



УДК 519.81+61:004

С. В. Никифоров, Б. И. Смирнов

Санкт-Петербургский государственный электротехнический
университет "ЛЭТИ" им. В. И. Ульянова (Ленина)

Модуль прогнозирования осложнений после трансплантации костного мозга у детей с лейкозами

Описан программный модуль, оценивающий вероятность возникновения осложнений после трансплантации костного мозга. В основу положена модель, построенная на основе логистической регрессии и использующая базу данных, содержащую записи 155 пациентов в возрасте от 5 до 21 года. Модуль содержит подробное графическое представление результатов в виде ROC-кривой, вычисленных значений истинного и ложного прогнозов, а также отношения правдоподобия полученных оценок. На графиках выделяются области достоверного и недостоверного прогнозов, а также область невозможности прогнозирования.

Трансплантация костного мозга, прогнозирование осложнений, логистическая регрессия, ROC-анализ, чувствительность, специфичность, истинный прогноз, ложный прогноз, отношение шансов, отношение риска, доверительный интервал, программный модуль

Трансплантация костного мозга (ТКМ) – процедура, используемая для лечения лейкозов. На этапе подготовки обычно используют химиотерапию, которая ослабляет иммунитет. В результате пациент в этот период подвержен различным инфекциям (бактериальным, вирусным, грибковым). Для борьбы с инфекциями пациенту делают профилактические процедуры [1].

Инфекционные осложнения часто являются причиной смерти, поэтому до ТКМ лечащий врач проводит собеседование, на котором принимается решение о трансплантации. Представленный в настоящей статье программный модуль наглядно демонстрирует вероятность возникновения осложнений, тем самым облегчает принятие решения и повышает его объективность. В частности, представление информации программным модулем облегчает ее понимание родителями или опекунами ребенка, дающими согласие на трансплантацию.

В большинстве известных прогностических калькуляторов [1]–[3] используются многомерные регрессионные модели, разработанные для получения оценок одной целевой переменной (ЦП), зависящей от нескольких переменных. В случае логистической регрессии ЦП имеет бинарный тип. Прогнозируется вероятность появления целевой категории в зависимости от независимых переменных.

В настоящей статье ЦП характеризует возможность появления осложнений. Она является бинарной со значениями "да" или "нет". Целью построения логистической регрессии является достижение ЦП значения "нет".

Логистическая регрессия – статистический подход, используемый для оценки связи дихотомической зависимой переменной от набора независимых переменных [4]. Благодаря тому что значения логистической функции заключены в интервале $[0, 1]$ (рис. 1), ее удобно использовать для описания вероятности прогноза целевой категории:

$$p(z) = \frac{1}{1 + e^{-z}}. \quad (1)$$

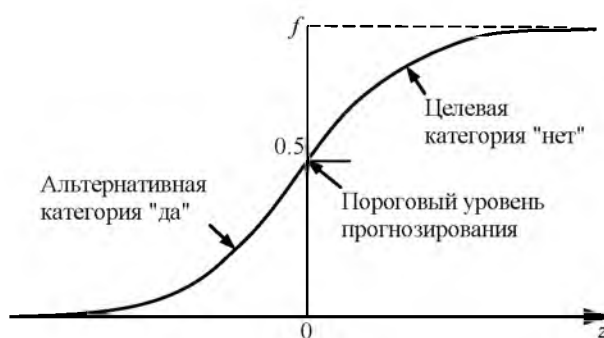


Рис. 1

Аргумент z регрессионного уравнения (1) записывается как взвешенная сумма всех независимых переменных X_i , включенных в модель:

$$z = \sum \beta_i X_i, \quad i = \overline{1, m}, \quad (2)$$

где β_i – весовые коэффициенты, определяемые в результате построения модели; m – количество независимых переменных.

Переменная z (2) строилась обычным пошаговым методом в процедуре логистической регрессии пакета SPSS [5]. Исходными данными являлся экспериментальный материал медико-биологических параметров: база данных из 155 записей детей в возрасте от 5 до 21 года. Перед построением модели был выполнен разведочный анализ [6], который позволил получить список потенциальных предикторов из 21 переменной, на основе которого сформирована логистическая модель из 6 переменных.

Предложенная модель прогнозирования отсутствия осложнений (состояние "нет") представляет собой линейное уравнение из бинарных независимых переменных $X_i, i = \overline{1, 6}$:

$$z = \ln \left[\frac{p(z)}{1-p(z)} \right] = 1.771X_1 + 0.984X_2 - 1.365X_3 - 1.853X_4 - 2X_5 - 2.75X_6. \quad (3)$$

Для приведения исходных данных к бинарным значениям $\{0, 1\}$ использовано внутреннее преобразование при построении регрессии. Весовые коэффициенты (3) получены в процессе подгонки модели.

Значение $z = 0$ используется как порог классификации: при $z > 0$, состояние классифицируется как целевая категория (в данном случае – состояние "нет"), а при $z < 0$ как альтернативная категория ("есть") (рис. 1). Полученная модель не единственная, возможны и другие варианты, например учитывающие двухфакторные связи [6].

С точки зрения радиотехники логистическая регрессия похожа на задачу распознавания двух сигналов, соответствующих различным состояниям индивида, причем распределения состояний могут различаться. Отношение "сигнал/шум", используемое для решения этой задачи, отражает полноту представления и качество извлечения информации из исходных данных.

Выражение (3) с учетом преобразования (1) доступно для анализа влияния указанного в нем набора переменных на ЦП и классификации осложнений конкретного индивида. Однако для

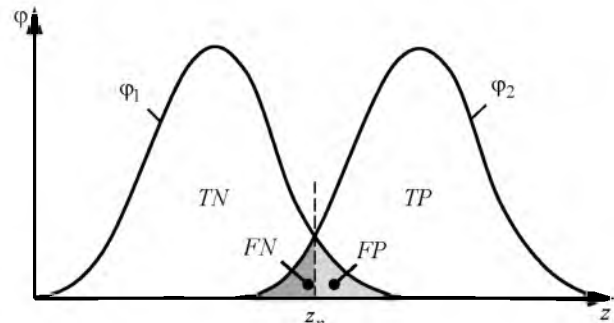


Рис. 2

применения модели в целях медицинской диагностики и проверки ее качества целесообразно перейти к общепринятым в указанной отрасли характеристикам прогнозирования: чувствительности и специфичности, которые, в свою очередь, можно получить из ROC-кривых¹ [3], [6], [7].

Чувствительность метода определяется как доля правильных решений среди всех положительных:

$$Sen = TP / (TP + FN),$$

где TP, FN – количества верных и ошибочных положительных решений соответственно.

Специфичность метода определяется как доля правильных отрицательных решений среди всех отрицательных:

$$Spe = TN / (TN + FP),$$

где TN, FP – количества верных и ошибочных отрицательных решений соответственно.

Если при отсутствии сигнала плотность вероятности значений анализируемого параметра z имеет вид $\varphi_1(z)$, а при его наличии – вид $\varphi_2(z)$ (рис. 2), вероятности принятия решений при пороге $z = z_n$ определяются указанными на рис. 2 площадями.

ROC-кривая – это график, построенный в координатах " $Sen - \overline{Spe}$ " ($\overline{Spe} = 1 - Spe$) (рис. 3), позволяющий оценить качество работы классификатора через уравнивание его чувствительности и специфичности. Классификатор обладает тем лучшим качеством, чем выше над биссектрисой первого квадранта l проходит его ROC-кривая 2 .

В медицинской практике для принятия решения о классификации конкретного набора независимых переменных, характеризующих индивида, обычно используется критерий идеального наблюдателя [7], [8], согласно которому ошибки первого и второго рода уравниваются. В ROC-пространстве (рис. 3) порог по критерию идеального

¹ ROC – receiver operating characteristic (рабочая характеристика приемника).

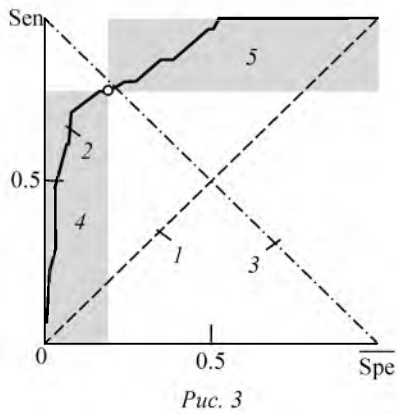


Рис. 3

наблюдателя выбирается как точка, наиболее близкая к пересечению ROC-кривой 2 с линией 3, в котором $\overline{Spe} = 1 - \overline{Sen}$ (на рис. 3 обозначена круглым маркером).

ROC-пространство на рис. 3 разбито на две части линией 1. Пространство, лежащее выше этой диагонали, относится к выбору целевой категории. Пространство ниже диагонали относится к ошибочному выбору категории. Сама диагональ соответствует случайному выбору или "нулевому прогнозу". Обычно используется первый вариант. Появление порога классификации позволяет разбить ROC-пространство еще на две части: пространство отсутствия осложнений 4 (целевая категория) и пространство осложнений 5 (альтернативная категория). В этих двух пространствах и должны быть представлены характеристики прогнозирования осложнений индивидов.

ROC-кривая – один из популярных видов представления информации о модели прогнозирования. По ней видно, как регрессионный анализ извлекает информацию о ЦП из экспериментальных данных, как оценивается качество прогнозирования.

Интегрально качество извлечения информации оценивается по площади между реальной ROC-кривой и линией нулевого прогноза (критерий AUC [7]). Кроме того, из ROC-кривой видны связь ошибок первого и второго рода модели, потенциальные значения \overline{Sen} и \overline{Spe} , однако по ней сложно определить вероятности прогнозов для реального индивида.

Поэтому второй вариант представления модели, который используется в разрабатываемом модуле, – график зависимости вероятностей прогнозов от значений регрессионного уравнения z (2). Предлагается ввести: вероятность истинного прогноза ($P_{ИП}$) – правильного предсказания состояния при его истинности и вероятность ложного прогноза ($P_{ЛП}$) – предсказания состояния при



Рис. 4

его ложности. Вместо значений чувствительности, специфичности и их дополнений предлагаются переменные: $P_{ИП1}$ – вероятность истинного прогноза состояния "нет", $P_{ЛП1}$ – вероятность ложного прогноза состояния "нет", $P_{ИП0}$ – вероятность истинного прогноза состояния "да", $P_{ЛП0}$ – вероятность ложного прогноза состояния "да".

Значения $P_{ИП}$ и $P_{ЛП}$ определяются по ROC-кривой в соответствии с выбранной областью (рис. 3). Для области прогнозирования отсутствия осложнений (рис. 3, 4) $P_{ИП1}$ определяется по оси \overline{Sen} как разность между пороговым значением и точкой состояния пациента, а $P_{ЛП1}$ – как разность по оси \overline{Spe} между порогом и точкой состояния пациента (рис. 4).

Для области прогнозирования осложнений (рис. 3, 5) вычисляются такие же разности, но для $P_{ИП0}$ по оси \overline{Spe} , а $P_{ЛП0}$ – по оси \overline{Sen} (рис. 5).

По значениям ИП и ЛП строится график прогнозов (рис. 6) в зависимости от значений регрессионного уравнения z , т. е. от состояния пациента. По этому графику более просто классифицировать осложнения (наличие – левее порога, отсутствие – правее) и определять числовое значение прогноза. На рис. 6 в области прогнозирования отсутствия осложнений сплошная кривая –



Рис. 5

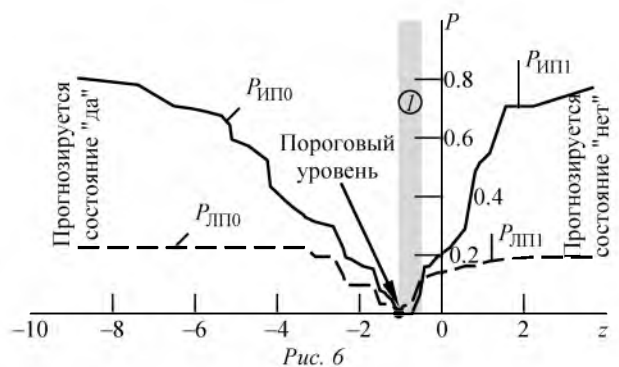


Рис. 6

вероятность $P_{ИП1}$ и штриховая кривая – вероятность $P_{ЛП1}$ переходят после порогового уровня в штриховую линию $P_{ЛП0}$ и сплошную линию $P_{ИП0}$ в области прогноза наличия осложнений соответственно. Как видно из рис. 6, условие нормировки выполняется.

По графику прогнозов (рис. 6) определилась область, где вероятность ЛП оказалась выше вероятности ИП (на рис. 6, 1). Эту область следует отнести к невозможности прогнозирования, возникновение которой обусловлено различием распределений вероятностей $\varphi_1(z)$ и $\varphi_2(z)$ вблизи порогового уровня и значительными величинами ошибок первого и второго рода.

Третий вариант представления характеристик прогнозирования (рис. 7) – отношения правдоподобия

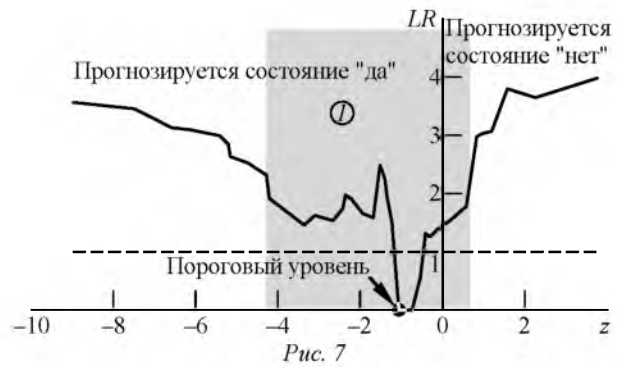
$$LR_1 = P_{ИП1}/P_{ЛП1}; LR_0 = P_{ИП0}/P_{ЛП0}$$

или относительный риск. Эта статистика широко используется для описания взаимосвязи номинальных переменных [5].

Оценки отношения правдоподобия можно задать в интервальном виде с учетом 95 %-го доверительного интервала (ДИ). Тогда вне зоны (рис. 7, 1) значения регрессионного уравнения (2) таковы, что весь ДИ лежит выше 1, что указывает границы, гарантирующие правильность прогноза с 95 %-й достоверностью. Эти значения могут быть использованы в лечебной практике как границы достоверного прогноза.

Если ДИ LR включает значение 1, это означает, что прогноз ненадежен и может наблюдаться как одно состояние, так и другое.

Так, на рис. 7 левая граница области 1 дает превышение оценки наличия осложнений над



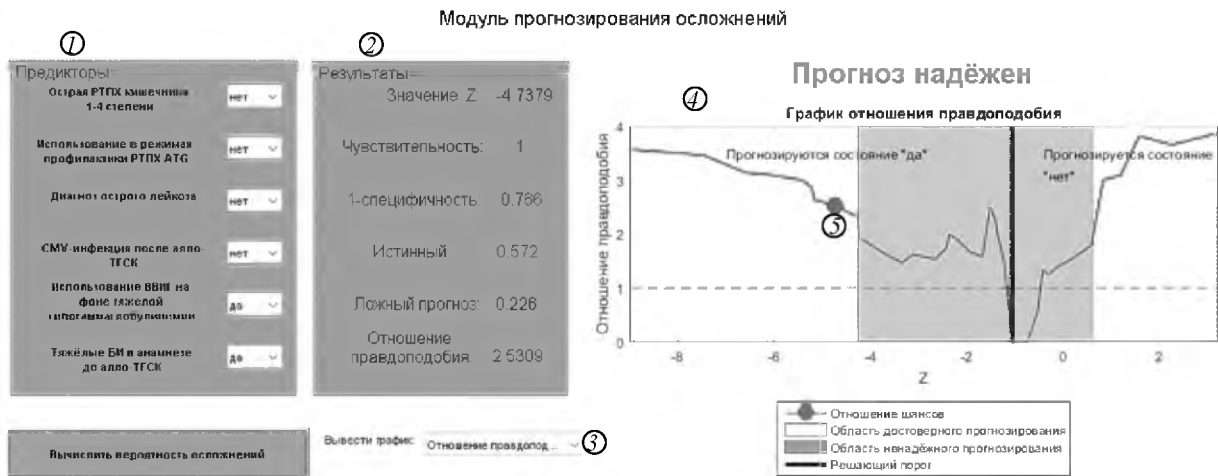
оценкой их отсутствия в 2.3 раза при 95 %-м ДИ, а правая граница этой области – превышение оценки отсутствия осложнений над оценкой их наличия в 1.8 раза поэтому справа и слева от 1 находятся области надежных прогнозов. В области 1 вероятность истинного прогноза может оказаться меньше вероятности ложного прогноза – это область ненадежного анализа.

По мере увеличения объема анализируемой выборки превышения верных оценок над ошибочными на границах области 1 будут снижаться.

Описание программного модуля. Программный модуль написан на языке Matlab с использованием графического интерфейса GUIDE. Модуль исполняется в среде Windows 7.

При исполнении модуля формируется главное окно (рис. 8), в котором расположены: блок выбора значений предикторов 1, блок вывода численных результатов 2 и окно графических представлений 4. Результаты в окне 4 могут быть отображены в формах, соответствующих рис. 3, 6, 7. Вариант представления задается в выпадающем меню в окне 3.

В блоке выбора значений предикторов 1 размещены независимые переменные $X_i, i = \overline{1, 6}$ (2), входящие в модель регрессии. Задавая набор пре-



дикторов состояний конкретного пациента, т. е. определяя вектор его состояния, можно получить числовые характеристики прогнозирования осложнений индивида и на их основании принять терапевтическое решение.

На рис. 8 приведен пример состояния пациента, определенный предикторами 1. Вычисленное значение $z = -4.7379$ соответствует прогнозированию осложнений (метка 5). На графике отношения шансов эта точка находится в области надежного прогнозирования (не менее 95 %). Вычисленное отношение шансов говорит о встречаемости осложнения 2.53 раза чаще, чем их отсутствия.

Приведенные в настоящей статье материалы показывают построение на основании модели логистической регрессии программного модуля прогнозирования осложнений после трансплантации костного мозга у детей с лейкозами. Программный модуль представляет три варианта визуализации решения проблемы прогнозирования. Модуль является инструментом информационной поддержки принятия решения в сложных вариантах лечения онкологических заболеваний. На основании полученных данных из модуля и личного мнения врача о пациенте будет вынесено решение о трансплантации.

СПИСОК ЛИТЕРАТУРЫ

1. Информационные технологии в здравоохранении: ОСИ (февр. 2013–февр. 2014) / сост. О. В. Смирнов. Казань: РОМНИ, 2014. 5 с. URL: [http://www.rmbic.ru/userfiles/OSI-2\(2\).pdf](http://www.rmbic.ru/userfiles/OSI-2(2).pdf) (дата обращения 25.12.2016).
2. Статистические методы анализа в клинической практике / П. О. Румянцев, В. А. Саенко, У. В. Румянцева, С. Ю. Чекин / ГУ "Медицинский радиологический научный центр РАМН". Обнинск, 2009. 96 с.
3. Файнзильберг Л. С., Жук Т. Н. Гарантированная оценка эффективности диагностических тестов на основе усиленного ROC-анализа // Управляющие системы и машины. 2009. № 5. С. 3–13.
4. Kleinbaum D. G., Klein M. Logistic Regression. New York: Springer-Verlag, 2002. 513 с.
5. Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках. СПб.: Питер, 2005. 416 с.
6. Маркелов О. А., Смирнов Б. И. Алгоритм построения моделей логистической регрессии при малом объеме данных // Изв. СПбГЭТУ "ЛЭТИ". 2010. № 10. С. 94–101.
7. Ланг Т. А., Сесик М. Как описывать статистику в медицине. Руководство для авторов, редакторов и рецензентов / пер. с англ. М.: Практическая медицина, 2011. 480 с.
8. Гуткин Л. С. Теория оптимальных методов радиоприема при флуктуационных помехах. 2-е изд. М.: Сов. радио, 1972. 448 с.

Статья поступила в редакцию 15 ноября 2016 г.

Для цитирования: Никифоров С. В., Смирнов Б. И. Модуль прогнозирования осложнений после трансплантации костного мозга у детей с лейкозами // Изв. вузов России. Радиоэлектроника. 2017. № 1. С. 74–79.

Никифоров Степан Викторович – бакалавр по направлению "Радиотехника" (2016), студент 1-го курса магистратуры Санкт-Петербургского государственного электротехнического университета "ЛЭТИ" им. В. И. Ульянова (Ленина). Сфера научных интересов – использование методов прикладной статистики в обработке медико-биологической информации.
E-mail: stepka_007@mail.ru

Смирнов Борис Иванович – кандидат технических наук (1977), доцент (1995) кафедры радиотехнических систем Санкт-Петербургского государственного электротехнического университета "ЛЭТИ" им. В. И. Ульянова (Ленина). Почетный работник высшего профессионального образования Российской Федерации (2005). Автор более 60 научных работ. Сфера научных интересов – разработка и применение микропроцессорных устройств; использование методов прикладной статистики в обработке медико-биологической информации.
E-mail: dmcs-2010@yandex.ru

S. V. Nikiforov, B. I. Smirnov
Saint Petersburg Electrotechnical University "LETI"

The Module for Predicting Complications after Bone Marrow Transplantation in Children with Leukemia

Abstract. The paper presents the program module that estimates the probability of complications after the bone marrow transplantation. The shown approach is based on a logistic regression. The medical database containing the records of 155 patients aged from 5 to 21 is used. The program module contains a detailed graphical representation of the results in the form of a ROC curve, the calculated values of true and false predictions as well as likelihood ratio of the obtained values. The areas of reliable and unreliable predictions as well as the area where predictions are impossible are highlighted on the chart.

Key words: Bone Marrow Transplantation, Complications Prediction, Logistic Regression, ROC-Analysis, Sensitivity, Specificity, True Forecast, False Forecast, Odds Ratio, Risk Ratio, Confidence Interval, Program Module

REFERENCES

1. *Informatsionnye tekhnologii v zdavookhranii: OSI Bibl. uk. (fevr. 2013–fevr. 2014) / sost. O. V. Smirnov* [Information technology in health care: OSI] (febr. 2013–febr. 2014). Kazan: ROMNI, 2014. 5. Available at: [http://www.rmbic.ru/userfiles/OSI-2\(2\).pdf](http://www.rmbic.ru/userfiles/OSI-2(2).pdf) (accessed: 25.12.2016). (In Russian)
2. Rumyantsev P. O., Saenko V. A., Rumyantseva U. V., Chekin S. Yu. *Statisticheskie metody analiza v klinicheskoi praktike* [Statistical methods of analysis in clinical practice]. Obninsk, 2009, 96 p.
3. Fainzil'berg L. S., Zhuk T. N. Guaranteed evaluation of diagnostic tests based on strengthened ROC-analysis. *Upravlyayushchie sistemy i mashiny*. 2009, no. 5, pp. 3–13. (In Russian).
4. Kleinbaum D. G., Klein M. *Logistic Regression*. New York, Springer-Verlag, 2002, 513 p.
5. Nasledov A. D. *Komp'yuternyi analiz dannykh v psikhologii i sotsial'nykh naukakh* [Computer analysis of data in psychology and social sciences]. St. Petersburg, *Piter*, 2005, 416 p. (In Russian).
6. Markelov O. A., Smirnov B. I. Algorithm for constructing a logistic regression model with a small amount of data. *Izvestiya SPbGETU "LETI"*. 2010, no. 10, pp. 94–101. (In Russian).
7. Lang T. A., Sesik M. *How to report statistics in Medicine*. American College of Physicians, Philadelphia, 2006, 478 p.
8. Gutkin L. S. *Teoriya optimalnykh metodov radiopriema pri fluktuatsionnykh pomehah; 2-e izd.* [The theory of optimal methods of radio receiving under fluctuation interferences; 2nd ed.] Moscow, *Sovetskoe radio*, 1972, 448 p. (In Russian)

Received November, 15, 2016.

For citation: Nikiforov S. V., Smirnov B. I. The Module for Predicting Complications after Bone Marrow Transplantation at Children with Leukemia. *Izvestiya Vysshikh Uchebnykh Zavedenii Rossii. Radioelektronika*. [Journal of the Russian Universities. Radioelectronics]. 2017, no. 1, pp. 74–79. (In Russian)

Stepan V. Nikiforov – Bachelor's Degree in Radio equipment (2016), the first year Master's Degree student of the of Saint Petersburg Electrotechnical University "LETI". Area of expertise – the use of applied statistics methods in the processing of biomedical information.

E-mail: stepka_007@mail.ru

Boris I. Smirnov – Ph.D. in engineering (1977), Associate Professor (1995) of the Department of Radio Engineering Systems of Saint Petersburg Electrotechnical University "LETI". Honorary Worker of higher professional education of the Russian Federation (2005). The author of more than 60 scientific publications. Area of expertise: microprocessor devices development and application; the use of applied statistics methods in the processing of biomedical information.

E-mail: dmcs-2010@yandex.ru