УДК 519.816
DOI: 10.25206/1813-8225-2018-160-179-184

А. С. СЕРОБАБОВ[1]
Е. В. ЧЕБАНЕНКО[1]
Л. А. ДЕНИСОВА[1]
Т. С. КРОЛЕВЕЦ[2]

[1]Омский государственный технический университет, г. Омск

[2]Омский государственный медицинский университет, г. Омск



# РАЗРАБОТКА ЭКСПЕРТНОЙ СИСТЕМЫ РАННЕЙ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ: ПРОГРАММНЫЕ СРЕДСТВА ПЕРВИЧНОЙ ОБРАБОТКИ И ВЫЯВЛЕНИЕ ЗАВИСИМОСТЕЙ

**Рассмотрены инструментальные средства описательной статистики для обработки биомедицинской информации данных клинических исследований болезней печени. Разработана структура программного комплекса первичной обработки данных, характеризующих состояние пациентов, включая результаты лабораторных исследований, информацию о сопутствующих заболеваниях, а также о физиологических параметрах пациентов. Получена карта взаимосвязей результатов исследования состояния пациентов, позволяющая выявлять зависимости между показателями болезней печени для разработки экспертной системы ранней диагностики заболеваний.**

**Ключевые слова: описательная статистика, экспертная система, ранняя диагностика, медицинская информация, корреляционный анализ.**

**Введение.** В настоящее время медицинские экспертные системы, базирующиеся на современных информационных технологиях обработки биомедицинской информации, находят все более широкое применение как в медицинских исследованиях, так и в реальной клинической практике [1–3].

В работе рассмотрены вопросы разработки экспертной системы ранней диагностики заболеваний печени, при построении которой необходимо использовать современные методы обработки данных для повышения эффективности процесса постановки диагноза. В качестве исходных данных для разработки экспертной системы использованы результаты, полученные при выполнении в Омском государственном медицинском университете исследований по неинвазивной оценке степени фиброза у пациентов с неалкогольной жировой болезнью печени (НАЖБП) [4].

Актуальность исследования обусловлена тем, что за последние несколько лет на фоне устойчивой тенденции роста распространённости среди населения избыточной массы и ожирения [4–6] неалкогольная жировая болезнь печени занимает лидирующее место среди причин заболевания печени. По данным исследования DIREG 2 [5] 37 % пациентов населения России, из лиц первичного или повторно обратившихся в лечебно-профилактическое учреждение, имеют подозрение на НАЖБП.

Предлагается подход к построению экспертной системы, предназначенной осуществлять прогнозную оценку течения болезни печени на основе обработки данных клинических исследований болезней печени, лабораторных исследований сопутствующих заболеваний и физических параметров пациентов в целях повышения качества ранней диагностики болезни.

**Постановка задачи разработки экспертной системы.** В качестве исходных данных для разработки экспертной системы использованы результаты, полученные при обследовании 149 пациентов с выявленной неалкогольной жировой болезнью печени. Каждый пациент с диагнозом НАЖБП был отобран в результате диспансеризации населения из различных поликлинических учреждений города Омска.

Представленные данные по пациентам состоят из четырех групп: лабораторные исследования, данные по сопутствующим заболеваниям, физиологические параметры пациентов, принимаемые препараты.

Необходимо провести статистические расчеты для каждой группы данных и представить результаты в виде таблиц и графиков.

В общем виде задача диагностики ставится следующим образом. Имеется выборка $X$ из $m$ объектов (пациентов с различными стадиями фиброза), характеризующихся $n$ переменными (параметрами,

взятыми из четырех составных групп данных по пациенту).

$$X = \left[x_{ij}\right] = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix},$$

где $i$ — номер объекта (пациента); $j$ — номер переменной (параметра).

В рассмотрение вводится вектор диагнозов $Y = (y_1, y_2, \dots y_l, \dots y_k)$, где $y_l$ — один из возможных диагнозов (стадий фиброза); $k$ — количество диагностируемых классов (возможных диагнозов).

Для определения диагноза пациенту необходимо отнести каждый $i$-й объект ($i = 1 \dots m$) с определенным набором значений параметров $j$ ($j = 1 \dots n$) к одной из имеющихся стадий фиброза $y_l$ ($l = 1 \dots k$).

Для этой цели проектируется экспертная система, позволяющая на основе выявленных взаимосвязей между параметрами, характеризующими состояние пациента (данные лабораторных исследований, информация по сопутствующим заболеваниям, физиологические параметры пациентов, принимаемые препараты), поставить предварительный диагноз пациенту, решив задачу классификации (определив стадию фиброза).

Проектирование экспертной системы реализуется на базе математического аппарата нечеткой логики [7, 8], используемого для построения набора продукционных правил, написанных на естественном языке качественных понятий, что обусловлено трудностью формализации диагностического процесса. Особые свойства нечеткой системы позволяют не только учитывать неопределенность, но и формировать собственные рассуждения на основе опыта эксперта — специалиста в области обработки биомедицинской информации.

Ввод исходных данных, характеризующих определенного пациента, производится с помощью интерфейса системы. Далее производится первичная обработка информации, результатом которой является построение таблиц с данными первичной обработки, гистограмм, отражающих распределение параметров, и регрессионных моделей для выявления зависимостей, а также подготовка данных для последующей визуализации.

Этап первичной обработки данных экспертной системы, кроме того, позволяет уменьшить размерность имеющихся данных на основе принятой гипотезы о связанности параметров [4, 5].

**Описание программного комплекса первичной обработки данных.** На рис. 1 представлена структура программного комплекса для первичной обработки данных пациента. На рисунке выделены три группы программных модулей, с помощью которых выполняются предварительная обработка данных, анализ информации и визуализация результатов.

На стадии предобработки данных неподготовленные сведения группируются по категориям в соответствии с принятыми критериями, проводится их очистка от аномальных данных. Результаты передаются на вход модулей анализа данных и определения статистических характеристик.

Модулями анализа производятся расчеты статистических показателей, включая статистическое описание данных физиологических параметров, лабораторных анализов и информации о наличии сопутствующих заболеваний пациентов. Полученные данные подвергаются преобразованию и подаются на вход модулей визуализации.

Модулями визуализации выполняется табличное представление результатов первичной обработки данных, построение диаграмм и графиков.

Программный комплекс для первичной обработки данных реализован на языке объектно-ориентированного программирования Python (обладающем набором библиотек, высокой гибкостью и динамичностью) [9] с использованием интерактивной оболочки Jupyter. При этом программный модуль предобработки данных разработан с использованием библиотеки Numpy, позволяющей совершать операции с большими объёмами данных и многомерными массивами. Модули анализа используют


**Рис. 1. Структура программного комплекса для первичной обработки данных**

**Таблица 1**

**Статистические характеристики физиологических параметров пациентов**

| Статистические характеристики | Физиологические параметры пациентов | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $A_p$, год | муж (1) жен (2) | $L_g$, мм | $T_g$, мм | $W_g$, мм | $H_p$, см | $P_p$, кг | $S_m$, мм$^2$ | $C_p$, см |
| Количество элементов | 149 | 149 | 110 | 81 | 110 | 138 | 138 | 115 | 113 |
| Среднее $M(K)$ | 48,49 | 1,23 | 75,18 | 2,15 | 28,19 | 173,4 | 98,62 | 35,86 | 108 |
| σ | 10,34 | 0,43 | 11,43 | 0,42 | 6,38 | 8,5 | 14,55 | 9,97 | 9,9 |
| $K_{min}$ | 23 | 1 | 46 | 1 | 18 | 152 | 64 | 17 | 87 |
| $Q_{25}$ | 41 | 1 | 66 | 2 | 23,25 | 168,3 | 89 | 29 | 101 |
| $Q_{50}$ | 46 | 1 | 75 | 2 | 28 | 174 | 98 | 35 | 107 |
| $Q_{75}$ | 57 | 1 | 83 | 2 | 31,75 | 179 | 106 | 40 | 115 |
| $K_{max}$ | 73 | 2 | 111 | 3 | 55 | 190 | 147 | 74 | 140 |

$C_p$ — обхват талии пациента; $H_p$ — рост пациента; $P_p$ — вес пациента; $A_p$ — возраст пациента; $S_m$ — площадь селезенки пациента; $W_g$ — ширина стенки желчного пузыря; $L_g$ — длина стенки желчного пузыря; $T_g$ — толщина желчного пузыря

**Таблица 2**

**Статистические характеристики сопутствующих заболеваний пациентов**

| Статистические характеристики | Сопутствующие заболевания пациентов | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Стеатоз (1) Гепатит (2) | $D_{СД2}$ | $D_{НТГ}$ | $D_{АГ}$ | $D_{ИБС}$ | $D_{О}$ | $D_{Б}$ | $D_{ОС}$ | $D_{НАСГ}$ |
| Количество элементов | 149 | 149 | 149 | 149 | 149 | 149 | 149 | 149 | 149 |
| Среднее $M(K)$ | 1,52 | 0,13 | 0,23 | 0,64 | 0,14 | 0,97 | 0,05 | 0,13 | 1,52 |
| σ | 0,5 | 0,34 | 0,42 | 0,48 | 0,35 | 0,16 | 0,23 | 0,34 | 0,5 |
| $K_{min}$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $Q_{25}$ | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| $Q_{50}$ | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| $Q_{75}$ | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| $K_{max}$ | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |

$D_{АГ}$ — артериальная гипертензия; $D_{ИБС}$ — ишемическая болезнь сердца; $D_{НТГ}$ — нарушенная толерантность к глюкозе; $D_{О}$ — ожирение у пациента; $D_{Б}$ — болезнь бронхов; $D_{ОС}$ — присутствие остеоартроза, $D_{СД2}$ — сахарный диабет 2 типа, $D_{НАСГ}$ — наличие заболевания неалкогольный стеатогепатит

библиотеку Pandas, визуализация производится библиотекой графического отображения Matplotlib и Sklearn.

**Результаты работы программного комплекса первичной обработки данных.** На этапе первичного анализа данных производится обработка массивов статистических данных с целью нахождения обобщающих характеристик элементов массива. Выполняется статистическое описание исходных совокупностей параметров с определением пределов варьирования переменных, анализ выбросов данных, восстановление пропущенных наблюдений. Проводится анализ основных статистических показателей.

В табл. 1–3 приведены результаты расчетов следующих статистических показателей [10]:

— $K_{min}$ ($K_{max}$) — минимальное (максимальное) значение исследуемого параметра;

— математическое ожидание (среднее значение исследуемого параметра) $M(K_j) = \frac{1}{n_j}\sum_{i=1}^{n_j} K_{ij}$, где $j$ — индекс параметра, $i$ — индекс элемента параметра, $K_{ij}$ — $i$-й элемент $j$-го параметра, $n_j$ — количество элементов $j$-го параметра;

Таблица 3

**Статистические характеристики лабораторных анализов пациентов**

| Статистические характеристики | Лабораторные анализы пациентов | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ТИМР1, нг/мл | ТИМР2, нг/мл | ММП9, нг/мл | HOMA-IR, у.е | ObR, нг/мл | Адипо Q, нг/мл | Лептин, нг/мл | ПТИ% | ГГТ (50), ед/л |
| Количество | 87 | 87 | 87 | 87 | 35 | 111 | 108 | 72 | 63 |
| Среднее $M(K)$ | 1464 | 127,3 | 391,2 | 6,76 | 9,42 | 18,89 | 21,55 | 97,49 | 81,63 |
| σ | 579,71 | 45,18 | 219,9 | 7,37 | 10,35 | 13,06 | 18,45 | 11,94 | 165,8 |
| $K_{min}$ | 570 | 70,5 | 61 | 0,12 | 2,46 | 0,07 | 1,35 | 18 | 5 |
| $Q_{25}$ | 1105 | 93,5 | 250,5 | 1,44 | 4,52 | 7,04 | 9,69 | 93,75 | 31,9 |
| $Q_{50}$ | 1345 | 113 | 342 | 4,54 | 7,03 | 18 | 16,31 | 99 | 53 |
| $Q_{75}$ | 1582 | 153,7 | 486 | 10,15 | 9,92 | 27,45 | 26,43 | 104 | 75 |
| $K_{max}$ | 4105 | 286 | 1636 | 43,64 | 64,32 | 61,2 | 108,8 | 116 | 200 |

ТИМР1 (2) — тканевой ингибитор матриксных протеиназ 1 (2); ММП9 — матриксная металлопротеиназа 9; HOMA-IR (Homeostasis Model, Assessment of Insulin Resistance) — индекс инсулинорезистентности; ПТИ — протромбиновый индекс; ГГТ (50) ед/л — гамма-глютамилтрансептидаза; Адипо Q — адипонектин

— дисперсия (характеризует меру изменчивости исследуемой величины) $D(K_j) = M(K_j^2) - (M(K_j))^2$, где $j$ — индекс параметра, $M(K_j)$ — математическое ожидание;

— среднеквадратическое отклонение (характеризует величину отклонений значений от среднего) $\sigma = \sqrt{D(K_j)}$, где $D(K_j)$ — дисперсия $j$-го параметра;

— квартили: $Q_{25}$ — нижний (первый) квартиль (значение случайной величины, ниже которого находится 25 % выборки); $Q_{50}$ — медиана (второй квартиль); $Q_{75}$ — верхний (третий) квартиль (значение случайной величины, выше которого находится 25% выборки).

Для визуализации областей наиболее достоверных значений, выделенных на основании анализа диаграмм распределения, выполнено построение диаграмм размаха. Диаграмма размаха, или т.н. «ящик с усами» (англ. *box-and-whiskers diagram*), представляет собой график, компактно изображающий одномерное распределение вероятностей [11]. Диаграмма показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.

Строится данная диаграмма на основании формулы $X_1 = Q_{25} - k(Q_{75} - Q_{25})$; $X_2 = Q_{75} - k(Q_{75} - Q_{25})$, где $X_1$ — нижняя граница уса; $X_2$ — верхняя граница уса; $k$ — коэффициент, наиболее часто употребляемое значение которого равно 1,5.

Анализ диаграммы размаха позволяет определить: область наиболее достоверных значений и выбросов, отличных от всей совокупности выборки.

На рис. 2 изображена диаграмма размаха для следующих пяти параметров пациентов: рост, вес, возраст, ObR, лептин. Как видно на рисунке, три из оцениваемых параметров (возраст, ObR, рост) не имеют четко выраженных выбросов. Однако у двух из оцениваемых параметров имеются области выброса (это вес пациента и лептин — гормон, регулирующий энергетический обмен).

По мнению эксперта, учитывая специфику заболевания, проявляющуюся у людей с ожирением, результаты выброса параметра «вес пациента» (Выброс 1) не стоит исключать из общей выборки. Необходимо принять во внимание принадежность параметра пациента к выбросу, для того чтобы учитывать при постановке диагноза.

Кроме того, анализ диаграммы позволяет выявить то, что прослеживается связь принадлежности некоторых пациентов к двум выборкам сразу. Это позволяет сделать предположение о том, что существует зависимость между параметрами лабораторных значений лептина и физиологичекого показателя «вес пациента».

В результате первичной обработки информации выполнено построение корреляционной матрицы данных пациентов. С помощью анализа корреляций эксперт может установить, существует ли зависимость между двумя величинами. Коэффициент корреляции двух случайных величин (параметров пациента $K_i$ и $K_j$) рассчитывается по формуле [10]

$$r_{K_iK_j} = \frac{\sum (K_i - M(K_i))(K_j - M(K_j))}{\sqrt{\sum (K_i - M(K_i))^2 \sum (K_j - M(K_j))^2}} \quad .$$

На рис. 3 представлена полученная в результате обработки данных карта взаимосвязей результатов исследования, соответствующая корреляционной матрице параметров пациентов. Она представляет собой симметричную квадратную матрицу размером $m \times m$ ($m$ — число параметров пациентов), главная диагональ которой заполнена единицами, а недиагональные элементы представляют собой коэффициенты корреляции. Для лучшей визуализации накладывается цветовой градиент, соответствующий степени взаимосвязей параметров.

Анализ карты взаимосвязей параметров пациентов позволяет выделить, например, следующие зависимости. Параметр 1 (Эласто F) имеет слабую

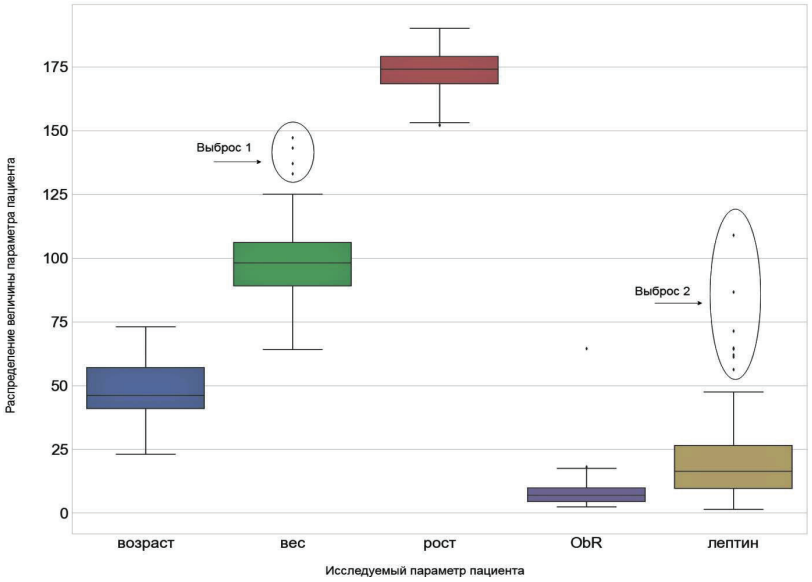


**Рис. 2. Диаграмма размаха параметров состояния пациентов**


**Рис. 3. Карта взаимосвязей результатов исследования:**
**оцениваемые параметры: 1 — Эласто F; 2 – ObR; 3 — дискомфорт; 4 — лептин;**
**5 — обхват талии; 6 — ТИМР 2; 7 — динамика массы; 8 — ожирение;**
**9 — НАСГ; 10 — ГГТ (50) ед/л; 11 — ММП9; 12 — время динамики;**
**13 — остеоартроз; 14 — метформин; 15 — АГ; 16 — стеатоз, гепатит;**
**17 — увеличение печени; 18 — гепатомегалия УЗИ**

корреляцию с параметрами 2 (ObR), 5 (обхват талии пациента), 6 (тканевой ингибитор матриксных протеиназ 2), 7 (динамика массы), 9 (неалкогольный стеатогепатит), 12 (время динамики), 14 (метформин), 16 (стеатоз, гепатит), 17 (увеличение печени), 18 (гепатомегалия УЗИ). Также, используя представленные на карте данные, можно установить наличие или отсутствие взаимосвязей между другими параметрами, характеризующими состояние пациентов. Таким образом, модули программного комплекса позволяют не только выполнить анализ основных статистических показателей, но и наглядно представить информацию пользователю — врачу-специалисту.

**Заключение.** В результате исследования разработана структура программного комплекса первичной обработки данных, характеризующих состояние пациентов, включая результаты лабораторных исследований, информацию о сопутствующих заболеваниях, а также о физиологических параметрах пациентов. Сформирована карта взаимосвязей результатов исследования состояния пациентов, позволяющая выявлять зависимости между показателями заболевания.

Представляется перспективным использовать полученные результаты в качестве входной информации экспертной системы ранней диагностики с целью автоматизации процесса обработки данных

и повышения точности при постановке диагноза. Разрабатываемая экспертная система на основе интеллектуального анализа данных позволит врачу не только проверить собственные диагностические предположения, но и получить информационную поддержку в трудных диагностических случаях.

**Библиографический список**

1. Афанасьева С. М., Токарев В. Л. Интеллектуальный анализ медицинской информации для принятия решений // Вестник новых медицинских технологий. 2006. Т. 13, № 1. С. 138—140.

2. Симанков В. С., Халафян А. А. Системный анализ и современные информационные технологии в медицинских системах поддержки принятия решений. М.: Бином, 2009. 362 с. ISBN 978-5-9518-0384-9.

3. Дюк В. А., Эмануэль В. Л. Информационные технологии в медико-биологических исследованиях. СПб.: Питер, 2003. 528 с. ISBN 5-94723-501-3.

4. Ливзан М. А., Кролевец Т. С., Лаптева И. В. [и др.]. Неалкогольная жировая болезнь печени у лиц с абдоминальным типом ожирения // Доказательная гастроэнтерология. 2014. № 4. С. 8—14.

5. Ивашкин В. Т., Драпкина О. М., Маев И. В. [и др.]. Распространенность неалкогольной жировой болезни печени у пациентов амбулаторно-поликлинической практики в Российской Федерации: результаты исследования DIREG 2 // Российский журнал гастроэнтерологии, гепатологии, колопроктологии. 2015. № 6. С. 31—41.

6. Vandevijvere S., Chow C. C., Hall K. D. [et al.]. Increased food energy supply as a major driver of the obesity epidemic: a global analysis // Bulletin of the World Health Organization. 2015. Vol. 93, Issue 7. P. 446—456. DOI: 10.2471/BLT.14.150565.

7. Штовба С. Д. Проектирование нечетких систем средствами MATLAB. М.: Горячая линия — Телеком, 2007. 288 с.

8. Meshcheryakov V., Denisova L. Computer-aided design of the fuzzy control system using the genetic algorithm // Dynamics of Systems, Mechanisms and Machines (Dynamics), Nov. 15—17, 2016. Omsk, 2016. DOI: 10.1109/Dynamics.2016.7819000.

9. Сузи Р. А. Язык программирования Python. М.: Бином, 2007. 328 с. ISBN 978-5-94774-711-9; 978-5-9556-0109-0.

10. Кремер Н. Ш. Теория вероятностей и математическая статистика. 2-е изд., перераб. и доп. М.: ЮНИТИ-ДАНА, 2004. 573 с. ISBN 5-238-00573-3.

11. Frigge M., Hoaglin D. C., Iglewicz B. Some Implementations of the Boxplot // The American Statistician. 1989. Vol. 43, Issue 1. P. 50—54.

**СЕРОБАБОВ Александр Сергеевич,** студент кафедры «Автоматизированные системы обработки информации и управления» Омского государственного технического университета (ОмГТУ).
Адрес для переписки: aserobabow95@mail.ru
**ЧЕБАНЕНКО Евгений Владимирович,** ассистент кафедры «Радиотехнические устройства и системы диагностики» ОмГТУ.
Адрес для переписки: evchebanenko@gmail.com
**ДЕНИСОВА Людмила Альбертовна,** доктор технических наук, доцент (Россия), профессор кафедры «Автоматизированные системы обработки информации и управления» ОмГТУ.
SPIN-код: 4926-3449
AuthorID (РИНЦ): 512788
Адрес для переписки: denisova@asoiu.com
**КРОЛЕВЕЦ Татьяна Сергеевна,** аспирантка кафедры «Факультетская терапия, профессиональные болезни» Омского государственного медицинского университета.

**Для цитирования**

Серобабов А. С., Чебаненко Е. В., Денисова Л. А., Кролевец Т. С. Разработка экспертной системы ранней диагностики заболеваний: программные средства первичной обработки и выявление зависимостей // Омский научный вестник. 2018. № 4 (160). С. 179—184. DOI: 10.25206/1813-8225-2018-160-179-184.

**Статья поступила в редакцию 20.06.2018 г.**