INSELSPITAL
HOPITAL DE L'ILE

UNIVERSITÄT
BERN

Prof. em. Martin Fey

Clinical trial statistics

- Significance (or is it clinical relevance?)
- Power to the trial (and to the people!)
- Hazard ratios (no need for insurance)
- Absolute AR and relative risk RR
- Numbers needed to treat NNT
- Confidence (not in Trump but in intervals)
- ITT or PP analysis?

Clinical trial statistics for uninterested dummies

Null hypothesis H0	The new treatment produces **NO** effect. Usually trialists want an effect and thus hope to reject the null hypothesis H0.
Type I error α	H0 is **falsely** rejected (i.e. the new treatment is truly no better). The result is **false positive.**
α (type I error rate)	arbitrary level usually set @ 0.05 = the probability of falsely rejecting H0 is $\leq$ 5%. Hence, the rate of false positive results $\leq$ 5%.

Type I error, or p-value < 5% or < 0.05

- One in 20 experiments will yield a SIGNIFICANT result just by chance.
- If you look at a wide range of potential outcomes, false positive results are virtually guaranteed.
- Medicine is dominated by type I error rates with little consideration of type II errors (see in a minute)
- P < 0.01 would mean fewer positive trials with larger, perhaps more clinically relevant effect sizes

Clinical trial statistics for uninterested dummies

Null hypothesis H0	The new treatment produces **NO** effect. Usually trialists want an effect and thus hope to reject H0 (equipoise).
Type II error β	H0 is in fact false (i.e. the new treatment has **indeed a positive** effect). H0 is not rejected although it should have been. The result is **false negative.**
$1-\beta$ (power of test)	is the statistical power to address H0 correctly.

Type II error or power of a trial

- The accepted level of power (1-β) is 80% which is totally arbitrary.
- Higher power would mean accrual of (considerably) more trial patients, longer trial duration, a longer waiting time before the trial chairman can grab a microphone @ ASCO, and delays in getting the drug to market and making lots of money.
- Nobody wants all that. Power stays @ 80%.

α or β – which is which?

- Type I al**p**ha false **p**ositive
- Type II b**e**ta false n**e**gative

Eselsbrücke «donkey bridge»

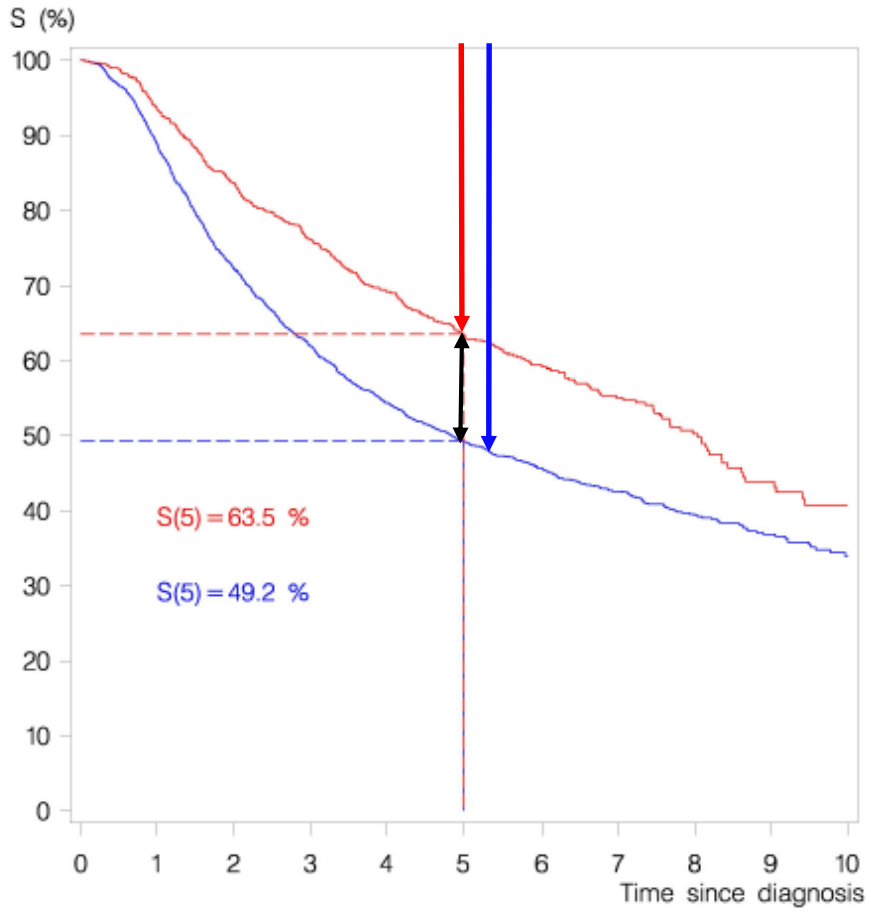
mnemonic aid

Mnemonyse Goddess of memory

Hazard ratios yield «survey data»

- HR, or **relative hazards** are typically used to compare time-to-event data between two treatment groups (new vs standard).
- HR 1 = no difference i.e. no Δ.
- HR may not be constant during follow-up.
- **However**, one assumes the HR represents the risk of an event in the experimental group compared with the risk of that event in the standard group **at any time** during the study period.

Absolute and relative risks



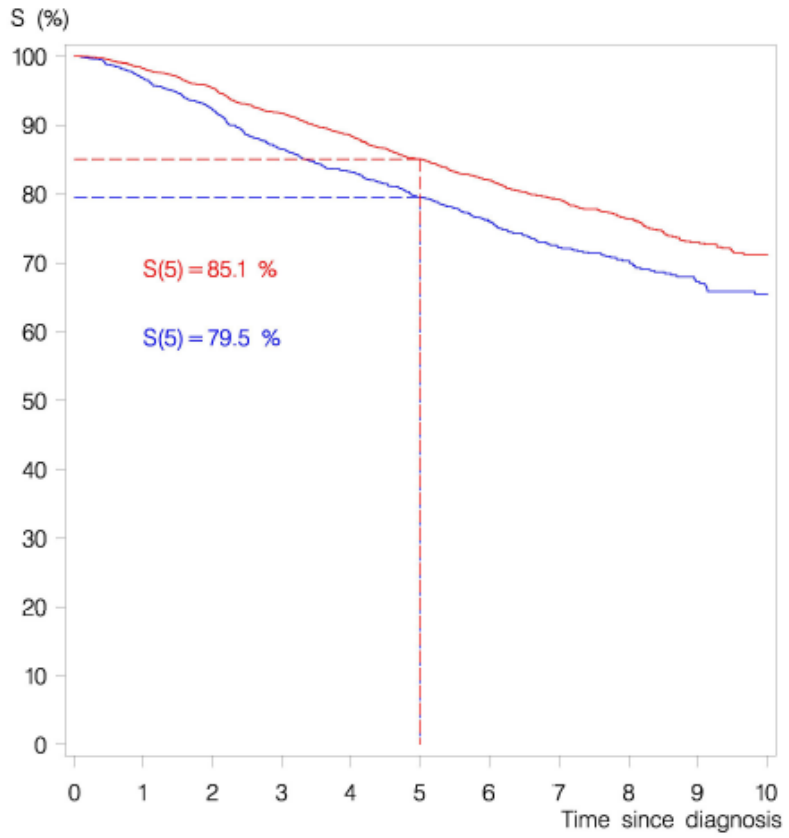
ARR = Absolute Risk Reduction
(100%- 49.2%)-(100%-63.5%)
= 14.3%

RR = Relative Risk at 5 yrs
(100-63.5)/(100-49.2) = 0.72

i.e.
the relative risk reduction is 28%

36.5/50.8 = 0.72 = 72%
100% - 72% = **28%**

Risks and numbers needed to treat



Absolute risk reduction
85.1%-79.5% = 5.6%

Relative risk reduction 27%

Number needed to treat
The number of pts who need to be treated to prevent one additional event

NNT = 1/ARR = 1/0.056 = 18

Absolute versus relative risk (reduction)

PARP INH for BRCA POS breast cancer (NEJM 2018)

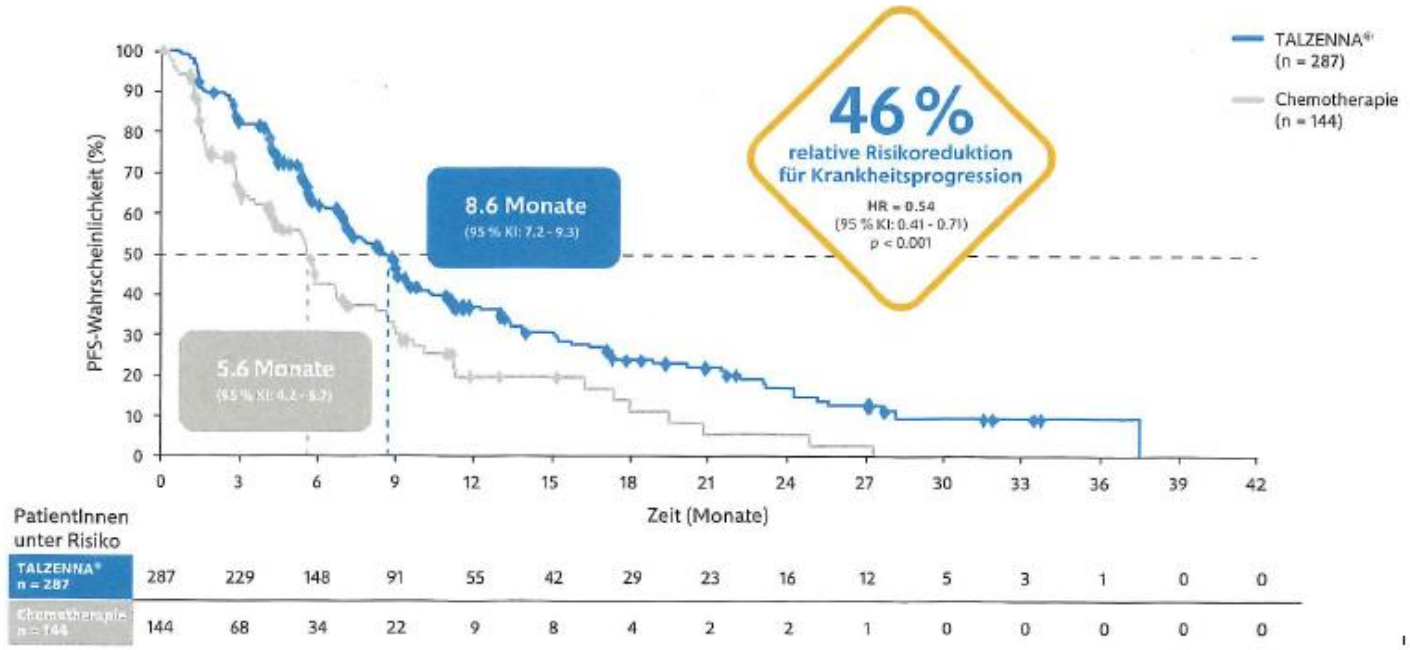


Abb. 1: Signifikante Verlängerung des medianen progressionsfreien Überlebens unter Talazoparib im Vergleich zur Standard-Chemotherapie mit Capecitabin, Eribulin, Gemcitabin oder Vinorelbin. Adaptiert nach (2)

PFS
Radiologic PFS

HA 0.54 (CI 0.4-0.7)
= 54%
RRR = 46%
ABS RR = 15%

NNT
ABS RR 1/0.15 = **7**
6 mts tx x 7 women
= CHF 252'000.- to get a nicer X-ray

95% confidence interval CI

- We estimate the magnitude of the difference between treatments on patient outcomes.
- Firstly, we obtain a point estimate, that is, the actual difference observed.
- Then we express the degree of uncertainty present in the data. The bigger the trial, the more precise the point estimate will be. Such uncertainty is usually expressed as a 95% CI.

Confidence intervals CI

- The 95% CI represents the inaccuracy of the sample in estimating the endpoint chosen.
- If the sample size of the trial was increased, the width of the 95% CI would decrease.
- A 95% CI is a trade off - a smaller % would not provide an interval estimate with enough certainty of including the endpoint, whereas a greater % would present an interval estimate that was too wide to be of practical benefit.

ITT versus Per Protocol

ITT

- Patient groups are maintained as allocated by randomisation
- Treatment analysis is done regardless of whether a patient started treatment, received modified treatment, or dropped out of the study at some stage
- **This reflects clinical practice!**
- ITT should be the standard trial analysis technique.

PP

- Analysis includes only those patients who completed the protocol treatment as assigned and planned («pp»).
- This reflects the treatment effect unaffected by protocol deviations or non-adherence.
- The original comparability of the treatment groups may not be maintained.
- Per protocol analysis is not illegal!

Intention-to-treat analysis, or Intention-to-analyse … treatment?

- **Per protocol**: analyse all patients who fulfill the protocol in terms of eligibility, treatment received (interventions), and outcome assessment

- **Per intention-to-treat**: analyse all eligible patients no matter what happend to them, or what doctors did to them

Randomised clinical trials

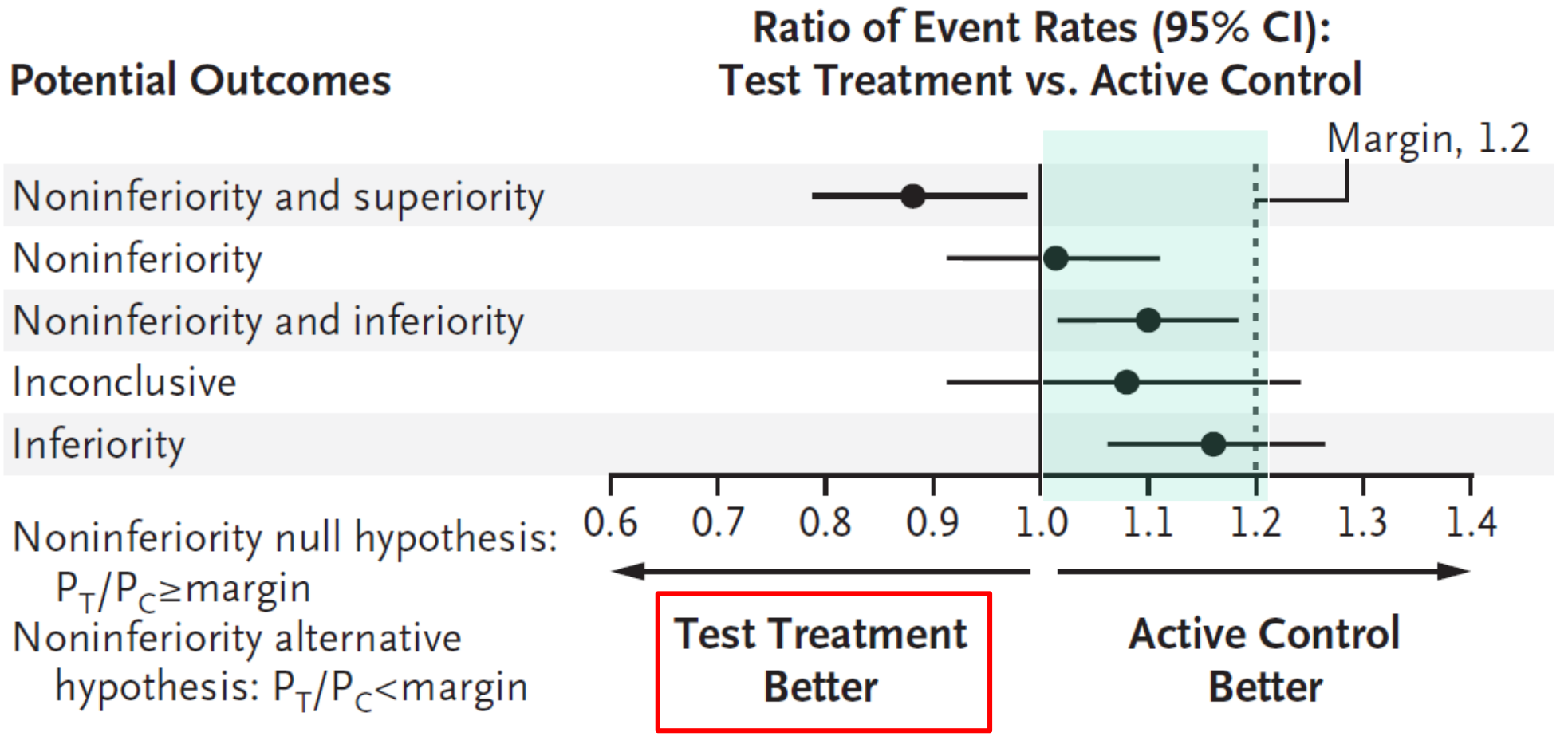
- Superiority design — **New** is better
- Non-inferiority design — **New** is at least not worse,

  and

  New has an added bloody obvious advantage

The non-inferiority margin

NEJM 2017; 377: 1357



Superiority Δ
Non-inferiority margin Δ

- Δ = acceptable difference stated **by clinical consensus** among investigators
  (absolute risk Δ, hazard ratio, odds ratio)
- Superiority: minimum of Δ expected (or more)
- Non-inferiority: maximum Δ accepted (or less)
- Non-inferiority margins in trials may vary widely (for example, absolute risk Δ between 0.4% to 25%, or HRs between 1.05 to 2.85).

Final summary

- Judge the Δ in superiority and non-inferiority trials based on your practical clinical experience and your oncological gut feeling.
- $P < 0.05$ and power @ 80% are (unfortunately) here to stay.
- Look at the direct clinical relevance of the primary endpoint.
- In papers reporting trials with an adaptive design it won't do just to read the abstract. If you do, don't believe what the authors summarise.