After Work Statistics

Dr. Jochen Kruppa
Institute of Biometry and
Clinical Epidemiology
jochen.kruppa@charite.de

Institute of Biometry and Clinical Epidemiology

running your research

We are…

- … open and helpful!
- … active in the statistical methodologic research and in medical research
- …active in teaching in many ways

Our Service Unit Biometry

- **Free biometrical consulting for all medical research projects,** registration online
- **“Statistik-Ambulanz” (Walk-in service):** Consultation without prior registration every Tuesday from 9am to 12pm
- **Training** in biometrical topics and statistical software
- Responsibility for project biometry within cooperation

For further information visit us online:
https://biometrie.charite.de/

Contact: Univ.-Prof. Dr. Geraldine Rauch (Head of Institute), Institut für Biometrie und Klinische Epidemiologie (iBikE)

Standort Mitte (Charité Campus Mitte)
Reinhardstraße 58, 10117 Berlin

Standort Mitte (Charité Campus Klinik)
Rahel-Hirsch-Weg 5, 10117 Berlin

Slot	Topic
1	So many tests! The agony of choice.
2	**So many questions! Multiple testing.**
3	So many patients? Sample size calculation.
4	What is it this odds ratio? Logistic regression.
5	Missing information? Dealing with missing data.
6	The right time? Survival analysis.
7	The variety of influences - Mixed models.
8	Who fits together? Patient matching.
1	So viele Tests! Die Qual der Wahl.
2	So viele Fragestellungen! Multiples Testen.
3	So viele Patienten? Fallzahlplanung.
4	Was ist dieses Odds Ratio? Logistische Regression.
5	Fehlende Information? Umgang mit fehlenden Daten.
6	Der richtige Zeitpunkt? Analyse von Ereigniszeiten.
7	Die Vielfalt der Einflüsse – Gemischte Modelle.
8	Wer passt zusammen? Matching von Patienten.

So many questions! Multiple Testing

Dr. Jochen Kruppa
Institute of Biometry and
Clinical Epidemiology
jochen.kruppa@charite.de

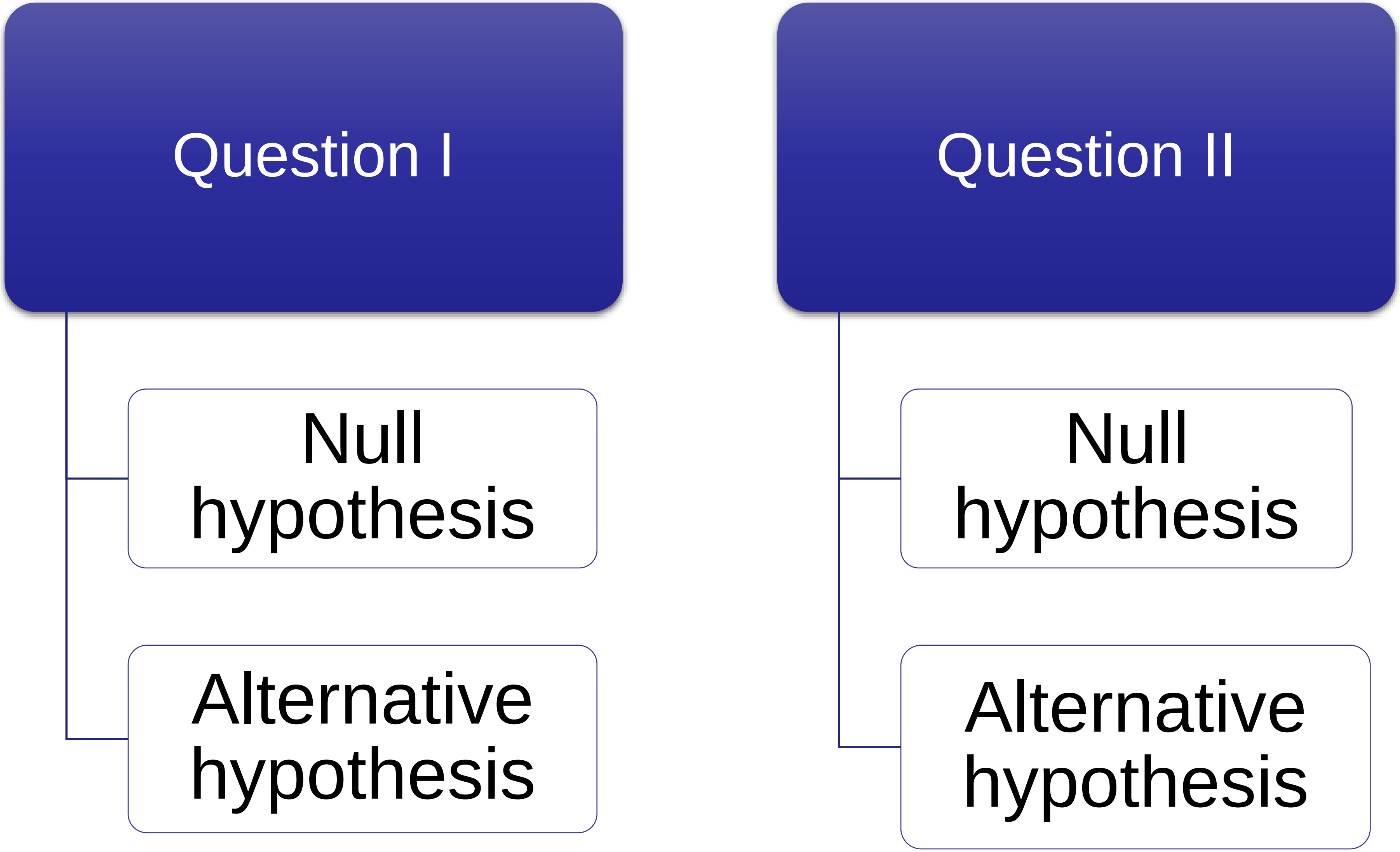
So much scientific questions!

- A **medical scientific question** can be very complex
  - What are the causes of prostate cancer?
  - Can the survival and quality of life of cancer patients be improved by better pain management?
  - What genetic influences are there on throat cancer?
  - Is there a difference in the methylation pattern between smokers and non-smokers?

So many hypotheses!

- A statistical hypothesis is very simple
  - Null hypothesis, H_0:
    - There is no difference between treatments
    - Stands for the **equality** of treatments
  - Alternative hypothesis, H_A or H_1:
    - There is a difference between the treatments
    - Stands for the **difference** of the treatments

Questions have hypothesis pairs



Question I

Null hypothesis

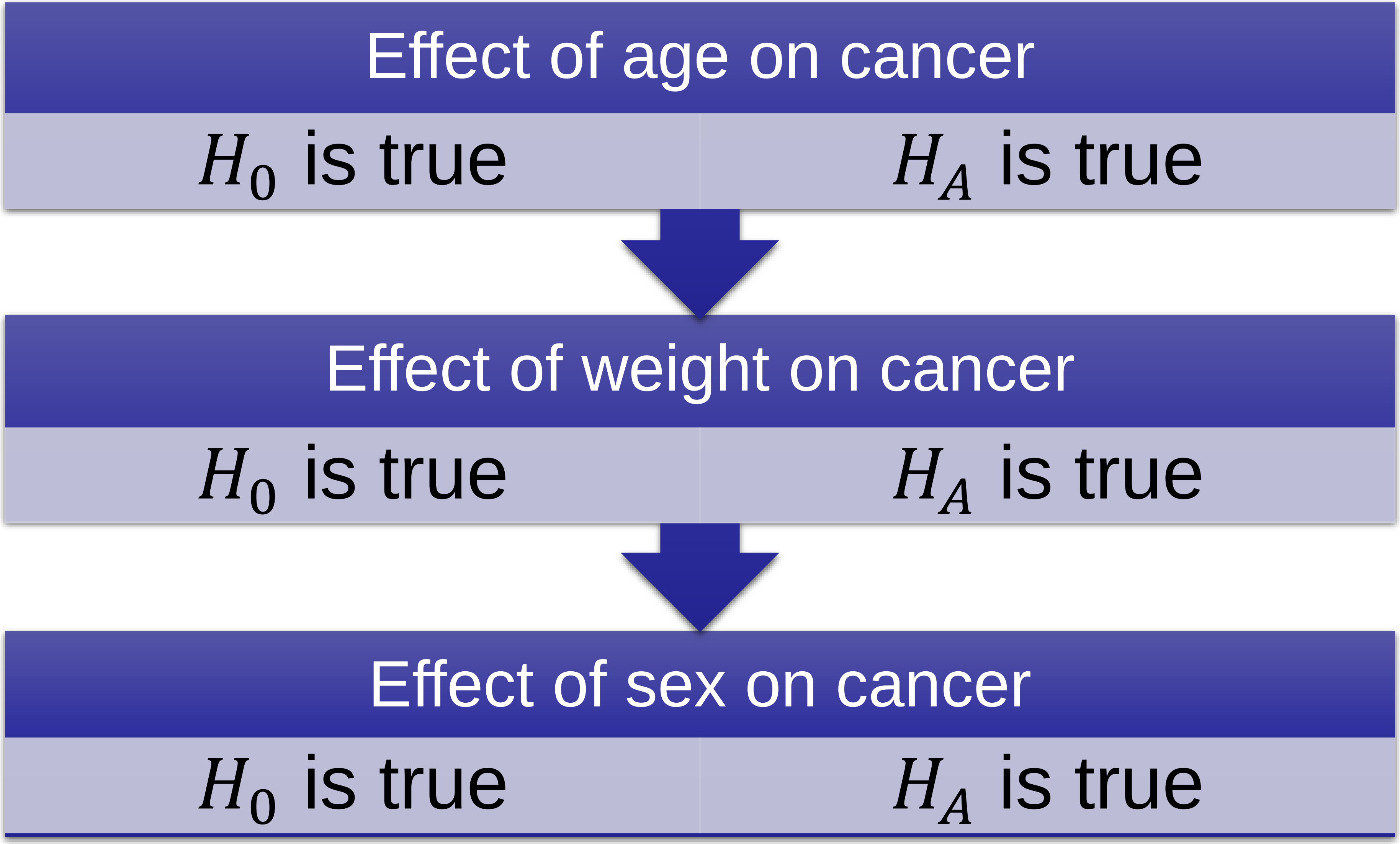
Alternative hypothesis

Question II

Null hypothesis

Alternative hypothesis

Every statistical test: a hypothesis



Effect of age on cancer

H_0 is true

H_A is true

Effect of weight on cancer

H_0 is true

H_A is true

Effect of sex on cancer

H_0 is true

H_A is true

Error 1. and 2. type

True Test decision	Null hypothesis applies	Research hypothesis applies
Null hypothesis is rejected	**Type 1 error** Probability max. α = level of significance ➔ *controlled by level of significance*	**Test decision correct** Probability 1-β = **Power/test quality**
Null hypothesis is retained	**Test decision correct**	**Type 2 error** Probability β ➔ *unknown, depending on unknown factors and the number of cases*

The ASA's Statement on p-Values: Context, Process, and Purpose

Q: Why do so many colleges and grad schools teach *p* = 0.05?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use *p* = 0.05?

A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: **"We teach it because it's what we do; we do it because it's what we teach."** This concern was brought to the attention of the ASA Board.

https://doi.org/10.1080/00031305.2016.1154108

The ASA's Statement on p-Values: Context, Process, and Purpose

Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. **Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect**. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

https://doi.org/10.1080/00031305.2016.1154108

The ASA's Statement on p-Values: Context, Process, and Purpose

A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Statistical significance is not equivalent to scientific, human, or economic significance. **Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect**. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

https://doi.org/10.1080/00031305.2016.1154108

The ASA’s Statement on p-Values: Context, Process, and Purpose

By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Researchers should recognize that a *p*-value without context or other evidence provides limited information. For example, a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, **a relatively large *p*-value does not imply evidence in favor of the null hypothesis**; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a *p*-value when other approaches are appropriate and feasible.

https://doi.org/10.1080/00031305.2016.1154108

Significance is not relevance!

- Whether a result is relevant in clinical practice or not cannot be seen from the p-value!

- The following is necessary to classify clinical relevance:

  – Estimated effect, e.g. relative risk, and associated confidence interval

  – Your assessment based on your medical experience and expertise

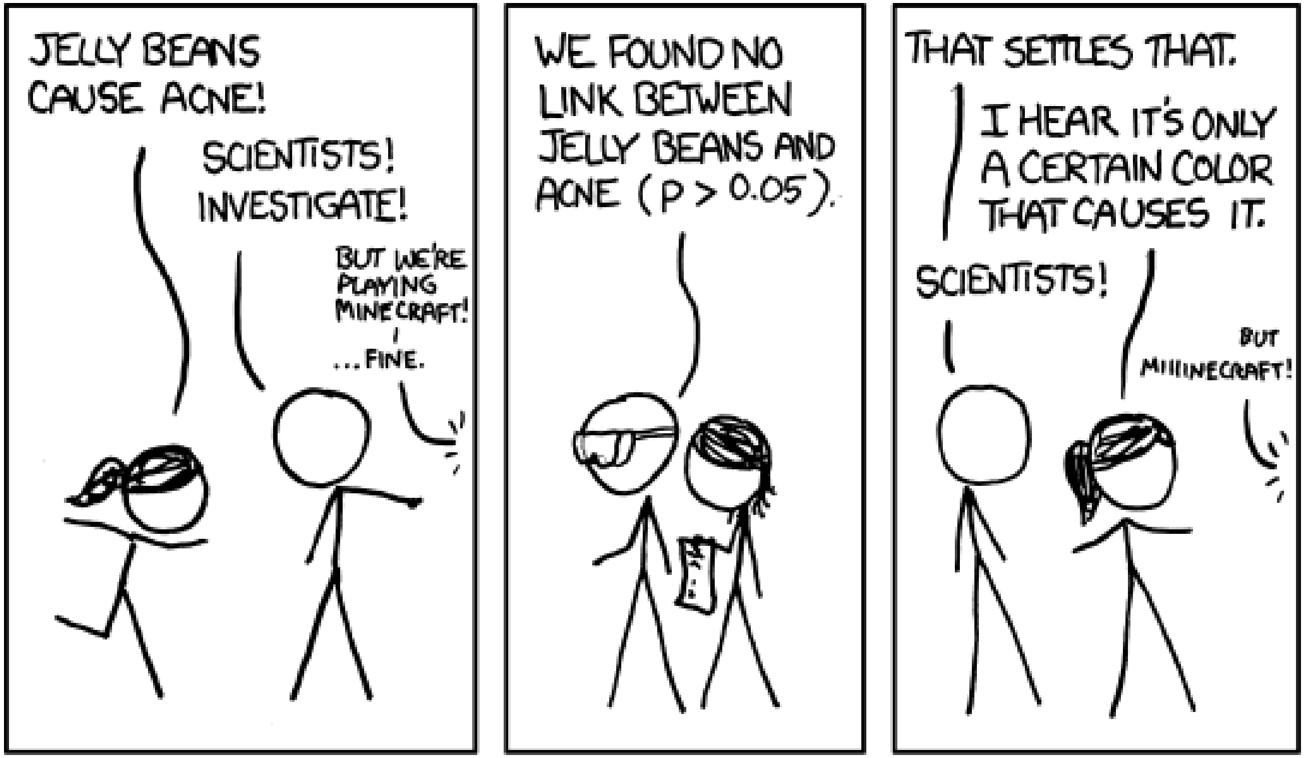
Key-Message 1:

When reporting the result of a statistical test, specify the p-value, the estimated effect, and the confidence interval for the effect.

Problems of multiple testing

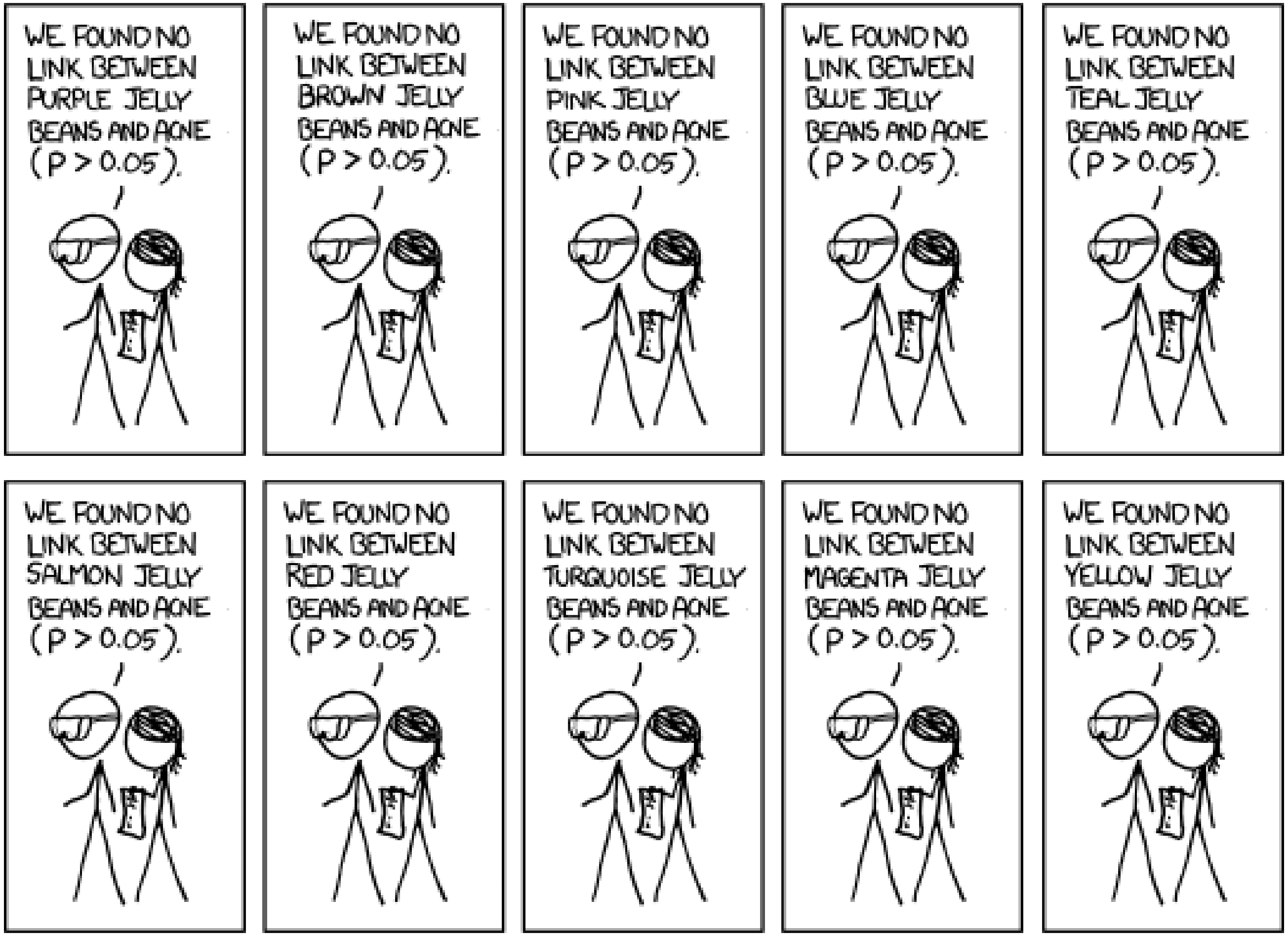
- **Wish**
  - Several questions are to be answered simultaneously within the same study
- **Inflation of α error or alpha error cumulation**
  - Simultaneous testing of multiple hypotheses leads to α error inflation
  - The probability that at least one null hypothesis is erroneously rejected is **no longer controlled** by the significance level α, but can become very high.

We test each pair of hypotheses with $\alpha = 0.05$


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

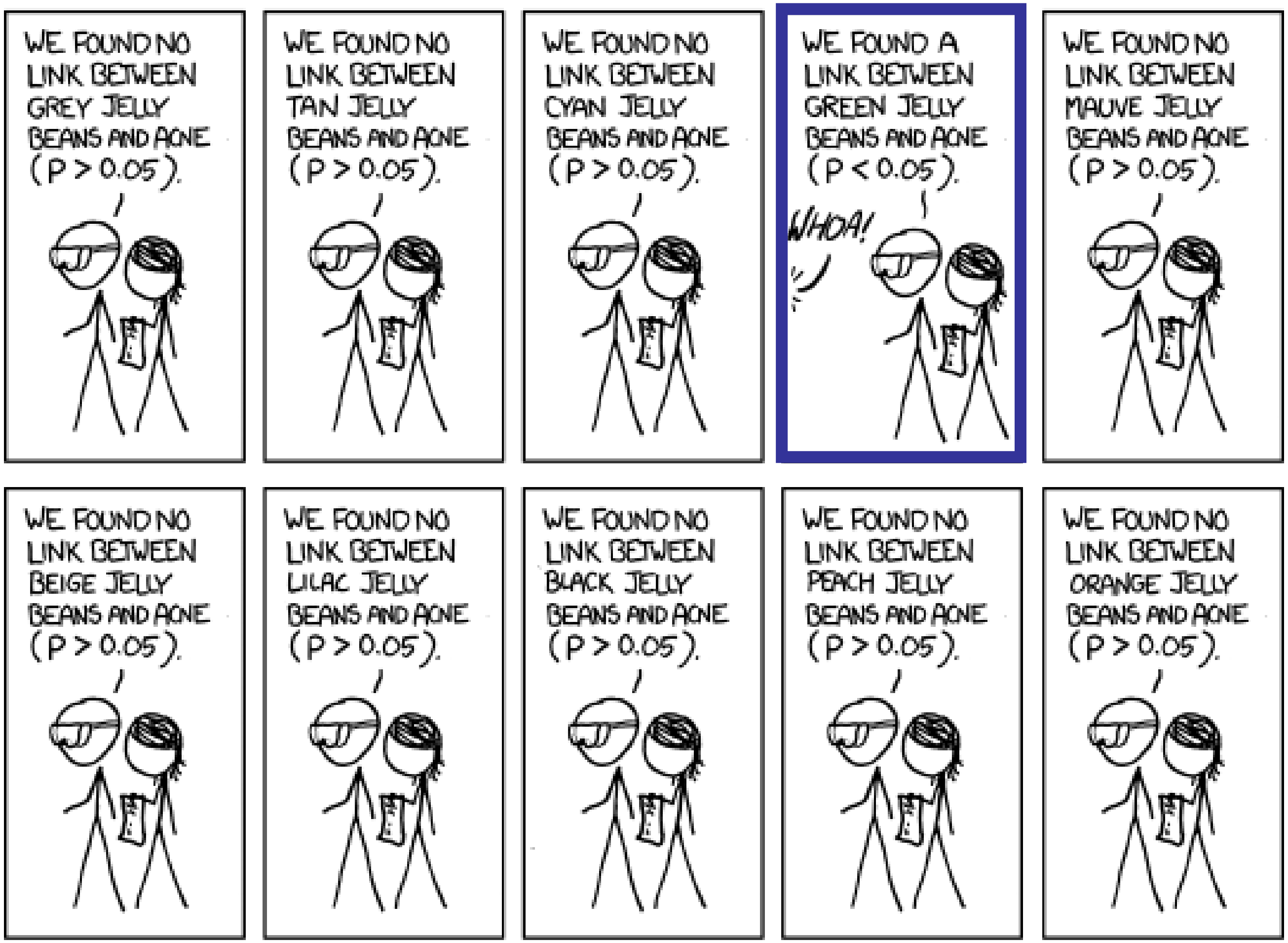
https://imgs.xkcd.com/comics/significant.png

We test each pair of hypotheses with $\alpha = 0.05$


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

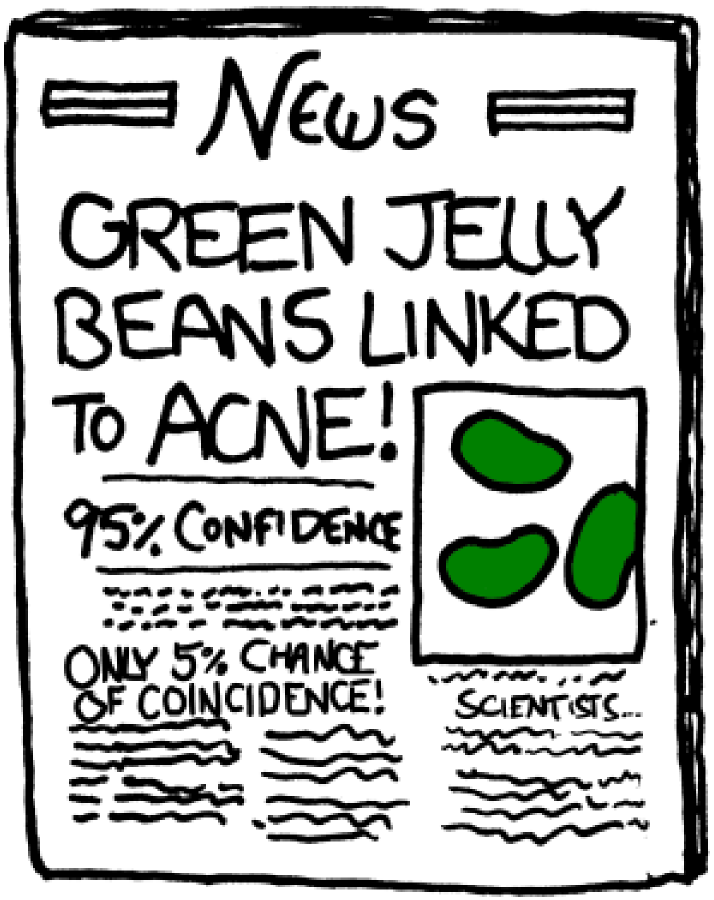
https://imgs.xkcd.com/comics/significant.png

We test each pair of hypotheses with $\alpha = 0.05$


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

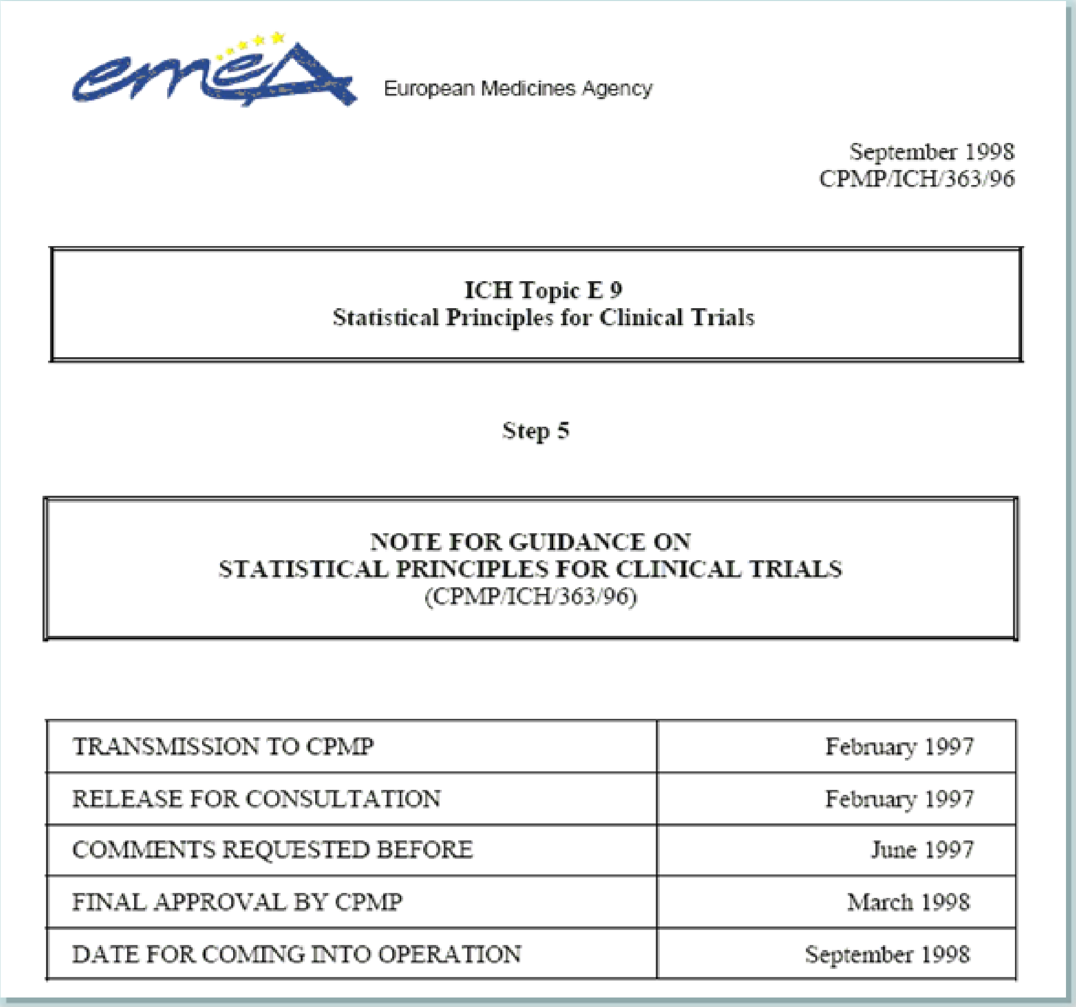
https://imgs.xkcd.com/comics/significant.png

We test each pair of hypotheses with $\alpha = 0.05$


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://imgs.xkcd.com/comics/significant.png

Medical guidelines



European Medicines Agency

September 1998
CPMP/ICH/363/96

ICH Topic E 9
Statistical Principles for Clinical Trials

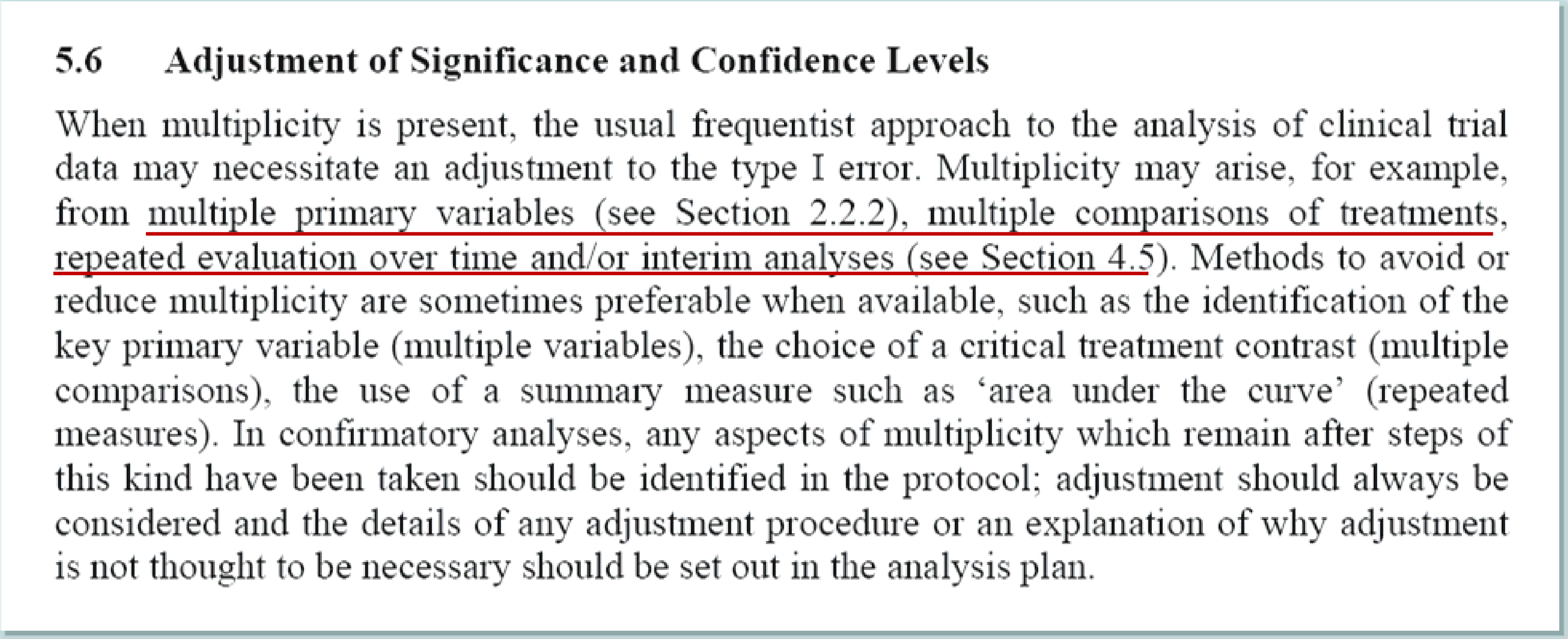
Step 5

NOTE FOR GUIDANCE ON
STATISTICAL PRINCIPLES FOR CLINICAL TRIALS
(CPMP/ICH/363/96)

TRANSMISSION TO CPMP	February 1997
RELEASE FOR CONSULTATION	February 1997
COMMENTS REQUESTED BEFORE	June 1997
FINAL APPROVAL BY CPMP	March 1998
DATE FOR COMING INTO OPERATION	September 1998

EMEA (1998): ICH Topic E9. Statistical Principles for Clinical Trials. CPMP/ICH/363/96, London.

Medical guidelines



5.6 Adjustment of Significance and Confidence Levels

When multiplicity is present, the usual frequentist approach to the analysis of clinical trial data may necessitate an adjustment to the type I error. Multiplicity may arise, for example, from multiple primary variables (see Section 2.2.2), multiple comparisons of treatments, repeated evaluation over time and/or interim analyses (see Section 4.5). Methods to avoid or reduce multiplicity are sometimes preferable when available, such as the identification of the key primary variable (multiple variables), the choice of a critical treatment contrast (multiple comparisons), the use of a summary measure such as 'area under the curve' (repeated measures). In confirmatory analyses, any aspects of multiplicity which remain after steps of this kind have been taken should be identified in the protocol; adjustment should always be considered and the details of any adjustment procedure or an explanation of why adjustment is not thought to be necessary should be set out in the analysis plan.

Inflation of the α error

Situation

➔ k Null and alternative hypotheses H_0^i

➔ all null hypotheses are tested to local level a=0.05

➔ in fact all null hypotheses are valid

➔ Acceptance of stochastically independent tests

Probability for a single test this correctly
to be rejected $(1-\alpha)$

Since the tests are independent, the probability is all k
correctly reject tests $(1-\alpha)^k$

The probability that at least one false null hypothesis will be rejected:

$$1-(1-\alpha)^k$$

Inflation of the α error

Probability that at least one null hypothesis is incorrectly rejected:

$$1 - (1 - \alpha)^k$$

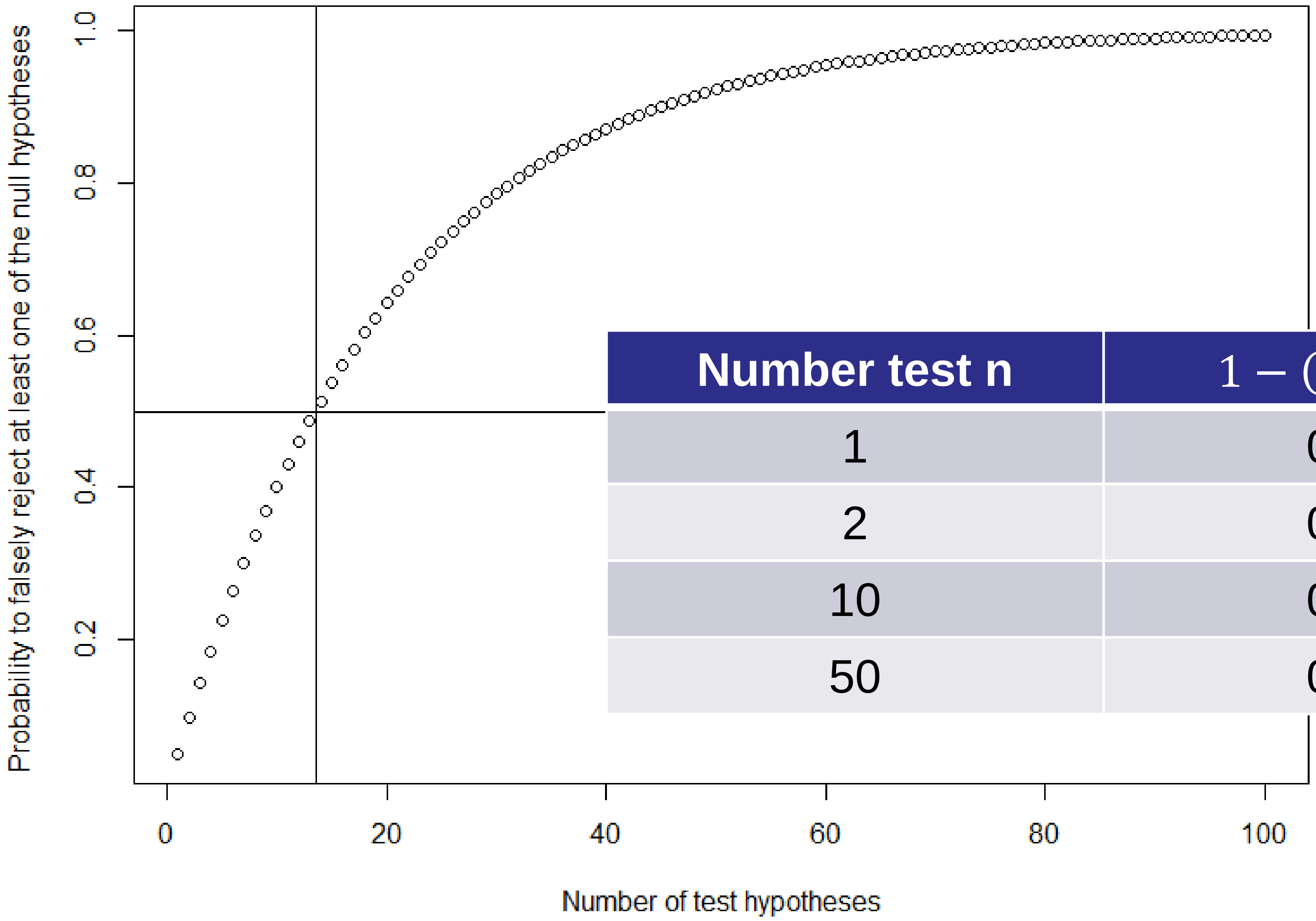
That means:

Number test n	$1 - (1 - \alpha)^k$
1	0,05
2	0,10
10	0,40
50	0,92

$\Rightarrow$ **Inflation of the α error**

If 50 hypotheses are tested, the probability of making at least one wrong test decision is almost 100%!

Inflation of the α error



Number test n	$1-(1-\alpha)^k$
1	0,05
2	0,10
10	0,40
50	0,92

Inflation of the α error

Expected number of incorrectly rejected null hypotheses:

$$\alpha \cdot k$$

That means:

Number test n	$\alpha \cdot k$
1	0,05
20	1
100	5
200	10

If 100 hypotheses are tested, 5 hypotheses are wrongly rejected on average.

Possibilities of correction

Bonferroni correction:

To ensure that the probability that at least one null hypothesis is incorrectly rejected is controlled by the global (and multiple) significance level α during simultaneous testing of k hypotheses, the individual hypotheses are tested for the local significance level $\boldsymbol{\alpha_{\text{lokal}} = \frac{\alpha}{k}}$.

Advantage: very easy to perform

Problem: is very conservative, i.e. the actual global (and multiple) level is clearly at α, i.e. the null hypotheses are too often maintained

Carlo Emilio Bonferroni wikimedia.org

Bonferroni Correction: Two Possibilities

- **Possibility I: Correction of the $\boldsymbol{\alpha}$ level**
  - The global α level is divided by the number of k statistical tests performed.
  - α/k = local α for the decision $\mathrm{p} < \alpha$

- **Possibility II: Correction of p-values**
  - The p-values are multiplied by the number of statistical tests carried out at k .
  - $p_{adjust} = p_{roh} * k$; with k equal to the number of comparisons
  - if $p_{adjust} > 1$, then p_{adjust} set to 1 (probability)

Bonferroni correction using an example

- **García-Arenzana et al. (2014)**
  - Association between food intake and Mammographic density (MD)
  - Risk factors for breast cancer
  - 25 risk factors representing a "food" were recorded

Dietary variable
Total calories
Olive oil
Whole milk
White meat
Proteins
Nuts
Cereals and pasta
White fish
Butter
Vegetables
Skimmed milk
Red meat
Fruit
Eggs
Blue fish
Legumes
Carbohydrates
Potatoes
Bread
Fats
Sweets
Dairy products
Semi-skimmed milk
Total meat
Processed meat

García-Arenzana, N. et al. (2014). Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925. https://doi.org/https://10.1002/ijc.28513

Alpha adjustment using an example

Bonferroni correction of alpha level

$$\alpha_{adj} = \frac{\alpha}{m} = \frac{0.05}{25} = 0.002$$

Dietary variable	Raw p-values	Bonferroni adjusted p-values
Total calories	**<0,001**	**0,025**
Olive oil	**0,008**	0,200
Whole milk	**0,039**	0,975
White meat	**0,041**	1,000
Proteins	**0,042**	1,000
Nuts	0,060	1,000
Cereals and pasta	0,074	1,000
White fish	0,205	1,000
Butter	0,212	1,000
Vegetables	0,216	1,000
Skimmed milk	0,222	1,000
Red meat	0,251	1,000
Fruit	0,269	1,000
Eggs	0,275	1,000
Blue fish	0,340	1,000
Legumes	0,341	1,000
Carbohydrates	0,384	1,000
Potatoes	0,569	1,000
Bread	0,594	1,000
Fats	0,696	1,000
Sweets	0,762	1,000
Dairy products	0,940	1,000
Semi-skimmed milk	0,942	1,000
Total meat	0,975	1,000
Processed meat	0,986	1,000

Alpha adjustment using an example

Dietary variable	raw p-values
Total calories	**<0,001**
Olive oil	**0,008**
Whole milk	**0,039**
White meat	**0,041**
Proteins	**0,042**
Nuts	0,060
Cereals and pasta	0,074
White fish	0,205
Butter	0,212
Vegetables	0,216
Skimmed milk	0,222
Red meat	0,251
Fruit	0,269
Eggs	0,275
Blue fish	0,340
Legumes	0,341
Semi-skimmed milk	0,942
Total meat	0,975
Processed meat	0,986

Bonferroni correction of alpha level

$$\alpha_{adj} = \frac{\alpha}{m} = \frac{0.05}{25} = 0.002$$

<u>Key-Message 2:</u>

A large list of unadjusted p-values is neither meaningful in terms of content nor statistically.

Alpha adjustment using an example

Bonferroni correction of alpha level

$$\alpha_{adj} = \frac{\alpha}{m} = \frac{0.05}{25} = 0.002$$

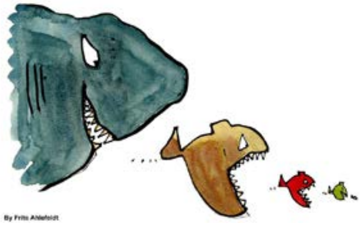
Dietary variable	raw p-values	Bonferroni adjusted p-values
Total calories	**<0,001**	**0,025**
Olive oil	**0,008**	0,200
Whole milk	**0,039**	0,975
White meat	**0,041**	1,000
Proteins	**0,042**	1,000
Nuts	0,060	1,000
Cereals and pasta	0,074	1,000
White fish	0,205	1,000
Butter	0,212	1,000
Vegetables	0,216	1,000
Skimmed milk	0,222	1,000
Red meat	0,251	1,000
Fruit	0,269	1,000
Eggs	0,275	1,000
Blue fish	0,340	1,000
[illegible]	0,341	1,000
Sweets	0,762	1,000
Dairy products	0,940	1,000
Semi-skimmed milk	0,942	1,000
Total meat	0,975	1,000
Processed meat	0,986	1,000

<u>Key-Message 3:</u>

In a confirmatory analysis, p-values must be adjusted.

Further options for correction

Hierarchically ordered hypotheses



➔ Determine a sequence of hypotheses before testing

➔ All hypotheses are tested at global significance level α

➔ Test until a null hypothesis can no longer be rejected.

➔ Cancel testing

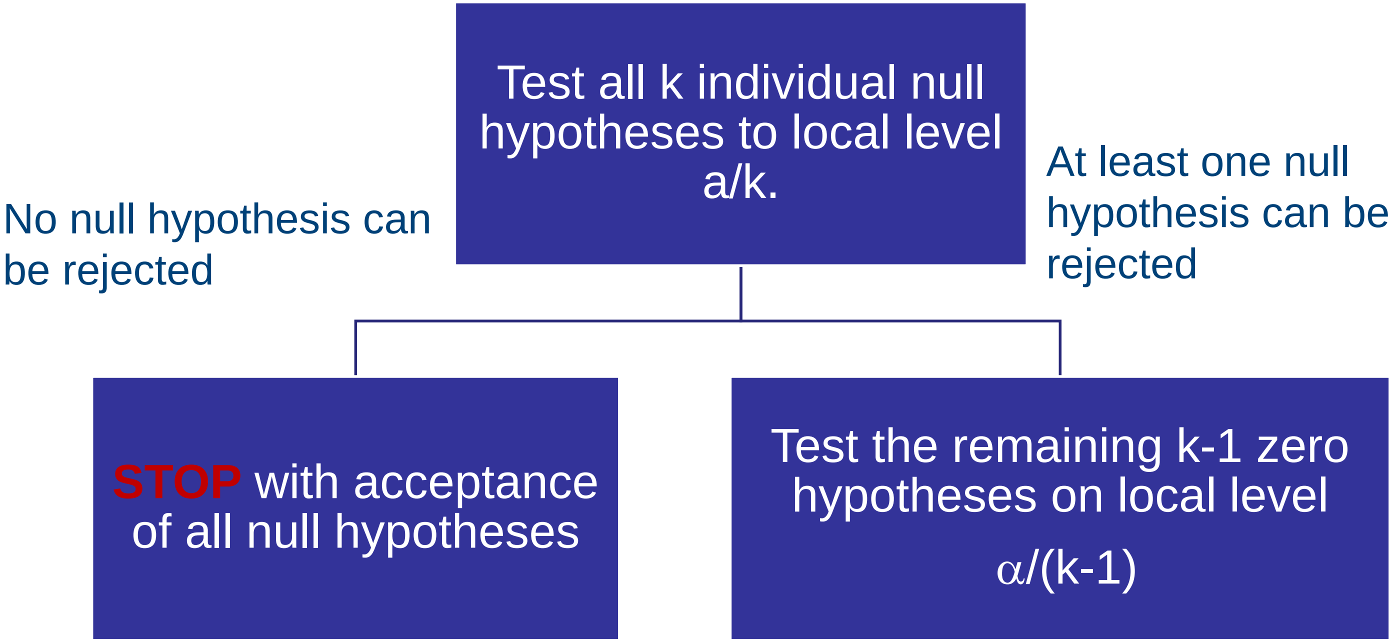
Problem: Sequence of hypotheses must be meaningful to interpret

Positiv: Less conservative than Bonferroni

There are many other arbitrarily complicated procedures to counteract α error inflation!

Source: https://pixabay.com/de/hierarchie-gruppe-essen-fisch-73335/

Bonferroni-Holm



Test all k individual null hypotheses to local level a/k.

No null hypothesis can be rejected

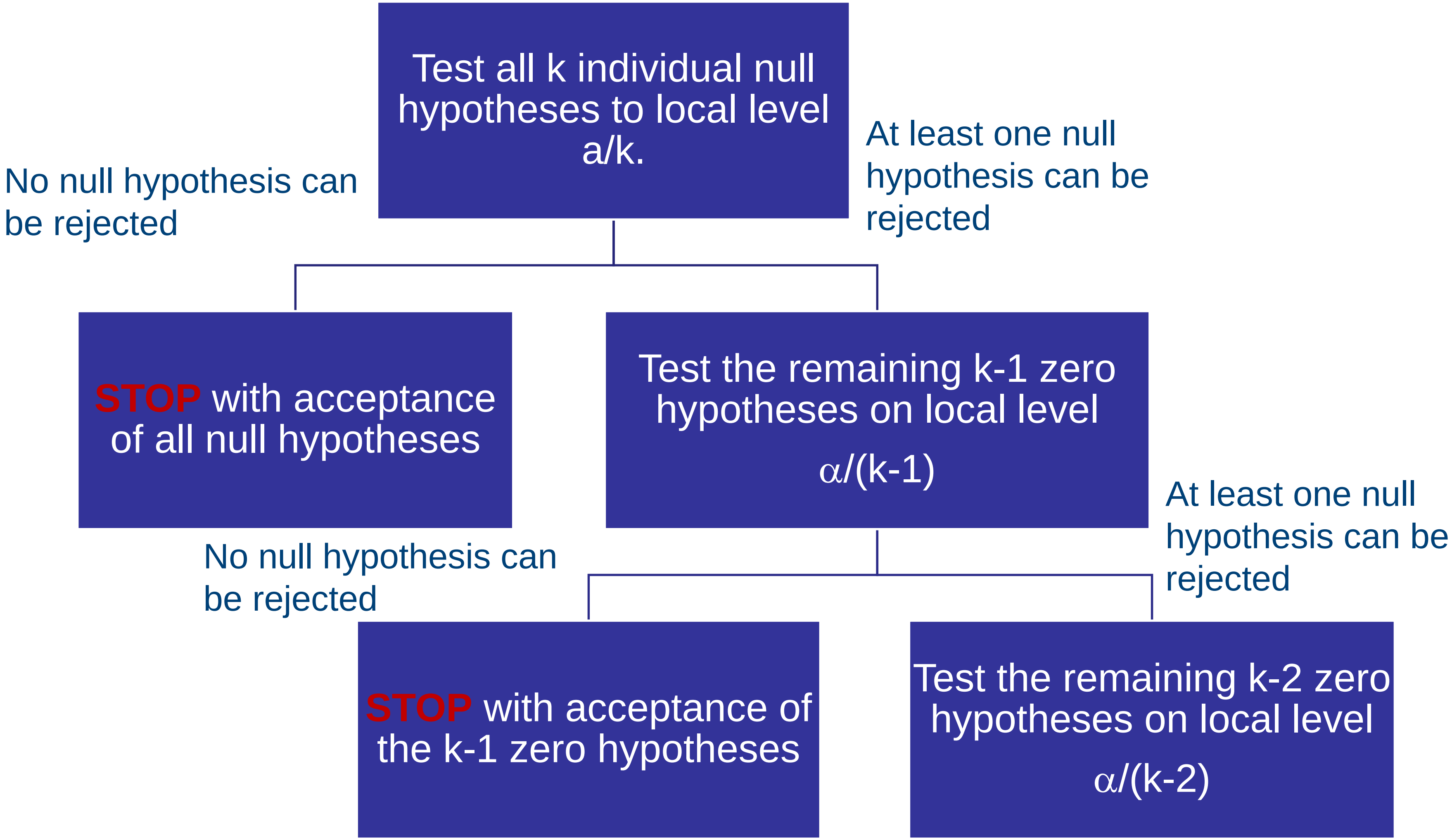
At least one null hypothesis can be rejected

STOP with acceptance of all null hypotheses

Test the remaining k-1 zero hypotheses on local level

α/(k-1)

Bonferroni-Holm



Test all k individual null hypotheses to local level a/k.

No null hypothesis can be rejected

At least one null hypothesis can be rejected

STOP with acceptance of all null hypotheses

Test the remaining k-1 zero hypotheses on local level

$\alpha/(k-1)$

No null hypothesis can be rejected

At least one null hypothesis can be rejected

STOP with acceptance of the k-1 zero hypotheses

Test the remaining k-2 zero hypotheses on local level

$\alpha/(k-2)$

Remarks on different adjustment methods

- The Bonferroni correction generally leads to an unnecessarily high number of cases.
- The presented procedures ignore a possible correlation of the test statistics.
- If the correlation is known, then the case number can be determined by the multivariate distribution, but there are usually no closed case number formulas for this anymore.
- If sequential test methods (such as Bonferroni-Holm) are used, the number of cases can only be determined by simulations. However, a significant reduction of the required falling number can often be achieved.

Descriptive p-values?

What does the p-value say? Actually not much...

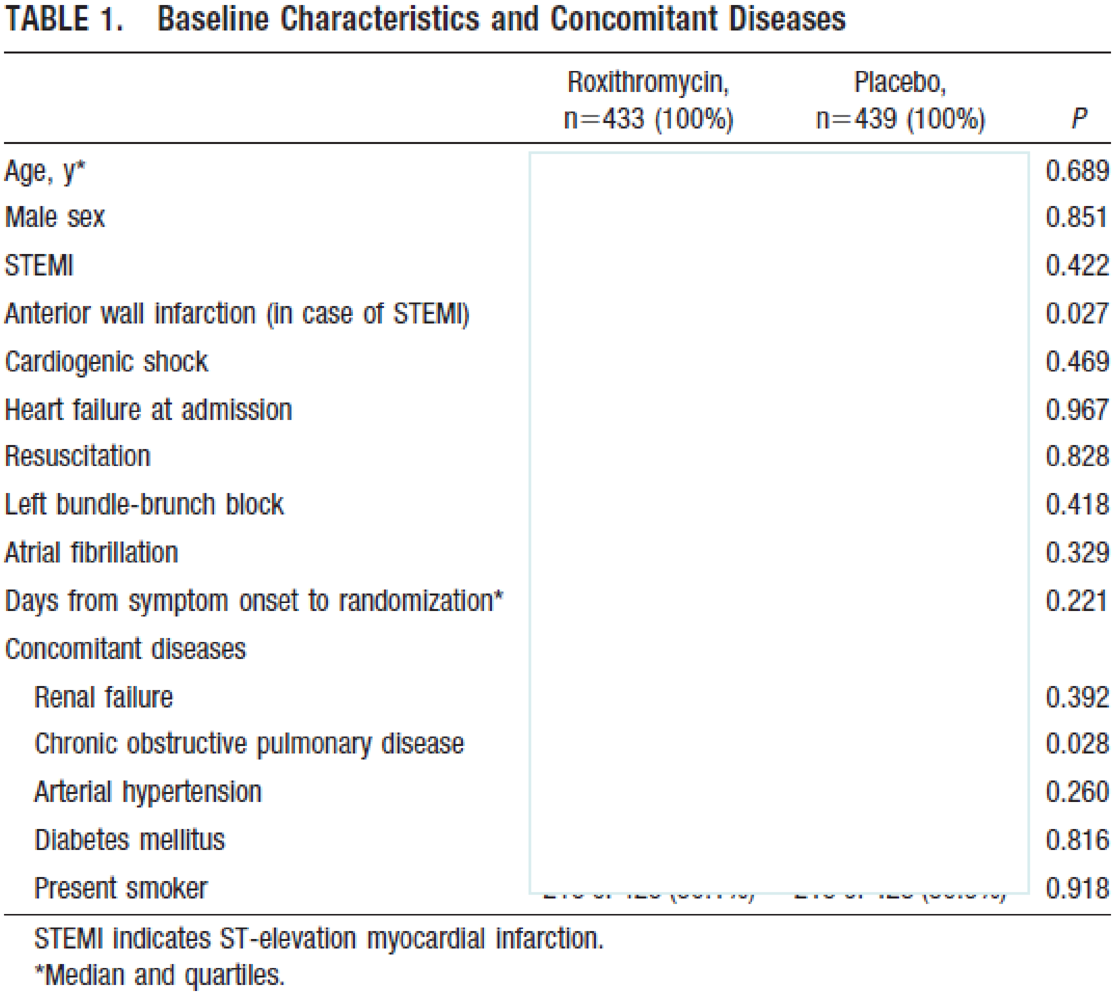


TABLE 1. Baseline Characteristics and Concomitant Diseases

	Roxithromycin, n=433 (100%)	Placebo, n=439 (100%)	*P*
Age, y*			0.689
Male sex			0.851
STEMI			0.422
Anterior wall infarction (in case of STEMI)			0.027
Cardiogenic shock			0.469
Heart failure at admission			0.967
Resuscitation			0.828
Left bundle-brunch block			0.418
Atrial fibrillation			0.329
Days from symptom onset to randomization*			0.221
Concomitant diseases			
Renal failure			0.392
Chronic obstructive pulmonary disease			0.028
Arterial hypertension			0.260
Diabetes mellitus			0.816
Present smoker	[illegible]	[illegible]	0.918

STEMI indicates ST-elevation myocardial infarction.
*Median and quartiles.

Descriptive p-values?

What does the p-value say? Actually not much... p

I'm mysterious.

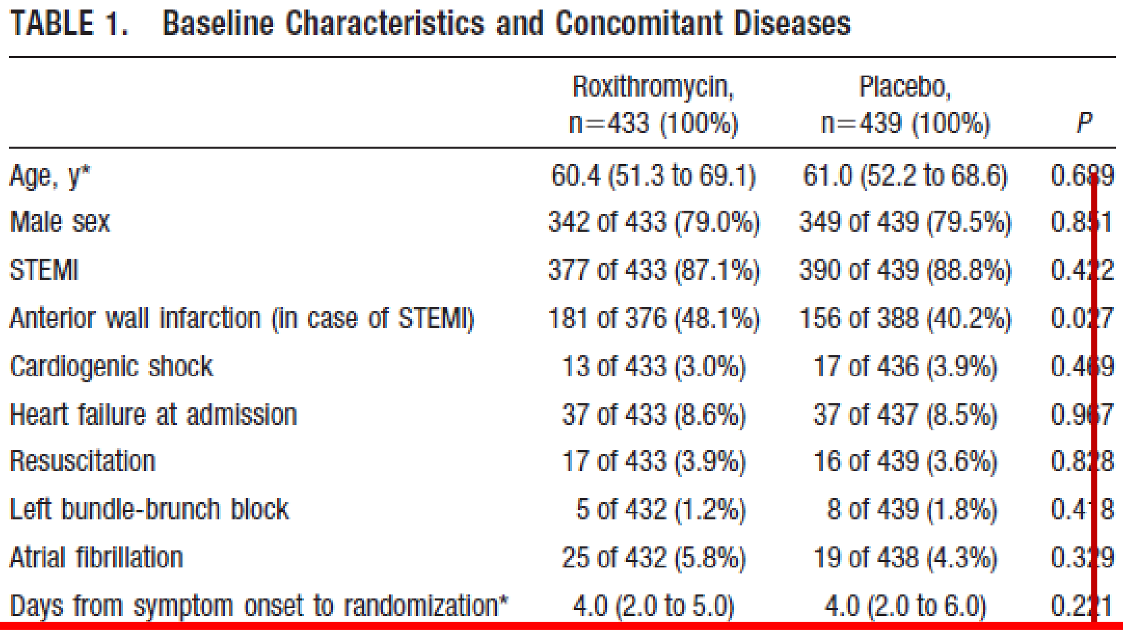


TABLE 1. Baseline Characteristics and Concomitant Diseases

	Roxithromycin, n=433 (100%)	Placebo, n=439 (100%)	*P*
Age, y*	60.4 (51.3 to 69.1)	61.0 (52.2 to 68.6)	0.689
Male sex	342 of 433 (79.0%)	349 of 439 (79.5%)	0.851
STEMI	377 of 433 (87.1%)	390 of 439 (88.8%)	0.422
Anterior wall infarction (in case of STEMI)	181 of 376 (48.1%)	156 of 388 (40.2%)	0.027
Cardiogenic shock	13 of 433 (3.0%)	17 of 436 (3.9%)	0.469
Heart failure at admission	37 of 433 (8.6%)	37 of 437 (8.5%)	0.967
Resuscitation	17 of 433 (3.9%)	16 of 439 (3.6%)	0.828
Left bundle-brunch block	5 of 432 (1.2%)	8 of 439 (1.8%)	0.418
Atrial fibrillation	25 of 432 (5.8%)	19 of 438 (4.3%)	0.329
Days from symptom onset to randomization*	4.0 (2.0 to 5.0)	4.0 (2.0 to 6.0)	0.221

Key-Message 4:

In a descriptive analysis, p-values are not very meaningful. A descriptive analysis should be limited to measures of location and dispersion and confidence intervals.

STEMI indicates ST-elevation myocardial infarction.
*Median and quartiles.

!

Recommended reading

e 26 Statistik | Statistics

Multiples Testen

– Artikel Nr. 12 der Statistik-Serie in der DMW –

Multiple testing

Autoren R. Bender[1] S. Lange[1] A. Ziegler[2]

Institut [1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln
[2] Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck

- Bender R, Lange S, Ziegler A (2007): Multiples Testen, *Dtsch Med Wochenschr,* 132:e26-e29.

  https://www.thieme-connect.com/ejournals/pdf/dmw/doi/10.1055/s-2007-959035.pdf (Stand Juli 2011)

Recommended reading

- Victor, A.; Elsäßer, A.; Hommel, G.; Blettner, M.: Wie bewertet man die p-Wert-Flut? Hinweise zum Umgang mit dem multiplen Testen. 2010 [online]. DOI: https://doi.org/10.3238/arztebl.2010.0050; Deutsches Ärzteblatt, Jg. 107, Heft 4, S.50–56
- Groß, Marcus: Multiples Testen und Confirmation Bias. last revised on 31.07.2018 [online]. URL: https://wikis.fu-berlin.de/display/fustat/Multiples+Testen+und+Confirmation+Bias
- Deutsches Netzwerk Evidenzbasierte Medizin: Glossar zur Evidenzbasierten Medizin. 2011 [online]. URL: https://www.ebm-netzwerk.de/pdf/publikationen/dnebm-glossar-2011.pdf
- Health Bridge Limited (t/a DrEd): Systolischer Blutdruck [online]. URL: https://www.dred.com/de/systolischer-blutdruck.html
- Black, K.: 10. Calculating p Values. 2015 [online]. URL: https://www.cyclismo.org/tutorial/R/pValues.html#calculating-many-p-values-from-a-t-distribution
- Yau, C.: R Tutorial. An R Introduction to Statistics. Pt. [online]. URL: http://www.r-tutor.com/category/r-functions/pt
- McDonald, J.H. (2014). Handbook of Biological Statistics. Multiple comparisons. last revised on 20.07.2015 [online]. URL: http://www.biostathandbook.com/multiplecomparisons.html; This web page contains the content of pages 254-260 in the printed version. Sparky House Publishing, Baltimore, Maryland.
- García-Arenzana, N. et al. (2014). Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925. https://doi.org/https://10.1002/ijc.28513