# Методология расчёта объёма выборки в сравнительных контролируемых клинических исследованиях с «неменьшей эффективностью»: сравнение двух пропорций в параллельных группах

**Ляшенко А.А.[1,2], Свищева М.С.[2]**

[1] — *Первый МГМУ им И.М. Сеченова, НИИ молекулярной медицины, г. Москва*

[2] — *ООО «Центр медицинских биотехнологий», г. Москва, www.cmbio.ru*

**Резюме.** Подавляющее большинство клинических исследований генерических препаратов направлены на сравнение терапевтической эффективности тестируемого препарата и препарата сравнения. При этом необходимо доказать, что тестируемый препарат не менее эффективен, чем препарат сравнения. Подобные дизайны получили название исследований «неменьшей эффективности»; в основном — это исследования III фазы.

Довольно часто в этих исследованиях в качестве первичного критерия эффективности используются частоты какого-либо изучаемого признака, как результат анализа бинарных данных. В свою очередь, бинарные данные сравниваются между собой методом сравнения пропорций. При этом одним из важнейших условий правильной интерпретации данных, полученных в ходе клинического исследования, является обеспечение исследования необходимым и достаточным объёмом экспериментальных данных.

Цель этой статьи — показать, как планировать эксперимент, обеспечивая его необходимым и достаточным объёмом выборки, для того, чтобы полученные результаты и выводы были бы достоверными; как интерпретировать данные в сравнительных контролируемых исследованиях, направленных на установление «неменьшей эффективности» при сравнении пропорций в параллельных группах.

**Ключевые слова:** исследования «неменьшей эффективности», граница «неменьшей эффективности», сравнительные контролируемые исследования в параллельных группах, доверительный интервал, формула подсчёта выборки, сравнение пропорций

**The methodology of calculation of sample size in "non-inferiority" comparative controlled clinical trials: a comparison of two proportions in parallel group**

Lyashenko A.A.[1,2], Svishcheva M.S.[2]

[1] — First MGMU named after I.M. Sechenov, Institute of Molecular Medicine, Moscow

[2] — LLC «Center of Medical Biotechnology,» Moscow, www.cmbio.ru

**Abstract.** The absolute majority of clinical trials of generic drugs aimed to compare the therapeutic efficacy of the tested drug and the drug of an active control. It is necessary to estimate that the test drug is not less effective (or non-inferior) than the control drug. The designs of the aforementioned trials are called "non-inferiority" study; often, these are phase III of clinical trials.

The primary criterions of effectiveness which are quite often used in the clinical trials are frequencies of signs, as a result of the analysis of binary data. Binary data are analyzed by comparing proportions. One of the most important conditions for a correct interpretation of the data obtained during the clinical trials — to provide necessary and sufficient sample size.

The purpose of this article is to show how to plan the study, how to provide necessary and sufficient sample size to ensure that the results and conclusions would be reliable; how to interpret the data in a comparative controlled study aimed to establish "non-inferiority" using proportions comparing in parallel groups.

**Key words**: "non-inferiority" trials, "non-inferiority" margins, comparative controlled study in parallel groups, the Confidence Interval, sample size calculating formulae, proportions comparison

Автор, ответственный за переписку:
*Ляшенко Алла Анатольевна — к.б.н., ведущий научный сотрудник НИИ молекулярной медицины 1 МГМУ им И.М. Сеченова, генеральный директор ООО «Центр медицинских биотехнологий», г. Москва. e-mail: allaliachenko@yandex.ru, тел: +7(916) 222-64-51, www.cmbio.ru*

В современных клинических исследований довольно часто в качестве первичного критерия эффективности выбирается частота какого-либо признака; в сравнительных клинических исследованиях эти частоты необходимо сравнить, чтобы получить вывод о преимуществе какого-либо из сравниваемых препаратов. При сравнении частот признака в таких случаях применяются методы сравнения пропорций. Традиционно частоты признака составляют бинарные данные. Примером бинарных данных могут быть следующие: «выздоровел/заболел», «есть симптом/нет симптома», «да/нет», «м/ж» и т.д. Т.е. вариантов изучаемых признаков может быть только два. Здесь мы не будем подробно останавливаться на том, как сравнивать пропорции и отсылаем читателя к любому программному обеспечению по статистике.

Цель этой статьи — показать, как планировать эксперимент, чтобы полученные результаты и выводы были бы достоверными, и как их интерпретировать в сравнительных контролируемых исследованиях, направленных на установление «неменьшей эффективности» при сравнении пропорций.

Одним из важнейших условий правильной интерпретации данных, полученных в ходе клинического исследования, является обеспечение исследования необходимым и достаточным объёмом экспериментальных данных. От того, насколько адекватным будет это число, зависит уровень «доверия» к полученным выводам и, следовательно, и их дальнейшее использование.

Нисколько не умаляя важность получения правильных выводов в «неклинических» исследованиях, необходимо помнить, что выводы из результатов, полученных из клинического исследования, являются наиболее ответственными среди всех других видов исследований, поскольку от того, насколько они верны, зависит здоровье и жизнь человека, принимающего изучаемое лекарство. Так как же обезопасить исследование от ошибочных выводов и сделать так, чтобы доверие к полученным результатам было максимальным?

Зачастую при планировании медицинских исследований исследователи (или спонсоры) пренебрегают (или не учитывают вовсе) такой важный аспект, как объём выборки. Объём выборки, строго говоря, подразумевает необходимое и достаточное количество пациентов в группах, при котором вероятность принятия правильного решения является максимальной. Зачем нужно планировать эксперимент и нельзя ли ограничиться привычным подходом (происхождение которого, впрочем, авторам неизвестно), который долгое время использовался в медицине? Например, взять 30 пациентов, провести с ними исследования и обработать полученные данные? Оказывается, нельзя. И причина здесь вовсе не только в том, что их может «не хватить». Основная причина в том, что при таком подходе мы можем быть уверены только в том, что полученные выводы распространяются именно на этих 30 пациентов. А «интересы» остальных пациентов с аналогичным диагнозом, которых существенно больше, при таком подходе были как бы «проигнорированы». Стало быть, полученные выводы являются случайными и не могут быть распространены на всех пациентов, для которых разработано изучаемое лекарственное средство. С другой стороны, привлечь всех пациентов с конкретной проблемой невозможно: их точное количество не знает никто, и это количество может варьироваться: одни будут поправляться, а другие — появляться вновь. Пользуясь статистической терминологией, мы не сможем привлечь к исследованиям всю «генеральную совокупность», которую составляют эти пациенты, но при этом хотим распространить на неё все полученные выводы. Для этого существует компромиссный и единственно правильный вариант, — тщательное планирование медицинских экспериментов. Планирование является ключевым шагом в доказательной медицине и гарантирует, с соблюдением несложных правил, получение выводов, которые можно с уверенностью распространить в дальнейшем на всю генеральную совокупность, которую в нашем случае составляют все потенциальные пациенты.

Для начала введём необходимые термины, которыми обычно оперируют при расчёте объёма выборки. В принципе, эти определения можно найти во множестве литературных источников, посвященных как классической статистике, так и доказательной медицине. Они приводятся в настоящей статье в несколько «адаптированном» виде для удобства чтения и восприятия дальнейшего материала.

«Ошибка 1 рода» (α) — это вероятность отвергнуть нулевую гипотезу, хотя на самом деле она истинна. Это своеобразная ложно-позитивная ошибка, например, вероятность обнаружить отличия между группами, если их на самом деле нет.

«Уровень значимости» — это допустимая вероятность ошибки 1 рода. Т.е. насколько мы готовы ошибиться, отклоняя нулевую гипотезу при условии, если она верна. Чаще всего в медико-биологических исследованиях уровень значимости при двусторонних тестах выбирают 5%. Другими словами, мы уверены на 95%,

что различия достоверны. В исследованиях «неменьшей эффективности» стандартным подходом при проверке нулевой гипотезы рекомендуется использовать односторонний тест при уровне значимости 2,5% (International Conference on Harmonization (ICH), E9, 1998). Другими словами, достоверность выводов составляет 97,5%. Иногда выбирают и другие, более уровни значимости (0.05, 0.01, 0.001), — их выбор зависит от целей исследования и от задач, которые оно решает.

«Ошибка 2 рода» (β) — это вероятность принять нулевую гипотезу, если она ложная. Это т.н. ложно-негативная ошибка, например, вероятность принять гипотезу о том, что отличий нет, хотя они на самом деле есть. Чаще всего, этот параметр не превышает 0,2.

«Статистическая мощность» — имеет смысл, обратный значению β, — вероятности совершить ошибку 2 рода. Статистическая мощность рассчитывается, как 1-β, и чаще всего выбирается не ниже 0,8.

*Za* и *Zв* — табличные критические значения нормального распределения, соответствующие заданным уровням ошибок 1 рода и выбранного уровня значимости α. Наиболее часто используемые значения величин *Za* и *Zв* приведены в табл. 1. Полные таблицы с расчётными критическими значениями Z можно найти в специализированной статистической литературе.

*Таблица 1*

**Критические значения Z нормального распределения (частичный пример)**

| α | 1-β | Z |
|---|---|---|
| 0,05 | 0,95 | 1,6449 |
| 0,025 | 0,975 | 1,96 |
| 0,1 | 0,9 | 1,2816 |
| 0,2 | 0,8 | 0,8416 |

«Нулевая гипотеза (H0)» — это гипотеза, которая традиционно проверяется в процессе статистического анализа. Пример: предположение отсутствия различий между сравниваемыми переменными, отсутствие корреляции и т.д. Если в процессе статистического анализа гипотеза H0 отвергается, то при этом необходимо принять альтернативную гипотезу.

*«Альтернативная гипотеза (H1)»*. В соответствии с предыдущим определением, эта гипотеза принимается, если гипотеза H0 отвергается. Например, в ходе клинического исследования, направленного на сравнение терапевтической эффективности сравниваемых препаратов, использовалась нулевая гипотеза об отсутствии различий между ними по основному критерию клинической эффективности X при уровне значимости $p<0.05$. По результатам данного исследования было получено, что имеются отличия между препаратами по данному признаку X ($p<0.05$). Это означает, что нулевая гипотеза H0 отвергается и принимается альтернативная H1 о существовании различий между препаратами по признаку X.

*Односторонний тест.* Односторонние тесты используются, если при анализе полученных данных принимается допущение, что показатель одной из сравниваемых групп лучше показателя другой. Т.е. заранее определяется направление «влияния вмешательства». Поэтому односторонние тесты используются в исследованиях, направленных как на установление «неменьшей эффективности», так и на установление превосходства.

*Двусторонний тест* используется, если заранее неизвестно о преимуществе сравниваемых показателей, либо нет смысла обозначать преимущество одного показателя перед другим. В симметричных исследованиях, направленных на установление эквивалентности (например, исследования биоэквивалентности дженериков), используются двусторонние тесты. Проверяемые двусторонние гипотезы в исследованиях эквивалентности выглядят следующим образом:

$$H0: p1=p2; H1: p1\neq p2$$

От того, какой подход будет выбран, зависит выбор критического значения уровня ошибки I рода Za, и, следовательно, и объём выборки.

*Исследования «неменьшей эффективности»* — это, в основном, сравнительные исследования с препаратом активного контроля, в которых предполагается, что тестируемый препарат (Т) не хуже, чем препарат сравнения (С). Для сравнения, в исследованиях превосходства устанавливается явное преимущество тестируемого препарата над плацебо.

При сравнении двух препаратов (методов лечения и т.п.) в исследовании «неменьшей эффективности» принимается допущение, что тестируемый препарат Т не хуже, чем препарат (или метод) активного контроля (С). При этом допускается, что он может быть немного хуже, но не больше, чем на некую величину. Эта некая величина, предел, который позволяет отвергнуть нулевую гипотезу и принять альтернативную о том, что сравниваемый препарат «не менее эффективен», чем препарат сравнения. В литературе её называют *пределом или границей «неменьшей» эффективности* (или «non-inferiority margins», δ). Для расчётов объёма выборки в исследованиях превосходства δ всегда положительна, в исследованиях «non-inferiority» — всегда отрицательна. Эту логику легко понять из рис. 1.

Проверяемые гипотезы для исследования «неменьшей эффективности» будут следующие:

$$H_0: P_T-P_C\leq -\delta; H_1: P_T-P_C > -\delta$$

С учётом вышесказанного, можно отметить, что наиболее логичным для таких исследований было бы название «не большей эффективности» (прим. авторов), но мы будем пользоваться устоявшейся для этого типа дизайна терминологией.

Для сравнения, проверяемые гипотезы для исследования превосходства выглядят следующим образом:

$$H_{0:}\ P_T-P_C\leq\delta; H_1: P_T-P_C >\delta$$

### Пример расчёта объёма выборки

Пусть необходимо провести расчёт объёма выборки для сравнительного контролируемого исследования двух препаратов, тестируемого (Т) и препарата сравнения (С). Пусть это будут антибиотики, основной критерий эффективности — доля полного выздоровления пациентов.

Выбираем входные условия, которые необходимы при расчёте:

— уровень значимости альфа: 2,5%; $Z_{0,025}$=1,96;
— уровень статистической мощности: 0,8; $Z_{0,8}$=0,84.
— $P_C$ и $P_T$ — клиническая эффективность препаратов, активного контроля и тестируемого, соответственно. Пусть известно, что $P_C$ (по первичному критерию эффективности) составляет 80%. Данный параметр берётся из предыдущих клинических исследований, в которых эффективность препарата сравнения, выступающего в данном клиническом исследовании в роли активного контроля, сравнивалась с плацебо, либо со стандартной терапией, и показал такой результат. Пусть желаемая клиническая эффективность тестируемого препарата $P_T$ = 60%, границу «неменьшей эффективности» выбираем в 10% (т.е. δ = -0.1). Проверяемые гипотезы:

$$H_0: P_T-P_C \leq -0,1; H_1: P_T-P_C > -0,1$$

Объём выборки в каждой группе при сравнении пропорций для исследований с неменьшей эффективностью рассчитывается по формуле (Dunnett & Gent,1977):

$$n=(Za+Zb)^2\times(pc\times(1-pc)+pt\times(1-pt))/(pc-pt-\delta)^2$$

Подставляем данные и получаем, что объём выборки n составит не менее 35 человек в каждой из сравниваемых групп. Т.е. всего для участия в клиническом исследовании необходимо вовлечь как минимум 70 пациентов. Более точная формулировка — «при уровне значимости 2,5% для сохранения статистической мощности в 0,8 клиническое исследование должны закончить не менее 70 пациентов». Поскольку существует риск, что пациенты по каким-то причинам могут отказаться от участия и выйти из клинического исследования, полученный объём выборки целесообразно увеличить на 25-30%.

С уменьшением предела неменьшей эффективности объём выборки существенно возрастает. Т.е. величины предела неменьшей эффективности и объёма выборки обратно пропорциональны: чем меньшее отличие мы стремимся установить, тем больший объём выборки требуется. Приведём здесь несколько готовых расчётов объёма выборки для демонстрации этой важной составляющей клинического исследования (табл. 2).

Несомненно, адекватный выбор границ клинической эффективности — очень важный параметр. Его величина гарантирует, что в процессе исследования мы не только оценим статистическую значимость результатов исследования, выдержав статистическую мощность, но и не потеряем клиническую значимость полученных выводов.

Какими правилами следует руководствоваться при выборе границ неменьшей эффективности? При сравнении пропорций в исследованиях неменьшей эффективности следует пользоваться рекомендациями от CPMP (2004 г.) и FDA (1992 г.). Данный документ от FDA рекомендует учитывать уровень первичного критерия эффективности препарата активного контроля, а CPMP — использовать константу, равную -10. Границы неменьшей эффективности представлены в табл. 3. Некоторые «готовые» решения по выбору границ неменьшей эффективности при анализе разницы пропорций можно также найти в работах Hou с соавт. (Hou et al., 2009 г.) и Julious (Julious, 2010 г.). Общие принципы изложены также и в рекомендациях EMEA (E9, 1998 г.; E10, 2001 г.).

### Интерпретация и представление результатов

После завершения клинического исследования необходимо провести статистический анализ полученных результатов. Традиционно для ответа на вопрос, доказана ли «неменьшая эффективность» тестируемого препарата по сравнению с препаратом активного контроля, кроме статистических критериев для проверки гипотез, используется оценка значения доверительного интервала (ДИ) разницы полученных эффектов: если нижняя граница расчётного ДИ больше, чем — δ, тогда $H_0$ отвергается и принимается альтернативная гипотеза о «неменьшей эффективности» тестируемого препарата по сравнению с препаратом сравнения. Используется либо двусторонний 95% ДИ, либо односторонний 97,5% ДИ. Иллюстрация возможных вариантов приведена на рис. 1 (по Schumi & Wittes, 2011 г.).

Пусть в результате клинического исследования получен односторонний 97,5% ДИ [-8,4; 1]. Выбранный предел «неменьшей эффективности» δ равен -10. В данном клиническом исследовании доказана «неменьшая эффективность» опытного препарата.

Допустим, что в результате проведённого клинического исследования был получен двусторонний 95% ДИ [-6; -1]. Выбранный предел «неменьшей эффективности» выберем -5. В данном случае «неменьшая эффективность» тестируемого препарата не подтверждена.

### Заключение

Таким образом, при расчёте объёма выборки при сравнении пропорций в параллельных группах в сравнительном контролируемом исследовании неменьшей эффективности необходимо выбрать эффективность препарата сравнения из предшествующих клинических исследований, определиться с эффективностью тестируемого препарата, установить границы неменьшей эффективности δ, выбрать уровень значимости и статистической мощности исследования. Интерпретация результатов клинического исследования проводится после определения границ доверительного

*Таблица 2*

**Расчёты объёма выборки (n) при сравнении двух пропорций для уровня значимости 2,5% и статистической мощности 0,8. Условные обозначения: PC — эффективность препарата сравнения; PT — эффективность тестируемого препарата; δ – предел неменьшей эффективности; n – объём выборки в каждой из сравниваемых групп.**

| $P_C$ | $P_T$ | δ | n | $P_C$ | $P_T$ | δ | n | $P_C$ | $P_T$ | δ | n |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0,90** | 0,85 | -0,05 | 171 | **0,80** | 0,75 | -0,05 | 273 | **0,70** | 0,65 | -0,05 | 343 |
| | | -0,10 | 76 | | | -0,10 | 121 | | | -0,10 | 153 |
| | | -0,15 | 43 | | | -0,15 | 68 | | | -0,15 | 86 |
| | | -0,20 | 27 | | | -0,20 | 44 | | | -0,20 | 55 |
| | 0,80 | -0,05 | 87 | | 0,70 | -0,05 | 129 | | 0,6 | -0,05 | 157 |
| | | -0,10 | 49 | | | -0,10 | 73 | | | -0,10 | 88 |
| | | -0,15 | 31 | | | -0,15 | 46 | | | -0,15 | 57 |
| | | -0,20 | 22 | | | -0,20 | 32 | | | -0,20 | 39 |
| | 0,75 | -0,05 | 54 | | 0,65 | -0,05 | 76 | | 0,55 | -0,05 | 90 |
| | | -0,10 | 35 | | | -0,10 | 49 | | | -0,10 | 57 |
| | | -0,15 | 24 | | | -0,15 | 34 | | | -0,15 | 40 |
| | | -0,20 | 18 | | | -0,20 | 25 | | | -0,20 | 29 |
| | 0,70 | -0,05 | 38 | | 0,60 | -0,05 | 50 | | 0,50 | -0,05 | 58 |
| | | -0,10 | 26 | | | -0,10 | 35 | | | -0,10 | 40 |
| | | -0,15 | 19 | | | -0,15 | 26 | | | -0,15 | 29 |
| | | -0,20 | 15 | | | -0,20 | 20 | | | -0,20 | 23 |

интервала разницы эффекта от сравниваемых препаратов и соотношения полученного ДИ (двустороннего 95% или одностороннего 97,5%) по отношению к выбранному пределу неменьшей эффективности. Если левая граница ДИ больше, чем — δ, нулевая гипотеза отвергается и признается альтернативная о том, что тестируемый препарат не менее эффективен, чем препарат сравнения.

*Таблица 3*

**Границы «неменьшей эффективности» по рекомендациям CPMP и FDA**

| Эффективность препарата активного контроля | FDA | CPMP |
|---|---|---|
| >90% | -10 | -10 |
| 80-89% | -15 | -10 |
| 70-79% | -20 | -10 |

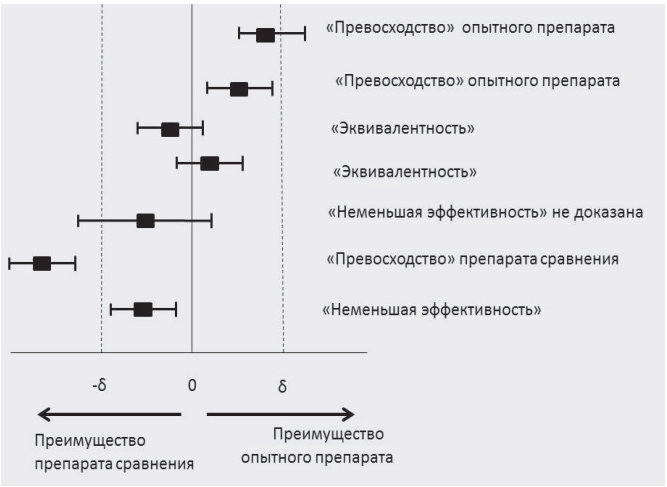


**Рис. 1.** Схематичное представление вариантов исходов клинического исследования в зависимости от значений доверительных интервалов разницы эффективности опытного препарата и препарата сравнения по отношению к пределу клинической эффективности δ

**Литература**


1. *Schumi J., Wittes J.T.* Through the looking glass: understanding non-inferiority. // Trials. 2011 May 3;12:106.
2. *Dunnett C.W., Gent M.* Significance testing to establish equivalence between treatments, with special reference to data in the form of 2X2 tables. // Biometrics.1977 Dec;33(4):593-602.
3. International Conference on (ICH) of technical requirements for registration of pharmaceuticals for human use. // Statistical principles for clinical trials, 1998.
4. Food and Drug Administration (FDA) (1992) Points to consider. Clinical evaluation of Anti-infective drug products.
5. *Hou Y., Wu X.Y., Li K.* Issues on the selection of non-inferiority margin in clinical trials. // Chin Med J (Engl). 2009 Feb 20;122(4):466-70.
6. Committee for proprietary medicinal products (CPMP) (2004). Points to consider on the choice of non-inferiority margin.
7. Committee for proprietary medicinal products (CPMP) (2000) Points to Consider on Switching Between Superiority and Non-Inferiority.
8. *Julious S.A.* Sample Sizes for Clinical Trials. CRC Press 2010, p 180.
9. ICH Note for Guidance E9 (Statistical Principles for Clinical Trials) (1998), CPMP/ICH/363/96.
10. ICH Note for Guidance E10 (Choice of Control Group) (2001), CPMP/ICH/364/96