British Medical Journal 2001; 322: 226–31

Отсеивание фактов: почему нас не удовлетворяют статистические критерии значимости?

Jonathan A. C. Sterne, George Davey Smith
Адрес для корреспонденции: Jonathan A. C. Sterne, Department of Social Medicine, University of Bristol, Bristol BS8 2PR.
E-mail: jonathan.sterne@bristol.ac.uk
Sifting the evidence — what's wrong with significance tests?
© 2001 BMJ. Printed by permission

Результаты научных исследований в области медицины часто оцениваются с большим скептицизмом, даже работы с явно корректным методологическим подходом и использованием соответствующих методов статистического анализа. Чаще всего это касается, по-видимому, результатов эпидемиологических исследований, которые показывают, что некоторые аспекты повседневной жизни отрицательно воздействуют на людей. Так, в недавно опубликованном популярном издании "Взлеты и падения современной медицины" (The Rise and Fall of Modern Medicine) журналист James Le Fanu, пишущий на медицинские темы, даже предлагает закрыть все научные отделы и кафедры, которые занимаются эпидемиологией, чтобы предотвратить вред, наносимый медицине [1].

В частности, этому способствует выраженная в медицинских публикациях тенденция усиливать положительные результаты: подтверждающие гипотезу факты имеют больше шансов быть опубликованными, чем отрицательные результаты [2–4]. Только в этой связи публикуется множество данных, основанных на случайных совпадениях, так как согласно общепринятым подходам 20 случайно полученных в исследованиях сочетаний позволяют говорить о результате, "значимом при $P = 0{,}05$". Если публикуются только положительные результаты, их могут ошибочно принимать как истинные, а не полученные вследствие случайного стечения обстоятельств на основе применения критериев достоверности, вытекающих из представления о статистической значимости. Поскольку во многих исследованиях применяются обширные опросники, аккумулирующие информацию о сотнях переменных, и оценивающие множество возможных вариантов, несколько ложноположительных результатов фактически гарантированы. Однако широкий спектр и часто противоречивый характер результатов медицинских научных исследований обусловлен не только систематической ошибкой, связанной с публикацией. Более существенной проблемой является весьма частое недопонимание сущности статистической значимости.

В этой статье рассматривается история разработки критериев значимости. Создатели статистического анализа не ставили перед собой цель произвольно разграничить "значимые" и "незначимые" результаты (в соответствии с общепринятым пороговым значением $P = 0{,}05$). Значения P должны быть гораздо меньше 0,05 для того, чтобы являться строгим доказательством, отвергающим нулевую гипотезу. Следовательно, необходимо проводить исследования с большей статистической мощностью. При составлении отчетов о научных исследованиях, которые проводятся в области медицины, необходимо переходить от оценки результатов как значимых или незначимых к интерпретации данных в контексте характера исследования и других доступных данных. Редакторы медицинских журналов имеют широкие возможности для того, чтобы поощрять эти перемены, и в заключительной части статьи мы предлагаем рекомендации по представлению и интерпретации результатов исследований.

Величина *P* и применение критериев статистической значимости: краткая историческая справка

Причиной путаницы, существующей в современном подходе к критериям проверки гипотез, являются противоречия, царившие более 60 лет назад среди основоположников статистического анализа [6–8]. Идею оценки статистической значимости предложил R. A. Fisher. Предположим, мы хотим составить мнение о том, повышает ли новый лекарственный препарат уровень дожития после перенесенного инфаркта миокарда. Исследовав группу пациентов, которые лечились этим препаратом, и сопоставимую группу лиц, принимавших плацебо, мы получили данные о том, что показатель смертности в группе лечившихся новым препаратом в два раза ниже такового в группе пациентов, получавших плацебо. Эти результаты обнадеживают, но возникает вопрос, не могут ли они быть случайными? Мы решаем этот вопрос, определяя величину P: вероятность повышения показателя дожития минимум в два раза в случае, если препарат в действительности не влияет на продолжительность жизни.

Fisher рассматривал показатель P как оценку силы доказательств ошибочности нулевой гипотезы (в нашем примере — гипотезы о том, что препарат не влияет на показатель дожития). Он поддерживал идею, что $P < 0{,}05$ (5%-ная значимость) является стандартным уровнем, позволяющим сделать вывод о том, что есть доказательства ошибочности нулевой гипотезы, но не считал это абсолютным правилом. "Если показатель $P = 0{,}1$–$0{,}9$, очевидно, нет причин сомневаться в истинности нулевой гипотезы. Если $P < 0{,}02$, то, скорее всего, нулевая гипотеза не может объяснить все полученные данные. Не будет

большой ошибкой, если мы согласимся в отношении условной границы в 0,05… .” [9]. Важно отметить, что Fisher настаивал на том, что окончательная интерпретация величины *P* остается за исследователем. Например, если *P* » 0,05, нельзя ни доказать, ни отвергнуть нулевую гипотезу, а следует проводить дополнительные исследования.

Итоговые положения

Величина *P* (уровень статистической значимости) оценивает силу доказательств ошибочности нулевой гипотезы; чем меньше величина *P*, тем сильнее доказательства ошибочности нулевой гипотезы.

Произвольное разделение результатов на “значимые” и “незначимые” в соответствии с величиной *P* не было целью основоположников статистического анализа.

Величина *P* = 0,05 не обязательно обеспечивает сильные доказательства ошибочности нулевой гипотезы, но допустимо считать, что *P* < 0,001 их обеспечивает. В разделе научной статьи, в котором отражены результаты исследования, должны быть приведены точные значения *P*, а не данные их сравнения с произвольно установленными порогами.

Результаты научных исследований в области медицины нельзя описывать как “значимые” или “незначимые”, их следует интерпретировать в контексте характера исследования и других имеющихся данных. При получении результатов с низкими значениями *P* необходимо всегда учитывать возможность систематической ошибки и вмешивающихся факторов.

Для того чтобы прекратить дискредитацию научных медицинских исследований публикациями случайных результатов, необходимо провести исследования с большей статистической мощностью.

	Истина	
Результат исследования	Нулевая гипотеза верна (лечение неэффективно)	Нулевая гипотеза ошибочна (лечение эффективно)
Отклонение нулевой гипотезы	Частота ошибки I типа	Мощность = 1 – частота ошибки II типа
Принятие нулевой гипотезы		Частота ошибки II типа

Учитывая отрицательное отношение к субъективности интерпретации, присущей этому подходу, Neyman и Pearson предложили проводить “проверку гипотезы”, заменив субъективную оценку силы доказательств ошибочности нулевой гипотезы на основании величины *P* объективной оценкой результатов экспериментов, основанной на выборе [10]. Neyman и Pearson считали, что существует два типа ошибок, которые могут быть допущены при интерпретации результатов эксперимента (табл. 1). Fisher в основном интересовался ошибками типа I: вероятность отвергнуть нулевую гипотезу (о том, что лечение неэффективно), хотя она на самом деле справедлива. Neyman и Pearson изучали также ошибки типа II: вероятность принять нулевую гипотезу (вследствие чего новый вид лечения не принимается), хотя в действительности она ошибочная (лечение эффективное). Планируя заранее фиксированную частоту ошибок обоих типов, можно уменьшить количество погрешностей в различных экспериментах. Эти мысли разделяют все, кто занимался статистическими расчетами с целью определения количества испытуемых, необходимого для клинического исследования. Цель таких расчетов — обеспечить достаточно большую выборку для того, чтобы частота ошибок типов I и II была невысокой.

Приведем высказывание Neyman и Pearson: “Ни один из методов оценки, основанных на теории вероятности, не может сам по себе предоставить убедительные доказательства правильности либо ошибочности той или иной гипотезы. Но мы можем рассматривать цель этого метода с другой точки зрения. Не надеясь установить, является ли каждая отдельная гипотеза истинной или ошибочной, мы можем определить правила, обусловливающие наше поведение по отношению к ним, следуя которому мы обеспечиваем определенные процедуры по ходу эксперимента, благодаря чему ошибки будут редкими” [10].

Таким образом, при использовании подхода Neyman-Pearson мы заранее выбираем правила принятия решений при интерпретации результатов нашего эксперимента, и результат нашего анализа представляет собой просто отклонение или подтверждение нулевой гипотезы. В отличие от более субъективного подхода Fisher, который был противником метода Neyman-Pearson [11], здесь не предпринимаются попытки интерпретировать величину *P* для того, чтобы оценить силу доказательств ошибочности нулевой гипотезы в конкретном исследовании.

Для того чтобы использовать метод Neyman-Pearson, необходимо точно сформулировать альтернативную гипотезу. Другими словами, недостаточно заявить, что лечение эффективно, мы

должны показать, насколько оно эффективно, например, что наш лекарственный препарат снижает смертность на 60%. Исследователь имеет право изменить правила оценки, уточнив альтернативную гипотезу и допустимую частоту ошибок типов I и II, но это следует делать до начала эксперимента. К сожалению, ученые с трудом усваивают эти идеи. До проведения своих исследований или анализа материалов, за исключением постановки основного вопроса при рандомизированных исследованиях, они редко ориентируются на определенный показатель эффективности лечения в качестве альтернативной гипотезы. В то же время широко применяется только более простая часть метода Neyman-Pearson — положение о том, что нулевая гипотеза может быть отклонена при $P < 0{,}05$ (частота ошибки типа I — 5%). Так возникло ошибочное мнение о том, что метод Neyman-Pearson подобен методу Fisher.

На практике, отчасти в связи с требованиями контролирующих органов и редакторов медицинских журналов [12], применение статистических методов в медицине стало в основном ограничиваться делением результатов на значимые и незначимые, при этом мало внимания обращается, или вовсе не обращается, на частоту ошибок типа II. Это порождает два общих и, по-видимому, серьезных последствия: возможно, клинически важные различия, наблюдаемые в небольших по объему исследованиях, оцениваются как незначимые и игнорируются, тогда как все значимые результаты считаются проявлением действия лечебного вмешательства.

Эти проблемы, замеченные давно [13] и наблюдаемые с тех пор много раз [14–17], способствовали успешному распространению в научных кругах практики представления данных статистического анализа с включением доверительного интервала в дополнение к величине *P* или вместо нее [18–20]. При использовании доверительного интервала, когда внимание акцентируется на результатах конкретного единичного сопоставления, мы отходим от механистической дихотомии: принять или отклонить. В небольших по объему исследованиях он может напомнить нам о том, что полученные нами результаты согласуются как с нулевой гипотезой, так и с важными положительными или вредными лечебными воздействиями (часто сочетающимися). Если значение $P \gg 0{,}05$, возможен меньший или больший эффект, чем полученный в процессе оценки. Однако доверительный интервал 95% предполагает 5%-ный порог, что приводит к путанице при его интерпретации, если он используется лишь как способ оценки значимости (в зависимости от того, включает ли доверительный интервал нулевое значение) вместо того, чтобы с его помощью оценивался возможный разброс данных в популяции. Мы считаем, что исследователи, работающие в области медицины, не должны рассматривать 5%-ную значимость ($P < 0{,}05$) как имеющую какое-либо особое значение. Для этого можно устанавливать другой стандартный доверительный интервал.

Ошибочная интерпретация
величины *P* и критериев значимости

К сожалению, величину *P* все еще часто понимают неправильно. Наиболее распространенная ошибочная интерпретация — мнение о том, что величина *P* отражает вероятность подтверждения нулевой гипотезы. В соответствии с этой точкой зрения значимый результат означает весьма низкую вероятность подтверждения нулевой гипотезы. Ниже, сделав два реальных предположения, мы покажем ошибочность этой интерпретации.

Таблица 2. Количество случаев, когда мы принимаем или отвергаем нулевую гипотезу при рассмотрении реальных предположений об условиях проведения медицинских исследований

Результат исследования	Нулевая гипотеза верна (лечение не эффективно)	Нулевая гипотеза ошибочна (лечение эффективно)	Всего
Отклонение нулевой гипотезы	855	50	905
Принятие нулевой гипотезы	[illegible]	50	95
Всего	900	100	1000

Во-первых, предположим, что процент нулевых гипотез, которые действительно соответствуют отрицательному результату, составляет 10%, т. е. 90% проверяемых гипотез ошибочные. Это согласуется с данными литературы по вопросам эпидемиологии: к 1985 году было выявлено около 300 факторов риска коронарной болезни сердца, вероятно, только незначительная часть из них действительно повышает риск этого заболевания [21]. Наше второе предположение заключается в том, что из-за часто очень небольшого объема исследований средняя статистическая мощность (1 минус частота ошибки типа II) работ, публикуемых в медицинской литературе, составляет 50%. Это совпадает с данными обзорных публикаций, посвященных объему выборок [22–24].

Теперь предположим, что мы проверяем гипотезы в 1000 исследований и отвергаем нулевую гипотезу при $P < 0{,}05$. Наше первое допущение означает, что в 100 исследованиях нулевая гипотеза в действительности ошибочная. Так как частота ошибки типа II составляет 50% (второе допущение), мы отвергаем нулевую гипотезу в 50 из этих 100 исследований. Для 900 исследований, в которых нулевая гипотеза верна (т. е., лечение не оказывает эффекта), мы используем 5%-ный уровень значимости и таким образом отвергаем нулевую гипотезу в 45 исследованиях (табл. 2, адаптированная из работы Oakes [25]).

Из 95 исследований, в которых были получены значимые результаты (т. е. при $P < 0{,}05$), в 45 (47%) нулевые гипотезы верны, т. е. получены “ложные сигналы тревоги”: мы отвергли нулевую гипотезу, хотя не должны были этого делать. Здесь существует прямая аналогия с исследованиями по скринингу популяции на наличие заболеваний. Если заболевание (ошибочная нулевая гипотеза) встречается редко, то специфичность скрининговых тестов должна быть высокой, чтобы истинные случаи заболеваний, выявленные с помощью теста, не перекрывались большим количеством ложноотрицательных случаев из той части популяции, где это заболевание отсутствует [26]. “Показатель положительной прогностической способности” значимого ($P < 0{,}05$) статистического анализа может в действительности быть низким (в приведенном примере — около 50%). Довольно часто ошибочно считают, что при уровне значимости 0,05 показатель положительной прогностической способности равен 95%.

Положения, отраженные в табл. 2, сходны по духу с байесовским подходом к статистическому анализу, при котором мы начинаем с априорных положений в отношении вероятности различных возможных значений лечебного эффекта и меняем эти положения в соответствии с полученными данными. Аргументы байесовского подхода используются для демонстрации того, что обычный порог $P < 0{,}05$ не обязательно является веским доказательством ошибочности нулевой гипотезы [27, 28]. Разные авторы в разное время предлагали шире использовать байесовские статистические методы, чтобы избежать ошибочной интерпретации $P < 0{,}05$ как показателя того, что истинность нулевой гипотезы маловероятна, либо как панацеи, которая значительно повысит качество медицинских исследований [26, 29–32]. Различия между доминирующим (“классическим” или “частотным”) и байесовским подходами к статистическому анализу отражены во вставке 1.

Вставка 1. Сравнение частотного и байесовского подходов к статистическому анализу

Предположим, мы хотим установить, повышает ли новый лекарственный препарат вероятность дожития года после инфаркта миокарда, используя данные плацебо–контролируемого клинического испытания. Для этого мы оцениваем показатель

относительного риска, получаемый делением частоты случаев смерти пациентов, получающих новый препарат, на частоту случаев смерти в контрольной группе. Если показатель относительного риска составляет 0,5, значит, новый препарат снижает риск смерти на 50%. Если показатель относительного риска составляет 1, значит, препарат неэффективен.

Частотные методы статистики

Подобно Mulder и Scully в “The X-files”, статистики, использующие частотные методы, считают, что “истина находится не здесь”. Данные используются для того, чтобы делать выводы в отношении истинного (но неизвестного) значения показателя относительного риска для популяции.

Доверительный интервал 95% позволяет получать достаточно достоверные величины показателя относительного риска для популяции; в 95% случаев этот ряд показателей будет содержать истинные (но неизвестные) показатели для популяции.

Величина *P* — это вероятность получения показателя относительного риска, по меньшей мере настолько удаленного от единицы, как и обнаруженный в нашем исследовании.

Методы байесовского статистического анализа

Сторонники байесовского статистического анализа предпочитают субъективный подход. Вначале составляется априорное мнение о показателе относительного риска, представляемое как распределение вероятностей (или частот). Полученные данные используются для коррекции этого мнения (мы получаем окончательное распределение вероятностей для показателя относительного риска, основываясь как на полученных данных, так и на априорном распределении).

Интервал надежности 95% означает, что вероятность того, что в него входит показатель относительного риска для популяции составляет 95%.

Окончательное распределение может использоваться для того, чтобы сделать прямой статистический вывод в отношении показателя относительного риска — например, вероятность того, что лекарственный препарат повышает риск смерти.

Если предварительное мнение в отношении показателя относительного риска неопределенное (предполагается равная вероятность получения весьма отличающихся друг от друга значений переменных), то результаты частотного анализа будут сходны с результатами байесовского статистического метода; оба подхода основываются на том, что статистики называют правдоподобием полученных данных:

· **доверительный интервал 95% — это то же самое, что и интервал надежности 95%, за исключением того, что последнему часто неправильно придается значение доверительного интервала;**

· **(одномерная) величина *P* представляет собой то же самое, что и байесовская апостериорная вероятность в отношении того, что лекарственный препарат повышает риск смерти (при исследовании возможного защитного эффекта препарата).**

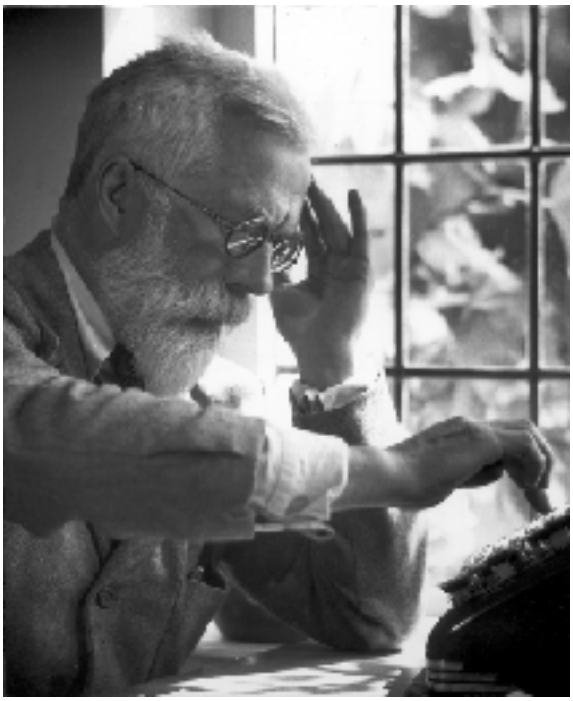
Однако оба метода будут давать разные результаты, если наша априорная оценка не очень приблизительная относительно значимости информации, которая содержится в полученных данных.

Насколько значима статистическая значимость?

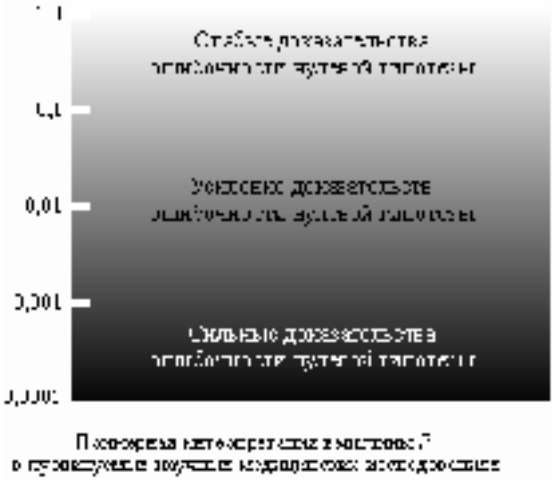
Когда в первые десятилетия XX века разрабатывались принципы статистического анализа, сфера научной деятельности была менее развитой, чем сегодня. В те времена, когда ежегодно проверялось, вероятно, лишь несколько сотен статистических гипотез, а вычисления были трудоемкими и выполнялись с помощью ручных механических калькуляторов (как на фотографии Fisher), казалось, были основания считать, что 5%-й показатель ложноположительных результатов выявит большинство случайных ошибок. Сегодня, когда тысячи журналов публикуют огромное количество проверяемых гипотез и мы располагаем простыми в использовании компьютерными статистическими программами, а доля значимых (в том смысле, что выявленный эффект достаточно велик, чтобы представлять интерес) гипотез явно снизилась, установлено, что $P < 0{,}05$ имеет низкую прогностическую способность, чтобы с уверенностью можно было отклонить нулевую гипотезу.

Часто можно легко повысить статистическую мощность исследований, увеличив либо размер выборки, либо точность измерений. В табл. 3 приводится прогностическая ценность разных пороговых значений величины *P* при различных вариантах как статистической мощности исследований, так и доли значимых гипотез. Для любого значения величины *P* процент значимых результатов, которые являются ложноположительными, значительно снижается при повышении статистической мощности исследования. Из табл. 3 следует, если мы не очень пессимистически оцениваем долю значимых гипотез, что величину $P < 0{,}001$ разумно рассматривать как убедительное доказательство ошибочности нулевой гипотезы.

В качестве одного из аргументов против изменения требований к силе доказательств ошибочности нулевой гипотезы приводится то, что в этом случае объем исследований должен в значительной степени возрасти. К удивлению, это не так. Для иллюстрации можно показать, выполнив стандартные расчеты статистической мощности, что размер выборки следует увеличить максимум лишь в 1,75 раза при переходе от $P < 0{,}05$ к $P < 0{,}01$ и в 2,82 раза при переходе от $P < 0{,}05$ к $P < 0{,}001$. Также возможно (и обычно предпочтительно) увеличить статистическую мощность, изменив подход к измерению, не увеличивая объем выборки [33]. Таким образом, проводя меньше исследований, но с большей статистической мощностью, можно предупредить дискредитацию научных исследований в области медицины. Потребность в обширных статистически корректных исследованиях подчеркивается в течение многих лет Richard Peto и его сотрудниками [34]. Однако качество научных исследований в области медицины не улучшится, если мы просто заменим один произвольно устанавливаемый порог величины *P* (0,05) другим (0,01).



R.A. Fisher, основатель статистического анализа, выполняет вычисления с помощью механического калькулятора



приходят к необоснованным выводам. В медицинской науке чаще всего игнорируются периодические призывы к полному использованию байесовских статистических методов. Главная причина — трудность квантификации априорных положений. Например, какой статистический вес следует придать конкретной совокупности полученных доказательств, если результаты исследования противоречат данным о международных различиях в распространенности заболевания?

Таким же образом прогностическую ценность $P < 0{,}05$ для значимой гипотезы легко подсчитать на основе предполагаемой доли значимых гипотез в исследуемом интервале, однако истинное значение этой доли знать невозможно. К сожалению, табл. 2 и 3 — это всего лишь теоретические построения. Если попытаться избежать проблемы квантификации априорных положений — а это сделает нашу гипотезу крайне неопределенной, — то результат анализа с помощью байесовского метода будет аналогичным результату стандартного подхода. С другой стороны, рационально интерпретировать $P = 0{,}008$ для основного эффекта, выявленного в клинических испытаниях, иначе, чем такую же величину P для одного из основных выводов эмпирического исследования на основании того, что в первом случае доля проверяемых значимых гипотез, по-видимому, больше, а систематическая ошибка и вмешивающиеся факторы менее вероятны.

Что необходимо сделать?

Существует три способа снизить частоту ошибок вследствие применяемой практики использования критериев значимости. Во-первых, в табл. 3 показано, что величина $P < 0{,}05$ не может рассматриваться как обеспечивающая абсолютные или даже убедительные доказательства ошибочности нулевой гипотезы. Во-вторых, ясно, что повышение доли значимых проверяемых гипотез также снизит частоту ошибок. К сожалению, этот способ трудно применить: вывод о том, что формулирование априорных гипотез полностью избавит нас от ошибочных действий, сам ошибочный. Если с целью изучения неэффективного метода лечения провести 100 рандомизированных клинических испытаний, каждое из которых будет проверять только одну гипотезу и применять только один метод статистической оценки гипотез, все значимые результаты окажутся ошибочными. Более того, невозможно с уверенностью заявить, что существующие гипотезы послужили причиной исследования полученных взаимосвязей. Этот процесс высмеян Philip Cole, который писал, что он с помощью компьютерного алгоритма сформулировал все возможные гипотезы в эпидемиологии, так что все статистические методы анализа считаются сейчас априорными гипотезами [39]. В-третьих, необходимо не изменить статистические парадигмы, а повысить качество исследований, увеличив объем выборки и точность измерений.

Хотя нет простого или единственного решения, есть возможность уменьшить риск совершения ошибки благодаря результатам проверки гипотез. Такой возможностью располагают, в частности, редакторы научных журналов. Существенные изменения в представлении данных статистического анализа произошли после появления в 80-х годах методических рекомендаций, требующих приводить в публикациях доверительные интервалы. Сейчас необходимы такие же изменения в отношении представления статистической проверки гипотез. Мы считаем, что редакторы журналов должны требовать от авторов научных публикаций следовать рекомендациям, которые содержатся во вставке 2.

Вставка 2. Предлагаемые рекомендации для представления результатов статистического анализа в медицинских журналах

1. Характеристика различий как статистически значимых недопустима.

2. Во всех случаях необходимо представлять доверительные интервалы для основных результатов, но предпочтительно использовать уровень 90%, а не 95%. Доверительные интервалы не должны использоваться как суррогатные способы оценки значимости на общепринятом уровне 5%. При интерпретации доверительного интервала следует обращать особое внимание на выводы (клиническое значение), которые можно сделать на основании ряда показателей, расположенных в этом интервале.

3. При наличии значимой нулевой гипотезы сила доказательств ее ошибочности должна подтверждаться величиной P. Чем меньше величина P, тем сильнее доказательства.

4. Поскольку невозможно существенно уменьшить объем поиска данных, в клинических испытаниях и эмпирических исследованиях авторы должны весьма скептически относиться к анализу подгрупп. В любом случае следует представлять данные о силе доказательств в пользу обнаруженного взаимодействия, т. е. о том, что эффект в различных подгруппах действительно различается. Выводы, сделанные на основе анализа

подгрупп, должны обосновываться более тщательно, чем выводы в отношении основного эффекта.

5. При проведении эмпирических исследований необходимо помнить о том, что учет вмешивающихся факторов и систематической ошибки, по меньшей мере, так же важен, как и аспекты, обсуждаемые в этой статье [40].

Таблица 3. Соотношение ложноположительных значимых результатов при трех различных уровнях статистической значимости

[illegible]	Доля [illegible] результатов, которые являются ложноположительными		
	[illegible]	[illegible]	[illegible]
[illegible]			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]			
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

Выражаем благодарность профессору S. Goodman, врачам
M. Hills и K. Abrams за полезные замечания к первому варианту рукописи; это не означает, что они поддерживают наши взгляды. Бристоль — ведущий центр по проведению совместных исследований служб органов здравоохранения.

В прошлом оба автора неправильно использовали термин “значимость” и, возможно, переоценили силу доказательств в пользу своей гипотезы.

ЛИТЕРАТУРА

1. Le Fann J. *The rise and fall of’ modern medicine*. New York: Little, Brown, 1999.

2. Berlin JA, Begg GB, Louis TA. An assessment of publication bias using a sample of published clinical trials. *J Am Stat Assoc* 1989; 84:381–92.

3. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991; 337: 867–72.

4. Dickersin K, Min YI, Meinert CI. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 1992; 263: 374–8.

5. Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. *Int J Epidemiol* 1988; 17, 680–5.

6. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* l993; 137: 485–96.

7. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc* 1993; 88: 1242–9.

8. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 1999; 130:995–1004.

9. Fisher RA. *Statistical methods for research workers*. London: Oliver and Boyd, 1950:80.

10. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans Roy Soc A* 1933; 231: 289–337.

11. Fisher RA. *Statistical methods and scientific inference*. London: Collins Macmillan, 1973.

12. Feinstein AR. *P*-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol* 1998; 51: 355–60.

13. Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc* 1942; 37: 325–35.

14. Rozeboom WW. The fallacy of the null-hypothesis significance test. *Psychol Bull* 1960; 57: 416–28.

15. Freiman JA, Chalmers TC, Smith HJ, Kuebler RR. 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.

16. Cox DR. Statistical significance test. *Br J Clin Pharmacol* 1982; 14: 325–31.

17. Rothman KJ. Significance questing. *Ann Intern Med* 1986; 105: 445–7.

18. Altman DG, Gore SM, Gardner MJ, Pocock, SJ. Statistical guidelines for contributors to medical journal. *BMJ* 1983:286: 1489–93.

19. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *BMJ* 1986; 292: 746–50.

20. Gardner MJ, Altman DG. *Statistics with confidence. Confidence intervals and statistical guidelines.* London: BMJ Publishing, 1989.

21. Hopkins PN, Williams RR. Identification and relative weight of cardiovascular risk factors. *Cardiol Clin* 1986; 4: 3–31.

22. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. In: Bailar JC, Mosteller F. ads. *Medical uses of statistics.* Boston, Ma: NEJM Books, 1992: 357–73.

23. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994:272: 122–4.

24. Mulward S, Gωtzsche PC. Sample size of randomized double-blind trials 1976–1991. *Dan Med Bull* l996; 13: 96–8.

25. Oakes M. *Statistical inference*. Chichester: Wiley, 1986.

26. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987; 257: 2459–63.

27. Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. *Psychol Rev* 1963; 70: 193–242.

28. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *J Am Stat Assoc* l987; 82: 112–22.

29. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996; 313: 603–7. 30. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA* 1995; 273: 871–5.

31. Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to *p* values. *J Epidemiol Community Health* 1998; 52: 318–23.

32. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999; 130: 1005–13.

33. Phillips AN, Davey Smith G. The design of prospective epidemiological studies: more subjects or better measurements? *J Clin Epidemiol* 1993; 46: 1203–11.

34. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984; 3: 409–22.

35. Eggcr M, Davey Smith G. Meta-analysis. Potentials and promise. *BMJ* 1997; 315: 1371–4.

36. Danesh J, Whincup P, Walker M, Lennon L, Thompson A, Appleby P, *et al* Chlamydia pneumoniae IgG titres and coronary heart disease: prospective study and meta-analysis. *BMJ* 2000; 321: 208–13.

37. Morris JN. *The uses of epidemiology*. Edinburgh: Churchill-Livingstone, 1975.

38. Davey Smith G. Reflections on the limits to epidemiology. *J Clin Epidemiol* (in press).

39. Cole P. The hypothesis generating machine. *Epidemiology* 1993; 4: 271–3.

40. Davey Smith G, Pliillips AN. Confounding in epidemiological studies: why "independent" effects may not be all they seem. *BMJ* 1992; 305: 757–9.