



Проверка статистических гипотез: общие подходы в практике медицинских исследований

А.Ю. Суворов, Н.М. Буланов✉, А.Н. Шведова, Е.А. Тао, Д.В. Бутнару,
М.Ю. Надинская, А.А. Заикин

ФГАОУ ВО «Первый Московский государственный медицинский университет им. И.М. Сеченова»
Минздрава России (Сеченовский Университет)
ул. Трубецкая, д. 8, стр. 2, г. Москва, 119991, Россия

Аннотация

Проверка статистических гипотез – один из ключевых этапов современных исследований в области медицины. На начальном этапе ученые выдвигают исследовательскую гипотезу, на основе которой формулируют статистическую гипотезу, которая поддается проверке с помощью статистических тестов. В руководстве представлены примеры составления нулевой и альтернативной статистических гипотез для разных исследовательских вопросов, представлен общий алгоритм их проверки на примере t -теста. Авторы разбирают концепции ошибок I рода, которые необходимы для интерпретации p -значений, полученных в статистических тестах, и ошибок II рода, которые используют для расчета мощности исследования. Существенное внимание уделено понятию размера эффекта и его оценке, различиям между статистически и клинически значимыми эффектами. Продемонстрирована взаимосвязь между размером эффекта, численностью выборки и величиной ошибки II рода.

Ключевые слова: статистика как раздел; нулевая гипотеза; ошибка I рода; ошибка II рода; размер эффекта; t -тест

Рубрики MeSH:

СТАТИСТИКА КАК ТЕМА

БИОМЕДИЦИНСКИЕ ИССЛЕДОВАНИЯ

НАУЧНЫХ ИССЛЕДОВАНИЙ ПЛАНИРОВАНИЕ

Для цитирования: Суворов А.Ю., Буланов Н.М., Шведова А.Н., Тао Е.А., Бутнару Д.В., Надинская М.Ю., Заикин А.А. Проверка статистических гипотез: общие подходы в практике медицинских исследований. Сеченовский вестник. 2022; 13(1): 4–13. <https://doi.org/10.47093/2218-7332.2022.426.08>

КОНТАКТНАЯ ИНФОРМАЦИЯ:

Буланов Николай Михайлович, канд. мед. наук, доцент кафедры внутренних, профессиональных болезней и ревматологии ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет)

Адрес: ул. Трубецкая, д. 8, стр. 2, г. Москва, 119991, Россия

Тел.: +7 (919) 100-22-79

E-mail: bulanov_n_m@staff.sechenov.ru

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Финансирование. Исследование не имело спонсорской поддержки (собственные ресурсы).

Поступила: 05.04.2022

Принята: 11.05.2022

Дата публикации онлайн: 16.06.2022

Дата печати: 23.06.2022

Statistical hypothesis testing: general approach in medical research

Alexander Yu. Suvorov, Nikolay M. Bulanov✉, Anastasia N. Shvedova, Ekaterina A. Tao,
Denis V. Butnaru, Maria Yu. Nadinskaia, Alexey A. Zaikin

Sechenov First Moscow State Medical University (Sechenov University)
8/2, Trubetskaya str., Moscow, 119991, Russia

Abstract

Statistical hypothesis testing is one of the key steps in modern medical research. Initially, scientists formulate a research hypothesis based on which the statistical hypothesis is then developed and statistically tested. This review provides the null and alternative hypotheses' compiling examples for different research questions and the general algorithm for their testing using t-test. The authors also describe type I errors, which are necessary to interpret *p-values* estimated from statistical tests, and type II errors, which are used to assess study power. The article focuses on effect size and its calculation methods, and the difference between statistically significant and clinically significant effects. The associations between effect size, sample size, and type II error are also discussed.

Keywords: statistics as topic; null hypothesis; type I error; type II error; effect size; *t*-test

MeSH terms:

STATISTICS AS TOPIC

BIOMEDICAL RESEARCH

RESEARCH DESIGN

For citation: Suvorov A.Yu., Bulanov N.M., Shvedova A.N., Tao E.A., Butnaru D.V., Nadinskaia M.Yu., Zaikin A.A. Statistical hypothesis testing: general approach in medical research. Sechenov Medical Journal. 2022; 13(1): 4–13. <https://doi.org/10.47093/2218-7332.2022.426.08>

CONTACT INFORMATION:

Nikolay M. Bulanov, Cand. of Sci. (Medicine), Associate Professor, Department of Internal, Occupational Diseases and Rheumatology, Sechenov First Moscow State Medical University (Sechenov University)

Address: 8/2, Trubetskaya str., Moscow, 119991, Russia

Tel.: +7 (919) 100-22-79

E-mail: bulanov_n_m@staff.sechenov.ru

Conflict of interests. The authors declare that there is no conflict of interests.

Financial support. The study was not sponsored (own resources).

Received: 05.04.2022

Accepted: 11.05.2022

Published Online: 16.06.2022

Date of publication: 23.06.2022

Список сокращений

H_0 – нулевая гипотеза

H_1 (H_A) – альтернативная (нулевой) гипотеза

АД – артериальное давление

ОР – относительный риск

САД – систолическое артериальное давление

Проверка гипотез является одним из краеугольных камней современных медицинских исследований. Цель многих исследований не ограничивается простым описанием данных, а включает в себя поиск различий между характеристиками тех или иных объектов наблюдений (пациентов, животных, клеточных культур) и оценку их значимости. Любое наблюдение за объектами реального мира с активным вмешательством исследователя или без него может называться экспериментом. В медицинской науке одним из способов проведения экспериментов является выполнение планируемых исследований [1]. Концепция планируемых исследований характеризуется четкими правилами проведения эксперимента с обозначением ряда жестких условий, которые должны быть обозначены и выполнены до его инициации:

- цель исследования сопровождается четко поставленным *вопросом исследования*;
- ясно сформулированы конкретные *задачи*, с помощью которых будет достигнута цель;

- обозначены одна или несколько исследовательских гипотез, требующих проверки в рамках вопроса исследования;
- дизайн исследования направлен на максимально эффективное и надежное достижение цели, получение достоверных и воспроизводимых результатов за счет снижения вероятности возникновения ошибок;
- критерии включения и невключения в исследование, а также критерии исключения обозначены однозначно;
- пошагово описаны методы статистического анализа, направленного на получение выводов по каждой задаче и поставленному вопросу исследования.

Таким образом, при планировании эксперимента исследователи моделируют некие идеальные условия, которые позволяют ответить на поставленный вопрос исследования. При этом планируемые исследования достаточно компактны по количеству участников и по времени проведения.

Примером несколько иного подхода являются наблюдательные поисковые исследования, основной задачей которых является не подтверждение заранее сформулированных гипотез о влиянии тех или иных факторов на исход, а поиск любых подобных взаимодействий и генерация гипотез. То есть в обсервационных исследованиях во главу угла ставят поиск любых важных с точки зрения цели исследования взаимодействий факторов внутри исследуемой популяции, при этом узкий основной вопрос исследования, требующий подтверждения, обычно отсутствует. Однако и в поисковых исследованиях должны быть заранее сформулированы: цель, задачи и четко обозначена исследуемая популяция (обычно довольно широкая по сравнению с выборками планируемых исследований).

Целью этого обзора является знакомство читателей с основными аспектами планируемых исследований и облегчение понимания связи между вопросом исследования и ответом на него с точки зрения методологии проведения исследований, ключевым аспектом которого является проверка статистических гипотез.

ИССЛЕДОВАТЕЛЬСКИЕ И СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ

Началом любого исследования (эксперимента) является гипотеза или научное предположение. Исследователи выдвигают такие предположения и пытаются их доказать или опровергнуть. Например, исследователи могут сделать предположение о том, что плохая экология региона оказывает влияние на здоровье жителей, или о том, что курение может быть сопряжено с повышенным риском сердечно-сосудистых заболеваний. Такие предположения получили название «исследовательские гипотезы». От исследовательской гипотезы следует отличать статистическую гипотезу. Последняя, по сути, представляет такую формулировку исследовательской гипотезы, которая может быть проанализирована с помощью статистических методов в рамках концепции дизайна эксперимента.

Статистическая гипотеза представляет собой некое суждение о параметрах, описывающих статистическую популяцию (генеральную совокупность), но не выборку из нее. В свою очередь, *статистическая популяция* представляет собой группу однородных элементов, например людей в группе риска, представляющих интерес в рамках настоящего исследования (эксперимента). Примером может служить население города или все пациенты стационара за определенный период времени. На начальном этапе исследователи выдвигают две гипотезы о возможной взаимосвязи наблюдаемых явлений (потенциальных факторов риска и исходов): нулевую и альтернативную.

Нулевая гипотеза утверждает, что наблюдаемые эффекты, явления или взаимодействия происходят в силу случайности, то есть связь между ними отсутствует. Нулевую гипотезу традиционно обозначают, как H_0 . *Альтернативная гипотеза*, наоборот, утверждает, что наблюдаемые явления неслучайны и между ними есть связь. Альтернативную гипотезу традиционно обозначают как H_1 или H_A .

Следует иметь в виду, что под связью в данном случае подразумевают любую ассоциацию, не обязательно причинно-следственную. Например, в небольшом исследовании оценивают различия средних значений (μ_1 и μ_2) вещественного числового признака – уровень общего холестерина плазмы крови – в двух группах: у пациентов с и без инфаркта миокарда в анамнезе. Исследовательская гипотеза может состоять в том, что группы как различаются, так и не различаются. При этом нулевая гипотеза будет утверждать, что уровень общего холестерина никак не связан с риском развития инфаркта миокарда, то есть истинных различий между средними значениями нет (наблюдаемые различия носят случайный характер):

$$H_0 : \mu_1 = \mu_2.$$

Альтернативной гипотезой является утверждение о том, что различия в уровне холестерина между группами существуют, они значимы и неслучайны:

$$H_1 : \mu_1 \neq \mu_2.$$

Нулевая и альтернативная гипотезы являются *взаимоисключающими*, то есть если верна H_0 , то неверна H_1 , и наоборот. Таким образом, чтобы подтвердить альтернативную гипотезу – наличие истинных различий между группами, – нам нужно отклонить нулевую гипотезу.

СТАТИСТИЧЕСКИЕ ТЕСТЫ

Статистические тесты – это методы статистического доказательства (построения статистического вывода), которые используют для принятия решения о том, можно ли отклонить нулевую гипотезу H_0 . Следует отметить, что формально статистические тесты не позволяют принять нулевую гипотезу. Они лишь помогают оценить, может ли она быть отклонена в пользу альтернативной или нет.

Каждый статистический тест является математической функцией, вычисляющей так называемую *тестовую статистику*. Тестовая статистика показывает, насколько близко наблюдаемая величина соответствует ожидаемому распределению величин при условии, что нулевая гипотеза не была отклонена. Чем больше величина тестовой статистики, тем больше несоответствие между наблюдаемым и ожидаемым распределениями.

Совокупность тестовых статистик каждого теста подчиняется определенному закону распределения (подробнее о распределениях см. в [2]). Например, *t*-тест вычисляет *t*-статистики, подчиняющиеся *t*-распределению. Для того чтобы отклонить нулевую гипотезу, нужно обозначить некий *порог среди распределения тестовых статистик*. Такую величину пороговой тестовой статистики называют *критическим значением*, а соответствующую ей вероятность в распределении тестовых статистик – *уровнем значимости*.

Уровень значимости, обозначаемый α , является вероятностью того, что при текущем значении тестовой статистики нулевая гипотеза будет отклонена несмотря на то, что она верна (истинна). Иными словами, величина α отражает вероятность ошибочно отклонить верную

нулевую гипотезу. Такая вероятность также получила название *ошибка I рода*. Величину α и соответствующее ей критическое значение статистического теста исследователи декларируют до проведения статистического теста, чтобы заранее определить вероятность ошибочного отклонения нулевой гипотезы. Чем меньше уровень значимости, тем более низкая вероятность отклонить нулевую гипотезу в случае, если она верна. Однако эта вероятность, пусть в ряде случаев и очень небольшая, существует всегда. Исследователи лишь могут выбрать такое пороговое значение, при котором эта вероятность будет чрезвычайно низкой. В разных областях знаний традиционно используют разные значения α ; в частности, в медицинских исследованиях наиболее часто применяют пороговое значение, равное 0,05. Исследователь может использовать и более низкие пороговые значения, например 0,01 или ниже в экспериментах, где критически важно снизить вероятность ошибочного отклонения нулевой гипотезы (обнаружения различий между группами там, где их нет).

После вычисления результата статистического теста в виде тестовой статистики исследователи получают соответствующую ей *вероятность получения таких же или более экстремальных по сравнению с наблюдаемыми результатами теста (то есть сильнее отклоняющихся от ожидаемого распределения), в случае если нулевая гипотеза верна*. Такая вероятность получила название *p-value* (*p*-значение). Малые значения *p-value* говорят о том, что если нулевая гипотеза верна, то вероятность получения таких же или более экстремальных результатов тестовой статистики крайне мала. Следовательно, имеется высокая вероятность того, что нулевая гипотеза неверна и может быть отклонена. Если $p \leq \alpha$, то есть статистика теста равна или превышает критическое значение, результат считается статистически значимым (рис. 1). Именно поэтому в описании статистического анализа в разделе «Материалы и методы» всегда указывают, при каком значении *p* результаты считают статистически значимыми. Фактически фраза «различия считали значимыми при $p < 0,05$ » означает, что исследователи выбрали для своей работы величину ошибки I рода $\alpha = 0,05$.

Представим себе, что мы проводим параллельное сравнительное исследование с двумя группами: одной группе назначаем антигипертензивный препарат А, другой группе назначаем плацебо. Наша нулевая гипотеза утверждает, что эффективность препарата А в отношении показателей артериального давления (АД) не отличается от плацебо, если мы оцениваем средние цифры АД по завершении исследования в обеих группах:

- μ_{TRT} – среднее значение АД среди пациентов, получавших препарат А (treatment, TRT);
- μ_{PLC} – среднее значение АД среди пациентов, получавших плацебо (placebo, PLC);
- $H_0 : \mu_{TRT} = \mu_{PLC}$ или $\mu_{TRT} - \mu_{PLC} = 0$;
- $H_1 : \mu_{TRT} \neq \mu_{PLC}$ или $\mu_{TRT} - \mu_{PLC} \neq 0$.

С помощью *t*-критерия проверяем, можем ли мы отклонить нулевую гипотезу о том, что $\mu_{TRT} - \mu_{PLC} = 0$. Мы можем построить график распределения *t*-статистик, соответствующего такой разнице (рис. 1, зеленые линии).

После завершения эксперимента мы получили разницу $\mu_{TRT} - \mu_{PLC}$, соответствующую *t*-статистике 2,5 в случае, если нулевая гипотеза верна (рис. 1, красная линия). Мы видим, что наблюдаемый результат лежит за пределами критических значений, что позволяет отклонить нулевую гипотезу. Соответствующее значение *p* составляет 0,0126, следовательно, нулевая гипотеза может быть отклонена при выбранном значении $\alpha = 0,05$, а различие между группами является статистически значимым.

РАЗМЕР ЭФФЕКТА

Исследовательские гипотезы наиболее часто связаны с поиском различий или ассоциаций между некоторыми показателями. Однако существенное значение имеет не столько сам факт различий, сколько их клиническая значимость. Например, в одном исследовании сравнивали доли достижения терапевтического эффекта в группах лечения и плацебо, и они составили 10 и 80% соответственно. Мы видим, что доли различаются, и довольно существенно, разница составляет 70%. В другом исследовании аналогичные доли составили 45 и 55%. Мы снова видим, что доли различаются, однако уже не так сильно, разница всего 10%. Еще в одном исследовании эффективность терапии в двух группах составила 75 и 80%. Доли различаются, но разница очень невелика – всего 5%. Другой пример связан с новыми антигипертензивными препаратами. Новый препарат X снижает систолическое артериальное давление (САД) в среднем на 15 мм рт. ст. Препарат Y также снижает САД, но в среднем на 8 мм рт. ст. Наконец, препарат Z снижает

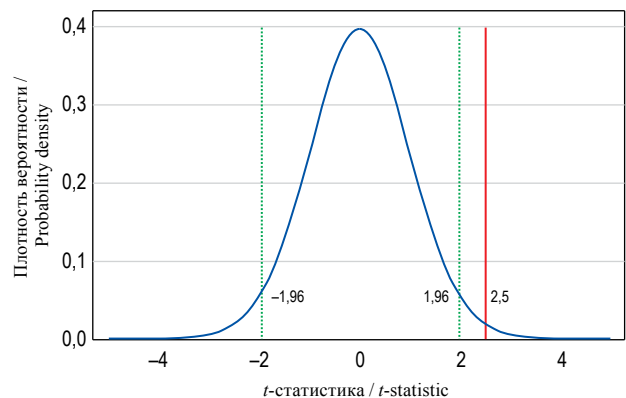


РИС. 1. T-распределение для 50 степеней свободы
FIG. 1. T-distribution for 50 degrees of freedom

Примечания: зелеными пунктирными линиями указаны критические значения -1,96 и 1,96, соответствующие 2,5-му и 97,5-му перцентилям – двустороннему уровню значимости $\alpha = 0,05$. Красной сплошной линией обозначена *t*-статистика = 2,5, которая превышает критическое значение 1,96 при двустороннем *t*-тесте ($p = 0,0126$ при заданном $\alpha = 0,05$). Таким образом нулевая гипотеза может быть отклонена и принята альтернативная.

Notes: green dotted lines indicate critical values of 1.96 and -1.96 corresponding to the 97.5th percentile and the 2.5th percentile, respectively, i.e. two-sided significance level $\alpha = 0,05$. B. Red solid line indicates *t*-statistics = 2.5, which exceeds critical value of 1.96 using two-tailed *t*-test ($p = 0.0126$ with specified $\alpha = 0.05$), thus the null hypothesis can be rejected, and we can accept the alternative one.

САД, но всего в среднем на 1 мм рт. ст. При использовании традиционной антигипертензивной терапии среднее снижение АД составило 1 мм рт. ст. Во всех примерах мы видим, что некий эффект есть, но он разный, в одних – больший, в других – меньший.

Размер эффекта – довольно широкое статистическое понятие, обозначающее некую статистику или показатель, показывающий величину различий или ассоциации между распределениями исследуемой величины в разных группах. Размер эффекта в медицине крайне важен: именно он привязан не только к статистике, но и к клинической значимости наблюдаемых в исследовании результатов. Например, мы используем среднее снижение САД в качестве размера эффекта. Среднее снижение САД в 1 мм рт. ст., скорее всего, не является клинически значимым – для пациента препарат Z не будет лучше традиционной терапии, таким образом, с практической точки зрения его назначение не дает преимуществ. Слишком большой размер эффекта (препарат X), напротив, может быть ассоциирован с развитием осложнений вследствие гипотензии. Препарат Y, вероятно, является оптимальным выбором среди новых препаратов в клинической практике – он имеет значимый и при этом не чрезмерный антигипертензивный эффект.

Крайне важно понимать, что, проверяя статистические гипотезы, мы пытаемся статистическими методами зафиксировать определенный размер эффекта. Абсолютная разница между средними редко бывает равна нулю, при этом разница может быть небольшой и клинически незначимой и принимать как положительные, так и отрицательные значения. С другой стороны, при проверке статистических гипотез необходимо четко ввести критерий наличия или отсутствия клинического смысла у того или иного эффекта. В примере с антигипертензивными препаратами среднее снижение САД

на ≤ 2 мм рт. ст. можно определить как отсутствие клинического эффекта, на 3–10 мм рт. ст. – как умеренный эффект и на >10 мм рт. ст. – как сильный. В качестве проверяемой исследовательской гипотезы мы хотим выяснить, достигнет ли среднее снижение САД хотя бы умеренного размера эффекта под влиянием препаратов X, Y и Z.

Как оценивается размер эффекта?

Огромный вклад в концепцию размера эффекта внес психолог и статистик Jacob Cohen, который в одной из своих поздних работ писал: «Основным результатом исследования являются одна или несколько оценок размера эффекта, а не р-значения» [3]. Сегодня существует большое количество статистик, позволяющих оценить размер эффекта, фактически при проверке любых гипотез.

Выделяют стандартизованные методы оценки размера эффекта и нестандартизованные. В отличие от последних, стандартизованные методы позволяют оценивать эффект для переменных не только с одинаковой, но и с разной размерностью (например, оценка коэффициента корреляции для переменных, измеряемых в разных единицах), для оценки совокупных результатов разных исследований (метаанализ и метарегрессия), при сравнении результатов исследований с использованием разных метрик переменных (например, при использовании г/л в одном исследовании и ммоль/л в другом) [4].

Выделяют следующие методы оценки размера эффекта (таблицы S1–4 в приложении):

- размер эффекта, оценивающий ассоциацию между распределениями числовых переменных или насколько распределение одной переменной вносит вклад в распределение другой переменной (коэффициент корреляции, коэффициент детерминации и др.);
- размер эффекта, оценивающий разницу между статистиками (Cohen's d, Glass' Δ , разница рисков и др.);

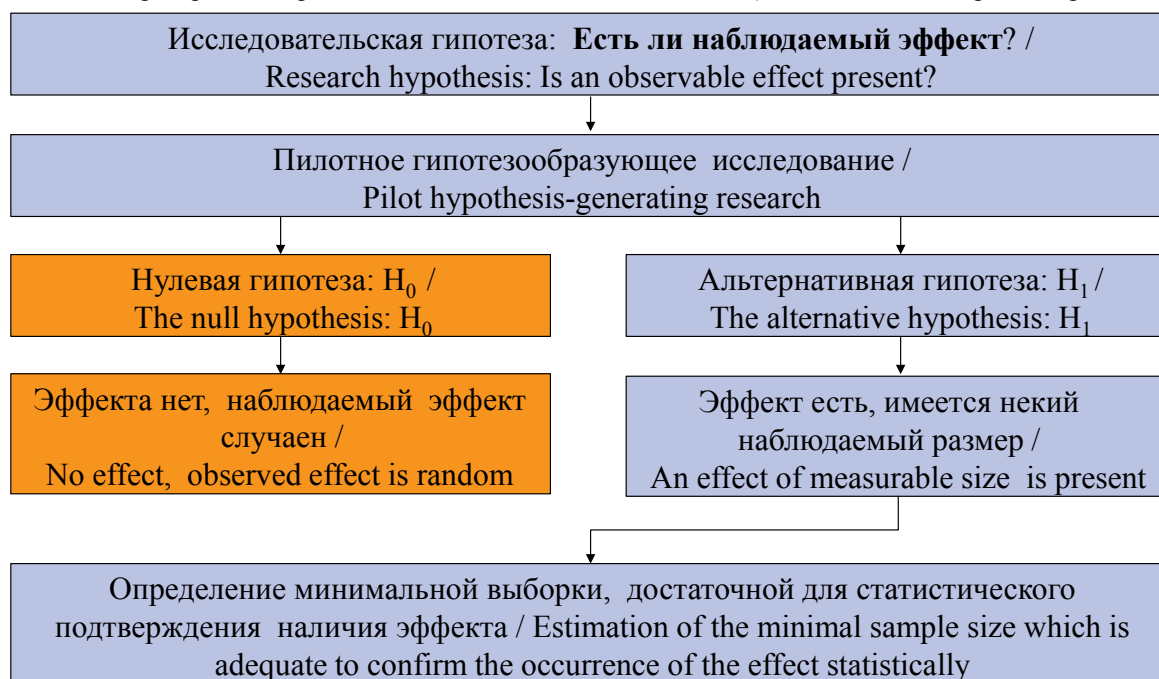


РИС. 2. Схема проведения гипотезообразующих исследований
FIG. 2. Flowchart of the hypothesis-generating studies

- размер эффекта, оценивающий ассоциацию между категориальными переменными (Cohen's h , отношение шансов и др.).

Исследовательские гипотезы и концепция размера эффекта

Размеру эффекта отводится ключевая роль при формировании исследовательских и статистических гипотез. Первоначально исследователи ставят вопрос о том, случаен ли наблюдаемый ими эффект? Например, различаются ли в действительности уровни общего холестерина в группе лечения новым препаратом и в контрольной группе? Наблюдаемые различия могут быть обусловлены случайностью. Для того чтобы проверить, существует ли эффект в действительности, проводят поисковые и пилотные исследования, основной целью которых является определение наличия эффекта или его отсутствия. Такие исследования получили названия гипотезообразующих (о различных подходах к оптимальному выбору дизайна исследований для различных целей – см. [5]). Безусловно, если удалось зафиксировать эффект, в пилотных исследованиях можно оценить его наблюдаемый или гипотетический размер. Однако пилотные исследования часто довольно компактны и дают возможность лишь ответить на вопрос наличия/отсутствия эффекта, но не позволяют достоверно определить его размер (рис. 2).

Следующий этап исследований после пилотных получил название подтверждающих исследований и направлен на то, чтобы зафиксировать эффект определенного размера. Например, в пилотном исследовании было установлено, что курение среди мужчин 35–45 лет, проживающих в городах, увеличивает риск развития сердечно-сосудистых заболеваний за 10 лет, а относительный риск (ОР) равен X . Перед исследователями

встает вопрос о влиянии курения на аналогичную группу мужчин, проживающих в сельской местности. Для того, чтобы спланировать такое исследование, мы можем опираться на полученные ранее результаты в городской популяции и исходить из того, что нам нужно зафиксировать размер эффекта (ОР) не менее X (рис. 3). Или, если по нашим исследовательским предположениям эффект будет менее выражен, например в n раз, мы можем спланировать исследование так, чтобы зафиксировать размер эффекта (ОР) не менее X/n .

Использование размера эффекта позволяет не проводить пилотные исследования каждый раз, а опираться на опыт предшествующих работ. Концепция размера эффекта требует от врачей понимания концепции порога размера эффекта, который они хотят зафиксировать статистически, если такой порог является целесообразным с точки зрения медицины. И такой порог требует именно медицинского обоснования. Например, при исследовании нового препарата для похудения у пациентов с весом выше 200 кг получено статистически значимое снижение веса в течение одного года, которое составило 1 кг. В результате эффект зафиксирован, он статистически значим, но с точки зрения помощи пациентам такой эффект абсолютно лишен всякого смысла: в течение 1 года диетологи (и сами пациенты) наверняка хотели бы наблюдать более выраженное снижение веса. Вероятно, более оправданным было бы введение порога размера эффекта в 10 или 15 кг.

ОДНОСТОРОННИЕ И ДВУСТОРОННИЕ СТАТИСТИЧЕСКИЕ ТЕСТЫ

Двусторонние тесты

Вернемся к примеру с антигипертензивной терапией. Предположим, что существует новый перспективный препарат, назовем его TRT (treatment), который

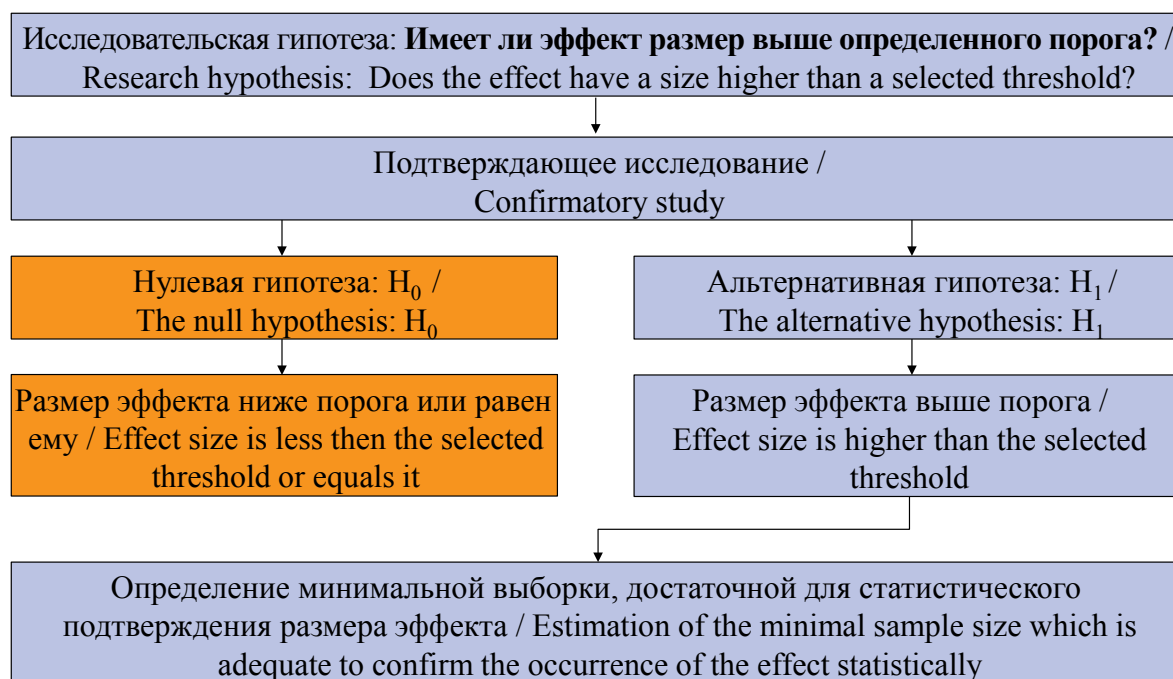


РИС. 3. Схема проведения подтверждающих исследований

FIG. 3. Flowchart of confirmatory studies

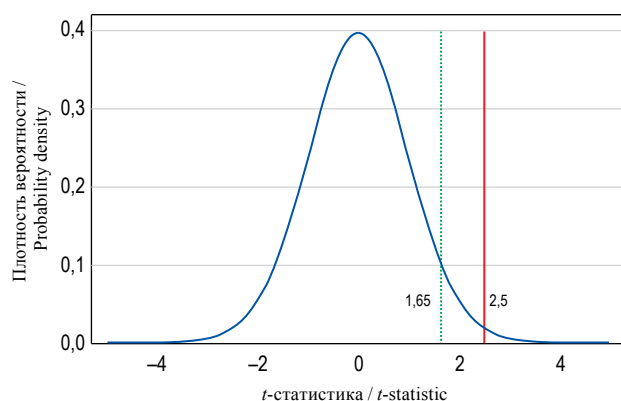


РИС. 4. Иллюстрация одностороннего t -теста
FIG. 4. One-sided t -test illustration

Примечания: зеленая линия обозначает критическое значение 1,65, соответствующее одностороннему уровню значимости $\alpha = 0,05$. Для статистики $t = 2,5$ (красная линия) p -value = 0,006: результат статистически значим при выбранном значении ошибки I рода, поскольку $p < \alpha$.
Notes: the green line indicates critical value of 1.65 which corresponds with one-sided significance level $\alpha = 0.05$; p -value is 0.006 for t -statistic $t = 2.5$ (red line): the result is clinically significant, given the chosen type I error value as $p < \alpha$.

должен снижать САД исходя из своего механизма действия, но как он себя покажет в клиническом эксперименте с пациентами, мы не знаем. Существует и традиционная антигипертензивная терапия, которая будет использована в контрольной группе (CTRL, control treatment). Измеренное среднее снижение САД в конце исследования в группе TRT будет равно μ_{TRT} , а в группе CTRL составит μ_{CTRL} .

Если мы формулируем вопрос исследования, как «какая терапия более эффективна?», нулевая гипотеза декларирует, что $H_0: \mu_{TRT} = \mu_{CTRL}$. А альтернативная гипотеза утверждает обратное $H_1: \mu_{TRT} \neq \mu_{CTRL}$, и, в свою очередь, может состоять из двух более простых утверждений:

$$H_1 = \begin{cases} \mu_{TRT} > \mu_{CTRL} \\ \mu_{TRT} < \mu_{CTRL} \end{cases}$$

То есть мы рассматриваем альтернативные гипотезы и для ситуации, когда новый препарат (TRT) оказался более эффективным, чем традиционное лечение (CTRL), и, наоборот, когда новый препарат (TRT) оказался менее эффективным. Тесты, используемые для такой проверки разнонаправленных по сути предположений, получили название *двусторонних*. Если рассмотреть диаграмму распределения,

например t -статистики, мы увидим 2 зеркальных критических значения с разным знаком: при уровне значимости $\alpha = 0,05$ для двустороннего теста критические значения будут -1,96 и 1,96 (рис. 1). При двусторонних тестах общий уровень значимости разделяется пополам и критические значения с каждой стороны соответствуют $\alpha/2$:

$$\begin{cases} \alpha_{right} = \alpha/2 \\ \alpha_{left} = \alpha/2 \end{cases}$$

При принятом уровне значимости в 5% нулевая гипотеза будет отклонена, если t -статистика наблюдаемого эффекта превысит любое из двух критических значений (зеленые линии) (рис. 1), соответствующие 2,5 и 97,5 процентиля.

Для чего исследователям двусторонние тесты? Так как мы не знаем истинного эффекта препарата TRT, двусторонний тест ответит на все варианты развития событий: TRT приблизительно одинаков по действию с CTRL, лучше или хуже CTRL.

Односторонние тесты

Если главный вопрос исследования «является ли новый препарат (TRT) лучшей альтернативой стандартному лечению (CTRL)?», наши гипотезы изменятся. Теперь нам важно зафиксировать только значимый размер эффекта, когда $\mu_{TRT} > \mu_{CTRL}$.

Таким образом, при формулировании нулевой и альтернативной гипотезы получаем:

$$\begin{cases} H_0 = \mu_{TRT} \leq \mu_{CTRL} \\ H_1 = \mu_{TRT} > \mu_{CTRL} \end{cases}$$

Для проверки такой гипотезы используется *односторонний тест*, позволяющий зафиксировать не только *определенный размер эффекта*, но и его *направление*. В данном случае нам важно проверить, превышает ли статистика теста критическое значение, расположенное на распределении справа (рис. 4).

ОШИБКИ I И II РОДА

Мы подробно разобрали концепцию ошибки I рода, однако получение и интерпретация результатов исследований связано не с одним, а с двумя типами ошибок (табл.).

Ошибка первого рода (α , ложноположительный результат) – ситуация, когда отклонена верная нулевая гипотеза. Принимается альтернативная гипотеза, которая неверна. Например, исследователи считают значимыми

Таблица. Ошибки I и II рода

Table. Type I and II errors

В статистической популяции / In statistical population	В ходе исследования / In the study	Результат проверки H_0 / After testing H_0	Вероятность / Probability
H_0 верна / H_0 true	H_0 не отклонена / H_0 not rejected	Решение не отклонять верное / Decision not to reject is correct	$P = 1 - \alpha$
H_0 верна / H_0 true	H_0 отклонена / H_0 rejected	Ошибочное отклонение, ошибка I рода / Incorrect (false) rejection, Type I error	$P = \alpha$
H_0 неверна / H_0 false	H_0 не отклонена / H_0 not rejected	Решение не отклонять ошибочное, ошибка II рода / Decision not to reject is incorrect (false), Type II error	$P = \beta$
H_0 неверна / H_0 false	H_0 отклонена / H_0 rejected	Верное отклонение / Correct rejection	$P = 1 - \beta$

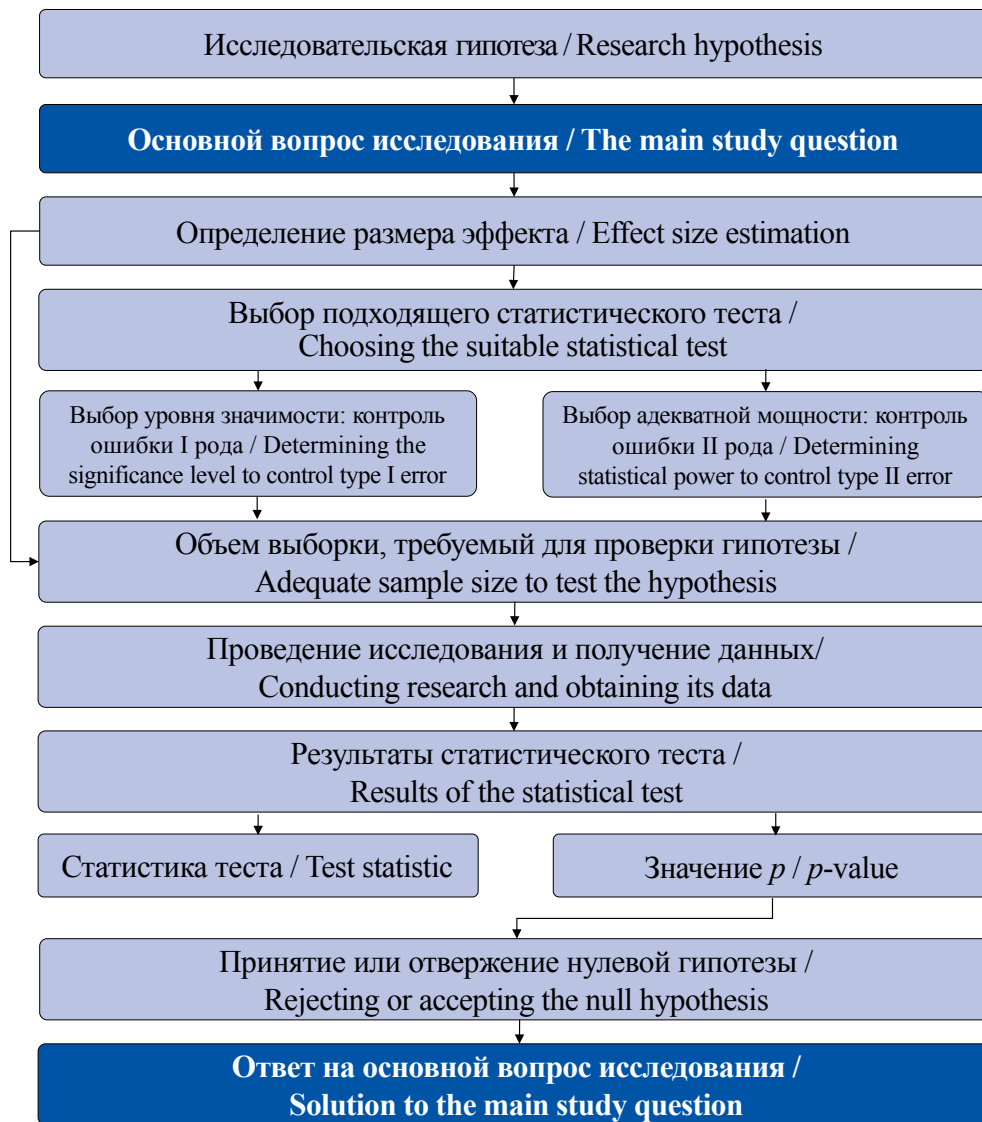


РИС. 5. Схема проверки статистических гипотез
FIG. 5. Flowchart of the statistical hypothesis testing

различия между группами, а на самом деле различия носят случайный характер.

Ошибка второго рода (β , ложноотрицательный результат) – ситуация, когда не отклонена ошибочная нулевая гипотеза. При этом верная альтернативная гипотеза отклоняется. Например, исследователи расценили как случайные различия между группами, которые на самом деле были значимы и не случайны.

Для планирования эксперимента важно попытаться минимизировать ошибки I и II рода. Ошибка I рода, как мы уже говорили, и является уровнем значимости теста, с которым сопоставляют величину p . Малое значение ошибки I рода позволяет с высокой вероятностью не отклонить нулевую гипотезу при условии, что она верна.

В свою очередь, ошибка II рода отражает возможность отклонить ошибочную нулевую гипотезу с вероятностью $1 - \beta$. Такая вероятность получила название *мощность статистического теста*: $power = 1 - \beta$. В ряде медицинских исследований общепринятая минимальная

мощность соответствует не менее 80% (то есть максимально допустимая ошибка II рода не превышает 20%).

ОБЪЕМ ТРЕБУЕМОЙ ВЫБОРКИ В КОНЦЕПЦИИ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Последним элементом, необходимым для проверки статистических гипотез, является минимальный объем требуемой выборки, необходимый для того, чтобы принять или отклонить нулевую гипотезу [6]. Таким образом, мы можем представить концепцию проверки статистических гипотез в виде схемы, представленной на рис. 5.

Сформулировав вопрос исследования и предположив ожидаемый размер эффекта, исследователь:

- выбирает наиболее подходящий статистический тест, связанный с законом распределения размера эффекта в статистической популяции;
- устанавливает подходящий уровень значимости и мощность исследования;

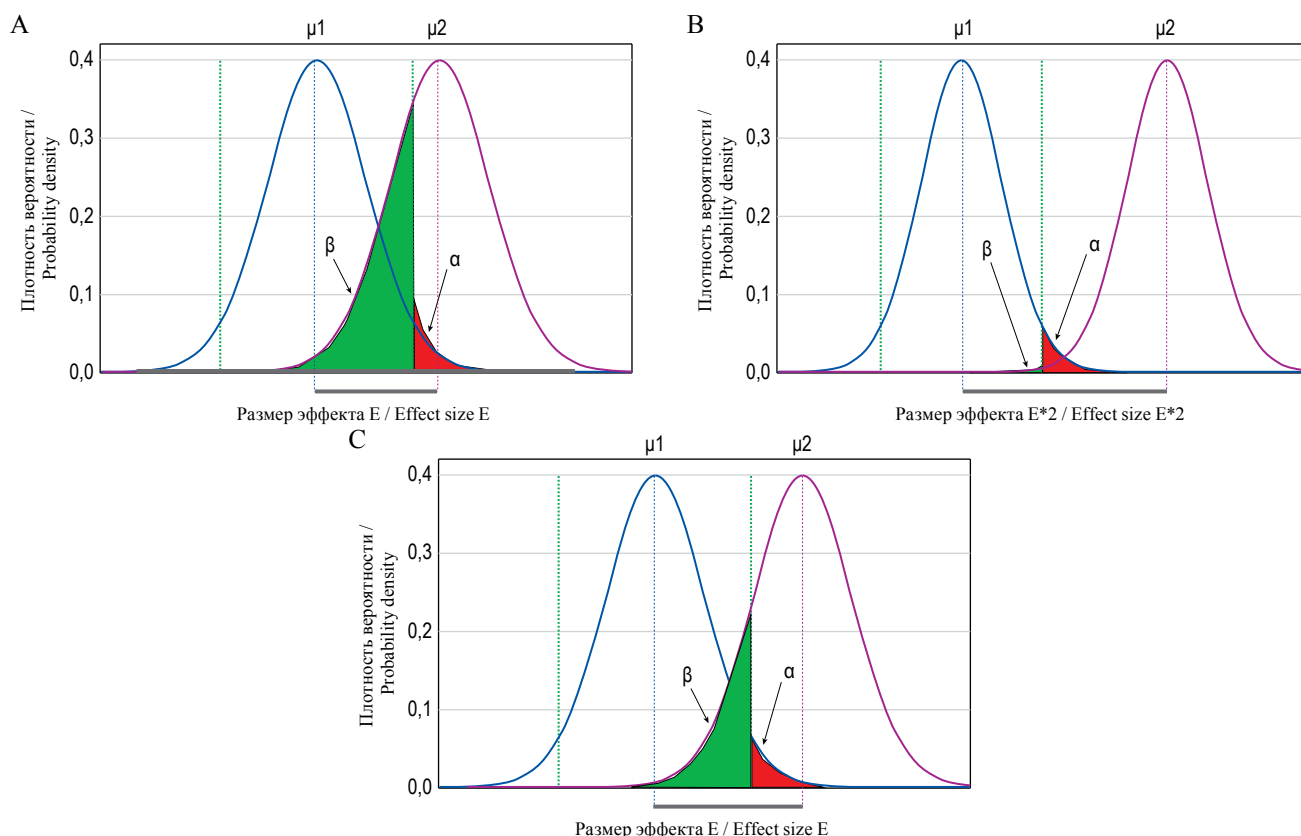


РИС. 6. Взаимосвязь между размером эффекта, ошибками I и II рода и размером выборки при проверке статистических гипотез: А. Взаимосвязь размера эффекта, ошибок I и II рода.

В. Изменение величины ошибки II рода при увеличении размера эффекта.

С. Изменение величины ошибки II рода при увеличении размера выборки.

FIG. 6. The relationship between effect size, type I and II errors, and sample size, when testing statistical hypotheses:

A. Effect size and type I and II errors.

B. Type II error changing after the effect size increasing.

C. Type II error changing after sample size increasing.

Примечание: μ_1 – среднее в группе 1, μ_2 – среднее в группе 2.

Note: μ_1 – mean in group 1, μ_2 – mean in group 2.

- после этого рассчитывает требуемый объем выборки.

Только после выполнения этих этапов можно переходить к выполнению статистического теста. Однако каким образом исследователь может связать мощность исследования, уровень значимости, размер эффекта и объем выборки? Представим себе, что мы пытаемся статистически зафиксировать размер эффекта E с мощностью $1 - \beta$, значимостью α (рис. 6А). Для этого нам потребуется объем выборки n [7]. Заранее отметим, что уровень значимости теста α должен оставаться фиксированным при любом развитии событий. Если исследователь хочет оставить объем выборки неизменным, но при этом повысить мощность, самое простое решение – предположить, что мы будем наблюдать больший размер эффекта, например в 2 раза (или $E \times 2$). При таких условиях мощность действительно увеличится (рис. 6В). Однако в реальной жизни исследователь не может по собственному желанию наблюдать больший или меньший эффект, более того, предположение о размере эффекта представляет из себя лишь исследовательскую гипотезу. В таком случае размер

эффекта также на самом деле не должен увеличиться, однако при увеличении размера выборки увеличивается мощность исследования (рис. 6С). Следовательно, при желании зафиксировать определенный размер эффекта со строгим уровнем значимости единственной возможностью снизить риск ложного принятия ошибочной нулевой гипотезы (β) является увеличение объема требуемой выборки.

ЗАКЛЮЧЕНИЕ

Грамотное формулирование исследовательских и статистических гипотез – важнейший навык врача-исследователя, без которого невозможно успешное планирование и проведение исследований в области медицины. Кроме того, концепции размера эффекта, ошибок I и II рода необходимы для интерпретации результатов своих собственных и опубликованных в литературе исследований. Эти идеи универсальны и применимы к любым статистическим тестам, более того, они имеют существенно большее значение для ученого, чем навык применения тех или иных частных методик.

ВКЛАД АВТОРОВ

А.Ю. Суворов, Н.М. Буланов, А.Н. Шведова в равной степени внесли вклад в эту работу и должны считаться первыми соавторами. А.Ю. Суворов, Н.М. Буланов, А.Н. Шведова, Е.А. Тао, А.А. Заикин и М.Ю. Надинская участвовали в написании текста рукописи. А.Ю. Суворов, Н.М. Буланов и А.Н. Шведова выполняли поиск и анализ литературы по теме обзора. А.Ю. Суворов и Д.В. Бутнару разработали общую концепцию статьи и осуществляли руководство ее написанием. Все авторы участвовали в обсуждении и редактировании работы. Все авторы утвердили окончательную версию публикации.

ДОПОЛНИТЕЛЬНЫЕ МАТЕРИАЛЫ

Дополнительные материалы, прилагаемые к этой статье, можно посмотреть в онлайн-версии по адресу: <https://doi.org/10.47093/2218-7332.2022.426.08.S>

ЛИТЕРАТУРА / REFERENCES

1. *Kestenbaum B.* Epidemiology and biostatistics: An introduction to clinical research. Springer, Cham, 2019, 246p. <https://doi.org/10.1007/978-3-319-96644-1>
2. *Bulanov N.M., Suvorov A.Yu., Blyuss O.B., et al.* Basic principles of descriptive statistics in medical research. *Sechenov Medical Journal.* 2021; 12(3): 4–16. <https://doi.org/10.47093/2218-7332.2021.12.3.4-16>
3. *Cohen J.* Things I have learned (so far) *Am Psychol.* 1990; 45(12): 1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>
4. *Kirkwood B.R., Sterne J.A.C.* Essential medical statistics.

AUTHOR CONTRIBUTIONS

Alexander Yu. Suvorov, Nikolay M. Bulanov, and Anastasia N. Shvedova contributed equally to this work and should be considered as co-first authors. Alexander Yu. Suvorov, Nikolay M. Bulanov, Anastasia N. Shvedova, Ekaterina A. Tao, Alexey A. Zaikin and Maria Yu. Nadinskaia, participated in writing the text of the manuscript. Alexander Yu. Suvorov, Nikolay M. Bulanov, and Anastasia N. Shvedova searched and analyzed the literature on the review topic. Alexander Yu. Suvorov and Denis V. Butnaru developed the general concept of the article and supervised its writing. All authors participated in the discussion and editing of the work. All authors approved the final version of the publication.

SUPPLEMENTARY MATERIALS

Supplementary materials associated with this article can be found in the online version at doi: <https://doi.org/10.47093/2218-7332.2022.426.08.S>

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Суворов Александр Юрьевич, канд. мед. наук, главный статистик Центра анализа сложных систем ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-2224-0019>

Буланов Николай Михайлович✉, канд. мед. наук, доцент кафедры внутренних, профессиональных болезней и ревматологии ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-3989-2590>

Шведова Анастасия Никитична, студентка 6-го курса клинического института детского здоровья им. Н.Ф. Филатова ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0001-7518-2273>

Тао Екатерина Александровна, канд. мед. наук, ассистент кафедры внутренних, профессиональных болезней и ревматологии ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-0621-7054>

Бутнару Денис Викторович, канд. мед. наук, проректор по научной работе ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0003-2173-0566>

Надинская Мария Юрьевна, канд. мед. наук, доцент кафедры пропедевтики внутренних болезней, гастроэнтерологии и гепатологии ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-1210-2528>

Заикин Алексей Анатольевич, канд. физ.-мат. наук, заместитель директора Центра анализа сложных систем ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0001-7540-1130>

Alexander Yu. Suvorov, Cand. of Sci. (Medicine), Chief Statistician, Centre for Analysis of Complex Systems, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-2224-0019>

Nikolay M. Bulanov✉, Cand. of Sci. (Medicine), Associate Professor, Department of Internal, Occupational Diseases and Rheumatology, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-3989-2590>

Anastasia N. Shvedova, 6th year student, Institute of Child's Health named after N. F. Filatov, Sechenov First Moscow State Medical University (Sechenov University)
ORCID: <https://orcid.org/0000-0001-7518-2273>

Ekaterina A. Tao, Cand. of Sci. (Medicine), Assistant Professor, Department of Internal, Occupational Diseases and Rheumatology, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-3989-2590>

Denis V. Butnaru, Cand. of Sci. (Medicine), Vice-rector for Research, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0003-2173-0566>

Maria Yu. Nadinskaia, Cand. of Sci. (Medicine), Associate Professor, Department of Internal Medicine Propaedeutics, Gastroenterology and Hepatology, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-1210-2528>

Alexey A. Zaikin, Cand. of Sci. (Phys. and Math.), Deputy Director, Centre for Analysis of Complex Systems, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0001-7540-1130>

✉ Автор, ответственный за переписку / Corresponding author