ДВАДЦАТЬ ОШИБОК СТАТИСТИЧЕСКОГО АНАЛИЗА, КОТОРЫЕ ВЫ САМИ МОЖЕТЕ ОБНАРУЖИТЬ В БИОМЕДИЦИНСКИХ СТАТЬЯХ*

Т. Ланг

«Критические обзоры убедительно свидетельствуют, что примерно в половине научных работ, выполняемых с использованием статистических методов анализа данных, этот анализ проводится с ошибками» [1].

«Исследования высокого методологического качества заслуживают соответствующего изложения материала, и хорошее представление результатов является важной частью исследования, такой же, как сбор и анализ данных. При чтении художественной литературы мы оцениваем мастерство автора. Давайте также признаем право научных знаний на достойное изложение» [2].

Первая попытка обсудить вопросы, связанные со статистической вероятностью, в медицинской литературе была предпринята в 30-х годах XX века [3]. С тех пор исследователи в разных областях медицины обнаружили множество ошибок при проведении статистического анализа даже в работах, результаты которых были опубликованы в наиболее авторитетных научных журналах [4—7]. Неверное отображение статистических данных представляет собой давнюю и широко распространенную проблему, чреватую серьезными последствиями. Проблема недостаточно хорошо осознана, несмотря на то что большинство ошибок возникает при использовании простейших статистических методов, и для того, чтобы избежать их появления, достаточно следовать нескольким рекомендациям [8].

С распространением движения доказательной медицины возросло понимание актуальности проблем, связанных с низким качеством изложения статистического материала. В основе доказательной медицины лежит использование опубликованных в медицинской литературе исследований, поэтому она зависит от методологического качества статей. Соответственно несколько авторских коллективов предложили свои руководства по представлению результатов различных исследований [9—11]; кроме того, появилась серия детальных руководств по изложению результатов статистического анализа [12].

В этой статье приведено 20 рекомендаций, затрагивающих наиболее часто встречающиеся в медицинской литературе аспекты статистического анализа. Они предназначены для авторов, редакторов и рецензентов, не являющихся специалистами в области статистики. Предлагаемый вниманию читателей материал представляет собой верхушку айсберга; при необходимости более подробные сведения можно получить, обратившись к соответствующим руководствам [12], а также к работам, указанным в библиографическом списке. Для облегчения знакомства с этой не очень увлекательной для врачей проблемой рекомендации приводятся в порядке возрастания значимости.

Ошибка 1. Количественные данные представлены с излишней точностью

Большинство из нас легче воспринимают количественные данные, представленные одной или двумя цифрами, чем тремя и более. Поэтому округление улучшает восприятие материала [13]. Рассмотрим пример, в котором количество участников (как мужчин, так и женщин) на момент окончания исследования примерно в 3 раза превышает таковое в начале, однако этот факт становится очевидным лишь после округления соответствующих показателей:

Число женщин возросло с 29 942 до 94 347, мужчин — с 13 410 до 36 051.

Число женщин возросло с 29 900 до 94 300, мужчин — с 13 400 до 36 000.

Число женщин возросло примерно с 30 000 до 94 000, мужчин — с 13 000 до 36 000.

Во многих случаях необязательно приводить максимально точные значения. Если масса тела больного составляет 60 кг, то использование показателя 60,18 кг только затруднит восприятие, несмотря на то что формально он соответствует действительности. По этой же причине наименьшая величина p, которую имеет смысл представлять, $p<0{,}001$.

Ошибка 2. Непрерывные данные представлены в виде порядковых без объяснения причин и способа преобразования

Для облегчения статистического анализа непрерывные данные можно представить в виде двух и более порядковых категорий, например рост в см как низкий, нормальный и высокий. Однако такое упрощение уменьшает точность результатов и вариабельность данных. Автор должен объяснить, почему он сделал это. Кроме того, он должен описать критерии выбора диапазона значений в рамках каждой из порядковых категорий, чтобы избежать возможности появления систематической ошибки [12]. В некоторых случаях преобразование непрерывных данных в порядковые имеет целью подгонку конечных результатов под желаемую схему (рис. 1).

* Переведено с разрешения издателя. Впервые опубликовано T. Lang. Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles. Croatian Medical Journal 2004;45(4):361—370.

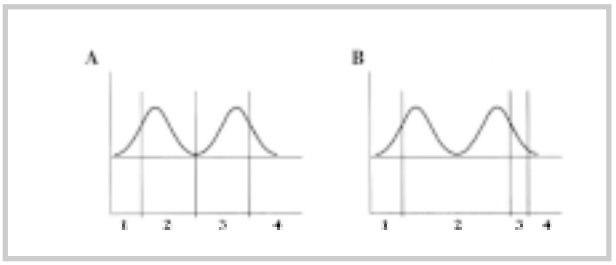

Рис. 1. **Чтобы избежать возникновения систематической ошибки, автор должен объяснить, почему и каким образом непрерывные данные были преобразованы им в порядковые.**

A. Преобразование выглядит оправданным.
B. Целесообразность преобразования требует объяснения.

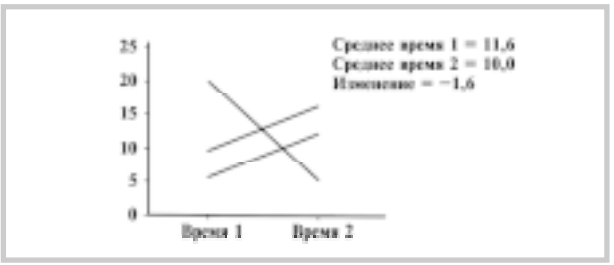

Рис. 2. **Парные данные должны быть представлены таким образом, чтобы были очевидными изменения как индивидуальных, так и групповых характеристик.**

В данном примере результат можно интерпретировать как среднее уменьшение показателя на 1,6 либо как его увеличение у 2 из 3 участников.

Ошибка 3. Представлены средние групповые значения для парных данных без сообщения размера изменений внутри групп

Данные, относящиеся к одному и тому же участнику исследования, называются парными. При последовательных измерениях признака величина как средних групповых, так и индивидуальных значений может различаться от измерения к измерению. Однако если в статье представлены только групповые средние значения, читатель может не заметить изменения индивидуальных показателей (рис. 2). Пока не будут отображены индивидуальные значения, альтернативный вариант трактовки данных будет неочевиден. Например, результаты, приведенные на рис. 2, можно интерпретировать как среднее уменьшение показателя в группе от момента 1 к моменту 2, либо как увеличение показателя у 2 из 3 участников. И то и другое соответствует истине, но если в статье представлен лишь один из этих выводов, читатель может сделать неверное заключение о результатах.

Ошибка 4. Неправильно используются статистические характеристики данных

При описании непрерывных данных наиболее часто используют такие понятия, как средняя величина и среднее квадратическое отклонение (СКО). Однако эти показатели применимы только при условии нормального или Гауссова распределения значений. При нормальном распределении в 68% случаев результаты измерений лежат в пределах ±1 СКО от среднего значения, в 95% случаев — в пределах ±2 СКО, в 99% случаев — в пределах ±3 СКО. При асимметричном распределении эта закономерность отсутствует, поэтому средняя величина и СКО не дают представления о характере кривой. Вместо них используют другие показатели, такие как медиана (50-й центиль, или точка, которая делит данные на две равные части) и межквартильный диапазон (обычно от 25-го до 75-го центиля) [14].

Хотя для определения средней величины и СКО достаточно результатов двух последовательных измерений, эти показатели недостаточно хорошо описывают данные исследований с небольшим числом участников (малые выборки). Кроме того, большинство биологических показателей не подчиняются нормальному распределению [15]. Исходя из этого, в медицинской литературе такие термины, как медиана, диапазон и межквартильный диапазон должны встречаться чаще, чем средняя величина и СКО.

Ошибка 5. Стандартная ошибка средней величины используется для описательного анализа данных или в качестве показателя точности оценки

Средняя величина и СКО описывают центральную тенденцию и вариабельность данных, подчиняющихся нормальному распределению, полученных в выборке. Средняя величина и стандартная ошибка средней величины (СОС) являются точечной оценкой (средняя величина) и показателем ее точности (СОС) для характеристики популяции. Однако СОС всегда меньше, чем СКО, поэтому иногда представляют именно ее, чтобы результаты измерений выглядели более точными [16]. Хотя СОС отражает точность измерения (в пределах средняя величина ±1 СОС лежит популяционная средняя с вероятностью 68%, т.е. это 68% доверительный интервал, ДИ), в медицинских исследованиях предпочтительно использовать 95% (ДИ) [17]. Таким образом, среднюю величину и СОС применяют для характеристики как выборки, так и популяции. Чтобы избежать путаницы, следует запомнить: среднюю величину и СКО предпочтительно использовать для обобщенной характеристики данных, подчиняющихся нормальному распределению, а среднюю величину и 95% ДИ — в качестве точечной оценки и уровня ее точности.

Например, если средняя масса тела у 100 мужчин составляет 72 кг, а СКО — 8 кг, то (при условии нормального распределения значений) примерно в $^2/_3$ (68%) случаев результат измерения будет лежать в диапазоне от 64 до 80 кг. Данный пример показывает правильное использование средней величины и СКО для характеристики распределения значений.

Средняя величина, составляющая 72 кг, также служит наиболее точным значением средней массы тела всех мужчин в популяции, из которой сформирована выборка. Используя формулу СОС=СКО/ЦN, и подставляя СКО=8 кг, а N=100 измерениям, получим СОС=0,8. Это означает, что при повторном определении массы тела в аналогичной (случайной) выборке мужчин из данной популяции примерно в 68% случаев средняя масса тела составит от 71,2 до 72,8 кг (диапазон значений в пределах ±1 СОС).

Для точечной оценки и определения уровня ее точности предпочтительно использовать среднюю величину и 95% ДИ (диапазон значений в пределах ±2 СОС). В рассмотренном чуть выше примере правильной будет фраза: средняя масса тела составляет 72 кг при 95% ДИ от 70,4 до 73,6 кг. Это означает, что при повторном измерении данного показателя в аналогичной (случайной) выборке мужчин в той же популяции примерно в 95% случаев средняя масса тела составит от 70,4 до 73,6 кг.

Ошибка 6. Для описания различий между группами используется только величина *p*

Использование величины *p* для оценки статистической значимости часто неоправдано [18]. Даже при условии корректного применения данный показатель имеет целый ряд ограничений. В большинстве случаев вместо величины *p* либо дополнительно к ней следует указывать абсолютное различие в частоте изучавшегося события между группами (относительное или выраженное в процентах различие может быть истолковано неверно) и его 95% ДИ. Ниже приводятся встречающиеся в статьях формулировки в порядке возрастания их методологического качества.

«Эффект от применения лекарственного средства оказался статистически значимым». Данная формулировка не позволяет определить ни величину эффекта, ни его клиническую, ни статистическую значимость. Читатель может заключить, что характеристика эффекта как «статистически значимого» в данной ситуации означает целесообразность использования препарата.

«Эффект от использования средства, заключающийся в снижении уровня диастолического артериального давления (АД), оказался статистически значимым ($p<0,05$)». И в этом случае отсутствует указание на величину эффекта, поэтому его клиническая значимость остается неясной. Кроме того, величина *p* может составлять 0,049; такое различие статистически значимо, но настолько близко к пороговой величине (0,05), что практически не отличается от, к примеру, 0,051, т. е. статистически незначимого уровня. Наличие подобной условной черты (0,05) представляет собой одну из проблем при использовании величины *p*.

«Среднее диастолическое АД в группе лечения уменьшилось со 110 до 92 мм рт. ст. ($p=0,02$)». Пожалуй, такая формулировка встречается наиболее часто. В ней отражены результаты измерений до и после вмешательства, однако не указано различие между ними. Среднее уменьшение показателя на 18 мм рт. ст. статистически значимо, однако это лишь точечная оценка. В отсутствие 95% ДИ нельзя определить, насколько она точна, и, следовательно, практически значима.

«Использование препарата привело к снижению уровня диастолического АД в среднем на 18 мм рт. ст. (со 110 до 82 мм рт. ст.) при 95% ДИ от 2 до 34 мм рт. ст. ($p=0,02$)». Границы ДИ свидетельствуют, что при использовании данного препарата в 100 выборках, аналогичных изучавшейся, среднее снижение АД в 95 из них будет лежать в пределах от 2 до 34 мм рт. ст. Уменьшение на 2 мм рт. ст. клинически незначимо в отличие от снижения на 34 мм рт. ст. Таким образом, хотя среднее уменьшение АД оказалось статистически значимым, эффект от использования препарата в других испытаниях может оказаться клинически незначимым, т. е. полученные в исследовании результаты не позволяют сделать окончательного вывода о целесообразности вмешательства.

Если оба показателя, определяющих как верхний, так и нижний пределы ДИ, клинически значимы, можно полагать, что вмешательство клинически эффективно. Когда оба показателя клинически незначимы, вероятнее всего, вмешательство неэффективно. Может оказаться, что клинически значим только один из показателей; в таком случае следует провести исследование с большим числом участников.

Ошибка 7. Отсутствует подтверждение того, что анализируемые данные соответствуют предположениям, лежащим в основе использованных статистических методов

Существуют сотни методов статистического анализа данных. В каждом конкретном случае можно выбрать несколько возможных вариантов анализа [19]. Однако при несоблюдении критериев использования того или иного метода полученный результат может оказаться неточным. По этой причине в тексте статьи должно присутствовать название использованного метода и подтверждение того, что он применим для анализа имеющихся данных.

Например: полученные результаты подчиняются нормальному распределению, что позволяет использовать t-тест.

Наиболее характерные ошибки:

— использование параметрических методов (основанных на предположении о нормальном распределении данных) для анализа данных, не подчиняющихся нормальному распределению (в частности, при сравнении двух групп нередко используют критерий Стьюдента, хотя более оправдано применение критерия Вилкоксона или другого непараметрического метода);

— использование методов, предназначенных для независимых выборок, при анализе парных данных (в этом случае нередко применяют критерий Стьюдента, а не парный t-тест).

Ошибка 8. Использование линейной регрессии без подтверждения линейного характера связи

В разделе 7 уже упоминалось, что в любой статье, включающей в себя описание статистического анализа, должно быть указано, применимы ли выбранные методы для анализа имеющихся данных [12]. Особенно это важно при использовании линейной регрессии, подразумевающей линейный характер связи между независимой переменной и исходом. В противном случае полученный результат окажется неверным.

Подтвердить линейный характер зависимости можно с помощью изучения остатков, т. е. различий между наблюдаемыми и прогнозируемыми при помощи модели величинами (рис. 3). Если при отображении в виде графика остатки представляют собой

прямую линию, а их значения приближаются к нулю, то можно говорить о линейном характере зависимости (рис. 4A). Если графическое изображение остатков имеет иной вид (рис. 4B, 4C, 4D), это свидетельствует о нелинейном характере зависимости. Изучение остатков необходимо, поскольку сам по себе график линейной регрессии не всегда позволяет верно оценить характер зависимости (рис. 5).

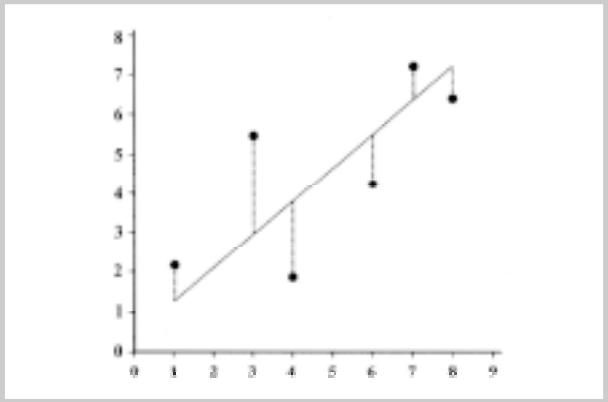

Рис. 3. **Остатки представляют собой различие между наблюдаемыми и прогнозируемыми при помощи регрессионной модели величинами.**

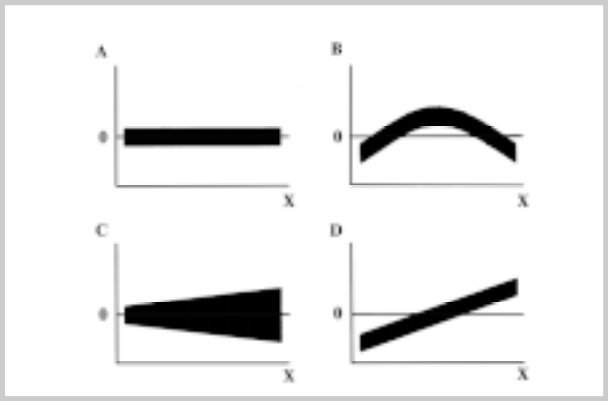

Рис. 4. **Когда график остатков свидетельствует, что их величины приближаются к нулю на протяжении всего диапазона значений, зависимость имеет линейный характер (A).**

Иной вид графического изображения остатков (B, C, D) свидетельствует о нелинейном характере зависимости, для описания которой линейная регрессия непригодна.

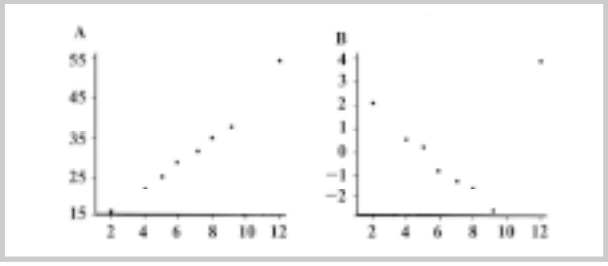

Рис. 5. **Впечатление о линейном характере зависимости может оказаться обманчивым.**

В данном примере зависимость выглядит линейной (A), однако в действительности это не так, о чем свидетельствует график остатков (B).

Ошибка 9. В анализ включены не все данные и не все участники

Пропуски в данных встречаются довольно часто и крайне отрицательно сказываются на общем впечатлении от статьи, поскольку у читателя может возникнуть предположение, что автор недостаточно внимателен или попросту ленив [20]. При обнаружении пропущенных данных возникают следующие вопросы:

— Причина пропуска данных. Включены ли в анализ минимальные и максимальные результаты? Пропущены ли данные из-за ошибки в лаборатории? Возможно, данные пропущены, поскольку они противоречат выводам автора?

— Воспроизводимость полученных результатов. Является ли указанный диапазон значений таковым в действительности? Так ли невелико число выбывших из исследования?

— Методологическое качество исследования в целом. Если итоговые данные не совпадают в исследовании, насколько автор был точен при описании других аспектов работы?

Одним из наиболее удобных способов отображения данных об участниках клинического испытания служат потоковые диаграммы (flow charts, рис. 6) [9, 12, 21]. Такое наглядное изображение позволяет читателю получить представление о количестве участников на каждом из этапов испытания, понять структуру исследования и визуально оценить соотношение между группами и подгруппами. Именно этот способ представления данных рекомендован в руководстве CONSORT [9].

Ошибка 10. Не указано, использовалась ли поправка на многократность проверки гипотез

В большинстве статей приводится несколько величин *p*, что повышает вероятность возникновения ошибки I рода (альфа-ошибки), т. е. ошибочного за-

Рис. 6. **Схема рандомизированного клинического испытания, представленная в виде последовательной диаграммы.**

Изображены вмешательства и исходы в обеих группах и количественное соотношение участников на каждом из этапов.

ключения об эффективности вмешательства, когда в действительности полученный результат случаен [22]. Предположим, что исследование включает шесть групп. Сравнение групп между собою требует проведения 15 парных статистических тестов, результатом которых будет определение 15 величин *p*. В отсутствие поправки вероятность возникновения ошибки I рода возрастает с 5 на 100 (обычный уровень вероятности альфа-ошибки составляет 0,05) до 55 на 100 (т. е. 0,55).

К проблеме множественных сравнений есть несколько подходов [12]:

— *проверка идентичности групп* путем определения различий в частоте исходных показателей (в надежде, что таких различий выявлено не будет);

— проведение *множества парных сравнений*, когда данные, полученные в трех и более группах, сравниваются отдельно между собой;

— многократное последовательное определение частоты исходов, на которые влияют одни и те же факторы;

— проведение вторичного анализа для оценки значимости связей между признаками в ходе наблюдения, не предусмотренного в плане исследования;

— проведение анализа данных в подгруппах, не включенных в первоначальную структуру исследования;

— проведение промежуточного анализа полученных данных (частота исхода, определявшаяся в различные сроки);

— последовательное сравнение характеристик групп в различные моменты времени при помощи серии сравнений отдельных групп.

Проведение серии сравнений во многих случаях можно признать целесообразным, но подобный поисковый анализ должен быть соответствующим образом обоснован и описан. Однако «перетряхивание» данных путем вычисления множества величин *p* с целью обнаружить какие-нибудь статистически значимые различия служит признаком низкого методологического качества исследования.

Ошибка 11. Ненужное сравнение исходных характеристик в рандомизированных клинических испытаниях

В рандомизированных клинических испытаниях (РКИ) каждый участник имеет равную вероятность оказаться как в группе вмешательства, так и в контрольной. Поэтому любое различие в исходных характеристиках групп случайно. Следовательно, наличие статистически значимых различий в исходных показателях (табл. 1) не свидетельствует о систематической ошибке (как в исследованиях другой структуры) [9]. Сравнение частоты исходных показателей может выявить некоторые различия между группами, которые, возможно, будет целесообразно учесть при дальнейшем анализе, однако величину *p* указывать при этом не обязательно [9].

Приняв во внимание, что вероятность альфа-ошибки составляет 0,05, в 5 сравнениях из 100 различие в исходных характеристиках окажется статистически значимым просто в силу случая. Однако в одном из исследований показано, что из 1076 сравнений исходных характеристик, проведенных в 125 РКИ, лишь в 2% были найдены различия, оказавшиеся статистически значимыми при $p<0,05$ [23].

Ошибка 12. Не указаны критерии нормы и отклонения от нормы при оценке эффективности диагностических методов

Значимость положительного или отрицательного результата при использовании любого диагностического метода зависит от того, какие критерии были выбраны для определения нормы и отклонения от нормы. В медицине существует шесть определений того, что представляет собой норма [24].

Диагностическая норма: диапазон значений, в пределах которого показатель свидетельствует об отсутствии заболевания, вне пределов которого — о вероятном его наличии. Такое определение представляется целесообразным, поскольку имеет клинический смысл.

Терапевтическая норма: диапазон значений, в пределах которого показатель свидетельствует об отсутствии показаний к назначению лечения, вне пределов которого — о целесообразности терапии. И это определение представляется оправданным.

Другие определения с практической точки зрения менее информативны, однако, к сожалению, нередко используются авторами:

Эпидемиологическое определение нормы: диапазон значений, в пределах которого показатель свидетельствует об отсутствии риска развития заболевания, вне пределов которого — о повышении риска. Данное определение подразумевает, что воздействие на фактор риска влияет на вероятность возникновения исхода. Например, в большинстве случаев высокий уровень холестерина в сыворотке крови сам по себе не представляет интереса; однако тот факт, что при этом повышается риск развития заболеваний сердца, заставляет считать высокий уровень холестерина отклонением от нормы.

Статистическое определение нормы: нормальным считается показатель, определенный у здоровых лиц. Данное определение подразумевает, что полученные результаты подчиняются нормальному распределе-

Таблица 1. **Статистические сравнения исходного состояния групп в РКИ. Различие в концентрации альбумина в крови случайно ($p=0,03$); оно не указывает на систематическую ошибку. В данном случае использование величины *p* необязательно**

Признак	Группа контроля ($n=43$)	Группа вмешательства ($n=51$)	Различие	*p*
Средний возраст, годы	85	84	1	0,88
Мужчины (*n*, %)	21 (49)	21 (51)	3%	0,99
Медиана концентрации альбумина в крови (г/л)	30,0	33,0	3,0 г/л	0,03
Сахарный диабет (*n*, %)	11 (26)	8 (20)	6%	0,83

нию, т. е. при графическом изображении кривая имеет вид колокола. При этом диапазон нормальных значений лежит в пределах ±2 СКО от средней величины, т. е. включает в себя 95% всех измерений. Однако оставшиеся 2,5% с каждой стороны диапазона (отклонение от нормы) не имеют клинического смысла, поскольку встречаются слишком редко. Следует учитывать, что многие результаты не подчиняются нормальному распределению.

Перцентильное определение нормы: нормальным считается показатель, лежащий в пределах диапазона. Например, любой показатель в пределах нижних 95% всех значений определяется как норма, а в пределах оставшихся верхних 5% — как отклонение от нормы. И в данном случае критерием служит частота показателя вне зависимости от клинической значимости.

Социальное определение нормы: нормальным следует называть показатель, который принято считать таковым. Например, желаемая масса тела или возраст, к которому ребенок должен научиться самостоятельно ходить. Подобные критерии не всегда клинически значимы.

Ошибка 13. Отсутствует объяснение, каким образом неопределенные (сомнительные) результаты учтены при вычислении операционных характеристик диагностического теста (таких, как чувствительность и специфичность)

Далеко не всегда использование диагностического метода позволяет получить однозначно положительный или отрицательный результат. Возможно, контрастное вещество было введено не полностью, результаты бронхоскопического исследования не позволяют ни подтвердить, ни опровергнуть наличие заболевания, врач может не согласиться с интерпретацией клинических признаков. Результаты, которые нельзя признать ни положительными, ни отрицательными, влияют на практическую значимость метода, поэтому их наличие и относительная частота должны быть приведены в статье.

Существует три варианта таких неопределенных результатов [25]:

Промежуточные результаты занимают промежуточное положение между отрицательными и положительными. Например, при микроскопическом исследовании ткани положительным результатом служит выявление клеток, окрашенных в синий цвет. Появление клеток, имеющих голубоватую окраску, не достигающую по интенсивности требуемого оттенка, в данном случае следует считать промежуточным результатом.

Неопределенные результаты такие, которые не позволяют сделать ни положительного, ни отрицательного заключения. Например, ответы, полученные при психологическом тестировании, из которых неясно, страдает ли обследуемый алкогольной зависимостью.

Не поддающиеся интерпретации результаты получены при использовании метода с несоблюдением существующих стандартов проведения исследования. Например, определение уровня глюкозы в крови после приема пищи.

В тексте статьи должно иметься объяснение того, каким образом подобные результаты были учтены при определении чувствительности и специфичности метода. Операционные характеристики будут зависеть от того, как неопределенные результаты учитывались: как положительные, отрицательные, либо не включались в анализ. В стандартной таблице сопряженности 2×2, использующейся для расчета чувствительности и специфичности диагностического метода, столбцы и строки для сомнительных результатов отсутствуют (табл. 2). Даже при условии высокой чувствительности или специфичности, но при наличии значительного процента сомнительных результатов, практическая значимость метода будет невелика.

Ошибка 14. Рисунки и таблицы используются лишь для «хранения» данных, а не с целью облегчить восприятие материала

При отображении, анализе и интерпретации данных огромное значение имеют таблицы и рисунки. Однако в научных статьях, помимо собственно «хранения» информации, они должны служить для облегчения восприятия материала [26]. Вследствие этого таблицы и рисунки в статьях могут отличаться от тех, которые были созданы автором исключительно для регистрации данных и проведения анализа. Так, таблица, включающая в себя 3 переменных, может выглядеть совершенно по-разному (табл. 3). Легче всего сравнивать данные, расположенные рядом друг с другом, поэтому оптимальным следует считать именно такую структуру таблицы, подсказывающую читателю то или иное сравнение.

В таблице и диаграммах на рис. 7 представлены одни и те же данные о распространенности заболевания в 9 регионах. Однако таблица позволяет наиболее информативно отобразить точные показатели распространенности, точечная диаграмма — сравнить показатели в различных регионах, а гистограмма — отразить пространственные взаимоотношения между регионами и распространенностью.

Ошибка 15. Несоответствие между внешним видом графика или диаграммы и данными, на которых они основаны

Информация, представленная в графическом виде, воспринимается легче, чем представленная в виде текста [27]. Поэтому очень важно, чтобы внешний вид графиков не искажал смысл данных, на которых они основаны. Одна из проблем возникает

Таблица 2. **Стандартная таблица для определения операционных характеристик диагностического теста***

Результат	Заболевание		Всего
	имеется	отсутствует	
Положительный	a	b	a+b
Отрицательный	c	d	c+d
Всего	a+c	b+d	a+b+c+d

Примечание. *** — чувствительность = a/(a+c), специфичность = d/(b+d). Данная таблица позволяет рассчитать также отношения правдоподобия. Таблица не включает в себя сомнительные результаты, которые нередко и не всегда оправданно игнорируют.**

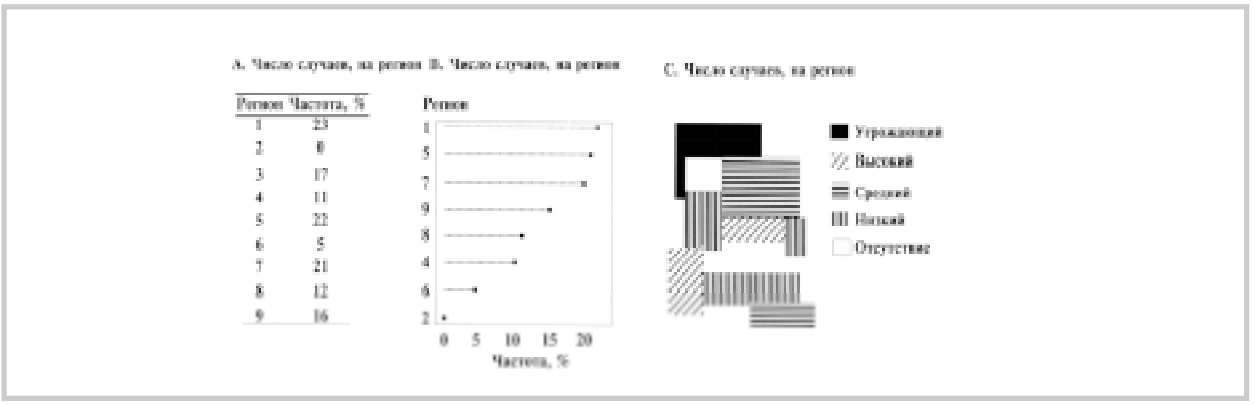

Рис. 7. **Таблицы и рисунки, помимо собственно «хранения» информации, должны служить для облегчения восприятия материала.**

A. Таблицы позволяют наиболее информативно отобразить точные количественные данные. B. Точечные (ленточные) диаграммы позволяют наиболее информативно отобразить общие закономерности и провести сравнение. C. Карты наиболее информативно отражают пространственные взаимоотношения.

при необходимости построения графиков, начальной точкой которых служат ненулевые значения. На гистограмме, представленной на рисунке 8A, визуально высота столбца 1 составляет менее половины высоты столбца 2. Однако такая картина вводит в заблуждение читателя, поскольку нулевые значения отсутствуют. При условии правильного построения гистограммы (рис. 8B) становится очевидным, что высота столбца 1 составляет около $^2/_3$ от высоты столбца 2. Чтобы избежать подобной ошибки, ось Y должна быть прерывистой для указания на отсутствие нулевых значений (рис. 8C).

Другая проблема заключается в «эластичности» графиков. Одна из осей может быть непропорционально сжата или растянута, что приводит к ошибочному восприятию данных (рис. 9). Аналогичные трудности возникают при использовании двойных осей. Если шкала справа не связана математическим отношением с левой, изменение масштаба на одной из осей может приводить к изменению впечатления о связи признаков (рис. 10).

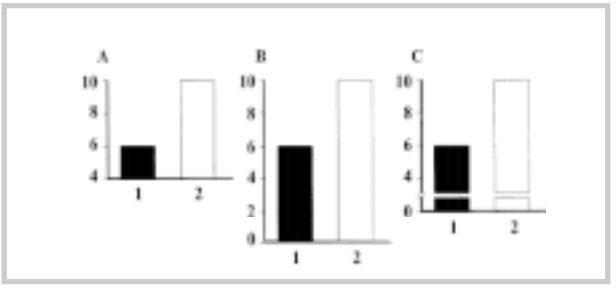

Рис. 8.

A. Гистограммы и графики с отсутствующими нулевыми значениями могут способствовать неверному восприятию материала.

B. Гистограмма построена правильно: высота обоих столбцов соответствует действительности.

C. При отсутствии возможности построения гистограммы, включающей в себя нулевые значения, ось должна быть разорвана.

Ошибка 16. Нечеткое определение понятия «объект исследования»

Термином «объект исследования» обозначают изучаемый предмет, событие или явление. Трудности возникают, если таким предметом служит не сам больной, а нечто иное. Например, если в ходе исследования изучены исходы в отношении 50 глаз, то сколько больных в нем участвуют? И что означает 50% эффективность лечения?

Если объектом изучения служит инфаркт миокарда, то выборка исследования, включающего 18 исходов у 1000 участников, составит 18, а не 1000. Тот факт, что инфаркт возник у 18 участников из 1000, может иметь значение, но на размер выборки это не повлияет, она по-прежнему будет составлять 18.

Если исходом диагностического исследования является заключение специалиста, может быть необходимым исследование выборки специалистов, а не просто выборки результатов исследования. В этом случае размером выборки является число специалистов, а не число полученных оценок.

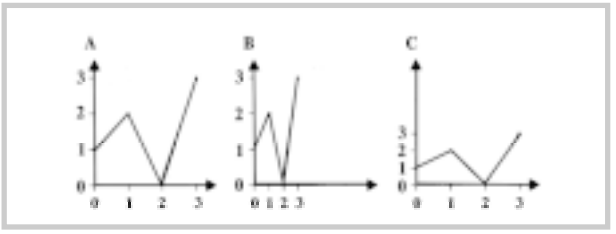

Рис. 9. **Неверно выбранный при построении графика масштаб может способствовать нарушению восприятия материала.**

Уменьшение масштаба по оси X (в данном случае отражающей время, B) приводит к тому, что изменения признака Y выглядят внезапными. Уменьшение масштаба по ординате приводит к впечатлению о постепенных изменениях Y. Предпочтительно использовать графики с одинаковым масштабом по каждой из осей.

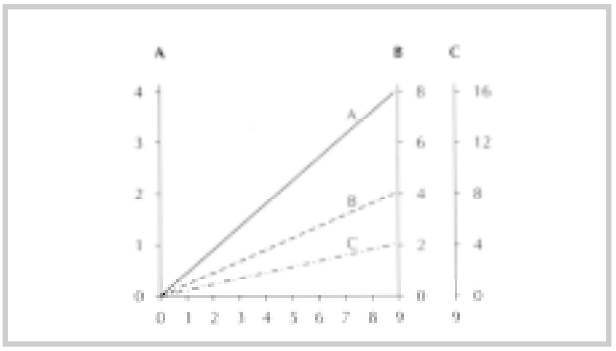

Рис. 10. **При использовании графиков с несколькими осями, каждая из которых служит для отображения своего показателя, связь между последними может искажаться. Линии A, B и C отображают одни и те же данные, но их восприятие зависит от масштаба, выбранного при построении графика. В данном примере подъем линии B представляется в 2 раза меньшим, чем линии A, а подъем линии C — в 4 раза меньшим.**

Ошибка 17. Интерпретация статистически незначимых или полученных в исследованиях с малой статистической мощностью результатов как отрицательных, а не недостаточных

Статистическая мощность представляет собой вероятность выявления статистически значимого различия при условии, что оно действительно существует. Статистически незначимые результаты, полученные в исследовании с малой статистической мощностью, неверно считать отрицательными; они недостаточны: «отсутствие гарантии не есть гарантия отсутствия». К сожалению, многие исследования, в которых получены статистически незначимые результаты, характеризуются малой статистической мощностью. Практическая ценность таких работ невелика, поскольку они не дают ответа на поставленный вопрос [28].

В некоторых ситуациях авторам желательно, чтобы результат оказался статистически незначимым. Например, при сравнении групп автор может стремиться доказать отсутствие различий в исходных характеристиках. Нередко подобные сравнения обладают недостаточной мощностью, поэтому результат не доказывает, что различие действительно отсутствует.

Ошибка 18. Непонимание различий между объяснительными (идеальная эффективность вмешательства) и прикладными (реальная эффективность вмешательства) исследованиями при планировании и интерпретации исследований

Задачей фундаментальных исследований служит объяснение патогенеза того или иного заболевания либо механизма действия лечебного вмешательства. Они проводятся в «идеальных» или «лабораторных» условиях, позволяющих осуществлять тщательный контроль за отбором участников, процессом лечения и наблюдения. Результаты таких исследований позволяют глубже понять биологические механизмы, однако не всегда применимы в клинической практике, не поддающейся столь тщательному контролю. Например, в ходе двойного слепого исследования можно оценить научную обоснованность применения диагностического метода. Но в реальной жизни врачи не ослеплены относительно информации о своих больных, поэтому результаты исследования могут быть нереалистичными.

Задачей прагматических исследований (оценивающих реальную эффективность вмешательства) является помощь в принятии решений в клинике. Они проводятся в обычных условиях, в которых осуществляется работа врачей. На конечный результат подобных исследований может влиять множество факторов, не поддающихся контролю, поэтому научная значимость полученных данных ограничена, однако практическая ценность велика. В отличие от участников фундаментального исследования, выбор которых ограничен жесткими критериями, больные, включаемые в прикладное исследование, как правило, более разнородны по своим характеристикам.

Во многих случаях авторы пытаются объединить оба подхода, но, в конечном счете, ни один из них не реализуется в полном объеме [29, 30]. Результаты исследования следует интерпретировать исходя из природы вопроса, для ответа на который оно предназначено (табл. 4).

Ошибка 19. Представление результатов не в клинически важных единицах

Во всех приведенных ниже примерах полученные результаты представлены клинически четко и грамотно, однако каждая из формулировок позволяет составить различное мнение об эффективности вмешательства [31, 32]. С клинической точки зрения наиболее информативно представление данных в виде связи между прилагаемыми усилиями и получаемым результатом, например, в виде числа пациентов, нуждающихся в лечении для получения одного положительного результата. Помимо прочего, такой способ представления данных позволяет сравнивать различные вмешательства с использованием единых критериев.

Результаты, представленные в абсолютных показателях. В Хельсинкском исследовании (мужчины с гиперхолестеринемией, продолжительность наблюдения 5 лет) инфаркт миокарда был отмечен у 84 (4,1%) из 2030 участников в группе плацебо по сравнению с 56 (2,7%) из 2051 — в группе получавших гемфиброзил (p<0,02); снижение абсолютного риска составило 1,4% (4,1%−2,7%=1,4%).

Результаты, представленные в относительных показателях. В Хельсинкском исследовании (мужчины с гиперхолестеринемией, продолжительность наблюдения 5 лет) частота возникновения инфаркта миокарда в группах плацебо и гемфиброзила составила 4,1 и 2,7% соотв. Абсолютное уменьшение риска на 1,4% соответствует снижению относительного риска развития инфаркта миокарда в группе вмешательства на 34% (1,4%/4,1%=34%).

Результаты, представленные в виде связи между усилиями и результатом. В Хельсинкском исследовании, включавшем 4081 мужчину с гиперхолестеринемией, показано, что для предотвращения 1 слу-

Таблица 3. **Варианты таблицы, включающей в себя 3 переменных (национальность, пол, возраст)**

Вариант 1

Возраст, годы	Мужчины		Женщины	
	США	Китай	США	Китай
0—21				
22—49				
50+				

Вариант 2

Возраст, годы	Китай		США	
	м	ж	м	ж
0—21				
22—49				
50+				

Вариант 3

	0—21		22—49		50+	
	м	ж	м	ж	м	ж
США						
Китай						

Вариант 4

	Мужчины (возраст, годы)			Женщины (возраст, годы)		
	0—21	22—49	50+	0—21	22—49	50+
США						
Китай						

Вариант 5

	0—21		22—49		50+	
	США	Китай	США	Китай	США	Китай
Мужчины						
Женщины						

Вариант 6

	США (возраст, годы)			Китай (возраст, годы)		
	0—21	22—49	50+	0—21	22—49	50+
Мужчины						
Женщины						

Вариант 7

	0—21 год	22—49 лет	50+
Мужчины:			
США			
Китай			
Женщины:			
США			
Китай			

Вариант 8

	0—21 год	22—49 лет	50+
США:			
мужчины			
женщины			
Китай:			
мужчины			
женщины			

чая инфаркта миокарда необходимо проводить лечение 71 участнику в течение 5 лет.

Результаты, представленные в виде связи между усилиями и результатом (другой вариант). В Хельсинкском исследовании, включавшем 4081 мужчину с гиперхолестеринемией, показано, что для предотвращения 1 случая инфаркта миокарда в течение 5 лет необходимо назначить около 200 000 доз гемфиброзила.

Результаты, представленные в виде отношений общей смертности. В Хельсинкском исследовании от инфаркта миокарда в группах гемфиброзила и контроля умерли 6 и 10 участников соответственно. Снижение абсолютного риска составило 0,2%, снижение относительного риска — 40%. Для предотвращения 1 случая смерти от инфаркта миокарда в течение 1 года необходимо назначить лечение 2460 мужчинам.

Ошибка 20. Смешение понятий статистической и клинической значимости

Даже несущественное различие, выявленное при сравнении больших групп, может оказаться статистически значимым, но не иметь при этом клинического значения [12, 33]. Так, при сравнении эффективности использования двух искусственных водителей ритма у нескольких тысяч больных среднее различие в 0,25 месяца на протяжении 5 лет клинически незначимо, даже если оно объясняется случайностью менее чем в 1 из 1000 случаев ($p<0,001$).

И наоборот, даже существенное различие, выявленное при сравнении небольших групп, может иметь клиническое значение, но не быть при этом статистически значимым. Если в ходе исследования, включающего несколько больных в терминальном состоянии, хотя бы один из участников в какой-либо из групп выживет, такой результат будет клинически значимым, хотя статистически значимое различие в частоте выживания между группами может отсутствовать.

ЗАКЛЮЧЕНИЕ

Главное решение проблемы ошибок статистического анализа данных состоит в изучении исследователями методологии исследований и статистического анализа. Статистикам следует проявлять большую активность в вопросе обучения авторов, редакторов и читателей. Необходимо, чтобы авторы привлекали статистиков на этапе планирования исследования, а не после его завершения. Редакторы должны систематически применять рекомендации по представлению статистических данных [12, 18, 19, 34—40]. Важно, чтобы в как

Таблица 4. **Различия между фундаментальным и прикладным исследованиями эффективности таблетированного препарата цинка при лечении простуды. Целью прикладного исследования было определить, способствует ли вмешательство уменьшению числа и длительности сохранения симптомов простуды у амбулаторных больных. В состав участников включали любых лиц, принимавших препарат. В ходе фундаментального исследования изучали эффективность цинка как противовирусного препарата; оно проводилось в более строгих экспериментальных условиях**

Характеристика исследования	Фундаментальное	Прикладное
Критерии диагностики	Получение культуры *Rhinovirus*	Наличие 3 из 10 симптомов
Критерии оценки эффективности (исходы)	Количество отделяемого из носа (подсчет салфеток)	Уменьшение числа и длительности сохранения симптомов
Условия проведения	Стационар	Амбулаторно
Прием препарата	Под контролем исследователя	Под контролем самих больных
Структура	Маскированное, плацебо-контролируемое	Маскированное, плацебо-контролируемое
Задача	Действенность цинка как противовирусного препарата	Эффективность использования цинка при лечении простуды

можно большем числе журналов тщательно проверяли статьи, в которых содержится статистический анализ. Читатели в свою очередь должны обучаться интерпретации статистических данных и требовать от авторов грамотного их представления.

Литература

1. *Glantz S.A.* Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation 1980;61:1—7.
2. *Evans M.* Presentation of manuscripts for publication in the British Journal of Surgery. Br J Surg 1989;76:1311—4.
3. *Mainland D.* Chance and the blood count. 1934 CMAJ 1993;148:225—7.
4. *Schor S., Karten I.* Statistical evaluation of medical journal manuscripts. JAMA 1966;195:1123—8.
5. *White S.J.* Statistical errors in papers in the British Journal of Psychiatry. Br J Psychiat 1979;135:336—42.
6. *Hemminki E.* Quality of reports of clinical trials submitted by the drug undustry to the Finnish and Swedish control autorities. Eur J Clin Pharmacol 1981;19:157—65.
7. *Gore S.M., Jones G., Thompson S.G.* The Lancet`s statistical review process: areas for improvement by authors. Lancet 1992;340:100—2.
8. *George S.L.* Statistics in medical journals: a survey of current policies and proposal for editors. Med Pediat Oncol 1985;13:109—12.
9. *Altman D.G., Schulz K.F., Moher D., Egger M., Davidoff F., Elbourne D., et al.* for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of parallel-group randomized trials. Ann Intern Med 2001;134:657—62; Lancet 2001;357:1191—4; JAMA 2001;285:1987—91.
10. *Stroup D., Berlin J., Morton S., Olkin I., Williamson G.D., Rennie D., et al.* Meta-analys of observation studies in epidemiology. A proposal for reporting. JAMA 2000;283:2008—12.
11. *Moher D., Cook D.J., Eastwood S., Olkin I., Rennie D., Stroup D.F.,* for the Quorum group. Improving the quality of reports of analises of randomised controlled trials: the Quorum statement. Lancet 1999;354:1896—900.
12. *Lang T., Secic M.* How to report statistics in medicine: annotated guideline for authors, editors, and reviewers. Philadelphia (PA): American Colleje of Physicians;1997.
13. *Ehrebeng A.S.* The problem of numeracy. Am Statistician 1981;286:67—71.
14. *Muray G.D.* The task of a statistical referee. Br J Surg 1988;75:664—7.
15. *Feinstein A.R. X* and iprP: an improved summery for scientifics communication. J Chronic Dis 1987;40:283—8.
16. *Feinstein A.R.* Clinical biostatistics XXXVII. Demeaned errors, confidence games, nonplused mineses, inefficient coefficients, and other statistical disruption of scientific communication. Clin Pharmacol Ther 1976;20:617—31.
17. *Gardner M.J., Altman D.* Confidence interval rather that P values estimation rather that hypothesis testing. BMJ 1986;292:746—50.
18. *Bailar J.C., Mosteller F.* Guidelines for statistical reporting in articles for medical journal. Ann Intern Med 1998;108:266—73.
19. *DerSimonian R., Charette L.J., McPeek B., Mosteller F.* Reporting on methods in clinical trials. N Engl J Med 1982;306:1332—7.
20. *Cooper G.S., Zangvill L.* An analysis of the quality of research reports in the Journal of General Internl Medicine. J Gen Inter Med 1989;4:232—6.
21. *Hampton J.R.* Presentation and analysis of the results of clinical trials in cardiovascular disease. BMJ 1981;282:1371—3.
22. *Pocock S.J., Hughes M.D., Lee R.J.* Statistical problems in the reporting of clinical trials/ A survey of three medical journals. N Engl J Med 1987;317:426—32.
23. *Altman D.G., Dore C.J.* Randomisation and baseline comparisons in clinical trials. Lancet 1990;335:149—53.
24. How to read clinical Journals: II. To learn about a diagnostic test. Can Med Assoc J 1981;124:703—10.
25. *Simel D.L., Feussner J.R., Delong E.R., Matchar D.B.* Intermediate, indeterminate and uninterpretable diagnostic test results. Med Decis Making 1987;7:107—14.
26. *Harris R.L.* Information graphics: a comprehensive illustrated reference. Oxford:Oxford University Press; 1999.
27. *Lang T., Tarelico C.* Improving comprehension: theories and research findings. In: American Medical Writers Association. Selected workshop in biomedical communications, Vol. 2. Bethesda (SD): American Medical Writers Association;1997.
28. *Gotzsche P.C.* Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antuunflammatory drugs in rheumatoid arthritis. Cont Clin Trials 1989;10:31—56.

29. *Schwartz D., Lellouch J.* Explantory and pragmatic attitudes in therapeutic trials. J Chron Dis 1967;20:637—48

30. *Simon J., Wagner E.,Voncroff M.* Cost-effectivenes comparisons using "real world" randomized trial: the case of new antidepressant drugs. J Clin Epidemiol 1995;48:363—73.

31. *Guyatt G.H., Sackett D.L., Cook D.J.* Users` guide to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patient? JAMA 1994;271:59—63.

32. *Brett A.S.* Treating hypercholesterolemia: how should practicing physicians interpret the published data for patients? N Engl J Med 1989;321:676—80.

33. *Ellenbaas R.M., Ellenbaas J.K., Cuddy P.G.* Evaluation the medical literature, part II: statistical analysis. Ann Emerg Med 1983;12:610—20.

34. *Altman D.G., Gore S.M., Gardner M.J., Pocock S.J.* Sttistical guidelines for contributors to medical joutnals. BMJ 1983;286:1489—93.

35. *Chalmers T.C., Smith H., Blackburn B., Silverman B., Schroeder B., Reitman D., et al.* A method for assessing the quality of a randomized control trial. Cont Clin Trials 1981;2:31—49.

36. *Gardner M.J., Machin D., Campbell M.J.* Use of checklist in assessing the statistical content of medical studies. BMJ 1986;292:810—2.

37. *Mosteller F., Gilbert J.P., McPeek B.* Reporting Standart and Research Strategies for Controlled Trials. Cont Clin Trials 1980;1:37—58.

38. *Murrey G.D.* Statistical guideline for the Dritish Journal of Surgery. Br J Surg 1991;78:782—4.

39. *Simon R., Wittes R.E.* Methodologic guidelines for reports of clinical trials/ Cancer Treat Rep 1985;69:1—3.

40. *Zelen M.* Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors/ J Clin Oncol 1983;1:164—9.