Analyzed dataset as many others applied medical researches includes binary signs. Logistic regression analysis is used in studies of variable dependence from one or several independent variables of arbitrary type. ROC-analysis is used in determination of classification problem based on logit regression. Logistic regression and ROC-analysis are the part of algorithm sets of data mining.

Logistic regression expresses data mining in the form of dependence $P\{Y=1 \mid X\} = f(X)$, to predict event rate $\{Y=1\}$, specified by values of independent variables X_1, X_2, ..., X_k. The task of logistic regression is to get prognosis of continuous variable with values on the segment [0, 1] at any values of independent variables instead of value prediction of binary variable.

Logistic regression expresses connection between the result and factors in the form of dependence [6], [7]

$$P\{Y=1 \mid X_1, X_2, ..., X_k\} = \frac{e^{\hat{Y}}}{1+e^{\hat{Y}}} = \frac{1}{1+e^{-\hat{Y}}}, \quad (1)$$

where $P\{Y=1 \mid X_1, X_2, ..., X_k\}$ – probability of event rate;

e = 2,718... – the base of natural logarithm;

$\hat{Y} = a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k$ – linear equation.

Logistic regression is based on logistic distribution function

$$F(x) = \frac{e^x}{1+e^x}, \qquad (2)$$

The model presented by this type of regression is in effect (per se) the distribution of this law; linear combination of independent variables is the function argument.

Ratio of event probability to incredibility, $\frac{P}{1-P}$ is called odds ratio (OD).

One more presentation of logistic regression is connected with this ratio. Solving (2) regarding $\hat{Y}$, is $\hat{Y} = \ln\left(\frac{P}{1-P}\right)$, where $P\{Y = 1 | X_1, X_2, ..., X_k\}$. Then odds ratio is written as follows

$$\frac{P}{1-P} = e^{a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k} = e^{a_0}\left(e^{a_1}\right)^{X_1}\left(e^{a_2}\right)^{X_2} ... \left(e^{a_k}\right)^{X_k}. \qquad (3)$$

Thus, if the model is correct in independent variables $X_1, X_2, ..., X_k$ variable X_j per unit will cause odds ratio change in y e^{a_j} times.

If we choose instead of distribution function $F(Y) = \frac{e^{\hat{Y}}}{1+e^{-\hat{Y}}}$ the normal law distribution function $F(Y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Y} e^{\frac{-x}{2}} dx$, then instead of logit-model we'll receive similar to it, probit-model.

Probit- and logit-models are used in binary value prognosis for qualitative variables assessment, where linear estimation is complicated by a variety of reasons. To determine in what cases logit-model should be used and when probit-model should be used at small samples is impossible because estimated coefficients of the model

$$P\{Y = 1 | X_1, X_2, ..., X_k\} = \frac{e^{\hat{Y}}}{1+e^{\hat{Y}}} = \frac{1}{1+e^{-\hat{Y}}}, \qquad (4)$$

practically differ only in constant factor.

There are several methods of coefficients evaluation of logistic regression. Maximum-likelihood procedure (method) is often used in practice. It is used in mathematical statistics for getting universe parameters estimation based on sampled data. The base of this method is likelihood function expressing probability of general occurrence of sampling results $Y_1, Y_2, ..., Y_k$:

$$L(Y_1, Y_2, ..., Y_k; \theta) = p(Y_1; \theta) \cdots p(Y_k; \theta). \qquad (5)$$

According to maximum-likelihood procedure (method) as the estimator of unknown parameter θ possess the value $\Theta = \Theta(Y_1, Y_2, ..., Y_k)$, that maximizes L function.

Multiple logit or probit-analysis is a natural continuation of binary and arises from the choice of more than two alternatives.

III. Data Selection and Description

This article uses the unique set of data based on questionnaire survey results of reproductive age women in Ternopil region (Ukraine) held during 2019. The questionnaire included some questions blocks. The first concerned the evaluation of menstrual function (MF) parameters: MF regularity and recurrence (cyclicity); menstrual duration and volume with large blood clots, inter-menstrual discharges, overall health and working capacity changes in premenstrual and menstrual periods. Pictogram was suggested for assessment of blood loss volume (Fig. 1).

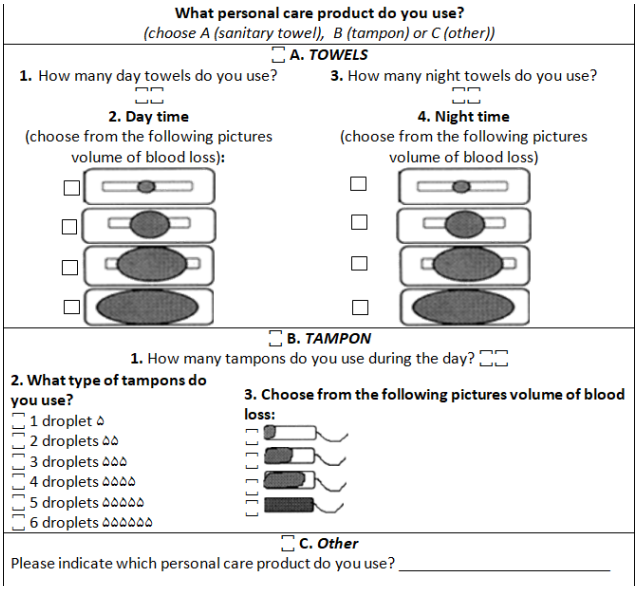

Fig. 1. Pictogram for assessment of blood loss volume

The second block of questions concerned migraine symptoms occurrence and their evaluation in the respondents, according to the recommendations of International Headache Society.

Analyzed dataset is prepared on the base of 1015 respondents personal data. Interviewees findings were analyzed using data mining algorithms to determine indices of parameter changes of MF, that are sensitive to migraine symptoms occurrence.

Such variables were used in empirical study: *MCD* (menstrual cycle duration): 0 – norm (24-38 days), 1 – short-cut cycle (less than 24 days), 2 – extended (more than 38days); *MD* (menstruation duration): 0 –norm (4-8days), 1 – short-term(less than 4 days), 2 – prolonged (more than 8 days); *LBV* (loss of blood volume): 0 – norm (up 80 ml), 1 – slight deviations (80–150 ml), 2 – significant deviations (more than 150 ml); *OBC* (occurrence of blood clots): 1 – yes, 0 – no; *RCMC* (regularity and cyclicity of menstrual

cycle): 1 – yes, 0 – no; *SM* (symptoms of migraine): 1 – yes, 0 – no. *SM* is dependent variable, other – independent.

IV. Empiracal Results and Discussion

We have formulated an equation of logistic regression for dependent variable *SM* by Forward Stepwise method.This is step-by-step method of variables sampling where inclusion check is based on the significance of Lagrange multiplier criterion.

On zero step the model is not built, all "predictable" values *Y* equal one (migraine syndromes are available), therefore all observations in $Y = 1$, are true "predictable", but observations, in $Y = 0$ (syndromes of migraine are absent) are false (Fig. 2).

Classification Table[a,b]

Observed		Predicted		
		SM		Percentage Correct
		0	1	
Step 0 SM	0	606	0	100,0
	1	409	0	,0
Overall Percentage				59,7

a. Constant is included in the model.
b. The cut value is ,500

Fig. 2. Initial step of binary logistic model realization

On the subsequent steps logistic model will provide odds and probability occurrence of unrealized projects based on independent variables of regression equation.

Approximation of regression model quality is assessed by means of likelihood function. Negative duplicate logarithm value of this function -2log is the measure of likelihood (Fig.3). The less is this value, the better is binary logistic model built.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1308,952[a]	,057	,077
2	1298,954[a]	,066	,090
3	1293,623[a]	,071	,096

Fig. 3. Model summary for binary logistic model

Obtained score for regression model, that includes only constants is used as the initial value for -2log. After summation of all significant variables the value of -2log equals 1293,623 and it is less than initial. The value decrease means model improvement. Difference is denoted as Chi-square value and is significant (Fig. 4).

It means that initial model after summation of powerful variables has become considerably better.

Cox & Snell R Square and Nagelkerke R Square (Fig. 2) are determination pseudo-coefficients obtained on the basis of functions ratio of models likelihood only with constant and all coefficients.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	59,658	1	,000
	Block	59,658	1	,000
	Model	59,658	1	,000
Step 2	Step	9,998	1	,002
	Block	69,656	2	,000
	Model	69,656	2	,000
Step 3	Step	5,331	1	,021
	Block	74,987	3	,000
	Model	74,987	3	,000

Fig. 4. Values of binary logistic model

We are going to prognosticate on the basis of logistic regression model – will the event take place or won't {*Y*–1}. The law of prognosis that on default is used in logistic regression procedure: if the predicted probability of the event is more than 0,5 – the event will take place; if it is less or equals 0,5 – the event won't take place.

The number of true and false predicted results in the category of analyzed variable and general percentage of correct prognosis were determined according to the classification table (Fig. 5).

Classification Table[a]

Observed		Predicted		
		SM_		Percentage Correct
		0	1	
Step 3 SM_	0	227	379	37,5
	1	109	300	73,3

a. The cut value is ,500

Fig. 5 Classification Table of Binary Logistic Model

Variables analysis of regression model obtained by Forward Stepwise method (Fig. 6).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	OBC	,520	,132	15,429	1	,000	1,682
	Constant	-,590	,082	51,233	1	,000	,554
Step 2[b]	LBV	,269	,089	9,093	1	,003	1,309
	OBC	,456	,135	11,478	1	,001	1,577
	Constant	-,714	,093	58,977	1	,000	,490
Step 3[c]	LBV	,261	,090	8,494	1	,004	1,298
	OBC	,433	,135	10,248	1	,001	1,542
	RCMC	,303	,150	4,101	1	,043	1,354
	Constant	-,777	,099	62,176	1	,000	,460

a. Variable(s) entered on step 1: OBC.
b. Variable(s) entered on step 2: LBV.
c. Variable(s) entered on step 3: RCMC.

Fig. 6 Model summary for binary logistic model

Note: B – coefficients a_j of regression equation;

S.E. – measure of coefficients a_j variability;

Wald – Wald test for regression coefficients. The higher its value is (together with the number of free degrees), the more is coefficient regression;

df – the number of free degrees;

Sig – significance using Wald test (H_0: $a_j = 0$);

Exp(B) – odds ratio e^{a_j} (are used for results interpretation).

Observed coefficients significance is calculated on the basis of Wald statistics. Its universality enables to assess significance not only independent variables but categorical variables in general, despite that they are deaggregated on index variables. Wald statistics has distribution χ^2. The number of freedom degrees equals one if hypothesis about coefficient's equal-zero indicator is checked in common or index variable and for categorical variable it equals to the number of values minus one (the number of proper index variables). Square root from Wald statistic equals approximately the ratio of coefficient value to its standard error(as for t-statistic in simple linear model of regression).

In coefficient table (Fig. 5.) *LBV* (loss of blood volume), *OBC* (occurrence of blood clots), *RCMC* (regularity and cyclicity of menstrual cycle) are significant variables. Other characteristics were not included in equation.

Subsequent equation of regression logistic model was worked out:

$$\hat{Y} = -0{,}777 + 0{,}261 \cdot LBV + 0{,}433 \cdot OBC + 0{,}303 \cdot RCMC \,. \quad (5)$$

If substitute appropriate values of independent variables in this equation, the result will be logarithm of odds ratio (OR) of migraine symptoms onset. To determine these odds it is necessary to raise number *e* (logarithm to the base e) to the power. It is necessary to calculate likelihood of migraine symptoms occurrence based on information about loss of blood volume, occurrence of blood clots, regularity and cyclicity of menstrual cycle, using the ratio:

$$OR = \frac{P}{1-P}, \quad (6)$$

where *P* is the likelihood of migraine symptoms occurrence;

We solve this equation concerning *P*:

$$P = \frac{OR}{1+OR}. \quad (7)$$

We used this equation to predict odds and likelihood of migraine symptoms occurrence in female-patient with normal loss of blood volume (*LBV* = 0), without occurrence of blood clots (*OBC* = 0) and regularity and cyclicity of menstrual cycle with no deviations (*RCMC* = 0):

$$Log(OR_{sm}) = \hat{Y} = -0{,}777 + 0{,}261 \cdot 0 + 0{,}433 \cdot 0 + 0{,}303 \cdot 0 = -0{,}777; \quad (8)$$

$$OR_{sm} = \exp(-0{,}777) = 0{,}459, \quad (9)$$

consistent likelihood of risk onset of migraine symptoms:

$$P_{sm} = \frac{0{,}459}{1+0{,}459} = 0{,}315. \quad (10)$$

Similar we calculate odds and likelihood of migraine symptoms occurrence in the patient with slight deviations of loss blood volume (*LBV* = 1):

$$Log(OR^1_{sm}) = -0{,}777 + 0{,}261 \cdot 1 + 0{,}433 \cdot 0 + 0{,}303 \cdot 0 = -0{,}516; \quad (11)$$

$$OR^1_{sm} = \exp(-0{,}516) = 0{,}596; \quad (12)$$

$$P^1_{sm} = \frac{0{,}596}{1+0{,}596} = 0{,}373 \quad (13)$$

and in the patient with significant deviations of loss blood volume (*LBV* =2):

$$Log(OR^2_{sm}) = -0{,}777 + 0{,}261 \cdot 2 + 0{,}433 \cdot 0 + 0{,}303 \cdot 0 = -0{,}255; \quad (14)$$

$$OR^2_{sm} = \exp(-0{,}255) = 0{,}755; \quad (15)$$

$$P^2_{sm} = \frac{0{,}755}{1+755} = 0{,}430. \quad (16)$$

Therefore, odds of migraine symptoms occurrence in the patient with slight deviations of loss blood volume increase from 0,32 to 0,37 (almost on 6%), and for the patient with significant deviations of loss blood volume increase from 0,32 to 0,43 (almost on 12%). Present views on menstrual inflammation pathogenesis are: the increase of blood loss volume and cycle irregularity cause inadequate endometrium reproduction and the development of local inflammation in target organs (brain particularly). X_j variation on one unit causes odds ratio variation in e^{a_j} times.

Obtained results by Forward Stepwise method were checked by Enter method (selection procedure when all changes are introduced in one step) and Backward Stepwise method (step-by-step selection and exception procedure based on probability of likelihood ratio usage). Similar models were obtained using these three methods. All variables in the equation are significant.

The procedure of ROC-analysis is used to set out the results of binary classification and the assessment of scheme classification effectiveness.

Classification table (Fig. 4) is based on model classification results (cut-off threshold equals 0,5) and different types of observations appliance (objective)- those possessing migraine syndrome (positive) and no possessing ones (negative).

Objective value of each binary classificatory is determined by sensitivity and specificity of the model.

Sensitivity (*Se*) measures true positives rate, that are identified correctly (in proposed model correctly identified patients percentage, possessing migraine symptoms):

$$Se = \frac{TP}{TP + FN} 100\% = \frac{300}{300 + 109} 100 = 73{,}3\% , \quad (17)$$

where *TP* – true positives cases, *FN* – false negative cases.

Specificity (*Sp*) – true negative rate, identified correctly by the model:

$$Sp = \frac{TN}{TN + FP} 100\% = \frac{227}{227 + 379} 100 = 37{,}5\% , \quad (18)$$

where *TN* – true negative cases, *FP* – false positives cases.

The developed model has high sensitivity and presents true result in prediction of positive cases – it gives risk prognosis of migraine symptoms occurrence that is based analyzed aspects of values. But it has low specificity and does not present reliable result in negative cases prognosis – migraine symptoms are absent. Such model is quite applicable for diagnostic of migraine symptoms occurrence in women of reproductive age. When the answer is negative about migraine symptoms even at the moment of inquiry, abnormalities in menstrual cycle parameters in the course of time (average duration of reproductive age is 27 years) may cause the development of pathological headache. Taking into account this fact, preventive measures are quite justified.

Quality assessment of the developed model was studied by surface estimation under ROC-curves. Theoretically the surface ranges from 0 to 1, but as the model is always characterized by the curve, located above the main diagonal, so it is said that changes are from 0,5 ("helpless classifier") to 1 ("ideal model").We can obtain this assessment directly by calculating polygon surface under experimentally received ROC-curve (Fig. 7).

Area Under the Curve

Test Result Variable(s):Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,655	,017	,000	,621	,688

Fig. 7. Surface assessment under ROC-curve

AUC is called surface numerical index. It is considered, the more AUC is the greater prognosis power the model has. But AUC is most likely intended for comparative analysis of several models and does not comprise any information about model sensitivity and specificity. The expert scale of AUC [8] values, found in publications, we can conclude about the model quality – the developed model is of an average value (AUC = 0.65). Ideal model has 100% of sensitivity and specificity. But it is impossible to achieve such result in practice: to increase sensitivity and specificity of the model simultaneously.

CONCLUSION

We have discussed some aspects of parameters assessment of menstrual cycle. They significantly affect probability of migraine symptoms occurrence. The results of study have proved the association of pathologic headache development with blood loss volume including large blood clots, regularity and menstrual cycle cyclicity.

The developed model enables to predict likelihood of migraine symptoms occurrence individually for each patient based on her values of MC parameters. The application of regression models elaborated on data mining algorithms making use of information about changes of MC parameters is of great practical importance for decision making in prophylaxis and timely warning to probable risk development of migraine headache which affects life quality of a woman. It assists in solving medical, social and economic aspects of this problem.

REFERENCES

[1] V. T. Martin, and R. B. Lipton, "Epidemiology and biology of menstrual migraine," *Headache*, vol. 48, no. 3, pp. 124-130, 2008.

[2] M. G. Munro, et al. "The two FIGO systems for normal and abnormal uterine bleeding symptoms and classification of causes of abnormal uterine bleeding in the reproductive years: 2018 revisions," *International Journal of Gynecology & Obstetrics*, vol. 143, no. 3, pp. 393-408, 2018.

[3] M. Berbic, and I. S. Fraser, "Immunology of normal and abnormal menstruation," *Women's Health*, vol. 9, no. 4, pp. 387-395, 2013.

[4] C. Wober, W. Brannath, K. Schmidt, M. Kapitan, E. Rudel, P. Wessely et al., "Prospective analysis of factors related to migraine attacks: the PAMINA study," *Cephalalgia*, vol. 27, no. 4, pp. 304-314, 2007.

[5] Y. G. Antypkin, Y. P. Vdovychenko, A. Graziottin, V. V. Kaminskyi, and T. F. Tatarchuk, "Uterine bleedings and quality of woman's life resolution of advisory board," *Reproductive endocrinology*, vol. 47, no. 3, pp. 13-18, 2019.

[6] S. W. Menard, *Applied Logistic Regression Analysis*, 2nd ed, London: New Delhi, 2002.

[7] "Classification: ROC Curve and AUC," in *Machine Learning Crash Course*. Google Developers, [online document], Available: Google's fast-paced, practical introduction to machine learning, https://developers.google.com. [Accessed Jan. 7, 2020].

[8] P. Pandey, "Simplifying the ROC and AUC metrics," *towardsdatascience.com*, Mar. 3, 2019. [Online]. Available: https://towardsdatascience.com/understanding-the-roc-and-auc-curves-a05b68550b69. [Accessed Jan. 5, 2020].