КАК ОРГАНИЗОВАТЬ СОБСТВЕННЫЕ ИССЛЕДОВАНИЯ И НЕ ЗАПУТАТЬСЯ В ИНТЕРПРЕТАЦИИ ЧУЖИХ: ИССЛЕДОВАНИЯ III ФАЗЫ (часть 3)

Н.В. Жуков
ФГУ ФНКЦ ДГОИ
Росздрава

Размер выборки, период наблюдения за больными и статистическая значимость

К сожалению, в большинстве отечественных исследований время финального (или промежуточного) анализа данных привязывается к потребностям автора. Чаще всего анализ проводится «по мере возникновения необходимости»: апробация или защита диссертации, крайний срок подачи тезисов на конференцию или проведение очередного отчета по НИР (научно-исследовательским работам). Таким образом, количество больных, включенных в исследование, и время наблюдения за ними после окончания лечения (follow-up) либо не планируются вообще, либо привязываются к «пропускной способности» клиники. К сожалению, такой подход делает результаты многих проспективных исследований нетрактуемыми. Для получения адекватных для анализа данных протоколы исследований III фазы должны заранее определять необходимое для набора количество пациентов и длительность периода наблюдения после закрытия набора больных. Только статистически обоснованное определение этих величин позволяет избежать высокого риска получения ложноположительных и ложноотрицательных результатов. Именно на основании этих параметров (количество включенных больных и длительность периода наблюдения) и определяется время, необходимое для проведения финального или промежуточного анализа (а не наоборот).

Методы планирования размера выборки базируются на предположении, что к окончанию заранее определенного периода наблюдения будет возможно проведение адекватного анализа на статистическую значимость, способного подтвердить или опровергнуть наличие предполагаемых различий в эффективности экспериментального и контрольного лечения. Что же такое анализ на статистическую значимость?

Уровень статистической значимости (в отечественной литературе называемый также статистической достоверностью) в 0,05 имеет следующее значение: в случае, если между исследуемыми методами лечения нет истинного различия в эффективности, то вероятность увидеть различие в результатах такой величины, как наблюдается в данном исследовании, составляет 0,05. Проще говоря, если в исследовании выявлена разница в 5-летней выживаемости, равная 20% при p=0,05, то шанс, что наблюдаемое 20% различие в выживаемости получено случайно (в результате стечения обстоятельств), а не является истинным показателем различия между методами лечения, составляет 5%. Важно помнить, что уровень статистической значимости в 0,05 не обозначает, что вероятность равенства эффективности исследуемого и контрольного метода лечения (вероятность верности нулевой гипотезы) равняется 0,05 (5%). В противном случае, включив в группы по 5 человек и получив закономерное p=0,85, можно было бы говорить, что с 85% вероятностью исследуемые виды лечения одинаковы. Разумеется, такое заключение бессмысленно, т.к. p=0,85 говорит нам не о вероятности равенства между группами, а лишь о вероятности ошибочного заключения.

Шанс выявления статистически значимых различий зависит от размера выборки и величины истинного различия сравниваемых показателей между контрольной и основной группами. Если в исследование включено небольшое количество пациентов, различие в наблюдаемых результатах должно быть крайне велико для того, чтобы получить статистическую значимость с высокой вероятностью. С другой стороны, вероятность получить статистически значимые результаты может быть очень мала, даже если наблюдается истинное, но незначительное по величине различие в эффективности лечения. Например, если вы решите сравнить выживаемость больных герминогенными опухолями на фоне шести курсов химиотерапии по схеме BEP и приема биологически активной добавки, то шанс не получить статистической значимости различий практически равен нулю.

Каков бы ни был дисбаланс прогностических факторов между сравниваемыми группами, результаты лечения будут различаться уже после включения небольшого количества больных, т.е. наблюдаемые в исследовании результаты лечения будут отражать наличие истинного различия. Трудно представить себе ситуацию, в которой различие не будет выявлено. Другое дело, если та же биологически активная добавка будет сравниваться с адьювантной терапией ингибиторами ароматазы при раке молочной железы. Разумеется, при благоприятном стечении обстоятельств существует шанс, что с самого начала исследования группы больных окажутся идеально сбалансированы по всем возможным факторам прогноза, и единственным влияющим фактором будет вид лечения. В такой ситуации кривые выживаемости «разойдутся» на ранних этапах исследования и «отразят» истинное различие. Однако может оказаться, что распределение не оговоренных заранее (или неизвестных) прогностических факторов между сравниваемыми группами будет неравномерным и нивелируется лишь в очень большом размере выборки, в результате чего небольшое, но истинное различие в выживаемости будет скрыто при малом размере выборки (наблюдаемое в исследовании различие будет меньше истинного), и при проведении теста на статистическую значимость мы не получим заветного $p<0,05$. Означает ли это, что сравниваемые методики одинаковы по эффективности? Нет. Просто при недостаточном количестве включенных в исследование больных мы не смогли выявить имеющееся истинное различие.

Вероятность получить статистически значимый результат в случае, если исследуемые методы лечения действительно отличаются по эффективности, называется статистической силой исследования. В основном при определении числа планируемых для включения в исследование больных ориентируются на статистическую силу исследования в 0,80 или 0,90 (шанс выявления истинного различия 80 или 90% соответственно). Статистическая сила исследования возрастает при увеличении размера выборки и удлинении интервала наблюдения. Однако в наибольшей степени статистическая сила исследования зависит от величины истинного различия эффективности между двумя методами лечения.

В настоящее время разработаны весьма точные методы планирования размеров выборки исследований, направленных на выявление различий в выживаемости в исследованиях III фазы. Для планирования большинства таких исследований принципиальным является даже не общее число больных, а количество неблагоприятных событий для данного вида выживаемости (смерти для общей выживаемости, рецидивы — для безрецидивной и т.д.). В табл. 4 представлены данные о числе неблагоприятных событий, которые должны произойти в обеих сравниваемых группах (суммарно) для того, чтобы исследование имело необходимую статистическую силу (90%) при выявлении определенного снижения риска развития неблагоприятного события. При экспотенциальном распределении процентное сокращение риска смерти может быть выражено и как соотношение медиан выживаемости, которые отражены во втором столбце табл. 4.

Таким образом, например, предполагается, что медиана общей выживаемости больных контрольной группы составит 10 мес, а экспериментальное лечение увеличит ее до 15 мес (соотношение медиан выживаемости 1,5). Снижение риска смерти в таком случае составит: (медиана выживаемости экспериментальной группы – медиана выживаемости контрольной группы)/медиана выживаемости экспериментальной группы = 33%. Исходя из данных в табл. 4 рекомендаций, для того чтобы с вероятностью в 90% не пропустить истинное различие между группами, необходимо проводить окончательный анализ после того, как из всех больных, включенных в исследование, умрут 257. Однако при планировании общего числа больных, включенных в исследование, необходимо учитывать не только эти данные, но и ожидаемое время до события (когда наблюдается наибольшее число событий). Например, при метастатическом раке подже-

Таблица **4**. *Требуемое число событий для выявления различий в выживаемости со статистической силой исследования 90%*

Планируемое для выявления снижение риска развития неблагоприятного события между контрольной и основной группами, %	Соотношение медиан выживаемости при экспотенциальном распределении	Требуемое количество неблагоприятных событий в обеих группах (суммарно)
20	1,25	846
30	1,43	330
33	1,50	257
40	1,67	162
50	2,0	88

лудочной железы, как это ни печально, планируемое количество больных будет мало отличаться от необходимого для финального анализа на статистическую значимость количества смертей. В случае же адъювантной химиотерапии при раке молочной железы I—II стадии количество включенных больных должно значительно превышать требуемое количество событий, так как в противном случае даже при потрясающей эффективности нового препарата исследователь рискует так и не получить при жизни точного ответа на вопрос, помогает ли он или нет. Таким образом, при планировании исследований III фазы окончательный анализ обычно проводится после того, как произойдет требуемое для необходимой статистической силы исследования количество событий, а не «привязывается» к конкретной календарной дате (например, время апробации диссертации). Необходимо лишь еще раз уточнить, для чего требуется большое количество событий (больных). Отнюдь не для того, чтобы исследование было «более достоверным». Принцип «чем больше, тем лучше» и слепая гонка за количеством не приносят ничего, кроме лишних расходов и траты времени (за которое новая методика уже может устареть или быть «заявлена» другими исследователями). Больных нужно включать «столько, сколько нужно». Планируемое заранее количество включенных больных (требуемое количество событий) позволяет быть с определенной вероятностью (при 90% статистической силе — с вероятностью в 90%) уверенным, что ожидаемое различие в выживаемости будет выявлено (разумеется, если оно действительно есть).

Однако во многих случаях при планировании размера выборки может быть удобнее не отталкиваться от медианы выживаемости (которая, кстати, при адъювантном лечении рака молочной железы может быть и не достигнута до срока, когда больные начнут погибать от старости), а ориентироваться на другие показатели. Например, можно оценивать долю больных, переживших без отрицательных событий определенный промежуток времени (например, 5-летняя выживаемость). В табл. 5 представлены данные о числе пациентов, требуемых для выявления различия в таких показателях.

К сожалению, такой подход гораздо менее гибок, так как подразумевает проведение финального анализа только когда все пациенты, у которых не произошло неблагоприятного события, будут прослежены до срока, отражаемого в названии выживаемости (5 лет для 5-летней выживаемости, 10 для 10-летней и т.д.). Как можно легко понять, взглянув на эту таблицу, в большинстве случаев выполнение адекватного исследования III фазы в одной клинике практически нереально, так как в современной онкологии различия между двумя методами лечения редко превышают 10—15%.

Еще раз хотелось бы остановиться на том, что размер выборки крайне важен для того, чтобы быть уверенным в результатах исследования. Причем касается это в большей степени результатов негативных (т.е. два вида лечения действительно не отличаются друг от друга). Как видно при анализе медицинской литературы, многие публикуемые «негативные» исследования на самом деле являются нетрактуемыми (т.е. их результат неясен) в связи со слишком малым размером выборки.

С другой стороны, особенно важно выявление небольших различий в исследованиях, сравнивающих стандартное лечение с более «консервативным» (органосохраняющие операции, сокращение интенсивности или длительности терапии и т.д.). В данном случае получение $p<0,05$ при сравнении выживаемости будет свидетельствовать о том, что консервативный подход снижает эффективность лечения. В такой ситуации даже небольшое уменьшение эффективности может все же быть медицински значимым, так как выживаемость или частота излечения не должны приноситься в жертву удобству оперирования или косметическим результатам. Такие исследования всегда должны иметь большую статистическую силу, необходимую для выявления даже небольших различий.

Таблица **5**. *Количество больных в каждой из сравниваемых групп, требуемое для выявления различий в выживаемости (двустороннее p=0,05, статистическая сила 90%)*

X-летняя выживаемость в контрольной группе, %	**Планируемая разница в выживаемости (X-летняя выживаемость экспериментальной группы — X-летняя выживаемость контрольной группы), %**				
	5	**10**	**20**	**30**	**50**
5	620	206	74	42	19
10	956	285	92	48	21
20	1502	411	118	57	22
30	1880	495	134	62	22
50	2132	537	134	45	17

Что же мы должны сделать, чтобы получить полноценные результаты исследования III фазы, на которые можно будет затем опираться в клинической практике?

1. Создать медицински правильный дизайн исследования (критерии отбора больных, проспективный характер, рандомизация, стратификация). Неправильное медицинское планирование исследования делает все дальнейшие шаги бессмысленными. Если дизайн исследования спланирован таким образом, что в одну группу могут попасть 90% больных со II стадией и ECOG=0, а во вторую — 90% с IV стадией и ECOG=3, то различия в выживаемости, скорее всего, будут получены. И *p* будет меньше 0,000001. И различия в выживаемости будут истинными. Только обусловлены они будут не лечением.
2. Определить целевой уровень статистической значимости при проведении финального анализа.
3. Определить целевую статистическую силу исследования.
4. Определить ожидаемую выживаемость в группе контроля (по историческим данным) и предполагаемую или желаемую выживаемость в экспериментальной группе.
5. На основании желаемого уровня статистической значимости, статистической силы исследования и планируемого различия в выживаемости определить количество событий, которое должно произойти до финального анализа.
6. Оценить, сколько больных и времени потребуется для того, чтобы такое количество событий произошло (с учетом ранее имевшихся данных о выживаемости больных, получающих подобное лечение).
7. Назначить время финального анализа и начать исследование. Или с сожалением констатировать, что для проведения исследования понадобится 30 лет (или 5000 больных), и отказаться от этой затеи.

Факториальный дизайн

Данный вид дизайна исследований достаточно часто стал использоваться в настоящее время. Основной целью таких исследований является желание «содрать 7 шкур с одной овцы», а именно сравнить за одно исследование 4 лечебные группы. При факториальном дизайне (также называемом 2 × 2), действительно, имеется 4 лечебные группы. Первый фактор представляет два альтернативных вмешательства, например ампутацию и резекцию. Второй — два альтернативных последующих вмешательства, например, проведение или непроведение адъювантной химиотерапии. В другом примере первый фактор может быть представлен лекарствами А или В, а второй — длительностью применения — 6 или 12 мес. Хотя на самом деле мы имеем дело с четырьмя лечебными группами, общий эффект каждого из факторов может быть оценен с использованием данных обо всех пациентах, подвергшихся его воздействию. Если взять первый пример, то мы можем сравнить ампутацию и резекцию, «не замечая» при этом вид последующего лечения (т.е. в одной сравниваемой группе будут больные, подвергшиеся ампутации с или без адъювантной химиотерапии, а во второй — больные, подвергшиеся резекции с или без адъювантной химиотерапии). К сожалению, исследования данного типа имеют одно весьма существенное ограничение, которое часто нельзя предвидеть заранее. Исследуемые методы должны быть «независимы» друг от друга. Например, если адъювантная терапия улучшает результаты и после ампутации, и после резекции, то данное различие будет выявлено при сравнении групп, получавших и не получавших адъювантную терапию. Однако если проведение адъювантной химиотерапии улучшает прогноз только больных, подвергшихся резекции, то это различие может остаться невыявленным из-за «смешения» с группой, подвергшейся ампутации, в которой адъювантная терапия выживаемость не улучшает. В итоге может быть сделано ошибочное заключение о неэффективности адъювантной терапии. К сожалению, обычно объем выработки при планировании исследований с факториальным дизайном подсчитывается исходя из предположения, что взаимосвязи между лечебными методиками нет. В связи с этим к данным исследованиям (особенно при получении негативных результатов) следует относиться с большой осторожностью.

Исследования на терапевтическую эквивалентность (non inferiority trial)

Как ни странно, данный вид исследований очень распространен в нашей стране, хотя многие исследователи, его применяющие, даже не знают этого названия. Наиболее часто он используется в хирургии (при разработке органосохраняющих операций), при испытании дженериков (структурных аналогов уже зарегистрированных препаратов), а также в диссертациях (для доказательства «одинакового прогноза» исследуемых групп). Учитывая, что в таких исследованиях основной задачей чаще всего является «неполучение» различия между группами (отсутствие *p* при сравнении выживаемости), они кажутся наиболее легкими в исполнении. Возьмем простой пример из жизни: на рынок выходит произведенный фирмой «Вселечин & Со» препарат, содержащий то же действующее вещество, что и уже давно зарегистрированный препарат известного производителя. Необходимо показать, что препарат этот не хуже. За исследование берется ис-

следователь Х из известной онкологической клиники, которому фирма благородно выделяет свой препарат и закупает препарат сравнения на лечение аж 40—50 больных. Проводится исследование, в результате которого при сравнении кривых выживаемости p=0,4. Победа! Можно подавать препарат на регистрацию с заключением из известного института о том, что он так же эффективен, как и брэндовый. Все бы хорошо, но «недостоверная» p может быть обусловлена просто малым количеством включенных в исследование больных (недостаточной статистической силой исследования). Если размер выборки «берется с потолка», то лучше всего было бы посоветовать в таких ситуациях включать в каждую группу по одному больному (и деньги, и время будут значительно сэкономлены, а p будет близка к единице).

Настоящей задачей исследований на терапевтическую эквивалентность является демонстрация того, что с определенной статистической силой (т.е. шансом не выявить различий определенной величины) новое лечение если и уступает по эффективности стандартной терапии, то не более чем на очень небольшой процент. Например, показать, что лечение препаратом фирмы «Вселечин & Со» с вероятностью в 90% (статистическая сила) может отличаться (хотя совсем не обязательно отличается) по эффективности от уже зарегистрированного препарата не более чем на 5%. И уж тем более данный вид исследований отличается от исследований на биологическую эквивалентность, в которых основной задачей является продемонстрировать эквивалентность концентрации активных метаболитов в сыворотке крови (при использовании тестируемого препарата и препарата сравнения).

Таким образом, исследования на терапевтическую эквивалентность представляются достаточно проблематичными, так как на самом деле продемонстрировать эквивалентность невозможно. В обычных исследованиях отклонение нулевой гипотезы (доказательство различия между лечебными группами) ведет к изменению терапии у пациентов в дальнейшем. Необходимость доказательства невозможности отклонить нулевую гипотезу является задачей, более тяжелой для интерпретации. К сожалению, как было сказано выше, невозможность отклонить нулевую гипотезу многими исследователями (в том числе и зарубежными) интерпретируется как демонстрация терапевтической эквивалентности и, соответственно, как повод для принятия нового метода лечения. Однако невозможность опровергнуть нулевую гипотезу может являться лишь свидетельством недостаточности размера выборки или неэффективности стандартного лечения для пациентов, участвующих в исследовании.

Для осмысленных полноценных исследований на терапевтическую эквивалентность обычно требуется очень большая выборка, так как необходимо выявить даже очень небольшое, но медицински важное различие. Например, необходима организация исследования, оценивающего резекцию опухоли в качестве альтернативы ампутации, являющейся стандартом терапии и приводящей к излечению большинства больных. По сравнению с ампутацией резекция опухоли «в пределах здоровых тканей» с последующим эндопротезированием может иметь явное преимущество в отношении качества жизни, однако очень немногие пациенты согласятся получить такой «подарок судьбы», если не будут уверены в том, что при проведении резекции шанс на уменьшение выживаемости будет очень незначительным.

К сожалению, до настоящего времени ни один из традиционных подходов к дизайну и анализу исследований на терапевтическую эквивалентность не является идеальным. Большинство исследований строится на возможности определения минимального уровня различий в эффективности (дельта), которое исследователю кажется приемлемым. К сожалению, ни одно из исследований не посвящено вопросу, каким образом должна определяться эта дельта. Другими словами, потенциальное 5% различие в общей выживаемости при раке поджелудочной железы — это достаточно малое различие, чтобы признать методы эквивалентными? А при раке яичка?

Анализ исследований III фазы

Анализ по «намерению лечить» (intention-to-treat analysis)

Одним из наиболее важных принципов анализа исследований III фазы является так называемый принцип анализа по включению в исследование, или намерению лечить (intention-to-treat principle). Этот принцип требует, чтобы все рандомизированные пациенты были включены в финальный анализ результатов исследования вне зависимости от того, какую часть от запланированной программы лечения они получили (и получили ли вообще). Для онкологических исследований это часто интерпретируется как «оцениваемые в отношении эффективности и безопасности» рандомизированные пациенты. Этот принцип, к сожалению, очень часто высмеивается или не соблюдается корифеями отечественной науки. Часто приходится встречаться с ситуацией, когда из анализа выживаемости «выкидываются» больные, погибшие от осложнений в послеоперационном периоде, или пациенты, получившие исследуемый препарат с «нарушением режима». Разумеется, такой подход позволяет автору получить гарантированный «положительный» результат, однако цена его невелика. В связи с тем

что критерии отбора могут быть достаточно неопределенными и неверифицируемыми внешними аудиторами, постфактум исключение «неподходящих под критерии отбора» пациентов само по себе может стать фактором, влияющим на результаты исследования. Например, оговоренный критерий исключения «тяжелые сопутствующие заболевания, препятствующие проведению лечения в запланированном объеме» в случае применения на стадии финального анализа может привести к исключению из исследования всех не подходящих под требуемые результаты пациентов. Кто сможет доказать, что исключение проведено не из-за того, что данный больной «портит статистику», а из-за тяжелого заболевания, не распознанного при включении в исследование? Кроме того, значительно исказить результаты может исключение из анализа пациентов в связи с отклонениями от плана лечения, ранней смертью или отказом больного от лечения [14]. Часто именно исключаемые под таким предлогом пациенты имеют наихудший прогноз по сравнению с больными, оставшимися в исследовании. Исключая таких больных из анализа, исследователи обычно мотивируют это тем, что плохие результаты лечения данного больного обусловлены отклонениями от лечебного плана, однако все может быть совсем наоборот. Например, в одном из кардиологических исследований 5-летняя смертность пациентов, «не выдерживавших» план лечения в группе плацебо, составила 28,3%, что было статистически значимо больше, чем у пациентов, которые «смогли» продолжить получение плацебо в жестких рамках протокола — 15,1% [15]. К примеру: больной И. не явился для очередного введения исследуемого препарата, мотивируя это слабостью и потерей веры в лечащего врача, а через 2 мес его полуживое тело с явными признаками бурного прогрессирования привезли родственники. Если исследование не организовано по принципу intention-to-treat, то, скорее всего, больной будет исключен из финального анализа (прогрессия из-за несоблюдения режима лечения). Однако вполне может оказаться, что прогрессирование началось уже на фоне лечения (слабость и потеря веры в лечащего врача были ее первыми признаками), но было выявлено лишь при последнем визите в клинику. Разумеется, если в одной группе из анализа будет исключено 50 «больных И.», а в другой только 5, то это может существенно исказить результаты исследования. В связи с этим исключение пациентов из анализа или их сепаратный анализ (что эквивалентно исключению) по любым причинам, кроме исходного несоответствия критериям отбора, обычно признается неприемлемым. И совсем уж нелепым выглядит исключение из анализа хирургических исследований больных, погибших от послеоперационных осложнений. Если в результате суперрадикальных операций в послеоперационной реанимации погибает и исключается из исследования 95% больных, но остальные 5% излечиваются (100% выживаемость), это не значит, что метод лучше более консервативного, при котором от осложнений погибает 5%, а из оставшихся 95% «лишь» половина излечивается (выживаемость «всего» 47,5%).

Анализ по намерению лечить для всех рандомизированных пациентов, отвечающих критериям отбора, должен быть основным в исследовании. Если выводы исследования основываются на анализе после исключения «неудобных» пациентов, они обычно находятся под очень большим сомнением. Во многом это обусловлено и тем, что в исследованиях оценивается весь план лечения в целом (на группе больных, соответствующих критериям отбора), а практически любое лечение не может быть проведено в полном объеме и без отклонений всем пациентам. В связи с тем что задачей исследований III фазы является выявление метода, рекомендованного для переноса в широкую клиническую практику (а не на группу отобранных пациентов с благоприятным прогнозом), все больные, отвечающие критериям включения, должны анализироваться.

Промежуточные анализы (interim analyses)

По мере набора больных у многих исследователей возникает «благая» мысль: давайте посчитаем результаты, вдруг уже получилось! При наличии современных компьютерных статистических программ обычно это не представляет никакой технической сложности и может повторяться хоть ежедневно. К сожалению, если тесты на статистическую значимость будут повторяться с большой частотой в течение времени проведения исследования, то вероятность выявить несуществующие «статистически достоверные различия» (при $p<0,05$) в одном из промежуточных анализов составляет гораздо более 5%. Например, T. Fleming и соавт. [16] показали, что шанс выявить несуществующие различия может составлять до 26% в случае, если тесты на статистическую значимость будут проводиться каждые 3 мес в исследовании, длящемся 3 года. Если посмотреть онкологическую литературу (причем не только отечественную), то можно обнаружить, что результаты многих исследований публикуются до набора планируемого количества пациентов и даже без указаний целевого объема выборки, необходимого для достаточной статистической силы исследования. В большинстве таких публикаций даже не приводятся данные, чем же является данный анализ (промежуточным тестом на статистическую значимость или заранее запланированным финальным анализом исследования). В этих случаях необходимо быть крайне осторожным при ин-

терпретации данных, и они в большинстве случаев должны расцениваться как нетрактуемые. Косвенно это подтверждается тем, что большинство таких «ранних публикаций» не получают своего продолжения (после одного из промежуточных «обнадеживающих» результатов различия вдруг исчезают и писать уже не о чем).

Таким образом, не запланированные и не обоснованные заранее промежуточные анализы могут часто вводить в заблуждение. К сожалению, публикации незапланированных промежуточных результатов, могут принести много вреда, особенно если интерпретируются врачами (и уж тем более пациентами), не имеющими образования в области статистики. Случайная «разница» в выживаемости, обусловленная неравномерным распределением прогностических признаков в небольшой группе включенных пациентов, может привести к полному прекращению включения больных в клиническое исследование и не дать врачу возможности убедить пациента в том, что в настоящее время не существует четких доказательств того, что один из видов лечения, используемых в исследовании, действительно лучше другого. Часто и сами врачи попадаются на эту «удочку», основывая свои рекомендации и предпочтения на данных опубликованных промежуточных исследований, а затем долго удивляются, увидев, что через год «статистически достоверная разница» куда-то исчезла. По этим причинам стандартом рандомизированных многоцентровых исследований стало то, что промежуточные анализы проводит комитет по мониторированию данных, а не участвующие врачи. Такой подход помогает предохранить пациентов и врачей от адаптации результатов поверхностно проведенного анализа, исследование — от разрушения, однако позволяет и вовремя отреагировать на действительно значимые изменения. В настоящее время в основном промежуточные результаты доступны только комитету по мониторингу данных, причем главные исследователи в данный комитет не входят, так как получение результатов может привести к конфликту интересов при продолжении исследования. Лишь комитет принимает решение о том, являются ли данные достаточно зрелыми для опубликования и дальнейшего применения. Однако необходимо отметить, что такой подход касается исключительно исследований III фазы.

В настоящее время используется большое количество подходов для оптимизации промежуточных анализов. Наиболее применимым и простым является предложение J. Haybittle [17]: при промежуточном анализе различия в выживаемости не принимаются в расчет, за исключением случаев, когда они являются статистически значимыми при двустороннем $p<0{,}0025$.

Уровень статистической значимости, тестирование гипотезы и конфиденциальные интервалы

Формирование медицинского мнения об эффективности или неэффективности новой методики достаточно сложно, и клиницисты достаточно часто неверно интерпретируют тест на статистическую значимость, видя в нем четкую грань между этими понятиями. Как было сказано выше, уровень статистической значимости при сравнении результатов отражает вероятность увидеть различие столь же большое, как наблюдается при анализе данных, в случае, если сравниваемые виды лечения на самом деле равны, а различие обусловлено другими факторами. Если различие в каждом из направлений оценки столь же велико в абсолютном значении, как то, которое действительно наблюдается, уровень статистической значимости называется двусторонним. Если вероятность рассчитывается только для различия того же направления, которое наблюдается при анализе, уровень значимости называется односторонним. Выбор между односторонним и двусторонним критерием формирования мнения является критичным, так как одностороннее $p=0{,}05$ является просто незначимым, если заранее предопределенная допустимая ошибка 1-го типа, равная 0,05, базируется на двустороннем тесте.

Определенный уровень статистической значимости может служить весьма значительной помощью при интерпретации результатов исследований, однако остается условно принятой величиной. В медицинских исследованиях в большинстве случаев он должен быть меньше 0,05. В ядерной физике, например, где счет идет на миллионы и миллиарды событий, он может составлять 0,0001. Таким образом, мы просто ставим себе границу «доверия»: меньше 0,05 — верю, что новое лечение лучше, больше 0,05 — не верю. Если в вашем исследовании различие достоверно при $p=0{,}000004$, а в исследовании коллеги при $p=0{,}04$, не стоит смотреть на коллегу свысока (мое исследование в 1000 раз достовернее). По принятым в медицине «правилам игры» оба различия достоверны и соотнести их напрямую не представляется возможным.

Уровень статистической значимости в большой степени находится под влиянием размера выборки, и невозможность отклонить нулевую гипотезу (равенства сравниваемых лечебных подходов) не является эквивалентом отсутствия различий в эффективности лечения (см. выше). К сожалению, простого «индекса доверия» при интерпретации результатов не существует, и как бы заманчиво ни выглядело $p=0{,}00001$, исследователь должен проходить через тщательный самостоятельный анализ данных и скептическую их оценку. В связи с этим можно настоятельно рекомендовать перед адаптацией данных исследования все же загляды-

вать в полнотекстовые первоисточники, не ограничиваясь поверхностным изучением абстрактов с приведенным уровнем статистической значимости. При прочтении вы с изумлением можете обнаружить, что различие, действительно, скорее всего, является истинным, но обусловлено причинами, далекими от проведенного лечения.

Гораздо больше информации о различии (или эквивалентности) сравниваемых методик дает величина, незаслуженно забытая отечественными онкологами, — конфиденциальный, или доверительный интервал. Конфиденциальный интервал к размеру различия в эффективности лечения представляет потенциально возможный «разброс» эффектов, обусловленный полученными данными. Уровень же статистической значимости не говорит ничего по поводу «размера» эффекта лечения, так как во многом зависит от размера выборки. В то же время именно «размер» лечебного эффекта в связи с конфиденциальным интервалом должен использоваться для «взвешенного» сопоставления цены и выигрыша принятого медицинского решения. Например, соотношение риска смерти между группами составляет 0,85, статистическая значимость 0,003. Казалось бы, выявлено статистически значимое снижение риска смерти на 15%. Однако 95% конфиденциальный интервал составляет 0,81—0,99, т.е. весьма велик шанс того, что хотя различие и существует, но истинный его размер гораздо меньше.

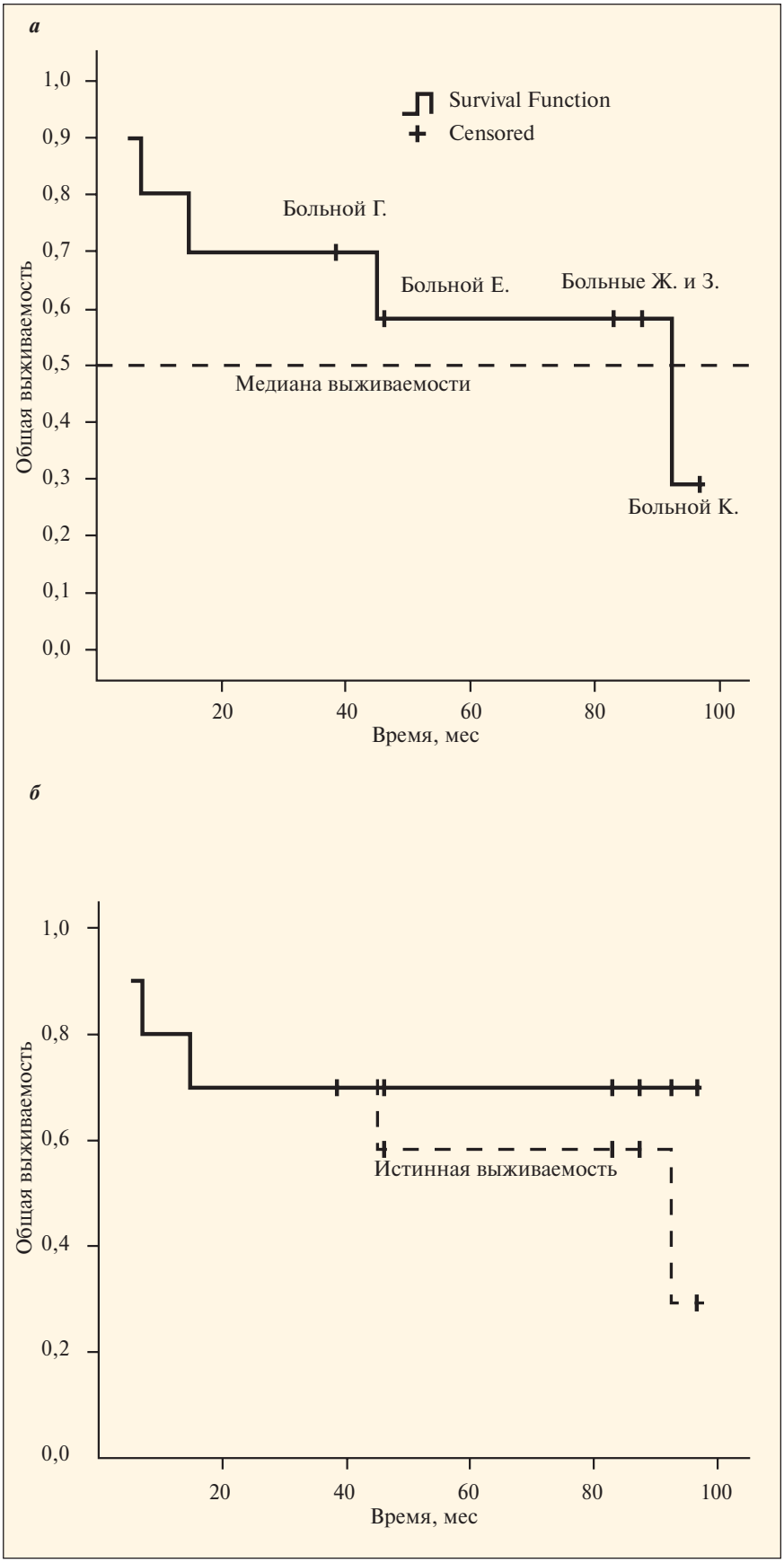
Расчет кривых выживаемости

В большинстве онкологических клинических исследований III фазы приведены результаты в виде кривых выживаемости. Кривые выживаемости отражают вероятность пережить определенное время без отрицательного события с представлением времени на горизонтальной оси. С использованием этой методики могут представляться и другие показатели, оценивающие время до события (восстановление нейтрофилов, время до госпитализации и т.д.). Обычные статистические методы неприемлемы для анализа выживаемости, так как они игнорируют тот факт, что к определенному временному интервалу (follow-up period) один пациент может не иметь отрицательного события («censored» — под наблюдением), а у другого оно может произойти в этом же временном интервале. Например, нам до сих пор приходится встречаться с отечественными исследователями, представляющими медиану выживаемости как простое распределение сроков наблюдения (жизни) по ранжиру и выбора из них центрального значения. В лучшем случае у живых на момент последнего наблюдения больных рядом со значением выживаемости стоит плюсик (или крестик). Наверное, и многим из вас знакомо такое представление данных (выживаемость больных составила 4, 5, 6+, 6, 7+ и 7 мес). Вот только 7 и 7+ это совсем разные вещи. 7+ может стать и 8 и 25, а просто «7» уже не станет никогда. Для этого и нужны способы расчета или прогнозирования выживаемости на основании ее динамики в настоящий момент. В противном случае для ее определения нам пришлось бы дожидаться смерти всех больных.

Наиболее удовлетворительным путем представления таких данных является определение функции выживаемости *S(t)*. Данная функция представляет вероятность прожить более чем *t* временных единиц. Время *t* измеряется от диагноза, начала лечения или других значимых временных точек. Для рандомизированных исследований лучшей точкой отсчета считается время рандомизации. Существует два основных метода определения *S(t)*. Один из них, так называемый метод life-table, или актуриальный метод Berkson и Gage [18], используется в настоящее время достаточно редко, так как для точности требует

Таблица **6**. *Данные для расчета кривых выживаемости по методу Каплана — Майера*

Больной	Дата включения в исследование	Дата последнего наблюдения	Время от включения до последнего наблюдения	Статус на момент последнего наблюдения
А	09.04.1995	12.09.1995	5,20	Умер
Б	01.11.1995	06.06.1996	7,27	Умер
В	17.08.1995	04.11.1996	14,83	Умер
Г	02.04.1995	02.06.1998	38,57	Жив
Д	19.09.1995	01.06.1999	45,03	Умер
Е	19.08.1995	01.06.1999	46,07	Жив
Ж	03.01.1996	30.10.2002	83,07	Жив
З	15.10.1995	20.12.2002	87,43	Жив
И	18.04.1995	20.11.2002	92,43	Умер
К	09.12.1994	16.11.2002	96,63	Жив



Определение медианы выживаемости с учетом (а) и без учета (б) статуса больного

очень большого количества пациентов. Другой метод Каплана и Майера может быть использован практически при любом количестве пациентов [19]. Для построения кривых необходимо знать как минимум 3 показателя — точка начала отсчета выживаемости (рандомизация, начало лечения и т.д.), точка окончания отсчета выживаемости (дата смерти, рецидива или дата последнего наблюдения) и статус больного на этот момент (произошло или не произошло отрицательное событие). Больные, у которых это событие на данный момент не произошло, считаются цензурируемыми. Приведем пример построения графика по методу Каплана — Майера. Предположим, что в нашем исследовании 10 больных и нам нужно определить некоторые критерии выживаемости. Точка начала отсчета выживаемости — дата включения в исследование, точка окончания отсчета выживаемости — дата последнего наблюдения, отрицательное событие — смерть от любой причины (табл. 6).

Попробуем определить медиану выживаемости «по старинке» — без учета статуса больного, а просто выстроив по ранжиру время от даты включения до последнего наблюдения. Медиана (не путать со средней!) — 45,5 мес. Для того чтобы понять, какова она на самом деле, посмотрим на рисунок, *а*.

На самом деле, медиана — временной интервал, через который отрицательное событие происходит у 50% больных. В данном случае — это 92,4 мес. Каждый уступ на графике выживаемости соответсвует отрицательному событию, произошедшему в момент времени, отображенный на горизонтальной шкале. Поперечный штрих к графику отображает, где в настоящее время «находятся» больные, у которых отрицательное событие еще не произошло (больные, остающиеся под на-

блюдением, или цензурированные случаи). В нашем примере умерли 5 больных, и на графике хорошо видны уступы, отражающие эти события. Если присмотреться внимательнее, то можно увидеть, что смерть первых четырех пациентов «опускала» график примерно на 10—15%, а вот смерть последнего больного (пациент И.) «съела» около 25—30%. Почему так произошло? Потому, что «цензурируемые» больные уже прожили то время, в течение которого наблюдались первые смерти. И каждая смерть опускала график примерно на 10% (т.е. как раз на 1/10). А вот чем дальше «по кривой», тем больше выживаемость становится прогнозируемой (расчетной), так как часть цензурируемых больных еще до этого времени «не дошли». В такой ситуации судьба больных с бóльшим сроком наблюдения «предсказывает» прогноз цензурируемых больных гораздо значительнее. Смерть больного И. свидетельствует о том, что и после 80 мес наблюдения могут случаться неблагоприятные события, а с учетом, что до этого срока дошли пока только больные И. и К., вероятность неблагоприятного события для последующих больных (Г., Е., Ж. и З.) оценивается методом Каплана — Майера весьма высоко. Однако прогноз может и измениться для пациентов Г. и Е., если, например, больные Ж. и З. переживут этот срок без отрицательного события.

Основной проблемой расчетной выживаемости по Каплану — Майеру является наличие «мертвых душ», или неинформативное цензурирование. В рандомизированных контролируемых исследованиях цензурирование означает, что пациент жив (или не имеет другого отрицательного события) и остается под наблюдением на момент анализа. Однако часто цензурированным больной становится на момент контакта, произошедшего очень задолго до окончательного анализа (потерян для наблюдения, не способен прибыть в клинику в связи с тяжелым состоянием и т.д.). Оставление этой ситуации без попытки выяснить судьбу «пропавшего» пациента может приводить к появлению в кривых выживаемости «вечно живых» больных. Возвращаясь к рисунку, можно представить, что было бы с графиком выживаемости, если больные И. и Д., например, просто исчезли бы из-под наблюдения на тех же сроках (а на самом деле умерли дома от прогрессирования, «не поставив об этом в известность» лечащего врача; см. рисунок, *б*).

Не правда ли, пунктирная линия выглядит гораздо хуже, чем сплошная с «вечно живыми» пациентами? Только вот пунктирная отражает истинное положение дел с эффективностью нашего лечения, а сплошная — неправильное (хотя, может, и очень желаемое).

Следовательно, перед проведением анализа крайне важно получить информацию о судьбе по возможности всех пациентов. Если с каким-либо пациентом не было контакта на протяжении многих месяцев и его статус неизвестен, эта информация должна быть получена до того, как будет выполняться анализ. Оценивая распределение времени с момента последнего контакта с пациентом, о котором неизвестно, является ли он умершим (произошло событие), можно получить адекватную информацию о полноценности наблюдения.

Множественные сравнения

Возможность проведения множественных сравнений в пределах одного исследования является и «счастьем», и «бедой» большинства диссертаций и многих отечественных и зарубежных исследований. С одной стороны, такой поиск позволяет получить вожделенное p, с другой — дает весьма большой шанс на выявление несуществующих различий. В табл. 7 представлена вероятность получить как минимум одну статистически значимую ($p<0,05$) «разницу» в результате нескольких независимых сравнений двух одинаковых по эффективности видов лечения.

Уже при пяти сравнениях вероятность случайного «выявления» несуществующей разницы составляет 22,6%. В случае, если число конечных целей исследования, промежуточных анализов и подгрупп пациентов в исследовании слишком велико, то этот шанс может вырасти еще более значительно [20].

Проведение поданализов лечения в отношении множества конечных целей и множественные промежуточные анализы (пусть каждый вспомнит количество выводов из собственной диссертации — вряд ли оно было менее пяти) являются частым источником неправильных выводов. Первичные цели исследования должны быть определены протоколом. Анализы по подгруппам (дополнительные анализы) и анализы вторичных целей исследования должны также быть определены заранее, а статистическая значимость должна заявляться только в случае, если уровень достоверности значительно превышает обычные 0,05. Наиболее простой способ избежать подобных ошибок — признавать результаты статистически значимыми только в случае, если уровень p составляет меньше $0,05/n$, где n — количе-

Таблица 7. *Вероятность получить статистически значимое «различие» двух одинаковых видов лечения*

Число сравнений	Вероятность выявления «различия», %
1	5
5	22,6
10	40
20	64,1

ство проведенных сравнений. Например, если n=10, то уровень p=0,005 должен быть признан статистически значимым для проведенного поданализа. Планируемое протоколом количество сравнений должно быть указано. Сравнения, которые не были заявлены заранее, не должны признаваться значимыми в любом случае и должны (в лучшем случае) являться лишь поводом для генерации гипотезы, которую планируется тестировать в последующих исследованиях.

В любом случае недопустимо проведение дополнительного анализа исходя из характеристик, которые выявляются уже после начала лечения (выполнимость процедур, полнота доз, токсичность).

Публикация результатов клинических исследований

Адекватная публикация результатов является важной и неотъемлемой частью хорошо организованного исследования. К сожалению, даже в развитых странах в большинстве обзоров отмечается плохое качество публикации результатов медицинских исследований [21]. S. Pocock и соавт. [22] пришли к выводу, что «в общем публикации результатов находятся под влиянием попыток преувеличить различие в эффективности лечения», а J. Barr и I. Tannock [23] дали прекрасную иллюстрацию того, как это легко выполняется. R. Simon and R. Wittes [24] разработали четкие методологические рекомендации по обнародованию результатов клинических исследований, и эти рекомендации приняты большинством онкологических журналов. Эти девять рекомендаций суммированы ниже.

1. Авторы должны коротко описать методы контроля качества полученных результатов (включая оценку ремиссии), чтобы быть уверенными, что данные являются полными и аккуратными.

2. Должно быть указано общее число пациентов, зарегистрированных в исследовании.

3. Доля «неоцениваемых» пациентов не должна превышать 15% для главной цели исследования.

4. В рандомизированных исследованиях отчет должен включать сравнение выживаемости и других основных конечных целей с использованием данных всех рандомизированных пациентов без исключений (кроме тех, которые не отвечают критериям отбора).

5. Размер выборки должен быть достаточным для создания осмысленного заключения по поводу наличия клинически значимого эффекта. Для «негативного» заключения при сравнении терапевтических подходов адекватность размеров выборки такому заключению должна быть подтверждена представлением конфиденциального интервала для истинного различия в эффективности лечения.

6. В отчете должен быть представлен начальный планируемый объем выборки. Также должно быть указано, какое количество промежуточных анализов было проведено и каким образом было принято решение по поводу остановки исследования и генерации отчета.

7. Заявление о терапевтической эффективности не должно основываться на результатах нерандомизированных исследований II фазы, за исключением случаев, когда заболевание является очень редким или прогноз заболевания настолько плох, что проведение контролируемого рандомизированного исследования невозможно. В таком случае результаты нерандомизированного исследования должны сравниваться с историческим контролем, в котором используются сравнимые по характеристикам пациенты, данные которых могут быть тщательно оценены. Сравнение выживаемости между ответившими и не ответившими на лечение пациентами не может являться доказательством терапевтической эффективности.

8. Пациенты, включенные в исследование, должны быть детально описаны. Возможность переноса заключений исследования на всех больных должна быть адекватно отражена в дискуссии. Заявление о различии в эффективности лечения по подгруппам должно быть тщательно документировано и подтверждено статистическими тестами.

9. Методы статистического анализа должны быть описаны в объеме, достаточном для того, чтобы обладающий знаниями читатель мог воспроизвести результаты исследования при наличии данных.

ЛИТЕРАТУРА

14. Barr J., Tannock I. Analyzing the same data two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. J Clin Oncol 1989;7:969.
15. Group C.D.R. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. N Engl J Med 1980;302:1038.
16. Fleming T.R., Green S.J., Harrington D.P. Considerations of monitoring and evaluating treatment effects in clinical trials. Control Clin Trials 1984;5:55.
17. Haybittle J.L. Repeated assessment of results in clinical trials of cancer treatment. J Radiol 1971;44:793.
18. Berkson J., Gage R.P. Calculation of survival rates for cancer. Proc Mayo Clin 1950;25:270.
19. Kaplan E.I., Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457.
20. Tannock I.F. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. J Natl Cancer Inst 1996;88:206.
21. Begg C.B. Quality of clinical trials. Ann Oncol 1990;1:319.
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.
23. Barr J., Tannock I. Analyzing the same data two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. J Clin Oncol 1989;7:969.
24. Simon R., Wittes R.E. Methodologic guidelines for reports of clinical trials. Cancer Treat Rep 1985;69:1.