Математическая биология и биоинформатика
2016. Т. 11. № 1. С. 46–63. doi: 10.17537/2016.11.46

==========ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ==========
==========ТЕХНОЛОГИИ В БИОЛОГИИ И МЕДИЦИНЕ==========

УДК: 519.24+004.931+616.12

Применение оптимальных разбиений для многопараметрического анализа данных в клинических исследованиях

©2016 Гулиев Р.Р.[1,*], Сенько О.В.[2], Затейщиков Д.А.[3,5], Носиков В.В.[1], Упоров И.В.[4], Кузнецова А.В.[1], Евдокимова М.А.[3,5], Терещенко С.Н.[6], Акатова Е.В.[7], Глезер М.Г.[8], Галявич А.С.[9], Козиолова Н.А.[10], Ягода А.В.[11], Боева О.И.[11], Шлык С.В.[12], Левашов С.Ю.[13], Константинов В.О.[14], Бражник В.А.[3,5], Варфоломеев С.Д.[1], Курочкин И.Н.[1,4]

[1]*Институт биохимической физики им. Н.М. Эмануэля РАН, Москва, Россия*
[2]*Вычислительный центр им. А.А. Дородницына РАН, Москва, Россия*
[3]*Центральная государственная медицинская академия управления делами президента Российской Федерации, Москва, Россия*
[4]*Московский Государственный Университет им. М.В. Ломоносова, Москва, Россия*
[5]*Городская клиническая больница № 51 ДЗМ, Москва, Россия*
[6]*Институт экспериментальной кардиологии РКНПК, Москва, Россия*
[7]*Московский государственный медико-стоматологический университет, Москва, Россия*
[8]*Московская медицинская академия им. И.М. Сеченова, Москва, Россия*
[9]*Казанский государственный медицинский университет, Казань, Россия,*
[10]*Пермский государственный медицинский университет им. академика Е.А. Вагнера, Пермь, Россия*
[11]*Ставропольский государственный медицинский университет, Ставрополь, Россия*
[12]*Ростовский государственный медицинский университет, Ростов-на-Дону, Россия*
[13]*Уральская государственная медицинская академия дополнительного образования, Челябинск, Россия*
[14]*Северо-Западный государственный медицинский университет им. И.И. Мечникова, Санкт-Петербург, Россия*

Аннотация. В данном исследовании, построена прогностическая модель, позволяющая оценить риск возникновения неблагоприятных исходов в первые полгода после перенесенного обострения ишемической болезни сердца (ИБС). Анализируемые данные, на основе которых строилась модель, собирались в течение семи лет в 16 клиниках семи городов России и содержат широкий набор клинических, биохимических и генетических показателей. Для построения модели использовались подходы, основанные на оптимальных разбиениях: метод оптимально достоверных разбиений (ОДР) и модифицированный метод статистически взвешенных синдромов (МСВС). Полученная система оценки риска имеет хорошую прогностическую силу (AUC = 0.72). Также показано, что она обладает

* glvrst@gmail.com

большей точностью предсказания по сравнению с моделями, полученными наиболее известными методами: логистическая регрессия, деревья решений, нейронные сети и др.

Ключевые слова: *острый коронарный синдром, ишемическая болезнь сердца, распознавание, коллективные методы, оптимальные разбиения, прогнозирование.*

1. ВВЕДЕНИЕ

Ишемическая болезнь сердца (ИБС) является весьма распространенным заболеванием. В развитых странах мира – это одна из основных причин смерти и утраты трудоспособности. По данным ВОЗ [1] за 2012 г., среди 10 ведущих причин смерти в мире, ИБС занимает первое место (7.4 млн. случаев).

Очевидно, что такой масштаб проблемы подталкивает на поиск дополнительных инструментов, которые позволили бы улучшить качество терапии больных ИБС. Одним из них служит прогностическая модель, позволяющая оценить риск возникновения неблагоприятных исходов у больных, перенесших обострение ИБС. Примером наиболее часто используемых моделей подобного рода являются системы стратификации риска TIMI [2, 3], PURSUIT [4], GRACE[5, 6].

Отличительной чертой данного исследования является то, что был существенно расширен спектр рассматриваемых признаков: добавилось значительное количество потенциально важных параметров, которые отражают генетические особенности пациента, его образ жизни, перенесённые заболевания.

Таким образом, задачей данной работы было построение прогностической модели по более широкому набору показателей. Для решения данной задачи мы использовали подходы, основанные на оптимальных разбиениях: модифицированный метод статистически взвешенных синдромов (МСВС) [7] и метод оптимальных достоверных разбиений (ОДР) [8, 9].

Данные подходы уже показали свою эффективность в ряде клинических исследований [7–12]. Они также обладают рядом свойств особенно полезных при построении прогностических моделей на основании клинических исследований с большим количеством параметров. В частности, они не чувствительны к пропускам и аномальным значениям в данных, что очень важно, так как увеличение числа анализируемых параметров на практике ведёт к заметному возрастанию числа пропусков в итоговой базе данных.

Для понимания качества построенной прогностической модели было также проведено сравнение нашего результата с другими наиболее широко известными методами [13]: логистическая регрессия, деревья решений, нейронные сети, дискриминантный анализ, метод опорных векторов.

2. МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Сбор данных

В сборе данных для исследования участвовали 16 центров в семи разных городах России: Москве, Казани, Перми, Ставрополе, Ростове-на-Дону, Челябинске и Санкт-Петербурге. Набор и последующее наблюдение за больными выполнялись в период с 2004 по 2010 годы.

В исследование включались больные на 10-й день от момента развития обострения ИБС (инфаркта миокарда или нестабильной стенокардии) при условии стабилизации клинического состояния. Больные, умершие в первые 10 дней после обострения, в исследование не включались. В случае развития в течение этого периода рецидива

инфаркта миокарда, повторных симптомов ишемии миокарда длительностью более 10 минут на фоне оптимальной медикаментозной терапии, повторных изменений электрокардиограммы, свидетельствующих об ишемии, повторного повышения уровня кардиоспецифических ферментов, включение откладывали еще на 10 суток. У всех больных определяли уровень креатинина, глюкозы, мочевой кислоты, исследовали липидный спектр крови, проводили генотипирование и эхокардиографию (ЭхоКГ).

В ходе наблюдения регистрировались следующие неблагоприятные исходы: нефатальный и фатальный инфаркт миокарда (ИМ), нефатальный и фатальный инсульт, потребовавшая госпитализации нестабильная стенокардия, внезапная сердечная смерть и смерть от других (некардиальных) причин. Наличие конечных точек устанавливалось при телефонных контактах или во время амбулаторного приема.

Таким образом, собранная база данных (БД) содержит разнообразную информацию: анамнез, физиологические характеристики пациента при поступлении и выписке, информацию о лекарствах назначенных больному в ходе лечения и на момент выписки, данные электрокардиограммы (ЭКГ) и ЭхоКГ, особенности стиля жизни, информацию о родственниках и генетических маркерах больного (30 параметров). На 1 октября 2010 года в БД содержалось 407 параметров для 1193 пациентов. Более подробно методика сбора данных описана в работах [14–17].

2.2. Преобразования базы

Как было упомянуто выше, используемый в нашей работе подход, основанный на оптимальных разбиениях, не чувствителен к пропускам и аномальным значениям в данных. Эта особенность метода значительно упростила задачу предобработки данных.

Таким образом, перед построением прогностической модели были проведены лишь следующие преобразования исходной базы:

– из рассмотрения были исключены признаки, которые (по техническим причинам) были заполнены лишь в нескольких центрах.

– часть связанных по смыслу показателей объединялась в один общий. Например, параметр «Сахарный диабет» был сформирован из трех исходных и характеризует не только наличие или отсутствие у пациента этого заболевания, но и тяжесть его течения. Аналогичным образом были сформированы параметры «Количество ИМ», «История ИБС у родителей», «Потребление алкоголя» и «Курение».

– были добавлены расчетные показатели: «Индекс массы тела» [18], «Скорость клубочковой фильтрации» (формула Кокрофта-Голта) [19], «Отношение окружности талии к окружности бедер» [20].

– все категориальные признаки были разбиты на несколько бинарных (по одному на каждую категорию).

В результате была получена база, содержащая 382 переменные. В качестве целевой переменной прогнозирования использовался бинарный признак: 1 – наступление в первые полгода после выписки одного из исходов: фатальный/нефатальный ИМ, фатальный/нефатальный инсульт, нестабильная стенокардия, госпитализация для лечения периферического атеросклероза; 0 – отсутствие осложнений, перечисленных для 1.

В таблице 1 представлено распределение зафиксированных осложнений в первые полгода после выписки пациента. Как мы видим, общее количество пациентов, у которых было зафиксировано осложнение, составляет чуть больше 10 % от всех наблюдавшихся пациентов.

Таблица 1. Распределение осложнений в первые полгода после выписки пациента

Зафиксированное осложнение в первые полгода после выписки	Количество пациентов (% от общей выборки)
Фатальный инфаркт	45 (4 %)
Фатальный инсульт	3 (0.3 %)
Нефатальный инфаркт	24 (2 %)
Нефатальный инсульт	6 (0.5 %)
Нестабильная стенокардия	55 (5 %)
Госпитализация для лечения периферического атеросклероза	3 (0.3 %)

2.3. Общее описание методики построения прогностической модели

В данном разделе приведено краткое описание методики, использовавшейся в данной работе для построения прогностической модели.

В основе использованного подхода лежит поиск закономерностей, которые задаются оптимальными разбиениями признакового пространства. Разбиения ищутся таким образом, чтобы максимально разделить объекты из сравниваемых групп.

На рисунке 1 представлены различные классы (модели) разбиений.

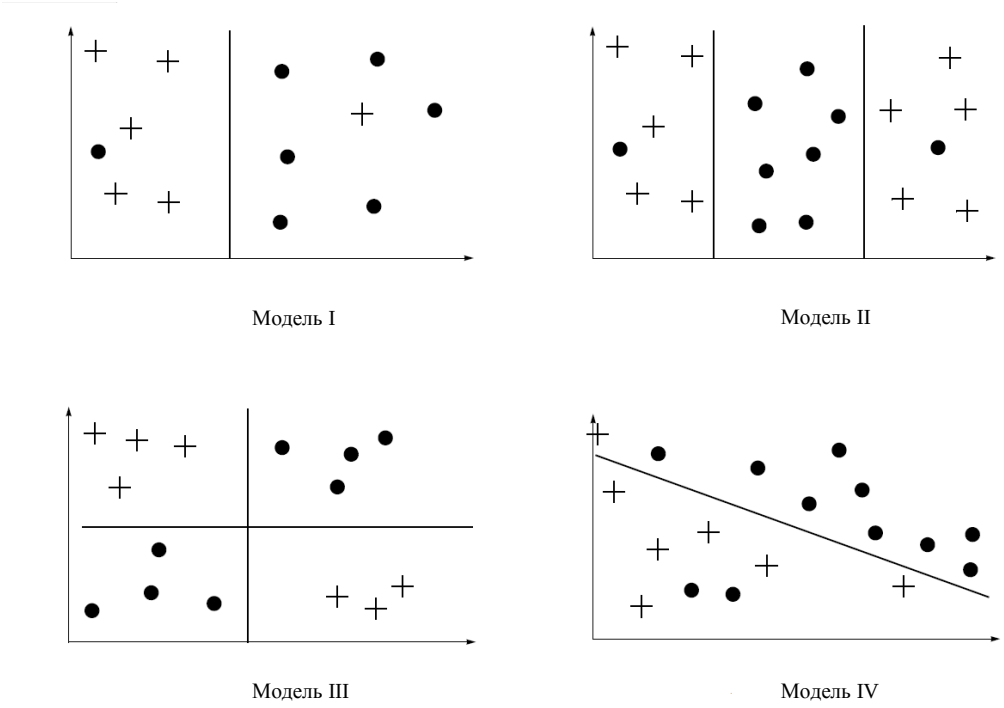

Рис. 1. Классы разбиений. Осями координат являются предиктивные признаки (например, возраст, пол, СКФ и т.д.). «●» и «+» – обозначены значения целевой переменной (например, болен / не болен, наблюдалось осложнение / не наблюдалось).

Классы I и II называют одномерными разбиениями, III и IV – двумерными.

В настоящей работе использовался метод оптимальных достоверных разбиений (ОДР) [8, 9] и метод мультимодельных статистически взвешенных синдромов (МСВС) [7]. Первый из использованных методов позволяет найти и верифицировать всевозможные одномерные и двумерные закономерности. В данном методе рассматриваются I–III модели. Второй метод строит собственно прогностическую

модель с помощью коллективного решения по разбиениям из всех представленных моделей I–IV.

Суть метода ОДР заключается в том, что (в рамках каждой рассматриваемой модели) для всех признаков/пар признаков (в зависимости от модели) выполняются следующие две процедуры:

1. *Поиск оптимального разбиения.* На этом шаге вычисляется граничное значение, максимизирующие заданный функционал качества (обычно это значение статистики χ^2). Вычисление граничных значений выполнятся путем полного перебора всех возможных вариантов. При этом рассматриваются только те наблюдения, в которых значение признака (или пары признаков) не пропущено, что позволяет методу эффективно работать с пропусками.

Вычисленное граничное значение называют оптимальным граничным значением (или оптимальной границей). Значение функционала качества при оптимальной границе (то есть максимально возможное значение при рассматриваемых данных) также называют оптимальным. Оптимальное значение функционала используется далее в качестве количественной оценки значимости найденной закономерности.

Например, при поиске оптимального разбиения модели I для признака «Возраст», вычисляется *граничное значение*, при котором максимально значение статистики χ^2 для пары «возраст не меньше *граничного значения*» и «целевая переменная» (в нашем случае - возникновение осложнения в первые полгода после выписки).

2. *Верификация найденной закономерности.* На этом шаге с помощью перестановочного теста производится проверка достоверности найденной закономерности.

Для модели I проверяется нулевая гипотеза о независимости пары «признак не меньше *граничного значения*» и «целевая переменная» (т.е. найденная закономерность является случайной). Под перестановочным тестом в данном случае понимается следующая процедура. Значения целевой переменной перемешиваются случайным образом (значения признака при этом не перемешиваются) и для полученной новой целевой переменной вычисляются новые оптимальная граница и значение функционала. Данный перерасчет повторяется некоторое количество раз (чем больше, тем достовернее оценка). Далее в качестве оценки вероятности (p-value) того, что нулевая гипотеза о независимости верна, используется доля перемешиваний, при которых новое оптимальное значение функционала оказалось не меньше, чем исходное.

В случае моделей II и III (модели с двумя граничными значениями) выполняется два перестановочных теста: по одному на каждое граничное значение. Суть перестановочного теста остается такой же, как и для модели I, с той лишь поправкой, что перемешивание значений целевой переменной происходит отдельно слева и справа от второй границы (той, которая не верифицируется).

Допустим, мы проверяем достоверность закономерности, которая получена с помощью модели III для пары признаков X_1 и X_2. Обозначим $ГЗ_1$ и $ГЗ_2$ соответствующие оптимальные граничные значения. Тогда перестановочный тест для верификации значимости разбиения по X_1 описывается следующим образом. Значения целевой переменной перемешиваются отдельно в областях «X_2 не меньше $ГЗ_2$» и «X_2 меньше $ГЗ_2$». Далее по перемешанным данным вычисляются новые оптимальное граничное значение признака X_1 и оптимальное значение функционала. Граничное значение переменной X_2 при этом никак не изменяется. Аналогично модели I, описанная процедура перерасчета повторяется несколько раз, и затем рассчитывается значение p-value – оценка вероятности того, что целевая переменная не зависит от X_1 (при фиксированной границе X_2). Повторив эту же процедуру для X_2, можно получить второе p-value – оценку вероятности того, что целевая переменная не зависит от X_2 (при

фиксированной границе X_1). В случае модели II процедура перестановочного теста аналогична.

Таким образом, при проверке достоверности закономерностей, полученных с помощью моделей II и III, рассчитываются два значения *p*-value: по одному на каждое граничное значение. Такой подход позволяет верифицировать обе границы и исключить из рассмотрения фиктивные закономерности, обусловленные только лишь наличием достоверного разбиения модели I.

В результате применения метода ОДР мы получаем для каждого признака/пар признаков:

– оптимальные (с точки зрения заданного функционала качества) граничные значения;

– оптимальное значение функционала качества – характеризует значимость найденной закономерности; чем больше данное значение, тем более значимой считается закономерность;

– значения *p*-value – характеризуют достоверность найденной закономерности; данную величину стоит понимать как оценку вероятности того, что целевая переменная не зависит от соответствующего признака (при фиксированной второй границе для моделей II и III), т.е. чем ближе данное значение к 0, тем более достоверной является закономерность.

Далее из найденных закономерностей отбираются наиболее значимые и достоверные, т.е. отбираются разбиения удовлетворяющие условию: значение функционала больше X_1 и/или *p*-value меньше X_2, где значения X_1 и X_2 задаются вручную.

Таким образом, на первом шаге с помощью метода оптимальных достоверных разбиений (ОДР) выявляются наиболее значимые достоверные закономерности (разбиения). Отметим, что метод ОДР позволяет эффективно отбирать признаки по величине статистики χ^2. В настоящей работе признаки отбирались с использованием простейшего одномерного разбиения метода ОДР по величине функционала χ^2. Наилучший результат получился при отборе признаков, для которых значение функционала больше 10. Такому набору признаков соответствовала наивысшая оценка точности в режиме хорошо известного скользящего контроля.

Следующим шагом является построение классифицирующей модели с помощью модифицированного метода статистически взвешенных синдромов (МСВС). В исходном методе статистически взвешенных синдромов [10] использовались границы, найденные с помощью одномерных разбиений. Позже был разработан модифицированный метод взвешенных синдромов (МСВС), использующий наряду с одномерными двумерные разбиения. Суть метода МСВС заключается в построении взвешенного голосования по наборам синдромов, где под синдромом понимается область пространства прогностических переменных с преобладанием объектов одного из классов целевой переменной. Пример одномерного синдрома: «Возраст не меньше 80», т.е. доли пациентов, у которых наблюдалось осложнение в первые полгода после выписки, в группах «старше 80 лет» и «моложе 80 лет» значительно отличаются. Пример двумерных синдромов: «Возраст не меньше 80, и СКФ не меньше 36» или «Возраст меньше 80, и СКФ не меньше 36» (это два разных синдрома, несмотря на то, что переменные используются одни и те же). В качестве синдромов используются области оптимальных достоверных разбиений отобранных признаков.

Подробное математическое обоснование методов ОДР и МСВС представлено в ранее опубликованных работах [21, 22].

2.4. Результаты применения описанного метода

Так как в качестве целевой переменной прогнозирования использовался бинарный признак: наличие или отсутствие (1 или 0) какого-либо осложнения в первые полгода после выписки, то для представления результата метода использовалась ROC-кривая (рис. 2). Это кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении [23]. ROC-кривая отображает зависимость чувствительности (доли верно классифицированных положительных случаев) от специфичности (доли неверно классифицированных отрицательных случаев) при варьировании порогового значения.

В качестве числовой характеристики оценивающей качество прогностической модели, использовался показатель AUC - площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0.5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). Данный показатель наиболее часто используется для оценки качества моделей и сравнения различных моделей между собой в случае применения бинарного классификатора [23].

В результате была построена прогностическая модель с предсказательной силой AUC = 0.72. Величина AUC была получена с применением метода скользящего контроля по 10 блокам (или, другими словами, методом 10-кратной перекрестной проверки (10-fold cross-validation) [24]).

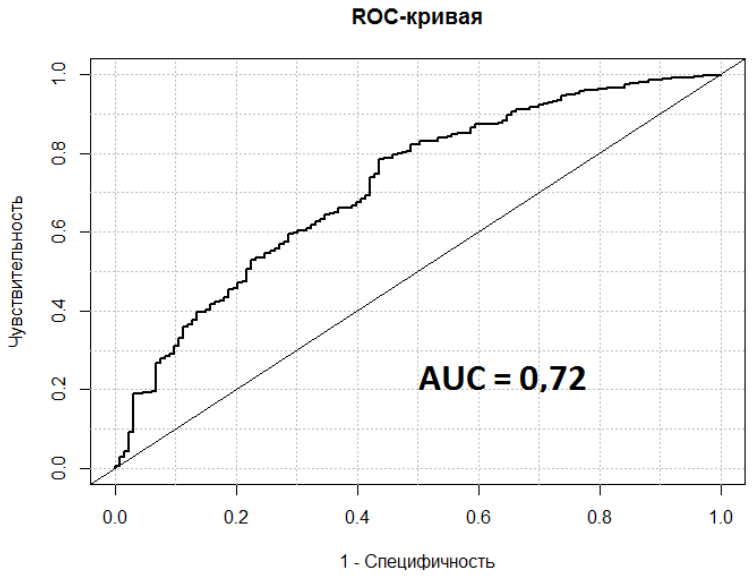

Рис. 2. ROC-кривая, полученная в результате применения описанной методики построения прогностической модели (с применением скользящего контроля по 10 блокам).

При использовании метода рассматривались разбиения классов I и III (см. рис. 1). Метод МСВС был реализован в виде отдельной программной системы. Для верификации закономерностей на первом шаге анализа, был использован вариант метода ОДР, реализованный в рамках программы «РАЗБИЕНИЯ» [21].

В методе СВС оценка распознавания за класс вычисляются в виде суммы вкладов всевозможных синдромов, к которым принадлежит распознаваемый объект. В случае большого числа признаков общее число синдромов может оказаться достаточно большим. В полученной модели общее число синдромов более 1000. Поэтому представление полученной модели мы предлагаем делать не в виде формулы (как,

например, в логистической регрессии), а в виде программы, которая будет оценивать вероятность наступления неблагоприятного исхода.

Наиболее значимые синдромы приведены в виде таблиц в приложениях к данной статье. В приложении 1 приведены наиболее значимые одномерные закономерности (значение функционала больше 15 и *p*-value меньше 0.025). Для каждого признака указаны:

– граничное значение – оптимальная граница, при которой значение функционала максимально. Для бинарных признаков – это всегда «1», для категориальных – одна из категорий (например, акинез);

– доля событий слева (меньше / «0») от граничного значения – доля пациентов, у которых в первые полгода после выписки наблюдалось осложнение, в группе пациентов, у которых значение соответствующего признака меньше граничного значения (в случае бинарного признака имеется в виду, что значение равно «0»; в случае категориального – равно любой другой категории, кроме граничного значения). В скобках указано общее количество пациентов, попавших в эту группу;

– доля событий справа (не меньше / «1») от граничного значения – аналогичный показатель для группы пациентов, у которых значение соответствующего признака не меньше граничного значения (в случае бинарного признака имеется в виду, что значение равно «1»; в случае категориального – равно граничному значению). В скобках также указано общее количество пациентов, попавших в эту группу;

– *p*-value – достоверность найденной закономерности, определенная с помощью перестановочного теста (количество перестановок 2000);

– значение функционала – значение статистики χ^2 при указанном граничном значении.

Так, например, по приведенным одномерным синдромам, можно заметить, что наличие у пациента перенесенного ИМ значительно повышает вероятность наступления неблагоприятного исхода в первые полгода после выписки. Также можно заметь, что шансы на возникновения рецидива повышаются при заниженной скорости клубочковой фильтрации (меньше 36.3 мл/мин/1.73 м^2).

Аналогично в приложении 2 приведены наиболее значимые двумерные закономерности (значение функционала больше 30 и *p*-value меньше 0.001). Из-за большого количества переменных перебор всех возможных пар и тестирование их с помощью перестановочного теста с 2000 перестановок представляет сложную задачу с вычислительной точки зрения. В связи с этим тестирование двумерных закономерностей осуществлялось в два этапа:

– На первом этапе с помощью теста с 200 перестановками были отобраны закономерности с достоверностью (максимальное из двух *p*-value) не хуже 0.1 по каждой из двух размерностей. В результате было отобрано 4521 пары.

– На втором этапе отобранные закономерности (4521) тестировались перестановочным тестом с 2000 перестановками.

Как можно заметить, состав значимых параметров достаточно разносторонний: он включает в себя клинические, генетические, биохимические показатели, информацию об образе жизни, предшествовавшем данной госпитализации, а также включает результаты ЭКГ и ЭхоКГ. Особенно следует отметить, что большое количество значимых закономерностей связано с результатами ЭхоКГ (боковая стенка верхушечный сегмент, переднебоковой средний сегмент, заднесредний сегмент и т.д.), что говорит о важности включения в рассмотрение подробной информации при оценивании рисков возникновения неблагоприятных исходов у больных, перенесших обострение ИБС. Также ниже (рис. 3-5) приведены наиболее наглядные иллюстрации результатов работы программы «РАЗБИЕНИЯ».

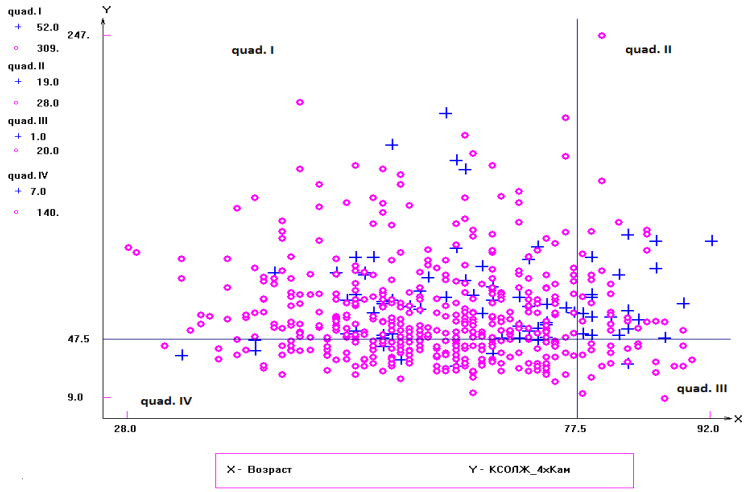

Рис. 3. Диаграмма рассеяния, отражающая связь возраста (ось *X*) и конечно-систолического объема левого желудочка (ось *Y*) с наличием («o») или отсутствием («+») у пациентов осложнений в первые полгода после выписки.

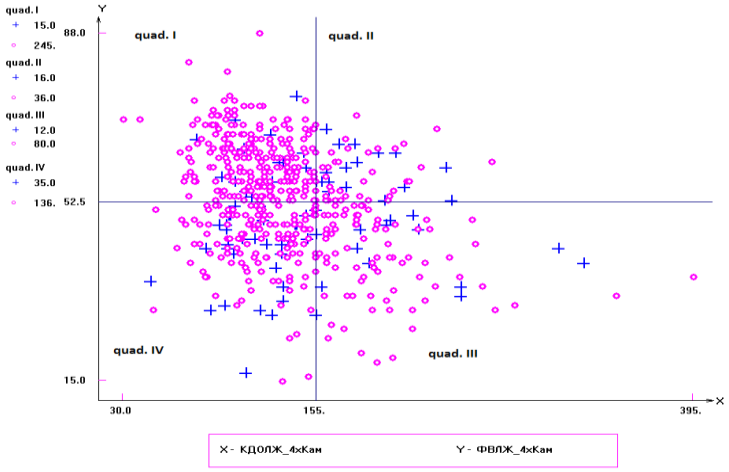

Рис. 4. Диаграмма рассеяния, отражающая связь конечно-диастолического объема ЛЖ (ось *X*) и фракций выброса ЛЖ (ось *Y*) с наличием («o») или отсутствием («+») у пациентов осложнений в первые полгода после выписки.

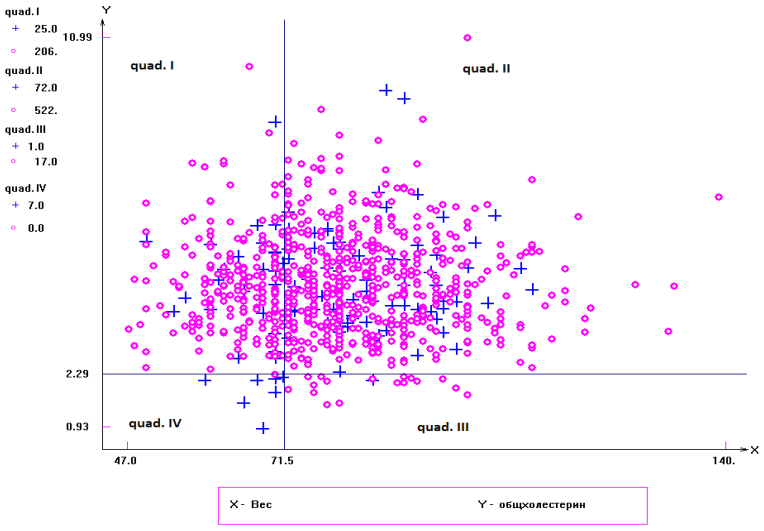

Рис. 5. Диаграмма рассеяния, отражающая связь веса (ось *X*) и холестерина (ось *Y*) с наличием («o») или отсутствием («+») у пациентов осложнений в первые полгода после выписки.

2.5. Сравнение с другими методами

Также в рамках данной работы было проведено сравнение предсказательной силы полученной модели с:

– моделями, полученными наиболее распространенными методами [13]: логистическая регрессия, дискриминантный анализ, метод опорных векторов, деревья решений;

– одной из наиболее часто используемых систем стратификации риска – GRACE [5, 6].

Предварительная обработка данных и реализация методов логистической регрессии, дискриминантного анализа, опорных векторов, деревья решений, нейронных сетей, байесовских сетей доверия проводили с использованием следующих программных пакетов IBM SPSS Statistics [25], IBM SPSS Modeler [26].

Следует отметить, что в этом случае предварительная подготовка анализируемой базы данных проводилась в соответствии с требованиями получения корректного результата для выбранного метода. Также для каждого метода проводился перебор значений основных параметров (полный список основных параметров для каждого метода приведен в [27]). В сравнении участвовала та конфигурация метода, прогностическая способность (по значению AUC) которой оказалась наибольшей.

Определение индекса GRACE для оценки риска полугодовой смертности совместно с инфарктом миокарда после выписки из стационара, проводили на основании следующих параметров: ЧСС, САД, возраст больных, наличие депрессии ST на ЭКГ, повышение уровня кардиоспецифических ферментов, наличие инфаркта миокарда и сердечной недостаточности в анамнезе, уровень креатинина крови и проведение вмешательства на коронарных артериях в связи с данным эпизодом ОКС.

В качестве оценки предсказательной способности модели использовали величину AUC для ROC кривой, построенной с применением скользящего контроля по 10 блокам. Результат сравнения перечисленных выше моделей приведен в таблице 4.

Таблица 4. Результаты сравнения полученной модели с другими методами

Метод	AUC
МСВС	0.72
Деревья решений	0.64
GRACE	0.63
Дискриминантный анализ	0.63
Логистическая регрессия	0.56
Метод опорных векторов	0.56
Нейронные сети	0.55

Как видно по результатам сравнения, модель, полученная с помощью оптимальных разбиений, обладает набольшей предсказательной силой. Также, принимая во внимание успешный опыт применения в ряде других клинических исследований [7–12], можно сказать, что подходы, основанные на оптимальных разбиениях – это один из наиболее подходящих методов многопараметрического анализа данных в клинических исследованиях.

3. ЗАКЛЮЧЕНИЕ

В результате исследования при участии 16 центров в семи различных городах России в период с 2004 по 2010 годы собрана база данных, содержащая информацию о 1193 пациентах. Полученные данные содержат широкий спектр признаков: анамнез,

физиологические характеристики пациента при поступлении и выписке, информацию о лекарствах назначенных больному в ходе лечения и на момент выписки, данные электрокардиограммы (ЭКГ) и ЭхоКГ, особенности стиля жизни, информацию о родственниках и генетических маркерах больного, а также состояние больного после выписки (всего 407 параметров).

На базе собранной информации, с помощью подходов, основанных на оптимальных разбиениях (методов ОДР и МСВС) получена прогностическая модель, оценивающая вероятность наступления неблагоприятного исхода в первые полгода после выписки. Где под неблагоприятным исходом понимается наступление любого сердечно-сосудистого события: фатальный / нефатальный ИМ, фатальный / нефатальный инсульт, нестабильная стенокардия, атеросклероз.

Качество полученной модели было оценено с помощью ROC-анализа: показатель AUC, полученный методом скользящего контроля по 10 блокам, составил 0.72. Достигнутая величина AUC представляется весьма значимой. Полученная модель сравнена с моделями, полученными другими широко распространенными методами [13], такими как логистическая регрессия, деревья решений, нейронные сети, дискриминантный анализ, метод опорных векторов, а также и с одной из наиболее часто используемых систем стратификации риска – GRACE [5, 6]. По результатам сравнения (табл. 4) видно, что модель, полученная методом МСВС, обладает наибольшей предсказательной силой.

В перспективе на основе построенной модели возможна реализация программы-калькулятора (аналогичной калькулятору GRACE 2.0 [28]) для автоматизированного расчета вероятности наступления неблагоприятного исхода в первые полгода после выписки. Использование подобного калькулятора на практике поможет врачам в оптимизации терапии больных.

Помимо построения прогностической модели, использованный подход, позволяет выделить наиболее значимые с точки зрения прогнозирования параметры. За счет данного свойства в работе также показана (табл. 2, 3) важность учета данных ЭхоКГ при оценке риска возникновения неблагоприятного исхода в первые полгода после выписки.

Также, принимая во внимание то, что данный метод обладает рядом свойств особенно полезных в рутинной клинической практике (нечувствительность к пропускам и аномальным значениям в данных; практически не требует предварительной обработки данных), можно предположить, что данный метод является потенциально наиболее подходящим инструментом для разработки тактики ведения таких больных.

Использование подходов, основанных на оптимальных разбиениях, открывает дополнительные перспективы для построения прогностических моделей на основе данных клинических исследований, содержащих разнообразную медицинскую, биохимическую, генетическую и др. информацию.

ПРИЛОЖЕНИЕ 1

Здесь приведены наиболее значимые одномерные закономерности (значение функционала больше 15 и *p*-value меньше 0.025). Для каждого признака указаны:

– граничное значение – оптимальная граница, при которой значение функционала максимально. Для бинарных признаков – это всегда «1», для категориальных – одна из категорий (например, акинез);

– доля событий слева (меньше / «0») от граничного значения – доля пациентов, у которых в первые полгода после выписки наблюдалось осложнение, в группе пациентов, у которых значение соответствующего признака меньше граничного значения (в случае бинарного признака имеется в виду, что значение равно «0»; в случае категориального – равно любой другой категории, кроме граничного значения). В скобках указано общее количество пациентов, попавших в эту группу;

– доля событий справа (не меньше / «1») от граничного значения – аналогичный показатель для группы пациентов, у которых значение соответствующего признака не меньше граничного значения (в случае бинарного признака имеется в виду, что значение равно «1»; в случае категориального – равно граничному значению). В скобках также указано общее количество пациентов, попавших в эту группу;

– *p*-value – достоверность найденной закономерности, определенная с помощью перестановочного теста (количество перестановок 2000);

– значение функционала – значение статистики χ^2 при указанном граничном значении.

Таблица 2. Наиболее значимые одномерные закономерности (значение функционала больше 15 и *p*-value меньше 0.025)

Признак	Граничное значение	Доля событий слева (меньше/нет) от граничного значения	Доля событий справа (не меньше/да) от граничного значения	*p*-value	Значение функционала
1	2	3	4	5	6
ИМ	Да	8% (782)	19% (366)	0.000	29.8
Прием гиполипидемических препаратов до госпитализации	Да	10% (1027)	25% (113)	0.000	21.3
Липопротеины низкой плотности	0.76	80% (5)	12% (850)	0.011	21.1
Скорость клубочковой фильтрации (формула Кокрофта-Гаулта)	36.31	35% (43)	11% (797)	0.003	21.1
КДР ЛЖ 4-х камерная	68.5	12% (805)	50% (16)	0.004	19.7
Возраст	81.5	11% (1107)	33% (40)	0.008	17.7
Боковая стенка средний сегмент	Акинез	11% (1129)	42% (19)	0.001	17.6
Пик А	0.28	55% (11)	12% (509)	0.008	17.3
Боковой верхушечный сегмент	Дискинез	11% (1139)	56% (9)	0.002	17.1

Продолжение таблицы 2. Наиболее значимые одномерные закономерности (значение функционала больше 15 и *p*-value меньше 0.025)

1	2	3	4	5	6
Применение других диуретиков на 10-й день	Да	10% (940)	20% (207)	0.000	17.1
КСР ЛЖ 4-х камерная	39.5	9% (512)	19% (291)	0.004	16.8
Креатинин	140.5	11% (776)	27% (79)	0.004	16.1
КСО ЛЖ 4-х камерная	47.5	5% (168)	17% (409)	0.010	16.0
Инвалидность до госпитализации	1 или 2 группа	9% (708)	16% (440)	0.001	15.9

ПРИЛОЖЕНИЕ 2

Здесь приведены наиболее значимые двумерные закономерности (значение функционала больше 30 и p-value меньше 0.001). Для каждой пары признаков указаны:

– $ГЗ_1$ – граничное значение первого признака ($П_1$);

– $ГЗ_2$ – граничное значение второго признака ($П_2$);

– Доля пациентов, у которых в первые полгода после выписки наблюдалось осложнение, в каждой из четырех областей, получаемых при разбиении признакового пространства соответствующими граничными значениями (см. модель III на рис. 1).

Таблица 3. Наиболее значимые двумерные закономерности (значение функционала больше 30 и *p*-value меньше 0.001)

$П_1$	$ГЗ_1$	$П_2$	$ГЗ_2$	$П_2$ меньше $ГЗ_2$		$П_2$ не меньше $ГЗ_2$	
				$П_1$ меньше $ГЗ_1$	$П_1$ не меньше $ГЗ_1$	$П_1$ меньше $ГЗ_1$	$П_1$ не меньше $ГЗ_1$
1	2	3	4	5	6	7	8
ИМ	Да	Боковая стенка средний сегмент	Акинез	8% (771)	18% (358)	18% (11)	75% (8)
ИМ	Да	Боковой верхушечный сегмент	Дискинез	8% (778)	18% (361)	25% (4)	80% (5)
Полиморфизм гена ФНО-308 (TNF-308)	AG	Боковая стенка средний сегмент	Акинез	9% (761)	15% (368)	17% (12)	86% (7)
Наличие документированных эпизодов нестабильной стенокардии	Да	Боковая стенка средний сегмент	Акинез	10% (947)	18% (182)	27% (15)	100% (4)
Стенокардия	Да	Боковая стенка средний сегмент	Акинез	7% (390)	13% (739)	11% (9)	70% (10)
Боковая стенка базальный сегмент	Гипокинез	Боковая стенка средний сегмент	Акинез	10% (1061)	25% (68)	53% (15)	0% (4)
Боковая стенка базальный сегмент	Гипокинез	Боковая стенка средний сегмент	Акинез	10% (1061)	25% (68)	53% (15)	0% (4)
Заднебазальный сегмент	Дискинез	Характер приступа при поступлении	Болевой	100% (5)	12% (631)	0% (7)	11% (505)
Гипертонические кризы в анамнезе	Да	Боковая стенка верхушечный сегмент	Дискинез	11% (776)	11% (361)	0% (6)	100% (5)
Физическая активность до госпитализации	Средняя или высокая	КСР ЛЖ 4-х камерная	39.5	11% (180)	8% (331)	32% (101)	13% (190)
Применение других диуретиков на 10-й день заболевания	Да	Боковой верхушечный сегмент	Дискинез	9% (932)	19% (206)	50% (8)	100% (1)

Продолжение таблицы 3. Наиболее значимые двумерные закономерности (значение функционала больше 30 и *p*-value меньше 0.001)

Окружность талии	73.5	Применение антиагрегантов в течение госпитализации	Да	100% (4)	0% (37)	11% (47)	12% (1027)
Инвалидность до госпитализации	1 или 2 группа	Переднеперегородочный средний сегмент	Норма	8% (479)	12% (289)	10% (229)	25% (151)
Стенокардия до госпитализации	Да	Боковая стенка верхушечный сегмент	Дискинез	7% (395)	13% (742)	0% (4)	71% (7)
Элевация ST при поступлении	Да	Переднебоковой средний сегмент	Акинез	12% (668)	10% (451)	83% (6)	0% (14)
ИБС	Да	Переднеперегородочный средний сегмент	Норма	7% (235)	10% (533)	7% (136)	22% (244)
Боковая стенка средний сегмент	Норма	Переднеперегородочный средний сегмент	Норма	9% (602)	8% (166)	34% (59)	13% (321)
Элевация ST при поступлении	Да	Заднебазальный сегмент	Дискинез	13% (670)	9% (458)	100% (4)	14% (7)
Наличие документированных эпизодов нестабильной стенокардии	Да	Боковой верхушечный сегмент	Дискинез	10% (955)	18% (184)	43% (7)	100% (2)
Стенокардия	Да	Переднеперегородочный средний сегмент	Норма	7% (253)	10% (515)	8% (146)	22% (234)
Наличие документированных эпизодов нестабильной стенокардии	Да	Задний базальный сегмент	Дискинез	10% (954)	18% (184)	38% (8)	100% (2)
Боковая стенка средний сегмент	Акинез	Задний средний сегмент	Гипокинез	10% (982)	44% (18)	20% (147)	0% (1)
Полиморфизм гена Лимфотоксина-альфа C804A (LTA)	CC	Задний средний сегмент	Гипокинез	13% (605)	7% (395)	13% (101)	34% (47)
Боковая стенка средний сегмент	Акинез	Заднесредний сегмент	Гипокинез	10% (960)	42% (19)	20% (169)	– (0)
Наличие документированных эпизодов нестабильной стенокардии	Да	Передний верхушечный сегмент	Норма	9% (633)	14% (136)	13% (329)	34% (50)
Боковая стенка средний сегмент	Норма	МЖП средний сегмент	Норма	9% (596)	11% (170)	32% (65)	12% (317)

.

СПИСОК ЛИТЕРАТУРЫ

1. World Health Organization. *The top 10 causes of death*. URL: http://www.who.int/mediacentre/factsheets/fs310/ (дата обращения: 03.02.2016).
2. Antman E.M., Cohen M., Bernink P.J., McCabe C.H., Horacek T., Papuchis G., Mautner B., Corbalan R., Radley D., Braunwald E. The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making. *Journal of the American Medical Association.* 2000. V. 284. № 7. P. 835–842. doi: 10.1001/jama.284.7.835.
3. Pollack C.V. Jr., Sites F.D., Shofer F.S., Sease K.L., Hollander J.E. Application of the TIMI Risk Score for Unstable Angina and Non-ST Elevation Acute Coronary Syndrome to an Unselected Emergency Department Chest Pain Population. *Academic Emergency Medicine.* 2006. V. 13. № 1. P 13–18. doi: 10.1197/j.aem.2005.06.031.
4. Boersma E., Pieper K.S., Steyerberg E.W., Wilcox R.G., Chang W., Lee K.L., Akkerhuis K.M., Harrington R.A., Deckers J.W., Armstrong P.W. et al. Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. *Circulation.* 2000. V. 101. № 22. P. 2557–2567. doi: 10.1161/01.CIR.101.22.2557.
5. Granger C.B., Goldberg R.J., Dabbous O., Pieper K.S., Eagle K.A., Cannon C.P., Van de Werf F., Avezum A., Goodman S.G., Flather M.D. et al. Predictors of hospital mortality in the global registry of acute coronary events. *Archives of Internal Medicine.* 2003. V. 163. № 19. P. 2345–2353. doi: 10.1001/archinte.163.19.2345.
6. Eagle K.A., Lim M.J., Dabbous O.H., Pieper K.S., Goldberg R.J., Van de Werf F., Goodman S.G., Granger C.B., Steg P.G., Gore J.M. et al. A validated prediction model for all forms of acute coronary syndrome. Estimating the risk of 6-month postdischarge death in an international registry. *Journal of the American Medical Association.* 2004. V. 291. № 22. P. 2727–2733. doi: 10.1001/jama.291.22.2727.
7. Senko O.V., Kuznetsova A.V. A recognition method based on collective decision making using systems of regularities of various types.*Pattern Recognition and Image Analysis*. 2010. V. 20. № 2. P. 152–162. doi: 10.1134/S1054661810020069.
8. Kuznetsova A.V., Kostomarova I.V., Senko O.V. Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. *Pattern Recognition and Image Analysis*. 2014. V. 24. № 1. P. 114–123. doi: 10.1134/S105466181401009X.
9. Senko O.V., Kuznetsova A.V. The Optimal Valid Partitioning Procedures. *InterStat.* 2006. April. № 2.
10. Кузнецов В.А., Сенько О.В., Кузнецова А.В., Семенова Л.П., Алещенко А.В., Гладышева Т.Б., Ившина А.В.. Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической нормы и хронической патологии. *Химическая физика*. 1996. Т. 15. № 1. С. 81–100.
11. Ivshina A.V., George J., Senko O.V., Mow B., Putti T.C., Smeds J., Lindahl T., Pawitan Y., Hall P., Nordgren H., Wong J.E.L., Liu E.T., Bergh J., Kuznetsov V.A., Miller L.D. Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer. *Cancer Research*. 2006. V. 66. № 21. P. 10292–10301. doi: 10.1158/0008-5472.CAN-05-4414.
12. Заковряшин А.С., Заковряшина С.Е., Доровских И.В., Сенько О.В., Кузнецова А.В., Козлов А.А. Прогнозирование отдаленных последствий психогенных расстройств у военнослужащих в остром периоде боевой психической травмы (с использованием логико – статистических методов). *Журнал неврологии и психиатрии имени С.С. Корсакова.* 2006. Т. 106. № 3. С. 31–38.

13. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning*: *Data Mining, Inference, and Prediction.* New York: Springer, 2009. 533 p.
14. Затейщиков Д.А., Волкова Э.Г., Гузь И.О., Евдокимова М.А., Асейчева О.Ю., Галявич А.С., Терещенко С.Н., Казиолова Н.А., Глезер М.Г., Ягода А.В. и др. Лечение больных, перенесших острый коронарный синдром, по данным российского многоцентрового проспективного наблюдательного исследования. *Фарматека.* 2009. № 12. С. 109-113.
15. Чумакова О.С., Селезнева Н.Д., Евдокимова М.А., Осмоловская В.С., Кочкина М.С., Асейчева О.Ю., Минушкина Л.О., Бакланова Т.Н., Талызин П.А., Терещенко С.Н. и др. Прогностическое значение аортального стеноза у больных, перенесших обострение ишемической болезни сердца. *Кардиология.* 2011. № 1. С. 23–28.
16. Благодатских К.А., Евдокимова М.А., Агапкина Ю.В., Никитин А.Г., Бровкин А.Н., Пушков А.А., Благодатских Е.Г., Кудряшова О.Ю., Осмоловская В.С., Минушкина Л.О. и др. Полиморфные маркеры G(-174)C гена *IL*6 и G(-1082)A гена *IL*10 и генетическая предрасположенность к неблагоприятному течению ишемической болезни сердца у больных, перенесших острый коронарный синдром. *Молекулярная биология.* 2010. Т. 44. № 5. С. 839–846.
17. Благодатских К.А., Никитин А.Г., Пушков А.А., Благодатских Е.Г., Осмоловская В.С., Асейчева О.Ю., Бакланова Т.Н., Талызин П.А., Терещенко С.Н., Джаиани Н.А. и др. Полиморфные маркеры G2667C, G3014A, C3872T, A5237G гена CRP и генетическая предрасположенность к неблагоприятному течению ишемической болезни сердца у больных, перенесших обострение ишемической болезни сердца. *Медицинская генетика.* 2011. Т. 10. № 4. С. 3–9.
18. BMI Classification. *World Health Organization: Global Database on Body Mass Index.* URL: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html (дата обращения: 03.02.2016).
19. Cockroft D.W., Gault M.H. Prediction of creatinine clearance from serum creatinine. *Nephron*. 1976. V. 16. № 1. P. 31–41.
20. Wing R.R., Matthews K.A., Kuller L.H., Meilahn E.N., Plantinga P. Waist to hip ratio in middle-aged women. Associations with behavioral and psychosocial factors and with changes in cardiovascular risk factors. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 1991. V. 11. № 5. P. 1250–1257. doi: 10.1161/01.ATV.11.5.1250.
21. Кузнецова А.В., Костомарова И.В., Сенько О.В. Логико-статистический анализ связи клинико-лабораторных показателей с возникновением нарушения мозгового кровообращения у пациентов пожилого возраста с хронической ишемией головного мозга. *Математическая биология и биоинформатика.* 2013. Т. 8. № 1. С. 182–224. doi: 10.17537/2013.8.182.
22. Кузнецова А.В., Костомарова И.В., Водолагина Н.Н., Малыгина Н.А., Сенько О.В. Изучение влияния клинико-генетических факторов на течение дисциркуляторной энцефалопатии с использованием методов распознавания. *Математическая биология и биоинформатика.* 2011. Т. 6. № 1. С. 115–146. doi: 10.17537/2011.6.115.
23. Паклин Н. Логистическая регрессия и ROC-анализ – математический аппарат. *BaseGroup Labs: Технологии анализа данных*. URL: https://basegroup.ru/community/articles/logistic (дата обращения: 03.02.2016).
24. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов. *Математические вопросы кибернетики.* 2004. Т. 13. С. 5–36.
25. *IBM SPSS Statistics 23 Documentation.* URL: http://www-01.ibm.com/support/docview.wss?uid=swg27043946 (дата обращения: 03.02.2016).
26. *IBM SPSS Modeler 17.0 Documentation*. URL: http://www-01.ibm.com/support/docview.wss?uid=swg27043831 (дата обращения: 03.02.2016).

27. *Узлы моделирования IBM SPSS Modeler 17*. URL: http://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.0/ru/ModelerModelingNodes.pdf (дата обращения: 03.02.2016).
28. *GRACE 2.0 ACS Risk Calculator*. URL: http://www.gracescore.org/WebSite/WebVersion.aspx (дата обращения: 03.02.2016).

Рукопись поступила в редакцию 14.01.2016, переработанный вариант поступил 08.02.2016.
Дата опубликования 23.03.2016