Математическая биология и биоинформатика
2018. Т. 13. № 2. С. 563–590. doi: 10.17537/2018.13.563

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

УДК: 004.852

Метод оптимальных разбиений для оценки влияния степени оксигенации гемоглобина на фактор роста эндотелия сосудов

Сенько О.В.*[1], Кодрян М.С.†[2], Кузнецова А.В.‡[3], Клименко Л.Л. §[4], Деев А.И.[4], Баскаков И.С.[4], Мазилина А.Н.[5]

[1]*Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия*
[2]*Московский государственный университет им. М.В.Ломоносова, Москва, Россия*
[3]*Институт биохимической физики им. Н.М. Эмануэля, Москва, Россия*
[4]*Институт химической физики им. Н.Н. Семенова, Москва, Россия*
[5]*Отделение неврологии КБ № 123 ФМБА России*

Аннотация. Целью работы является исследование связи фактора роста эндотелия сосудов в сыворотке крови и гипоксией в группах пациентов, страдающих тяжелыми неврологическими заболеваниями. Имеющиеся в литературе данные свидетельствуют об активации синтеза этого белка при гипоксии. Вместе с тем, стандартный корреляционный анализ не позволил достоверно выявить по клиническим данным наличие связи между уровнем фактора роста эндотелия сосудов и параметрами оксиметрии, объективно характеризующими снабжение организма кислородом. В статье представлена методика, позволившая статистически достоверно доказать существование указанной связи, сводящейся к увеличении корреляции между VEGF и комплементом C4, а также между VEGF и белками S100 при снижении уровня оксигенации ниже некоторого порогового значения. Методика включает поиск таких границ для показателей оксиметрии, при которых достигаются максимальные различия по уровню корреляции между VEGF и некоторым дополнительным фактором Z в образуемых группах, а также по возможности максимальный уровень корреляции между VEGF и Z в одной из групп. Использовалась оригинальная методика статистической верификации, основанная на использовании перестановочных тестов, которая позволила не только установить статистическую достоверность эффектов, связанных с отдельными показателями оксиметрии, но и рассчитать совокупную значимость по всем таким показателям. В связи с исследованием возможности использования в качестве дополнительных факторов Z большого числа показателей, потребовалась коррекция с целью учета множественного тестирования.

Ключевые слова: *фактор роста эндотелия сосудов, гипоксия, верификация, перестановочный тест, множественное тестирование.*

*senkoov@mail.ru
†max-kodr@rambler.ru
‡azfor@yandex.ru
§klimenkoll@mail.ru

ВВЕДЕНИЕ

В настоящее время существует достаточно большое количество работ, анализирующих связь уровня оксигенации крови с ростом и регрессией сосудов. В работе [1] представлена общая схема влияния различных факторов на процесс ангиогенеза. К снижению тканевой оксигенации и увеличению ангиогенеза приводят: рост тканей, физические упражнения, гипертиреоидизм, повреждение сосудов, пониженное содержание атмосферного кислорода; в то же время гиподинамия и повышенное содержание атмосферного кислорода приводят к обратному эффекту – регрессии ангиогенеза.

Ключевым фактором, регулирующим процесс ангиогенеза, признан фактора роста эндотелия сосудов (Vascular Endothelial Growth Factor – VEGF), который представляет собой гомодимерный гликопротеин с молекулярной массой 45 кДа, содержащий 26 аминокислот. Известно, что VEGF выполняет функцию поддержания гомеостаза эндотелиального барьера, разделяющего кровь и ткани, влияет на развитие коллатералей и устранение ишемии в органах, а также обладает васкулопротективными свойствами [2]. Имеются данные, что VEGF является не только фактором ангиогенеза, но также участвует в нейрогенезе и нейропротекции [3].

В норме VEGF содержится в тканях в незначительном количестве (10–246 пг/мл), но экспрессия его гена значительно активируется при гипоксии, доходя в нашем исследовании до крайне высокого значения 3175.894 пг/мл. Примером аналогичного эффекта является корреляция уровня VEGF в сыворотке крови со встречаемостью телеангиэктазии у жителей высокогорья [4]. Тканевая гипоксия вызывает активацию генов семейства HIF-1 или гипоксия-индуцибельного фактора (hypoxia-inducible factor) [5]. Активация гена HIF-1 происходит в физиологически важных местах регуляции кислородных путей, обеспечивая быстрые и адекватные ответы на гипоксический стресс, в первую очередь – ответ генов, регулирующих процесс ангиогенеза и способствующих образованию VEGF, продуцируемого различными типами клеток – макрофагами, фибробластами, лимфоцитами, полиморфноядерными клетками и т.д. [5]. Следует отметить, что, с одной стороны, клетки иммунной системы способны секретировать VEGF и регулировать процессы ангиогенеза, а с другой стороны они имеют специфические рецепторы для распознавания VEGF и сами могут подвергаться его действию, например, при росте опухолей. Таким образом, VEGF является одним из центральных факторов физиологической иммунорегуляции.

Несмотря на сложившееся понимание, что VEGF является важнейшим инструментом физиологической регуляции, в настоящее время существует ограниченное число работ, в которых влияние гипоксии и других биологических факторов на активацию синтеза VEGF оценивается по клиническим данным, что может быть связано со сложным и многоуровневым механизмом регуляции, делающих выявление статистически достоверных эффектов по отдельным переменным затруднительным. Например, стандартный корреляционный анализ не позволил достоверно выявить по анализируемым в настоящей работе клиническим данным наличие связи между уровнем VEGF и показателями оксиметрии, объективно характеризующими снабжение организма кислородом. Однако, как это показывается далее, успех при поиске закономерностей, связанных с регуляцией уровня VEGF, может быть достигнут при учете взаимодействия различных показателей. Поиск таких закономерностей затрудняет сложный и существенно нелинейный характер взаимодействия.

В этих условиях исследование процесса регуляции синтеза VEGF как реакцию на гипоксию, целесообразно проводить через поиск математических моделей, связывающих уровень VEGF с показателями оксиметрии в сочетании с некоторым

дополнительным фактором, принадлежащим набору разнообразных клинических, биохимических или инструментальных показателей, присутствующих в базе данных. При этом модели строятся в соответствии с возникающими предположениями о характере зависимостей. Важнейшей составляющей исследования является верификация выявленных эмпирических закономерностей, которая должна включать невозможность простого объяснения последних только одномерными эффектами. Верификация закономерностей осложняется также несостоятельностью гипотез о нормальности распределений для большинства биомедицинских показателей. Для верификации моделей, целесообразно использовать перестановочные тесты, не требующих априорных предположений о распределениях и являющихся универсальным инструментом статистической верификации.

ЦЕЛИ, МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЙ

Цели

Целью исследования является изучение механизма образования VEGF в связи с совокупностью комплекса показателей оксиметрии. Показатели оксиметрии:
1) индекс сатурации sO2, т.е. доля гемоглобина, связанного с кислородом (%);
2) pO2 – парциальное напряжение кислорода в артериальной крови (мм рт ст.);
3) pCO2 – парциальное давление углекислого газа в крови (мм рт ст.);
4) FO2Hb – фракция оксигемоглобина (%);
5) FCOHb – фракция карбоксигемоглобина (%);
6) FMetHb – фракция метгемоглобина (%);
7) FHHb – фракция дезоксигемоглобина (%).

Итак, предположение о существовании такой связи основано на биологической необходимости компенсации эффекта гипоксии, связанной с нарушением кислород-транспортирующей функции гемоглобина. Такая компенсация может быть достигнута через повышение уровня васкуляризации, которое стимулируется увеличением концентрации VEGF.

Данные

Статистические исследования проводились на основе базы данных, содержащей значения клинических, лабораторных и инструментальных показателей для 88 пациентов с возрастом от 33 до 88 лет, страдающих неврологическими заболеваниями: острое нарушение мозгового кровообращения (ишемический инсульт) и транзиторная ишемическая атака. База данных содержит значения 146 показателей, включая концентрацию VEGF (пг/мл), перечисленные выше параметры оксиметрии, количественные значения биохимических показателей, показатели медленной электрической активности коры головного мозга – уровень постоянного потенциала, (УПП, мВ), концентрация микроэлементов в сыворотке крови (мкг/г).

Исследование проведено в клинической больнице № 123 ФМБА России. Определение концентрации биохимических, гематологических и иммунологических показателей выполнялось в клинико-диагностической лаборатории на автоматическом биохимическом анализаторе «RX Imola» фирмы «Randox» (Великобритания), автоматическом биохимическом анализаторе «Сапфир-400», автоматическом гематологическом анализаторе «Medonic MC-15», «МЕК 7222», автоматическом иммуноферментном анализаторе «Лазурит» фирмы «Вектор-бест» (Россия), с использованием реагентов фирмы «Randox» (Великобритания), «CORMAY» (Польша), фирмы «Юнимед» (Россия), фирмы «Вектор-бест» (Россия) соответственно. Пробы крови брали свободным истечением из локтевой вены утром натощак через 12–14 часов

после приема пищи.

Для определения концентрации нейроспецифических белков VEGF (пг/мл) и S100 (нг/л), а также белков сывороточного комплемента С3 и С4 (г/л) с помощью прибора «Лазурит» использовали метод иммуноферментного анализа ELISA (enzyme linked immunosorbent assay). Пусть g – число анализируемых показателей оксиметрии, n – число показателей в базе, отличных от VEGF и показателей оксиметрии, которые могут оказаться потенциальными дополнительными факторами. Далее содержания VEGF будет обозначаться через Y, показателей оксиметрии – через $U_1, \ldots, U_g$, остальные показатели, вошедшие в базу и являющиеся потенциальными дополнительными факторами, будут обозначаться через $Z_1, \ldots, Z_n$. Анализируемая выборка может быть представлена в виде $\{(y_1, \mathbf{u}_1, \mathbf{z}_1, \ldots, (y_m, \mathbf{u}_m, \mathbf{z}_m)\}$, где y_j – значение переменной Y, $\mathbf{u}_j$ – вектор значений переменных $U_1, \ldots, U_g$, $\mathbf{z}_j$ – вектор переменных $Z_1, \ldots, Z_n$, $j = 1, \ldots, m$.

Методы анализа данных

Результаты использования стандартного корреляционного анализа представлены в таблице (табл. 1). Как видно из таблицы значения коэффициентов корреляции не позволяют сделать вывод о существовании линейной связи между VEGF и показателями оксиметрии. Однако отсутствие линейной связи не обязательно противоречит существованию более сложных нелинейных закономерностей, в которых связь уровня VEGF с показателями оксиметрии проявляется в сочетании с дополнительным фактором, представленным в базе данных. Для поиска закономерностей на первом этапе был использован метод оптимальных достоверных разбиений (ОДР) [16] с использованием в качестве целевой переменной бинарного показателя уровня VEGF, аналогичного бинарному показателю, используемому в работе [12]. Однако технология, используемая в работе [12] не может быть без изменений использована в настоящем исследовании. Последнее связано с необходимостью проводить анализ различий в группах с высоким и низким уровнями VEGF с одновременным учетом нескольких показателей оксиметрии. Такой учет может быть достигнут путем комбинирования результатов верификации различий между группами, выявляемых двумерными моделями метода ОДР по отдельным показателям оксиметрии и дополнительному фактору. Для оценивания интегральной значимости по группам показателей принято использовать технологии непараметрических комбинаций или NPC технологии, подробно обсуждаемой в работе [9].

В ходе поиска сложных закономерностей для каждого из показателей, вошедших в базу данных, последовательно проверяется существование сложной закономерности, в которой этот показатель является дополнительным фактором. Вероятность случайного возникновения конфигурации данных, для которых нулевые гипотезы об отсутствии закономерностей формально оказываются отвергнутыми, возрастает при росте числа проверяемых показателей. Ошибочное опровержение нулевых гипотез просто в результате большого числа проводимых проверок носит название эффекта множественного тестирования (ЭМТ). Для гарантирования достоверности результатов применения тестов необходимо проводить дополнительную коррекцию p-значений, получаемых после применения тестов, верифицирующих отдельные закономерности. В настоящее время существуют ряд технологий, позволяющих проводить коррекцию точнее, чем это позволяло использование простых и весьма консервативных оценок по методу Бонферрони [7].

Одним из подходов является использование рандомизированных перестановочных тестов [13, 10, 14, 15, 11]. В работе [12] представлен способ коррекции достоверности закономерностей, найденных с помощью метода ОДР. Метод основан на сравнении

Таблица 1. Коэффициенты корреляции между показателями оксиметрии и содержанием VEGF

показатель	коэффициент корреляции	значимость
sO2	–0.158	не знач.
pO2	–0.059	не знач.
pCO2	–0.035	не знач.
pCOHb	–0.058	не знач.
FO2Hb	–0.146	не знач.
FMetHb	0.13	не знач.
FHHb	0.196	не знач.

достоверности закономерностей, найденных в исходной выборке с достоверностью закономерностей, найденных в наборе случайных выборок, полученных из исходной выборки с помощью случайных перестановок позиций целевой переменной при фиксированных позициях векторов объясняющих переменных. Такой метод коррекции в отличие от наиболее распространенного подхода [13, 10, 11], основанного на прямом сравнении величин статистик критериев в исходной и случайных выборках, не требует одинаковой распределенности таких статистик. Однако, он требует применения перестановочных тестов также для оценки достоверности закономерностей, найденных в генерируемых случайных выборках, что приводит к огромным объемам вычислений. Отметим, что метод, основанный на повторном использовании перестановочных тестов, может быть использован и для оценивания интегральной достоверности по группам показателей. Последняя задача при прямом применении подхода, разработанного в [12], также оказывается весьма трудоемкой.

В приложении представлена модификация метода ОДР, позволившая использовать значительно менее трудоемкие процедуры оценок интегральной достоверности по группам показателей и коррекции значимости с целью учета ЭМТ. Использование таких процедур позволило установить достоверность связи VEGF с показателями оксиметрии при использовании S100 в качестве дополнительного фактора. Однако достоверность связи на приемлемом уровне устанавливается только при использовании в качестве целевой переменной бинарного показателя уровня содержания VEGF в сыворотке крови, задаваемого пороговым значением 750 пг/мл. Слабая биологическая обоснованность выбора именно такого порогового значения снижает достоверность вывода о существовании связи между бинарным показателем и сочетанием показателей оксиметрии и S100. Было визуально замечено существенное различие между уровнями корреляции VEGF с S100 слева и справа от порогового значения для показателя оксиметрии sO2. В связи с этим был предложен метод поиска достоверных условно-линейных закономерностей (ДУЛЗ), соответствующих наличию достоверных различий между коэффициентами корреляции целевой переменной Y с дополнительным фактором Z в группах, задаваемых оптимальным пороговым значением для показателя X. Как и в методе ОДР, верификация в методе ДУЛЗ производится с помощью перестановочных тестов. Для метода были разработаны процедуры поиска и оценок интегральной достоверности по группам показателей, а также коррекции значимости с целью учета ЭМТ приемлемой трудоемкости.

АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОПТИМАЛЬНЫХ ДОСТОВЕРНЫХ РАЗБИЕНИЙ

Поиск и верификация двумерных закономерностей, связывающих бинарный показатель уровня VEGF с показателями оксиметрии и различными дополнительными факторами

В настоящей работе нашей целью является изучение возможности существования сочетанных эффектов, связывающих VEGF с перечисленными выше семью показателями оксиметрии в сочетании с каким-то дополнительным фактором из числа показателей, присутствующих в анализируемой базе данных. Иными словами нашей целью является поиск для каждого дополнительного фактора Z эмпирических закономерностей, связывающих уровень VEGF с одним из 7 оксиметрических показателей. Подобно тому, как это делалось в работе [12] перейдем от исходного непрерывного показателя содержания VEGF к бинарному показателю $VEGF_b$. Перевод показателя содержания VEGF в бинарную форму производился с помощью простого порогового правила: бинарный показатель $VEGF_b$ считался равным 1 при содержании VEGF ниже 750 и $VEGF_b$ считался равным 2 при содержании VEGF выше 750.

Будем искать эмпирические закономерности, описываемые двумерными оптимальными разбиениями совместных областей допустимых пар признаков. Поиск и верификации таких закономерностей может производится с помощью технологии оптимальных достоверных разбиений (ОДР). Технологии ОДР подробно описаны в работах [12, 16, 17]. Используемая в настоящей работе модификация метода изложена в приложении.

При верификации двумерной закономерности, описывающей зависимость целевой переменной Y от сочетания переменных X_1 и X_2 в методе ОДР вычисляется две пары характеристик: параметров (p_1, h_1), оценивающих вклад в закономерность переменной X_1; параметров (p_2, h_2), оценивающих вклад в закономерность переменной X_2. В нашем случае в качестве целевой переменной Y выступает показатель $VEGF_b$. В качестве переменных X_1 и X_2 одна из переменных $Z_1, \ldots, Z_n$ и одна из переменных $U_1, \ldots, U_g$. Два оптимальных двумерных разбиения являются очевидно тождественными, если одно из них получено из другого переменой местами X_1 и X_2. Поэтому каждому дополнительному фактору Z_* может быть поставлено в соответствие g оптимальных разбиений, построенных в сочетании с каждым из показателей оксиметрию. Для указанных разбиений вычисляется множество характеристик значимости $\{(p^i_{ox}, h^i_{ox}), (p^i_z, h^i_z) \mid i = 1, \ldots, g\}$, где номеру i соответствует оптимальное двумерное разбиение совместной области допустимых значений дополнительного фактора z и показателя оксиметрии с номером i.

Интегральная оценка значимости по группам показателей

Более объективную оценку значимости связи $VEGF_b$ с показателями оксиметрии совместно с дополнительным фактором Z могут дать оценки, учитывающие все g оптимальных разбиений. Для интегральной оценки вклада показателей оксиметрии в совместную с дополнительным фактором z связь с $VEGF_b$ очевидно может быть использована сумма соответствующих p-значений $P_{ox} = \frac{1}{g}\sum_{i=1}^{g} p^i_{ox}$. Также для интегральной оценки вклада показателей оксиметрии может быть получена сумма h-значений $H_{ox}(\widetilde{S}_0) = \frac{1}{g}\sum_{i=1}^{g} h^i_{ox}$. Одновременно для интегральной оценки вклада дополнительного фактора Z в совместную с показателями оксиметрии связь с $VEGF_b$ очевидно могут быть использованы суммы $P_z = \frac{1}{g}\sum_{i=1}^{g} p^i_z$ и $H_z = \frac{1}{g}\sum_{i=1}^{g} h^i_z$.

Низкие значения величин P_z и P_{ox} и высокие значения величин H_z и H_z свидетельствуют о статистической значимости совместной связи с $VEGF_b$ оксиметрии

и фактора Z. Однако сами по себе они не имеют прямой вероятностной интерпретации. Оценивание статистической значимости значений P_z и P_{ox}, рассчитанных по обучающей выборке $\widetilde{S}_0$, будет производиться через сравнение их со значениями аналогичных показателей, рассчитанных по выборкам из набора $\widetilde{\mathbb{S}}^{gr}_{RP}$. Набор $\widetilde{\mathbb{S}}^{gr}_{RP}$ состоит из случайных выборок, которые независимо друг от друга генерируются из исходной обучающей выборки $\widetilde{S}_0$ путем случайных перестановок значений целевой переменной ($VEGF_b$) относительно фиксированных позиций векторов, компонентами которых являются показатели оксиметрии и дополнительный фактор Z. Для каждой выборки из $\widetilde{S}_r \in \widetilde{\mathbb{S}}^{gr}_{RP}$ ищутся оптимальные двумерные разбиения, связывающих случайную целевую переменную Y с одним из показателей оксиметрии в сочетании с Z. Далее производится верификация каждого из построенных разбиений с использованием технологии перестановочных тестов, описанной в приложении. Иными словами для каждой выборки из $\widetilde{S}_r \in \widetilde{\mathbb{S}}^{gr}_{RP}$ согласно изложенной в приложении схеме с помощью случайных перестановок целевой переменной относительно фиксированных позиций вектора объясняющих переменных генерируется набор случайных выборок $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$. Верификация разбиений, построенных по $\widetilde{S}_r$ производится путем сравнения с соответствующими разбиениями, построенными по выборкам из $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ (см. приложение).

Очевидно, что такая процедура требует огромных объемов вычислений пропорциональных $N_{gt}N_r$, где $N_{gt} =\mid \widetilde{\mathbb{S}}^{gr}_{RP} \mid$, $N_r =\mid \widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r) \mid$ (считается, что размер $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ не зависит от $\widetilde{S}_r$). Однако каждая из выборок из $\widetilde{S}_{rr} \in \widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ получается из исходной выборки $\widetilde{S}_0$ с помощью композиции случайной перестановки g, переводящей $\widetilde{S}_0$ в $\widetilde{S}_r$, и перестановки f, переводящей $\widetilde{S}_r$ в $\widetilde{S}_{rr}$. Пусть $\{f_j \mid j = 1, \ldots, N_r\}$ – множество перестановок по которым набор $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ был построен из выборки $\widetilde{S}_r$. Откуда следует, что набор $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ может быть получен из $\widetilde{S}_0$ с помощью перестановок из множества $\{g * f_j \mid, j = 1, \ldots, N_r\}$, которое является всего лишь множеством случайных перестановок. Поэтому распределения произвольной статистики на $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ должно сходиться к распределению аналогичной статистики на наборе $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_0)$, полученным из $\widetilde{S}_0$ с помощью произвольного набора из N_r случайных перестановок, при $N_r \to \infty$.

Приведенные рассуждения позволяют заменить изложенную выше схему верификации величин P_z и P_{ox} на существенно менее трудоемкую. Обсуждаемые выше величины $p^i_z, h^i_z, p^i_{ox}, h^i_{ox}, P_z, H_z, P_{ox}, H_{ox}$ далее будем обозначать $p^i_z(\widetilde{S}_0), h^i_z(\widetilde{S}_0), p^i_{ox}(\widetilde{S}_0), h^i_{ox}(\widetilde{S}_0), P_z(\widetilde{S}_0), H_z(\widetilde{S}_0), P_{ox}(\widetilde{S}_0), H_{ox}(\widetilde{S}_0)$, подчеркивая, что они были рассчитаны по исходной обучающей выборке $\widetilde{S}_0$.

Верификация производится по двум независимым наборам случайных выборок $\widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0)$ и $\widetilde{\mathbb{S}}^{t}_{RP}(\widetilde{S}_0)$. Для произвольной выборки $\widetilde{S}_r \in \widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0)$ по выборке $\widetilde{\mathbb{S}}^{t}_{RP}(\widetilde{S}_0)$ по схеме совершенно аналогичной той, которая использовалась при вычислении $\{[p^i_{ox}(\widetilde{S}_0), h^i_{ox}(\widetilde{S}_0)], [p^i_z(\widetilde{S}_0), h^i_z(\widetilde{S}_0)] \mid i = 1, \ldots, 7\}$,вычисляется множество значений $\{[p^i_{ox}(\widetilde{S}_r), h^i_{ox}(\widetilde{S}_r)], [p^i_z(\widetilde{S}_r), h^i_z(\widetilde{S}_r)] \mid i = 1, \ldots, 7\}$. Использовавшиеся в работе методы вычисления данных параметров представлены в приложении. Вычислим по этим значениям величины:

$P_z(\widetilde{S}_r) = \frac{1}{7}\sum_{i=1}^{7} p^i_z(\widetilde{S}_r)$,
$H_{ox}(\widetilde{S}_r) = \frac{1}{7}\sum_{i=1}^{7} h^i_{ox}(\widetilde{S}_r)$,
$P_z(\widetilde{S}_r) = \frac{1}{7}\sum_{i=1}^{7} p^i_{ox}(\widetilde{S}_r)$,
$H_z(\widetilde{S}_0) = \frac{1}{7}\sum_{i=1}^{7} h^i_z(\widetilde{S}_r)$.

Статистическая значимость интегральной оценки вклада дополнительного фактора z в совместную с показателями оксиметрии связь с $VEGF_b$, очевидно, может быть оценена

с помощью p-значений, рассчитываемых по формуле:

$$p_z^{gp} = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0) | P_z(\widetilde{S}_r) \le P_z(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)|}, \quad (1)$$

если оценивание значимости основывать на величинах p_z^i, и по формуле:

$$p_z^{gh} = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0) | H_z(\widetilde{S}_r) \ge H_z(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)|}, \quad (2)$$

если оценивание значимости основывать на величинах h_z^i.

Соответственно, статистическая значимость интегральной оценки вклада показателей оксиметрии может быть оценена с помощью p-значений, рассчитываемых по формуле:

$$p_{ox}^{gp} = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0) | P_{ox}(\widetilde{S}_r) \le P_{ox}(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)|}, \quad (3)$$

если оценивание значимости основывать на величинах p_{ox}^i, и по формуле:

$$p_{ox}^{gh} = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0) | H_{ox}(\widetilde{S}_r) \ge H_{ox}(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)|}. \quad (4)$$

Учет эффекта множественного тестирования при интегральном оценивании

Высокая исходная размерность данных требует учета при оценке статистической достоверности эффекта множественного тестирования (ЭМТ). Суть данного эффекта заключается в резком повышении вероятности случайно отвергнуть по крайней мере одну из проверяемых нулевых гипотез, если количество проводимых проверок велико.

ЭМТ приводит к существенному завышению уровней статистической значимости, полученных с помощью стандартных статистических тестов, как правило основанных на проверке одной единственной нулевой гипотезы. Для решения данной проблемы был разработан ряд достаточно простых методы коррекции уровней значимости, включая наиболее простые поправки Бонферрони, Холма и др. [7, 8]. Однако практически все они приводят к недооценке значимости. Особенно заметной недооценка оказывается в данных высокой размерности и с сильно коррелирующими переменными. В связи с этим получают популярность методов оценки ЭМТ, основывающихся на использовании перестановочных тестов [9, 10, 11, 13], когда статистическая значимость закономерностей, найденных по реальной выборке, сравнивается со статистической значимостью закономерностей, найденных в случайных выборках, которые генерируются из реальной выборки путем случайных перестановок позиций целевой переменной относительно фиксированных позиций векторов объясняющих переменных. Такой подход во многом соответствует изложенному выше способу учета группового эффекта.

Коррекция с целью учета ЭМТ должна производится также и при оценивании интегральной значимости по группам показателей. В этом случае, как это видно из описанного ранее, последовательно рассматриваются группы двумерных разбиений $\widetilde{R}(z, \widetilde{G}, \widetilde{S}_0)$, которые строятся по исходной выборке $\widetilde{S}_0$ и описывают связь целевой переменной Y с одним из показателей из группы $\widetilde{G}$ и некоторым дополнительным фактором из z. В настоящем случае нас интересует группа $\widetilde{G}_{ox}$ состоящая из показателей оксиметрии. Значимость вклада показателей оксиметрии оценивается с помощью величин $(p_{ox}^{gp}, p_{ox}^{gh})$, рассчитываемых по формулам (3, 4). Значимость вклада

дополнительного фактора z оценивается с помощью величин (p_z^{gp}, p_z^{gh}), рассчитываемых по формулам (1,2). Следует подчеркнуть, что перечисленные показатели значимости оценивают вклад в построенные по $\widetilde{S}_0$ модели из $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S}_0)$. Поэтому справедливым оказывается применение обозначений $p_z^{gp}(\widetilde{S}_0), p_z^{gh}(\widetilde{S}_0)$ и $p_{ox}^{gp}(\widetilde{S}_0), p_{ox}^{gh}(\widetilde{S}_0)$.

Выберем один из трех показателей значимости в качестве показателя корректирования по ЭМТ при оценивании значимости вклада фактора z. Например, в качестве показателя корректирования по ЭМТ может быть выбран p_z^{gh}. Пусть $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$ – набор двумерных разбиений, описывающих совместную связь оксигенации гемоглобина и дополнительного фактора Z с $VEGF_b$. Пусть $\widetilde{\mathbb{S}}_{rand}$ – множество выборок, которые сходны по размеру и составу переменных с $\widetilde{S}_0$, но генерируются гипотетическим случайным процессом исходя из независимости $VEGF_b$ как от показателей оксиметрии, так и от предполагаемых дополнительных факторов. Обозначим через $\widetilde{\mathbb{S}}_{rand}^{phz}(\alpha, z)$ множество выборок из $\widetilde{\mathbb{S}}_{rand}$, для которых выполнено неравенство

$$p_z^{gh}(\widetilde{S}) < \alpha. \tag{5}$$

То есть $\widetilde{\mathbb{S}}_{rand}^{phz}(\alpha, z)$ является множеством выборок из $\widetilde{\mathbb{S}}_{rand}$, для которых интегральный вклад дополнительного фактора z в $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$ значим на уровне α.

Обозначим через $\widetilde{\mathbb{S}}_{rand}^{phox}(\alpha, z)$ множество выборок из $\widetilde{\mathbb{S}}_{rand}$, для которых выполнено неравенство

$$p_{ox}^{gh}(\widetilde{S}) < \alpha, \tag{6}$$

то есть $\widetilde{\mathbb{S}}_{rand}^{phz}(\alpha, z)$ является множеством выборок из $\widetilde{\mathbb{S}}_{rand}$, для которых интегральный вклад оксигенации гемоглобина в $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$ значим на уровне α. Пусть $\widetilde{Z}$ – множество потенциальных дополнительных факторов. Через $\widetilde{\mathbb{S}}_{rand}^{phz}(\alpha, \widetilde{Z})$ обозначим объединение всевозможных множеств $\widetilde{\mathbb{S}}_{rand}^{phz}(\alpha, z)$ при $z \in \widetilde{Z}$. Для вычисления скорректированного с учетом эффекта множественного тестирования p-значения с помощью множества независимых случайных перестановок позиций целевой переменной $VEGF_b$ относительно фиксированных позиций векторов объясняющих переменных формируется набор случайных выборок $\widetilde{\mathbb{S}}_{RP}^{mt}$. В число объясняющих переменных входят оксиметрические показатели из группы G_{ox} и всевозможные потенциальные дополнительные факторы, в число которых входят всевозможные показатели из анализируемой базы за исключением VEGF, $VEGF_b$ и оксиметрических показателей из группы G_{ox}. Обозначим множество всевозможных дополнительных факторов через $\widetilde{Z}$.

Для каждой выборки из $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{mt}$ и для каждого $z \in \widetilde{Z}$ построим множество двумерных разбиений совместных областей допустимых значений z и каждого из показателей из G_{ox}, которое обозначим $\widetilde{R}(z, G_{ox}, \widetilde{S}_r)$, и вычислим для него набор показателей значимости $[p_z^{gp}(\widetilde{S}_r), p_z^{gh}(\widetilde{S}_r)]$ и $[p_{ox}^{gp}(\widetilde{S}_r), p_{ox}^{gh}(\widetilde{S}_r)]$. Перечисленные показатели могут быть рассчитаны с помощью аналога описанной ранее процедуры оценки интегральной по группам показателей значимости для исходной выборки $\widetilde{S}_0$, но с использованием двух независимых наборов случайных выборок $\widetilde{\mathbb{S}}_{RP}^{t}(\widetilde{S}_r)$ и $\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_r)$. Однако, из-за трудоемкости расчетов с отдельной генерацией таких наборов для каждой выборки $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{mt}$ вычисление скорректированных с учетом ЭМТ p-значений целесообразно проводить по общим для всех $\widetilde{S}_r$ наборам $\widetilde{\mathbb{S}}_{RP}^{t}(\widetilde{S}_0)$ и $\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)$.

Мы будем считать, что интегральный вклад дополнительного фактора Z в $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$,

который по отдельности значим на уровне α, значим с учетом ЭМТ на уровне β, если

$$\frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0) | \widetilde{S}_r \in \widetilde{\mathbb{S}}^{phz}_{rand}(\alpha, \widetilde{Z})|}{|\widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0)|} < \beta. \quad (7)$$

Также мы будем считать, что интегральный вклад фактора оксигенации в $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$, который по отдельности значим на уровне α, значим с учетом ЭМТ на уровне β, если

$$\frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0) | \widetilde{S}_r \in \widetilde{\mathbb{S}}^{phox}_{rand}[\alpha, \widetilde{Z}]|}{|\widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0)|} < \beta. \quad (8)$$

Результаты

Описанная выше модификация метода ОДР была использована для изучения связи группы g показателей оксиметрии с $VEGF_b$ в сочетании с каждым из n потенциальных дополнительных факторов. То есть рассматривалось множество из n наборов $\{(z_j, G_{ox}) \mid j = 1, \ldots, 136\}$. Для каждого набора было построено g двумерных закономерностей, с использованием описанной в приложении модификации перестановочного теста оценены характеристики значимости вклада как дополнительного фактора, так и соответствующего показателя оксиметрии. То есть для каждой двумерной закономерности для соответствующего показателя оксиметрии и дополнительного фактора рассчитаны p-значения и h-значения. Далее для каждого набора $\{(z_j, G_{ox}) \mid j = 1, \ldots, n\}$ рассчитаны интегральные значения P_z, P_{ox}, H_z, H_{ox} с использованием описанной выше технологии оценки интегральной значимости связи $VEGF_b$ с оксигенацией гемоглобина в сочетании дополнительным фактором.

Оценки проводились согласно формулам (1,2,3,4). Единственным дополнительным фактором, в сочетании с которым связь оксигенации гемоглобина с $VEGF_b$ оказалось значимой на уровне $p < 0.05$ оказалось содержание S100. Результаты приведены в таблице (табл. 2). В первом столбце таблицы приведено название интегральной характеристики (ИХ). Во втором столбце приведены величины соответствующих ИХ. В третьем столбце представлена интегральная значимость.

Таблица 2. Интегральная значимость вкладов оксигенации гемоглобина и содержания S100 в их совместную связь с $VEGF_b$

Интегральная характеристика	Значение ИХ	p-значение
P_{ox}	0.09	0.0055
H_{ox}	0.63	0.00015
P_z	0.0062	0.0007
H_z	0.798	0.00005

Из таблицы (табл. 2) видно, что более значимой связь оказывается, если ее оценивать по интегральным характеристикам, рассчитываемым по h-значениям, то есть по H_z и H_{ox}. При этом оказывается, что значимость вклада оксигенации гемоглобина оценивается на уровне $p = 0.00015$, а значимость вклада содержания S100 оценивается на уровне $p = 0.00005$.

Связь между исходной и скорректированной с учетом ЭМТ значимостью, полученная с использованием формул (8,7), представлена в таблице (табл. 3). Расчеты проводились по

Таблица 3. Связь между исходной и скорректированной значимостью

α	β_{s100}	β_{ox}
0.000025	0.00375	0.00675
0.00005	0.0065	0.0135
0.000075	0.0065	0.0135
0.0001	0.00995	0.02025
0.0002	0.0151	0.03345
0.0003	0.0209	0.046
0.0004	0.02665	0.0581
0.0005	0.0329	0.07

множеству $\widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0)$, включающему 20000 случайных выборок. В столбце, обозначенном β_{ox}, представлена доля выборок с интегральной значимостью по вкладу оксигенации на уровне $p < \alpha$, то есть $\beta_{ox} = \frac{|\widetilde{\mathbb{S}}^{phox}_{rand}[\alpha,\widetilde{Z}]|}{|\widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0)|}$. В столбце, обозначенном β_{s100}, представлена доля выборок с интегральной значимостью по вкладу оксигенации на уровне $p < \alpha$, то есть $\beta_{s100} = \frac{|\widetilde{\mathbb{S}}^{phz}_{rand}[\alpha,\widetilde{Z}]|}{|\widetilde{\mathbb{S}}^{mt}_{RP}(\widetilde{S}_0)|}$. Из таблицы (табл. 3) видно, что исходная значимость вклада S100, оцениваемая согласно таблице (табл. 2) на уровне $p = 0.00005$, соответствует $\beta_{s100} = 0.0065$. То есть скорректированная с учетом ЭМТ интегральная значимость вклада S100 оценивается согласно таблице (табл. 3) на уровне $p < 0.01$. Исходная значимость вклада уровня оксигенации, оцениваемая согласно таблице (табл. 2) на уровне $p = 0.00015$, соответствует $\beta_{ox} = 0.03345$. То есть скорректированная с учетом ЭМТ интегральная значимость вклада оксигенации оценивается согласно (табл. 3) на уровне $p < 0.035$. Связь $VEGF_b$ с S100 и показателями оксиметрии sO2 и FHHb иллюстрируется приведенными ниже диаграммами, представляющими соответствующие оптимальные разбиения.

Данная связь показана на трех приведенных ниже диаграммах рассеяния с использованием следующих условных обозначений:
+ – для VEGF > 750 (29 случаев),
o – для VEGF < 750 (59 случаев).

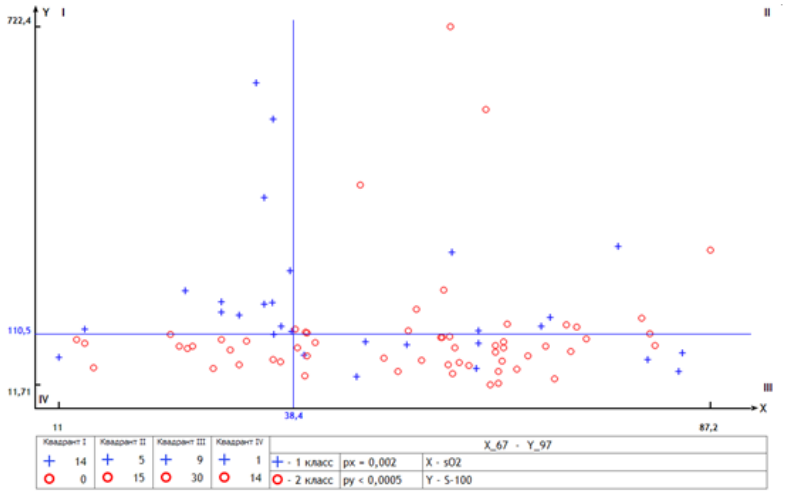

Рис. 1. Двумерная закономерность, связывающая VEGF с sO2 и S100.

В квадранте I, для которого выполняются условия S100 > 110.5 и sO2 < 38.4, содержатся только 14 случаев с VEGF > 750 и отсутствуют случаи с VEGF < 750, среднее

значение уровня VEGF составляет 1406; в квадранте II, для которого выполняются условия S100 $>$ 110.5 и sO2 $>$ 38.4, содержатся 5 случаев с VEGF $>$ 750 и 15 случаев с VEGF $<$ 750, среднее значение уровня VEGF составляет 688; в квадранте III, для которого выполняются условия S100 $<$ 110.5 и sO2 $>$ 38.4, содержится 9 случаев с VEGF $>$ 750 и 15 случаев с VEGF $<$ 750, среднее значение уровня VEGF составляет 583.7; в квадранте IV, для которого выполняются условия S100 $<$ 110.5 и sO2 $<$ 38.4, содержится только 1 случай с VEGF $>$ 750 и 15 случаев с VEGF $<$ 750, среднее значение уровня VEGF составляет 669. Значимость закономерности оценивалась с помощью перестановочного теста на уровне $p = 0.002$ для sO2, $p < 0.0005$ для S100.

Таким образом, средний уровень VEGF при содержании S100 $>$ 110.5 и sO2 $<$ 38.4 более чем в два раза превосходит средний уровень VEGF в остальной части выборки. Иными словами, увеличение уровня VEGF при снижении кислород-транспортирующей способности гемоглобина, происходит при повышении уровня S100.

Заметный рост уровня VEGF отмечается при содержании S100 $>$ 116.3 и FHHb $<$ 38.4. В связи с тем, что показатели FHHb и sO2 являются противоположными по их биохимическому смыслу, то закономерность из рисунка (рис. 2) находится в полном соответствии с закономерностью для пары sO2 и S100, изображенной на рисунке (рис. 1). Достоверность закономерности, связывающей бинарный показатель уровня VEGF с содержаниями sO2 и S100, подтверждается статистическими исследованиями с учетом эффекта множественного тестирования, представленными в работе [12].

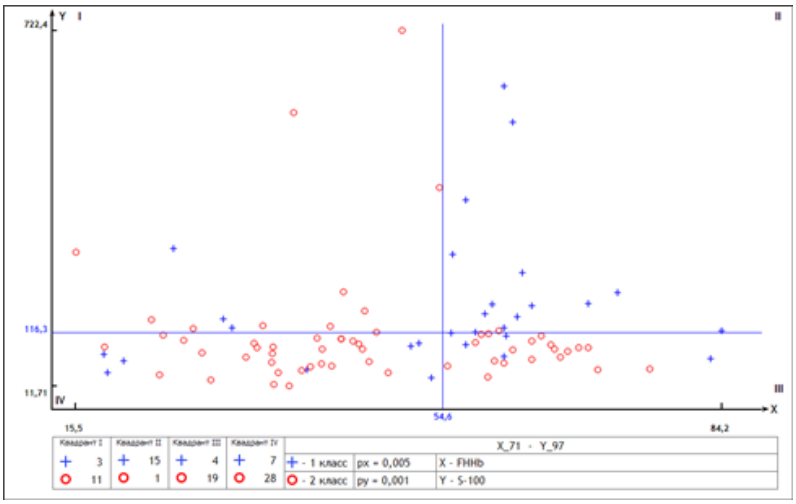

Рис. 2. Двумерная закономерность, связывающая VEGF с FHHb и S100.

В квадранте I, для которого выполняются условия S100 $>$ 116.3 и FHHb $<$ 54.6, содержатся 11 случаев с VEGF $<$ 750 и 3 случая с VEGF $>$ 750, среднее значение уровня VEGF составляет 743; квадрант II, для которого выполняются условия S100 $>$ 116.3 и FHHb $>$ 54.6, содержит 15 случаев с VEGF $>$ 750 и только 1 случай с VEGF $<$ 750, среднее значение уровня VEGF составляет 1310; квадрант III, для которого выполняются условия S100 $<$ 116.3 и FHHb $>$ 54.6, содержит 4 случая с VEGF $>$ 750 и 19 случаев с VEGF $<$ 750, среднее значение уровня VEGF составляет 666; квадрант IV, для которого выполняются условия S100 $<$ 116.3 и FHHb $<$ 54.6, содержит 7 случаев с VEGF $>$ 750 и 28 случаев с VEGF $<$ 750, среднее значение уровня VEGF составляет 559. Значимость закономерности оценивалась с помощью перестановочного теста на уровне $p = 0.005$ для FHHb и $p = 0.001$ для S100.

Обоснованность и биологическую ясность вывода о зависимости характера связи $VEGF_b$ с показателями оксиметрии и S100 снижает слабая биологическая

обоснованность использования границы 750 при переводе показателя содержания VEGF к бинарному виду. Для получения статистически значимых доказательств существования значительного влияния уровня оксигенации крови на характер связи между VEGF и S100 может быть использован метод, основанный на сочетании построения оптимальных разбиений с корреляционным анализом.

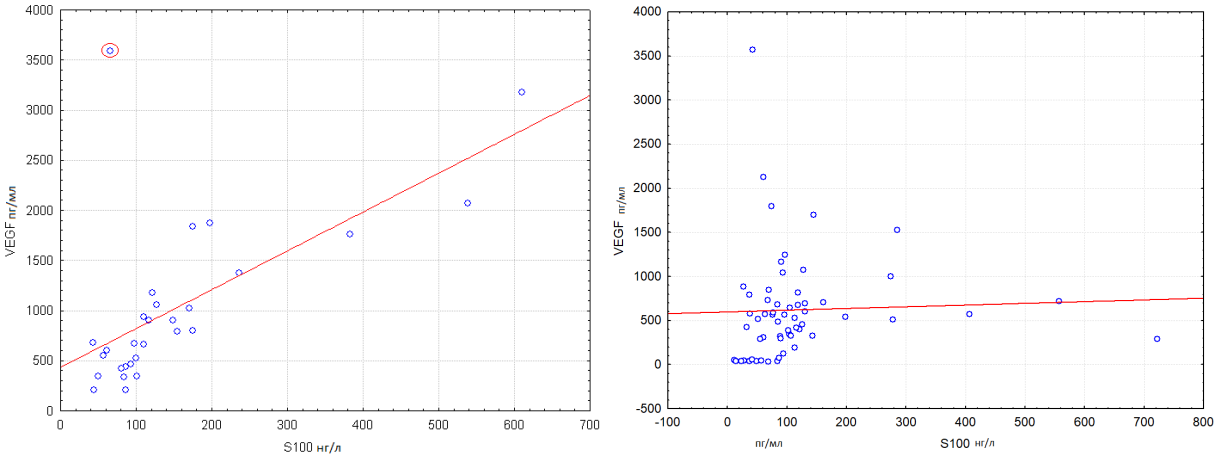

Рис. 3. Сравнение корреляции между S100 и VEGF слева и справа от границы 38.4 % для индекса сатурации sO2.

На рисунке (рис. 3) сравнивается корреляция между содержаниями VEGF и S100 в группах с содержанием $sO2 < 38.4$ и $sO2 > 38.4$. Слева зависимость содержания VEGF от содержания S100 в группе из 29 случаев, для которых $sO2 < 38.4$. Коэффициент корреляции равен 0.635. После исключения выпадающего наблюдения, которое на рисунке обведено красным кружком, коэффициент корреляции возрастает до 0.888. Справа изображена зависимость содержания VEGF от содержания S100 в группе из 59 случаев, для которых $sO2 > 38.4$. Коэффициент корреляции равен 0.039.

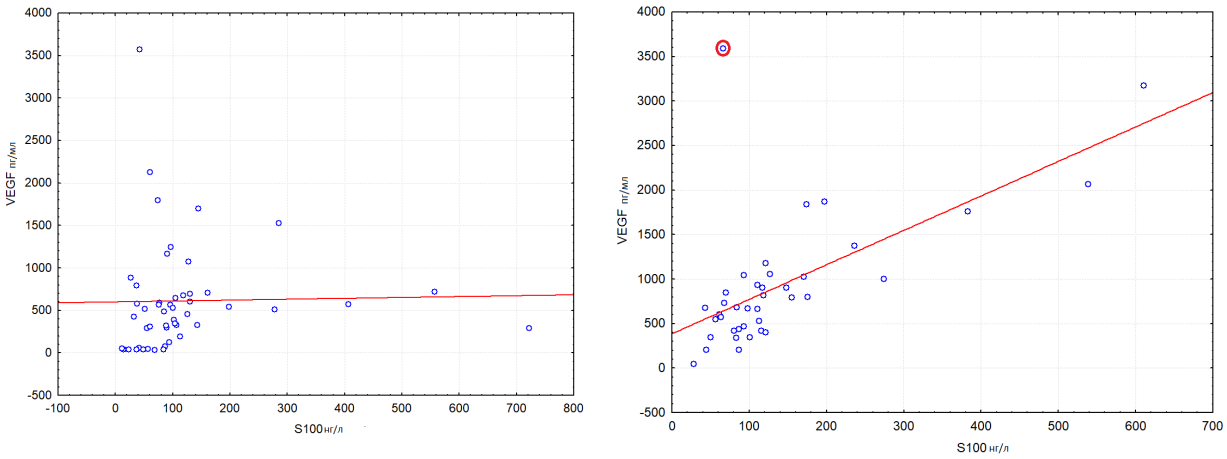

Рис. 4. Сравнение корреляции между S100 и VEGF слева и справа от границы 54.6 % для фракции дезоксигемоглобина.

На правой части рисунка (рис. 4) изображена зависимость содержания VEGF от содержания S100 в группе из 39 случаев, для которых $FHHb > 54.6$. Видно наличие выраженной линейной связи за исключением выделенного красным кружком выпадающего случая с содержанием VEGF выше 3500; после исключения выпадающего наблюдения коэффициент корреляции возрастает до 0.865. Коэффициент корреляции между VEGF и S100 составляет 0.639. В левой части рисунка (рис. 4) изображена

зависимость содержания VEGF от содержания S100 в группе из 49 случаев, для которых FHHb < 54.6. Видно отсутствие линейной связи: коэффициент корреляции равен 0.02.

Визуально наблюдаемое различие между уровнями корреляции VEGF и показателем S100 в левых и правых сегментах рисунков (рис. 3, 4) может указывать на реальное существование эффекта появления выраженной линейной связи между содержанием VEGF и S100, отражающего характер протекающих биохимических процессов. Вместе с тем статистическое обоснование такого вывода требует разработку нового метода, позволяющего находить статистически достоверные различия между уровнями корреляции переменных Y и Z в различных интервалах значений переменной X. По аналогии с методом кусочно-линейной регрессии назовем разрабатываемую модификацию ОДР методом достоверных кусочно-линейных связей (МДУЛЗ).

МЕТОД ДОСТОВЕРНЫХ УСЛОВНО-ЛИНЕЙНЫХ ЗАКОНОМЕРНОСТЕЙ (МДУЛЗ)

Оптимальные разбиения, максимизирующие корреляционные различия

В настоящей работе будет рассматриваться простейший вариант МДУЛЗ с одной граничной точкой b_x, разбивающих интервал значений переменной U на два подмножества q_l и q_r таким образом, чтобы различия между q_l и q_r по уровню корреляции между Y и Z были бы по возможности максимальными. Такие конструкции далее будем называть оптимальные разбиения, максимизирующие различия по уровням корреляции(ОРМКР). Предположим, что поиск оптимального разбиения и его последующая верификация производится по обучающей выборке: $\widetilde{S} = \{(y_i, z_i, x_i), i = 1, \ldots, m\}$. Очевидно, что разбиение интервала значений переменной U на подмножества q_l и q_r индуцирует разбиение $\widetilde{S}$ на две подвыборки: $\widetilde{S}_{sl} = \{(y_i, z_i, x_i) \in \widetilde{S} | x_i \in q_l\}$ и $\widetilde{S}_{sr} = \{(y_i, z_i, x_i) \in \widetilde{S} | x_i \in q_r\}$. Пусть ρ_{sl} – выборочный коэффициент корреляции, рассчитанный по выборке $\widetilde{S}_{sl}$, ρ_{sr} – выборочный коэффициент корреляции, рассчитанный по выборке $\widetilde{S}_{sr}$. Для описания различий между ρ_{sl} и ρ_{sr} предлагается использовать следующий функционал:

$$Q_{2\rho}(\widetilde{S}, b_u) = \frac{m_{sl} m_{sr} \, ||\rho_{sl}| - |\rho_{sr}||}{1 - \max(\rho_{sl}^2, \rho_{sr}^2)},$$

где m_{sl} и m_{sr} число объектов, со значением признака U соответственно меньшим и большим порога b_u. Данный функционал, очевидно, достигает максимальных значений, когда один из коэффициентов ρ_l и ρ_r достигает высоких по модулю значений при большой разности между модулями. Использование множителя $m_l m_r$ соответствует увеличению $Q_{2\rho}(\widetilde{S}, b_u)$, когда выборки $\widetilde{S}_r$ и $\widetilde{S}_l$ близки по размерам. Закономерности, изображенные на рисунках соответствуют высоким значениям $Q_{2\rho}(\widetilde{S}, b_x)$. На левой части рисунка (рис. 3) и на правой части рисунка (рис. 4) видны помеченные красными кружками выпадающие объекты (ВО), значительно отклоняющиеся от линейной зависимости между VEGF и S100, которой хорошо соответствуют все остальные объекты. Исключение ВО приводит к значительному росту коэффициента корреляции внутри соответствующих подвыборок, что, в свою очередь, приводит к увеличению значимости различий корреляции выше и ниже границы. Вместе с тем оценивание различий по выборке с исключенными ВО приводит к значительной переоценке достоверности. Альтернативным корректным подходом является использование так называемых робастных коэффициентов корреляции в рамках технологии перестановочных тестов.

Вычисления робастного коэффициента корреляции (РКК) между переменными V_1 и V_2 по выборке $\widetilde{S}_{2v} = (v_{11}, v_{12}), \ldots, (v_{1k}, v_{2k})$, где v_{ij} – значение переменной V_i

Математическая биология и биоинформатика. 2018. Т. 13.№ 2. doi: 10.17537/2018.13.563

для наблюдения j, $j = 1,\ldots,k$, $i = 1,2$. На первом шаге с использованием метода наименьших квадратов строится модель линейной регрессии переменной V^2 по переменной V^1: $V^2 = \alpha + \beta V^1 + \epsilon$. Пусть $\sigma = \sqrt{\frac{\sum_{i=1}^{k}(v_i^2-\alpha-\beta*v_i^1)^2}{k}}$. Далее из выборки удаляются явно не вписывающиеся в модель объекты, для которых $|v_{2j} - \alpha - \beta * v_{1j}| > 3\sigma$.

Верификация кусочно-линейных закономерностей

Для верификации ОРМКР был использован вариант перестановочного теста, предназначенный для проверке нулевой гипотезы I о независимости переменной Y от двумерного вектора (Z,U). Следует отметить, что кусочно-линейные закономерности на самом деле описывают связь целевой переменной Y с сочетаниями факторов. Поэтому при их верификации следует учитывать значимость вклада каждого фактора, как это и делается при верификации двумерных закономерностей в методе ОДР (см. приложение). Вклад различных факторов может быть оценен через проверку нескольких нулевых гипотез, включая гипотезу о независимости U от сочетания Y и Z, а также гипотезу о независимости Z от сочетания Y и U. Подробнее необходимость полной верификации рассмотрена ниже в соответствующем разделе. Гипотезу о независимости переменной Y от двумерного вектора (Z,U) далее будем называть нулевой гипотезой I.

Оценивание статистической значимости производится как и в подробно рассматриваемом в приложении методе ОДР через сравнение величины функционала $Q_{2\rho}$ для ОРМКР, построенному по обучающей выборке $\widetilde{S}_0$, с величинами $Q_{2\rho}$ для ОРМКР, построенных тем же самым способом по случайным выборкам, полученным из исходной выборки путем случайных независимых перестановок значений Y относительно фиксированных значений векторов переменных U и Z. Обозначим через $Q^o_{2\rho}(\widetilde{S})$ значение $Q_{2\rho}$ для ОРМКР, построенного по произвольной выборке $\widetilde{S}$. Предположим, что $\widetilde{F} = \{f_r | r = 1,\ldots,N\}$ множество случайных перестановок натуральных чисел из множества $\{1,\ldots,m\}$. Пусть $\widetilde{\mathbb{S}}^t_{RP} = \{\widetilde{S}_r \mid r = 1,\ldots,N_{per}\}$, где $\widetilde{S}_r = (y_{f_r(j)}, z_j, x_j) \mid j = 1,\ldots,m\}$. Величина p-значения вычисляется согласно формуле

$$p = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP} | Q^o_{2\rho}(\widetilde{S}_r) \geq Q^o_{2\rho}(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}^t_{RP}|}. \quad (9)$$

Наряду с p-значениями могут вычисляться также h-значения (см. приложение) по формуле

$$h = \frac{Q^o_{2\rho}(\widetilde{S}_o)}{\max_{\widetilde{\mathbb{S}}^t_{RP}} Q^o_{2\rho}(\widetilde{S}_r)}. \quad (10)$$

Следует отметить, что вероятность приближения модуля РКК к 1, оказывается достаточно высокой даже при справедливости нулевой гипотезы о полной независимости V_2 от V_1, если вычисления РКК производятся по выборке небольшого объема. Результаты исследования, подтверждающего справедливость данного утверждения, представлены на гистограммах, входящих в рисунок (рис. 5). Исходя из предположения о независимости переменных V_1 и V_2 с с использованием технологии Монте–Карло было сгенерировано по 10000 выборок вида $\widetilde{S}_{2v}$, состоящих из 5, 10, 15, 20, 25 и 30 объектов. Каждая из гистограмм описывает распределение сгенерированных случайных выборок по величинам коэффициентов корреляции между переменными.

Из рисунка видно, что РКК могут случайным образом превышать значение 0.85 при размере выборок ниже 25. Такое повышение РКК в свою очередь вызывает значительные случайные повышения величин функционала $Q^o_{2\rho}(\widetilde{S})$ при таких величинах порога b_u когда

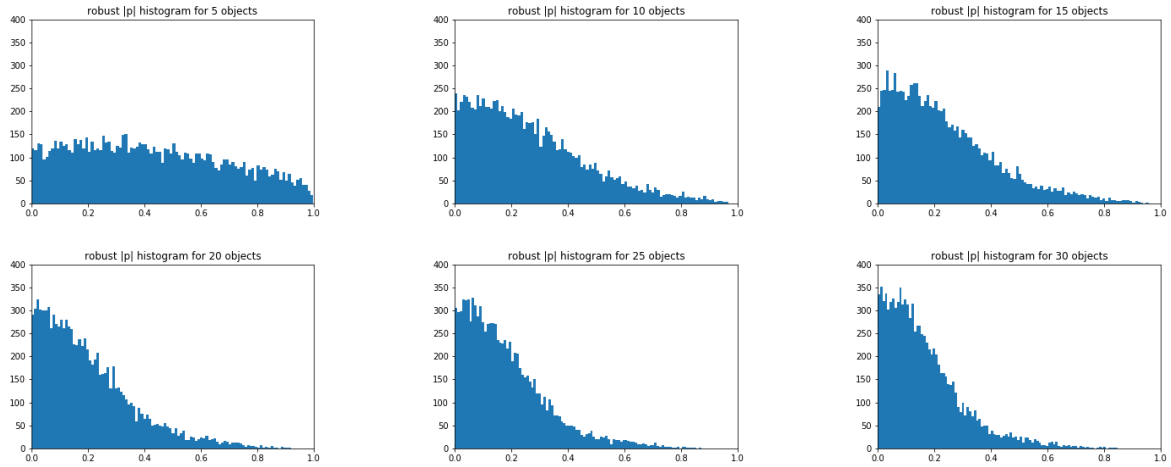

Рис. 5. Гистограммы распределения модуля робастного коэффициента корреляции (РКК) для выборок разного размера.

$|\widetilde{S}_{sl}| < 25$ или $|\widetilde{S}_{sr}| < 25$ даже при генерации выборки $\widetilde{S}$ из $\widetilde{S}_o$ с помощью случайных перестановок позиций y относительно фиксированных позиций векторов (Z, U). Чисто случайные скачки значений $Q^o_{2\rho}$ приводят к существенному снижению мощности теста, основанного на вычислении p-значений по формуле (9). Однако повышение мощности может быть достигнуто за счет сужения интервала поиска границы b_u. Так поиск оптимальных границ может проводится в интервале (b_{m_l}, b_{m_r}). Величина b_{m_l} является таким минимальным порогом, что число объектов из $\widetilde{S}_o$, удовлетворяющих условию $u_j < b_{m_l}$, оказывается равным или превышает m_l. Величина b_{m_r} является таким максимальным порогом, что число объектов из $\widetilde{S}_o$, удовлетворяющих условию $u_j < b_{m_r}$, оказывается равным или превышает m_r. На результаты поиска границы, оптимально разделяющих группы объектов с различными уровнями корреляции между Y и Z, а также на значимость выявленной закономерности, оказывают влияние m_l и m_r. Максимизация значимости достигается подбором наилучших значений m_l и m_r.

Перебор всевозможных значений пар (m_l, m_r) требует чрезвычайно большого объема вычислений и может приводить к связанному с эффектом множественного тестирования завышению уровня значимости. Мы проводили исследования исходя из условия равенства m_l и m_r, значения которых выбирались из множества $\{25, 30, 35\}$.

Интегральные оценки по группам показателей

Оценки, вычисляемые по формулам (9,10) характеризуют значимости кусочно-линейных связей для отдельных показателей оксиметрии. Очевидно, что более объективную и надежную оценку связи содержания VEGF с уровнем оксигенации гемоглобина в сочетании с дополнительным фактором может дать совокупность кусочно-линейных зависимостей с участием нескольких показателей оксиметрии.

Предположим, что найден набор кусочно-линейных моделей $\widetilde{R}(z, G_{oz})$, соответствующих значимым различиям между уровнем корреляции VEGF и Z в различных интервалах значений одного из показателей оксиметрии из группы G_{oz}. Данные закономерности характеризуются множеством характеристик значимости $\{(p_i, h_i) \mid i = 1, \ldots, g\}$, рассчитанных по формулам (9,10) для каждого из g показателей в группе G_{oz}.

В качестве интегральных характеристик значимости по группе G_{oz} могут быть использованы величины $P_{gz} = \sum_{i=1}^{g} p_i$ и $H_{gz} = \sum_{i=1}^{g} h_i$. Однако эффективность таких характеристик снижается из-за зашумляющего влияния показателей из G_{oz} со

слабой связью с целевой переменной. Для увеличения эффективности целесообразно использовать наиболее информативные показатели из группы. Предположим, что используется k показателей. Тогда в качестве интегральных характеристик значимости могут использоваться величины $P_{gz}^{kb} = \frac{1}{k}\sum_{i=1}^{k} p_i^r$ и $H_{gz}^{kb} = \frac{1}{k}\sum_{i=1}^{k} h_i^r$, где $k \leq g$, $\{p_i^r | i = 1,\ldots,g\}$ – последовательность отранжированных по мере возрастания p-значений, $\{p_i^r | i = 1,\ldots,g\}$ – последовательность отранжированных по мере убывания h-значений.

Альтернативными характеристиками значимости могут быть использована медианная характеристика значимости P_{gz}^{med}, определяемая как p-значение из множества $\{(p_i, h_i) \mid i = 1,\ldots,g\}$ такое, что число p-значений из $\{(p_i, h_i) \mid i = 1,\ldots,g\}$, не превышающих P_G^{med} равно числу p-значений не ниже, чем P_g^{med}. Аналогичным образом определяется величина H_g^{med}. Низкие значения величин P_{gz}^{kb} и P_{gz}^{med} и высокие величин H_z^{kb} и H_z^{med} свидетельствуют о статистической значимости совместной связи с VEGF оксиметрии и фактора z. Однако сами по себе они не имеют прямой вероятностной интерпретации. Верификацию величин P_{gz}^{kb},H_{gz}^{kb},P_{gz}^{med} и H_{gz}^{med}, рассчитанных для набора ОРМКР, построенных по обучающей выборке $\widetilde{S}_0$, с величинами P_{gz}^{kb},H_{gz}^{kb},P_{gz}^{med} и H_{gz}^{med}, рассчитанных для наборов ОРМКР, построенным по случайным выборкам из набора $\widetilde{\mathbb{S}}_{RP}^{gr}$, которые независимо друг от друга генерируется из исходной обучающей выборке $\widetilde{S}_0$ путем случайных перестановок значений целевой переменной (VEGF) относительно фиксированных позиций векторов, компонентами которых являются показатели из группы G_{ox} и дополнительный фактор z. Для каждой выборки из $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gr}$ строятся ОРМКР, каждая из которых связывает случайную целевую переменную y с одним из показателей оксиметрии в сочетании с z. Далее производится верификация каждого из построенных разбиений с использованием технологии перестановочных тестов, описанной в приложении. Иными словами для каждой выборки из $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gr}$ согласно изложенной в приложении схеме с помощью случайных перестановок целевой переменной относительно фиксированных позиций вектора показателей из G_{ox} и дополнительного фактора z генерируется набор случайных выборок $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$.

Верификации разбиений, построенных по $\widetilde{S}_r$ производится путем сравнения с соответствующими ОРМКР, построенными по выборкам из $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ (см. приложение). Очевидно, что такая процедура требует огромных объемов вычислений пропорциональных $N_{gt}N_r$, где $N_{gt} = \mid \widetilde{\mathbb{S}}_{RP}^{gr} \mid$, $N_r = \mid \widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r) \mid$ (считается, что размер $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ не зависит от $\widetilde{S}_r$). Однако каждая из выборок из $\widetilde{S}_{rr} \in \widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ получается из исходной выборки $\widetilde{S}_0$ с помощью композиции случайной перестановки g, переводящей $\widetilde{S}_0$ в $\widetilde{S}_r$, и перестановки f, переводящей $\widetilde{S}_r$ в $\widetilde{S}_{rr}$. Пусть $\{f_j \mid j = 1,\ldots,N_r\}$ множество перестановок по которым набор $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ был построен из выборки $\widetilde{S}_r$. Откуда следует, что набор $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ может быть получен из $\widetilde{S}_0$ с помощью перестановок из множества $\{g * f_j \mid, j = 1,\ldots,N_r\}$, которое является всего лишь множеством случайных перестановок. Поэтому распределения произвольной статистики на $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_r)$ должно сходиться к распределению аналогичной статистики на наборе $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_0)$, полученным из $\widetilde{S}_0$ с помощью произвольного набора из N_r случайных перестановок, при $N_r \to \infty$.

Приведенные рассуждения позволяют заменить изложенную выше схему верификации величин P_z и P_z^{med} или H_z и H_z^{med} на существенно менее трудоемкую. Характеристики значимости $p_i, h_i, P_{gz}^{kb}, H_{gz}^{kb}, P_{gz}^{med}, H_{gz}^{med}$ далее будем обозначать $p_i(\widetilde{S}_0), h_i(\widetilde{S}_0), P_{gz}^{kb}(\widetilde{S}_0), H_{gz}^{kb}(\widetilde{S}_0), P_{gz}^{med}(\widetilde{S}_0), H_z^{med}(\widetilde{S}_0)$, подчеркивая что они были рассчитаны по исходной выборке $\widetilde{S}_0$.

Верификация производится по двум независимым наборам случайных выборок $\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)$ и $\widetilde{\mathbb{S}}_{RP}^{t}(\widetilde{S}_0)$. Для произвольной выборки $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)$ по выборке $\widetilde{\mathbb{S}}_{RP}^{t}(\widetilde{S}_0)$

по схеме совершенно аналогичной той, которая использовалась при вычислении $\{[p^i_{ox}(\widetilde{S}_0), h^i_{ox}(\widetilde{S}_0)], [p^i_z(\widetilde{S}_0), h^i_z(\widetilde{S}_0)] \mid i = 1, \ldots, 7\}$, вычисляется множество значений $p^i(\widetilde{S}_r), h^i(\widetilde{S}_r), P_z(\widetilde{S}_r), H_z(\widetilde{S}_r), P^{med}_z(\widetilde{S}_r), H^{med}_z(\widetilde{S}_r)$. Использовавшиеся в работе методы вычисления данных параметров представлены в приложении.

Статистическая значимость интегральной оценки связи VEGF с оксигенацией гемоглобина в сочетании с дополнительным фактором Z очевидно может быть оценена с помощью p-значения, рассчитываемых по формуле

$$p^{gp}_z = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0) | P_z(\widetilde{S}_r) \le P_z(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0)|}, \tag{11}$$

если оценивание значимости основывать на величинах p^i, и по формуле

$$p^{gh}_z = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0) | H_z(\widetilde{S}_r) \ge H_z(\widetilde{S}_0)\}|}{|\widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0)|}, \tag{12}$$

если оценивание значимости основывать на величинах h^i_z.

Наряду с интегральными значениями p-значениями могут использоваться также интегральные h-значения. Статистическая значимость интегральной оценки вклада дополнительного фактора z очевидно может быть оценена с помощью h-значения, рассчитываемых по формуле

$$h^{gh}_z = \frac{H_z(\widetilde{S}_0)}{\max_{\widetilde{S}_r \in \widetilde{\mathbb{S}}^{gt}_{RP}(\widetilde{S}_0)} H_z(\widetilde{S}_r)}. \tag{13}$$

ПРОБЛЕМА МНОЖЕСТВЕННОГО ТЕСТИРОВАНИЯ

Коррекция с целью учета ЭМТ должна производится также и при оценивании интегральной значимости по группам показателей. Предположим, что $\widetilde{R}(z, \widetilde{G}_{ox})$ является набором ОРМКР, которые строятся по исходной выборке $\widetilde{S}_0$, которые описывают связь целевой переменной с одним из показателей из группы $\widetilde{G}_{ox}$ и некоторым дополнительным фактором из z. Значимость вклада показателей оксиметрии оценивается с помощью величин $[p^{gp}_z, p^{gh}_z, h^{gh}_z]$, рассчитываемых по формулам (11,12,13). Следует подчеркнуть, что перечисленные характеристики оценивают интегральную значимость по всем ОРМКР из $\widetilde{R}(z, \widetilde{G}_{ox})$, построенным по $\widetilde{S}_0$. Поэтому справедливым оказывается применение обозначений $p^{gp}(\widetilde{S}_0), p^{gh}(\widetilde{S}_0)$. Выберем один из показателей значимости в качестве показателя, по которому производится корректирование по ЭМТ при оценивании значимости вклада фактора z. Например, в качестве такого показателя корректирования может быть выбран p^{gh}_z или h^{gh}_z.

Введем дополнительные обозначения. Пусть $\widetilde{\mathbb{S}}_{rand}$ – множество выборок, которые сходны по размеру и составу переменных с $\widetilde{S}_0$, но генерируются гипотетическим случайным процессом исходя из независимости VEGF как от показателей оксиметрии так и от предполагаемых дополнительных факторов. Обозначим через $\widetilde{\mathbb{S}}^{phzc}_{rand}(\alpha, z)$ множество выборок из $\widetilde{\mathbb{S}}_{rand}$, для которых выполнено неравенство

$$p^{gh}_z(\widetilde{S}) < \alpha. \tag{14}$$

Через $\widetilde{\mathbb{S}}^{phz}_{rand}(\alpha, \widetilde{Z})$ обозначим объединение всевозможных множеств $\widetilde{\mathbb{S}}^{phzc}_{rand}(\alpha, z)$ при $z \in \widetilde{Z}$. В свою очередь, через $\widetilde{\mathbb{S}}^{hhzc}_{rand}(\gamma, z)$ множество выборок из $\widetilde{\mathbb{S}}_{rand}$, для которых выполнено

неравенство

$$h_z^{gh}(\widetilde{S}) > \gamma, \tag{15}$$

а через $\widetilde{\mathbb{S}}_{rand}^{hz}(\alpha, \widetilde{Z})$ обозначим объединение всевозможных множеств $\widetilde{\mathbb{S}}_{rand}^{hzc}(\alpha, z)$ при $z \in \widetilde{Z}$.

Вычисление скорректированных с учетом эффекта множественного тестирования p-значений. Для вычисления скорректированного с учетом ЭМТ p-значения с помощью множества независимых случайных перестановок позиций целевой переменной VEGF относительно фиксированных позиций векторов объясняющих переменных формируется набор случайных выборок $\widetilde{\mathbb{S}}_{RP}^{mt}$. В число объясняющих переменных входят оксиметрические показатели из группы G_{ox} и всевозможные потенциальные дополнительные факторы, в число которых входят всевозможные показатели из анализируемой базы за исключением VEGF и оксиметрических показателей из группы G_{ox}. Обозначим множество всевозможных дополнительных факторов через $\widetilde{Z}$.

Для каждой выборки из $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{mt}$ и для каждого $z \in \widetilde{Z}$ построим множество ОРМКР, и вычислим для него интегральные показателей значимости $[p_z^{gp}(\widetilde{S}_r), p_z^{gh}(\widetilde{S}_r)]$. Перечисленные показатели могут быть рассчитаны с помощью аналога описанной ранее процедуры оценки интегральной по группам показателей значимости для исходной выборки $\widetilde{S}_0$, но с использованием двух независимых наборов случайных выборок $\widetilde{\mathbb{S}}_{RP}^{t}(\widetilde{S}_r)$ и $\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_r)$. Однако из-за трудоемкости расчетов с отдельной генерацией таких наборов для каждой выборки $\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{mt}$ вычисление скорректированных с учетом ЭМТ p-значений целесообразно проводить по общим для всех $\widetilde{S}_r$ наборам $\widetilde{\mathbb{S}}_{RP}^{t}(\widetilde{S}_0)$ и $\widetilde{\mathbb{S}}_{RP}^{gt}(\widetilde{S}_0)$.

Предположим, что в качестве показателя корректирования выбран p_z^{gh}. Тогда мы будем считать, что интегральный вклад фактора вклад дополнительного фактора z в $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$, который по отдельности значим на уровне α, значим с учетом ЭМТ на уровне β, если

$$\frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{mt}(\widetilde{S}_0) | \widetilde{S}_r \in \widetilde{\mathbb{S}}_{rand}^{phz}(\alpha, \widetilde{Z})|}{|\widetilde{\mathbb{S}}_{RP}^{mt}(\widetilde{S}_0)|} < \beta. \tag{16}$$

Предположим, что в качестве показателя корректирования выбран h_z^{gh}. Тогда мы будем считать, что интегральный вклад фактора вклад дополнительного фактора z в $\widetilde{R}(z, \widetilde{G}_{ox}, \widetilde{S})$, который по отдельности значим на уровне γ, значим с учетом ЭМТ на уровне β, если

$$\frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP}^{mt}(\widetilde{S}_0) | \widetilde{S}_r \in \widetilde{\mathbb{S}}_{rand}^{hhz}(\gamma, \widetilde{Z})|}{|\widetilde{\mathbb{S}}_{RP}^{mt}(\widetilde{S}_0)|} > \gamma. \tag{17}$$

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ С ПОМОЩЬЮ МДУЛЗ

Описанный выше метод поиска верифицированных условно-линейных зависимостей (УЛЗ) был использован для изучения связи группы из трех информативных показателей оксиметрии, напрямую характеризующих степень оксигенации гемоглобина, с VEGF в сочетании с каждым из n потенциальных дополнительных факторов. Группа показателей оксиметрии включала sO2, FHHb и FO2Hb.

Исследования показали наличие двух достоверных УЛЗ, описывающих связь уровня VEGF со степенью оксигенации в сочетании с содержанием в сыворотке крови C4 и S100. В таблице (табл. 4) для обоих дополнительных факторов приведены величины медианные характеристики значимости P_G^{med} и H_G^{med}, введенные ранее. Также в таблице (табл. 4) приведены статистические оценки интегральной значимости p_z^{gp} и p_z^{gh}, рассчитываемые соответственно по формулам (11) и (12), а также интегральное

h-значение h_z^{gh}, рассчитываемое по формуле (13).

Таблица 4. Значения статистических параметров, характеризующих условно линейную связь VEGF с S100 и C4

Статистический параметр	S100	C4
P_{gz}^{med}	0.0006	0.0001
H_{gz}^{med}	0.63	0.975
p_z^{gp}	0.0	0.0
p_z^{gh}	0.0001	0.0
h_z^{gh}	1.016	1.056

Коррекция на множественное тестирование может производится как согласно формуле (16), то есть при использовании в качестве показателей коррекции параметров p_z^{gh}, так и согласно формуле (17), то есть при использовании в качестве показателей коррекции параметров h_z^{gh}. Отметим, что в последнем случае скорректированные значимость чаще оказывается лучше, чем в первом. Для вычисления скорректированных значений при использовании в качестве показателя корректирования величин h_z^{gh} будем использовать следующую процедуру.

А) Выделим набор, например, из 8 пороговых значений $\{\delta_i^{gh}|i=1,\ldots,8\}$ для величин h_z^{gh}.

Б) Для каждого порога δ_i^{gh} подсчитаем долю выборок из $\widetilde{\mathbb{S}}_{RP}^{mt}$, в которых хотя бы для одного из дополнительных факторов выполняется неравенство $h_z^{gh} > \delta_i^{gh}$. Обозначим эти доли через $\{\nu_i^{gh}|i=1,\ldots,8\}$.

В) В качестве скорректированного значимости закономерности, найденной на исходной выборке $\widetilde{S}_o$ с $h_z^{gh} = h_z^{gh}(\widetilde{S}_o)$ будем использовать долю ν_i^{gh}, соответствующую минимальному δ_i^{gh}, удовлетворяющему условию $h_z^{gh} = h_z^{gh}(\widetilde{S}_o) \geq \delta_i^{gh}$.

Таблица 5. Связь между исходной и скорректированной значимостью

h_z^{gh}	1.35	1.3	1.25	1.2	1.15	1.1	1.05	1.0
β	0.0018	0.002	0.0028	0.0037	0.0053	0.0066	0.0082	0.012

Из таблицы (табл. 5) видно, что при использовании содержания S100 в качестве дополнительного параметра z было получено значение h_z^{gh}, равное 1.016. Ближайшее снизу пороговое значение равно 1.0. Доля выборок в которых хотя бы для одного из дополнительных факторов выполняется неравенство $h_z^{gh} > 1.0$ составляет 0.012. Таким образом, p-значение, соответствующее проверке нулевой гипотезе I о независимости содержания VEGF от сочетания факторов оксигенации крови и содержания S100, удовлетворяет неравенству $p < 0.012$.

При использовании содержания C4 в качестве дополнительного параметра z было получено значение h_z^{gh}, равное 1.056. Ближайшее снизу пороговое значение равно 1.05. Доля выборок, в которых хотя бы для одного из дополнительных факторов выполняется неравенство $h_z^{gh} > 1.0$, скорректированное p-значение составляет 0.0082. Таким образом, p-значение, соответствующее проверке нулевой гипотезе I о независимости содержания VEGF от сочетания факторов оксигенации крови и содержания C4, удовлетворяет неравенству $p < 0.0082$.

Иллюстративное описание выявленных достоверных связей. На рисунке (рис. 4) сравнивается корреляция между содержаниями VEGF и S100 в группах с содержанием sO2 ниже и выше полученного с помощью МДУЛЗ порога равного 39.75 %, что оказывается несколько выше порога 38.4 %, рассчитанного ранее с помощью метода ОДР.

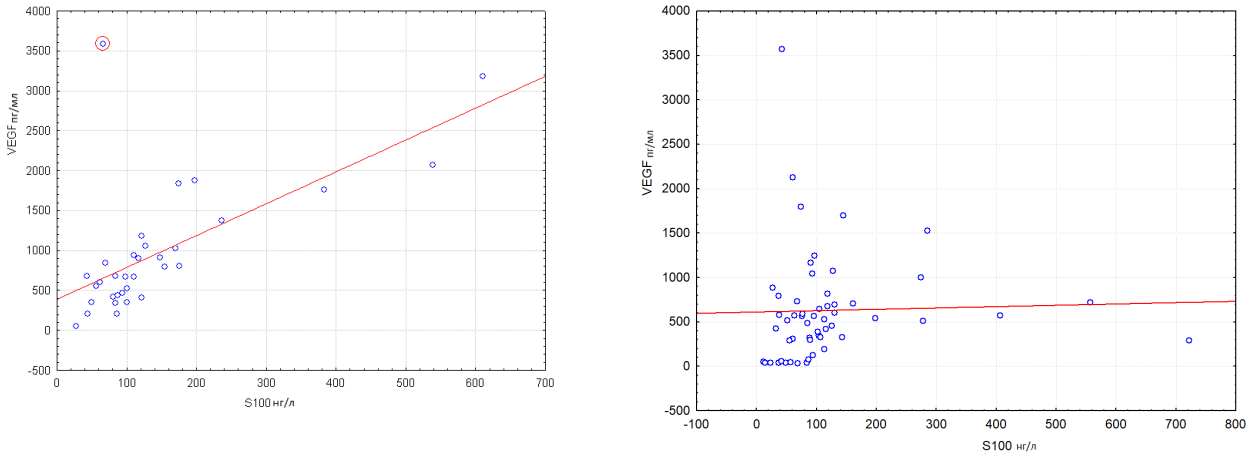

Рис. 6. Сравнение корреляции между S100 и VEGF слева и справа от границы 39.75 % для индекса сатурации sO2.

Слева представлена связь содержания VEGF и S100 в группе из 33 случаев, для которых sO2 $<$ 39.75 %. Коэффициент корреляции в группе слева от границы равен 0.64. После удаления обведенного красным кружком выпадающего наблюдения робастный коэффициент корреляции (РКК) возрастает до 0.88. Справа представлена зависимость содержания VEGF от содержания S100 в группе из 55 случаев, для которых sO2 $>$ 39.75 %. Коэффициент корреляции равен 0.03 с возрастанием РКК до 0.12.

На рисунке (рис. 7) сравнивается корреляция между содержаниями VEGF и S100 в группах с содержанием FHHb ниже и выше полученного с помощью МДУЛЗ порога равного 56.3 %, что оказывается несколько выше порога 54.6 %, рассчитанного ранее с помощью метода ОДР.

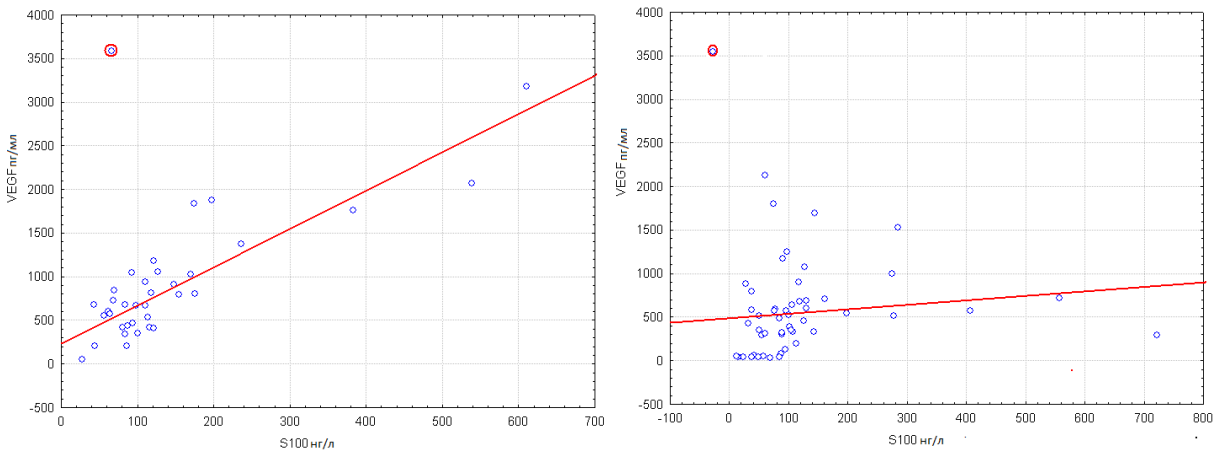

Рис. 7. Сравнение корреляции между S100 и VEGF слева и справа от границы 56.3 % для фракции дезоксигемоглобина.

Слева зависимость содержания VEGF от S100 в группе из 52 случаев, для которых FHHb $<$ 56.3 %. Коэффициент корреляции в группе слева от границы равен 0.04, а РКК возрастает до 0.13. Справа представлена связь содержания VEGF и S100 в группе из

36 случаев, для которых FHHb > 56.3 %. Коэффициент корреляции в группе справа от границы равен 0.64, а РКК возрастает до 0.871.

На рисунке (рис. 8) сравнивается корреляция между содержаниями VEGF и C4 в группах с содержанием sO2 ниже и выше полученного с помощью МДУЛЗ порога равного 39.25 %. Слева представлена связь содержания VEGF и C4 в группе из 31 случая, для которых sO2 < 39.25 %. Коэффициент корреляции равен в группе слева от границы равен 0.47, а РКК превышает 0.76. Справа зависимость содержания VEGF от содержания C4 в группе из 57 случаев, для которых sO2 > 39.25 %. Коэффициент корреляции в группе справа от границы равен -0.11, а РКК равен 0.05.

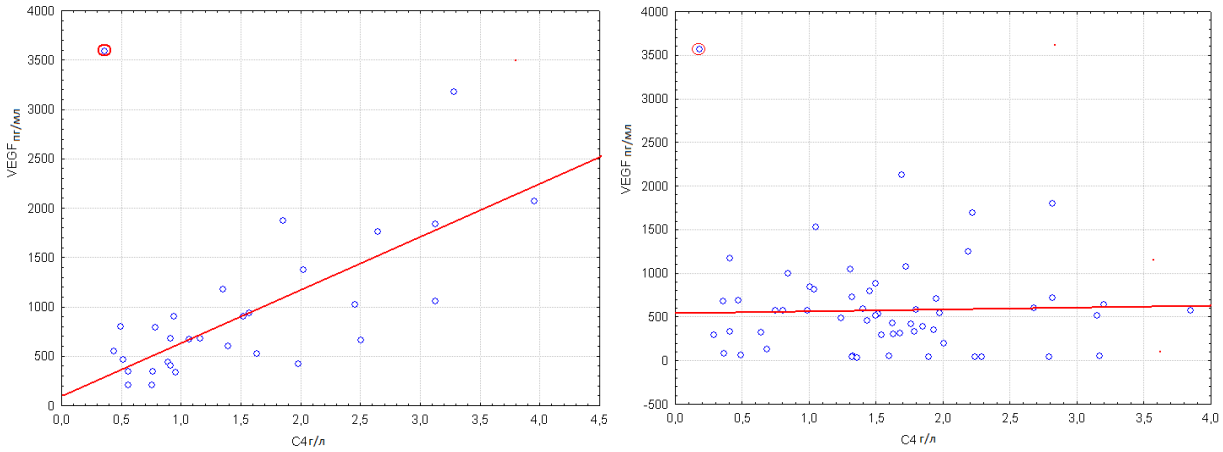

Рис. 8. Сравнение корреляции между C4 и VEGF слева и справа от границы 39.25 % для индекса сатурации sO2.

На рисунке (рис. 9) сравнивается корреляция между содержаниями VEGF и C4 в группах с содержанием FHHb ниже и выше полученного с помощью МДУЛЗ порога равного 56.3 %. Слева зависимость содержания VEGF от S100 в группе из 52 случаев, для которых FHHb < 56.3 %. Коэффициент корреляции в группе слева от границы равен -0.1, а РКК возрастает до 0.07. Справа представлена связь содержания VEGF и S100 в группе из 36 случаев, для которых FHHb > 56.3 %. Коэффициент корреляции в группе справа от границы равен 0.45, а РКК возрастает до 0.73.

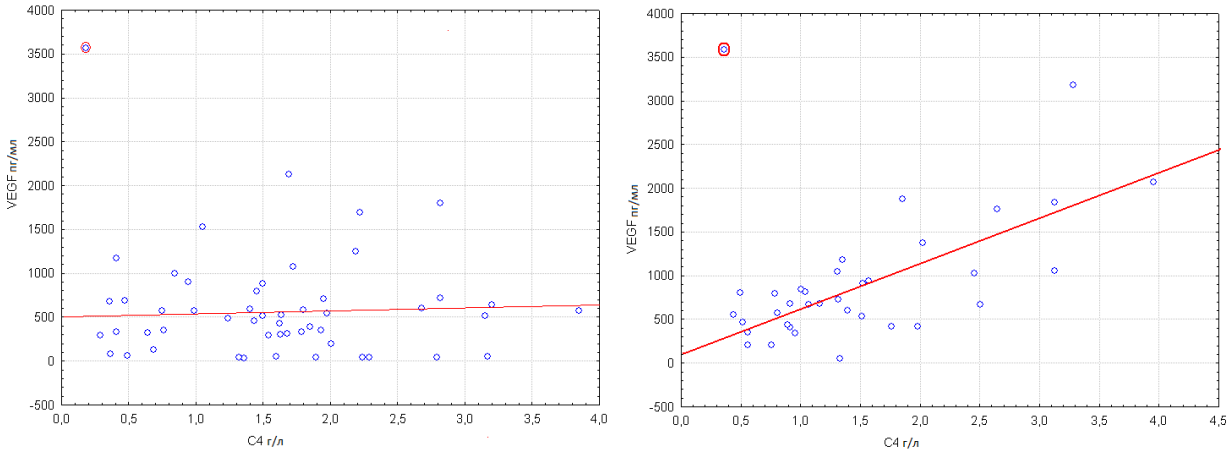

Рис. 9. Сравнение корреляции между C4 и VEGF слева и справа от границы 56.3 % для фракции дезоксигемоглобина FHHb.

ПОЛНАЯ ВЕРИФИКАЦИЯ УСЛОВНО-ЛИНЕЙНЫХ ЗАКОНОМЕРНОСТЕЙ

Вывод о достоверности условно-линейных закономерностей основывался на проверке нулевой гипотезы I о независимости VEGF от сочетания показателей оксиметрии и дополнительного фактора. Однако проверки только одной этой нулевой гипотезы может оказаться недостаточно для уверенного вывода о действительной необходимости использования всех участвующих в описании переменных. Например, выразительные сложные зависимости, видимые на рисунках (рис. 5–7), могут оказаться результатом только одной реально существующей связи между VEGF и дополнительным фактором, то есть S100 или C4. Также связи из рисунков (рис. 5–7) могут быть результатом лишь одной на самом деле существующей связи между дополнительными факторами показателями оксиметрии.

Возможность объяснения видимых на рисунках конфигураций только лишь реальным существованием связи между VEGF и дополнительным фактором мы предлагаем оценивать через проверку нулевой гипотезы о независимости показателей оксиметрии от сочетания VEGF и дополнительного фактора, которую мы далее будем называть нулевой гипотезой II. Для проверки этой нулевой гипотезы может быть использованы технологии основанные на перестановочных тестах, которые фактически повторяют методологию, использованную при проверке нулевой гипотезы о независимости VEGF сочетания дополнительного фактора и показателей оксиметрии. Однако наборы $\widetilde{\mathbb{S}}_{RP}^{gr}$ и $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_o)$ формируются из $\widetilde{S}_0$ путем случайных перестановок позиций векторов значений VEGF и дополнительного фактора относительно фиксированных значений векторов показателей оксиметрии.

В отличие от исследования по оцениванию справедливости нулевой гипотезы I о независимости VEGF от сочетания показателей оксиметрии и дополнительного фактора в настоящем случае к несколько более значимому результату привело использование в качестве групповых интегральных оценок величин P_{gz}^{kb} и H_{gz}^{kb} при $k = 3$. Результаты исследования по оцениванию справедливости нулевой гипотезы II приведены в таблице (табл. 6).

Таблица 6. Значения статистических параметров, характеризующих условно линейную связь VEGF с S100 и C4 при проверке нулевой гипотезы II

статистический параметр	S100	C4
P_{gz}^{kb}	0.0392	0.00126
H_{gz}^{kb}	0.315	0.536
p_z^{gp}	0.0415	0.00054
p_z^{gh}	0.0376	0.0007

Из таблицы (табл. 6) видно, что нулевая гипотез II, предполагающая независимость сочетания VEGF и дополнительного фактора от показателей оксиметрии, отвергается как при дополнительном факторе C4, так и при дополнительном факторе S100.Однако в случае S100 нулевая гипотеза отвергается только на уровне 0.05. Однако данный результат можно считать окончательным. Дополнительной коррекции с учетом ЭМТ не требуется, поскольку оценивание значимости нулевой гипотезы II проводилось только для тех дополнительных факторов, которые оказались значимы с учетом ЭМТ нулевой гипотезы I.

Нулевой гипотезой III мы будем называть гипотезу о независимости дополнительного фактора от сочетания VEGF и показателей оксиметрии. При проверке этой гипотезы

наборы $\widetilde{\mathbb{S}}_{RP}^{gr}$ и $\widetilde{\mathbb{S}}_{RP}(\widetilde{S}_o)$ формируются из $\widetilde{S}_0$ путем случайных перестановок позиций векторов значений VEGF и показателей оксиметрии относительно фиксированных значений дополнительного фактора.

Таблица 7. Значения статистических параметров, характеризующих условно линейную связь VEGF с S100 и C4 при проверке нулевой гипотезы III

статистический параметр	S100	C4
P_{gz}^{kb}	0.000053	0.000056
H_{gz}^{kb}	0.833	0.819
p_z^{gp}	0.00001	0.00002
p_z^{gh}	< 0.00001	0.00004

Из таблицы (табл. 7) видно, что нулевая гипотеза III отвергается с очень высоким уровнем значимости при использовании в качестве дополнительного фактора как уровня C4, так и уровня S100.

ЗАКЛЮЧЕНИЕ

Несмотря на литературные данные о процессе активации синтеза VEGF в ответ на гипоксию, стандартный корреляционный анализ клинических результатов не выявил прямой связи между уровнем VEGF в сыворотке крови и показателями оксиметрии. Однако применение методов интеллектуального анализа данных позволило обнаружить и верифицировать более сложные эффекты, заключающиеся во влиянии показателей оксиметрии на характер связи между VEGF и S100, а также VEGF и C4. При этом для интервалов значений показателей оксиметрии, соответствующих пониженной способности крови транспортировать кислород, характерен высокий уровень корреляции VEGF и S100, а также VEGF и C4. Для обнаружения и верификации закономерностей, связанных с различием уровней корреляции двух переменных внутри интервалов значений третьей переменной, был предложен новый вариант метода оптимальных достоверных разбиений, названный методом достоверных условно-линейных закономерностей. Разработанная технология позволяет учитывать интегральный эффект по группам переменных, а также проводить коррекцию, связанную с учетом множественного тестирования.

ПРИЛОЖЕНИЕ 1.

МОДИФИЦИРОВАННЫЙ ВАРИАНТ МЕТОДА ПОИСКА ПОЛНОСТЬЮ ВЕРИФИЦИРОВАННЫХ ЗАКОНОМЕРНОСТЕЙ

Метод оптимальных достоверных разбиений позволяет находить закономерности, связывающие целевую переменную Y с объясняющими переменными $X_1, \ldots, X_n$ и задаваемые одномерными или двумерными оптимальными разбиениями пространства объясняющих переменных. Поиск оптимальных разбиений производится по обучающей выборке $\widetilde{S}_o$ внутри фиксированных множеств определенного вида и определенного уровня сложности через максимизацию функционала качества Q, характеризующего разделение наблюдений с различными уровнями значений Y. Обучающая выборка имеет вид

$$\widetilde{S}_o = \{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}.$$

Пусть разбиение R включает элементы $q_1, \ldots, q_r$. Пусть m_i – число объектов из $\widetilde{S}_o$, для которых $\mathbf{x}_j \in q_i$. Тогда функционал Q задается в виде

$$Q(R, \widetilde{S}) = \max_{i \in \{1,\ldots,r\}} m_i \times (\hat{y}_i - \hat{y}_o)^2,$$

где $\hat{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_j$, $\hat{y}_i = \frac{1}{m_i} \sum_{\mathbf{x}_j \in q_i}^{m_i} y_j$.

Пусть максимальное значение функционала Q на выборке $\widetilde{S}$ достигается для некоторого разбиения R_o. Обозначим это оптимальное значение через $Q_o(\widetilde{S})$. Статистическая верификация закономерности, описываемой разбиений производится с помощью перестановочного теста, состоящего в сравнении оптимальной величины функционала Q на обучающей выборке $\widetilde{S}_0$ с тем же самым способом рассчитанными оптимальными величинами Q на случайных выборках, полученных из исходной выборки путем случайных независимых перестановок значений Y относительно фиксированных значений объясняющих переменных. Достоверность простейших закономерностей оценивается с помощью p-значений, вычисляемых как отношение числа случайных выборок, для которых оптимальное значение Q превосходит оптимальное значение Q, достигнутое на $\widetilde{S}_0$.

Предположим, что $\widetilde{\mathbb{S}}_{RP}$ – множество случайных выборок, задаваемых перестановками. Величина p-значения вычисляется по формуле

$$p = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbb{S}}_{RP} | Q(\widetilde{S}_r) \geq Q_o(\widetilde{S}_o)\}|}{|\widetilde{\mathbb{S}}_{RP}|} \tag{18}$$

Использование p-значений, рассчитываемых по формуле (18), не позволяет полностью охарактеризовать значимость выраженных закономерностей, когда $Q(\widetilde{S}_o)$ превышает максимальное значение $Q(\widetilde{S}_r)$, достигнутое на случайных выборках из $\widetilde{S}_{RP}$. В этих случаях p_{δ_i} очевидно оказывается равным 0. Это означает, что значимость закономерности по пороговому значению для переменной X_i соответствует уровню значимости α, меньшему $\frac{1}{N_{per}}$, где N_{per} – число перестановок. Однако α может быть как близко к $\frac{1}{N_{per}}$, так и существенно меньше $\frac{1}{N_{per}}$. Уточнение уровня α может быть достигнуто через увеличение N_{per}, но это требует слишком большого увеличения объема вычислений. Альтернативным способом более полной характеризации точности является

использование h-показателей, вычисляемых по формуле

$$h = \frac{Q_o(\widetilde{S}_o)}{\max_{\widetilde{S}_r} Q_o(\widetilde{S}_r)}. \tag{19}$$

При верификации более сложных закономерностей производится верификация необходимости включения всех их элементов. Таким образом реализуется концепция полностью верифицированной закономерности. Например, при верификации двумерной закономерности, которая задается границами, параллельными координатным осям, производится статистическая верификация необходимости включения в нее обеих границ. Предположим, что двумерная закономерность описывает зависимость целевой переменной Y от переменных X_1 и X_2. Для оценивания необходимости включения в закономерность границы для переменной X_2 быть использованы следующие подходы.

Предположим, что оптимальное на выборке $\widetilde{S}_0$ двумерное разбиение R^o_{II}, задается пороговым значением δ_1 для переменной X_1 и δ_2 для переменной X_2.

Первый подход основан на попытке опровержение гипотезы об исчерпывающем описании данных простой одномерной закономерностью с одним пороговым значением для переменной X_1. В случае, если одномерная закономерность из семейства I с порогом δ_1 по переменной X_1 исчерпывающе описывает существующую в данных зависимость, то Y не зависит от переменных X_1 и X_2 внутри областей, задаваемых неравенствами $X_1 < \delta_1$ и $X_2 < \delta_2$.

Проверку гипотезы о независимости проведем через сравнение значения $Q(R^o_{II}, \widetilde{S}_0)$ с оптимальными значениями функционала Q при построении оптимальных двумерных разбиений из семейства II по случайным выборкам, полученных из исходной истинной выборки $\widetilde{S}_0$ с помощью случайных перестановок значений Y относительно фиксированных позиций векторов переменных X_1 и X_2. При этом генерируются только перестановки без обмена значений Y между подвыборок $\widetilde{S}_l$ и $\widetilde{S}_r$ выборки $\widetilde{S}_0$. Подвыборка $\widetilde{S}_l$ включает все объекты $\widetilde{S}_0$, для которых $X_1 < \delta_1$. Подвыборка $\widetilde{S}_r$ включает все объекты $\widetilde{S}_0$, для которых $X_1 > \delta_1$.

Второй подход к оценке достоверности основан на улучшении точности аппроксимации. Пусть $R_I(\delta_1)$ – одномерное разбиение из семейства I порогом δ_1 по переменной X_1. Очевидно, что приращение качества аппроксимации при использовании двумерного разбиения из семейства III по отношению с простейшим разбиением из семейства I может оцениваться с помощью разницы

$$\Delta(\widetilde{S}_0, X_1, \delta_1) = Q(R^o_{II}, \widetilde{S}_0) - Q[R_I(\delta_1), \widetilde{S}_0].$$

Статистическую достоверность данного приращения будем оценивать через сравнение с приращением, достижимым на случайных выборках, генерируемых из $\widetilde{S}_0$ с помощью случайных перестановок значений Y относительно фиксированных позиций векторов переменных X_1 и X_2. В качестве статистической достоверности необходимости включения в закономерность границы δ_2 для переменной X_2 используется отношение

$$p_{\delta_2} = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbf{S}}_{RP} | \Delta(\widetilde{S}_r, X_1, \delta_1) \geq \Delta(\widetilde{S}_0, X_1, \delta_1)\}|}{|\widetilde{\mathbf{S}}_{RP}|} \tag{20}$$

В качестве статистической достоверности необходимости включения в

закономерность границы δ_1 для переменной X_1 используется отношение

$$p_{\delta_1} = \frac{|\{\widetilde{S}_r \in \widetilde{\mathbf{S}}_{RP} | \Delta(\widetilde{S}_r, X_2, \delta_2) \geq \Delta(\widetilde{S}_0, X_2, \delta_2)\}|}{|\widetilde{\mathbf{S}}_{RP}|} \tag{21}$$

Мы будем считать, что двумерная закономерность полностью значима на уровне α при выполнении неравенств $p_{\delta_1} < \alpha$ и $p_{\delta_2} \leq \alpha$. Минимальное значение α, при котором двумерная закономерность оказывается полностью значимой назовем индексом полной значимости по p-значениям и обозначим $p^c = \max_{i\in\{1,2\}} p_{\delta_i}$.

Как и в случае простых одномерных моделей, двумерные закономерности могут быть более точно охарактеризованы с помощью h-значений, которые вычисляются для пороговых значений по каждому из признаков по формуле

$$h_{\delta_i} = \frac{\Delta(\widetilde{S}_0, X_i, \delta_i)}{\max_{\widetilde{S}_r} \Delta(\widetilde{S}_r, X_i, \delta_i)}. \tag{22}$$

Использование h-значений позволяет повысить эффективность оценивания значимости именно для более сложных моделей, поскольку в этом случае увеличение размеров множества случайных выборок приводи к сильному увеличению объемов вычислений. Наряду с индексом полной значимости по p-значениям p^c вводим дополнительно индекс полной значимости по h-значениям: $h^c = \min_{i\in\{1,2\}} h_{\delta_i}$.

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 17-07-01362, 18-01-00557.

СПИСОК ЛИТЕРАТУРЫ

1. Adair T.H., Montani J.P. *Angiogenesis.* San Rafael, Calif.: Morgan & Claypool Life Science, 2011.
2. Киселева Е.П., Крылов А.В., Старикова Э.А., Кузнецова С.А. Фактор роста сосудистого эндотелия и иммунная система. *Успехи современной биологии.* 2009. Т. 129. № 4. С. 336–347.
3. Рудько А.С., Эфендиева М.Х, Будзинская М.В., Карпилова М.А. Влияние фактора роста эндотелия сосудов на ангиогенез и нейрогенез. *Вестник офтальмологии.* 2017. Т. 133. № 3. С. 75–81.
4. Ma L., Jin G., Yang Y., Ga Q., Ge R.L. Vascular endothelial Growth Factor as a Prognostic Parameter in Subjects with ”Plateau Red Face”. *High Alt. Med. Biol.* 2015. V. 16. № 2. P. 147–153
5. Захарова Н., Воскресенская О. Тарасова Ю. Ангиогенез и фактор роста эндотелия сосудов при цереброваскулярной патологии. *Врач. Научно-практический журнал.* 2014. № 10. С. 12–14.
6. Haymond S. Oxygen Saturation. *Clinical laboratory News.* 2006. P. 10–12.
7. Dunn O.J. Multiple Comparisons Among Means. *Journal of the American Statistical Association.* 1961. V. 56. № 293. P. 52–64.
8. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* 1979. V. 6. P. 65–70.
9. Pesarin F., Salmaso L. *Permutation Tests for Complex Data. Theory, Applications and Software.* John Wiley and Sons, Ltd., 2010. 450 p.
10. Dudoit S., Popper Shaffer J., Boldrick J. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science.* 2003. V. 18. P. 71–103.

11. Bretz F., Hothorn T., Westfall P. *Multiple Comparisons Using R.* Taylor & Francis Group, 2011. 186 p.
12. Сенько О.В., Морозов А.М., Кузнецова А.В., Клименко Л.Л. Оценка эффекта множественного тестирования в методе оптимальных достоверных разбиений. *Машинное обучение и анализ данных.* 2016. Т. 2. № 1.
13. Tusher V.G., Tibshirani R., Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*. 2001. V. 98. P. 5116–5121.
14. Ge Y., Sealfon S.C.S., Speed T.P. Multiple testing and its applications to microarrays. *Stat. Methods Med. Res.* 2009. V.18. № 6. P. 543–563.
15. Goeman J.J., Solari A. Multiple hypothesis testing in genomics. *Statistics in Medicine.* 2014. V. 33. № 11. P. 1946–1948.
16. Kuznetsova A.V., Kostomarova I.V., Sen'ko O.V. Modification of the method of optimal valid partitioning for comparison of patterns related to the occurence of ischemic stroke in two groups of patients. *Pattern Recognition and Image Analysis.* 2014. V. 24. P. 114–123.
17. 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. *Матем. биология и биоинформ.* 2013. Т. 8. № 1. С. 182–224. doi: 10.17537/2013.8.182

Рукопись поступила в редакцию 28.08.2018.
Дата опубликования 27.12.2018.