Кризис воспроизводимости в науке: почему большая часть научных исследований ложна?

Александр Пушин, Арсения Шихова, БПМИ142

14 ноября 2016г.

Статья: "Why Most Published Research Findings Are False"
John P. A. Ioannidis

Утверждается что вероятность того, что утверждение верно, может зависеть от:

- Статистической мощности ($1 - \beta$, где β - вероятность ошибки второго рода)
- Априорной вероятности его достоверности (до проведения исследования)
- Наличия систематических ошибок (плохой дизайн исследования)
- Количества других исследований по тому же вопросу
- Явной финансовой заинтересованности или другой предвзятости
- Вовлечённости большего количества исследовательских групп в научной области в погоне за статистической значимостью результата
- И других факторов...

Первая проблема - использвование в качестве метрики принятия гипотезы исключительно p-value

Рассмотрим задачу поиска зависимостей между разными величинами.

- α и β — вероятности ошибок *I* и *II* рода при проверке одной гипотезы. Вероятность обнаружить в ходе исследования реальную связь, отражает мощность $1 - \beta$.

Первая проблема - использвование в качестве метрики принятия гипотезы исключительно p-value

Рассмотрим задачу поиска зависимостей между разными величинами.

- α и β — вероятности ошибок *I* и *II* рода при проверке одной гипотезы. Вероятность обнаружить в ходе исследования реальную связь, отражает мощность $1 - \beta$.
-

	Есть связь	Нет связи
Исследование нашло связь	*TP*	*FP*
Исследование не нашло связи	*FN*	*TN*

Первая проблема - использвование в качестве метрики принятия гипотезы исключительно p-value

Рассмотрим задачу поиска зависимостей между разными величинами.

- α и β — вероятности ошибок *I* и *II* рода при проверке одной гипотезы. Вероятность обнаружить в ходе исследования реальную связь, отражает мощность $1 - \beta$.

	Есть связь	Нет связи
Исследование нашло связь	*TP*	*FP*
Исследование не нашло связи	*FN*	*TN*

- c — количество гипотез, которые мы проверяем

Первая проблема - использвование в качестве метрики принятия гипотезы исключительно p-value

Рассмотрим задачу поиска зависимостей между разными величинами.

- α и β — вероятности ошибок *I* и *II* рода при проверке одной гипотезы. Вероятность обнаружить в ходе исследования реальную связь, отражает мощность $1 - \beta$.

	Есть связь	Нет связи
Исследование нашло связь	*TP*	*FP*
Исследование не нашло связи	*FN*	*TN*

- c — количество гипотез, которые мы проверяем
- $R = \frac{TP+FN}{FP+TN}$ - отношение "реальных связей"к числу "отсутствующих связей"

Первая проблема - использвование в качестве метрики принятия гипотезы исключительно p-value

Рассмотрим задачу поиска зависимостей между разными величинами.

- α и β — вероятности ошибок *I* и *II* рода при проверке одной гипотезы. Вероятность обнаружить в ходе исследования реальную связь, отражает мощность $1 - \beta$.
-

	Есть связь	Нет связи
Исследование нашло связь	*TP*	*FP*
Исследование не нашло связи	*FN*	*TN*

- c — количество гипотез, которые мы проверяем
- $R = \frac{TP+FN}{FP+TN}$ - отношение "реальных связей"к числу "отсутствующих связей"
- $\frac{R}{R+1} = \frac{TP+FN}{c}$ - отношение "реальных связей"ко всем связям (априорная вероятность)

Первая проблема - использвование в качестве метрики принятия гипотезы исключительно p-value

Рассмотрим задачу поиска зависимостей между разными величинами.

- α и β — вероятности ошибок *I* и *II* рода при проверке одной гипотезы. Вероятность обнаружить в ходе исследования реальную связь, отражает мощность $1 - \beta$.
-

	Есть связь	Нет связи
Исследование нашло связь	*TP*	*FP*
Исследование не нашло связи	*FN*	*TN*

- c — количество гипотез, которые мы проверяем
- $R = \frac{TP+FN}{FP+TN}$ - отношение "реальных связей"к числу "отсутствующих связей"
- $\frac{R}{R+1} = \frac{TP+FN}{c}$ - отношение "реальных связей"ко всем связям (априорная вероятность)
- PPV (positive predictive value) $= \frac{TP}{TP+FP}$ — мера качества исследования.

Результаты исследования	Реальная связь		
	Да	Нет	Итог
Да	$c(1-\beta)R/(R+1)$	$c\alpha/(R+1)$	$c(R+\alpha-\beta R)/(R+1)$
Нет	$c\beta R/(R+1)$	$c(1-\alpha)/(R+1)$	$c(1-\alpha+\beta R)/(R+1)$
Итог	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t001

Результаты исследования	Реальная связь		
	Да	Нет	Итог
Да	$c(1-\beta)R/(R+1)$	$c\alpha/(R+1)$	$c(R+\alpha-\beta R)/(R+1)$
Нет	$c\beta R/(R+1)$	$c(1-\alpha)/(R+1)$	$c(1-\alpha+\beta R)/(R+1)$
Итог	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t001

Почему это так (пример для TP):

$$c(1-\beta)\frac{R}{R+1} = c(1-\beta)\frac{TP+FN}{c} = (1-\beta)(TP+FN) = TP$$

Не сложно показать что:

$$PPV = \frac{(1-\beta)R}{R - \beta R + \alpha}$$

В таком случае, вероятность истинности исследования больше чем вероятность того что оно будет ложным если

$$(1-\beta)R > \alpha$$

А т.к. обычно $\alpha = 0.05$, то должно быть выполнено условие

$$(1-\beta)R > 0.05$$

Следующая проблема - смещение

Смещение - сочетание различных факторов, связанных с планом исследования, данными, анализом и представлением результатов, приводящее к выводам, к которым исследователи не должны были приходить.

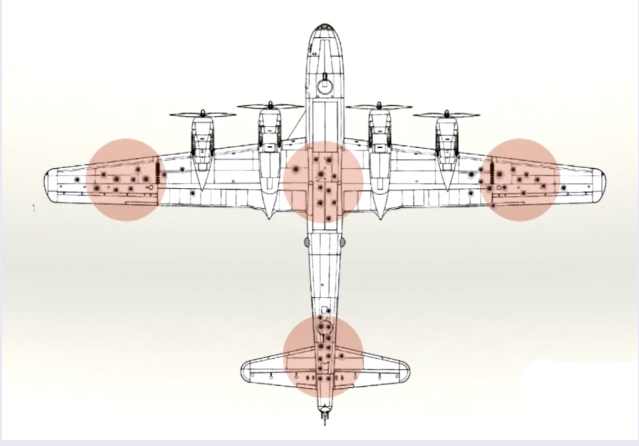
Пример

По данным интернет-голосования 100% людей пользуются интернетом.

Пример


Chicago Daily Tribune
HOME
DEWEY DEFEATS TRUMAN
G.O.P. Sweep Indicated in State; Boyle Leads in City

Пример



Введем $u = \frac{FP}{TP+FP}$ как долю исследованных анализов, которые не должны были стать "результатами исследования". Можно понимать как $1 - precision$.

Введем $u = \frac{FP}{TP+FP}$ как долю исследованных анализов, которые не должны были стать "результатами исследования". Можно понимать как $1 - precision$.

Результаты исследования	Реальная связь: Да	Нет	Итог
Да	$(c[1-\beta]R + uc\beta R)/(R+1)$	$c\alpha + uc(1-\alpha)/(R+1)$	$c(R+\alpha-\beta R+u-u\alpha+u\beta R)/(R+1)$
Нет	$(1-u)c\beta R/(R+1)$	$(1-u)c(1-\alpha)/(R+1)$	$c(1-u)(1-\alpha+\beta R)/(R+1)$
Итог	$cR/(R+1)$	$c/(R+1)$	c

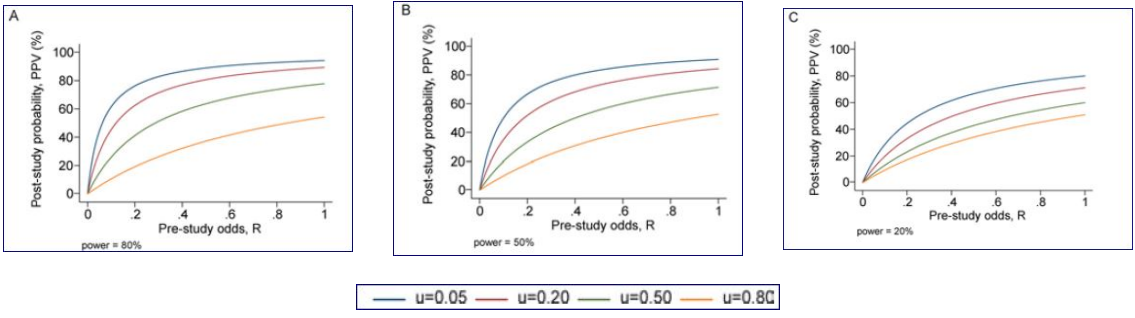
DOI: 10.1371/journal.pmed.0020124.t002

Аналогично предыдущей таблице, получаем:

$$PPV = \frac{(1-\beta)R + u\beta R}{R + \alpha - \beta R + u - u\alpha + u\beta R}$$

PPV существенно уменьшается при увеличении u пока $1-\beta \leq \alpha$, то есть, $1-\beta \leq 0.05$ в большинстве случаев.
Таким образом, с возрастанием смещения шанс, что результат исследования будет верен, существенно снижается. И наоборот, достоверные результаты исследования могут случайно быть аннулированы из-за обратного смещения.

Зависимость *PPV* от априорной вероятности для разной мощности в зависимости от *u*.



Проверка несколькими независимыми группами

n - число независимых групп. Если результаты рассматриваются изолированно, как это зачастую происходит на практике, то:

Результаты исследования	Реальная связь: Да	Нет	Итог
Да	$cR(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n)/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1)$
Нет	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n+R\beta^n)/(R+1)$
Итог	$cR/(R+1)$	$c/(R+1)$	c

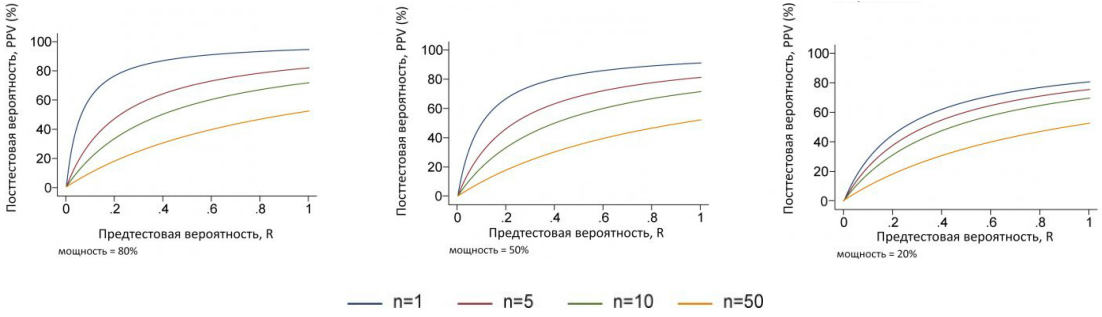
DOI: 10.1371/journal.pmed.0020124.t003

В таком случае (без учета смещений)

$$PPV = \frac{R(1-\beta^n)}{R+1-[1-\alpha]^n-R\beta^n}$$

При увеличении числа независимых исследований PPV снижается до $1-\beta<\alpha$, что в большинстве случаев $1-\beta<0.05$

Зависимость *PPV* от априорной вероятности для разной мощности в зависимости от *n*.



Выводы

- Чем меньше исследования в определённой области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. малая выборка влияет на мощность.

Выводы

- Чем меньше исследования в определённой области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. малая выборка влияет на мощность.
- Чем меньше величина эффекта в исследуемой области, тем меньше вероятность того, что результат достоверен. Мощность исследования также зависит от величины эффекта.

Выводы

- Чем меньше исследования в определённой области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. малая выборка влияет на мощность.
- Чем меньше величина эффекта в исследуемой области, тем меньше вероятность того, что результат достоверен. Мощность исследования также зависит от величины эффекта.
- Чем больше количество и чем меньше отбор тестируемых связей, выявленных в научной области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. это влияет на априорную вероятность.

Выводы

- Чем меньше исследования в определённой области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. малая выборка влияет на мощность.
- Чем меньше величина эффекта в исследуемой области, тем меньше вероятность того, что результат достоверен. Мощность исследования также зависит от величины эффекта.
- Чем больше количество и чем меньше отбор тестируемых связей, выявленных в научной области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. это влияет на априорную вероятность.
- Чем больше гибкость плана проведения исследования и предвзятость, тем меньше вероятность получить достоверный результат т.к. это влияет на смещение.

Выводы

- Чем меньше исследования в определённой области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. малая выборка влияет на мощность.
- Чем меньше величина эффекта в исследуемой области, тем меньше вероятность того, что результат достоверен. Мощность исследования также зависит от величины эффекта.
- Чем больше количество и чем меньше отбор тестируемых связей, выявленных в научной области, тем меньше вероятность того, что результаты исследования будут достоверны, т.к. это влияет на априорную вероятность.
- Чем больше гибкость плана проведения исследования и предвзятость, тем меньше вероятность получить достоверный результат т.к. это влияет на смещение.
- Чем большая активность проявляется в области (чем больше независимых групп исследователей вовлечено), тем меньше вероятность того, что результаты будут достоверны.

Практические примеры исследований и их достоверность

$1-\beta$	R	u	Пример	PPV
0.80	1:1	0.10	Рандомизированное контролируемое исследование (РКИ) адекватной мощности с небольшим смещением и предтестовой вероятностью 1:1	0.85
0.95	2:1	0.30	Проверочный мета-анализ рандомизированных контролируемых исследований высокого качества	0.85
0.80	1:3	0.40	Мета-анализ небольших исследований, не позволяющих сделать окончательный вывод	0.41
0.20	1:5	0.20	Фаза I/II РКИ, не обладающая достаточной мощностью, но качественно проведенная	0.23
0.20	1:5	0.80	Фаза I/II РКИ, не обладающая достаточной мощностью и некачественно проведенная	0.17
0.80	1:10	0.30	Эпидемиологические поисковые исследования, обладающие достаточной мощностью	0.20
0.20	1:10	0.30	Эпидемиологические поисковые исследования, не обладающие достаточной мощностью	0.12
0.20	1:1,000	0.80	Фундаментальные поисковые исслдования с обширным числом тестируемых объектов	0.0010
0.20	1:1,000	0.20	Как в предыдущем примере, но с меньшим смещением (более стандартизированные исследования)	0.0015

Что делать?

- Тщательнее отбирать тестируемые связи. Не создавать тенденцию, когда в области считается нормальным тестировать все подряд.

Что делать?

- Тщательнее отбирать тестируемые связи. Не создавать тенденцию, когда в области считается нормальным тестировать все подряд.
- Составлять четкие планы проведения исследований, не "подстраивающиеся"под результаты.

Что делать?

- Тщательнее отбирать тестируемые связи. Не создавать тенденцию, когда в области считается нормальным тестировать все подряд.
- Составлять четкие планы проведения исследований, не "подстраивающиеся"под результаты.
- Ограничивать влияние заинтересованных лиц на исследование.

Что делать?

- Тщательнее отбирать тестируемые связи. Не создавать тенденцию, когда в области считается нормальным тестировать все подряд.
- Составлять четкие планы проведения исследований, не "подстраивающиеся"под результаты.
- Ограничивать влияние заинтересованных лиц на исследование.
- Держать результаты исследований в общем доступе, чтобы другие люди могли их проверять. Делать группы внутри одной области менее независимыми.

Что делать?

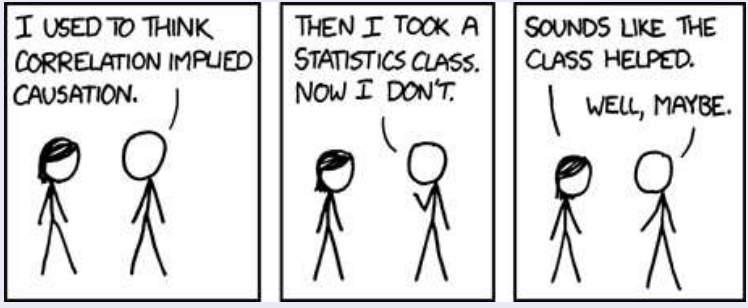
- Тщательнее отбирать тестируемые связи. Не создавать тенденцию, когда в области считается нормальным тестировать все подряд.
- Составлять четкие планы проведения исследований, не "подстраивающиеся"под результаты.
- Ограничивать влияние заинтересованных лиц на исследование.
- Держать результаты исследований в общем доступе, чтобы другие люди могли их проверять. Делать группы внутри одной области менее независимыми.
- Учитывать априорную вероятность при получении результатов исследований.

Что значит «исследование ложно» и насколько все плохо? — Leek, J. T., Jager, L. R. Is most published research really false?

Современные технологии — в первую очередь интернет — позволяют публиковать что угодно

Готовые инструменты для анализа данных позволяют заниматься статистикой непрофессионалам, допускающим элементарные ошибки

Пример


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Свойства исследования:

- Reproducibility - если взять код и данные, выложенные авторами исследования, то получим те же результаты вычислений, таблицы и графики

Свойства исследования:

- Reproducibility - если взять код и данные, выложенные авторами исследования, то получим те же результаты вычислений, таблицы и графики
- Replicability - можно провести аналогичные эксперименты и получить аналогичные результаты

Свойства исследования:

- Reproducibility - если взять код и данные, выложенные авторами исследования, то получим те же результаты вычислений, таблицы и графики
- Replicability - можно провести аналогичные эксперименты и получить аналогичные результаты
- True (false) dicovery - результат исследования является правильным (неправильным) ответом на вопрос, поставленный исследователями

По разным оценкам, от 14 до 80% статей содержат как приложение данные и код, больше всех — в биоинформатике

По разным оценкам, от 14 до 80% статей содержат как приложение данные и код, больше всех — в биоинформатике

Спойлер: некоторые из авторов, подсчитывавших этот процент, сами не публиковали свой код и список исследованных статей

По оценкам разных исследователей, можно повторить 33-45% результатов в психологии

Почти не исследований, оценивающих процент неверных результатов исследований, только рассуждают о причинах

Что нужно улучшать / повышать:

- Инструменты для анализа данных и оформления результатов исследования

Что нужно улучшать / повышать:

- Инструменты для анализа данных и оформления результатов исследования
- Нормы для проведения исследований и публикации результатов

Что нужно улучшать / повышать:

- Инструменты для анализа данных и оформления результатов исследования
- Нормы для проведения исследований и публикации результатов
- Уровень знаний исследователей

Инструменты для оформления результатов:

- knitr/markdown

Инструменты для оформления результатов:

- knitr/markdown
- ipython notebook

Инструменты для оформления результатов:

- knitr/markdown
- ipython notebook
- galaxy (специфичный инструмент для биоинформатики и не-программистов?)

Ресурсы для публикации датасетов:

- Figshare
- Open Science Framework
- Dataverse

Как учить людей проводить научные исследования:

- Организация открытых семинаров / вебинаров

Как учить людей проводить научные исследования:

- Организация открытых семинаров / вебинаров
- MOOC – онлайн-курсы

Как учить людей проводить научные исследования:

- Организация открытых семинаров / вебинаров
- MOOC – онлайн-курсы
- Что ещё?

Спасибо за внимание! Вопросы?