GRASP

APPROFONDIMENTO

Analisi per sottogruppi

Paolo Bruzzi, Genova

Analysis of RCTs: Reference criteria

STATISTICAL

PREDETERMINED ANALYSES

METHODOLOGICAL

INTENTION TO TREAT

Analysis of RCTs: Reference criteria

STATISTICAL

All statistical analyses must be explicitly predetermined (endpoint, transformations, test, timing, subgroups)

MULTIPLICITY

Multiplicity

With an increasing number of analyses, the probability of finding, **BY CHANCE**, some noteworthy difference increases

- Five consecutive reds at the roulette wheel
- Two cases with the same inherited mutation
- Three long-term survivors with advanced NSCLC

Multiplicity

If I look for any possible treatment effect, **BY CHANCE**,
I will always find some difference:

- Overall mortality, cause-specific mortality (50 causes)
- QoL (six different domains)
- Incidence of AEs (50 possible) and favorable events (50 possible)

If, afterward, I focus on the one(s) showing a difference,
I can always demonstrate that a treatment is effective (less toxic, etc.)

Possible sources of multiplicity in a clinical trial

Multiple endpoints

Transformation of endpoints
(e.g. cumulative incidence at 3, 5, 10 years vs survival curve)

Statistical test

Summary effect measure

Missing data

Interim analyses

Subgroup analyses

Critical distinction: Planned multiple tests vs data-derived tests (post hoc analyses)

PLANNED MULTIPLE TESTS

- Predetermined (study protocol)
- Finite number
- Statistical correction possible

POST HOC ANALYSES

- Number potentially infinite
  - The observation of an association induces a test of significance
  - Intensive crosstabulations in search of associations
- **Lack of any statistical rationale/validity**

Multiplicity: General rules

- Before the start of the study, the number, time, and types of analyses are declared
  - E.g. two analyses in subjects <50 or >50 years old, or three analyses after 100, 200, and 300 events (final)
- A set of rules is established to decide if the study has led to a positive result (or to stop the study)
- These rules are built in such a way that the **overall** probability of an α error is the desired one (e.g. 5%)

Analysis of results: Strategy

- **Primary analysis (p value)**
- **Interim analyses**
- **Secondary analyses**
  - **Other endpoints**
  - Subpopulations
  - **Subgroups (interactions)**
  - Multivariate analyses

PROTOCOL

For each analysis, the statistical plan establishes the statistical method to be used, when and how it will be conducted, and the decision rules (stop/go, positive/negative, p levels, etc.)

Methods to control multiplicity

Adjusted p values

Hierarchical test procedures

Closed test procedures

Adjusted p values

NOTE

Adjusted p values

Alpha spending function

Alpha split

Others

SAME MEANING

In each analysis, a p value <0.05 is used to ensure that the overall probability of a "significant" result (rejecting the null hypothesis when true) is <5%

Adjusted p values

MODIFIED SIGNIFICANCE LEVELS (all <α) that make the overall probability of a false-positive result equal to the desired α level (usually 5%)

EXAMPLE

If:

- Desired significance level = 0.05 (5%)
- Four analyses are planned
- Study is positive if in at least one of these four analyses p <0.05/4 = 0.0125 (Bonferroni)

NB the correct formula is:

Significance level for n analyses at the overall significance level p:
required p at each analysis = $1-(1-0.05)^{1/n}$

Therefore, for n=4 and p=0.05, the required p is 0.127

Adjusted p values

SINGLE-STEP METHODS

Examples: Bonferroni, Simes, Dunnett…

STEPWISE METHODS

The rejection or non-rejection of a particular hypothesis may depend on the decision made on other hypotheses

Examples: Holm, Hochberg, step-down Dunnett…

Adjusted p values

ALL METHODS INVOLVE A LOSS OF STATISTICAL POWER

- Limit the number of analyses
- Increase the sample size
- Use hierarchical procedures
- The Bonferroni method is the most conservative (least powerful)
  - Other more complex methods generally used

Methods to control multiplicity

Adjusted p values

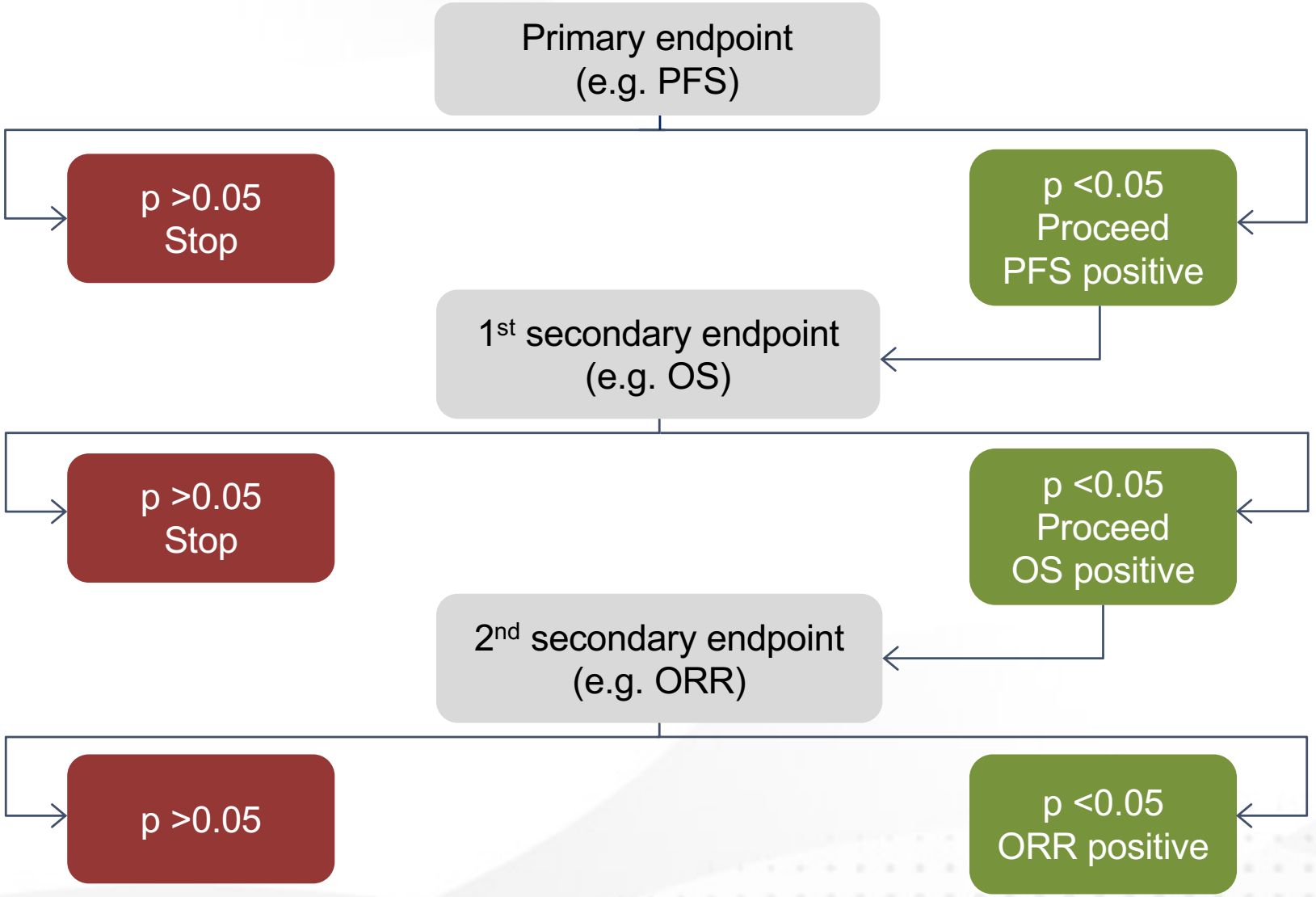
Hierarchical test procedures

Closed test procedures

Hierarchical test procedures

- Hypotheses are ordered in sequence and tested at level α until the first non-rejection
- In practice, first test:
  - If p >0.05 → stop (negative study)
  - If positive → second test
- **NO CORRECTION OF THE SIGNIFICANCE LEVEL IS REQUIRED**
- Sequence based on relevance, power, plausibility, etc.

Hierarchical test procedures (example)



Hierarchical test procedures

- No loss of power for the first analysis
- Frequently used for multiple endpoints
- Risk of missing relevant treatment effects if the order of tests is incorrect (e.g. OS then PFS)

Methods to control multiplicity

Adjusted p values

Hierarchical test procedures

Closed test procedures

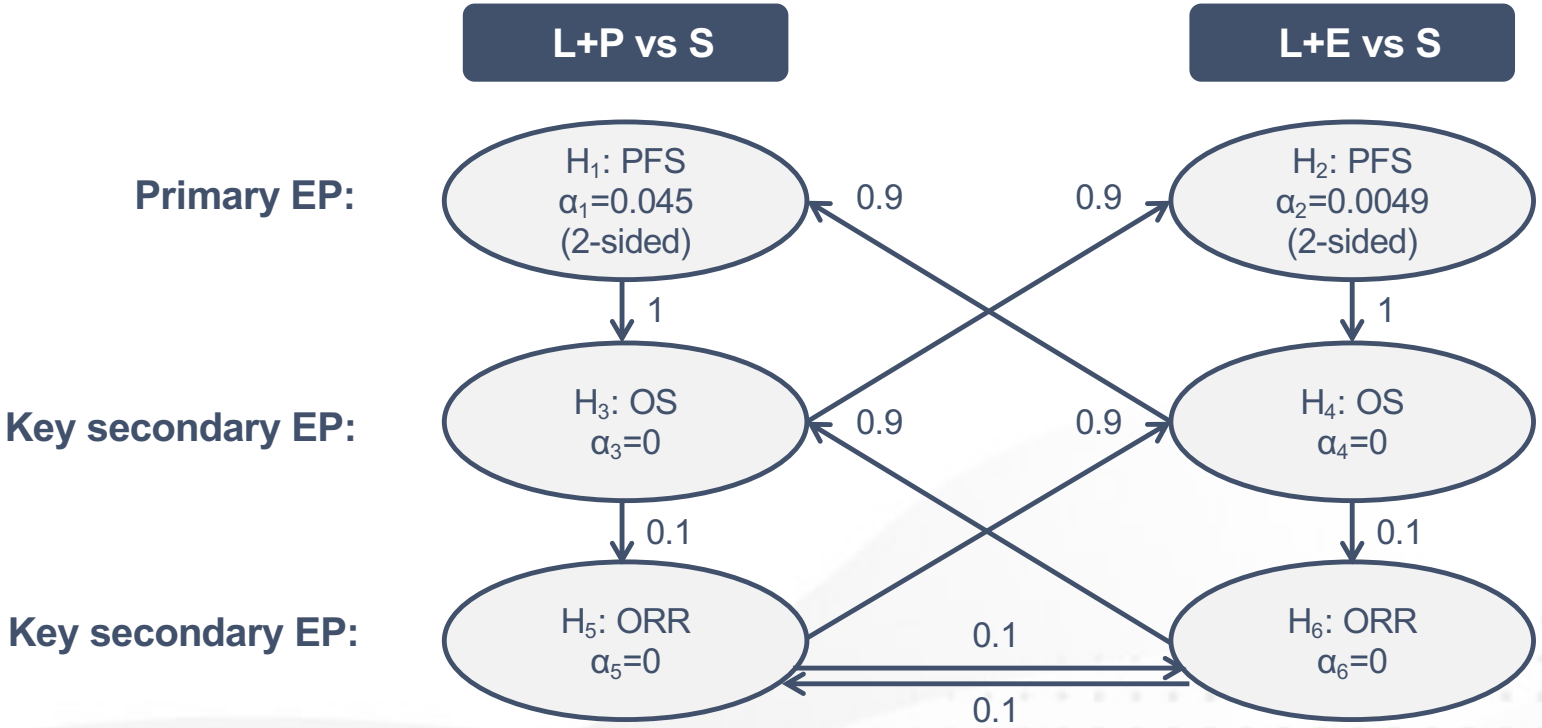
Closed test procedures

- General principle to build procedures for multiple tests
- Used to protect the α error while maintaining efficiency (reduced loss of power)
- Many of the previously mentioned procedures (e.g. Holm, hierarchical) are based on this principle
- Widely used: Maurer & Bretz graphical method

Statistical Plan AMENDMENT 07 (2020)

MULTIPLICITY

To adjust for multiplicity and control the overall FWER, the graphical approach of Maurer and Bretz (Maurer et al., 2013) will be used in the primary endpoint of PFS and the key secondary efficacy endpoints (OS and ORR). No multiplicity adjustment will be made for other secondary endpoint analyses.



Mod. da Eisai. Clinical Study Protocol E7080-G000-307, Protocol Amendment 07: 06 August 2020

Analysis of results: Strategy

- **Primary analysis (p value)**
- **Interim analyses**
- **Secondary analyses**
  - **Other endpoints**
  - Subpopulations
  - **Subgroups (interactions)**
  - Multivariate analyses
- Unplanned analyses: **merely exploratory aims**
  - **Used to plan other studies**

PROTOCOL

Possible sources of multiplicity in a clinical trial

Multiple endpoints

Transformation of endpoints
(e.g. cumulative incidence at 3, 5, 10 years vs survival curve)

Statistical test

Summary effect measure

Missing data

Interim analyses

Subgroup analyses

Subgroup analyses

The aim of these is to provide information on the opportunity to treat different groups of patients differently, informing the development of:

PERSONALIZED THERAPIES

Leading toward the era of

PRECISION MEDICINE

Prognostic and predictive factors

Subgroup analyses can inform potential prognostic and predictive factors

PROGNOSTIC FACTORS

- Predict outcome (with the same treatment)
- Do not require a randomized trial to identify
- Used in clinical decision-making (informing risk/benefit and cost/benefit)

FOR EXAMPLE:

Nodal status in early-stage breast cancer

- Strong prognostic effect (HR: 2)
- All adjuvant therapies have the same effect, regardless of nodal status

PREDICTIVE FACTORS

- Predict the efficacy of the treatment in different patients
- Identified solely by subgroup analyses from randomized trials

FOR EXAMPLE:

- Hormone receptors: efficacy of hormonal therapy in breast cancer
- PD-L1 expression: efficacy of immunotherapy in solid tumors
- Tumor grade (differentiation): efficacy of chemotherapy in NHL

Issues with subgroup analyses: Methodological

There are several methodological factors that can affect the validity of a subgroup analysis:

Retrospective vs prospective

Not very important

Planned vs unplanned

- Must always be included in the statistical plan
- If not planned = scientific exercise

Bias

- Blinded classification/analyses
- Selection of compliers, responders, and treated patients (?)
- Always compare subgroup HRs with those from the overall population

STUDY PROTOCOL!

Issues with subgroup analyses: Statistical

The common statistical issues with subgroup analyses are:

IMPROPER SIGNIFICANCE TESTING

MULTIPLICITY

Widely used method (wrong)

The test for significance is repeated in each subgroup at the conventional significance level

Hypothetical trial with 120 patients/arm

72 responses (60%) exp. therapy vs 48 (40%) control

p <0.002

72 responses (60%) exp. therapy vs 48 (40%) control
p <0.002

3 SUBGROUPS:

<30 years	**p=0.01**
30–50 years	**p=0.2 n.s.**
>50 years	**p=0.2 n.s.**

Conclusions?

72 responses (60%) exp. therapy vs 48 (40%) control
p <0.002

3 SUBGROUPS:

<30 years	**48/80 (60%) vs 32/80 (40%)** p=0.01
30–50 years	**12/20 (60%) vs 8/20 (40%)** p=0.2
>50 years	**12/20 (60%) vs 8/20 (40%)** p=0.2

72 responses (60%) exp. therapy vs 48 (40%) control
p <0.002

3 SUBGROUPS:

<30 years	**p=0.07 n.s.**
30–50 years	**p=0.07 n.s.**
>50 years	**p=0.07 n.s.**

Conclusions?

72 responses (60%) exp. therapy vs 48 (40%) control
p <0.002

3 SUBGROUPS:

<30 years	**24/40 (60%) vs 16/40 (40%)** p=0.07 n.s.
30–50 years	**24/40 (60%) vs 16/40 (40%)** p=0.07 n.s.
>50 years	**24/40 (60%) vs 16/40 (40%)** p=0.07 n.s.

Subgroup analyses

SUBGROUP SPECIFIC P: MEANINGLESS E MISLEADING

- Correct analysis: Test of interaction (Heterogeneity of the effect)
  - New null hypothesis: The effect is the same in all subgroups
  - The observed variation in the effect is compared to that expected by chance alone (small subgroups: large variations)

72 responses (60%) exp. therapy vs 48 (40%) control
OR=2.25

<30 years	60% vs 40%	**OR=2.25**
30–50 years	60% vs 40%	**OR=2.25**
>50 years	60% vs 40%	**OR=2.25**

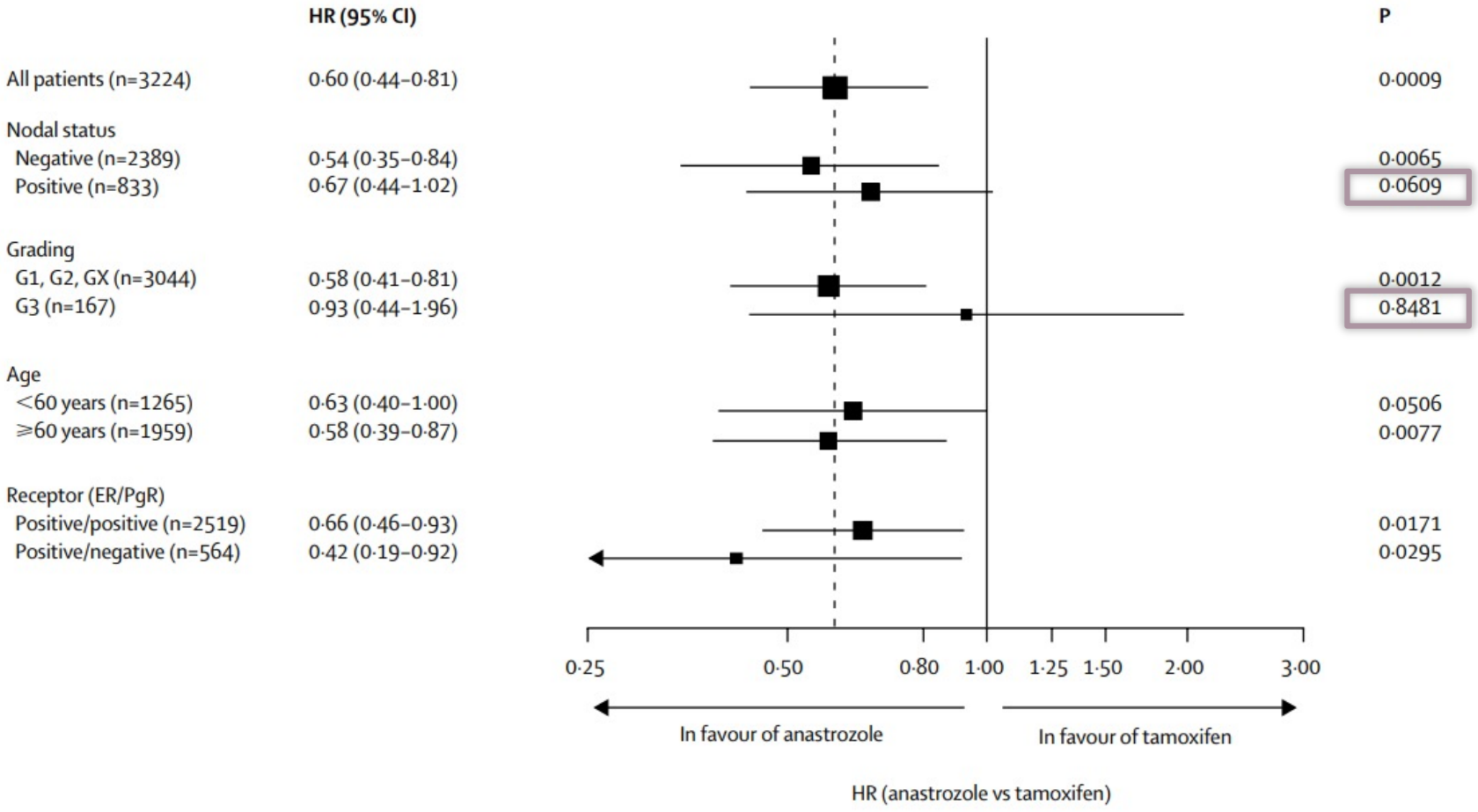
No evidence of interaction
p=1

72 responses (60%) exp. therapy vs 48 (40%) control
OR=2.25

<30 years	50% vs 50%	**OR=1**
30–50 years	60% vs 40%	**OR=2.25**
>50 years	70% vs 30%	**OR=5.3**

Evidence of interaction
p <0.05

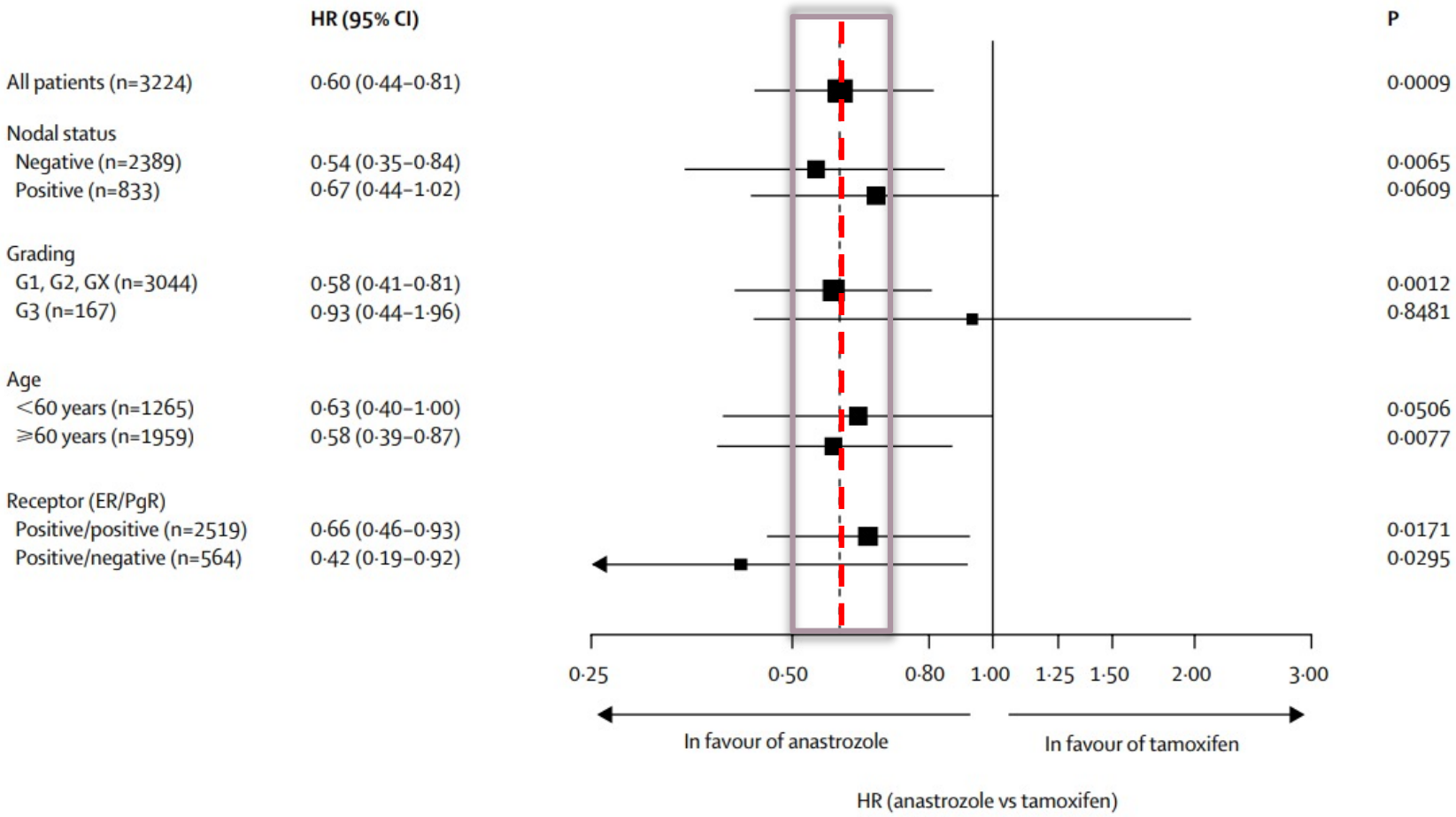
Improper significance testing



In this forest plot, the subgroup-specific p values are incorrect

Jakesz R, et al. Lancet 2005; 366: 455–462

Proper subgroup analysis



Including the overall treatment effect level makes it easier to see if the effect in a subgroup differed significantly from the overall treatment effect

Jakesz R, et al. Lancet 2005; 366: 455–462

Improper significance testing

A **TEST OF INTERACTION** is required to correctly analyze subgroup analyses; this assesses the heterogeneity of the treatment effect

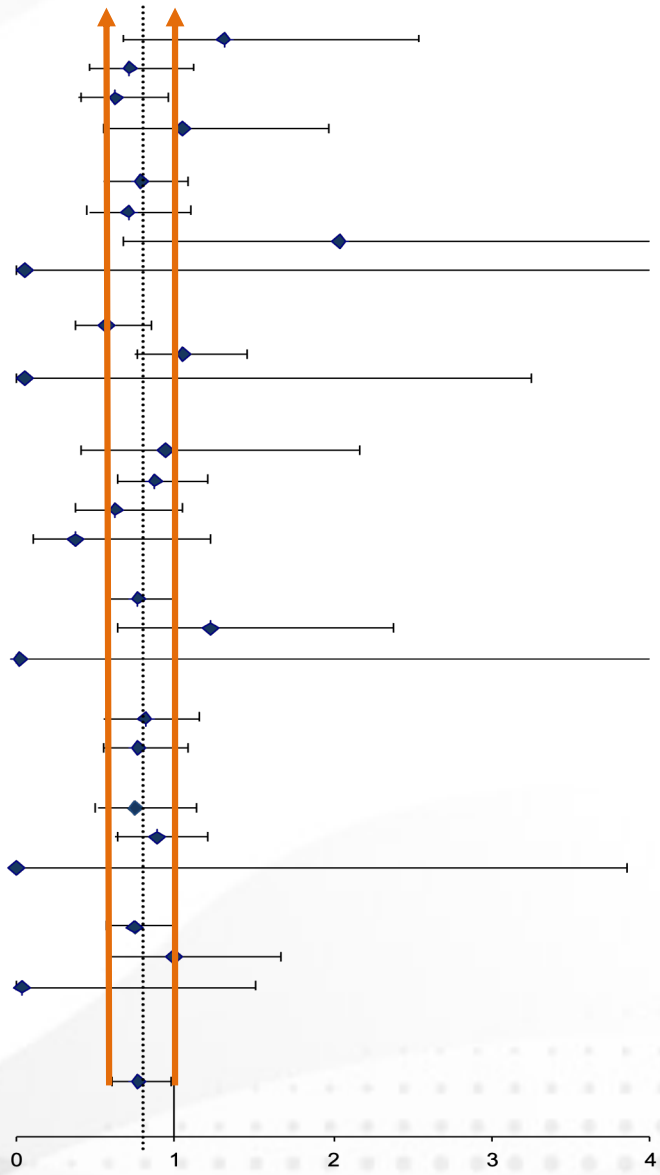
TEST OF INTERACTION

New null hypothesis: The treatment effect is the same across all subgroups

- The observed variation in the treatment effect is compared with that expected by chance alone
- Small subgroups, large variations

Subgroup-specific p values can be misleading and meaningless

Appropriate analysis – Overall survival

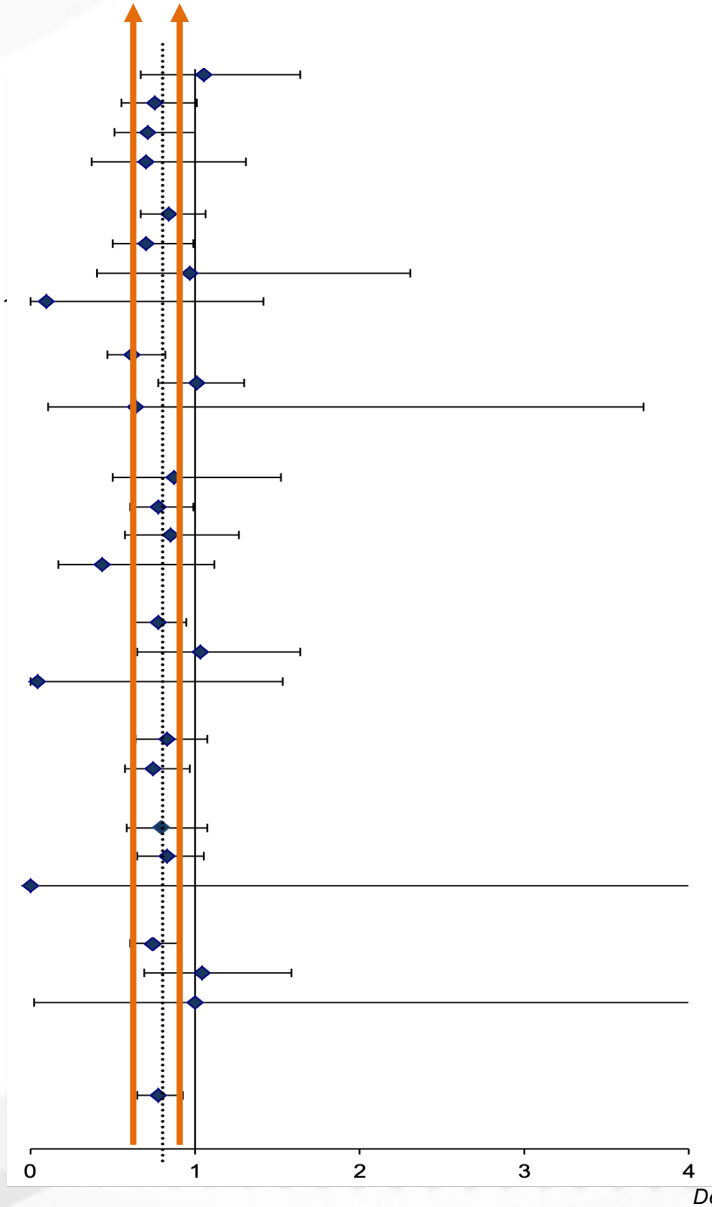


survival ITT2056		events/patients control	events/patients extended	Hazard ratio (95% CI)	P interaction
Age (years)	<55	16/275	22/271	1,31(0,676-2,536)	0,243
	55-64	49/393	36/386	0,718(0,462-1,117)	
	65-75	57/304	36/313	0,626(0,408-0,961)	
	>=76	25/58	22/56	1,044(0,553-1,973)	
Tumor size	pT1	83/704	69/703	0,789(0,571-1,091)	0,299
	pT2	52/261	35/252	0,707(0,454-1,102)	
	pT3-4	8/34	11/43	2,032(0,668-6,176)	
	pTx	4/31	1/28	0,053(0,001-5,298)	
Nodal status	pN0	64/581	39/568	0,567(0,379-0,849)	0,049
	pN+	75/411	76/428	1,056(0,763-1,462)	
	pNx	8/38	1/30	0,057(0,001-3,259)	
Histological Grade	G1	11/156	13/161	0,944(0,411-2,168)	0,391
	G2	81/564	74/589	0,877(0,637-1,209)	
	G3	42/221	25/213	0,628(0,372-1,058)	
	Gx	13/89	4/63	0,371(0,112-1,229)	
Hormone receptor	ER+a nd PR+	122/855	95/866	0,77(0,587-1,01)	0,274
	ER +o r PR+	20/153	21/146	1,229(0,636-2,375)	
	Unknown	5/22	1/14	0,028(0-2539)	
Previous Tamoxifen	<=2.5 years	80/555	51/509	0,809(0,565-1,158)	
	>2.5 years	67/475	65/517	0,77(0,546-1,087)	
Previous chemotherapy	NO	57/455	41/450	0,753(0,5-1,132)	0,163
	YES	85/557	74/565	0,886(0,645-1,217)	
	Unknown	5/18	1/11	0(0-3,855)	
BMI	<30	116/819	85/821	0,747(0,562-0,992)	0,773
	>=30	30/174	31/177	0,988(0,587-1,662)	
	Unknown	1/37	0/28	0,038(0-1,498)	
All patients		147/1030	116/1026	0,77(0,604-0,983) *p=0,036*	

Del Mastro L, et al. Lancet Oncol 2021; 22: 1458–1467

Appropriate analysis – Event-free survival

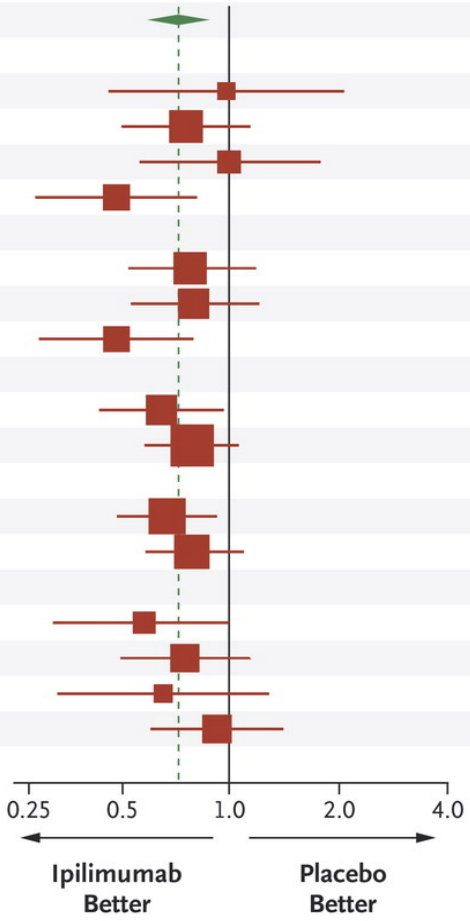
efs ITT 2056		events/patients control	events/patients extended	Hazard ratio (95% CI)	P interaction
Age (years)	<55	38/275	44/271	1,052(0,674-1,641)	0,666
	55-64	102/393	82/386	0,753(0,558-1,015)	
	65-75	94/304	63/313	0,718(0,515-1,001)	
	>=76	28/58	23/56	0,701(0,376-1,308)	
Tumor size	pT1	153/704	135/703	0,846(0,669-1,069)	0,32
	pT2	86/261	59/252	0,707(0,502-0,996)	
	pT3-4	15/34	16/43	0,969(0,406-2,314)	
	pTx	8/31	2/28	0,098(0,007-1,413)	
Nodal status	pN0	122/581	87/568	0,625(0,473-0,827)	0,031
	pN+	127/411	122/428	1,01(0,783-1,302)	
	pNx	13/38	3/30	0,639(0,11-3,723)	
Histological Grade	G1	28/156	27/161	0,874(0,501-1,526)	0,838
	G2	148/564	129/589	0,78(0,613-0,991)	
	G3	66/221	48/213	0,855(0,576-1,27)	
	Gx	20/89	8/63	0,435(0,169-1,118)	
Hormone receptor 5/13 C 1/11 L missing	ER+ and PR+	217/855	173/866	0,778(0,635-0,952)	0,23
	ER + or PR+	39/153	37/146	1,033(0,65-1,642)	
	Unknown	6/22	2/14	0,044(0,001-1,54)	
Previous Tamoxifen	<=2.5 years	145/555	100/509	0,833(0,642-1,08)	0,679
	>2.5 years	117/475	112/517	0,747(0,575-0,97)	
Previous chemotherapy	NO	96/455	80/450	0,797(0,591-1,076)	0,404
	YES	159/557	131/565	0,829(0,654-1,052)	
	Unknown	7/18	1/11	0,002(0-9,743)	
BMI	<30	2/37	161/821	0,745(0,605-0,918)	0,469
	>=30	212/819	51/177	1,047(0,693-1,584)	
	Unknown	48/174	0/28	1(0,019-52,748)	
All patients		262/1030	212/1026	0,777(0,648-0,931)	*p=0,006*



Del Mastro L, et al. Lancet Oncol 2021; 22: 1458–1467

Appropriate significance testing

Subgroup	Ipilimumab	Placebo	Hazard Ratio (95% or 99% CI)	P Value
	no. of deaths/total no.			
All patients	162/475	214/476	0.72 (0.59–0.88)	0.001
Disease stage				0.07
IIIA	24/98	22/88	0.98 (0.46–2.09)	
IIIB	68/213	85/207	0.75 (0.50–1.14)	
IIIC with 1–3 positive lymph nodes	34/69	45/83	1.00 (0.56–1.80)	
IIIC with ≥4 positive lymph nodes	36/95	62/98	0.48 (0.28–0.81)	
No. of positive lymph nodes				0.09
1	65/217	82/220	0.79 (0.52–1.21)	
2 or 3	61/163	70/158	0.83 (0.53–1.30)	
≥4	36/95	62/98	0.48 (0.28–0.81)	
Type of positive lymph node				0.21
Microscopic	54/210	76/193	0.61 (0.39–0.96)	
Macroscopic	108/265	138/283	0.80 (0.58–1.11)	
Ulceration				0.29
Yes	73/197	110/203	0.64 (0.44–0.94)	
No	79/257	88/244	0.80 (0.54–1.20)	
Lymph-node and ulceration status				0.35
Microscopic and ulceration	28/99	43/88	0.54 (0.29–0.99)	
Macroscopic and ulceration	45/98	67/115	0.76 (0.46–1.23)	
Microscopic and no ulceration	21/104	29/97	0.62 (0.30–1.29)	
Macroscopic and no ulceration	58/153	59/147	0.90 (0.56–1.45)	



Eggermont AMM, et al. N Engl J Med 2016; 375: 1845–1855

Subgroup analysis methodology

In modern trial design, subgroup analyses are carefully designed to ensure validity of the results

SUBGROUP ANALYSIS METHODOLOGY

- Careful planning to prevent selection and assessment biases
- Test for interaction: **H0=the (lack of) effect is the same in all subgroups**
  - No subgroup-specific p values should be calculated
- Controlling for multiplicity:
  - Planned vs post hoc analyses
  - Exploratory vs confirmatory analyses
  - p value corrections

Larger data sets allow more POWERFUL subgroup analyses

Strategies for addressing multiplicity

POSITIVE PRIMARY RESULTS

If all preceding statistical analyses are positive, **planned subgroup analyses remain valid**

NEGATIVE PRIMARY RESULTS

If any preceding statistical analyses were negative, **planned subgroup analyses will be invalid** if not corrected for multiplicity

α split required

e.g. 2% × primary analysis
1% each × 3 interaction tests

STATISTICAL ANALYSIS PLAN!

Conclusions

- Most modern trials are adequately designed to avoid problems of multiplicity in the statistical analyses
- However, these analyses are focused on statistical testing rather than estimation. Modern statistical and medical sciences are moving away from p values and are more interested in estimating the effects of the treatments

Multiplicity-corrected, statistically significant treatment effect estimates are biased, as they represent overestimations of the true treatment effects