



МОНИКИ

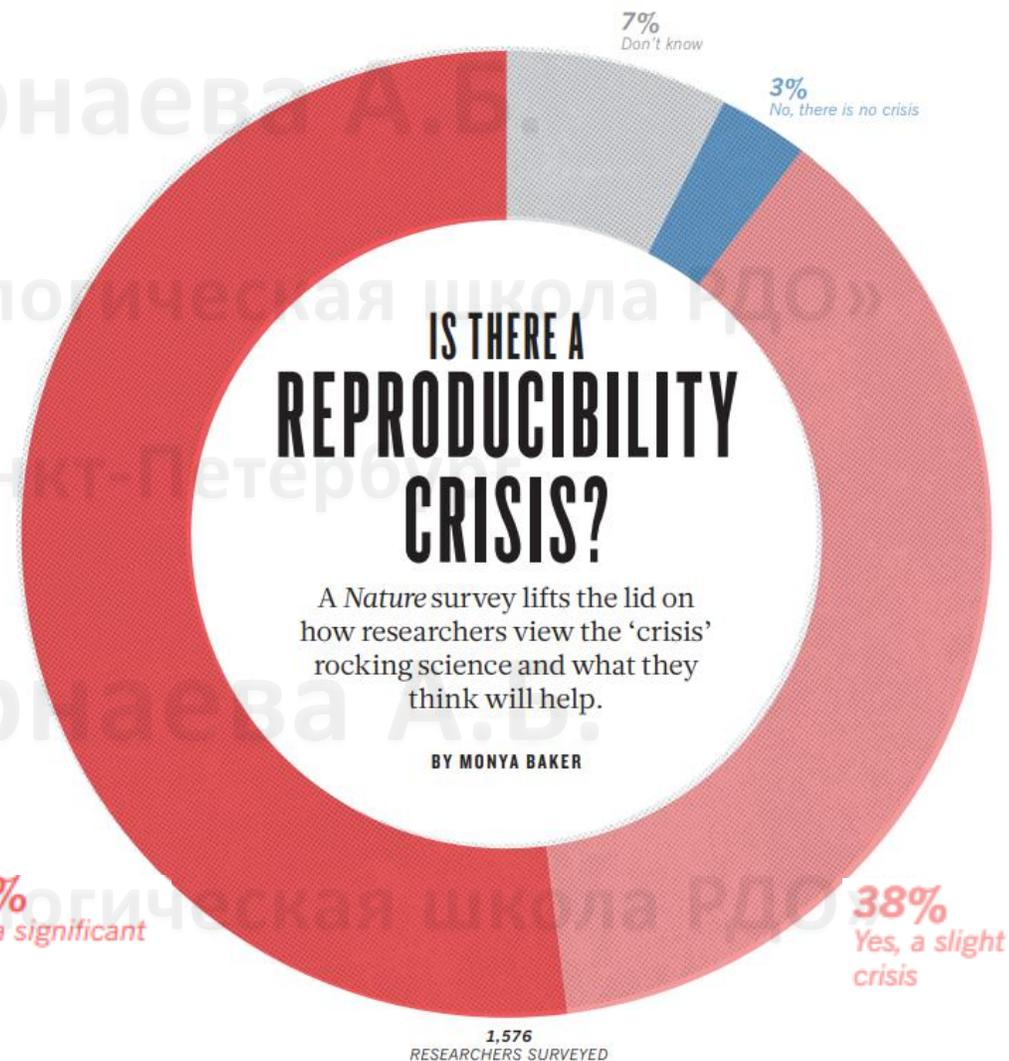
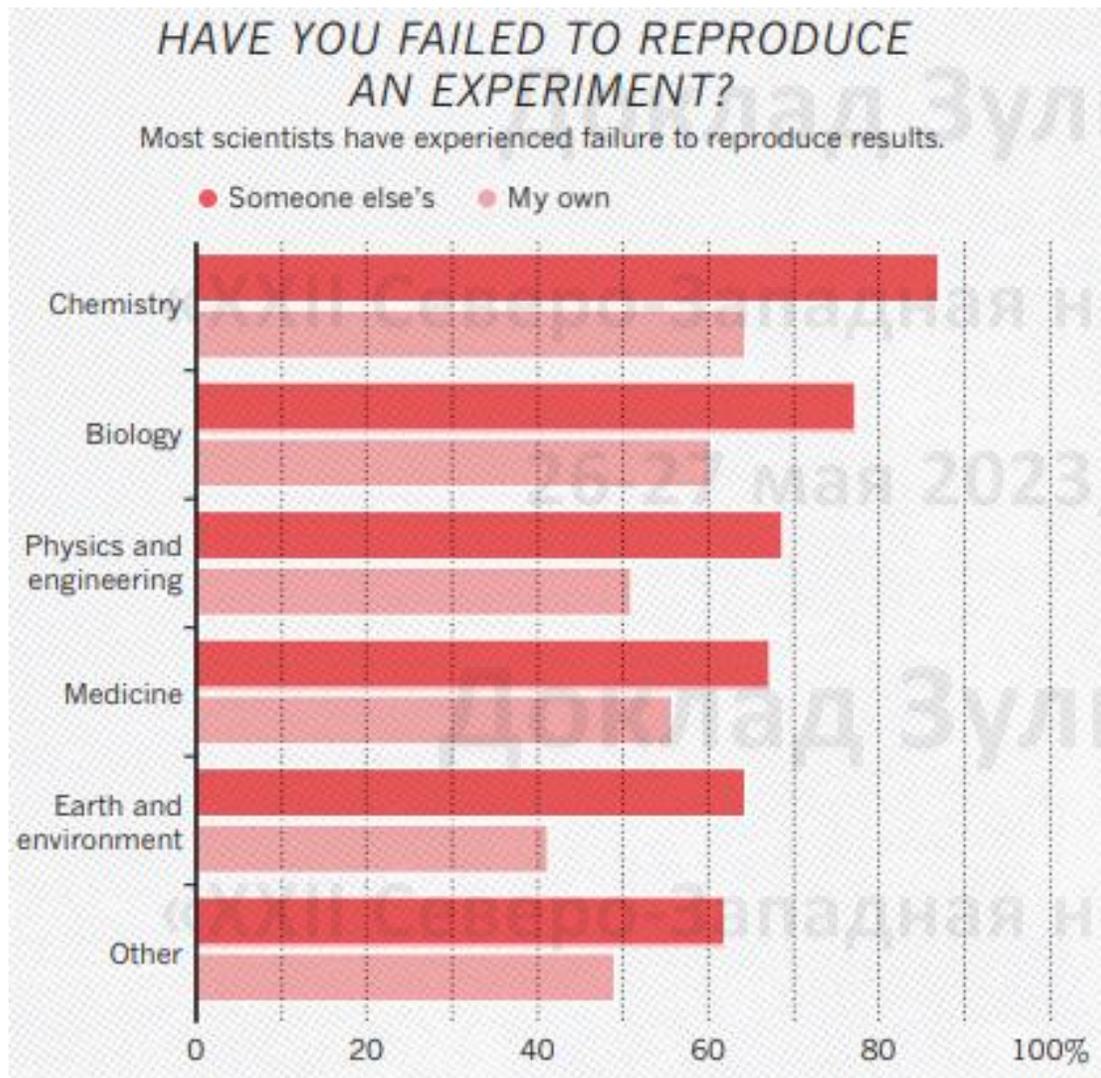
1775

Статистические мифы и опасные заблуждения

Зулькарнаев А. Б.

ГБУЗ МО МОНИКИ им. М.Ф. Владимирского

2023 г.



Доклад Зулькарнаева А.Б.

<https://www.equator-network.org/>

Reporting guidelines for main study types



equator
network

Enhancing the QUALity and
Transparency Of health Research

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

Миф № 1

26-27 мая 2023, г. Санкт-Петербург

*«Я прочитал и понял Стентона Гланца, у меня есть SPSS /
Statistica, я теперь сам себе статист!»*

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Стентон Гланц

Медико-биологическая

СТАТИСТИКА

Перевод с английского
доктора физ.-мат. наук
Ю. А. Данилова
под редакцией
Н. Е. Бузикашвили
и Д. В. Самойлова



п р а к т и к а
Москва 1999

P есть вероятность того, что значение критерия окажется не меньше критического значения при условии справедливости нулевой гипотезы об отсутствии различий между группами.

Определение можно сформулировать и по-другому.

P есть вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий.

Упрощая, можно сказать, что P — это вероятность справедливости нулевой гипотезы. Часто говорят также, что P — это вероятность ошибки. В общем, и это верно, однако несколько неточно. Дело в том, что существует два рода ошибок. Ошибка

Стентон Гланц

Медико-биологическая СТАТИСТИКА

Перевод с английского
доктора физ.-мат. наук
Ю. А. Данилова
под редакцией
Н. Е. Бузикашвили
и Д. В. Самойлова



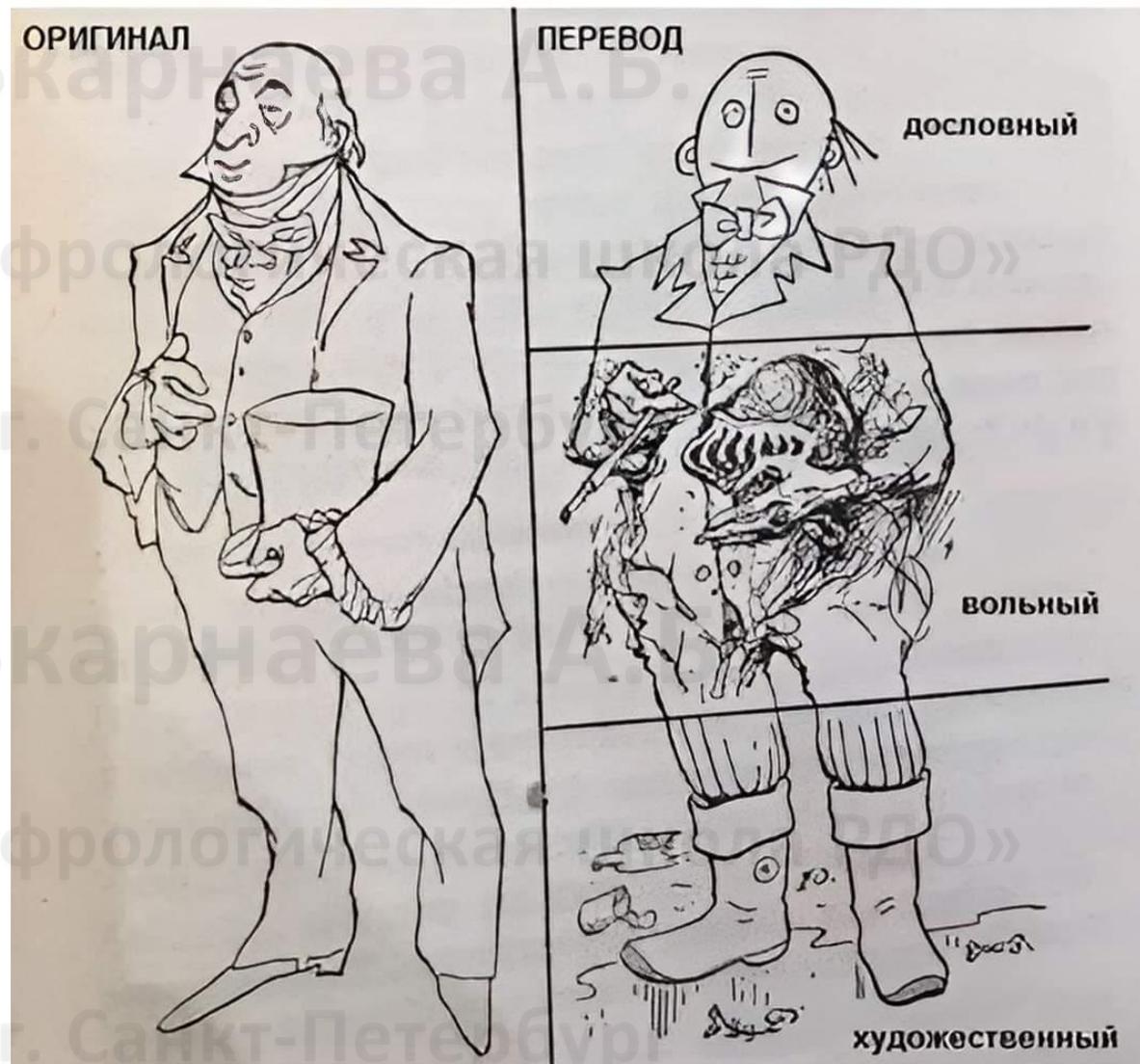
практика
Москва 1999

Р есть вероятность того, что значение критерия окажется не меньше критического значения при условии справедливости нулевой гипотезы об отсутствии различий между группами.

Определение можно сформулировать и по-другому.

Р есть вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий.

Упрощая, можно сказать, что *Р* — это вероятность справедливости нулевой гипотезы. Часто говорят также, что *Р* — это вероятность ошибки. В общем, и это верно, однако несколько неточно. Дело в том, что существует два рода ошибок. Ошибка



Стентон Гланц

Медико-биологическая СТАТИСТИКА

Перевод с английского
доктора физ.-мат. наук
Ю. А. Данилова
под редакцией
Н. Е. Бузикашвили
и Д. В. Самойлова



п р а к т и к а
Москва 1999

P есть вероятность того, что значение критерия окажется не меньше критического значения при условии справедливости нулевой гипотезы об отсутствии различий между группами.

Определение можно сформулировать и по-другому.

P есть вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий.

Упрощая, можно сказать, что *P* — это вероятность справедливости нулевой гипотезы. Часто говорят также, что *P* — это вероятность ошибки. В общем, и это верно, однако несколько неточно. Дело в том, что существует два рода ошибок. Ошибка

Primer of BIOSTATISTIC

FOURTH EDITION

Stanton A. Glantz, Ph.D.

Professor of Medicine
Member, Cardiovascular Research Institute
Member, Institute for Health Policy Studies
University of California, San Francisco

STANTON
A.
GLANTZ

PRIMER
OF

BIOSTATISTICS
FOURTH EDITION

The P value is the probability of obtaining a value of the test statistic as large as or larger than the one computed from the data when in reality there is no difference between the different treatments.

Or, in other words,

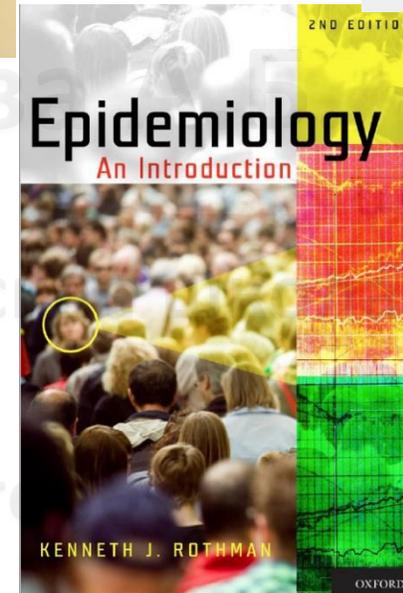
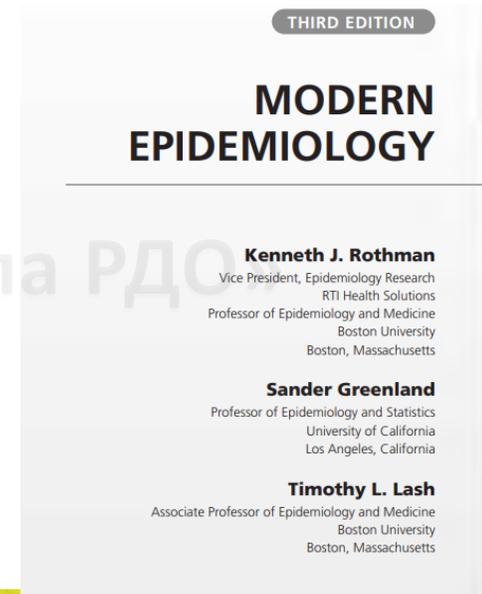
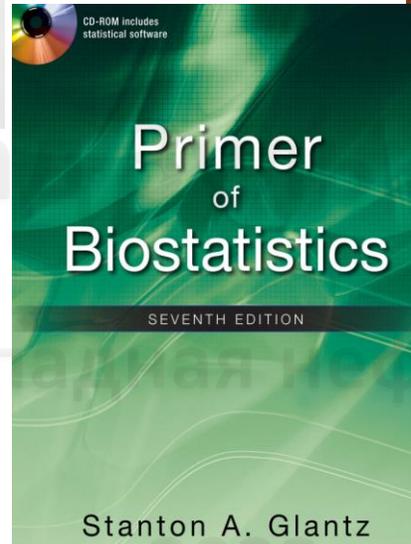
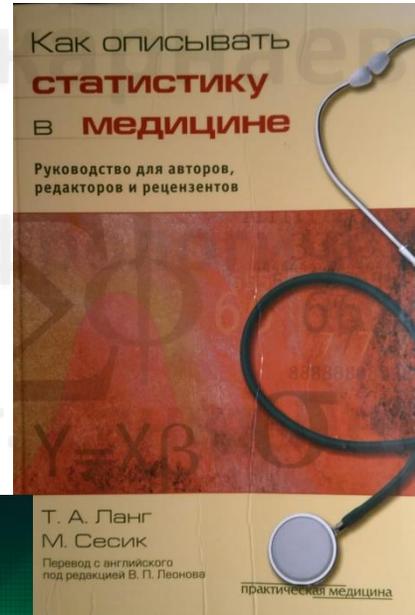
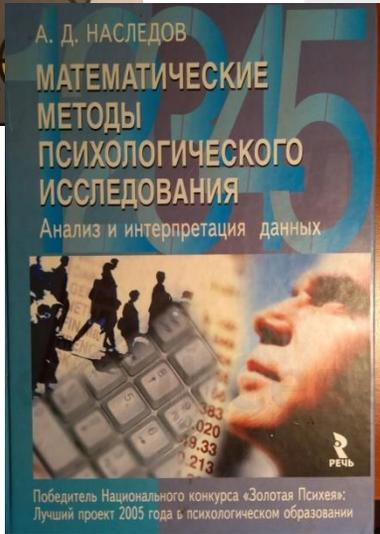
The P value is the probability of being wrong when asserting that a true difference exists.

If we are willing to assert a difference when $P < .05$, we are tacitly agreeing to accept the fact that, over the long run, we expect 1 assertion of a difference in 20 to be wrong.*

It is commonly believed that the *P* value is the probability of making a mistake. There are obviously two ways an investigator can reach a mistaken conclusion based on the data, reporting that the treatment had

Что делать?

- Читайте хорошие книги



Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

Миф № 2

26-27 мая 2023, г. Санкт-Петербург

*«Статистику в диссертации необходимо считать
самостоятельно»*

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

...и насчитать вот это:

Оценка степени причинно-следственной связи осуществлялась следующим

образом:

- нулевая $0 < RR \leq 1$
- малая $1 < RR \leq 1,5$;
- средняя $1,5 < RR \leq 2$;
- высокая $2 < RR \leq 3,2$;
- очень высокая $3,2 < RR \leq 5$;
- почти полная $RR > 5$.

Относительный риск (RR)	0,64 [0,52; 0,95]	0,57 [0,39; 0,87]	0,58 [0,43; 1,12]	0,63 [0,39; 0,92]
Критерий χ^2	0,52	0,49	0,62	0,58
P-уровень	1,25	1,02	1,15	0,95
Связь между состоянием ПСД и показателем общего анализа крови	нулевая	нулевая	нулевая	нулевая

Симптом	Показатели		
	p	ОШ	95% ДИ
Гемоглобин ниже 100 г/л	0,04	4,79	0,56-29,84
Гематокрит	0,04	2,1	1,00-7,84
Гематокрит ниже 37% у мужчин	0,05	0,86	0,49-3,85
Скорость оседания эритроцитов выше 19 мм/ч	0,03	0,59	0,69-3,68
Щелочная фосфатаза	0,05	0,9	0,31-3,57
Альбумин 2 и менее г/л	0,04	0,46	1,05-5,76
Уровень креатинина в сыворотке крови выше 0,13 ммоль/л	0,01	4,16	1,12-12,52
Креатинин в моче выше 0,13 ммоль/сут	0,03	2,38	1,12-10,13
Уровень гемоглобина в моче	0,02	2,78	0,74-7,59
Уровень гемоглобина в моче 30 мг/сут	0,03	1,1	0,4-9,71
Креатинин в моче выше 30 мг/сут	0,003	3,09	1,2-29,28
Уровень гемоглобина в моче 30 мг/сут	0,04	1,38	0,1-7,47
Уровень гемоглобина в моче 30 мг/сут	0,001	2,37	1,48-30,26
Уровень гемоглобина в моче 30 мг/сут	0,03	2,98	1,1-8,15
Уровень гемоглобина в моче 30 мг/сут	0,01	1	0,16-9,64

Доклад Зулъкарнаева А.Б.

УИИ Северо-Западная нефрология «УИИ «Северо-Западная нефрология» РДО»

26-27 мая 2023, г. Санкт-Петербург

26-27 мая 2023, г. Санкт-Петербург

- Читайте положение ВАК
- Там нет ни слова про «читайте сами», а там есть про «пишите сами»:

«XXII Северо-Западная нефрологическая школа РДО»
10. Диссертация должна быть написана автором самостоятельно, обладать внутренним единством, содержать новые научные результаты и положения, выдвигаемые для публичной защиты, и свидетельствовать о личном вкладе автора диссертации в науку.

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

- Статистик сделал это:

```
TENDERA_data_comp1FU <- TENDERA_data %>% filter(comp1_FU == "complete")
TENDERA_data_comp1FU$sex <- factor(TENDERA_data_comp1FU$sex,
  levels = c("M", "F"),
  labels = c("male", "female"))
#set(TENDERA_data_comp1FU)
tbl_summary(TENDERA_data_comp1FU,
  by = access_type,
  missing = "ifany",
  missing_text = "Unknown",
  digits = all_continuous() ~ 1,
  statistic = all_continuous() ~ c("{median} [{p25}; {p75}], {min} to {max}"),
  include = c(age,
    sex,
    BMI,
    brush_circum,
    height,
    cholesterol,
    NSTEMI_unstan,
    art_hypert,
    eGFR_CKD_EPI,
    kidney_function,
    DM,
    smoking,
    type_of_AC_therapy),
  label = list(age ~ "Age, years",
    BMI ~ "BMI, kg/m^2",
    brush_circum ~ "brush_circum, cm",
    height ~ "Height, cm",
    cholesterol ~ "Cholesterol, mmol/l",
    eGFR_CKD_EPI ~ "eGFR (CKD-EPI), ml/min/m^2",
    DM ~ "Diabetes mellitus") %>%
  add_p(pvalue_fun = ~ style_pvalue(.x, digits = 3),
    group = access_type,
    include = everything(),
    test = list(all_continuous() ~ "wilcox.test",
      all_categorical() ~ "fisher.test")) %>%
  bold_p(t = 0.05, q = FALSE) %>%
  as_gt() %>%
  tab_options(row_stripping.include_table_body = TRUE)
```

- Дал вам это:

Characteristic	distal, N = 393 ¹	traditional, N = 402 ¹	p-value ²
Age, years	63.0 [56.0; 69.0], 34.0 to 88.0	64.0 [56.2; 69.0], 30.0 to 88.0	0.714
sex			0.824
male	255 (65%)	264 (66%)	
female	138 (35%)	138 (34%)	
BMI, kg/m ²	28.7 [25.5; 31.8], 18.3 to 47.4	28.5 [26.0; 32.0], 19.6 to 44.5	0.627
brush_circum, cm	19.0 [17.5; 20.7], 14.0 to 24.0	19.0 [17.5; 20.5], 13.0 to 24.6	0.875
Unknown	0	3	
Height, cm	170.0 [164.0; 176.0], 148.0 to 192.0	171.0 [164.0; 178.0], 143.0 to 193.0	0.507
Unknown	1	0	
Cholesterol, mmol/l	4.8 [3.9; 5.8], 2.1 to 10.8	4.7 [3.8; 5.6], 2.3 to 8.0	0.501

Пишите на здоровье!

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

Ошибка № 3

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Все излишне упрощать

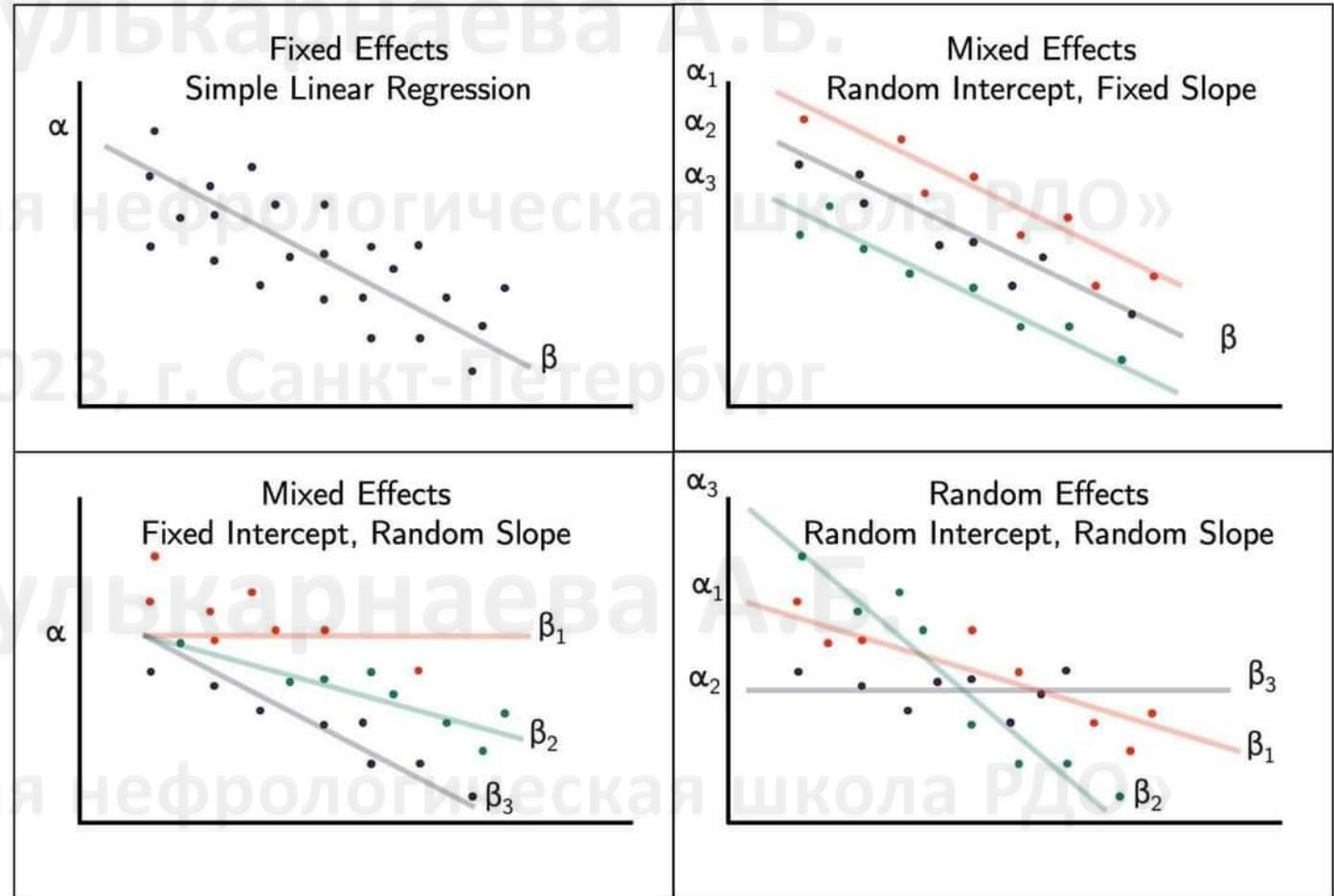
Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург



Если вы н/с и работаете по этой схеме, вы никогда на испытаете эти муки выбора



Почему все не так просто?

- Перед анализом идея трансформируется в «клиническую» гипотезу
- Для ее тестирования формулируется тестируемая «статистическая» гипотеза
- Для тестирования «статистической» гипотезы проводится при условии соблюдения ряда условий (т.н. допущений – «assumptions»)
- Почти все критерии (или может быть даже все) имеют в своей основе определенные предположения.

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Предположения общих линейных моделей (ANOVA, линейная регрессия)

- Линейная зависимость между независимыми переменными и зависимой переменной
- Отсутствие мультиколлинеарности
- Независимость наблюдений
- Гомоскедастичность остатков
- Нормальное распределение остатков
- Отсутствие влиятельных наблюдений

- Читателям

Если в статье используются любые эти ваши регрессии и нет указаний, как проверялись условия их применимости, насторожитесь!

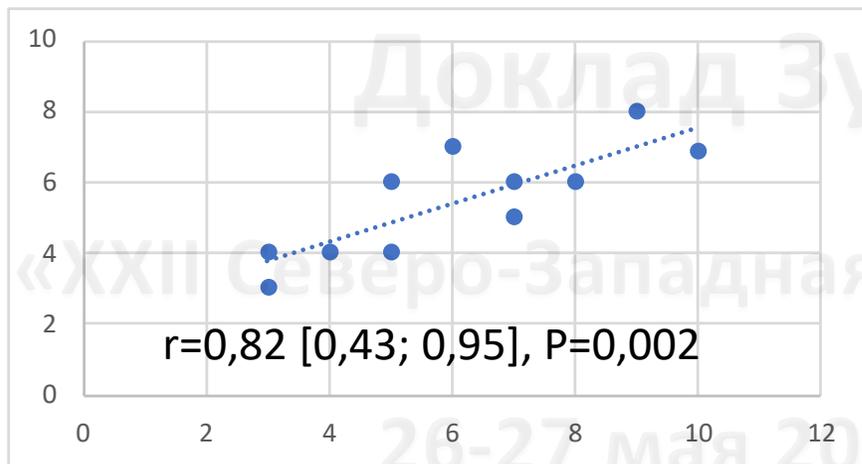
When you ignore linear regression assumptions

- Писателям

Всегда проверяйте условия применимости методов

Хотя бы загуглите «ANOVA assumptions testing SPSS»

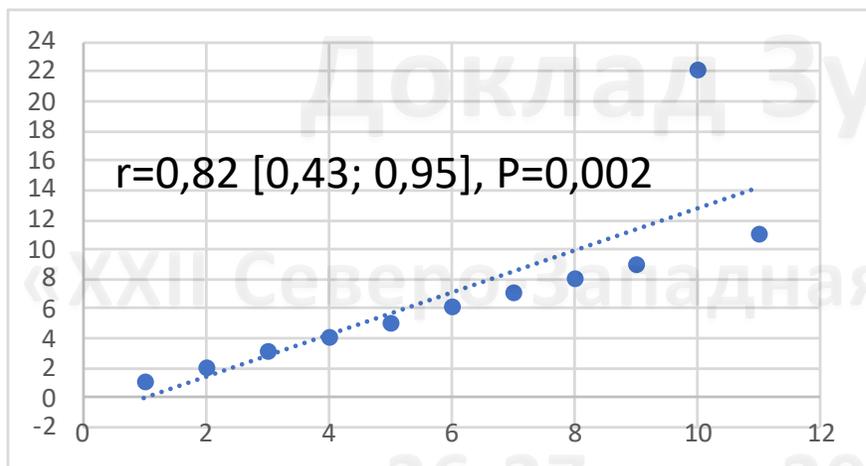
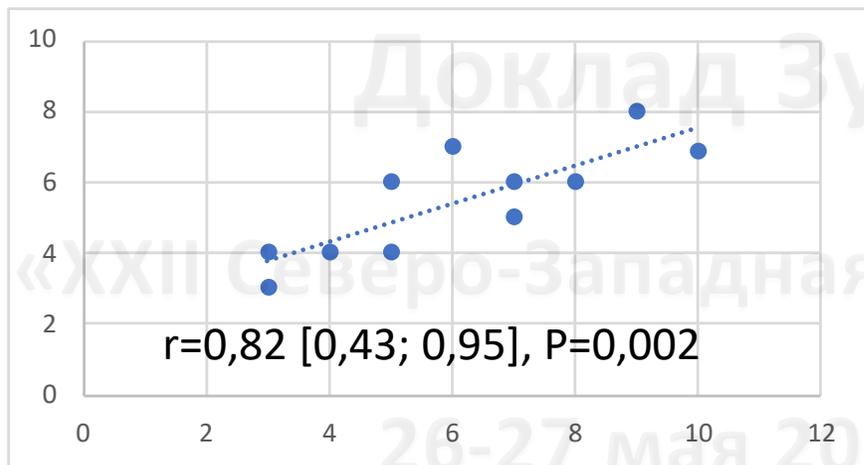


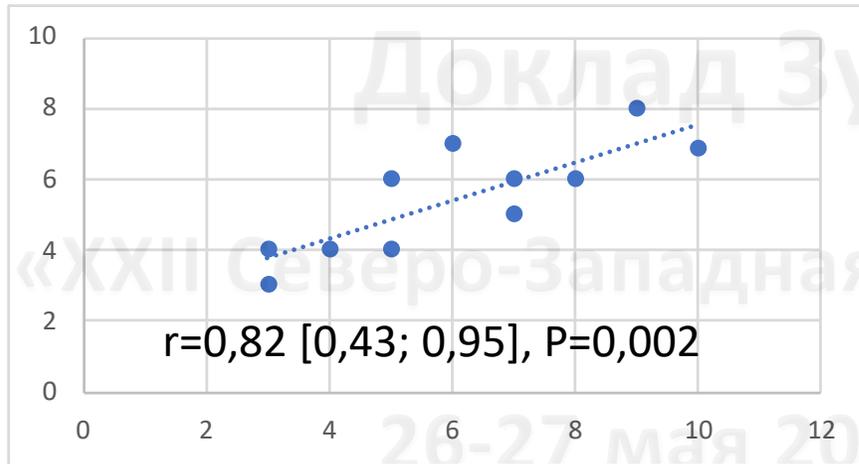


Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

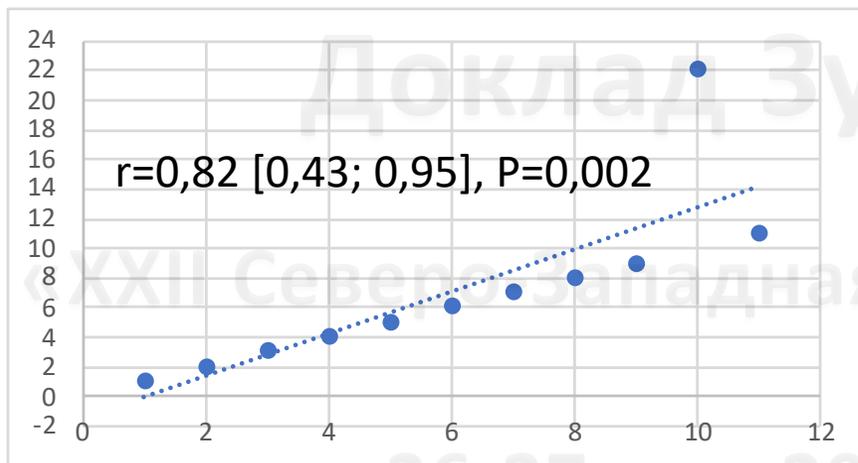
26-27 мая 2023, г. Санкт-Петербург





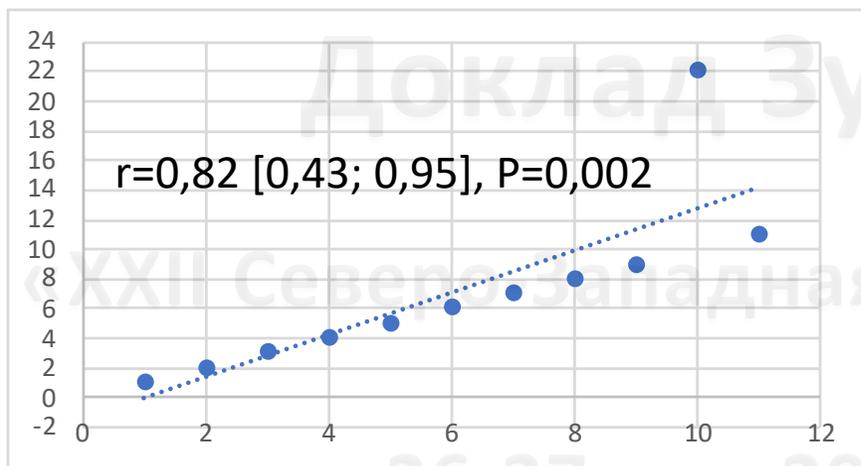
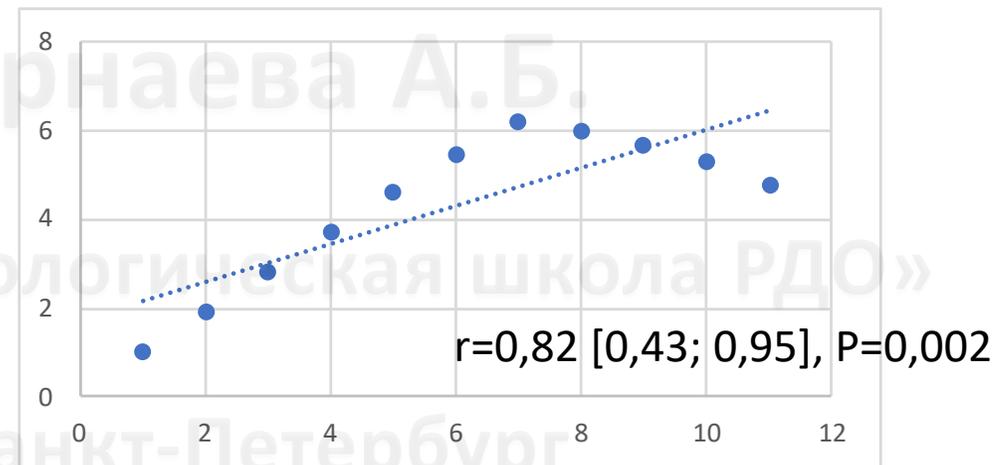
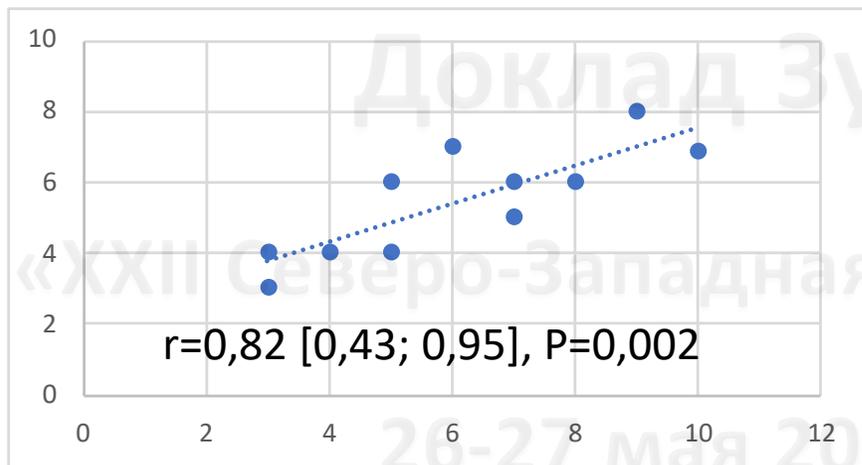
R: Your data has outliers

Me:

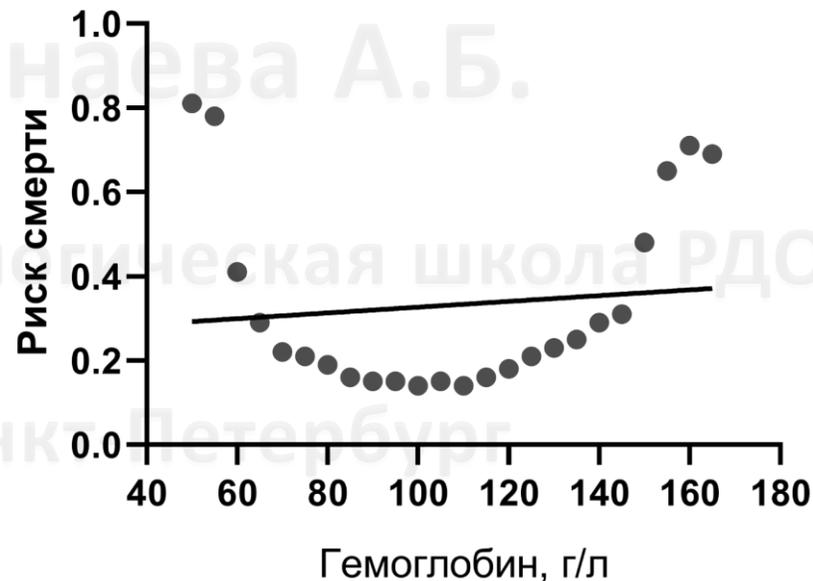
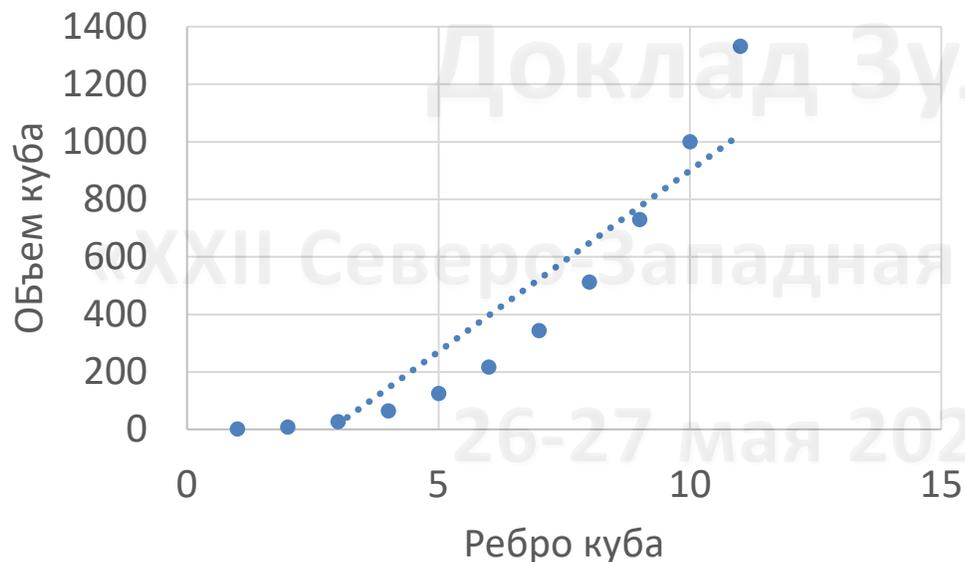


- Не все выбросы – влиятельные наблюдения
- Это повод для более детального анализа
- Часто нельзя просто удалить их
- Нередко удаление выбросов порождает новые выбросы
- Но автор обязан прокомментировать их наличие и описать свою тактику в отношении них

Простой пример



Часто мы имеем дело с нелинейными зависимостями

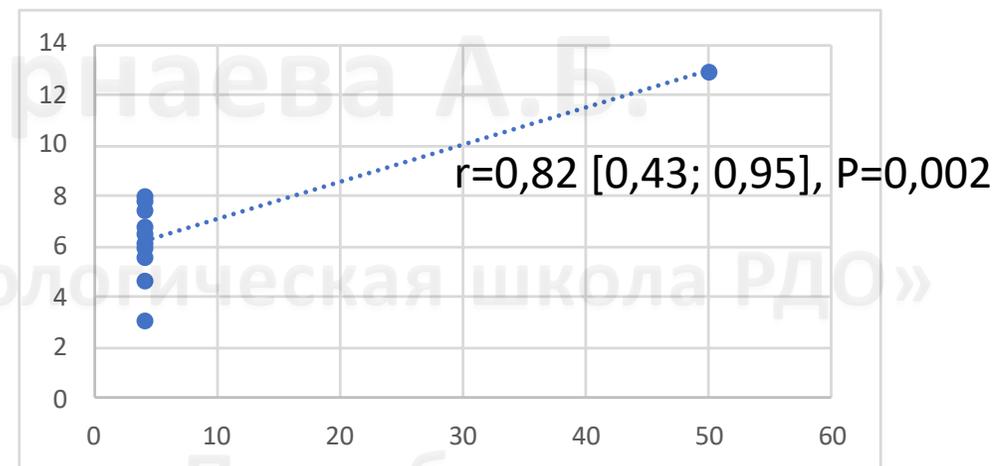
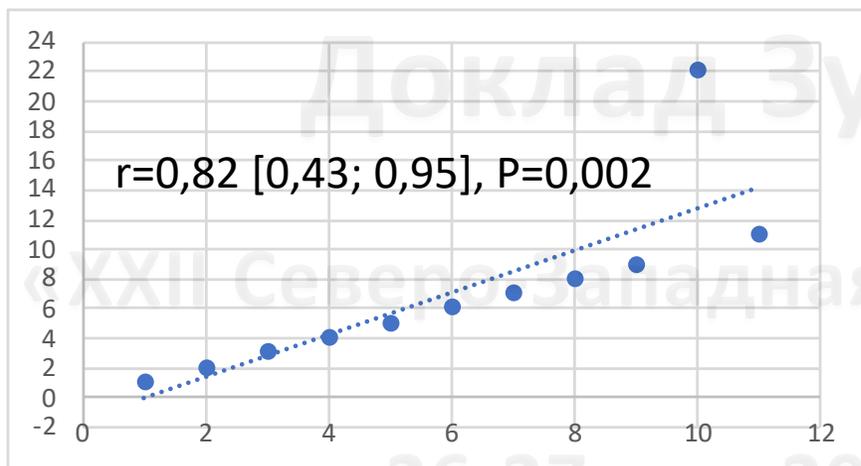
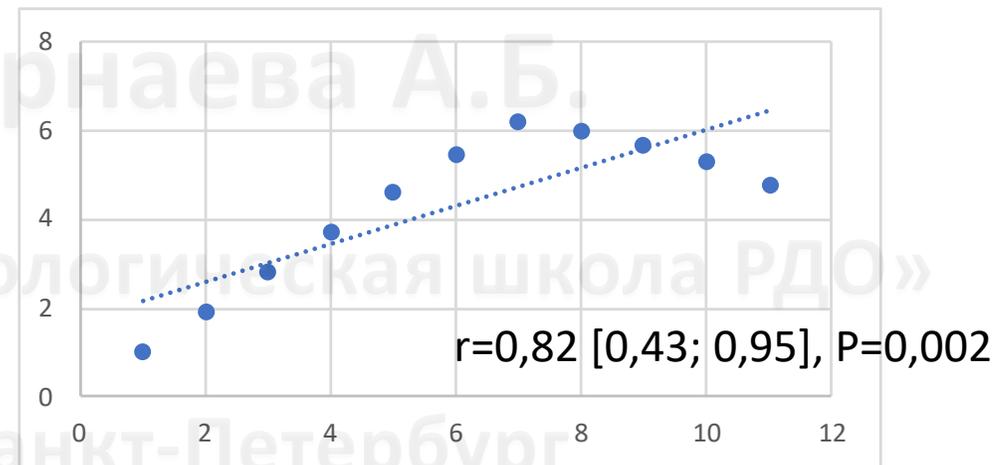
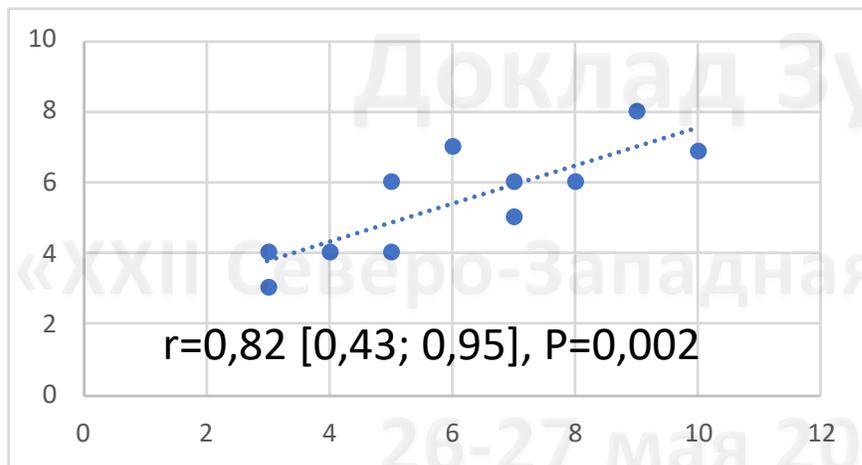


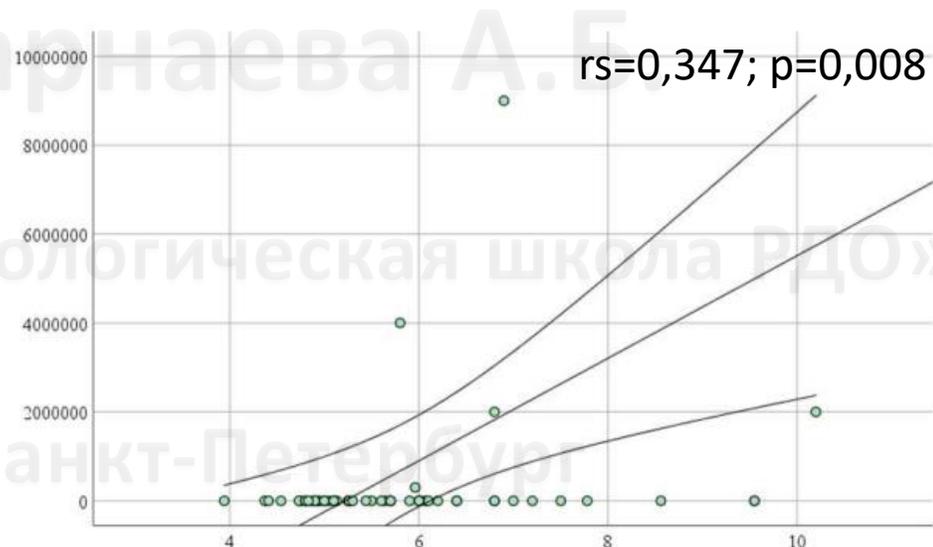
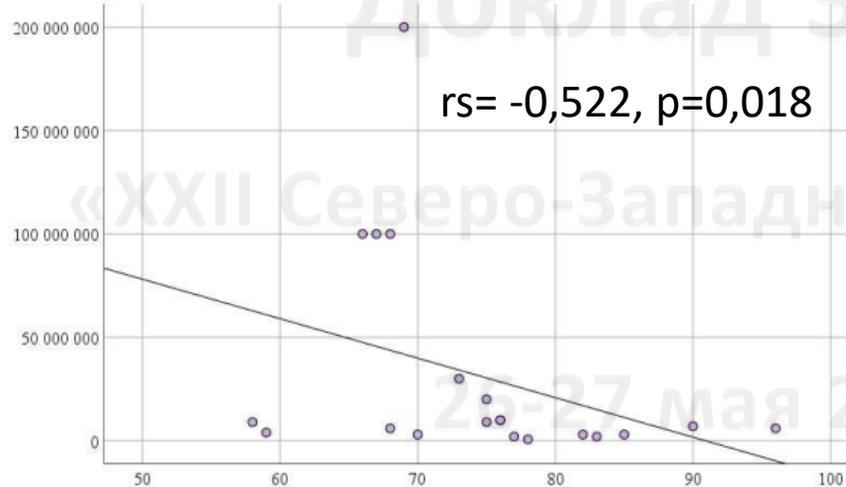
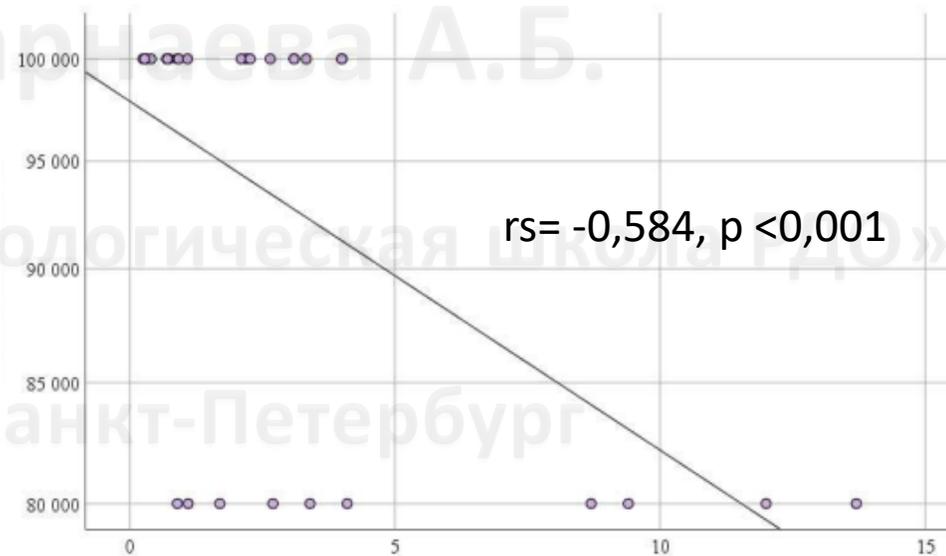
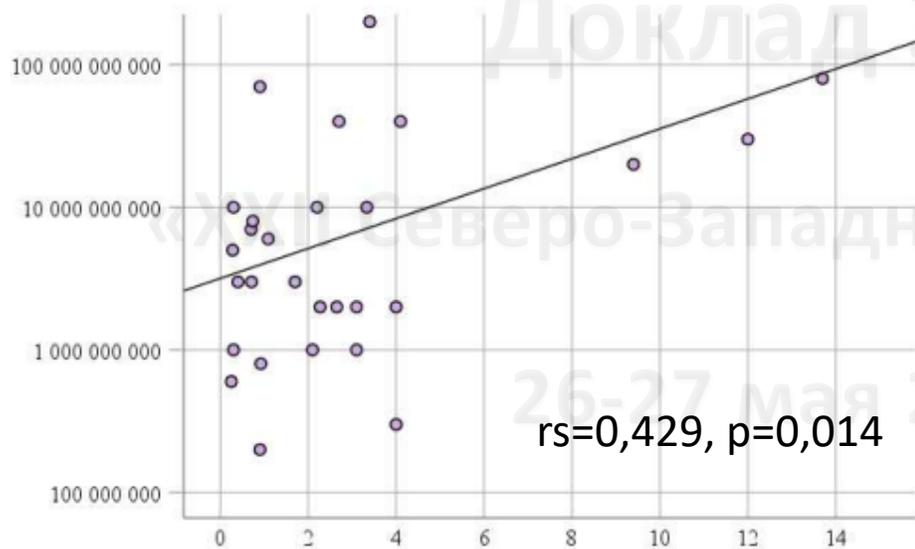
Доклад Зулькарнаева А.Б.

Степенное преобразование

Gender	Creatinine concentration	Formula
Woman	≤ 0.7	$GFR = 144 \times (Cr/0.7)^{-0.329} \times (0.993)^{age}$
	> 0.7	$GFR = 144 \times (Cr/0.7)^{-1.209} \times (0.993)^{age}$
Man	≤ 0.9	$GFR = 141 \times (Cr/0.9)^{-0.411} \times (0.993)^{age}$
	> 0.9	$GFR = 141 \times (Cr/0.9)^{-1.209} \times (0.993)^{age}$

Простой пример





Доклад Зулькарнаева А.Б.

Ошибка № 4

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Оценивать не то, что интересно на самом деле

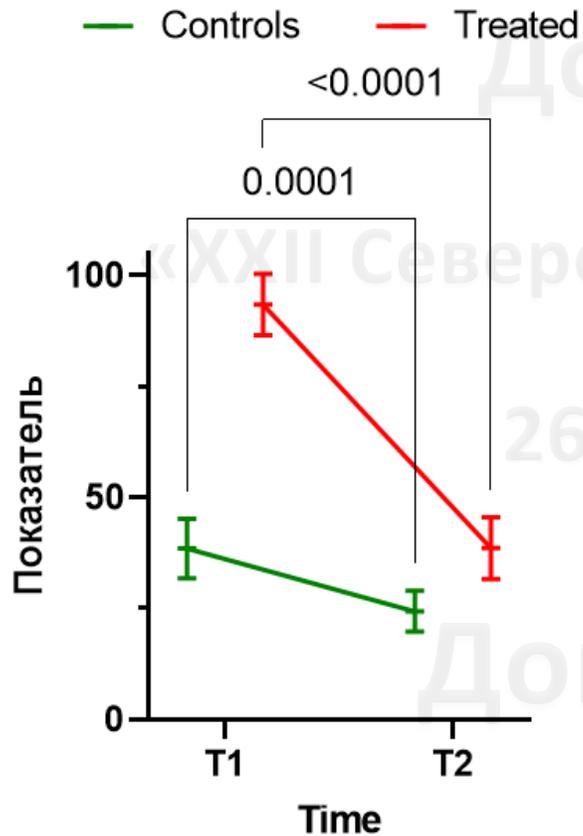
Доклад Зулькарнаева А.Б.

Додумывать за исследователя

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Какой вопрос задает исследователь?



В обеих группах была динамика.
Но это не означает, что динамика
в группах различалась!

Доклад Зулькарнаева А.Б.

«XVII Северо-Западная нефрологическая школа РДО»

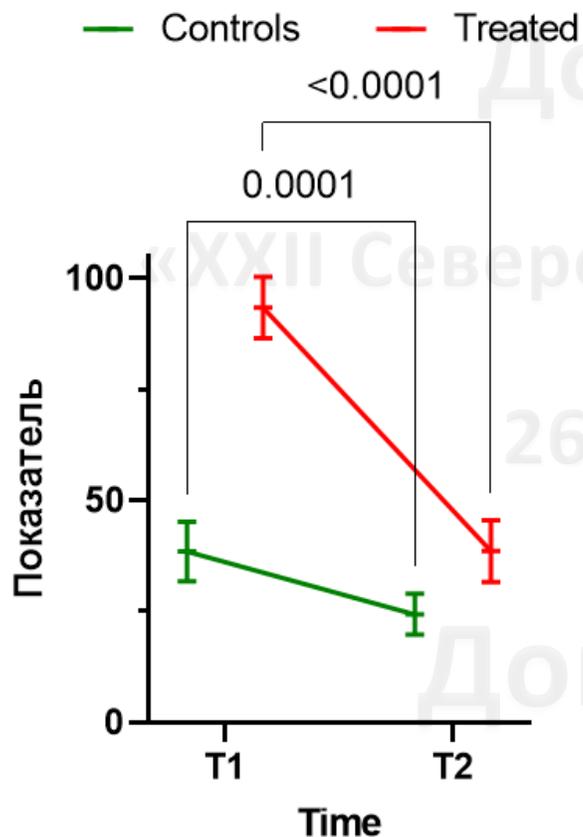
26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

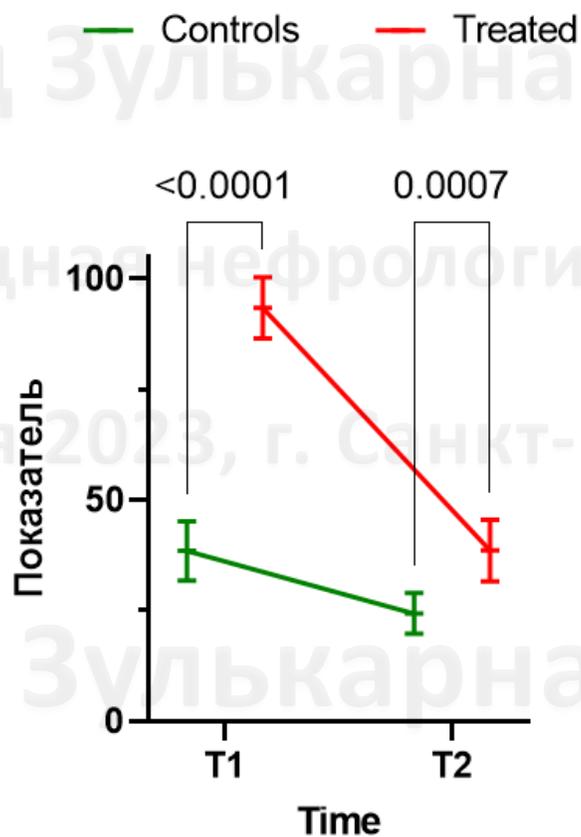
«XVII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Какой вопрос задает исследователь?

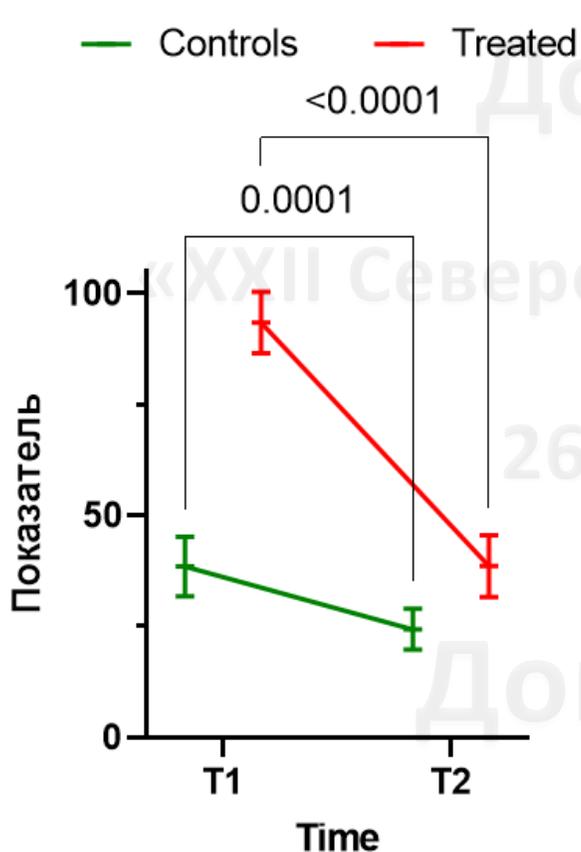


В обеих группах была динамика.
Но это не означает, что динамика
в группах различалась!

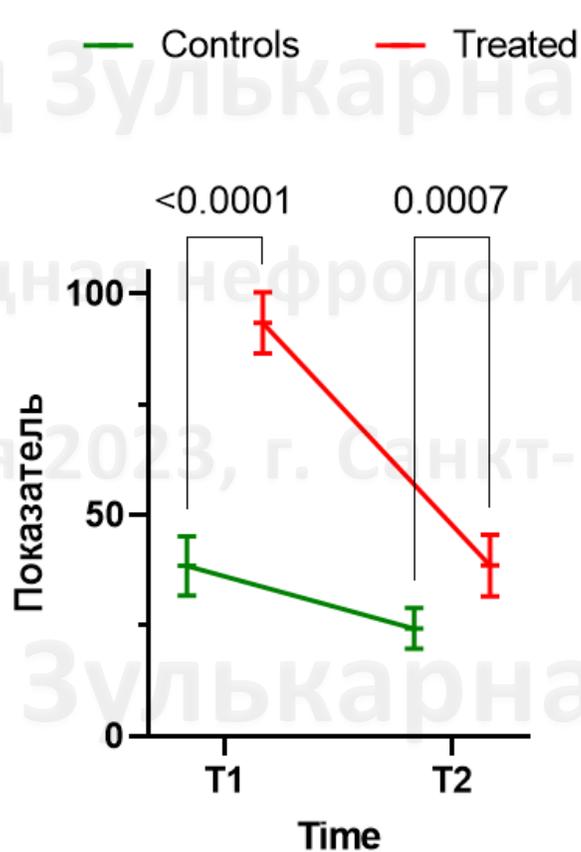


На обоих этапах группы различались.
Но автор не сравнивает динамику в
группах!

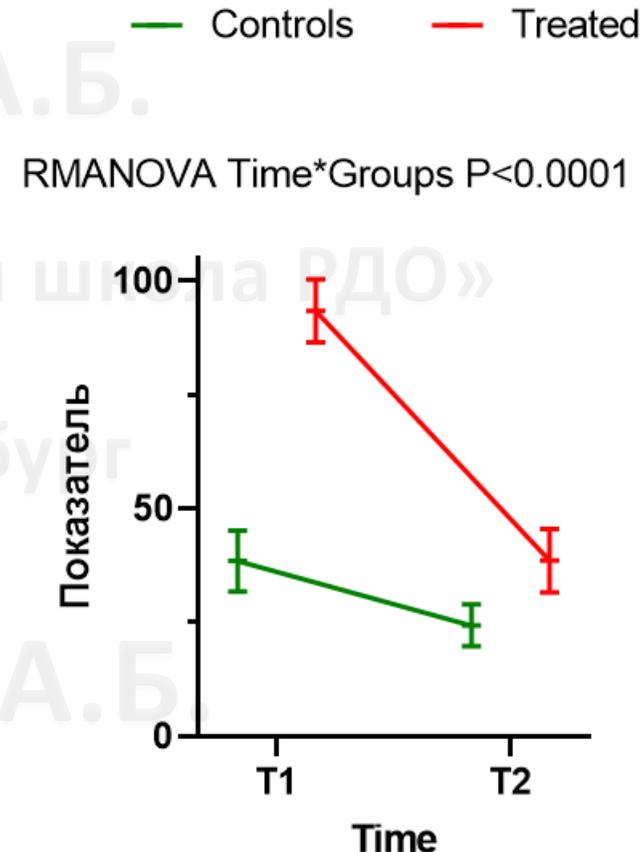
Какой вопрос задает исследователь?



В обеих группах была динамика. Но это не означает, что динамика в группах различалась!



На обоих этапах группы различались. Но автор не сравнивает динамику в группах!



Динамика в группах различалась

- Читателям и писателям

Всегда помните, какую именно гипотезу тестирует тот или иной тест

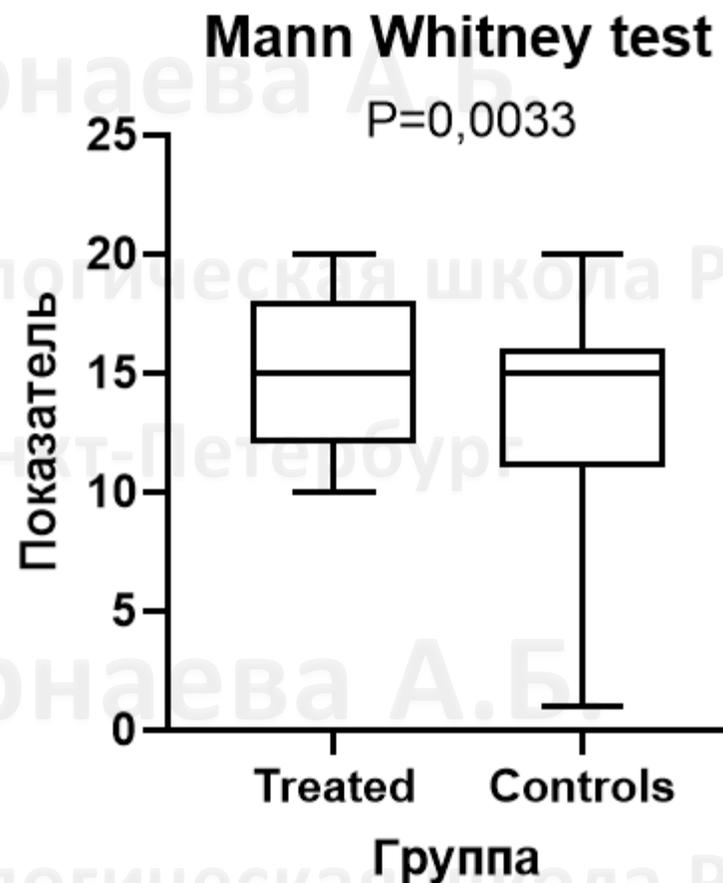
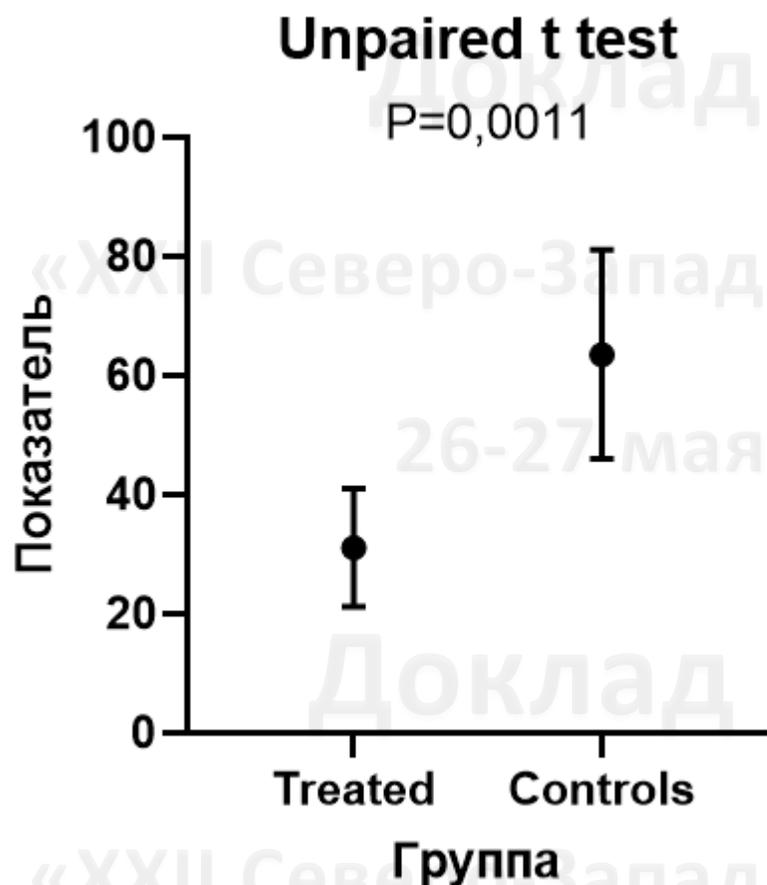
«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург



Средние в двух группах различаются

Difference between means: 32,43 [95%CI 15,83 to 49,03]

Распределения в двух группах различаются

Me [IQR]: 15 [12; 18] vs 15 [11; 16]

Difference between medians (Hodges-Lehmann): 1 [95%CI 0; 1]

При этом в обоих случаях вы можете сделать вывод, что фактор «группа» ассоциирован с количественным признаком

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

Заблуждение № 5

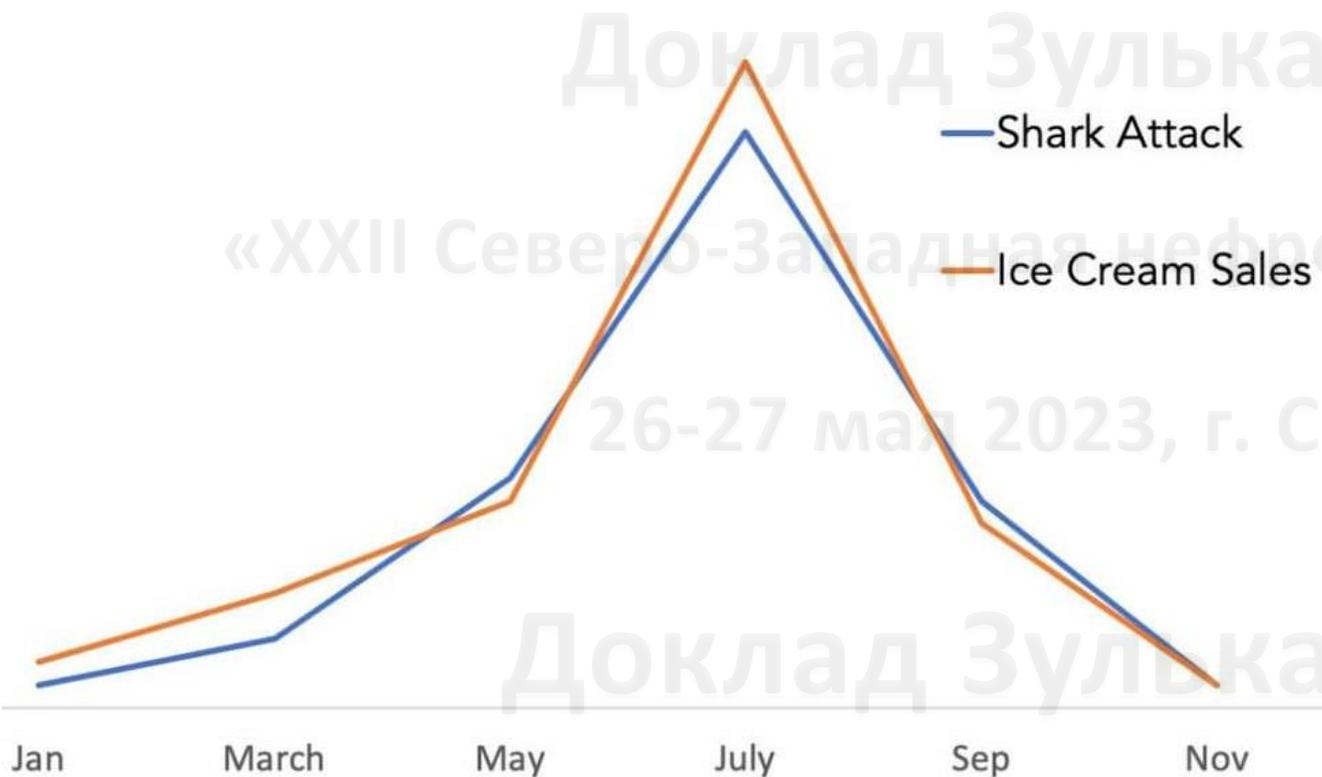
26-27 мая 2023, г. Санкт-Петербург

Если установлена связь между фактором риска и исходом, то она причинно-следственная

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург



Актуальный для нас с вами пример:

- ЦВК ассоциированы с худшей выживаемостью по сравнению с АВФ
- ЦВК это плохо, но в долгосрочной перспективе
- В течение первых пяти лет больные с ЦВК умирают чаще не из-за осложнений ЦВК, а из-за того, что ЦВК чаще используются у пациентов с более отягощенным коморбидным фоном

- Читателям

Факт связи между двумя событиями совершенно не означает причинно-следственный ее характер

- Писателям

Планируйте исследование, выбирайте его дизайн и потрудитесь доказать именно причинно-следственный характер связи

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»
Заблуждение № 6

26-27 мая 2023, г. Санкт-Петербург

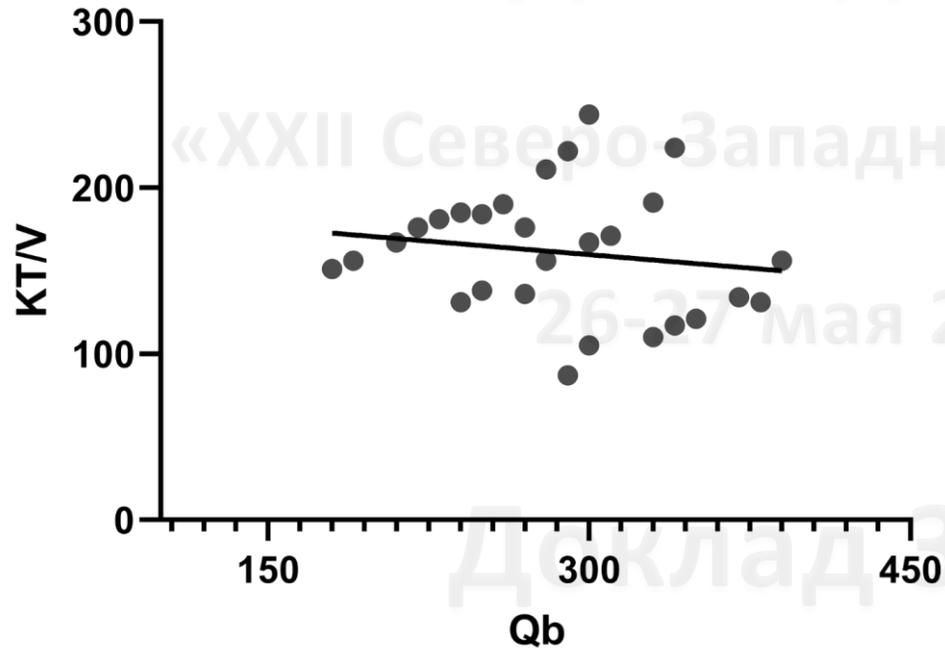
Сходу верить статье, где были получены парадоксальные
результаты

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.



«XXII Северо-Западная нефрологическая школа РДО»

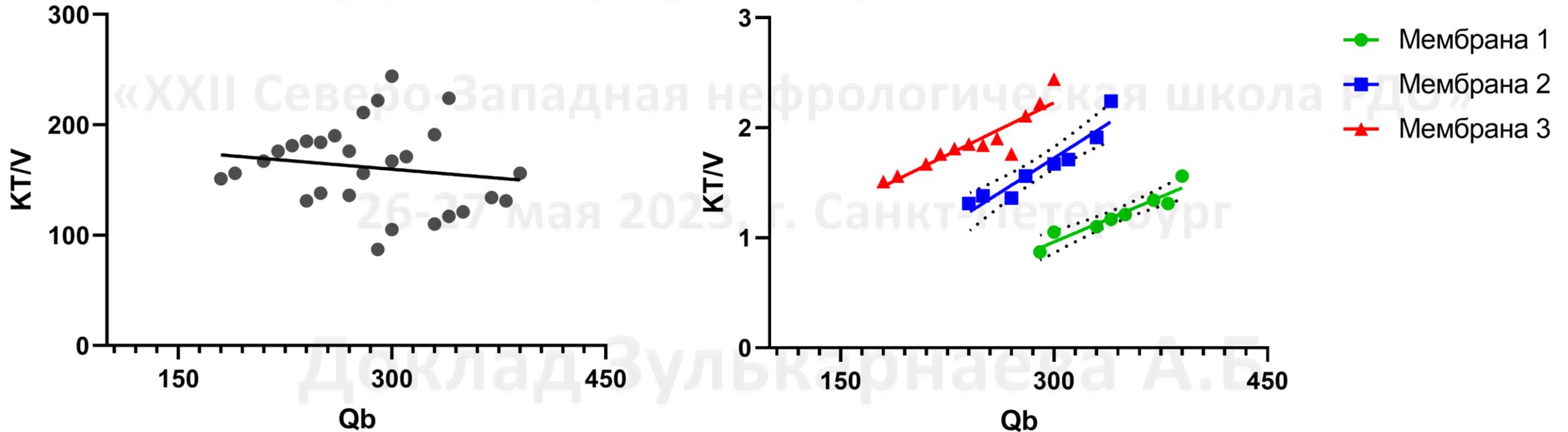
26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

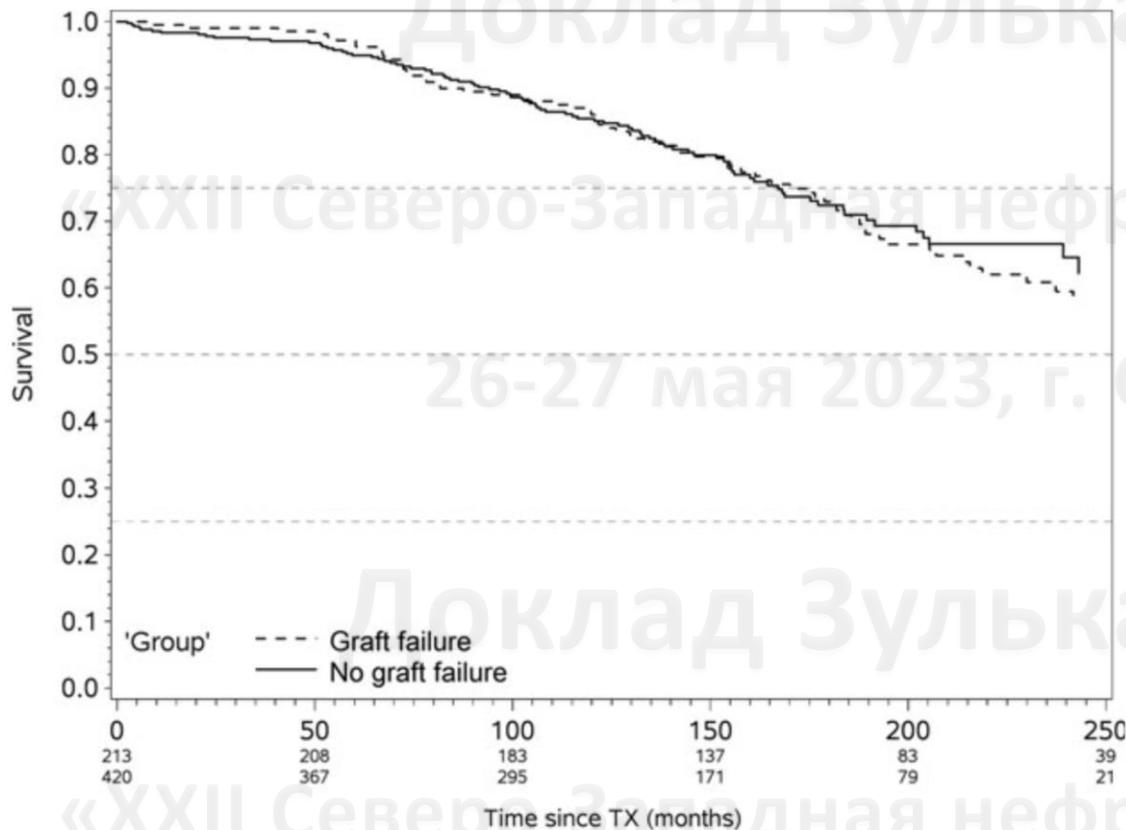
26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.



«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

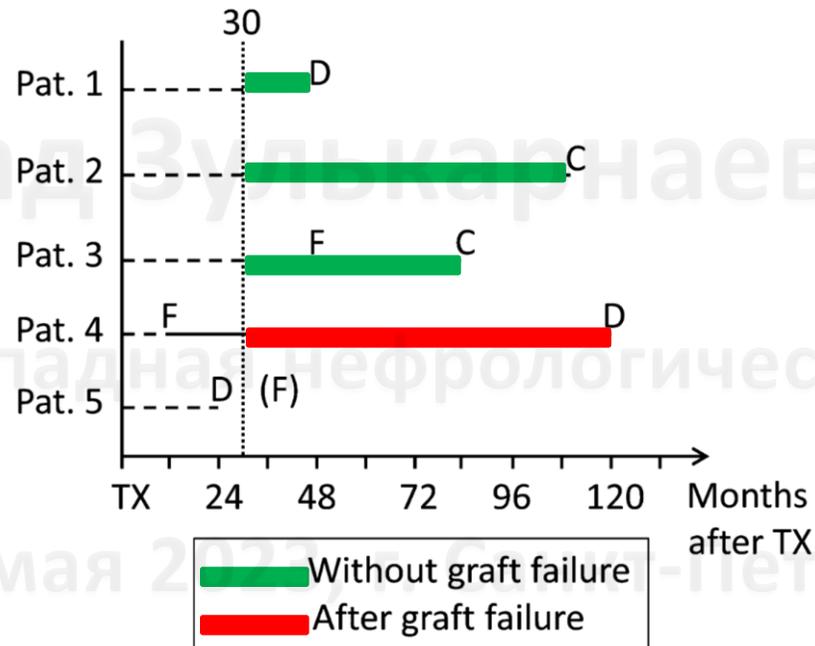
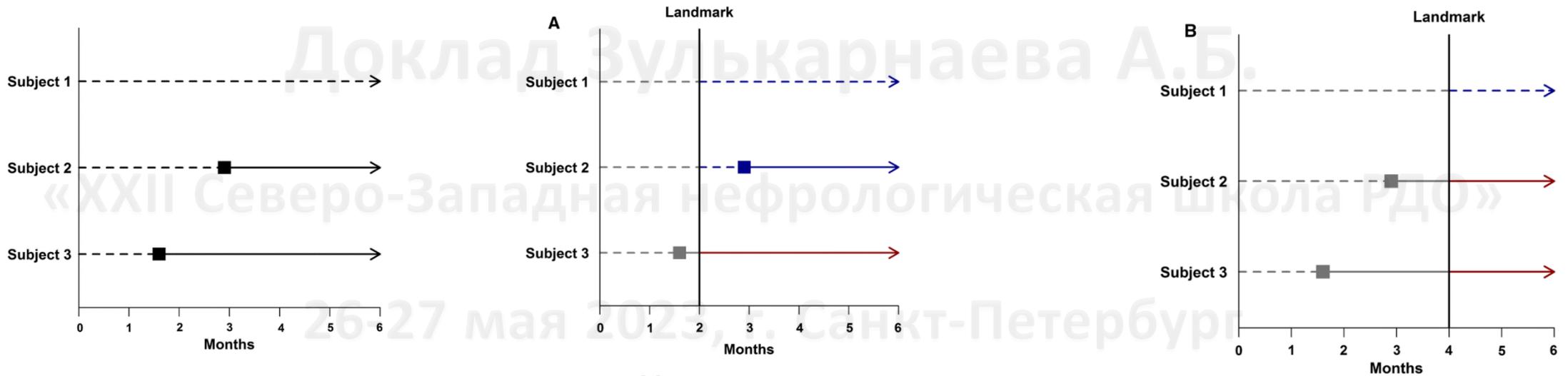


Анализируется связь утраты функции почечного трансплантата и смерти от всех причин.

В одно- и многофакторном анализе на нативных данных подтвердить наличие связи не удалось

Figure 1 Naïve Kaplan–Meier estimates (numbers at risk per group indicated below time axis).

Landmark analysis



Доклад Зулькарнаева А.Б.

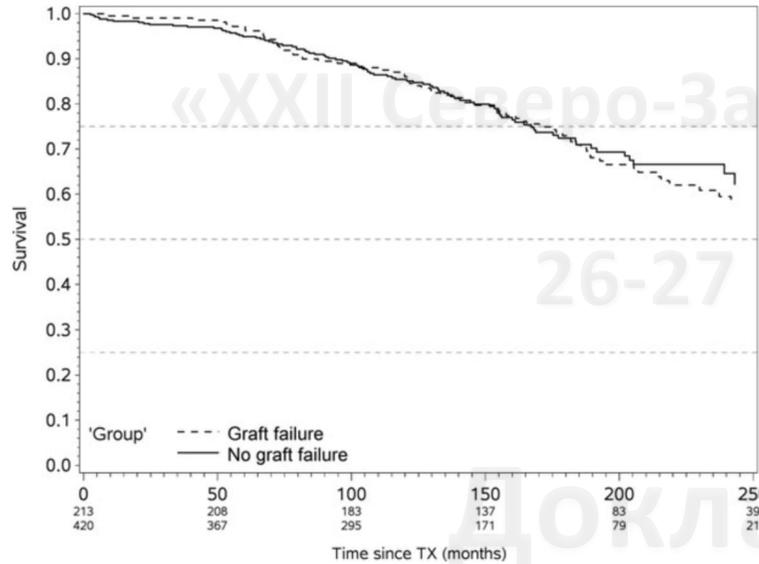
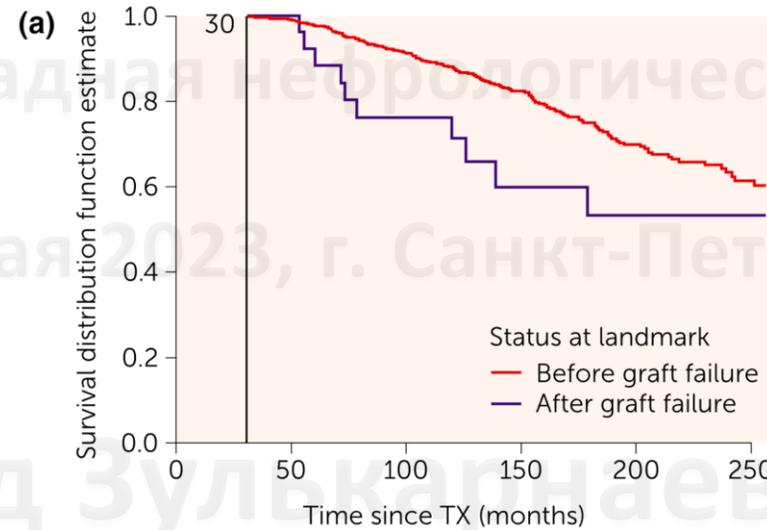
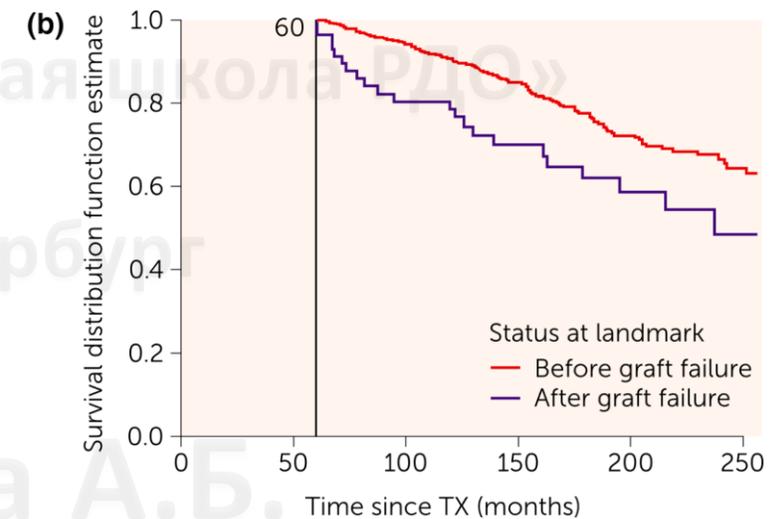


Figure 1 Naïve Kaplan–Meier estimates (numbers at risk per group indicated below time axis).



After graft failure	27	26	17	10	7	3
Before graft failure	573	549	461	298	155	57



After graft failure	58	44	29	17	6
Before graft failure	496	434	279	145	54

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

- Читателям:

Если в статье приведены парадоксальные результаты, прежде, чем верить ей, ознакомьтесь с известными источниками смещений для использованного метода анализа

26-27 мая 2023, г. Санкт-Петербург

- Писателям:

Сделайте то же самое, но ДО проведения исследования, на этапе планирования

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

Миф № 7

26-27 мая 2023, г. Санкт-Петербург

«Я построил регрессионную модель, которая позволяет
предсказывать риск ОПП»

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Если ничего не получается, а научник требует результат

```
set.seed(42)
df <- as.data.frame(matrix(rbinom(n=20*800,
                                size=1,
                                prob=0.4),
                            ncol=20))

set.seed(760)
outcome <- sample(c(rep(0, 400), rep(1, 400)))
```

- Генерируем набор данных:
 - 800 наблюдений
 - 20 бинарных предикторов с частотой 0,4
 - бинарный исход с частотой 0,5

26-27 мая 2023, г. Санкт-Петербург

Иными словами, у нас 400 больных, 400 здоровых субъектов, ни один из 20 предикторов в действительности не ассоциирован с заболеванием.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Однофакторный анализ

- Значимости нет и в помине

Variable	Outcome		p-value ²
	0, N = 400 ¹	1, N = 400 ¹	
V1	156 (39%)	147 (37%)	0.560
V2	144 (36%)	159 (40%)	0.308
V3	182 (46%)	165 (41%)	0.254
V4	160 (40%)	162 (40%)	0.943
V5	160 (40%)	177 (44%)	0.252
V6	164 (41%)	168 (42%)	0.830
V7	177 (44%)	161 (40%)	0.283
V8	150 (38%)	144 (36%)	0.714
V9	153 (38%)	173 (43%)	0.172
V10	149 (37%)	155 (39%)	0.716
V11	162 (40%)	158 (40%)	0.829
V12	162 (40%)	167 (42%)	0.774
V13	165 (41%)	171 (43%)	0.720
V14	160 (40%)	166 (42%)	0.719
V15	165 (41%)	158 (40%)	0.666
V16	171 (43%)	161 (40%)	0.518
V17	158 (40%)	151 (38%)	0.663
V18	161 (40%)	175 (44%)	0.352
V19	149 (37%)	147 (37%)	0.942
V20	155 (39%)	133 (33%)	0.122

¹ n (%)

² Fisher's exact test

Многофакторный анализ

```
model_1 <- glm(outcome ~ ., data = df, family = binomial)  
summary(model_1)
```

Characteristic	OR [†]	95% CI [†]	p-value
(Intercept)	0.99	0.58, 1.68	0.965
V1	0.91	0.68, 1.21	0.514
V2	1.19	0.89, 1.59	0.252
V3	0.83	0.62, 1.10	0.192
V4	1.01	0.76, 1.35	0.954
V5	1.18	0.89, 1.58	0.258
V6	1.02	0.76, 1.37	0.888
V7	0.86	0.64, 1.14	0.291
V8	0.93	0.69, 1.24	0.614
V9	1.28	0.96, 1.71	0.099
V10	1.12	0.84, 1.50	0.435
V11	0.94	0.71, 1.26	0.688
V12	1.06	0.79, 1.41	0.709
V13	1.07	0.80, 1.42	0.667
V14	1.09	0.81, 1.45	0.579
V15	0.93	0.70, 1.24	0.616
V16	0.91	0.68, 1.21	0.509
V17	0.93	0.69, 1.24	0.603
V18	1.15	0.87, 1.53	0.336
V19	1.00	0.75, 1.35	0.986
V20	0.79	0.59, 1.06	0.116

[†] OR = Odds Ratio, CI = Confidence Interval

- Значимости нет и в помине

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулъкарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

- А вдруг есть какое-то неблагоприятное сочетание факторов?
- Отсортируем предикторы по мере убывания p-value и посмотрим на них внимательно

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Если ничего не получается, а научник требует результат

Создаем варианты неблагоприятных сочетаний
топовых предикторов

```
df <- df %>%  
  mutate(v_comb_2 = case_when(  
    v20 == 0 &  
    v9 == 1 ~ 1,  
    TRUE ~ 0))  
  
df <- df %>%  
  mutate(v_comb_3 = case_when(  
    v20 == 0 &  
    v9 == 1 &  
    v5 == 1 ~ 1,  
    TRUE ~ 0))  
  
df <- df %>%  
  mutate(v_comb_4 = case_when(  
    v20 == 0 &  
    v9 == 1 &  
    v5 == 1 &  
    v3 == 0 ~ 1,  
    TRUE ~ 0))  
  
df <- df %>%  
  mutate(v_comb_5 = case_when(  
    v20 == 0 &  
    v9 == 1 &  
    v5 == 1 &  
    v3 == 0 &  
    v7 == 0 ~ 1,  
    TRUE ~ 0))
```

	Outcome		
Variable	0, N = 400 ¹	1, N = 400 ¹	p-value ²
V20	155 (39%)	> 133 (33%)	0.122
V9	153 (38%)	< 173 (43%)	0.172
V5	160 (40%)	< 177 (44%)	0.252
V3	182 (46%)	> 165 (41%)	0.254
V7	177 (44%)	> 161 (40%)	0.283
V2	144 (36%)	159 (40%)	0.308
V18	161 (40%)	175 (44%)	0.352

	Outcome		
Variable	0, N = 400 ¹	1, N = 400 ¹	p-value ²
V20	155 (39%)	133 (33%)	0.122
V9	153 (38%)	173 (43%)	0.172
V5	160 (40%)	177 (44%)	0.252
V3	182 (46%)	165 (41%)	0.254
V7	177 (44%)	161 (40%)	0.283
V2	144 (36%)	159 (40%)	0.308
V18	161 (40%)	175 (44%)	0.352
V16	171 (43%)	161 (40%)	0.518
V1	156 (39%)	147 (37%)	0.560
V17	158 (40%)	151 (38%)	0.663
V15	165 (41%)	158 (40%)	0.666
V8	150 (38%)	144 (36%)	0.714
V10	149 (37%)	155 (39%)	0.716
V14	160 (40%)	166 (42%)	0.719
V13	165 (41%)	171 (43%)	0.720
V12	162 (40%)	167 (42%)	0.774
V11	162 (40%)	158 (40%)	0.829
V6	164 (41%)	168 (42%)	0.830
V19	149 (37%)	147 (37%)	0.942
V4	160 (40%)	162 (40%)	0.943

¹ n (%)

² Fisher's exact test

Если ничего не получается, а научник требует результат

Создаем варианты неблагоприятных сочетаний
топовых предикторов

```
df <- df %>%  
  mutate(v_comb_2 = case_when(  
    v20 == 0 &  
    v9 == 1 ~ 1,  
    TRUE ~ 0))  
  
df <- df %>%  
  mutate(v_comb_3 = case_when(  
    v20 == 0 &  
    v9 == 1 &  
    v5 == 1 ~ 1,  
    TRUE ~ 0))  
  
df <- df %>%  
  mutate(v_comb_4 = case_when(  
    v20 == 0 &  
    v9 == 1 &  
    v5 == 1 &  
    v3 == 0 ~ 1,  
    TRUE ~ 0))  
  
df <- df %>%  
  mutate(v_comb_5 = case_when(  
    v20 == 0 &  
    v9 == 1 &  
    v5 == 1 &  
    v3 == 0 &  
    v7 == 0 ~ 1,  
    TRUE ~ 0))
```

	Outcome		
Variable	0, N = 400 ¹	1, N = 400 ¹	p-value ²
V20	155 (39%)	> 133 (33%)	0.122
V9	153 (38%)	< 173 (43%)	0.172
V5	160 (40%)	< 177 (44%)	0.252
V3	182 (46%)	> 165 (41%)	0.254
V7	177 (44%)	> 161 (40%)	0.283
V2	144 (36%)	159 (40%)	0.308
V18	161 (40%)	175 (44%)	0.352

	Outcome		
Variable	0, N = 400 ¹	1, N = 400 ¹	p-value ²
V20	155 (39%)	133 (33%)	0.122
V9	153 (38%)	173 (43%)	0.172
V5	160 (40%)	177 (44%)	0.252
V3	182 (46%)	165 (41%)	0.254
V7	177 (44%)	161 (40%)	0.283
V2	144 (36%)	159 (40%)	0.308
V18	161 (40%)	175 (44%)	0.352
V16	171 (43%)	161 (40%)	0.518
V1	156 (39%)	147 (37%)	0.560
V17	158 (40%)	151 (38%)	0.663
V15	165 (41%)	158 (40%)	0.666
V8	150 (38%)	144 (36%)	0.714
V10	149 (37%)	155 (39%)	0.716
V14	160 (40%)	166 (42%)	0.719
V13	165 (41%)	171 (43%)	0.720
V12	162 (40%)	167 (42%)	0.774
V11	162 (40%)	158 (40%)	0.829
V6	164 (41%)	168 (42%)	0.830
V19	149 (37%)	147 (37%)	0.942
V4	160 (40%)	162 (40%)	0.943

¹ n (%)

² Fisher's exact test

- Оценим связь неблагоприятных сочетаний предикторов с исходом

```
df %>%  
  select(outcome, v_comb_2, v_comb_3, v_comb_4, v_comb_5) %>%  
  tbl_uvregression(  
    method = glm,  
    y = outcome,  
    method.args = list(family = binomial),  
    exponentiate = TRUE,  
    estimate_fun = ~ style_ratio(.x, digits = 2),  
    pvalue_fun = ~ style_pvalue(.x, digits = 3)) %>%  
    bold_p(t = 0.05, q = FALSE) %>%  
    bold_labels() %>%  
    modify_header(label ~ "***variable**")
```

Variable	N	OR ¹	95% CI ¹	p-value
V_comb_2	800	1.35	0.98, 1.86	0.064
V_comb_3	800	2.10	1.34, 3.36	0.001
V_comb_4	800	1.60	0.92, 2.82	0.099
V_comb_5	800	2.38	1.15, 5.29	0.025

¹ OR = Odds Ratio, CI = Confidence Interval

- Поздравляю, теперь вы р-хакер!
- Вы доказали, что если фамилия начинается с гласной буквы (V20), пациент госпитализирован в среду (V5) и у него телефон на андроид (V5), то риск ОПП после аппендэктомии возрастает.

Если б не одно «но»: оно так не работает

Тестируем модель на новой выборке, меняем только seed ГПСЧ.

```
set.seed(42)
df <- as.data.frame(matrix(rbinom(n=20*800,
                                size=1,
                                prob=0.4),
                            ncol=20))
set.seed(760)
outcome <- sample(c(rep(0, 400), rep(1, 400)))
```

```
set.seed(1789)
df <- as.data.frame(matrix(rbinom(n=20*800,
                                size=1,
                                prob=0.4),
                            ncol=20))
set.seed(1015)
outcome <- sample(c(rep(0, 400), rep(1, 400)))
```

Variable	N	OR ¹	95% CI ¹	p-value
V_comb_2	800	1.35	0.98, 1.86	0.064
V_comb_3	800	2.10	1.34, 3.36	0.001
V_comb_4	800	1.60	0.92, 2.82	0.099
V_comb_5	800	2.38	1.15, 5.29	0.025

¹ OR = Odds Ratio, CI = Confidence Interval

Variable	N	OR ¹	95% CI ¹	p-value
V_comb_2	800	0.80	0.53, 1.19	0.275
V_comb_3	800	1.05	0.63, 1.71	0.862
V_comb_4	800	0.72	0.33, 1.48	0.382
V_comb_5	800	8,263,528	0.00, NA	0.971

¹ OR = Odds Ratio, CI = Confidence Interval

Для реализации парадокса Фридмана необязательно иметь большие выборки!

```
set.seed(1)
df <- as.data.frame(matrix(rbinom(n=10*100,
                                size=1,
                                prob=0.4),
                            ncol=10))

set.seed(1)
outcome <- sample(c(rep(0, 50), rep(1, 50)))
df$outcome <- outcome
colSums(df)
str(df)
```

100 наблюдений
50 с исходом, 50 без исхода
10 предикторов

Однофакторный анализ

Characteristic	0, N = 50 ¹	1, N = 50 ¹	p-value ²
V10	14 (28%)	26 (52%)	0.024
V3	9 (18%)	19 (38%)	0.044
V8	16 (32%)	25 (50%)	0.103
V7	17 (34%)	22 (44%)	0.412
V9	17 (34%)	22 (44%)	0.412
V1	23 (46%)	20 (40%)	0.686
V2	19 (38%)	21 (42%)	0.838
V5	19 (38%)	21 (42%)	0.838
V4	18 (36%)	19 (38%)	>0.999
V6	23 (46%)	24 (48%)	>0.999

¹ n (%)

² Fisher's exact test

Многофакторный анализ

Characteristic	OR ¹	95% CI ¹	p-value
V10	2.58	1.06, 6.46	0.038
V3	2.55	0.96, 7.21	0.066
V8	1.97	0.81, 4.90	0.137
V1	0.61	0.23, 1.51	0.289
V7	1.56	0.63, 3.94	0.339
V9	1.36	0.54, 3.44	0.512
V4	0.80	0.31, 2.05	0.646
V2	1.18	0.48, 2.95	0.715
V5	1.17	0.47, 2.91	0.732
V6	0.93	0.37, 2.32	0.882

¹ OR = Odds Ratio, CI = Confidence Interval

- «Я построил регрессионную модель, которая позволяет **предсказывать** риск ОПП»
- Вы можете говорить, что модель предсказывает что-то, только когда она начнет именно предсказывать
- Качество модели на обучающей выборке = соответствие модели данным
- Качество модели на тестовой выборке = предсказательная способность

Писателям и читателям

Доказано, что модель может что-то предсказывать только тогда, когда она «увидит» новые данные!

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

Миф № 8

26-27 мая 2023, г. Санкт-Петербург

«Больше графиков хороших и разных. Во-первых, это красиво.»

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

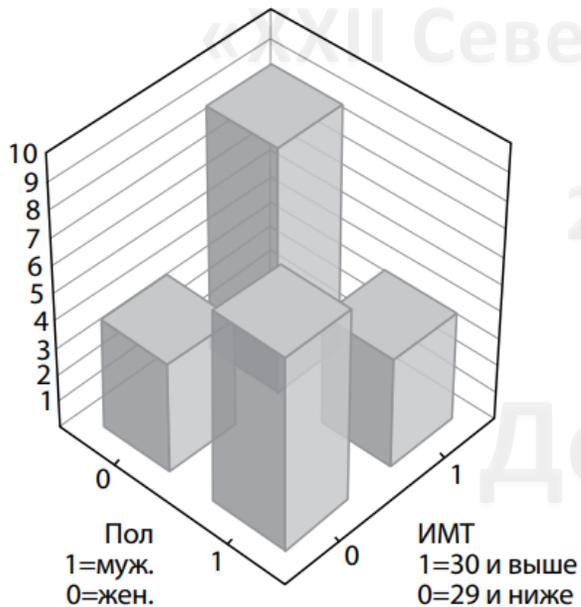
Ой, все!



26-27 мая 2023, г. Санкт-Петербург

Зачем нужны графики?

- Чтобы сделать изложение более наглядным
- Дополнить (но не дублировать!) текст



Этот график затрудняет восприятие.

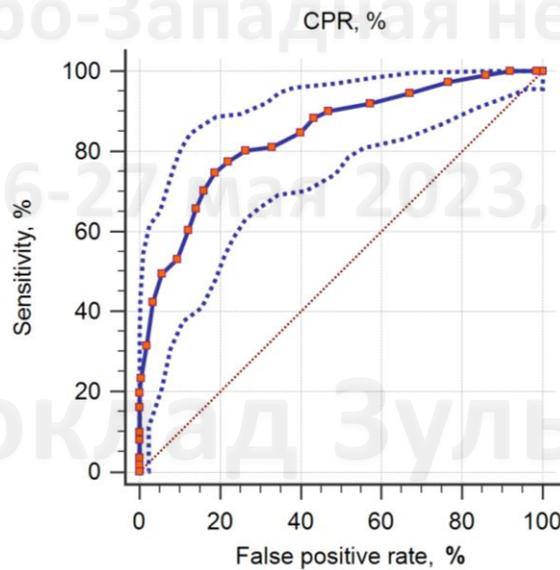
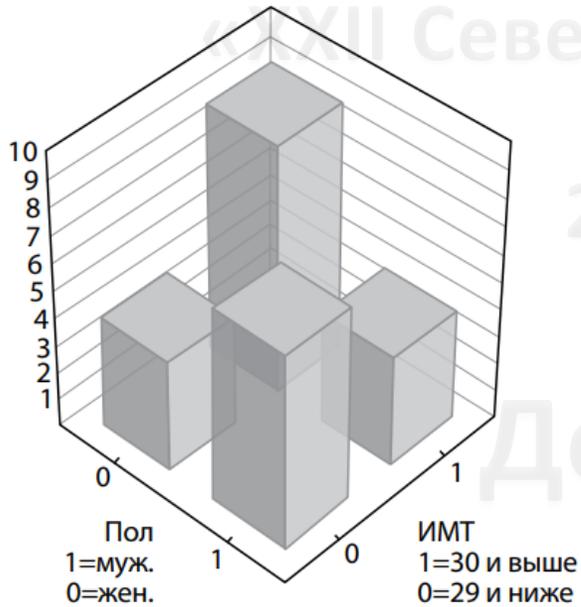
Его можно заменить одним предложением:

«Среди М лиц с ИМТ было X (%), среди Ж – Y (%)»/

График не нужен

Зачем нужны графики?

- Чтобы сделать изложение более наглядным
- Дополнить (но не дублировать!) текст



Этот график затрудняет восприятие.

Его можно заменить одним

предложением:

«Среди М лиц с ИМТ ≥ 30 было X (%),

среди Ж – Y (%)»/

График не нужен

Этот график можно заменить одним

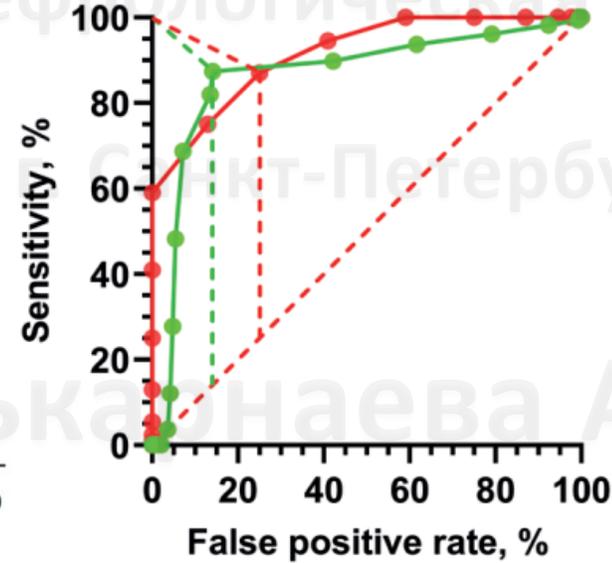
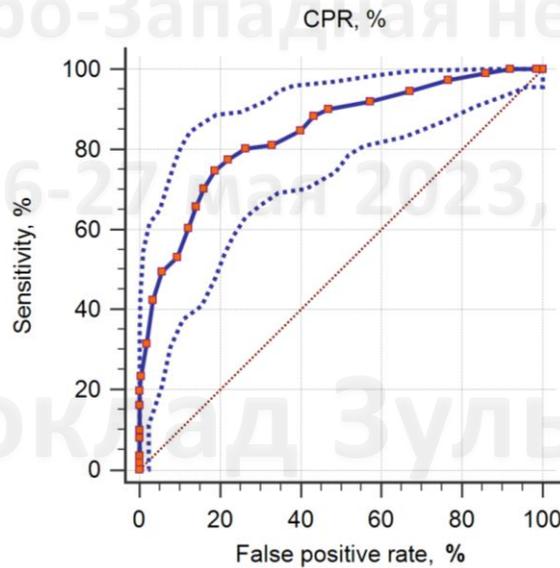
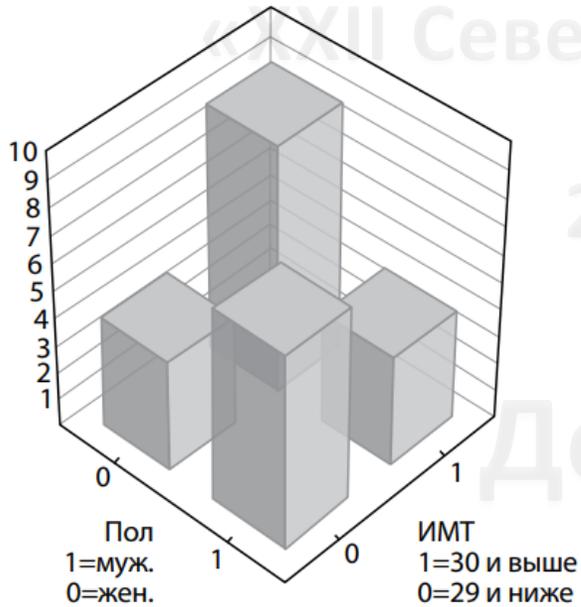
предложением:

«AUC-ROC 0.842 [95%CI 0.797; 0.88], $p < 0.0001$ ».

График не нужен

Зачем нужны графики?

- Чтобы сделать изложение более наглядным
- Дополнить (но не дублировать!) текст



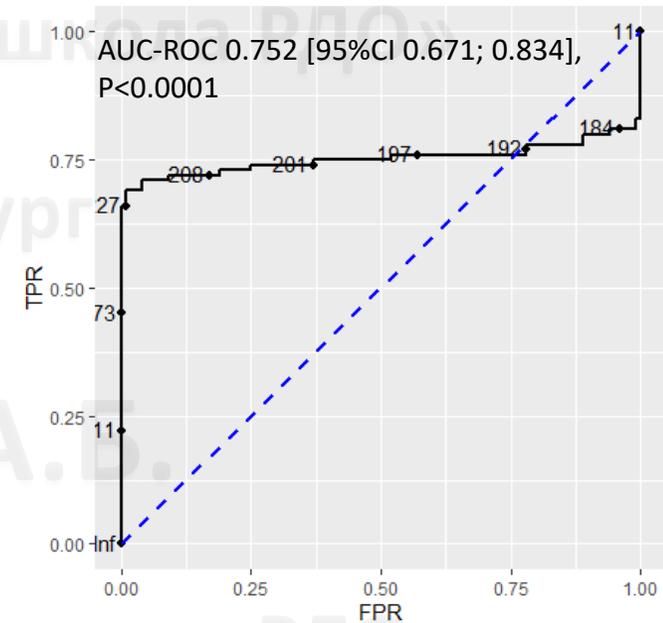
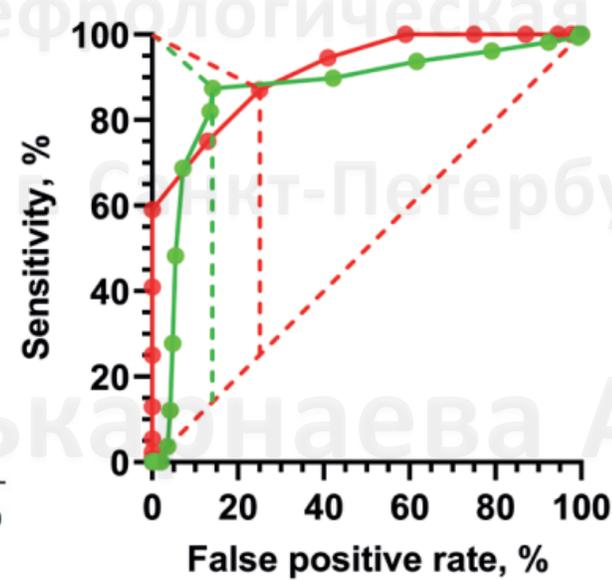
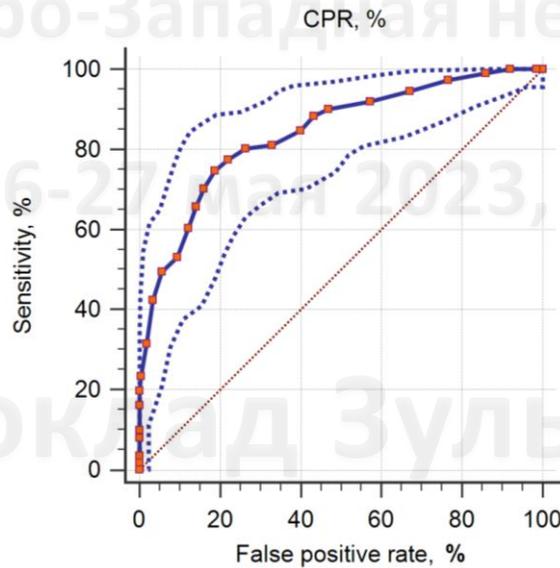
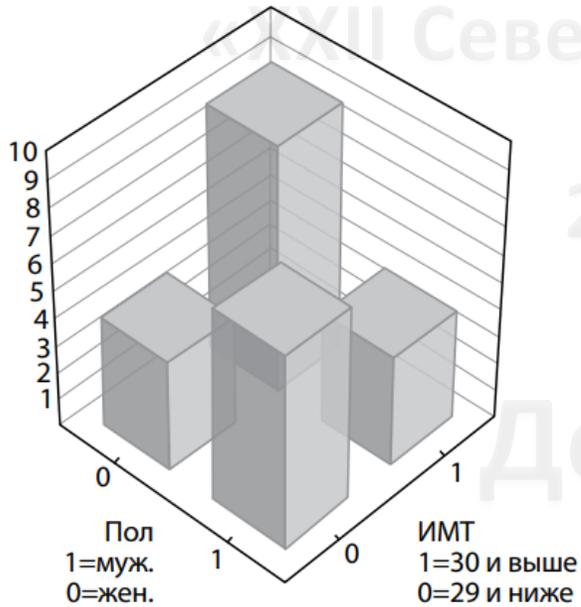
Этот график затрудняет восприятие.
Его можно заменить одним предложением:
«Среди М лиц с ИМТ было X (%), среди Ж – Y (%)»/
График не нужен

Этот график можно заменить одним предложением:
«AUC-ROC 0.842 [95%CI 0.797; 0.88], $p < 0.0001$ ».
График не нужен

Этот график делает изложение более наглядным

Зачем нужны графики?

- Чтобы сделать изложение более наглядным
- Дополнить (но не дублировать!) текст



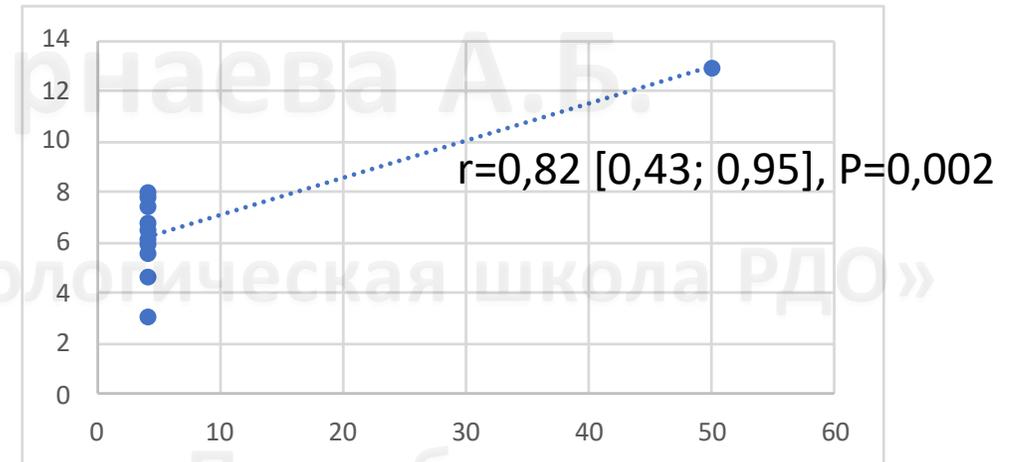
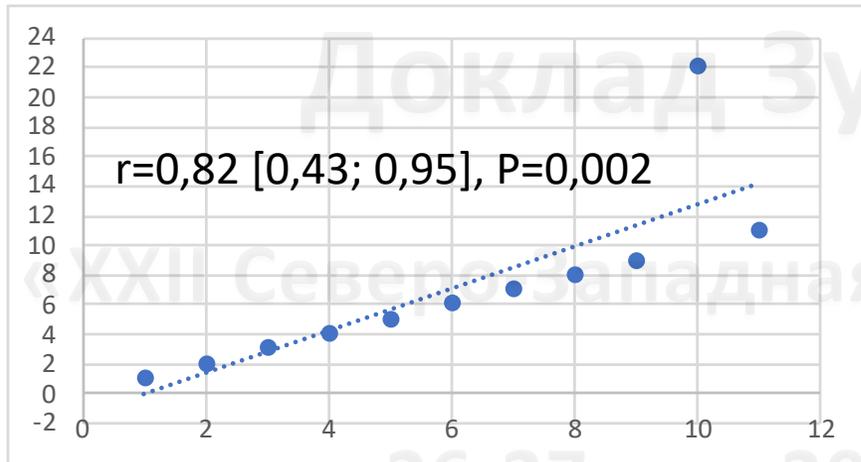
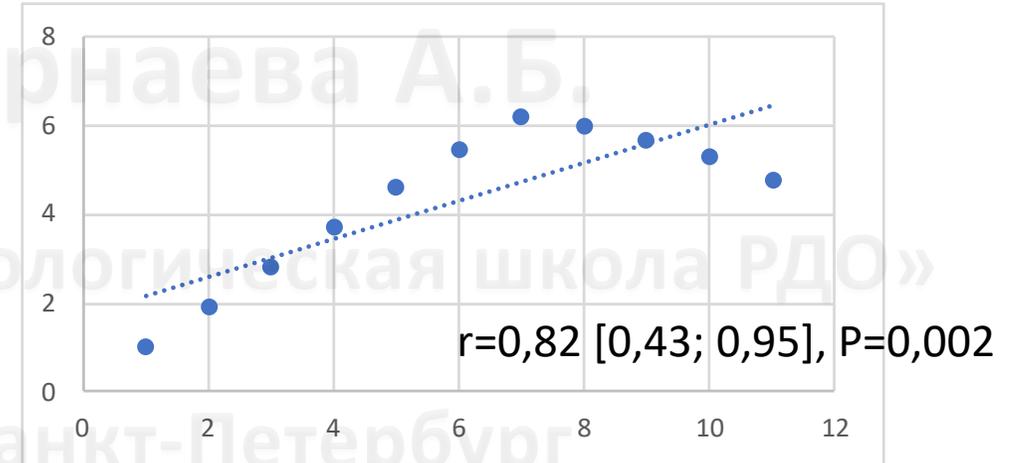
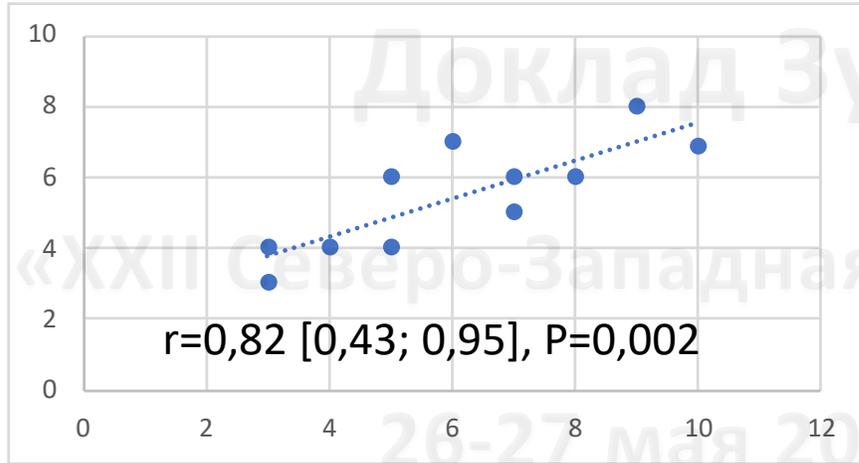
Этот график затрудняет восприятие.
Его можно заменить одним предложением:
«Среди М лиц с ИМТ было X (%), среди Ж – Y (%)»/
График не нужен

Этот график можно заменить одним предложением:
«AUC-ROC 0.842 [95%CI 0.797; 0.88], p<0.0001».
График не нужен

Этот график делает изложение более наглядным

Этот график дополняет изложение

Когда графики очень нужны?



Доклад Зулькарнаева А.Б.

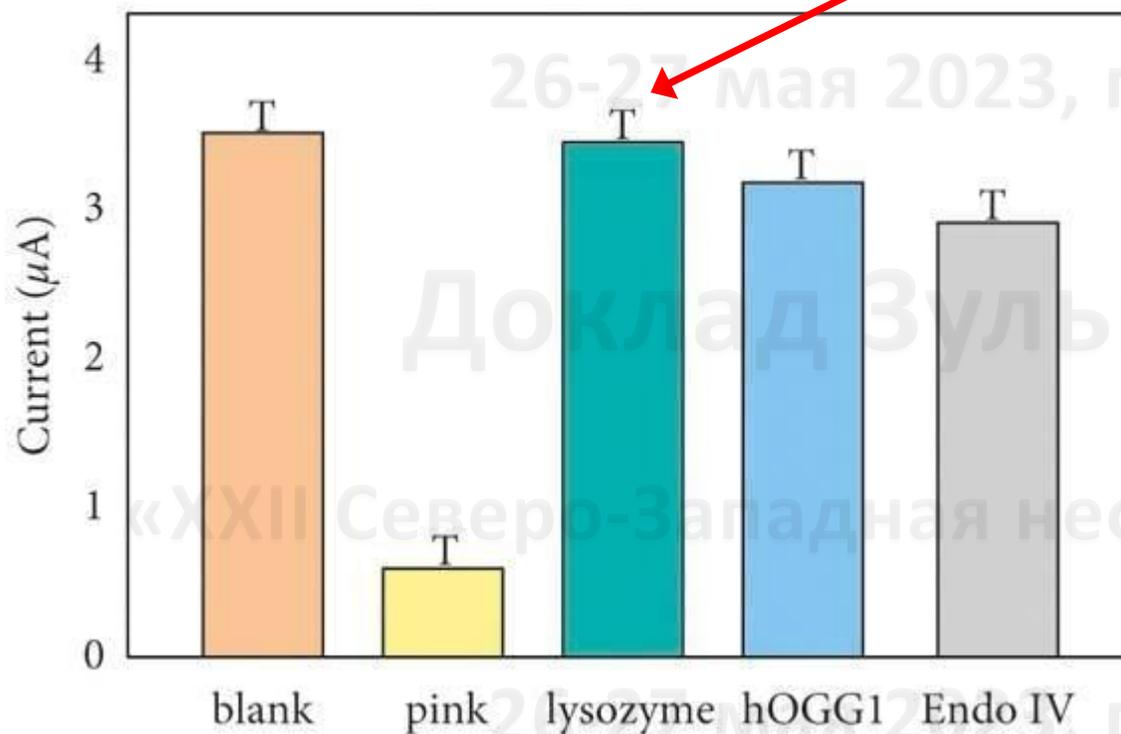
«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

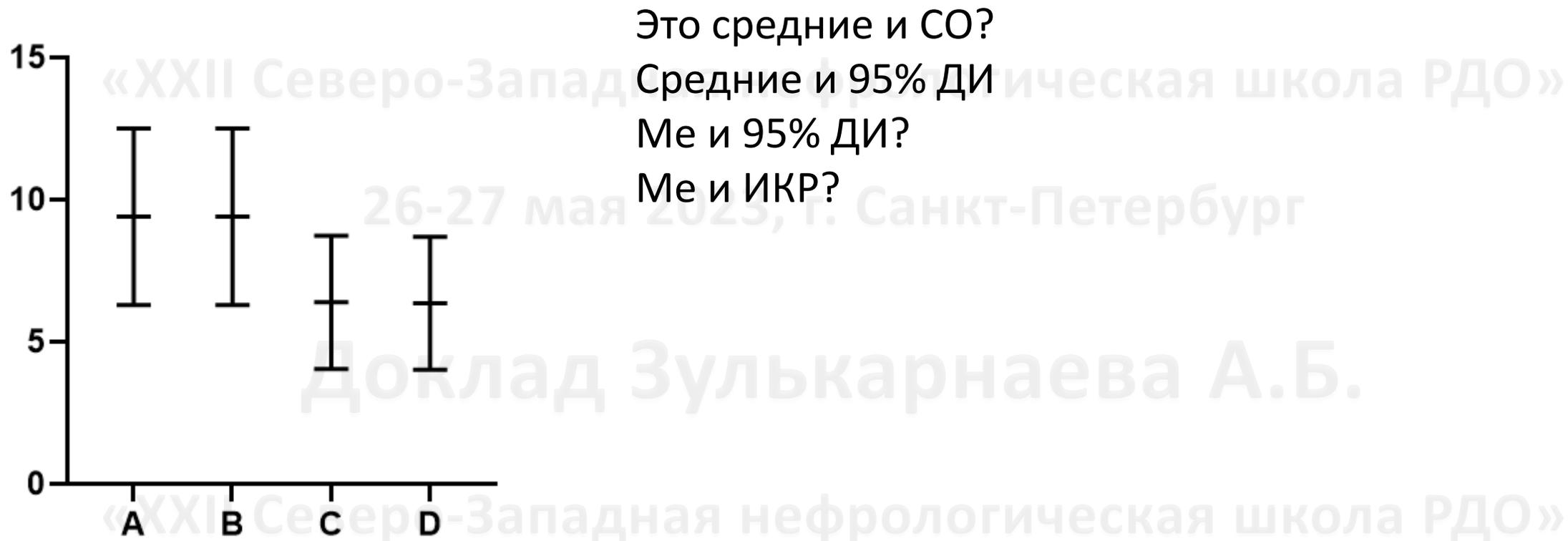
Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург



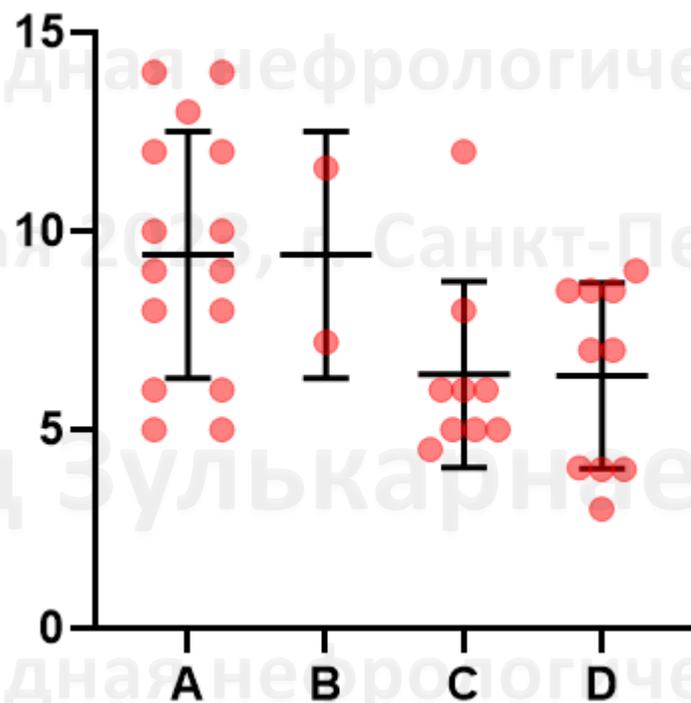
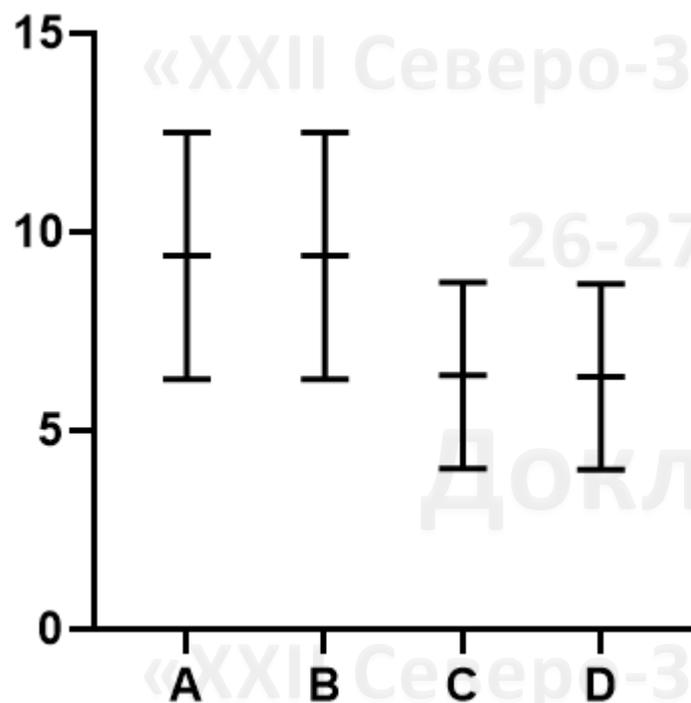
Доклад Зулькарнаева А.Б.



Доклад Зулькарнаева А.Б.

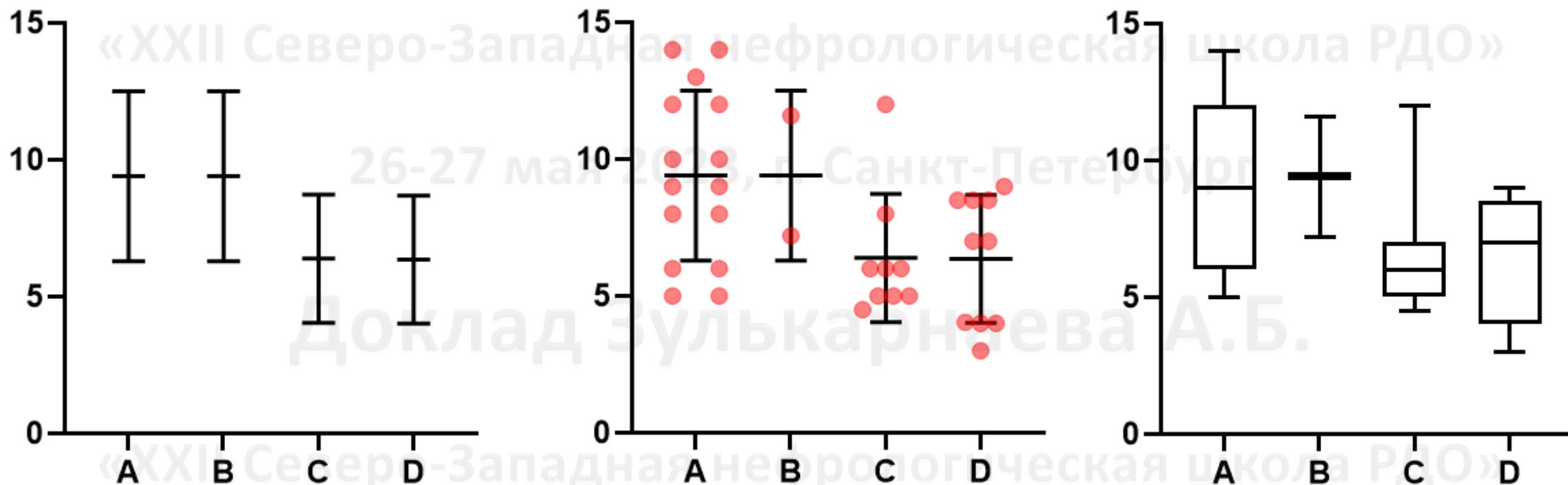
26-27 мая 2023, г. Санкт-Петербург

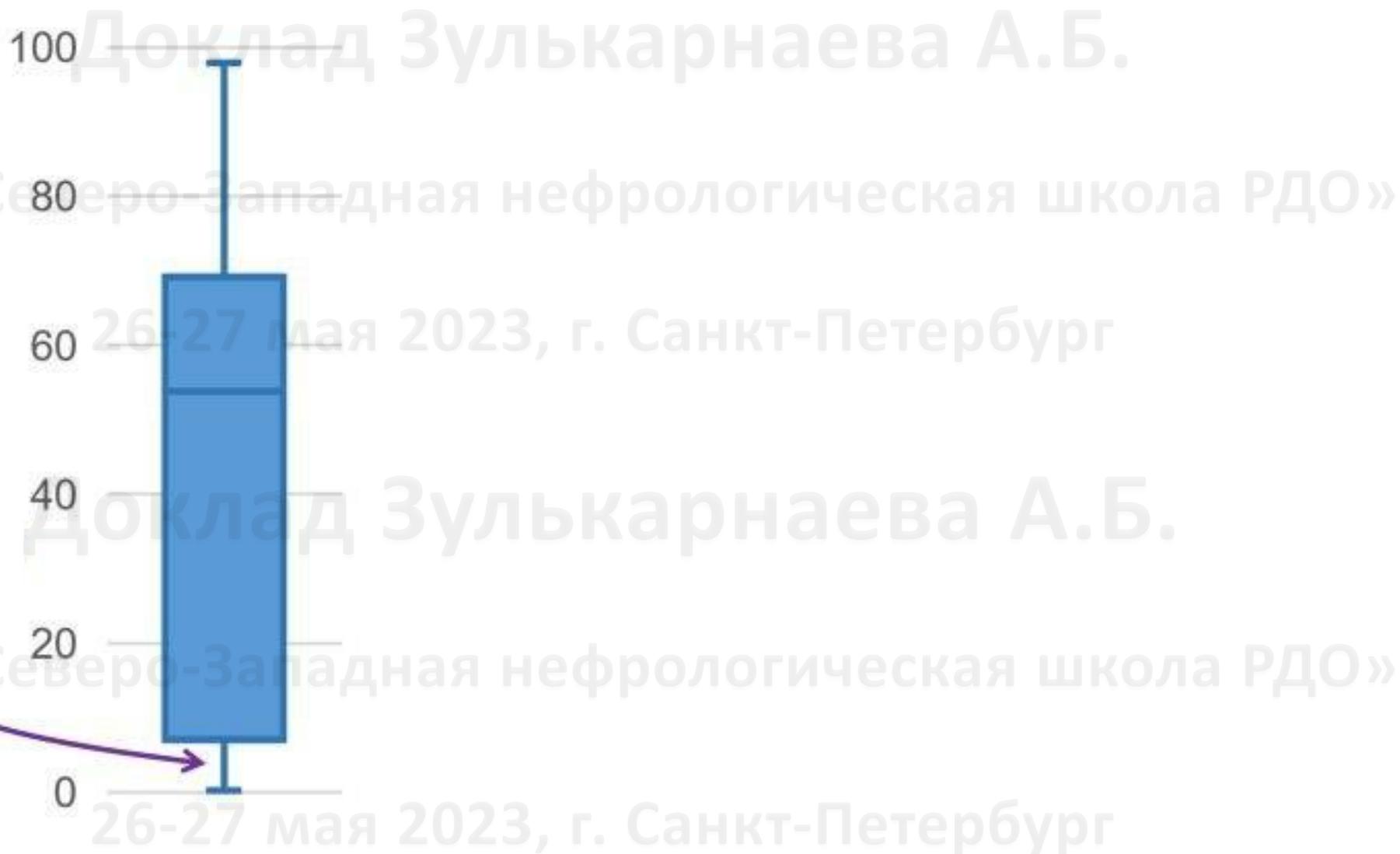
Доклад Зулькарнаева А.Б.

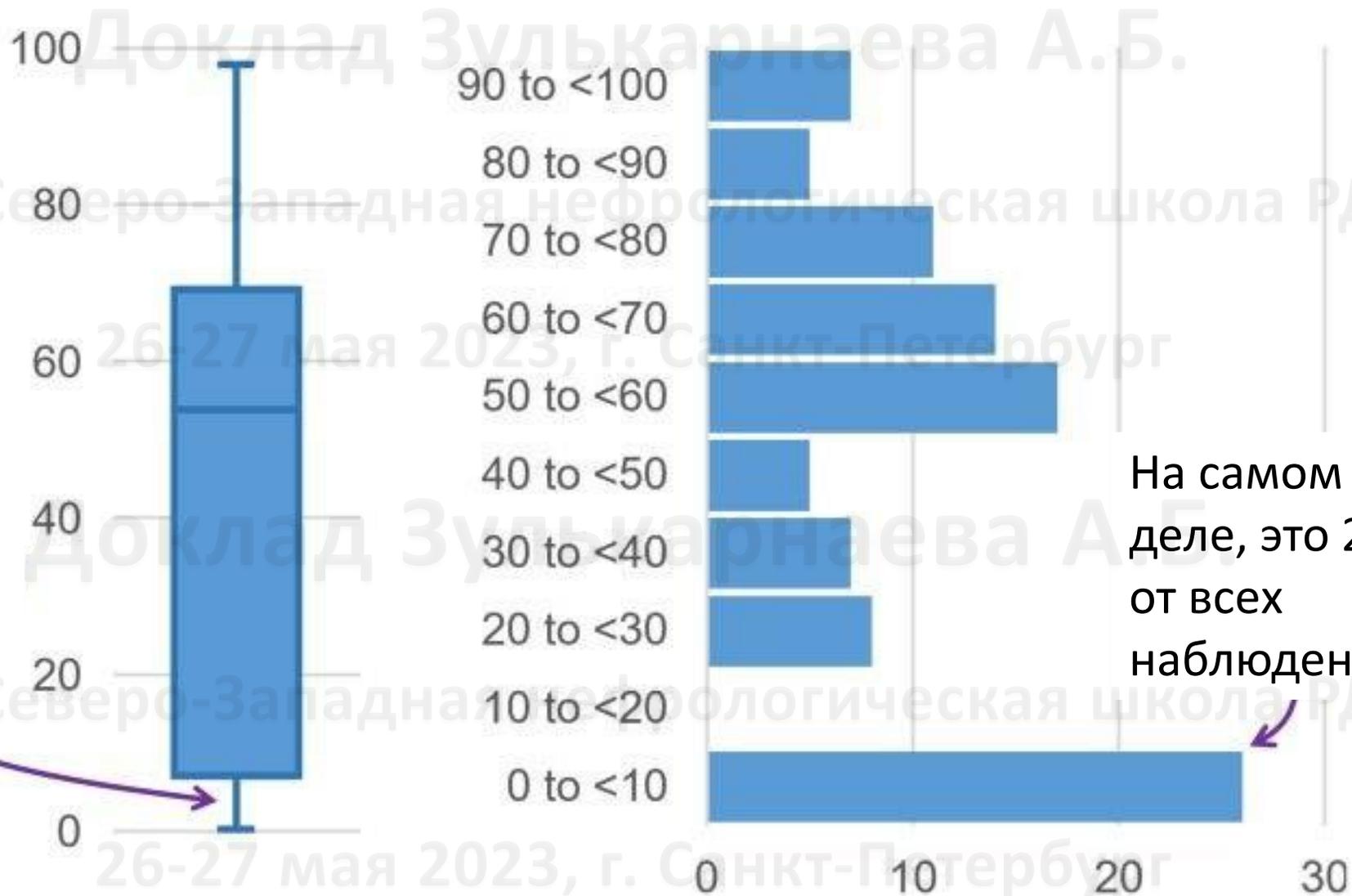


26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б. Это min-max boxplot или Tukey boxplot?

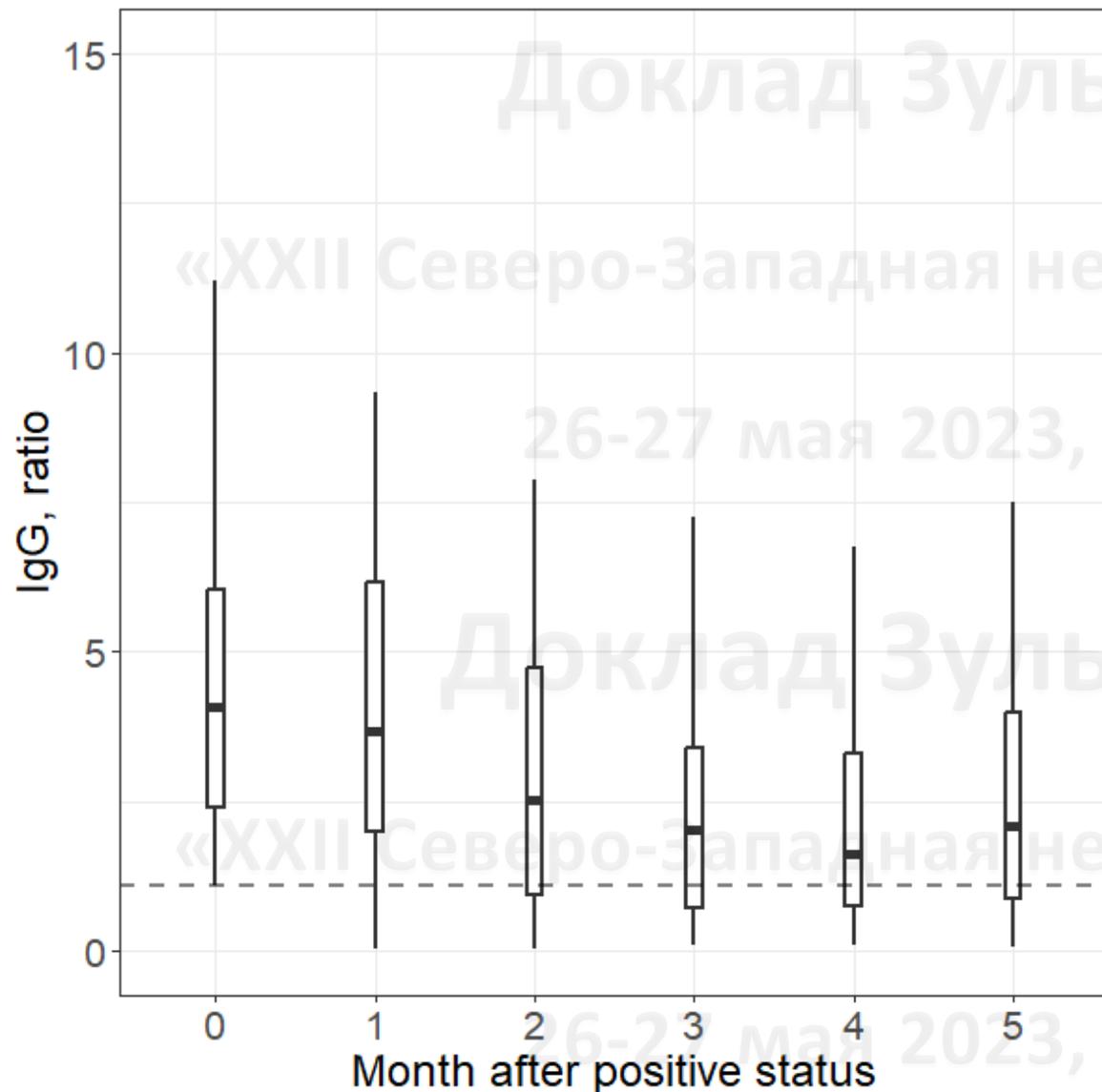




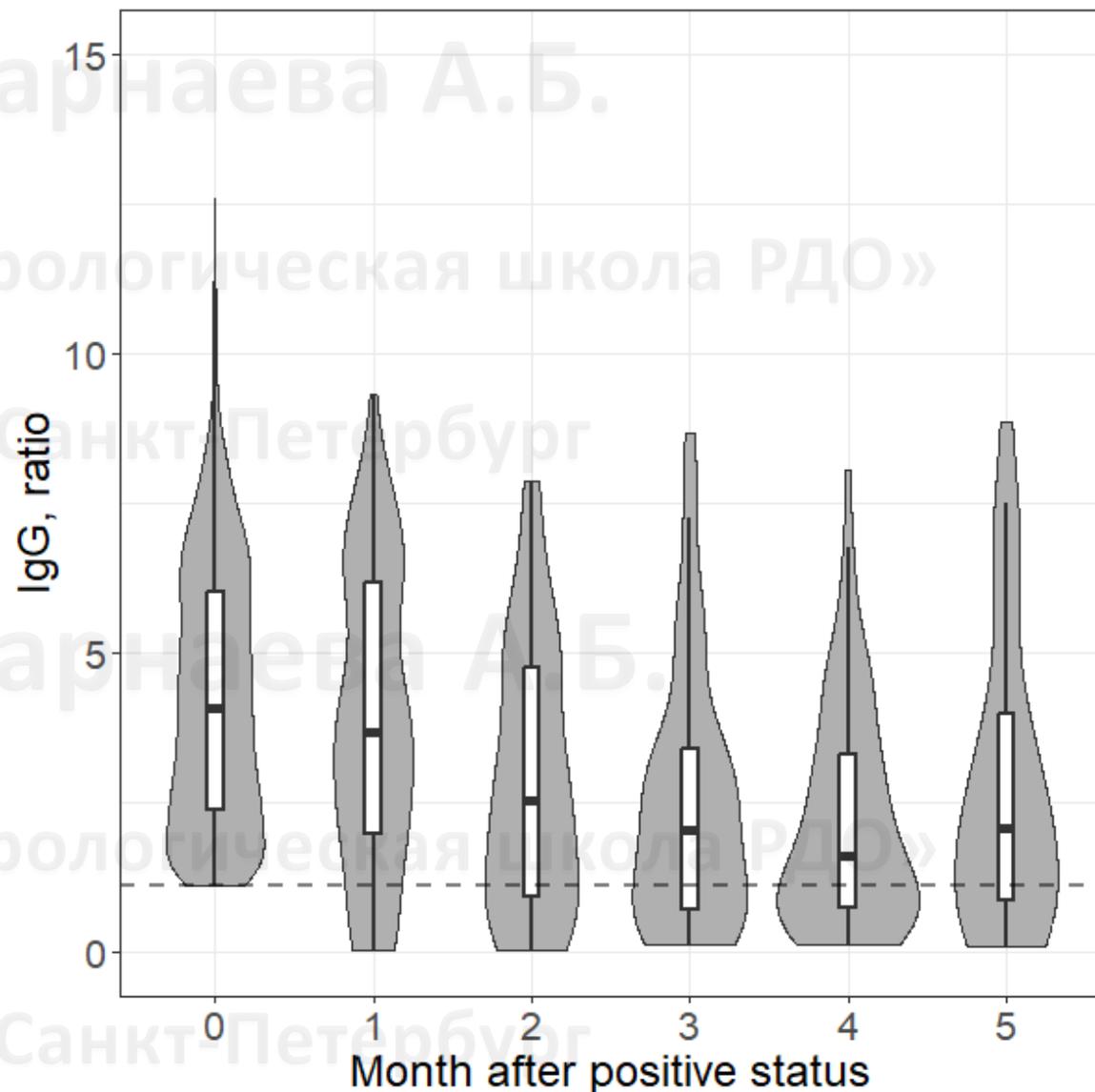
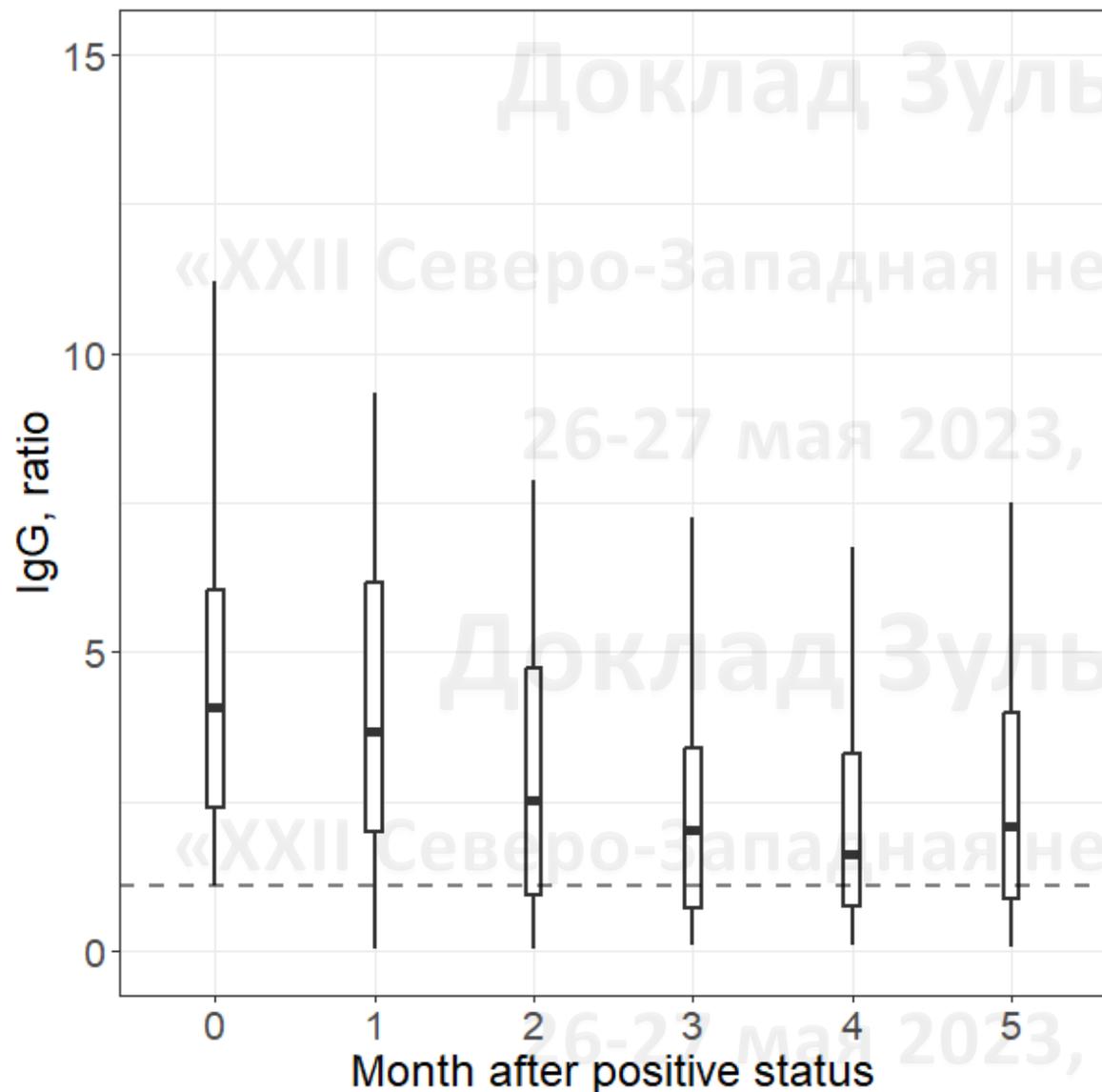


Кажется, что тут мало наблюдений

На самом деле, это 25% от всех наблюдений



Правильно «читайте» графики



Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

Ошибка № 9

26-27 мая 2023, г. Санкт-Петербург

Отдавать предпочтение точечным, а не интервальным оценкам

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

- Три коэффициента корреляции:

$$r_s = 0,9058, P=0,0942$$

$$r_s = 0,9058, P=0,0019$$

$$r_s = 0,9058, P<0,0001$$

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

- Три коэффициента корреляции:

$$r_s = 0,9058 \text{ [95\%ДИ } -0,427; 0,998], P=0,0942$$

$$r_s = 0,9058 \text{ [95\%ДИ } 0,5561; 0,983], P=0,0019$$

$$r_s = 0,9058 \text{ [95\%ДИ } 0,6913; 0,9736], P<0,0001$$

Доклад Зулькарнаева А.Б.

«XXII Северо-Западная нефрологическая школа РДО»

26-27 мая 2023, г. Санкт-Петербург

- Читателям:

Не интерпретируйте точечные оценки в отрыве от интервальных

- Писателям

Всегда дополняйте точечные оценки интервальными

Ой, все!

