

ОЦЕНКА СТАТИСТИЧЕСКОЙ И КЛИНИЧЕСКОЙ ЗНАЧИМОСТИ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

УО «Белорусский государственный медицинский университет»

В статье представлена оценка статистической и клинической значимости в медико-биологических исследованиях. Рассмотрены типичные ошибки при анализе и представлении результатов исследования. Показано, что некорректное использование оценки статистической и клинической значимости результатов медико-биологических исследований может приводить к неверным выводам.

Ключевые слова: статистическая значимость, клиническая значимость, медико-биологическое исследование.

I. N. Moroz

EVALUATION OF STATISTICAL AND CLINICAL SIGNIFICANCE IN BIOMEDICAL RESEARCH

The article describes the evaluation of statistical and clinical significance in biomedical research. Typical errors in the analysis and presentation of research results are considered. It is shown that the incorrect use of the assessment of the statistical and clinical significance of the results of biomedical research can lead to incorrect conclusions.

Key words: statistical significance, clinical significance, biomedical research.

Многие специалисты в области статистической методологии серьезно обеспокоены тем, что в последнее время увеличивает

ся количество работ, в которых необоснованно применяется проверка значимости нулевой гипотезы и неверная интерпретация результатов

этой проверки [1–38]. В медицинских изданиях увеличивается количество работ, содержащих положительные результаты (подтверждающие гипотезу), чем отрицательные результаты, но и работ, в которых отмечается нарушения методологии проведения исследования, связанные с выбором дизайна исследования, методами статистического анализа, не соответствующие целям и задачам исследования, а также недопонимание сущности статистической и клинической значимости. По данным Kline R. B. (2004), количество критических публикаций, касающихся корректности использования статистической методологии, увеличивается экспоненциально [14, 30].

Как известно, одним из факторов, определяющим качество исследования, является понимание, владение и использование исследователями научной терминологии. Некорректное использование понятий «клиническая значимость» и «статистическая значимость» наиболее часто приводит к неправильному способу действий и оценке результатов.

В международных исследованиях для обозначения практической важности эффекта лечения используется понятие клинической значимости. Под термином «клиническая значимость (clinical significance) понимают различия в размере эффекта, которые эксперты-клиницисты считают определяющими для решений, принимаемых в клинике, независимо от уровня их статистической значимости». Некоторые авторы считают, что предпочтительней употреблять термин «clinical importance (клиническая важность), поскольку он позволяет избежать путаницы со статистической значимостью [1–37].

Клинически значимое заключение – это характеристика исходов для пациента (последствий диагностики, лечения, реабилитации). Хорошо известно критическое отношение практиков к результатам исследований, и эта критичность отчасти связана с различными критериями важности результатов (Kline R. B., 2004). Социальная, практическая и клиническая значимость имеют мало общего со значимостью статистической (Thompson, 2002) [30, 37]. Так, при оценке клинической значимости обращают внимание, например, на долю пациентов, состояние которых улучшилось, и на индивидуально-ориентированные показатели, но не на величину р-значения [1–38].

Статистически значимое заключение – это заключение, основанное на вероятности. Под термином «статистическая значимость (statistical significance) понимают статистические методы, которые позволяют оценить вероятность наблюдаемой или более высокой степени ассоциации между независимыми и зависимыми переменными при справедливости нулевой гипотезы (или степени различия между выборками и т. д.)» [1–38].

Статистическая значимость не подразумевает важность или практическую ценность результатов исследования. То есть, статистически значимое заключение может иметь мало общего с практикой медицины и не иметь значения для лечения. При учете лишь статистической значимости исследователи могут сообщить о результатах, на практике не имеющих ценности. В связи с чем рекомендуется всегда указывать величину эффекта (полученного результата) и р-значение. Показатель величины эффекта количественно выражает интенсивность эффекта (т. е. указывает на медицинскую эффективность) и упрощает принятие решения о его практическом значении и применении [1–38].

При проведении медико-биологических исследований устанавливаются различные варианты статистической и клинической значимости полученных результатов, требующих различной оценки и действий (таблица 1). При анализе вариантов мы должны помнить, что клинически важное (значимое) заключение – это заключение, у которого есть последствия для лечения пациента, которое не всегда может быть статистически значимым. Клинически важное заключение в единственном случае, вероятнее всего, не устанавливает биологическую связь (статистически значимую) [1–38].

Таблица 1. Варианты установления статистической и клинической значимости при медико-биологических исследованиях

Клиническая значимость (clinical significance)	Статистическая значимость (statistical significance)	
	значимо	незначимо
значимо	A	B
незначимо	C	D

Вариант «А». Заключение, которое клинически и статистически значимо (как правило, характерно для 5–25 % всех возможных связей),

весьма важно и ценно, в связи с чем мы, вероятно, будем полагать, что это есть результат биологического процесса, общего для группы пациентов, и что он, возможно, поддается измерению, объяснению, предсказанию и управлению. Результаты свидетельствуют о наличии эффекта, получены на достаточном объеме выборки.

Вариант «В». При сравнении небольших групп можно выявить существенное различие показателей, имеющих клиническое значение, но при этом не быть статистически значимым. Однако, не стоит слепо доверять клинически значимым результатам, полученным на ограниченной выборке пациентов, без подтверждения статистической значимости вывода. Особое внимание следует обратить на размер выборки. Исследования, проведенные с участием 10 пациентов и 10 000 пациентов, имеют разную ценность. Чаще всего, чем больше выборка, тем достовернее результаты. Именно поэтому на этапе планирования исследования, разработке его дизайна необходимо рассчитать оптимальный объем выборки, обеспечивающий получение достоверных данных. При определении объема выборки необходимо руководствоваться рекомендациями специалиста в области статистики (использовать специальные формулы, номограмму Альтмана и др.). Применять мета-анализ для получения баз данных, основанных на объединении результатов, полученных в разных исследованиях, что обеспечит увеличение объема выборки пациентов и существенно повысит уровень доказательности результатов [1–38].

Исследователю необходимо помнить (вариант «С»), что незначительное изменение параметра можно доказать, как статистически значимое при большом размере выборки (т. е. статистически значимое различие может не быть клинически важным). Нельзя подменять клиническую значимость величиной статистической значимости несущественного эффекта. Такие результаты могут быть обусловлены отсутствием эффекта в объективной реальности; либо связаны с недостаточным объемом выборки, или выборка не соответствует по своим характеристикам генеральной совокупности. Результаты сравнения эффективности использования 2-х искусственных водителей ритма у нескольких тысяч пациентов, описанные Т. Lang (2004), показали, что «среднее различие в 0,25 месяца на протяже-

нии 5 лет наблюдения клинически незначимо, даже если оно объясняется случайностью не менее чем в 1 из 1000 случаев ($p < 0,001$)». В этом примере автор показал, что даже несущественное различие, выявленное при сравнении больших групп, может оказаться статистически значимым, но не иметь при этом клинического значения [7, 9, 11].

Таким образом, незначительные различия параметров, выявленные при больших группах, могут оказаться статистически значимыми, но не иметь при этом клинического значения (т. е. клинически бессмысленными), даже если он подтверждается случайностью менее чем в 1 из 1000 случаев ($p < 0,001$) [1–38].

Исследователь должен понимать, что на основании p -величины нельзя судить ни о степени этого различия, ни о его клинической важности. И наоборот, даже существенное различие, выявленное при сравнении небольших групп, может иметь клиническое значение, но не быть при этом статистически значимым (пример: если в ходе исследования, включающего несколько пациентов в терминальном состоянии, хотя бы один из участников какой-либо группы выживет, то статистически значимое различие уровня летальности или частоты выживания между группами может отсутствовать).

Вариант «D». Заключение, которое клинически и статистически незначимо. Результаты свидетельствуют об отсутствии эффекта в объективной реальности о наличии эффекта, получены на недостаточном объеме выборки.

Статистическая значимость является одним из основных аспектов проверки гипотезы. Статистическая значимость – максимально допускаемая исследователем вероятность ошибочного отклонения нулевой гипотезы (об отсутствии различия между группами, об отсутствии взаимосвязи признаков и т. п.) на основании результатов применения статистического теста. Большинство проверяемых гипотез в медико-биологических исследованиях основаны на сравнении между собой групп людей, которые испытывают влияние различных факторов (например, сравнение эффективности двух видов лечения, обеспечивающих снижение 5-летней смертности пациентов от рака молочной железы). Проверка гипотезы может подтвердить или отклонить утверждение о том, что наблюдаемые результа-

ты не случайны, а отражают связь между переменными. Общий подход таков: выдвигается нулевая гипотеза (H_0) о том, что различий в результате применения разных технологий лечения нет (например, 5-летняя смертность от рака молочной железы одинакова в группах с разными методами лечения). Альтернативная гипотеза (H_1) заключается в том, что различия есть, и они не случайны (двухсторонняя альтернативная гипотеза будет означать, что различие эффектов не равно нулю). Выдвигаются две противоречащие друг другу гипотезы: нулевая гипотеза (H_0) о том, что связь (зависимость) случайна, альтернативная (H_1) ей противоречит. Доказав, что гипотеза H_0 несостоятельна, подтверждается, что альтернативная гипотеза верна [1–38].

При проверке гипотез возникают ошибки, которые подразделяют на ошибки 1-го и 2-го типа (таблица 2). Ошибка 1-го типа (обозначается α) определяется как вероятность обнаружить различие, тогда как в действительности оно отсутствует («ложноположительный результат»). Другими словами, это вероятность неправомерно отбросить нулевую гипотезу (H_0) в пользу альтернативной (H_1).

Таблица 2. Типы ошибок и статистическая мощность исследования

Результаты проверки гипотезы	Истинный, но неизвестный характер взаимодействия	
	Гипотеза H_0 не верна	Гипотеза H_0 верна
Отвергнуть гипотезу H_0	Корректное решение (достаточная статистическая мощность)	Ошибка 1 типа (α)
Принять гипотезу H_0	Ошибка 2 типа (β)	Корректное решение

Уровень значимости (significance level) – уровень вероятности, на котором принимается решение об отклонении нулевой гипотезы (чаще всего принимается равным 0,05), когда она верна. Другими словами, это допустимая (с точки зрения исследователя) вероятность совершения статистической ошибки 1-го типа – ошибки того, что различия сочтены существенными, а они на самом деле случайны. Обычно используют уровни значимости (обозначаемые α), равные 0,05, 0,01 и 0,001 [1–38]. В настоящее время в связи с развитием компьютерных технологий используется точное значение ошибки. Согласно теории вероятностей и кривой Гаусса, для медико-биологических исследований уровень значимости равен 0,0456 (округленно 0,046). При описании необходимо применять значение, которое дает статистическая программа, а зна-

чение 0,046 при оценке результатов исследования стало определяющим.

В медико-биологических исследованиях в качестве критического порога значимости традиционно выбирается уровень 0,05, что допускает наличие ошибки 1-го типа 5 раз в 100 сравнениях. При $p \leq \alpha$ различия принимаются статистически значимыми. И чем меньше р-значение, тем меньше подобных ошибок: например, при $p = 0,01$ считается, что ошибка 1-го типа возможна 1 раз в 100 сравнениях, при $p = 0,001$ – 1 раз в 1000 сравнениях. Однако в разведочных/пилотных исследованиях допускается уровень значимости $\alpha = 0,1$ для выявления намечающихся различий и/или взаимосвязей с целью дальнейшего планирования на их основе новых исследований с достаточной значимостью [1–38].

Различные значения α -уровня имеют свои достоинства и недостатки. Меньшие α -уровни дают большую уверенность в том, что уже установленная альтернативная гипотеза подтверждена, но при этом есть больший риск не отвергнуть ложную нулевую (или отвергнуть истинную альтернативную) гипотезу (ошибка второго типа,

или «ложноотрицательное решение»), и таким образом меньшая статистическая мощность. Выбор α -уровня неизбежно требует компромисса между значимостью и мощностью, и, следовательно, между вероятностями ошибок 1-го и 2-го типа. Наиболее часто в биомедицинских исследованиях предельно допустимый предел ошибки 1-го типа обычно устанавливается на уровне 5 %, а ошибка 2-го типа – не более 20 % ($\alpha = 0,05$; $\beta \leq 0,2$) [1–38].

Ошибка 2 типа (обозначается β) – это вероятность сделать вывод об отсутствии различия, в то время как фактически оно имеется («ложноотрицательный результат»), т. е. неправомерно принять H_0 . В медико-биологических исследованиях в качестве критического порога принимается значение ошибки 2-го типа $\beta = 0,1$ или $\beta = 0,2$, что в виде статистической мощнос-

ти $(1 - \beta)$, выраженной в процентах, равно 90 % или 80 %, чаще всего – 80 %: вероятность того, что из 100 в 80 случаях действительно существующее различие будет выявлено и в 20 случаях – нет [1–38].

При планировании исследования необходимо определить статистическую мощность (statistical power), которая вычисляется как $(1 - \beta)$ и означает вероятность сделать заключение о наличии различия, в то время как оно имеется на самом деле (т. е. получить «истинно положительный результат») [1–38].

Статистическая мощность зависит от размера выборки (поскольку связана со стандартной ошибкой, которая в свою очередь уменьшается с увеличением размера выборки), а также от степени различия, которое ожидается обнаружить. Выявление больших различий требует меньшего числа наблюдений и, наоборот, для определения небольших различий потребуется более многочисленная выборка [1–38]. Если планируемая численность выборки не обеспечивает приемлемый уровень статистической мощности ($\geq 80\%$), чтобы убедительно отвергнуть нулевую гипотезу или согласиться с ней, результаты исследования не будут доказательными. Например, если исследователь хочет определить различие «в среднем весе тела между двумя группами (получавшими и не получавшими препарат, снижающий аппетит) и хочет доказать разницу в 1 кг при стандартном отклонении 10 кг в контрольной и изучаемой группах, то при $\alpha = 0,05$ и мощности 80% необходимо иметь не менее 1570 людей в каждой группе. Однако, если необходимо оценить различие в 5 кг при стандартном отклонении 10 кг, достаточно включить в группы по 64 человека» [10–12]. Расчет объема выборки для мощности 80 % и 2-стороннего критерия уровня значимости $\alpha = 0,05$ осуществлялся с использованием формулы Лера [11]:

$$\frac{16}{(\text{стандартизованная разность})^2} = \\ = \frac{16}{(5 \sqrt{10})^2} = 64.$$

В процессе проверки гипотезы вычисляют значение статистического критерия и уровень статистической значимости p , который сравнивают с заданным априори пороговым уровнем

значимости p_0 . При $p < p_0$ нулевая гипотеза (об отсутствии различий групп и т. п.) отклоняется и принимается альтернативная гипотеза о том, что различия групп существуют и являются статистически значимыми. Обращаем внимание, что если по результатам теста нулевая гипотеза не отклоняется, то это не означает, что различия групп отсутствуют. Причин может быть две: недостаточные объемы выборок и/или отсутствие эффекта. При проверке гипотез нельзя забывать о так называемой проблеме множественных сравнений, которая заключается в следующем: чем больше статистических гипотез проверяется на одних и тех же данных, тем более вероятно заключение о наличии различий между группами (либо наличии статистической связи признаков), в то время как на самом деле верна нулевая гипотеза об отсутствии различий/связей [1–11]. Реброва О. Ю. и другие авторы считают, что учет множественных сравнений следует начинать в тех случаях, когда число рассчитываемых значений p превышает 10 [1–13, 16]. В таких случаях рекомендуется использовать поправку Бонферрони (вычисление порогового уровня статистической значимости с учетом числа проверенных гипотез), специальные методы множественного сравнения групп, дисперсионный анализ [1–13, 15, 16].

Для принятия решений о том, какую из гипотез (нулевую или альтернативную) следует принять, используют статистические критерии, которые включают в себя методы расчета определенного показателя (p -значения), на основании которого принимается решение об отклонении или принятии гипотезы, а также правила (условия) принятия решения.

Решение, какой статистический критерий (тест) использовать для проверки гипотезы, зависит от цели и задач исследования, дизайна, типа переменной и распределения данных, объема выборки. Для данных, полученных в метрических шкалах (интервальных или относительных) при распределениях, близких к нормальным, используют параметрические методы, основанные на таких показателях, как среднее и стандартное отклонение. В частности, для определения статической значимости различия средних для двух выборок применяют критерий Стьюдента, а для того чтобы судить о различиях между тремя или большим числом выборок, – F-тест

или дисперсионный анализ (ANOVA). Данные, полученные в неметрических (номинативных или порядковых) шкалах или выборки слишком малы для уверенности в том, что генеральные совокупности, из которых они взяты, подчиняются нормальному распределению, оценивают, используя непараметрические методы – критерий Chi-square (χ^2 , хи-квадрат), Манна-Уитни, Уилкоксона и др. Выбор статистического метода также зависит от того, являются ли выборки независимыми (т. е. взятыми из двух разных групп исследуемых) или зависимыми (т. е. отражающими результаты одной и той же группы исследуемых до и после воздействия) [1–38].

Кроме того, для проверки статистической гипотезы можно использовать доверительный интервал (ДИ) – интервал значений признака, рассчитанный по выборке для какого-либо параметра распределения или показателя величины эффекта и с определенной вероятностью (например, 95 % в случае 95 % ДИ). Ширина ДИ зависит от объема выборки и вариабельности в ней: чем шире ДИ, тем менее точной является выборочная оценка. При увеличении числа наблюдений ДИ сужается, и точность оценки увеличивается [1–16, 23].

Вместе с тем, необходимо отметить различия двух указанных подходов: проверка гипотез позволяет оценить вероятность неверного отклонения нулевой гипотезы (однако ничего не говорит о вероятности альтернативной гипотезы), а ДИ позволяет при фиксированном уровне допустимой ошибки оценить размер эффекта и его точность.

В последнее время для проверки гипотез вместо частотного подхода начинает широко использоваться байесовский подход, который рассматривают как способ переоценки научных представлений с помощью вновь полученных данных. Goodman S. N., описывая метод Байеса, указал, что его можно использовать для количественной оценки доказательности данных, основой которой является байесовский критерий, который в простейшей его форме называют также отношением правдоподобия [22]. Минимальный байесовский критерий представляет собой объективный показатель, способный с успехом заменить величину p . В отличие от величины p теоретическое обоснование и интерпретация байесовского критерия позволяют исполь-

зовать его как в процессе проверки гипотез, так и в процессе принятия решений. Байесовский фактор показывает, что оценка данных с помощью величины p преувеличивает доказательство, опровергающие нулевую гипотезу [7, 11, 22]. И самое главное, байесовский критерий подразумевает включение в анализ прошлого опыта в виде вероятности, что тот или иной вывод правилен либо ошибочен. Байесовский критерий отделяет логическое умозаключение от данных опыта и в то же время дает исследователю возможность комбинировать старую и новую информацию [7, 11, 22].

Величина – p -значение является количественной характеристикой только лишь статистической, но не клинической значимости. Чем это значение больше, тем больше вероятность отсутствия различий в пользу нулевой гипотезы, и чем оно меньше, тем больше вероятность наличия различий в пользу альтернативной гипотезы.

Результаты научных исследований в области медицины нельзя описывать как «значимые» или «незначимые», их следует интерпретировать в контексте характера исследования и других имеющихся данных. При описании результатов исследования необходимо указывать не только величину эффекта, но должны быть приведены точные значения p , а не данные их сравнения с произвольно установленными порогами. Исследователь должен помнить, что бы ни показывали статистические данные, важно уметь аргументировать окончательные результаты и анализировать все данные в совокупности, а не сознательно выбранные для обсуждения.

Литература

1. Банержи, А. Медицинская статистика понятным языком: вводный курс / А. Банержи. – М.: Практическая медицина, 2007. – 287 с.
2. Власов, В. В. Эпидемиология / В. В. Власов. – М.: Гэотар-Мед, 2004. – С. 68, 250–253, 332.
3. Власов, В. В. Значение научных публикаций в специализированных журналах / В. В. Власов // Росс. вестник акушера-гинеколога. – 2010. – № 10 (4). – С. 72–75.
4. Гланц, С. Медико-биологическая статистика / С. Гланц. – М.: Практика, 1999. – 334 с.
5. Гржибовский, А. М. Выбор статистического критерия для проверки гипотез / А. М. Гржибовский // Экология человека. – 2008. – № 11. – С. 48–57.

6. Гринхальх, Т. Основы доказательной медицины / Т. Гринхальх. – М.: ГЭОТАР-МЕД, 2004. – 240 с.
7. Ланг, Т. А., Сесик М. Как описывать статистику в медицине. Аннотированное руководство для авторов, редакторов и рецензентов / Т. А. Ланг, М. Сесик; пер. с англ. под ред. В. П. Леонова. – М.: Практическая медицина, 2011. – 480 с.
8. Ласта, Джон М., Власов В. В., Вербицкая Е. В. Эпидемиологический словарь. – М., 2009. – 316 с.
9. Леонов, В. П. Ошибки статистического анализа биомедицинских данных / В. П. Леонов // Международный журнал медицинской практики. – 2007. – № 2. – С. 19–35, <http://www.biometrica.tomsk.ru/error.htm>.
10. Павлова, В. Ю. Основные вопросы статистического анализа в медицинских исследованиях / В. Ю. Павлова // Клиническая онкогематология. – 2009. – Т. 2, № 4. – С. 374–377.
11. Петри, А. Наглядная медицинская статистика / А. Петри, К. Сэбин. – М.: ГЭОТАР-МЕД, 2010. – 169 с.
12. Реброва, О. Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ Statistica / О. Ю. Реброва. – М.: МедиаСфера, 2006. – С. 166–176.
13. Реброва, О. Ю. Описание статистического анализа данных в оригинальных статьях. Типичные ошибки / О. Ю. Реброва // Медицинские технологии. – 2011. – № 4. – С. 36–40.
14. Сивуха, С. В., Козьяк, А. А. О реформе статистического вывода в психологии: сомнительная значимость статистической значимости / С. В. Сивуха, А. А. Козьяк // Психология. Журнал Высшей школы экономики. – 2009. – Т. 6, № 4. – С. 66–86.
15. Флетчер, Р., Флетчер С., Вагнер Э. Клиническая эпидемиология. Основы доказательной медицины. – М.: Медиа Сфера, 2004. – 297 с.
16. Altman, D. G. Statistical guidelines for contributors to medical journals / D. G. Altman, S. M. Gore, M. J. Gardner, S. J. Pocock // British Medical Journal. – 1983. – № 286. – P. 1489–1493.
17. Berger, J. O., Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence // J Am Stat Assoc. – 1987; 82: 112–22.
18. Brophy, J. M., Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA. 1995; 273: 871–5.
19. Burton, P. R., Gurrin L. C., Campbell M. J. Clinical significance not statistical significance: a simple Bayesian alternative to pvalues // J Epidemiol Community Health. 1998; 52: 318–23.
20. Gardner, M. J., Altman. D. G. Confidence intervals rather than P values: estimation rather than hypothesis testing. / M. J. Gardner, D. G. Altman // BMJ. – 1986. – 292: 746–50.
21. Goodman, S. N. Toward evidence-based medical statistics. 1: The P value fallacy // Ann Intern Med. 1999; 130:995–1004.
22. Goodman, S. N. Towards evidence-based medical statistics: 1: The Bayes factor // Ann Intern Med. – 1999;130:1005–13.
23. Feinstein, A. R. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998; 51:355–60.
24. Fisher, R. A. Statistical methods and scientific inference. – London: Collins Macmillan, 1973.
25. Fisher, R. A. Statistical methods for research workers. – London: Oliver and Boyd, 1950. – 80 p.
26. Freiman, J. A., Chalmers T. C., Smith H. J., Kuebler R. R. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 “negative” trials // N Engl J Med. 1978: 299: 690–1.
27. Freiman, J. A., Chalmers T. C., Smith H., Kuebler R. R. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. In: Bailar J. C., Mosteller F. eds. Medical uses of statistics. – Boston, Ma: NEJM Books, 1992. – S. 357–73.
28. Easterbrook, P. J., Berlin J. A., Gopalan R., Matthews D. R. Publication bias in clinical research // Lancet. – 1991; 337: 867–72.
29. Edwards, W., Lindman H., Savage L. J. Bayesian statistical inference for psychological research // Psychol Rev. 1963; 70:193–242.
30. Kline, R. B. Beyond significance testing: Performing data analysis methods in behavioral research. – APA, 2004.
31. Lang, T. Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles // Croatian Medical Journal. – 2004. – № 45(4). – P. 361–370.
32. Lang, T., Secic M. How to report statistics in medicine: annotated guideline for authors, editors, and reviewers. – Philadelphia (PA): American College of Physicians, 1997.
33. Lehmann, E. L. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? // J. Am Stat Assoc. – 1993; 88: 1242–9.
34. Moher, D., Dulberg C. S., Wells G. A. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272: 122–4.
35. Mulward, S., Gutzsche P. C. Sample size of randomized double-blind trials 1976–1991 // Dan Med Bull. 1996; 13: 96–8.
36. Sterne, J. A. C., Smith, G. D. Sifting the evidence – what’s wrong with significance tests? / J. A. C. Sterne, G. D. Smith // British Medical Journal. – 2001. – Vol. 322. – P. 226–31.
37. Thompson, B. «Statistical», «practical» and «clinical»: How many kinds of significance do counselors need to consider? // Journal of Counseling and Development. 2002. 80. 1. 64–71.
38. Yancy, J. M. Ten rules for reading clinical research reports / J. M. Yancey // Am J Surg. – 1990. – № 159(6). – P. 533–9.

Поступила 09.01.2020 г.