**К вопросу о воспроизводимость классического эффекта Бергера открывания глаз на изменении когерентности ЭЭГ человека**

**ВВЕДЕНИЕ (постановка проблемы).**

По мере нарастания темпов накопления нейрофизиологических данных, все больше отмечается плохая воспроизводимость ранее наблюдаемых эффектов не только в изучении «тонких» когнитивных функций, но классических ЭЭГ реакций. Одним из таких ЭЭГ-эффектов безусловно является эффект Бергера – при открывании глаз, по сравнению с закрытыми глазами, отмечается падение мощности альфа ритма. Говорят, о так называемой десинхронизации альфа-ритма. Классическим проявлением эффекта Бергера при анализе когерентности является \_\_\_\_\_\_.

Вместе с тем, даже при рассмотрении данного классического эффекта, отражающего весьма выраженную разницу в физиологическом состоянии оперативного покоя, отмечается значительная вариабельность проявлений этого эффекта. При анализе групповых данных, позволяющем экстраполировать выявляемые эффекты на ограниченной выборки испытуемых на всю популяцию, такая вариабельность проявляется как зависимость статистически значимых эффектов от размера выборки.

Следует отметить, что проиллюстрированная проблема варьирования эффекта весьма значима, поскольку в электрофизиологических исследования не так часто выборки испытуемых превышают 100-150 человек, а изучаемые физиологические и психологические состояния гораздо более тоньше, чем разница между состояниями открытых и закрытых глаз ().

Прежде всего каковы причины вариабельности: 1) при относительно малых выборках (20-30-40 человек) снижается мощность статистических критериев, возникают отклонения от нормальности распределения эффектов по группе испытуемых, что также вносит вклад в «стабильность» статистических тестов, 2) в силу множественности сравнений неизбежно увеличивается количество ложно-положительных значений, 3) мы используем стат. тесты, которые не являются оценкой вероятности нахождения эффекта, хотя зачастую таковыми воспринимаются.

Исходя из вышесказанного, возникает методическая проблема оценки надежности выявляемых, на относительно малочисленных объемах выборок, эффектов. Данная работа ставит перед собой цель оценки эффективности доступных на текущий методов и подходов для выявления эффектов, основанных на анализе когерентности и применения лучшего подхода для анализа различий в синхронизации состояний при открывании глаз. А именно, используя базу данных записей ЭЭГ открытые – закрытые глаза, собираемую на протяжении длительного периода времени и насчитывающую 177 человек, получены оценки воспроизводимости значимых различий между состояниями открытые и закрытые глаза в непересекающихся подгруппах разного размера при использования наиболее популярных подходов корректировки на множественные сравнения, в том числе многомерный непараметрический подход; для каждого из подходов, проведено сравнение воспроизводимости обнаруженных эффектов в полной подгруппе по сравнению с малыми подвыборками разного размера; а также предложен подход ранжирования обнаруженных эффектов, основанного на построении бутстрап интервала размера эффекта.

Работа организована следующим образом: в первой части обсуждаются основные вопросы связанные с выбором состояния глаз при проведении исследований и основные различия этих состояний и проблема воспроизводимости в нейровизуализационных исследованиях; во второй части приводится методика анализа, описываются критерии воспроизводимости и метрики используемые в работе, в третьей части приводятся результаты полученные для базы данных открытые и закрытые глаза.

1. **ЛИТ ОБЗОР, АКТУАЛЬНОСТЬ ИССЛЕДОВАНИЯ**

**Про открытые закрытые глаза**

В основе любого исследования, связанного с исследованиями головного мозга является сравнение характеристик состояния испытуемого во время какого-либо действия с базовым состоянием. Таким базовым состоянием является состояние покоя, характеризующееся отсутствием какого-либо стимула во время фиксации активности головного мозга. Можно выделить два подобных состояния: с открытыми (rsEO) и закрытыми глазами (rsEC); при этом, будучи "максимально близкими по поведению и субъективному восприятию, состояния объективно отличаются паттерном активации нейронов головного мозга (Danko, 2006). Интерпретация результатов изменений в паттернах активации осложняется из-за того, что "когнитивное содержание независимых от стимулов мыслей нелегко связать с объективно измеримыми показаниями" (Snyder & Raichle, 2012) Широкий спектр использования rsEEG, а также сложности в его интерпретации, приводят к важному вопросу воспроизводимости результатов такого исследования, поскольку их неустойчивость на протяжении n-ного количества экспериментов снижает статистическую мощность (способность обнаружить действительные эффекты) (Duan et al., 2021)

Анализ различий состояний с открытыми и закрытыми глазами имеет как методическую, так и практическую ценность. С методической точки зрения, эти два состояния характеризуются целым спектром различий, отличающихся по масштабу, и способность какой-то конкретной методики обнаружить данные различия могут быть использованы для ее верификации. С практической точки зрения анализ различий состояний покоя может помочь в понимании множества аспектов, связанных с человеческим восприятием, когнитивной функцией и даже патологическими состояниями.

*?Так, переход от состояния открытых глаз к закрытым глазам (и наоборот) может помочь понять, как изменяется восприятие окружающего мира и информации при изменении входных условий. Сравнение состояний глаз может помочь в изучении сна и бодрствования, а также изучения когнитивных процессов, таких как внимание, память и решение задач, психических процессов. Изменения в активности мозга и психофизиологические реакции могут различаться в зависимости от того, открыты глаза у человека или закрыты.?*

**Проблема воспроизводимости**

Широкий спектр использования rsEEG, а также сложности в его интерпретации, приводят к важному вопросу о воспроизводимости результатов такого исследования (Duan et al., 2021). Только воспроизводимые в определенных условиях исследования эффекты обеспечивают признаваемый научным сообществом прогресс в научных исследованиях и область психофизиологии здесь не может быть исключением. Однако, получаемые результаты, даже при повторении исследований с другими группами испытуемых в одной лаборатории, показывают далеко не полную воспроизводимость, особенно при работе с малыми группами испытуемых, что достаточно характерно для условий исследований в психофизиологии. Соответственно воспроизводимость получаемых данных не может не быть среди основных проблем научного сообщества психофизиологов, о чем свидетельствуют специализированные выпуски авторитетных научных журналов.

Данная работа посвящена анализу воспроизводимости ЭЭГ исследований использующих анализ различий в синхронизации между парами отведений при повторных измерениях. Особенностью нейровизуализационных исследований, в том числее ЭЭГ - исследований является небольшое количество участников: согласно работам (Chen et al., 2022) (Marek et al., 2022) указываются медианные размеры выборок от 12 до 36. При этом, выводы, к которым приходят в данных исследованиях, заключаются в необходимости увеличить этот размер, а также провести ряд других процедур для повышения воспроизводимости результатов. Относительно исследований, где размер выборки может исчисляться тысячами участников (например, в социологических опросах), небольшой размер групп участников нейровизуализационных экспериментов объясняется, как сложностью в наборе испытуемых, так как исследования могут идти много часов, так и стоимостью проведения эксперимента. Кроме того, чем сложнее процедура, тем больше вероятность каких-то накладок, часть данных может стать непригодной из-за помех различной природы, возникающих при записи. Размер выборки напрямую связан с достоверностью и надёжностью полученных результатов так как недостаточное количество участников эксперимента приводит к низкой мощности исследования (Vozzi et al., 2021). Несмотря на то, что в последние годы появляются крупные нейровизуализационные исследования, где размеры выборок доходят до сотен тысяч (например, Human Connectome Project или UK Biobank), однако большинство экспериментов всё же придерживаются более скромных размеров (менее 50 участников) (Smith & Nichols, 2018)

Большое количество данных, полученных во время эксперимента, помогает улучшить пространственное и временное разрешение, вместе с тем создавая проблему множественных сравнений, которая приводит к недостоверным результатам. В работе (Bennett et al., 2009) авторы указывают на множественное тестирование как на источник большого количества ложноположительных результатов, подчёркивая необходимость найти баланс между ошибками первого и второго рода: «если наши критерии слишком консервативные, то мы не сможем обнаружить значимые результаты (= исследованию будет не хватать мощности). Если наши пороги слишком либеральны, наши результаты будут загрязнены избытком ложноположительных результатов». Авторы рассматривают наиболее популярные корректировки на множественные сравнения: FWER и FDR и делают вывод, что именно проблема множественных сравнений приводит к низкой мощности исследования и преувеличению значимости эффектов либо к их полному игнорированию, а также к невозможности воспроизвести полученные результаты (Bennett et al., 2009; Cremers et al., 2017; Maris et al., 2007; Meyer et al., 2021; Puoliväli et al., 2020)

В работе Cremers et al. на основе анализа фМРТ данных проведён анализ воспроизводимости слабых, но пространственного распространённых, эффектов, и сильных, но локализованных, эффектов. Так как авторами было показано, что слабые эффекты воспроизводились в меньшей степени, на основании этого предложено несколько способов для увеличения мощности и, следовательно, повышения воспроизводимости. Первым они называют увеличение размера выборки, которое приводит к увеличению мощности благодаря центральной предельной теореме (ЦПТ). Авторы также предлагают использовать неклассические пороги, позволяющие найти баланс между ошибкой первого и второго рода, учитывая специфические характеристики исследования. В работе так же упоминается мощностный анализ, который направлен на определение минимального размера выборки в зависимости от желаемого уровня значимости, размера эффекта и мощности. Также предлагается проведение мета-анализа как части преэкспериментального анализа, для выявления неких паттернов активации отдельных частей сетей головного мозга, которые появляются среди большого количества исследований (Cremers et al., 2017).

**Основные подходы коррекции на множественные сравнения**

Наиболее часто используемые подходы FWER и FDR. FWER -вероятность совершения хотя бы одной ошибки первого рода при множественном тестировании гипотез, а FDR - ожидаемая доля правдивых нулевых гипотез, которые были отвергнуты. FWER равное 0.05 будет означать, что существует 5%-ная вероятность обнаружения одного или больше ложноположительных результатов среди всех тестов; либо ограничение значения FDR, при котором установка значения в 0.05 означает, что в среднем 5% полученных результатов являются ложноположительными. Контроль FDR — это более слабая мера контроля наличия ложноположительных результатов, так как он не стремится полностью исключить такие результаты, а скорее дать оценку их присутствию и контролировать их “устойчивость”, то есть присутствие среди результатов. При таком более слабом контроле, однако, удаётся повысить мощность исследования и обнаружить значимые эффекты, однако эти значения должны быть включены в итоговый обзор результатов работы для демонстрации того, насколько исследование подвержено ошибкам первого рода. Оба этих подхода основаны на коррекции значимостей индивидуальных тестов на сравнение, то есть используют только набор p\_value полученных для всех тестируемых эффектов, при этом они никак не учитывают структуру данных, неоднородности в распределении параметров. Кроме того, с помощью данных тестов можно провести коррекцию на количество проведенных сравнений, а не на количество доступной информации в тестах (Cohen, 2014).

Для решения выше-обозначенных проблем достаточно широко используются непараметрические подходы. Главное преимущество последних заключается в том, что они не подразумевают, что данные должны подчиняться определённому закону распределения. Более того, статистики, которые получаются в процессе зависят исключительно от структуры данных. Среди непараметрических подходов различаются две основные техники: основанные на перемешивании и на бустрапе. При этом, в отличии от именованных параметрических тестов, в каждом конкретном эксперименте можно организовывать процесс уникальным образом, главное определить, как именно организовывать процедуру вычисления статистик и конструировать нулевую гипотезу. Для анализа как ЭЭГ так и фМРТ достаточно популярны непараметрические подходы основанные на оценке размеров кластеров (cluster-based permutation) (Cohen, 2014), а для сравнения матриц синхронизации и корреляций – на оценке размера связных компонент (network based statistic) (Zalesky et al., 2010). Эти подходы основаны на перемешивании для построении нулевой гипотезы и используются для данных с очень высоким разрешением таких как карты частотно-временного анализа, матрицы функциональной связности большого размера, то есть когда значимые различия не могут быть изолированы и обязательно должны наблюдаться в целой области. Недостатком этих подходов является их неспособность определить изолированные эффекты, которые в случае анализа изменений когерентности при использовании системы EEG10-20 уже могут наблюдаться. Для таких случаев предложено многомерное непараметрическое тестирование (Galán et al., 1997; Herrera-Díaz et al., 2016; Pesarin, 1992).

В случае с использованием непараметрических подходов остается вопрос о мощности критерия и контролированной доле ошибок второго рода. При этом, в данном случае нет стандартизированных процедур для получения таких оценок. При этом стоит отметить, что и для параметрических тестов оценки мощности поправок и контролируемой доле ложноотрицательных гипотез чаще всего используются методы, верифицированные для случая, когда все гипотезы независимы. При этом данное предположение существенно нарушается при использовании их для анализа синхронизации между парами отведений в ЭЭГ данных для набора частотных диапазонов. Данные теряют независимость как при переходе от одного частотного диапазона к другому, так и в пространстве: электроды меряют сигналы на поверхности, поступающие от источников внутри головного мозга, информация распределяется по поверхности и в разные отведения может поступать импульс от одной и той же зоны, кроме того, области головного мозга физически связаны между собой, поэтому нельзя считать каждую связь между отведениями отдельной независимой единицей. В связи с этим необходима верификация оценок наиболее часто встречающихся подходов к коррекции на множественные сравнения.

Кроме того, полученных в результате статического анализа (любым способом, параметрическим или нет) значений p-value оказывается недостаточным для того, чтобы судить о их значимости. Основным недостатком p-value является его зависимость от размера выборки, кроме того известно что, очень небольшие эффекты по размеру могут иметь очень низкие значения p-value (Cohen, 2014), а вопрос практической значимости остается за границами статистического анализа. Отсюда следует, что кроме p-value необходимо публиковать также другие релевантные статистик, одной из которых является размер эффекта. При этом, для использования выбранной статистики в качестве критерия силы эффекта, также необходимо иметь информацию о неопределенности статистики или иметь оценку доверительного интервала для нее.

1. **МЕТОДЫ И РЕЗУЛЬТАТЫ**

**Данные**

Данные представляют собой ЭЭГ записи состояния покоя агрегированные в течение разных физиологических экспериментов, проведённых на базе Института мозга человека им Н.П. Бехтеровой. Всего база данных насчитывает 177 человек в возрасте от 17 до 30 лет . Для каждого испытуемого записывались ЭЭГ сигналы в состоянии покоя с открытыми и закрытыми глазами. Участники сидели в затемнённой комнате в комфортном кресле. Точка для фиксации взгляда (при rsEO) была расположена на мониторе, расположенном на расстояния 1.2 метра. Испытуемым было запрещено впадать в состояния дремоты; если такое происходило (отмечалось на ЭЭГ), участнику вновь проводили инструктаж и возобновляли эксперимент. Если такое происходило во второй раз, то данные ЭЭГ, полученные для этого человека, не использовались. Было использовано 19 электродов, соответствующих международной системе расположения электродов "10-20\%". После удаления артефактов был проведён количественный анализ: для каждого испытуемого в заданном состоянии (rsEO или rsEC) рассчитывали средние значения функции когерентности по каждой паре отведений. Эти оценки производились для спектральных компонентов, усреднённых по частотным диапазонам (7 диапазонов общей частотой от 1,5 Гц до 40 Гц). Матрицы оценок функции когерентности (фазовой синхронизации) были нормализованы. Далее проводился статистический анализ этих матриц синхронизации для обнаружение значимых различий между параметрами ЭЭГ в сравниваемых состояниях.

**Коррекция на множественные сравнения**

Добавить про результаты Maris

В данной работе мы предлагаем адаптированную методику многомерного непараметрического анализа для поиска упорядоченных по важности значимых различий между параметрами синхронизации ЭЭГ, полученных в различных состояниях. Данная методика основана на следующих работах (Herrera-Díaz et al., 2016; Maris et al., 2007; Pesarin, 1992), включает в себя небольших модификации критерия значимости и дополнена ранжированием найденных эффектов по значимости с использованием бустрап доверительного интервала для размера эффекта. Теоретическое обоснование подхода приведено в работе (Pesarin, 1992), применение к анализу ERP кривых можно найти у (Galán et al., 1997), а к сравнению изменений ЭЭГ корреляции между группой и патологией в работе (Herrera-Díaz et al., 2016). Приведем экспериментальную схему теста из последней работы.

Для каждого испытуемого получено **p** пар электродов и **f** частотных диапазонов для матрицы функциональной связности. При общей гипотезе **H0** о равенстве средних значений между обеими группами t-статистика была разложена на две маргинальные (частные) гипотезы: (a) максимум над набором частотных полос для каждой пары электродов. (b) максимум по совокупности пар электродов в определенной полосе частот. В обоих случаях использование распределения t-макс статистики для каждой краевой гипотезы позволяет найти пороги значимости с поправкой на множественные сравнения по тому измерению, в котором был взят максимум.

Перестановочные тесты состояли из следующих шагов:

1. Случайная перестановка наблюдений значений функциональной связности (когерентностью) между группами 10 000 раз. В каждом повторе вычисляются обе статистики t-max по каждой размерности для частных гипотез
2. Оценка эмпирического нулевого распределения для статистик, вычисленных на предыдущем шаге
3. Установка уровней значимости и вычисление значения p для статистики t-max эмпирического нулевого распределения
4. Отклонение нулевой гипотезы (контроль ошибки типа I) для тех статистик t-max исходной выборки, которые выше порогов значимости (т.е. значение p = 0,05)

Схематично процедура теста изображена на Рисунке 1.

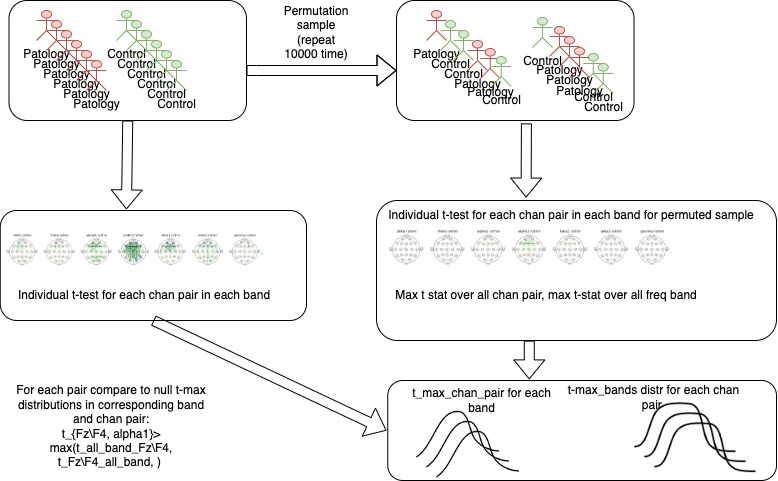


Рис. 1. Схема двумерного непараметрического перестановочного теста для выявления значимых связей матрицы функциональной ЭЭГ связности предложенной в работе (Herrera-Díaz et al., 2016)

В данной работе стояла задача не межгруппового сравнения параметров синхронизации, а сравнение при переходе из состояния открытых в состояние закрытых глаз, то есть сравнение повторных измерений. В связи с этим был модифицирован первый шаг для получения нулевого распределения в отсутствии эффекта. Для каждого испытуемого мы случайным образом изменяли состояние открытых и закрытых глаз, то есть в новой перестановочной выборке для каких-то испытуемых состояния были маркированы верно, а для других изменены. Далее также вычислялись частных распределения максимальных статистик для каждой пары отведений вдоль всех частотных полос и для всех отведений внутри одной частотной полосы. Таким образом для каждой пары отведений были получены два критических уровня значимости вдоль каждой из размерности, на основе комбинации этих уровней принимается решение отвергнуть или принять гипотезу. В оригинальной работе предлагалось использовать в качестве критерия максимальное значение статистики, то есть вдоль каждой размерности вычисляется как доля значение превышающих эмпирическое значение в нулевом распределении, а затем в качестве итогового берется максимальное. То есть, даже если вдоль одной размерности показатель оказывается очень маленьким, а вдоль другой выше заданного уровня (0.05) то гипотеза не будет отвергнута, и эффект будет рассматриваться как не значимый. Мы использовали его при тестировании подходов, однако дополнили еще взвешенной комбинацией значений. Размерность пар отведений и частотных полос сильно отличаются, так в нашей конфигурации 178 пар отведений и 7 частотных полос, кроме того, ожидается что пространственная корреляция должна быть больше частотной. В связи с этим мы использовали взвешенное среднее двух критериев, в котором больший вес применялся к статистике, полученной вдоль всех пар отведении, а именно

Данный подход сравнивался с наиболее часто используемыми параметрическими методами, а именно:

- FWER коррекция Бонферрони и Холма-Бонферрони

- FDR коррекция Бенджамини-Хохберга и Бенджамини- Иекутиели (Benjamini-Hochberg, Benjamini-Yekutiel)

Также использовались неоткорректированные пороги в демонстрационных целях, в качестве базы для сравнения с остальными подходами. Для каждой пары отведений проводился параметрический t-test с повторным измерениями или что тоже самое t-test на равенство разницы между двумя состояниями нулю, затем полученные значения p-value корректировались одним из вышеописанных подходов. Для непараметрический подхода была разработана реализация на языке Python, с использованием библиотек spipy, numpy код размещен в свободный доступ: <https://github.com/iknyazeva/EEGAnalysis>. Для остальных видов коррекции на множественные сравнения была использована открытая библиотека statsmodels v0.14.0 для языка Python.

FWER-корректировки предусматривают более строгий контроль, поскольку они направлены на снижение общей вероятности совершить хотя бы одну ошибку первого рода (Austin et al., 2014). FDR-корректировки допускают большую долю ложных эффектов и направлены на снижение их пропорции относительно верно обнаруженных эффектов.

С помощью вышеописанных подходов тестировалась нулевая гипотеза и обнаруживались значимые эффекты. Далее естественным образом возникает вопрос ранжирования эффектов по значимости. В данной работе в качестве критерия была выбрана величина размера эффекта. Она является универсальной характеристикой «масштаба» обнаруженного эффекта и позволяет оперировать не только статистической значимостью, но и «практической» ценностью. Для решения данной задачи можно использовать для каждой пары отведений точечного размера эффекта, вычисляемого просто как модуль отношения средней разницы в когерентности к стандартному отклонению, однако для более устойчивого результата предлагается использовать не точечную оценку, а среднюю из доверительного интервала. Получить такую оценку можно с помощью техники бутсрап. Идея использования бустрап интервалов различных характеристик используется довольно часто (Martin, 2007). Для построения распределения размеров эффекта использовалась техника бутстрап с повторением, где для каждого эксперимента симулировалось 1000 экспериментов со случайным добавлением или удалением испытуемых. Далее, из полученного распределения размера эффекта вычислялся доверительный 95 процентный интервал и среднее значение, которое и использовалось для ранжирования. Нижнюю границу доверительного интервала можно использовать также как дополнительный критерий значимости эффекта.

**Статистические критерии оценки воспроизводимости**

При построении выводов о наличии или отсутствии эффектов важной задачей является оценка уровня неопределенности и устойчивости полученных результатов. В одномерном случае, например задаче сравнения средних при повторных измерениях есть устоявшиеся критерии качества, это ошибки первого и второго рода, оценки для таких ошибок известны для каждого теста. В случае с многомерными данными ситуация усложняется, так уже упоминались подходы FWER иFDR коррекции, кроме того, можно вводить дополнительные критерии воспроизводимости.

В том случае, когда есть достаточно большое количество наблюдений, например, как в нашем случае, можно провести серию экспериментов с формированием подвыборок меньшего размера из общей выборки. Для анализа воспроизводимости можно ставить следующие вопросы:

1. Какова вероятность при последующих экспериментах получить те-же самые результаты, то есть насколько устойчивы результаты при повторных измерениях
2. Какова чувствительность и специфичность используемого метода: насколько полно и насколько точно метод способен выявить существующие эффекты

Для ответа на первый вопрос, мы предлагаем использовать индивидуальную частоту воспроизводимости значимых эффектов (разница в ЭЭГ когерентности) для каждой пары отведений. Частота воспроизводимости для каждой пары отведений вычисляется как доля эффектов, которые оказались для этой пары значимы, при многократном повторении исследования. Данная характеристика позволяет сделать индивидуальный вывод по каждой паре отведений для того, чтобы сделать вывод о картине в целом предлагается использовать коэффициентСёренсена или Dice коэффициент. Коэффициент Сёренсена (или Dice coefficient, DC) - мера сходства, определенная следующим образом:

где a- размер одной популяции, b - размер другой популяции, c - количество элементов, принадлежащих обоим популяциям в математических терминах c можно выразить как пересечение двух множеств. В контексте данной работы, DC позволяет оценить степень воспроизводимости паттерна путём сравнения сходства двух множеств значимых пар отведений. Чем выше значение DC, тем больше сходство между двумя множествами, то есть тем больше пар отведений сохраняется при повторении эксперимента. Выбор данной оценки обусловлен, во-первых, её простотой и интуитивной интерпретацией, а во-вторых, её широким применением как метрики сходства для различных наборов данных (например, при оценке сегментации и классификации изображений (Guindon & Zhang, 2017)}, а также при конструировании и оценке картирования нейронных сетей с использованием фМРТ и ЭЭГ (Rojas et al., 2018; Sair et al., 2016) DC представляет собой внешнюю оценку, не требующую дополнительных критериев или параметров для своего использования.

Для ответа на второй вопрос можно сравнивать насколько хорошо воспроизводится паттерн, обнаруженный в полной выборке в подгруппах меньшего размера, это опять же можно оценить с помощью Dice коэффициента. Также можно оценить долю выявленных эффектов в малых подвыборках относительно полной выборки, это будет некой оценкой мощности метода (power), а также долю ошибочно найденных эффектов среди всех обнаруженных, то есть тех, которые были обнаружены методом в подвыборке, но отсутствуют при использовании полной выборки, аналог fdr. Безусловно, ответы на эти вопросы зависят от размера выборки, поэтому для каждого фиксированного размера нужно проводить отдельное исследование.

**Описание экспериментов**

В данной работе, для оценки воспроизводимости результатов с помощью метрик, описанных в предыдущей части была проведено несколько экспериментов. В каждом эксперименте была общая часть: выявление значимых различий в ЭЭГ когерентности для каждой пары отведений и вычисление размера эффекта. Для определения значимости эффекта использовались методы, описанные в главе «Коррекция на множественные сравнения». А именно, для каждой пары отведений в каждом ритме был проведен параметрический тест на равенство среднего значения нулю для разницы состояний Открытые – Закрытые глаза. Далее были использованы FWER и FDR поправки для коррекции полученных p-value на множественные сравнения. Отдельно, для этой же выборки был проведен двумерный непараметрический тест. Далее, с помощью техники бутстрап с повторениями был получен доверительный интервал для размера эффекта для каждой пары отведений и среднее значение полученного распределения использовалось для ранжирования эффектов по всем парам отведений во всех частотных диапазонах.

***Эксперимент 1: устойчивость эффектов.***

Для тестирования воспроизводимости эффектов была реализована следующая процедура: для каждого исследуемого размера выборки, вся выборка случайным образом разбивалась на непересекающиеся группы этого размера. Далее, для каждого рассматриваемого метода для каждой подгруппы определялись значимые эффекты, результат представлялся в виде бинарного вектора, где 1 соответствовала тому, что нулевая гипотеза об отсутствии эффекта отвергалась. Далее для каждой пары подгрупп исследуемого размера вычислялся коэффициент перекрытия или Dice коэффициент, показывающий на сколько паттерн воспроизводится целиком, а также частота воспроизводимости эффекта (число раз когда пара отведений оказывалась значимой) для каждой пары отведений. Затем процедура повторялась: выборка перемешивалась случайным образом, проводилось переразбиение на непересекающиеся подгруппы и снова вычислялись коэффициенты воспроизводимости. С помощью данного эксперимента можно получить ответ на вопрос «Насколько похожи эффекты, найденные в подгруппах одного размера между собой».

***Эксперимент 2: близость результатов к результатам на полной выборке.***

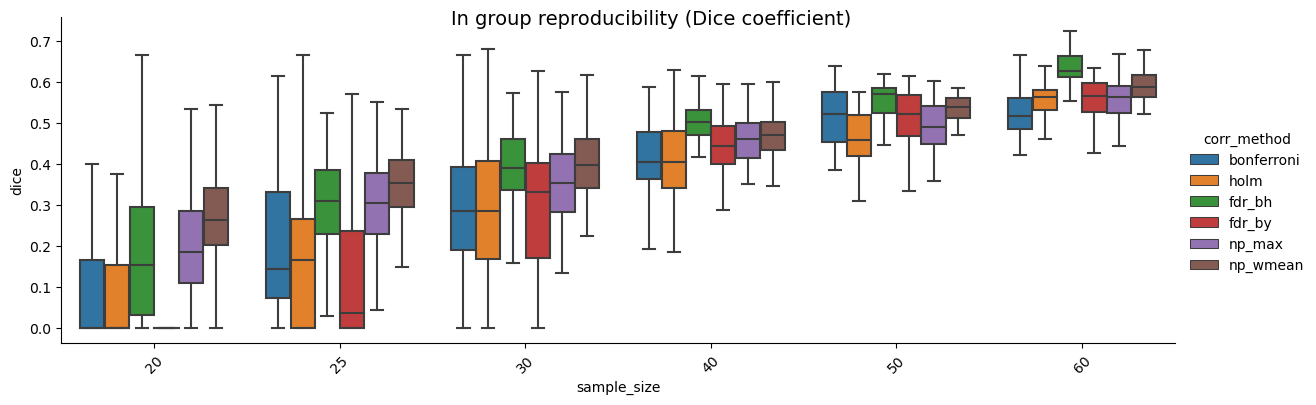
Второй эксперимент заключался в том, что все найденные эффекты во всех подгруппах разного размера сравнивались с эталоном или ground true. В качестве эталона принимались те эффекты, которые для каждого изучаемого метода, получались на полной выборке. Далее, с помощью техники бустрап без повторения, формировались подгруппы меньшего, заранее определенного размера. Для каждой подгруппы вычислялись метрики перекрытия (dice), доли выделенных эффектов (power) и доля ложных эффектов (fdr) относительно эталона.

**Эксперимент 3: анализ частоты воспроизводимости для разных размеров эффектов.**

Частота воспроизводимости вычислялась отдельно для каждой пары отведений, для того, чтобы получить какую-то интегральную характеристику и оценить качество ранжирования с помощью размера эффекта мы использовали среднюю частоту воспроизводимости для следующих диапазонов размеров эффектов : сильный (более 0.8), умеренный (от 0.5 до 0.8) и слабый (0.2 -0.5)

**Результаты**

Первый эксперимент: устойчивость эффектов или воспроизводимость результатов между непересекающимися подгруппами. На малых подгруппах лучше всего себя показывают непараметрических подход (np\_max, np\_wmean) и Бенджамини-Хохберга (fdr\_bh). Начиная с 40 результаты практически идентичны. При этом абсолютное значение коэффициента около 0.4 на малых подвыборках и приближается к 0.6 при размерах выборки 60. То есть даже при достаточно больших размерах выборки результаты воспроизводимости далеки от 1 (идентичный паттерн).



**Рисунок 2.** Воспроизводимость полного паттерна эффектов при повторных экспериментах

Второй эксперимент: близость результатов к результатам на полной выборке. С точки зрения доли выявленных ответов в малых подгруппах относительно полной выборки для малых размеров выборки (20, 30) лучше всего показывает себя непараметрический подход с использованием усреднения (np\_wmean), а при увеличении размера выборки результаты оказываются лучше у Бенджамини-Хохберга (fdr\_bh) (Рисунок 3а). Но эту характеристику надо рассматривать одновременно с оценкой доли ложно обнаруженных эффектов, и для Бенджамини-Хохберга это значение существенно превышает как заявленные уровень 0.05, так и уровень ощибок, которые показывают другие методы. Таким образом, если брать в совокупности эти характеристики, то опять лучше других показывают себя непараметрические подходы. При этом, для самого маленького размера выборки 20 уровень ошибочных гипотез кажется большим, но на самом деле это вызвано тем, что число всех обнаруженных эффектов не превышает 50 и каждая ложно обнаруженная связь вносит существенный вклад в ошибку, так уровню fdr 0.1 будет соответствовать 5 ложных связей, при общем количестве эффектов равных 50. При этом уже для размера выборки 30, в случае использования Бенджамини-Хохберга число обнаруженных эффектов превышает 100, а fdr превышает 5%. Дополнительно информация о числе обнаруженных эффектов представлена на Рис.3с

|  |
| --- |
| a |
| b |
| c |

**Рисунок 3.** Статистика эффектов обнаруженных в малых подвыборках относительно полной выборке при использовании различных методов коррекции на множественные сравнения. (а) Доля эффектов обнаруженных в малых подвыборках, которые сохранились в полной выборке (b) Доля эффектов, обнаруженных в малых подвыборках, которые не сохранились в полной выборке (с) Абсолютное число эффектов обнаруженных в малых подгруппах

Мы уже проанализировали воспроизводимость паттерна целиком, независимо от уровня важности эффекта, дополнительный вопрос, который ставился в этом исследовании, заключался в том, чтобы оценить вероятность воспроизводства значимого эффекта в зависимости от важности, в качестве которого использовалось среднее значение распределения размера эффекта, полученное с помощью бустрап подстановок. В результате первого эксперимента для каждой пары отведений вычислялась частота воспроизводимости эффекта. Затем, все пары отведений во всех частотных диапазонах были упорядочены в соответствии с размером эффекта, далее были выбраны три пороговых диапазона: сильный (размер эффекта больше 0.8), средний (0.5-0.8) и слабый (уровень меньше 0.5) и частота воспроизводимости усреднялась в каждом диапазоне. Опять же оказалось, что лучшим подходом, который обеспечивает высокий уровень воспроизводимости даже для малых выборок оказался непараметрический с функцией усреднения более 0.8 по сравнению с 0.6 у Бенджамини-Хохберга и не превышающий 0.2 у остальных) . Меньшая разница наблюдалась для среднего размера эффекта и еще меньшая для маленьких эффектов. Тем не менее во всех экспериментах с точки зрения частоты воспроизводимости непараметрические подходы дают лучший результат.

|  |
| --- |
|  |
|  |
|  |

Рисунок 4. Средняя частота воспроизводимости эффектов усредненная внутри трех диапазонов размера эффекта: сильный, средний и слабый

Проекция эффектов полученных на полной выборке непараметрическим методом с усреднением на схему электродов изображена на Рисунке 5, отдельно изображены сильные и средние и слабые эффекты.

|  |
| --- |
|  |
|  |

Рисунок 5. Проекция значимых эффектов полученных методом np\_wmean  на полной выборке.

Наверное надо как-то прокомментировать находки

Что показано:

1. Использование непараметрического многомерного перестановочного теста приводит к лучшим результатам по всем критериям воспроизводимости
2. Использование fwer методов коррекции приводит к тому, что на малых выборках даже для сильных эффектов невозможно отвергнуть нулевую гипотезу
3. FDR коррекция Бенджамини-Хохберга дает большое количество ложных эффектов и не гарантирует 5% уровень ошибок.
4. Оценка воспроизводимости эффекта различий когерентности между состояниями открытых и закрытых глаз для конкретной пары отведений показала, что даже для малых выборок сильные эффекты (с точки зрения размера эффекта) воспроизводятся достаточно уверенно, если использовать непараметрический перестановочный подход предложенный в работе
5. Установлены диапазоны и группы пар отведений, в которых изменения когерентности в условиях эффекта Бергера наиболее воспроизводимы – альфа 1-2 и бета-1 диапазоны. Наиболее вариабельны показатели когерентности в гамма диапазоне.

Полученные данные могут быть использованы в качестве опорных для дальнейших исследований по анализу когерентности в состоянии оперативного покоя – где можно ожидать эффект (например, влияния лекарственных средств, индивидуальных характеристик и т.д.), а также для работ методической направленности по развитию методов оценки когерентности (например, оценивать вклад тех или иных методических инноваций

**Пункты дискуссии.**

**Выводы по пунктам дискуссии.**

Austin, S. R., Dialsingh, I., & Altman, N. S. (2014). Multiple Hypothesis Testing: A Review. *J Indian Soc Agric Stat*, *68*(2).

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience*, *4*(4), 417–422. https://doi.org/10.1093/scan/nsp053

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., Taylor, P. A., & Haller, S. P. (2022). Hyperbolic trade-off: The importance of balancing trial and subject sample sizes in neuroimaging. *NeuroImage*, *247*, 118786. https://doi.org/10.1016/j.neuroimage.2021.118786

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press. https://doi.org/10.7551/mitpress/9609.001.0001

Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLOS ONE*, *12*(11), e0184923. https://doi.org/10.1371/journal.pone.0184923

Danko, S. G. (2006). The reflection of different aspects of brain activation in the electroencephalogram: Quantitative electroencephalography of the states of rest with the eyes open and closed. *Human Physiology*, *32*(4), 377–388. https://doi.org/10.1134/S0362119706040013

Duan, W., Chen, X., Wang, Y.-J., Zhao, W., Yuan, H., & Lei, X. (2021). Reproducibility of power spectrum, functional connectivity and network construction in resting-state EEG. *Journal of Neuroscience Methods*, *348*, 108985. https://doi.org/10.1016/j.jneumeth.2020.108985

Galán, L., Biscay, R., Rodríguez, J. L., Pérez-Abalo, M. C., & Rodríguez, R. (1997). Testing topographic differences between event related brain potentials by using non-parametric combinations of permutation tests. *Electroencephalography and Clinical Neurophysiology*, *102*(3), 240–247. https://doi.org/10.1016/S0013-4694(96)95155-3

Guindon, B., & Zhang, Y. (2017). Application of the Dice Coefficient to Accuracy Assessment of Object-Based Image Classification. *Canadian Journal of Remote Sensing*, *43*(1), 48–61. https://doi.org/10.1080/07038992.2017.1259557

Herrera-Díaz, A., Mendoza-Quiñones, R., Melie-Garcia, L., Martínez-Montes, E., Sanabria-Diaz, G., Romero-Quintana, Y., Salazar-Guerra, I., Carballoso-Acosta, M., & Caballero-Moreno, A. (2016). Functional Connectivity and Quantitative EEG in Women with Alcohol Use Disorders: A Resting-State Study. *Brain Topography*, *29*(3), 368–381. https://doi.org/10.1007/s10548-015-0467-x

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

Maris, E., Schoffelen, J.-M., & Fries, P. (2007). Nonparametric statistical testing of coherence differences. *Journal of Neuroscience Methods*, *163*(1), 161–175. https://doi.org/10.1016/j.jneumeth.2007.02.011

Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. *Computational Statistics & Data Analysis*, *51*(12), 6321–6342. https://doi.org/10.1016/j.csda.2007.01.020

Meyer, M., Lamers, D., Kayhan, E., Hunnius, S., & Oostenveld, R. (2021). Enhancing reproducibility in developmental EEG research: BIDS, cluster-based permutation tests, and effect sizes. *Developmental Cognitive Neuroscience*, *52*, 101036. https://doi.org/10.1016/j.dcn.2021.101036

Pesarin, F. (1992). A resampling procedure for nonparametric combination of several dependent tests. *Journal of the Italian Statistical Society*, *1*(1), 87–101. https://doi.org/10.1007/BF02589052

Puoliväli, T., Palva, S., & Palva, J. M. (2020). Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and Python-based software. *Journal of Neuroscience Methods*, *337*, 108654. https://doi.org/10.1016/j.jneumeth.2020.108654

Rojas, G. M., Alvarez, C., Montoya, C. E., de la Iglesia-Vayá, M., Cisternas, J. E., & Gálvez, M. (2018). Study of Resting-State Functional Connectivity Networks Using EEG Electrodes Position As Seed. *Frontiers in Neuroscience*, *12*. https://www.frontiersin.org/articles/10.3389/fnins.2018.00235

Sair, H. I., Yahyavi-Firouz-Abadi, N., Calhoun, V. D., Airan, R. D., Agarwal, S., Intrapiromkul, J., Choe, A. S., Gujar, S. K., Caffo, B., Lindquist, M. A., & Pillai, J. J. (2016). Presurgical brain mapping of the language network in patients with brain tumors using resting-state fMRI: Comparison with task fMRI. *Human Brain Mapping*, *37*(3), 913–923. https://doi.org/10.1002/hbm.23075

Smith, S. M., & Nichols, T. E. (2018). Statistical Challenges in ‘Big Data’ Human Neuroimaging. *Neuron*, *97*(2), 263–268. https://doi.org/10.1016/j.neuron.2017.12.018

Snyder, A. Z., & Raichle, M. E. (2012). A brief history of the resting state: The Washington University perspective. *NeuroImage*, *62*(2), 902–910. https://doi.org/10.1016/j.neuroimage.2012.01.044

Vozzi, A., Ronca, V., Aricò, P., Borghini, G., Sciaraffa, N., Cherubino, P., Trettel, A., Babiloni, F., & Di Flumeri, G. (2021). The Sample Size Matters: To What Extent the Participant Reduction Affects the Outcomes of a Neuroscientific Research. A Case-Study in Neuromarketing Field. *Sensors*, *21*(18), Article 18. https://doi.org/10.3390/s21186088

Zalesky, A., Fornito, A., & Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage*, *53*(4), 1197–1207. https://doi.org/10.1016/j.neuroimage.2010.06.041