How to interpret a trial: Design and statistics

Paolo Bruzzi, MD, MPH, PhD
Retired, formerly Head of Clinical Epidemiology, San Martino Hospital, Italy

February 2024 | 0124--MRC-047

Disclosures

Speaker fees, research and educational grants: Alexion Pharmaceuticals, Astellas Pharma Europe Ltd., AstraZeneca SPA, BeiGene, BioMarin Pharmaceutical Inc., Daiichi Sankyo, Eisai Co., Ltd., Eli Lilly, Fullcro, Genomic Health, Inc., Gilead Sciences, Inc., GSK, Johnson & Johnson Innovative Medicine (formerly Janssen), Menarini Group, Merck Group, MSD, Pierre Fabre, Roche, Rottapharm Biotech, Sanofi, Servier.

Introduction

- How to interpret a trial
  - Internal validity
  - External validity
  - Relevance
- Multiplicity
  - General
  - Multiple endpoints, interim analyses, subgroup analyses
- Indirect comparisons

When you interpret the results of a study, you should ask two questions

1. Are the results true?

2. So what?

Slide courtesy of Paolo Bruzzi.

Question 1: Are the results true?

The results of a trial are always (more or less) wrong!

↓

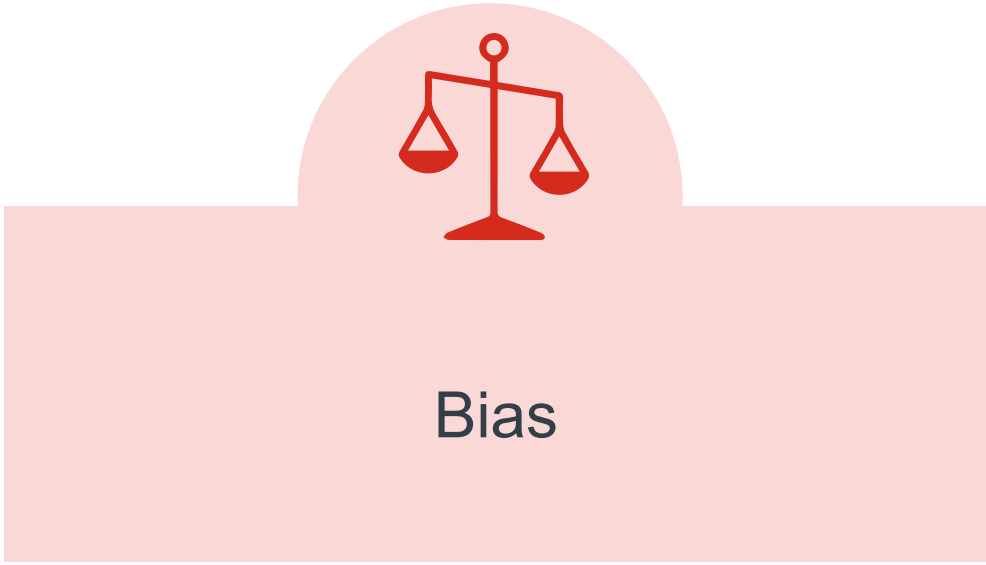
Why?

Slide courtesy of Paolo Bruzzi.

Sources of error

Chance

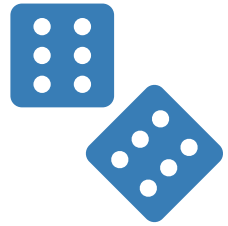
Usually referred to by statisticians as **sampling error**



Bias

Slide courtesy of Paolo Bruzzi.

The effect of chance



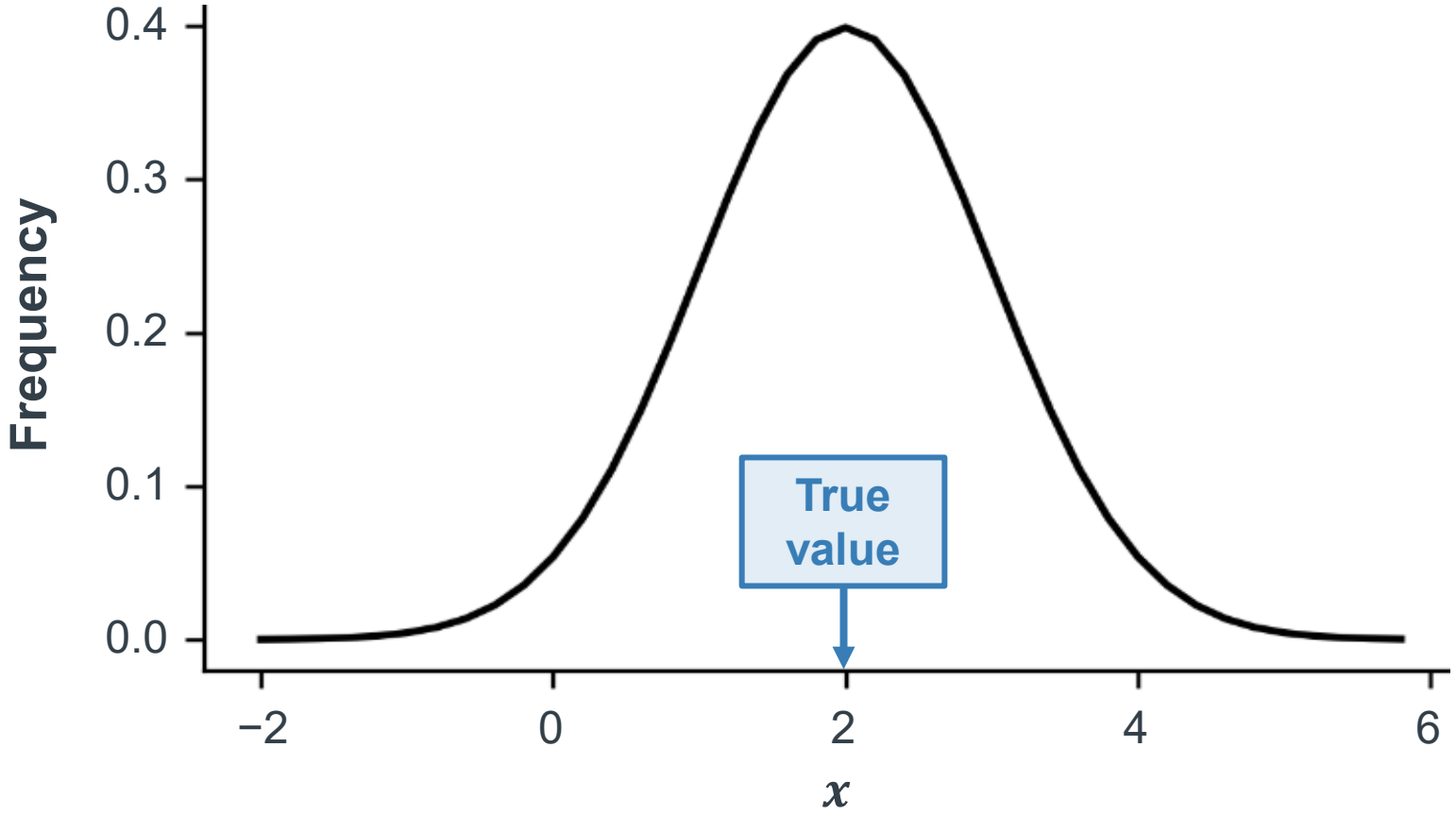
- Even if the same phenomenon is observed several times under **exactly** the same conditions, the results of the observations are never identical

- Some examples include:
  - **Individual measurements:** weight, blood pressure, tumor size
  - **Group measurements:** mean blood pressure, mean weight, ORR, mOS

- **If there is no bias**, the results of several observations of the same phenomenon under **exactly** the same conditions follow a normal distribution, whose **mean** is equal to the **true value**

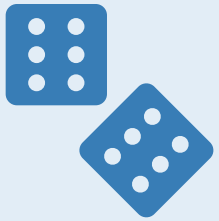
mOS, median overall survival; ORR, objective response rate.
Slide courtesy of Paolo Bruzzi.

Normal distribution



Slide courtesy of Paolo Bruzzi.

Sources of error



Chance
(sampling error)

Bias

Because of flaws in the design/conduct of the observation, the results may be **distorted**

Slide courtesy of Paolo Bruzzi.

Sources of bias

Measurement

Selection

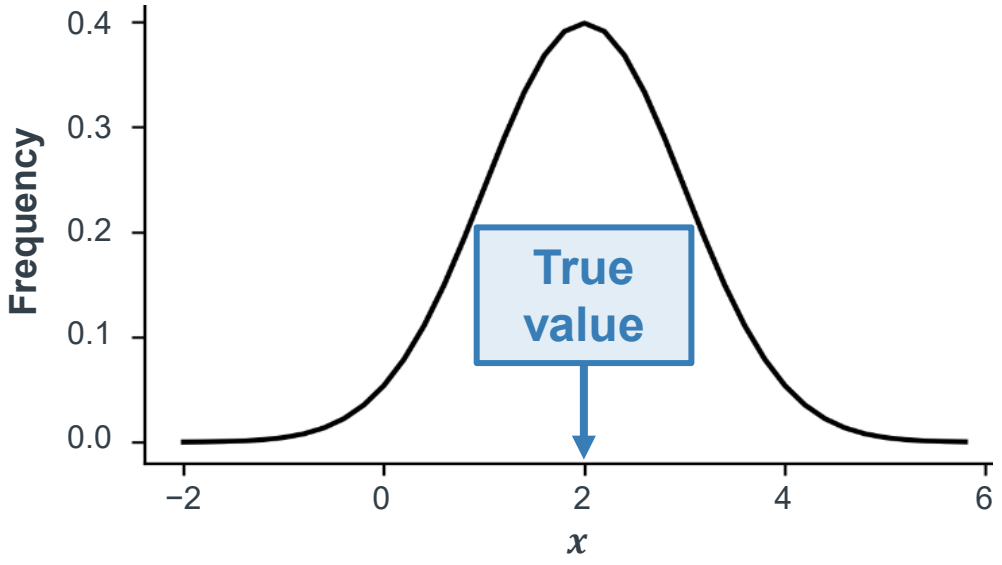
Reporting

Slide courtesy of Paolo Bruzzi.

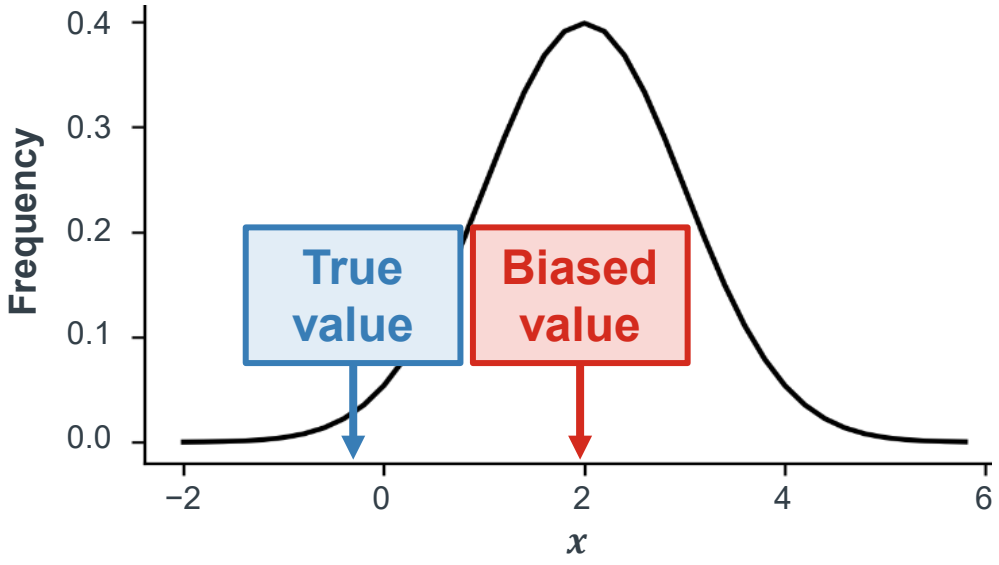
The effect of bias

- Because of flaws in the design/conduct of the observation, the results may be **distorted**
- The results of several ‘biased’ observations of the same phenomenon under exactly the same conditions follow a normal distribution that is centered on a **wrong** value

Unbiased observation

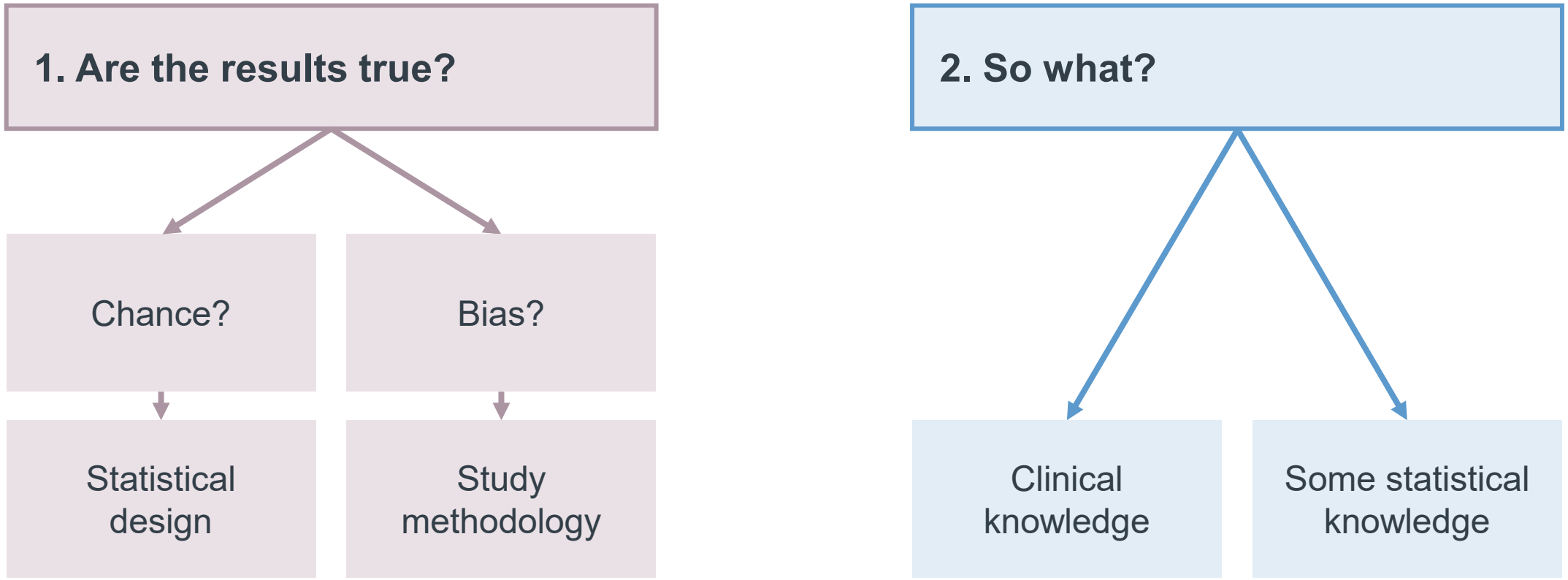


Biased observation



Slide courtesy of Paolo Bruzzi.

When you interpret the results of a study, you should ask two questions



Slide courtesy of Paolo Bruzzi.

1. How to interpret a trial

Interpretation of a trial

Reading scientific papers

Writing scientific papers

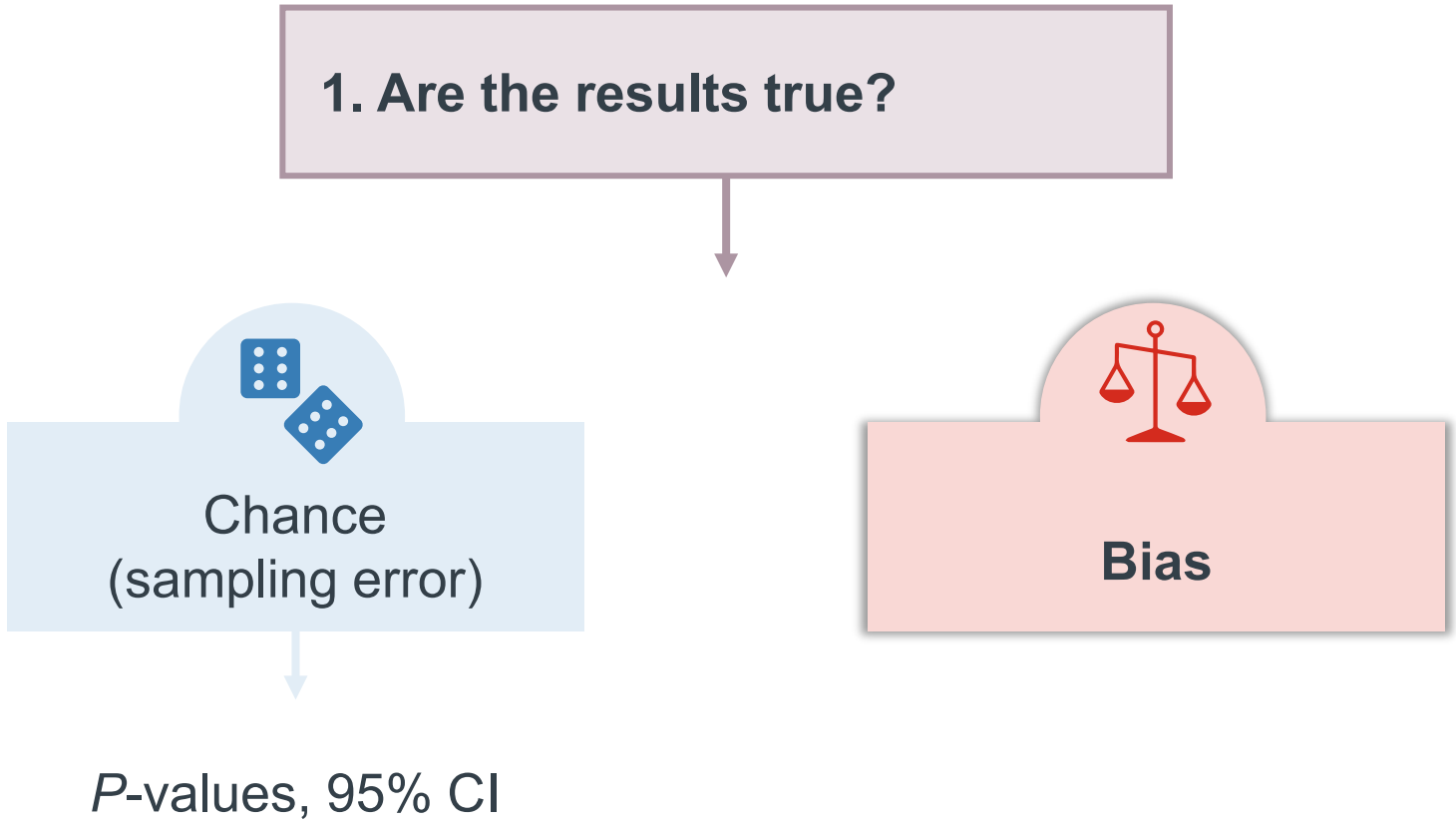
Designing studies

Preparing/reviewing grant applications

These require the same critical tools that we will be reviewing today

Slide courtesy of Paolo Bruzzi.

When you assess the results of a study, you should ask two questions



CI, confidence interval.
Slide courtesy of Paolo Bruzzi.

When you assess the results of a study, you should ask two questions

1. Are the results true?

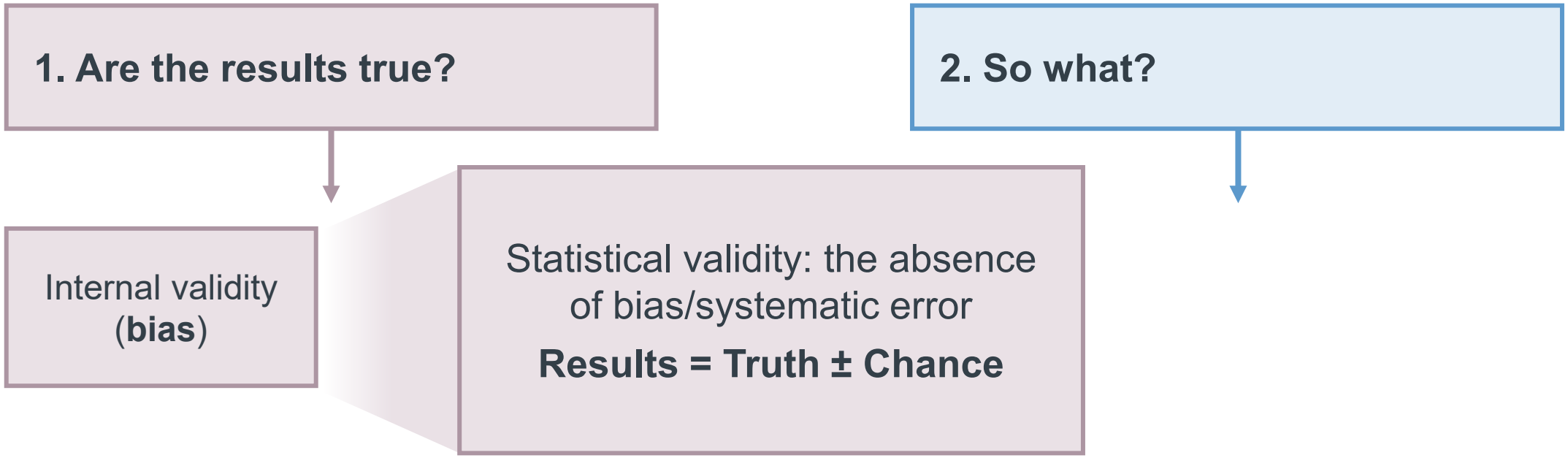
- Internal validity (**bias**)
- Internal consistency
- External consistency
- Plausibility

2. So what?

- External validity
- Relevance

Slide courtesy of Paolo Bruzzi.

Internal validity



Slide courtesy of Paolo Bruzzi.

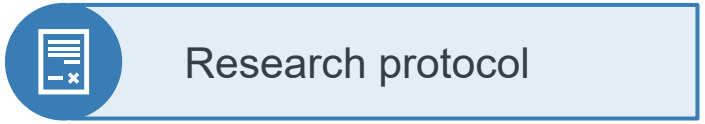
Internal validity

Checklist:

- ☑ Research protocol (rationale)
- ☑ Primary aim
- ☑ Study design
- ☑ Randomization
- ☑ Endpoint selection, assessment, and masking
- ☑ Inclusion/exclusion criteria
- ☑ Treatment protocol
- ☑ Statistical plan (including sample size power, *P*-values, and associated CIs)
- ☑ Endpoint assessment (including intention-to-treat population and FU protocol)

CI, confidence interval; FU, follow-up.
Slide courtesy of Paolo Bruzzi.

Research protocol



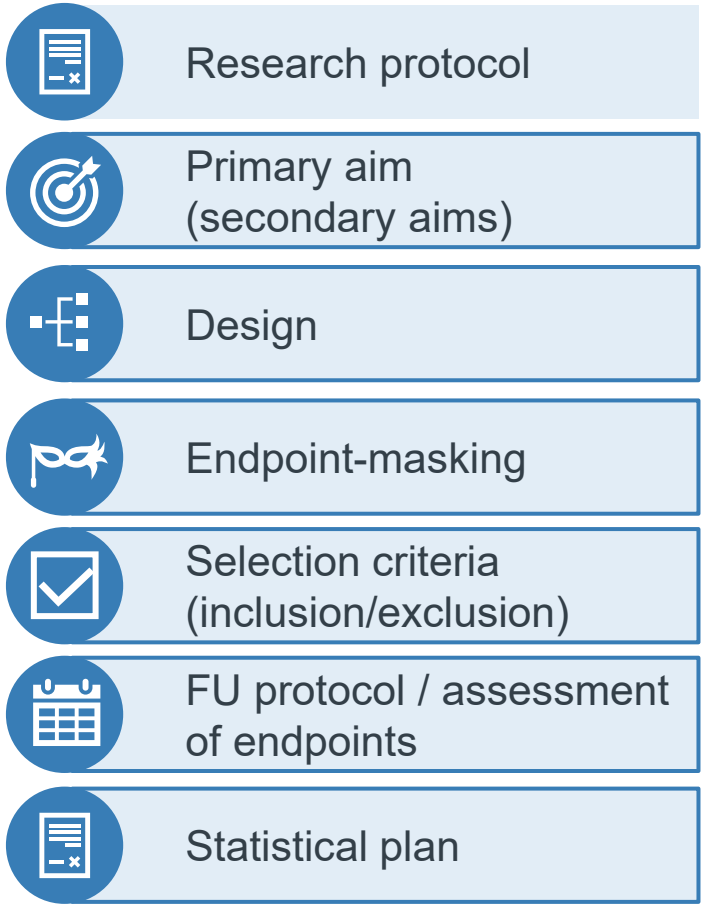
Research protocols should:

Be available for consultation

Contain explicit reference to registration of the trial

Slide courtesy of Paolo Bruzzi.

Study protocol



Slide courtesy of Paolo Bruzzi.

Internal validity

Research protocol

Primary aim

Primary aims should be stated in an **explicit statement**

Slide courtesy of Paolo Bruzzi.

Internal validity

- Research protocol
- Primary aim
- Randomization procedures

Within the research protocol, there should be a detailed description of the procedures used to ensure that randomization is free from bias

Slide courtesy of Paolo Bruzzi.

Internal validity

- Research protocol
- Primary aim
- Randomization procedures
- Endpoint-masking

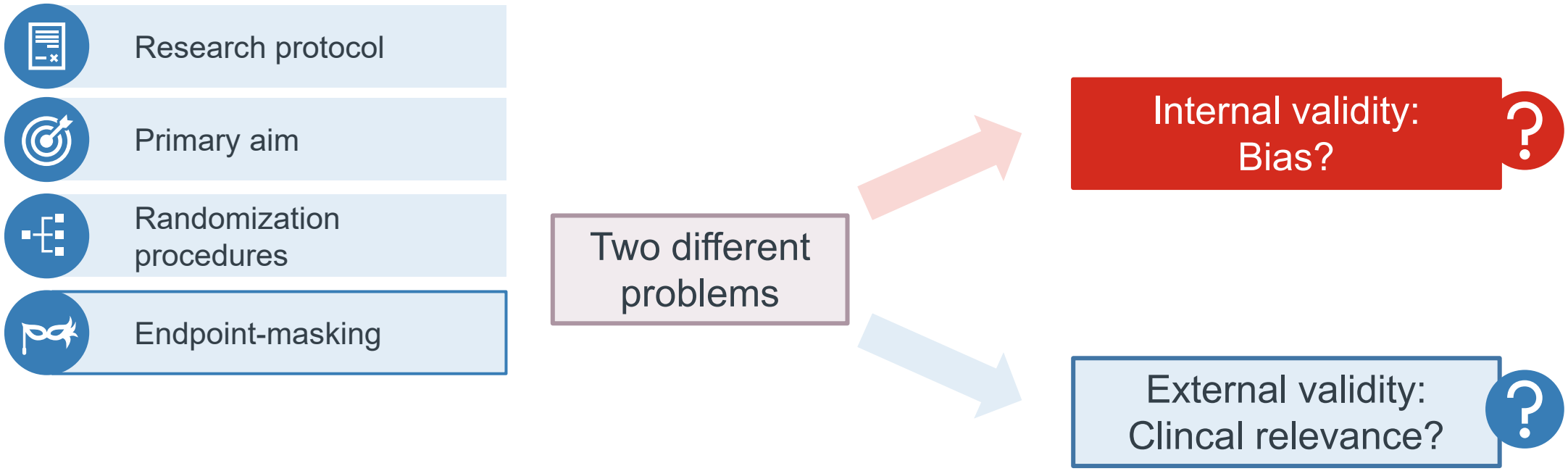
Endpoints should:

Have an unequivocal definition and assessment procedures

Be assessed using methods to limit potential bias, e.g., masking procedures

Slide courtesy of Paolo Bruzzi.

Internal validity



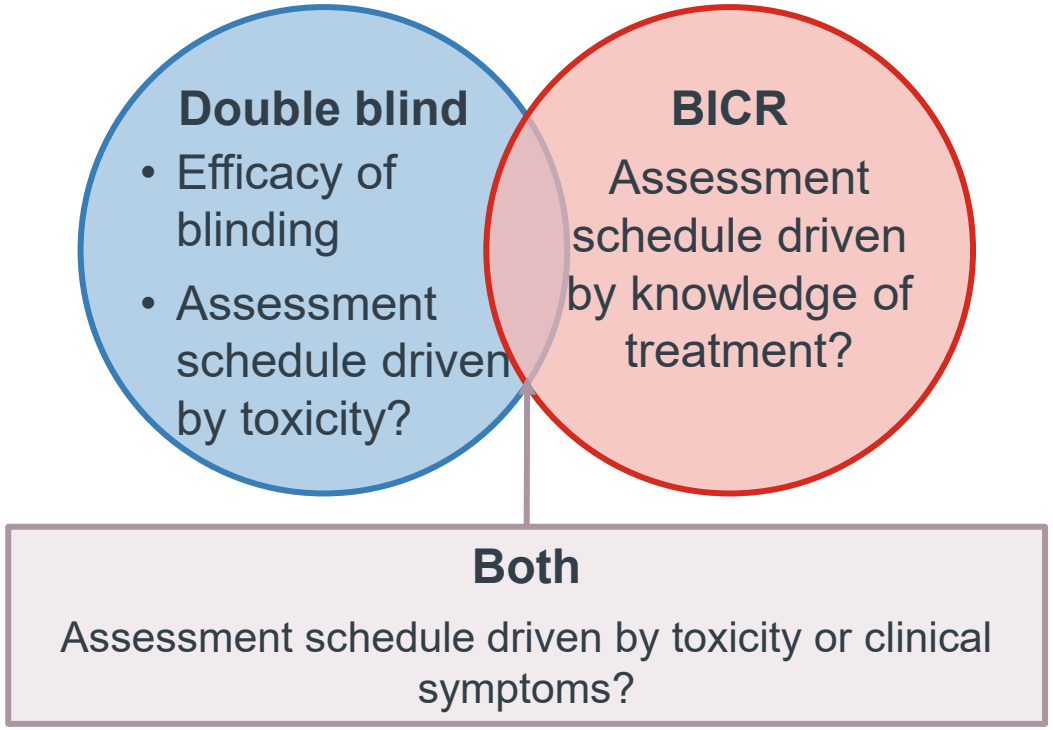
Slide courtesy of Paolo Bruzzi.

Internal validity: Endpoint selection, assessment, and masking

Considerations for endpoint selection

Endpoint	Objectivity	Assessment	Blinding requirements
Survival	+++	++	Not required
Objective response	++/−	+	BICR
Event-free survival (e.g. PFS, RFS)	+/−−	+?	BICR / double blind
QoL scoring	−−−	−−−	Double blind

Considerations for blinding selection



BICR, blinded independent central review; PFS, progression-free survival; QoL, quality of life; RFS, relapse-free survival.
Slide courtesy of Paolo Bruzzi.

Internal validity

- Research protocol
- Primary aim
- Randomization procedures
- Endpoint-masking
- Statistical plan

A statistical plan should contain the following elements:

Primary and secondary endpoints

Including, but not limited to:

- Details of associated tests and decision rules Planned interim and subgroup analyses
- Statistical powering and sample size calculations
- Analysis of the intention-to-treat population

Details of any protocol amendments

Including, but not limited to:

- Any associated conditions (e.g. results blinding)
- Date of amendments

Slide courtesy of Paolo Bruzzi.

Internal validity

- Research protocol
- Primary aim
- Randomization procedures
- Endpoint-masking
- Statistical plan
- Lost to FU / not evaluated

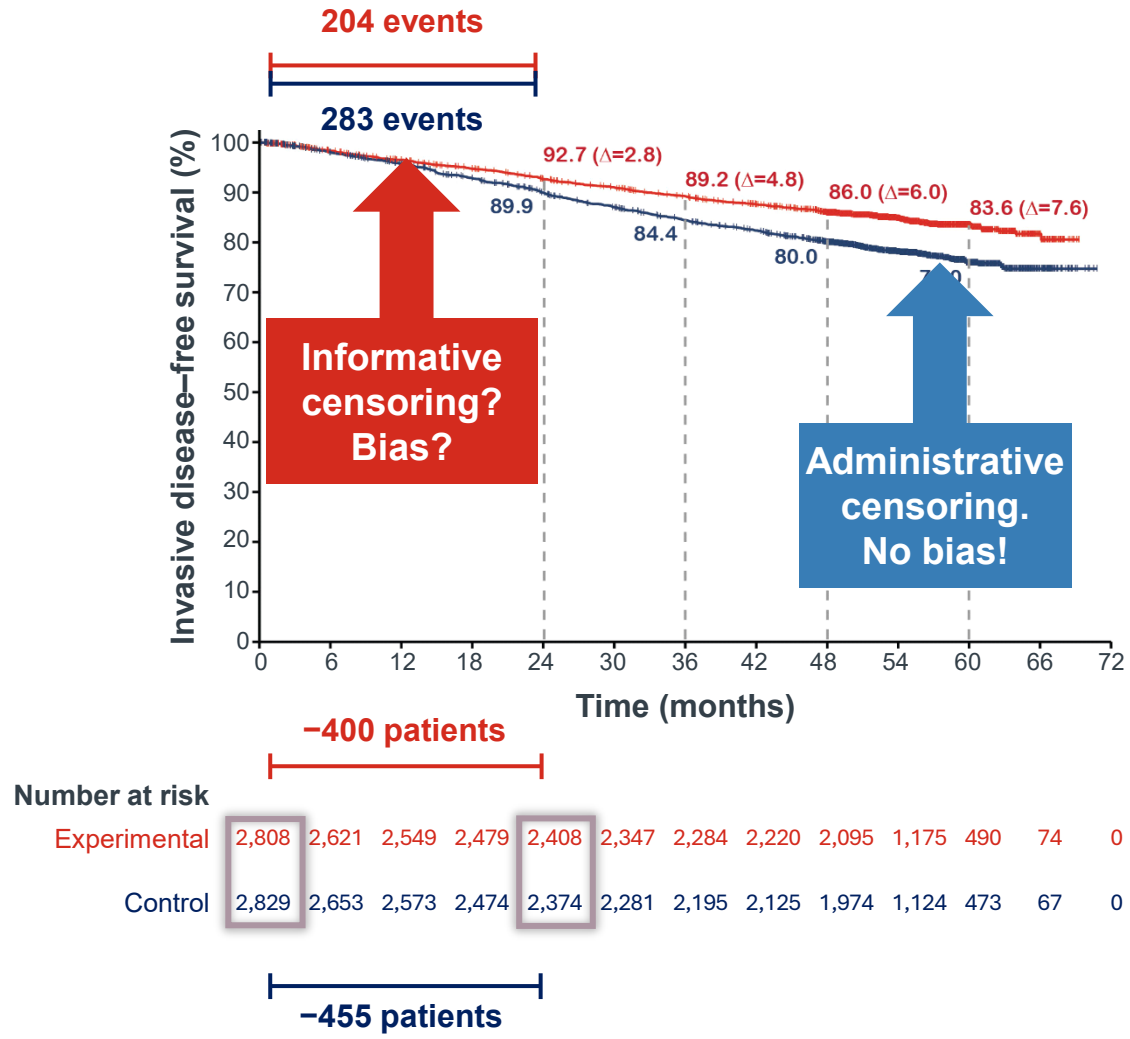
Patients that were lost to FU should have:

The **timing** of discontinuation noted

The **reason** for discontinuation noted

FU, follow-up.
Slide courtesy of Paolo Bruzzi.

Informative censoring: Kaplan–Meier curves



Calculating early censoring:

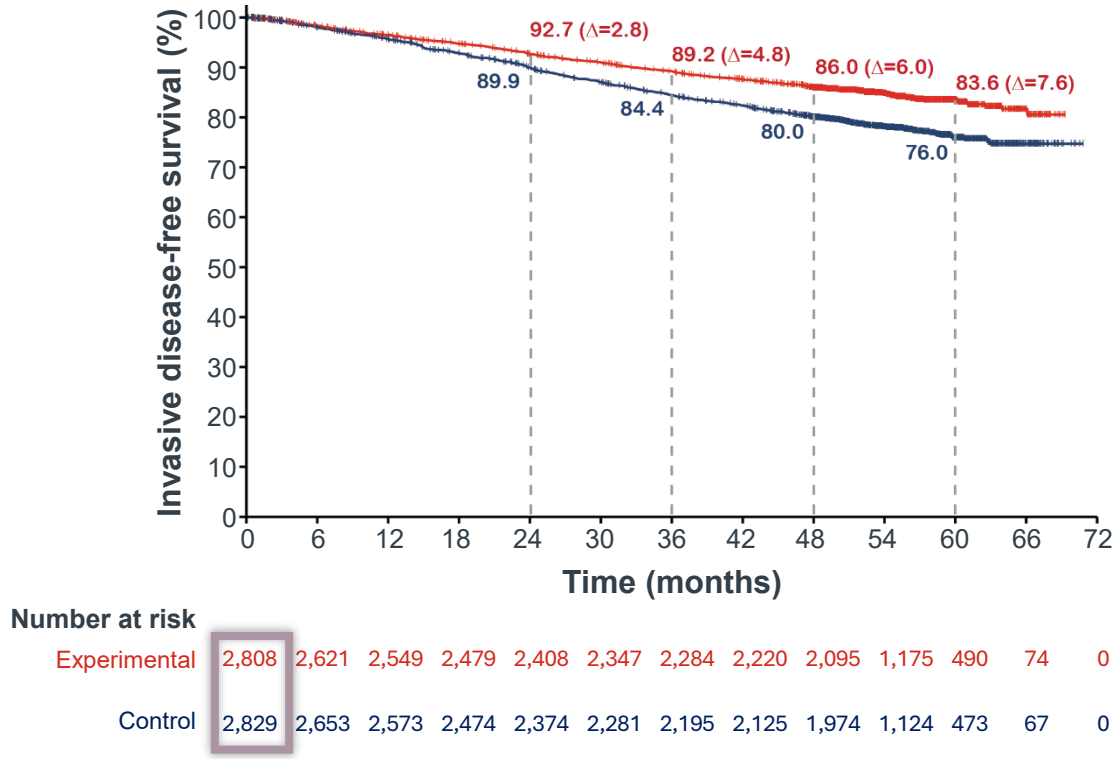
Experimental arm:
400 patients – 204 events = >194 censored

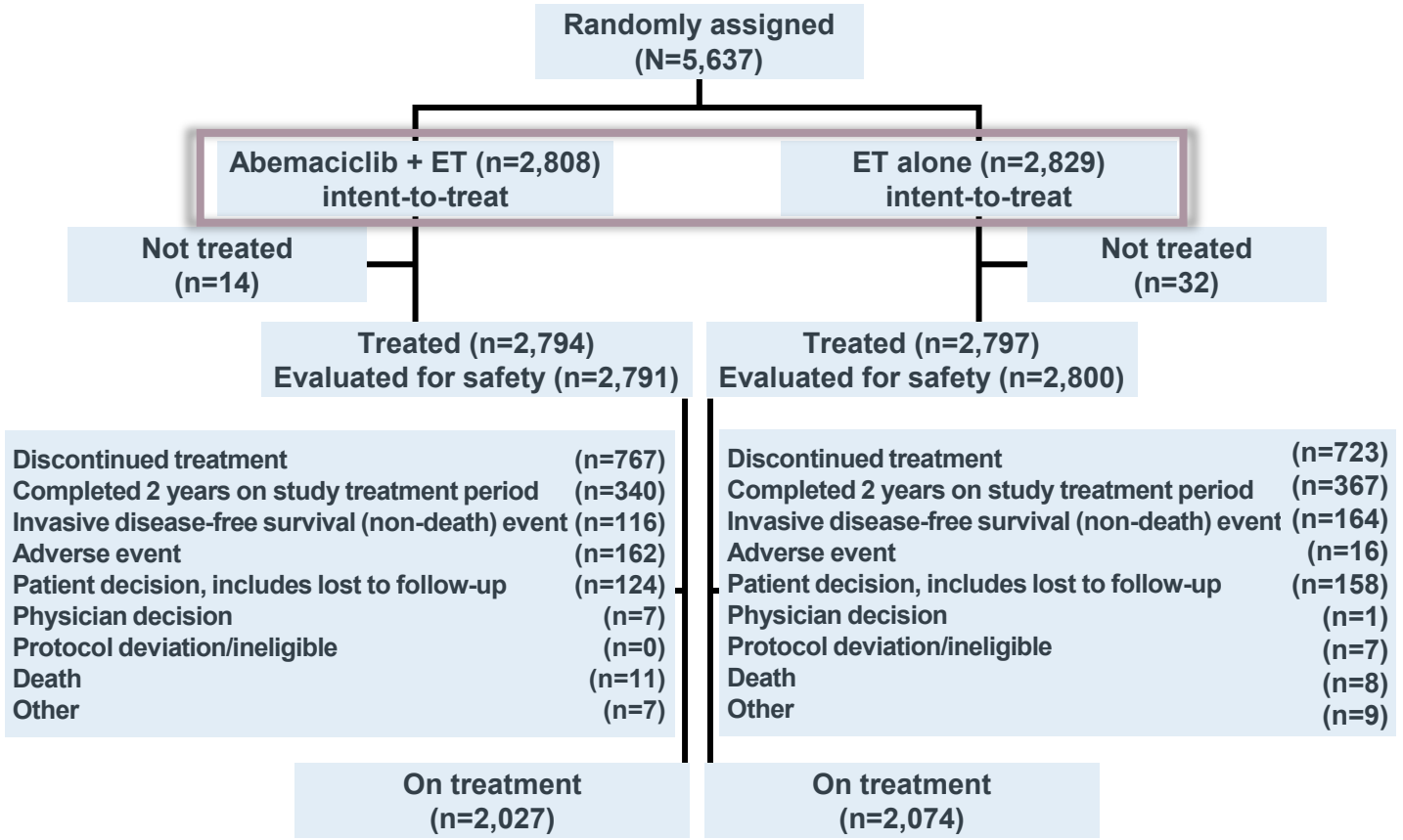
Control arm:
455 patients – 283 events = >172 censored

Figure included for illustrative purposes only. Rastogi P *et al. J Clin Oncol* 2024: JCO2301994.

Intention-to-treat population

The intention-to-treat population should also be reflected in the CONSORT diagram[1,2]

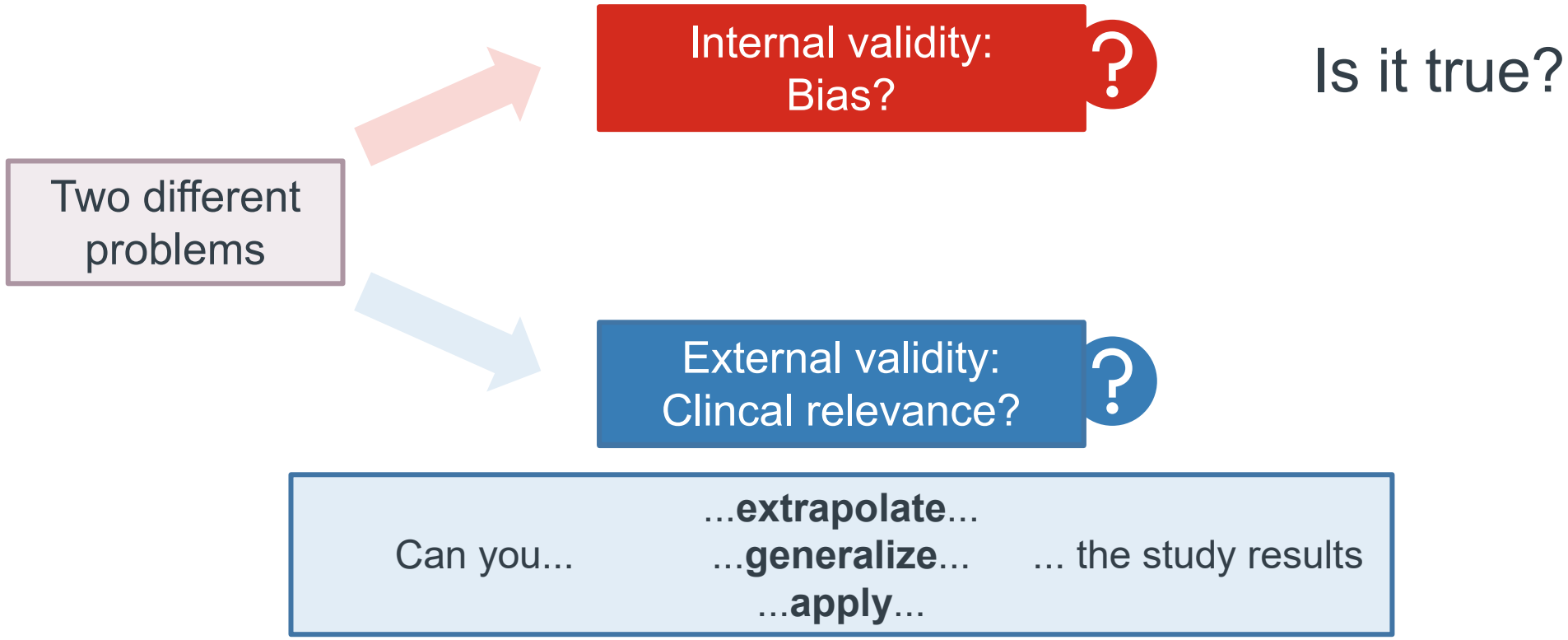




ET, endocrine therapy.

Figures included for illustrative purposes only. 1. Rastogi P *et al. J Clin Oncol* 2024: JCO2301994. 2. Adapted from Johnston SRD *et al. J Clin Oncol* 2020; 38 (34): 3987–3998.

Interpretation



Is it true?

Slide courtesy of Paolo Bruzzi.

Generalization

In clinical trial interpretation, **generalization** has two meanings:

Proof of principle

Is the approach feasible?

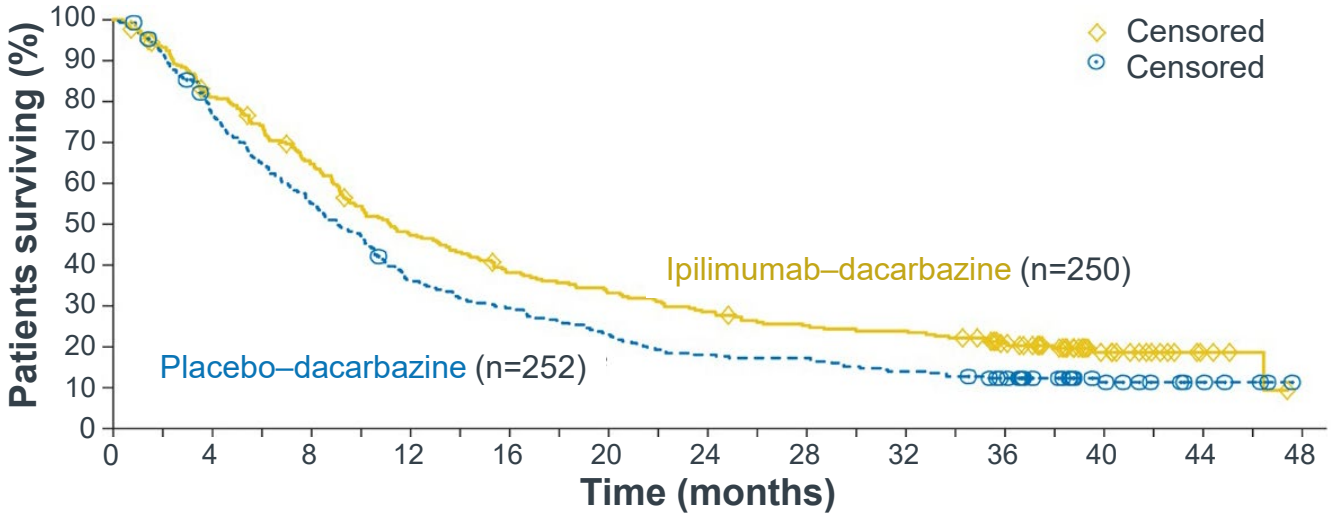
Applicability of the results

Could these data influence clinical practice?

Slide courtesy of Paolo Bruzzi.

Generalization: Proof of principle

Clinical trials can be designed to **validate the utility** of
a regimen for a particular indication and patient population

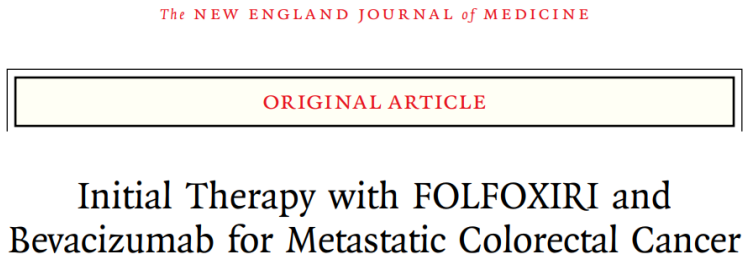


Endpoint	DTIC + ipilimumab (n=250)	DTIC + placebo (n=252)	Significance test
1-year OS rate, %	47.3	36.3	–
2-year OS rate, %	28.5	17.9	–
3-year OS rate, %	20.8	12.2	–
Median OS, months	11.2	9.1	HR: 0.72 ***P* <0.001**

"…significant improvement in OS among patients… who received ipilimumab plus dacarbazine as compared with dacarbazine plus placebo"

DTIC, dacarbazine; HR, hazard ratio; OS, overall survival.
Robert C *et al*. *N Engl J Med* 2011; 364 (26): 2517–2526.

Generalization: Use in clinical practice



5 **FOLFOXIRI + bevacizumab**

Five-drug regimen

- Fluorouracil
- Leucovorin
- Oxaliplatin
- Irinotecan
- Bevacizumab

vs.

4 **FOLFIRI + bevacizumab**

Four-drug regimen

- Fluorouracil
- Leucovorin
- Irinotecan
- Bevacizumab

"FOLFOXIRI plus bevacizumab, as compared with FOLFIRI plus bevacizumab, improved the outcome in patients with metastatic colorectal cancer and increased the incidence of some adverse events."

Loupakis F *et al*. *N Engl J Med* 2014; 371 (17): 1609–1618.

External validity

2. So what?

- Phase 2 data will inform the decision to initiate a Phase 3 trial
- Proof-of-principle trials may determine the decision to start as well as the design of other trials

A pragmatic trial may be important for:

Registration

Guidelines

Clinical decisions

Organization of health services

Slide courtesy of Paolo Bruzzi.

External validity

FU, follow-up.
Slide courtesy of Paolo Bruzzi.

External validity

Internal validity

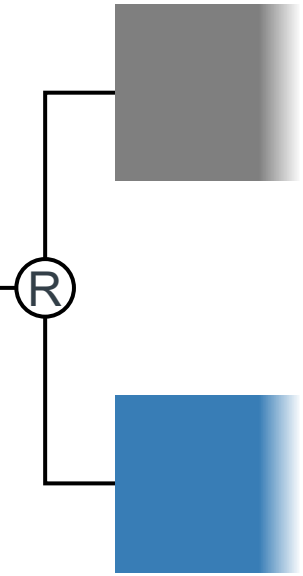
An **open-label, uncontrolled** trial has shown that drug X has a beneficial effect on QoL

QoL, quality of life.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)

A **randomized**, **double-blind** trial has shown that drug A is more effective than **placebo** in the treatment of acute leukemia



R, randomization.
Slide courtesy of Paolo Bruzzi.

Study design (contrast)

Experimental	vs.	Nil/placebo	Only if **no approved effective** therapy
Experimental + SoC	vs.	SoC	Only if **no interactions between treatments** and **different MoAs**
Experimental	vs.	SoC	Ethically acceptable?

Always consider:
Is the control treatment an acceptable standard?

MoA, mechanism of action; SoC, standard of care.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)
- Selection criteria

In an RCT, eligible patients:

- Were women
- Had early breast cancer
- Were T1N0
- Were ER+ HER2−
- Were <40 years of age

Drug X was shown to be more effective than standard chemotherapy…

ER, estrogen receptor; HER, human epidermal growth factor receptor; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)
- Selection criteria /
  - Patient characteristics

In an RCT, eligible patients:

- Were women (any)
- Had early breast cancer

Drug X was shown to be more effective than standard chemotherapy…

However

Because of competing studies, only patients to which the following applied were recruited:

- <40 years of age
- T1N0
- ER+ HER2− tumors

ER, estrogen receptor; HER, human epidermal growth factor receptor; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)
- Selection criteria
- Participating centers

An RCT conducted at the Mayo Clinic has shown that plastic surgery of aortic insufficiency is more effective than standard prosthesis implantation…

RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

External validity

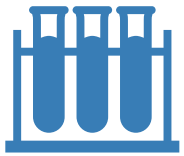
- Internal validity
- Study design (contrast)
- Selection criteria
- Participating centers
- Treatment/FU protocol

An RCT has shown that FU with monthly PET-CT and radioguided surgery of liver relapses in operated colon cancer…

FU, follow-up; PET-CT, positron emission tomography–computed tomography; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

External validity

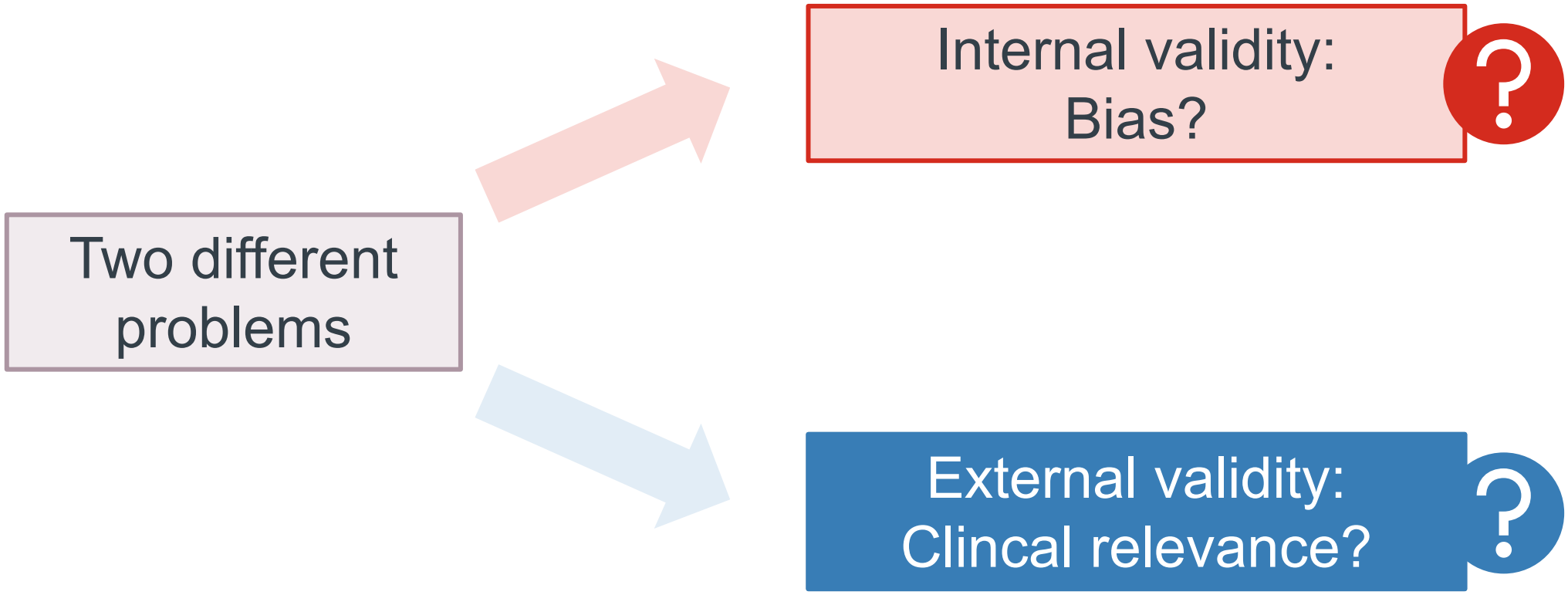
- Internal validity
- Study design (contrast)
- Selection criteria
- Participating centers
- Treatment/FU protocol
- Endpoints



…lead to a reduction of CEA levels

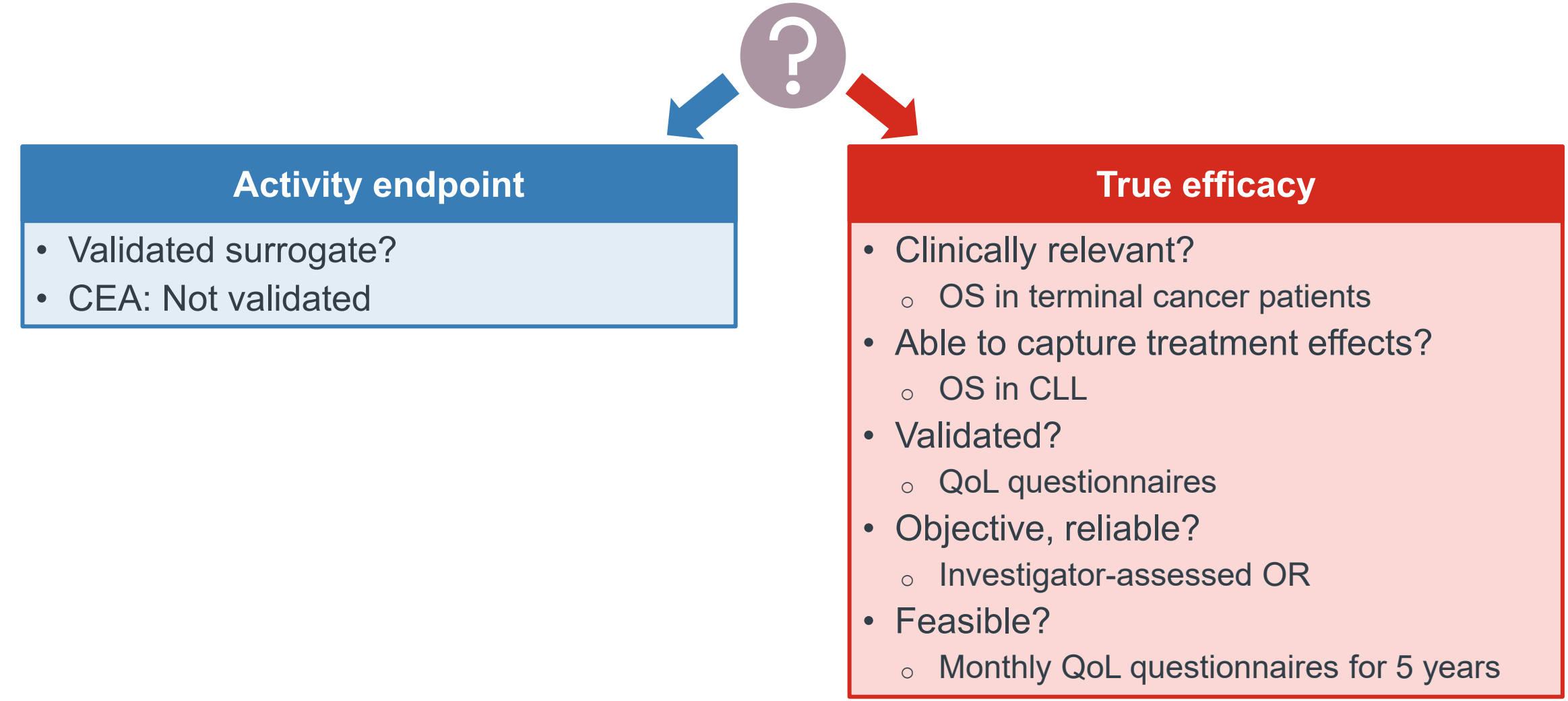
CEA, carcinoembryonic antigen; FU, follow-up.
Slide courtesy of Paolo Bruzzi.

Endpoints in clinical trials



Slide courtesy of Paolo Bruzzi.

Endpoints in Phase III studies



CEA, carcinoembryonic antigen; CLL, chronic lymphocytic leukemia; OR, objective response; OS, overall survival; QoL, quality of life.
Slide courtesy of Paolo Bruzzi.

Design of a Phase III trial: Endpoints and blinding

	OS	PFS

Benefit	• Natural endpoint
	  ○ If a benefit is observed, then no discussion is required
	  ○ Blinding is not required for assessment
		• Shorter FU than OS
		• Stronger effect than OS

Considerations	• Longer FU than PFS • (Almost) always weaker effect than PFS • Risk of false-negative results (crossover); (competing risks)	• Blinding required • Dependent on FU protocol • No cross-trial comparisons • **Interpretation questionable** • Double blinding is often overvalued and inefficient (toxicities?)

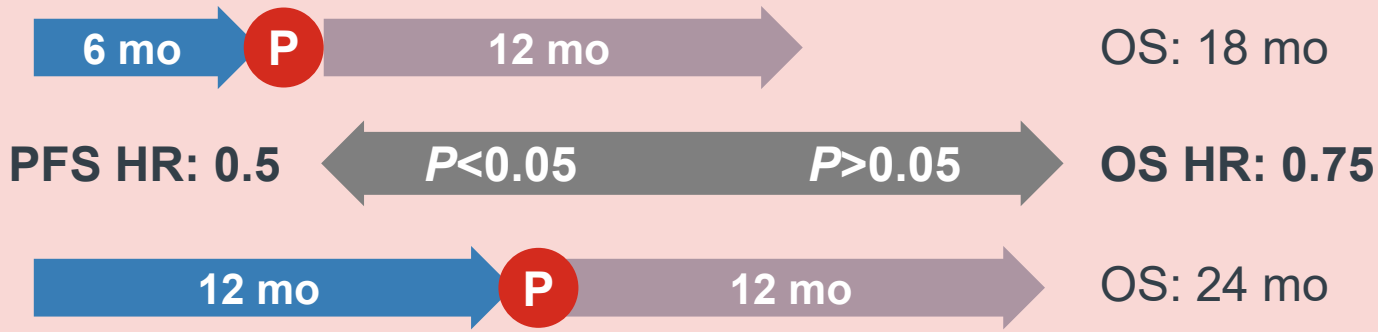
FU, follow-up; OS, overall survival; PFS, progression-free survival.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)
- Selection criteria
- Participating centers
- Treatment/FU protocol
- Endpoints

Note:

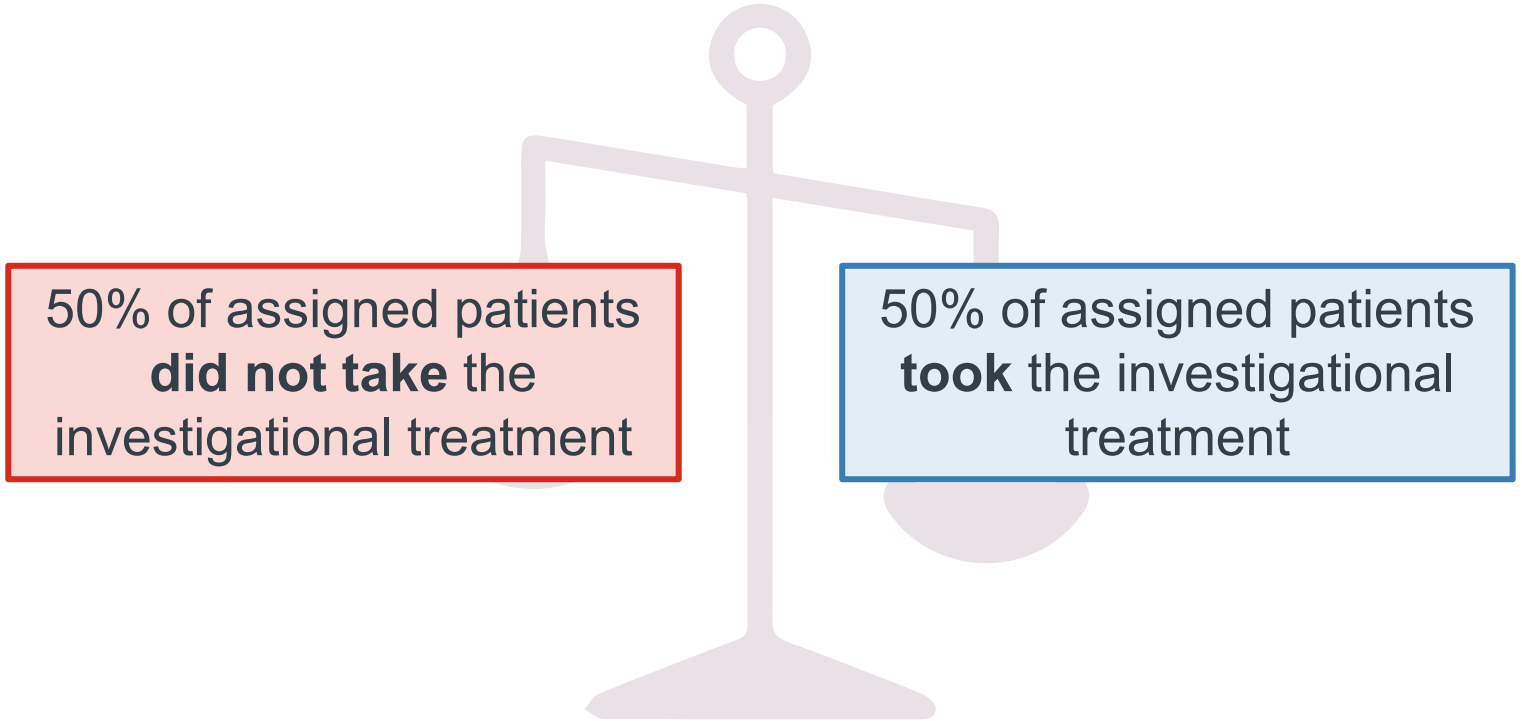
For mathematical (and biological) reasons, the effect of *most* anticancer treatments on PFS is much stronger than their effect on OS



FU, follow-up; HR, hazard ratio; mo, months; OS, overall survival; P, progression; PFS, progression-free survival.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)
- Selection criteria
- Participating centers
- Treatment/FU protocol
- Endpoints
- Compliance/contamination



FU, follow-up.
Slide courtesy of Paolo Bruzzi.

External validity

- Internal validity
- Study design (contrast)
- Selection criteria
- Participating centers
- Treatment/FU protocol
- Endpoints
- Compliance/contamination

Note:

Low compliance:

- Makes the two treatment groups more similar (HR biased toward unity)
- Decreases power* in superiority trials
- Makes it easier to spuriously demonstrate non-inferiority (sensitivity per-protocol analyses)

*Power is the probability to detect a significant effect in a clinical trial.
FU, follow-up; HR, hazard ratio.
Slide courtesy of Paolo Bruzzi.

External validity

FU, follow-up.
Slide courtesy of Paolo Bruzzi.

When you assess the results of a study, you should ask two questions

1. Are the results true?

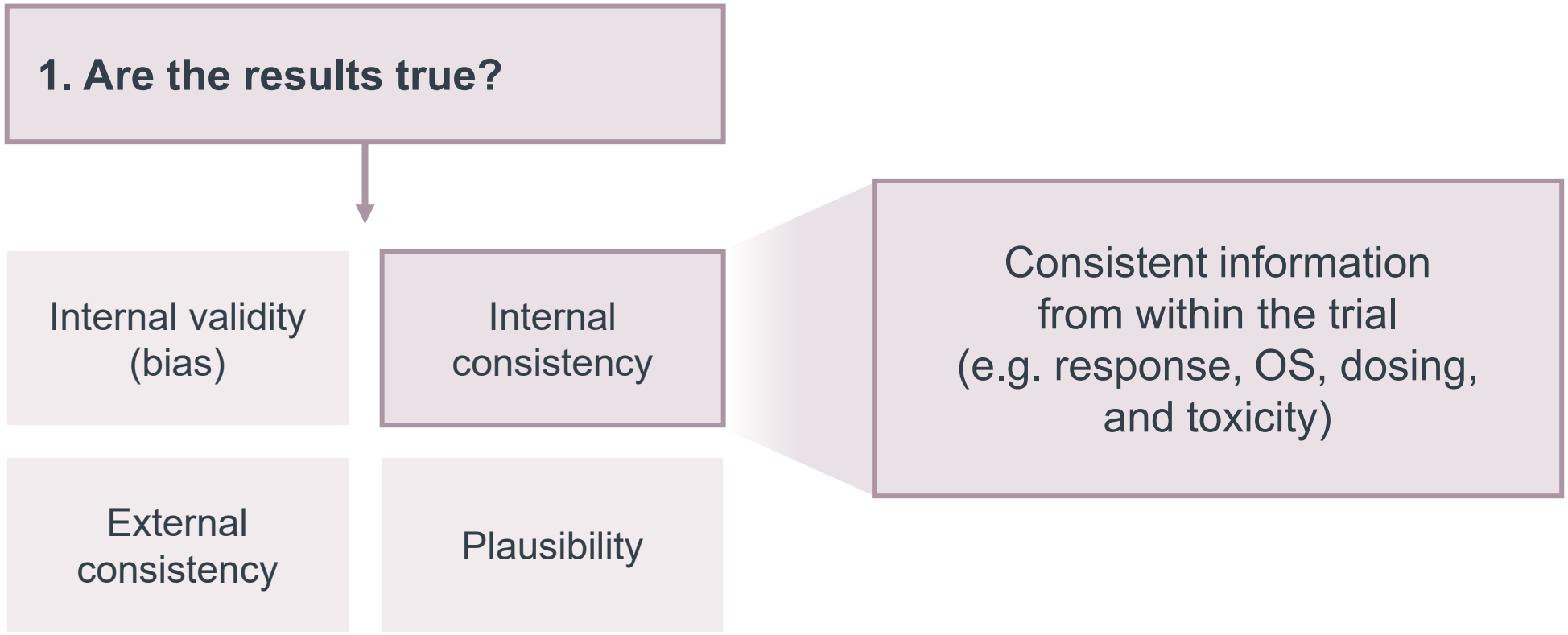
- Internal validity (bias)
- Internal consistency
- External consistency
- Plausibility

2. So what?

- External validity
- Relevance

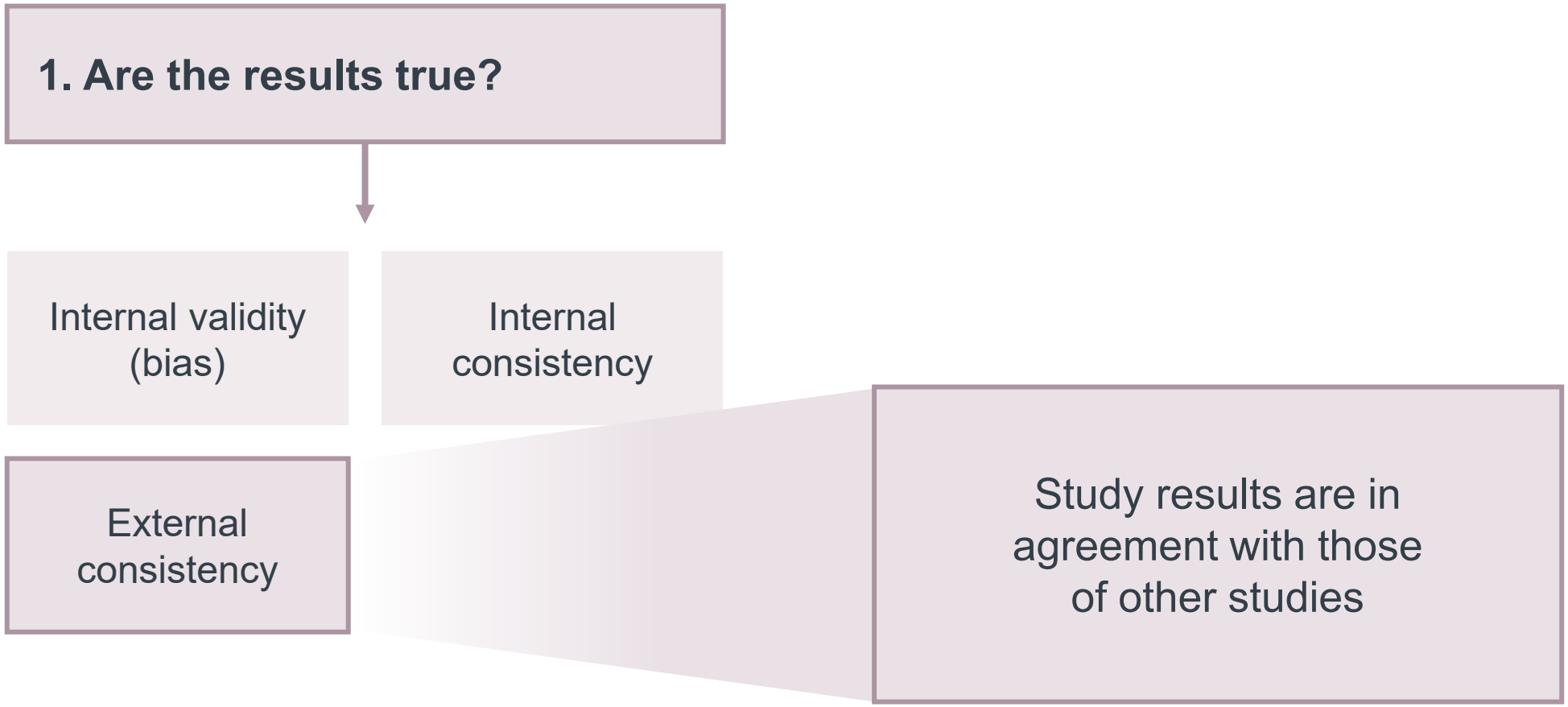
Slide courtesy of Paolo Bruzzi.

When you assess the results of a study, you should ask two questions



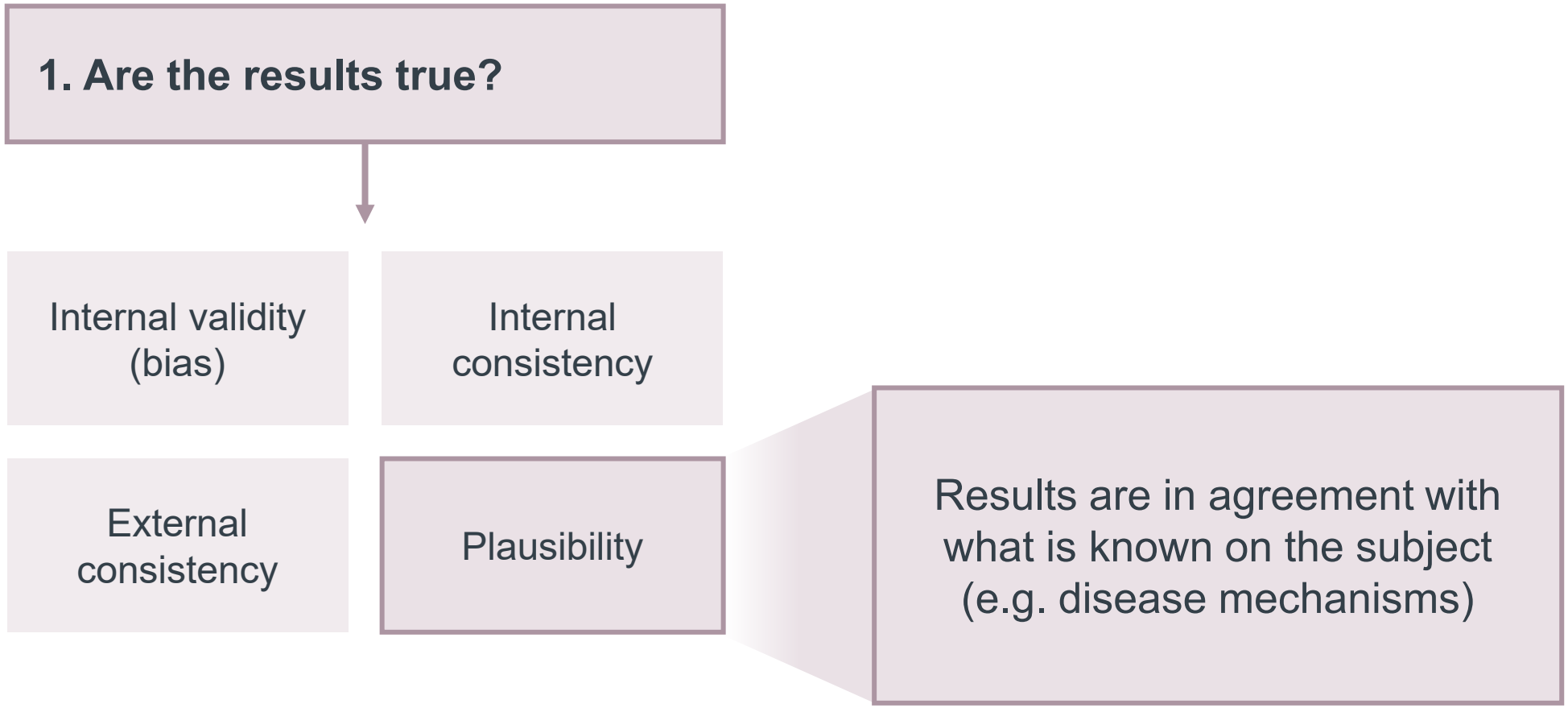
OS, overall survival.
Slide courtesy of Paolo Bruzzi.

When you assess the results of a study, you should ask two questions



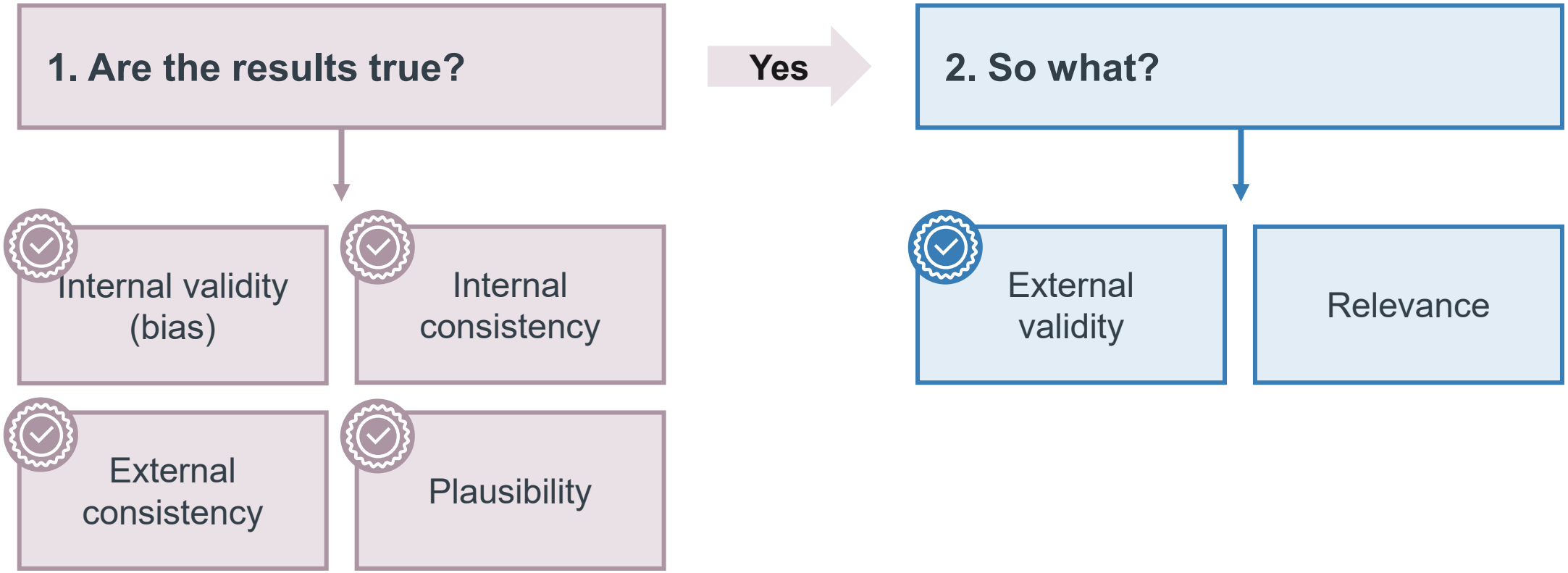
Slide courtesy of Paolo Bruzzi.

When you assess the results of a study, you should ask two questions



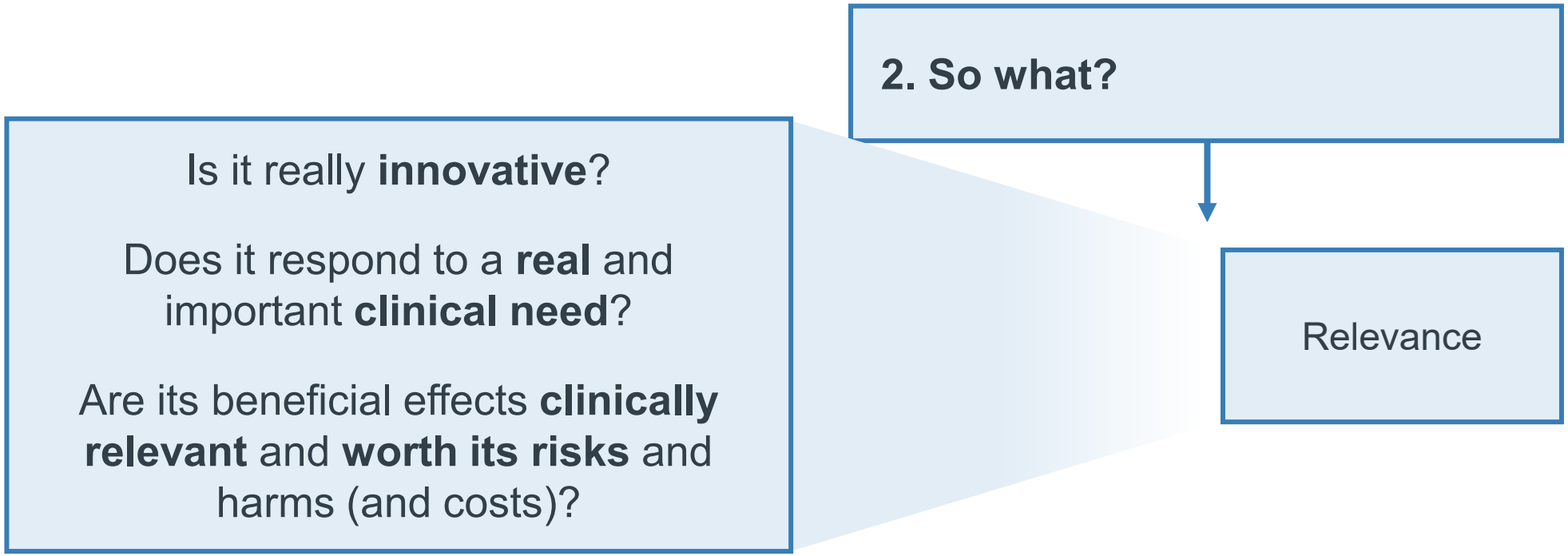
Slide courtesy of Paolo Bruzzi.

Process for understanding clinical results



Slide courtesy of Paolo Bruzzi.

Relevance of a new technology



Slide courtesy of Paolo Bruzzi.

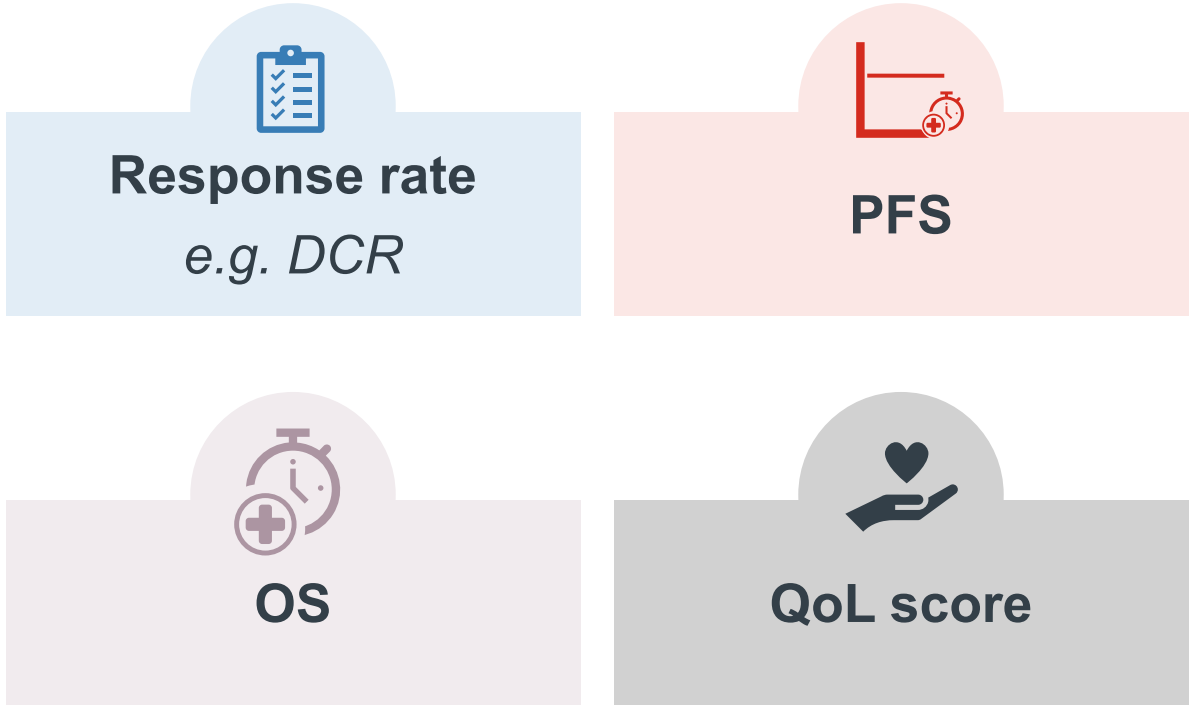
Two questions

1. Which endpoint?

2. Which summary indicator?

Slide courtesy of Paolo Bruzzi.

Endpoints in trials on advanced tumors



DCR, disease control rate; OS, overall survival; PFS, progression-free survival; QoL, quality of life.
Slide courtesy of Paolo Bruzzi.

Two questions

1. Which endpoint?

2. Which summary indicator?

- Increase in median OS, PFS, etc.?
- HR?
- Increase in percentage of long-term survivors?

HR, hazard ratio; OS, overall survival; PFS, progression-free survival.
Slide courtesy of Paolo Bruzzi.

Advanced disease

Treatment aims	Summary indicator
Palliate symptoms / reduce toxicity	Mean **QoL score**, etc.?
Postpone disease **progression and death**	**HR** **Increase in time to…**
Increase probability of **long-term survival**	**% progression-free** or **% alive** ‘long term’

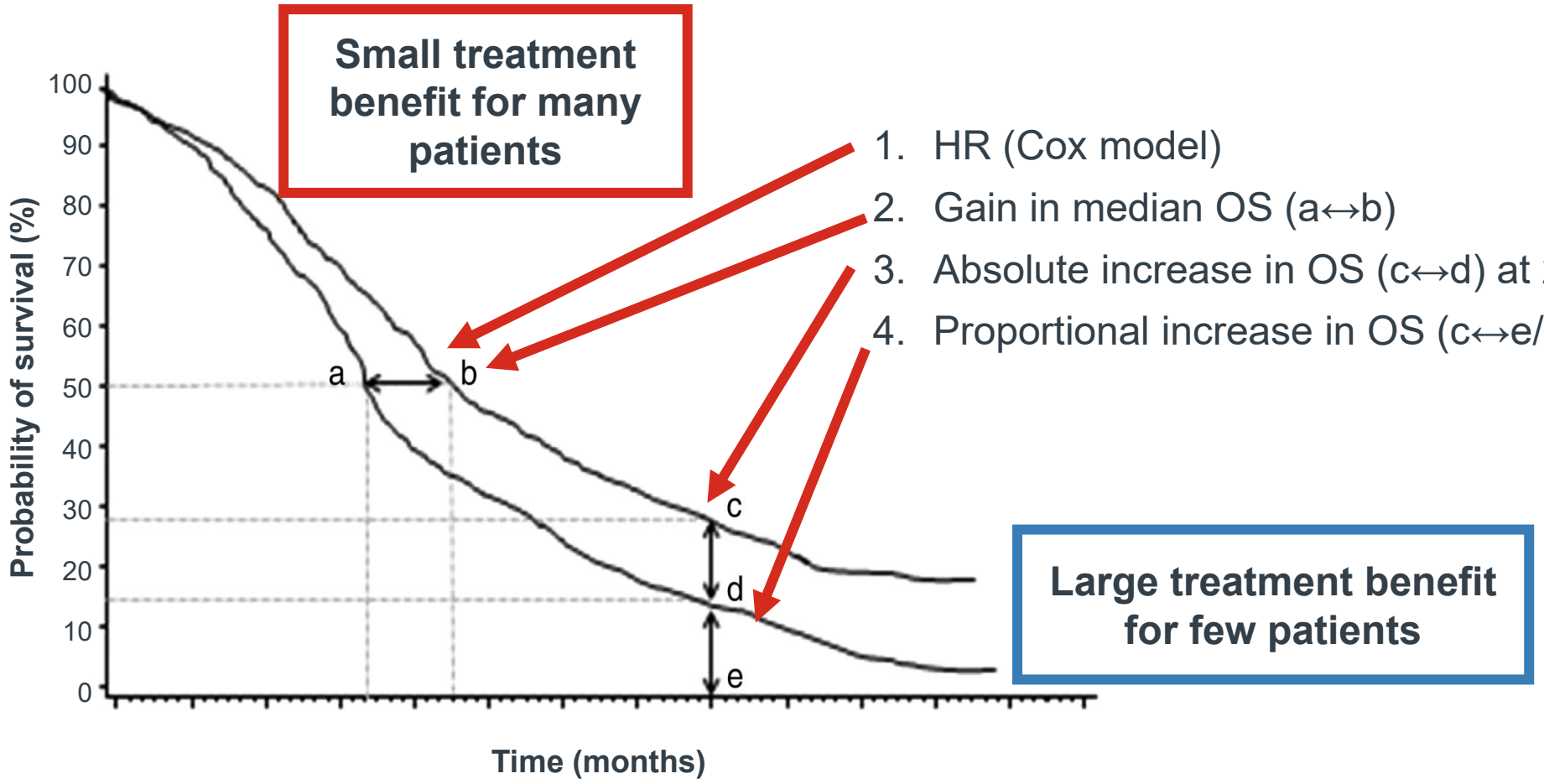
HR, hazard ratio; QoL, quality of life.
Slide courtesy of Paolo Bruzzi.

What does each summary indicator tell us?

Summary indicator	Meaning
Small treatment benefit for many patients	
Increase in **median time to event** (in restricted mean survival time)	**All** patients have the same (≈) **absolute** gain (e.g. 3 years)
HR	**All** patients have the same (≈) **proportional gain** (e.g. +50%)
Large treatment benefit for few patients	
Increase in the **proportion of long-term survivors** (e.g. OS, PFS)	**Few** patients have **large benefit** (become long-term survivors); most do not have any benefit

HR, hazard ratio; OS, overall survival; PFS, progression-free survival.
Slide courtesy of Paolo Bruzzi.

Understanding OS related parameters



1. HR (Cox model)
2. Gain in median OS (a↔b)
3. Absolute increase in OS (c↔d) at 2–3 years
4. Proportional increase in OS (c↔e/d↔e) at 2–3 years

Large treatment benefit for few patients

HR, hazard ratio; OS, overall survival.

Sobrero AF *et al*. *Clin Cancer Res* 2015; 21 (5): 1036–1043.

Small treatment benefit for many patients

Two types of effect:

1. KM curves **diverge early** and then get closer and **become parallel** or even **converge** (banana-like curves)

Increase in median (mean) survival is an appropriate measure of treatment effect

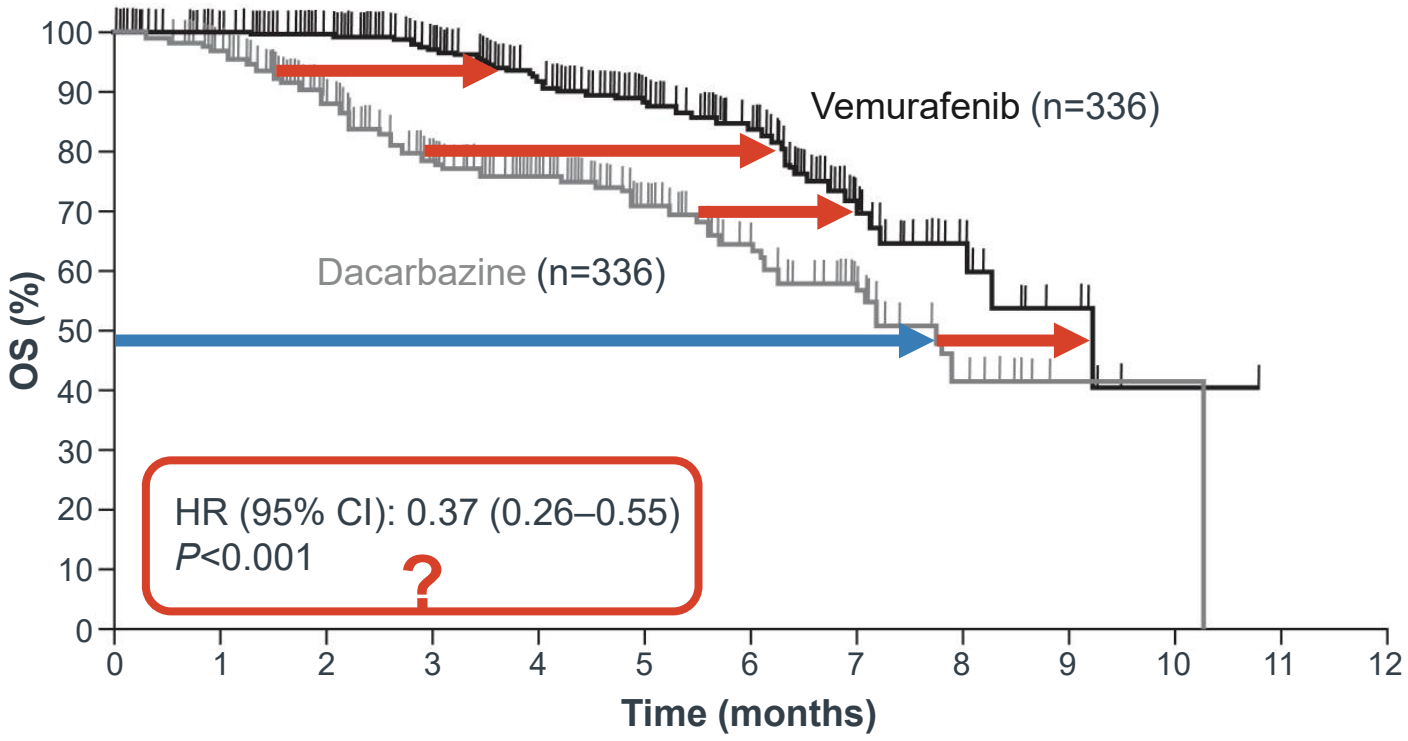
However:

- HR is inappropriate, as it does not remain constant
- The difference in percentage of progression-free/alive patients is an inappropriate measure, as it does not remain constant

HR, hazard ratio; KM, Kaplan–Meier.
Slide courtesy of Paolo Bruzzi.

Small treatment benefit for many patients

OS is prolonged by ≈2–3 months in ≈60% of patients



CI, confidence interval; HR, hazard ratio; OS, overall survival.
Chapman PB *et al*. *N Engl J Med* 2011; 364 (26): 2507–2516.

Small treatment benefit for many patients

Two types of effect:

1. KM curves **diverge early** and then get closer and **become parallel** or even **converge** (banana-like curves)

2. KM curves **continue to diverge**

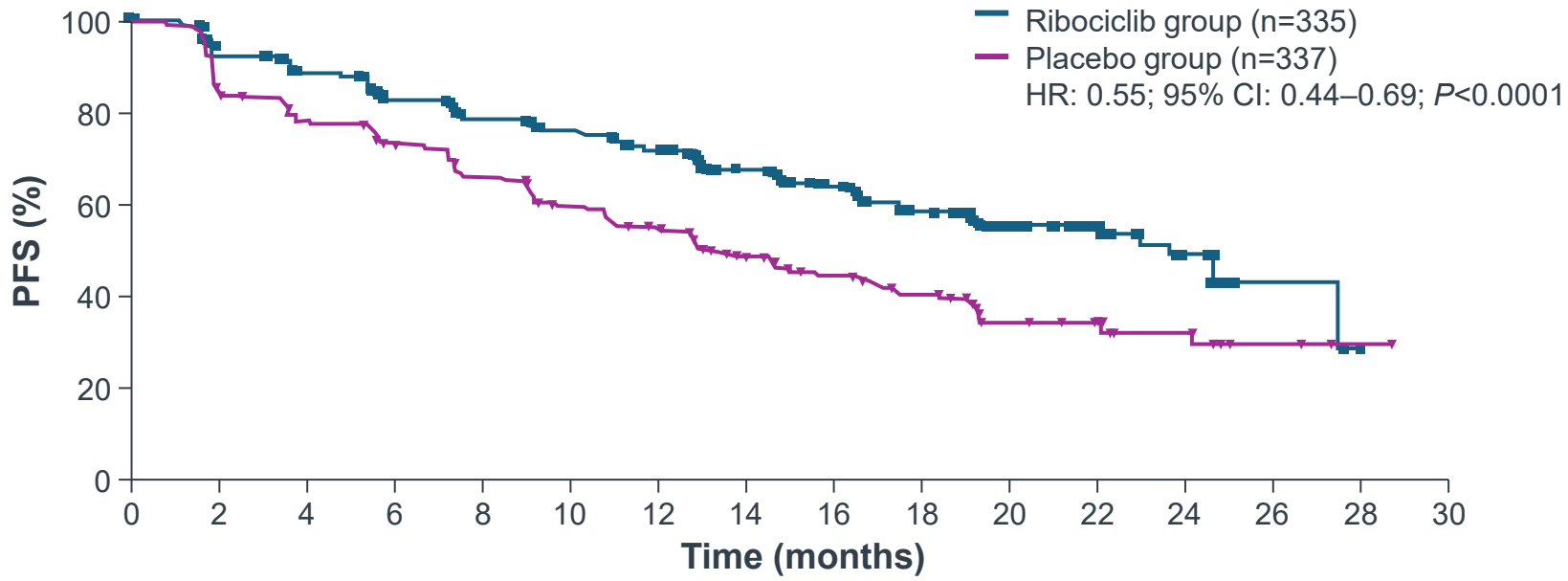
HR is an appropriate measure of treatment effect

However:

- An increase in median (mean) survival is not representative of treatment effect
- The difference in percentage of progression-free/alive patients is an inappropriate measure, as it does not remain constant

HR, hazard ratio; KM, Kaplan–Meier.
Slide courtesy of Paolo Bruzzi.

Constant HR



CI, confidence interval; HR, hazard ratio; PFS, progression-free survival.
Tripathy D *et al*. *Lancet Oncol* 2018; 19 (7): 904–915.

NOTE

If HR is (relatively) constant:

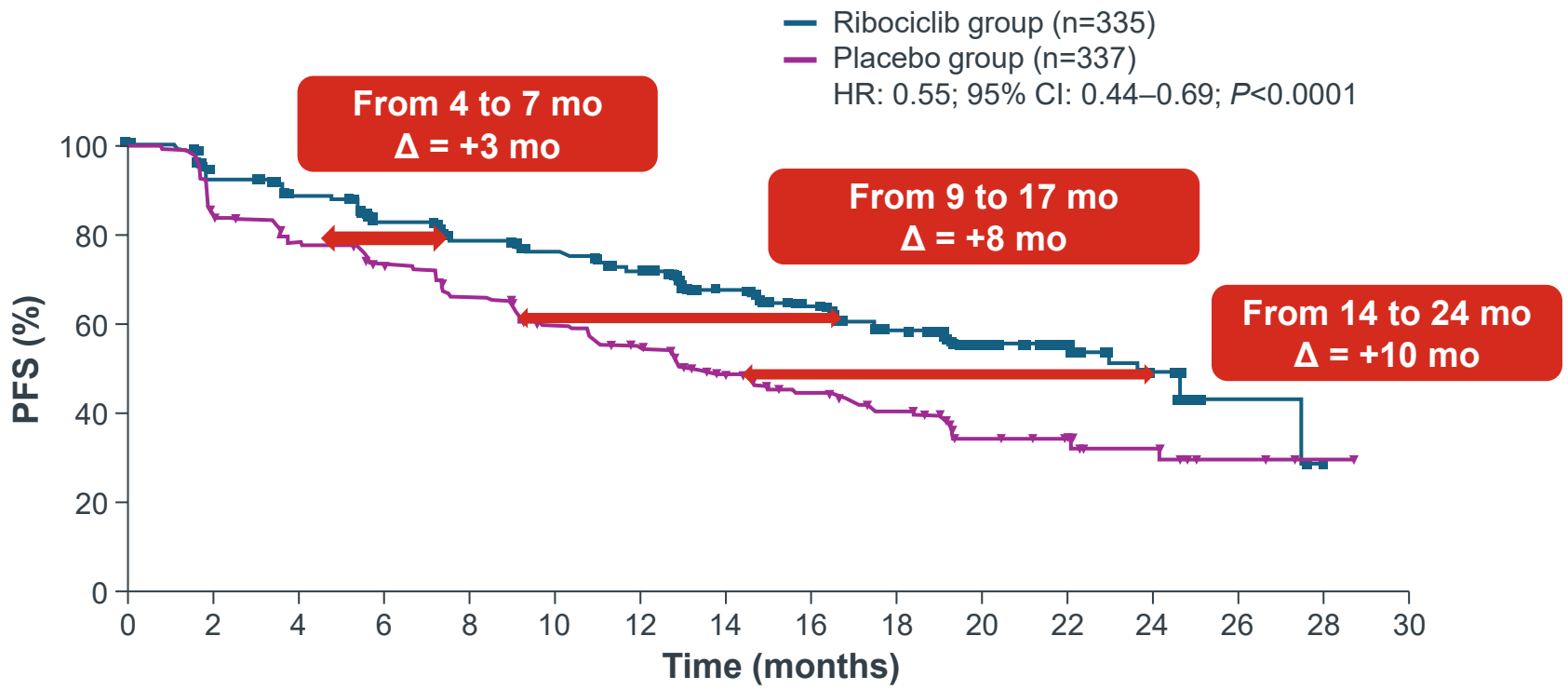
The **absolute gain in survival** increases progressively over time

The **proportional gain is constant**, and **equal to 1/HR**

HR = 0.5,	1/0.5 = 2	→	OS doubles
HR = 0.66,	1/0.66 = 1.5	→	50% gain in OS (from 6 to 9 mo, from 12 to 18 mo)
HR = 0.75,	1/0.75 = 1.33	→	33% gain in OS (from 6 to 8 mo, 12 to 16 mo)

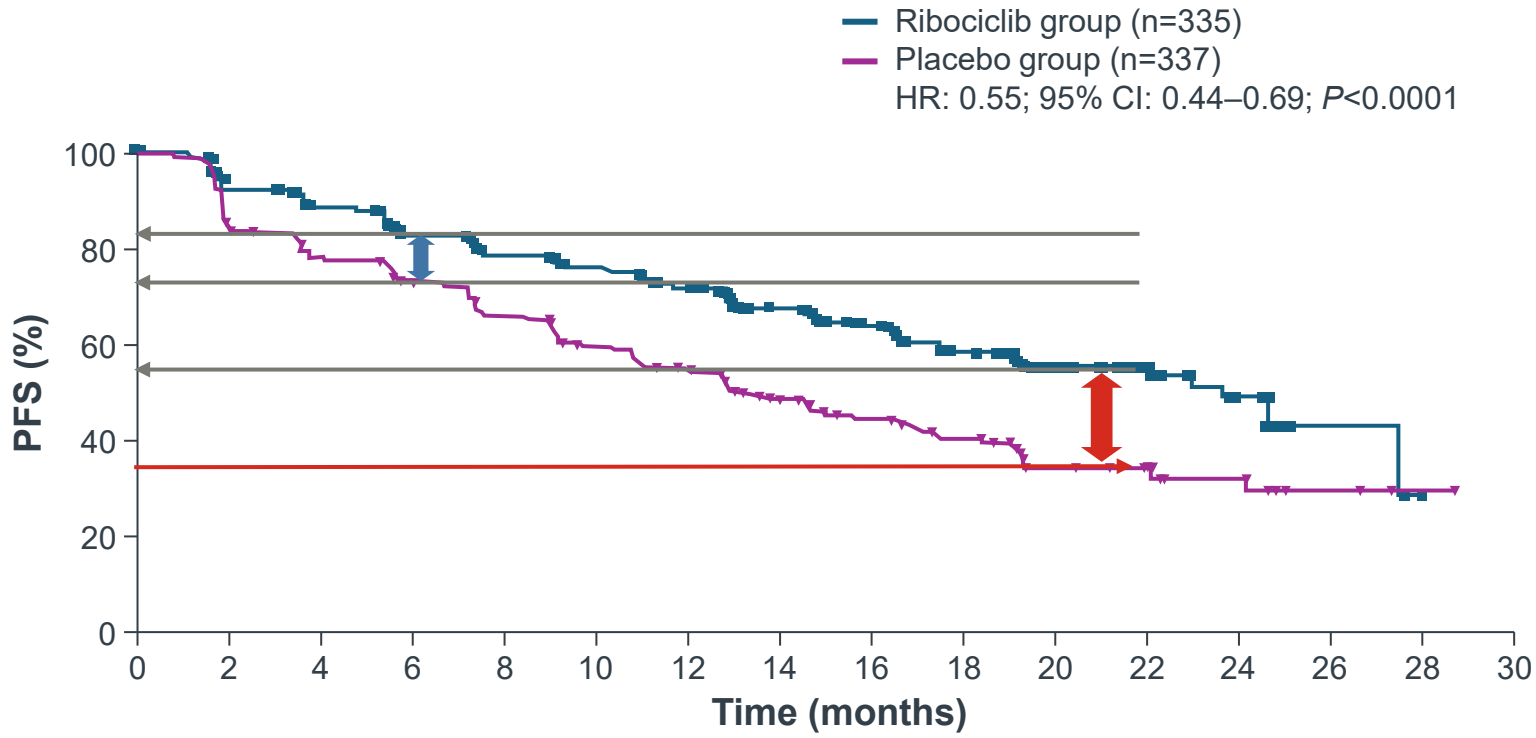
HR, hazard ratio; mo, months; OS, overall survival.
Slide courtesy of Paolo Bruzzi.

PFS: Absolute gain in time to progression?



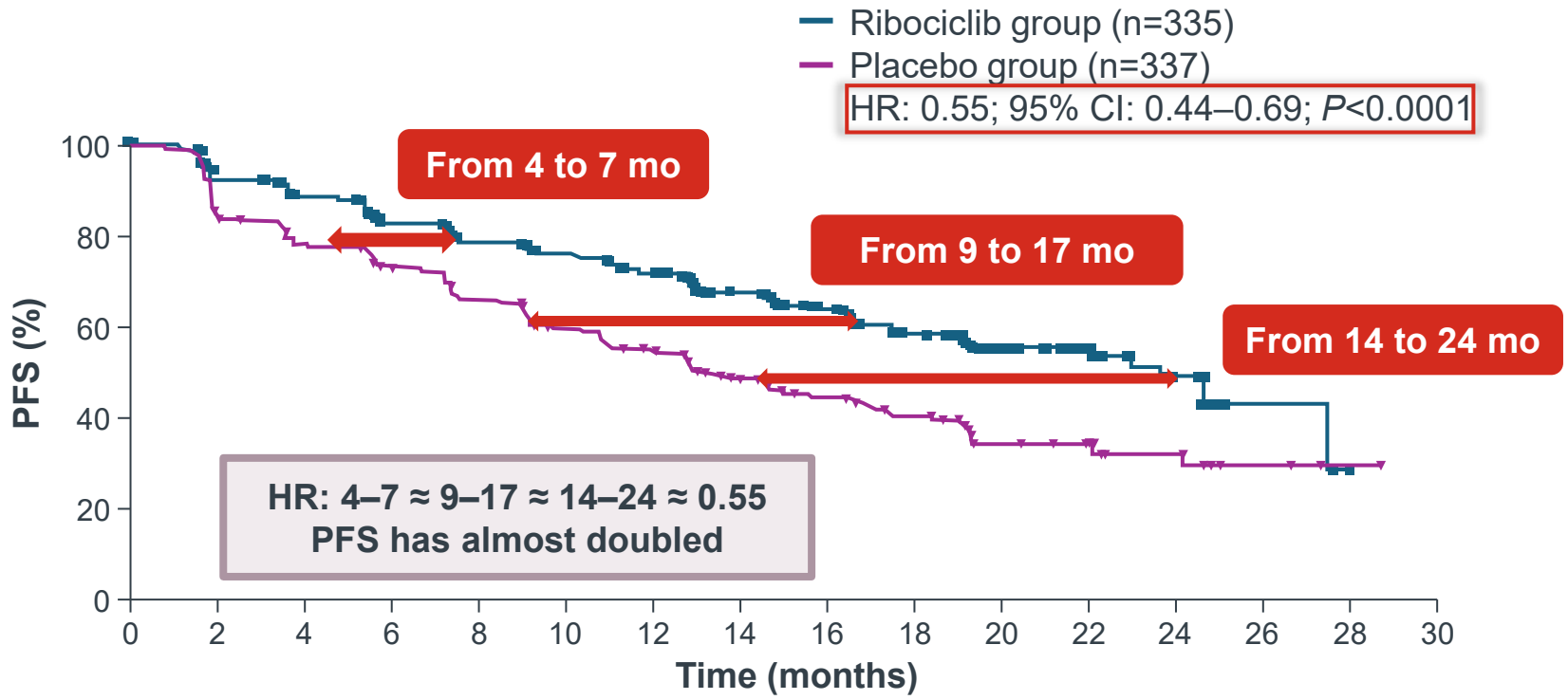
CI, confidence interval; HR, hazard ratio; mo, months; PFS, progression-free survival.
Tripathy D *et al*. *Lancet Oncol* 2018; 19 (7): 904–915.

Increase in PFS? Varies over time



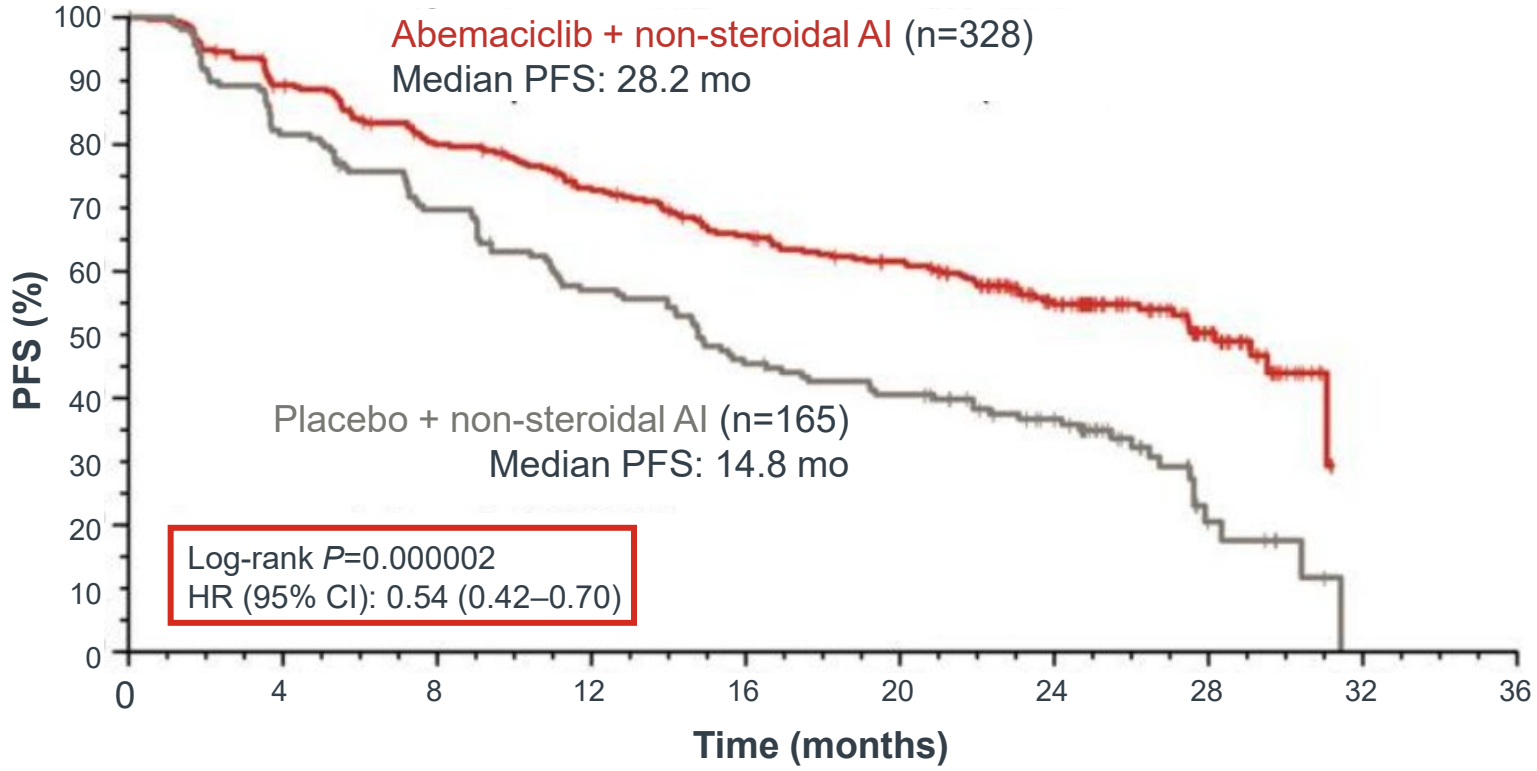
CI, confidence interval; HR, hazard ratio; PFS, progression-free survival.
Tripathy D *et al*. *Lancet Oncol* 2018; 19 (7): 904–915.

PFS: HR vs. gain in time to progression?



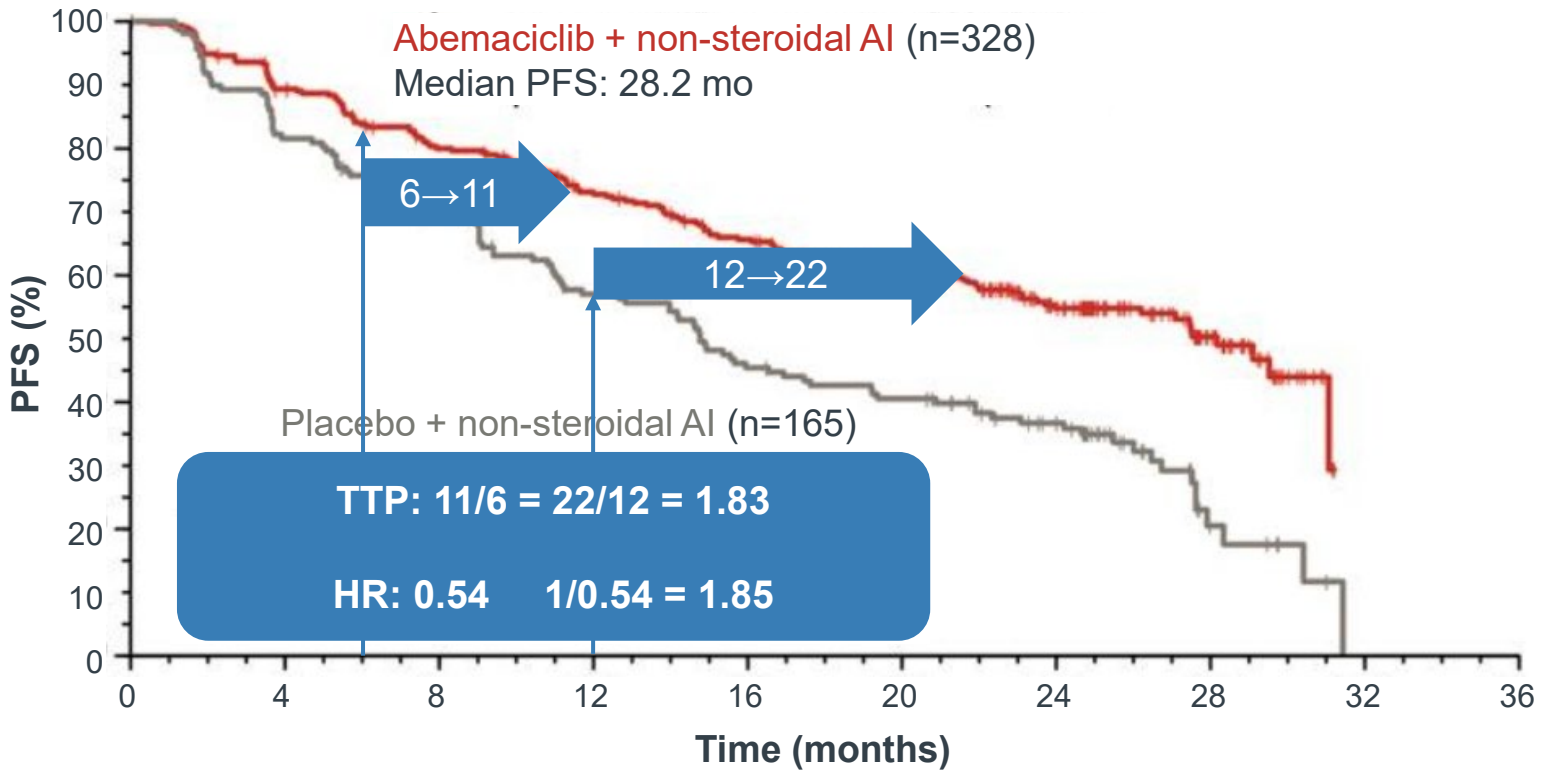
CI, confidence interval; HR, hazard ratio; mo, months; PFS, progression-free survival.
Tripathy D *et al*. *Lancet Oncol* 2018; 19 (7): 904–915.

Example: Final PFS in MONARCH 3 – a randomized study of abemaciclib as initial therapy for advanced breast cancer



AI, aromatase inhibitor; CI, confidence interval; HR, hazard ratio; mo, months; PFS, progression-free survival.
Johnston S *et al*. *NPJ Breast Cancer* 2019; 5: 5.

Example: Final PFS in MONARCH 3 – a randomized study of abemaciclib as initial therapy for advanced breast cancer



AI, aromatase inhibitor; HR, hazard ratio; mo, months; PFS, progression-free survival; TTP, time to progression.
Johnston S *et al*. *NPJ Breast Cancer* 2019; 5: 5.

Large treatment benefit for few patients

1. KM curves become parallel after a while (and ideally flat)

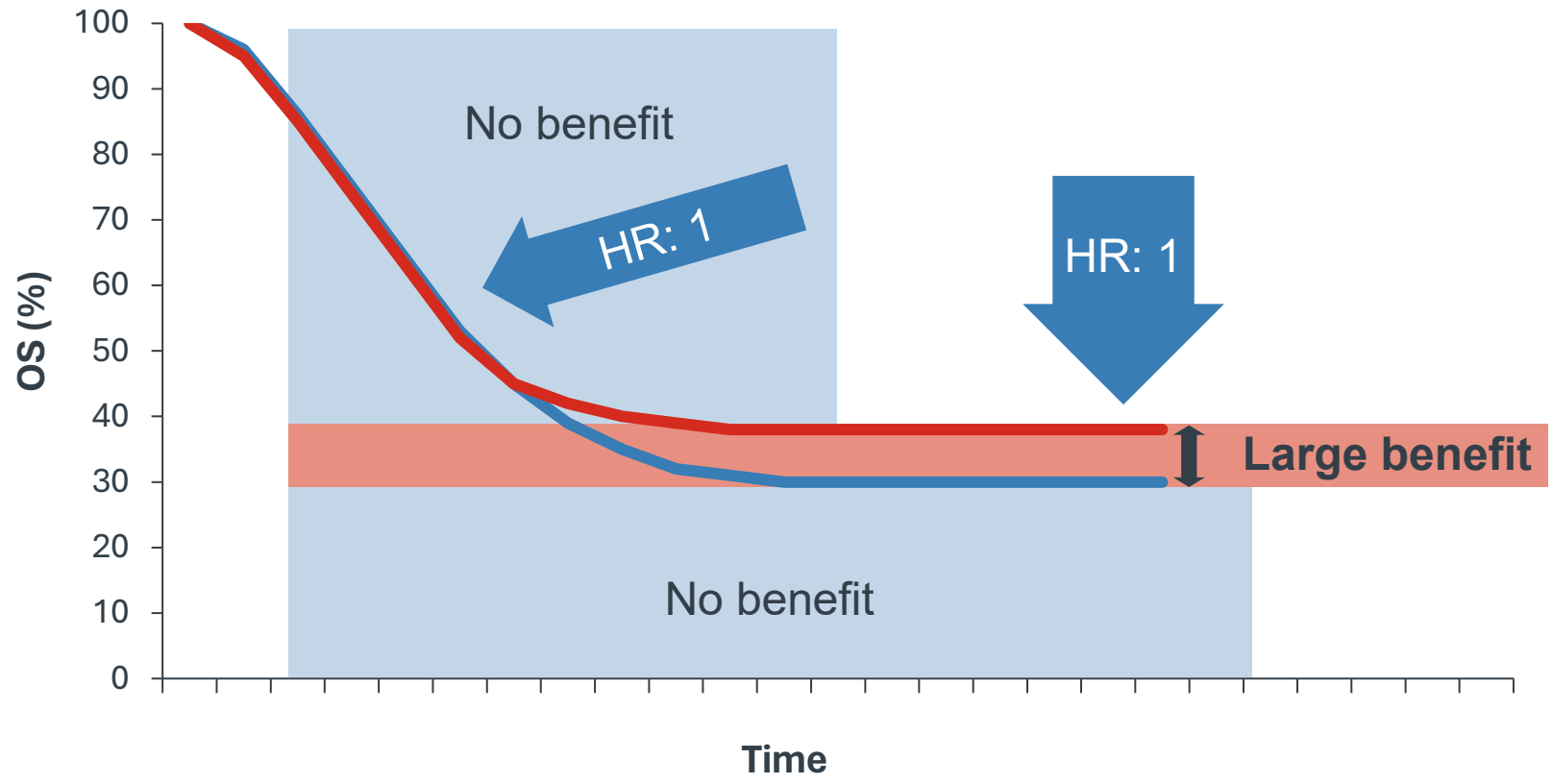
The difference in percentage of progression-free/alive patients is an appropriate measure of treatment effect

However:

- **HR** is **not appropriate** when the difference in the proportion of progression-free/alive patients is not constant
- An **increase in median/mean survival** is **not representative** of the treatment effect across the whole patient population

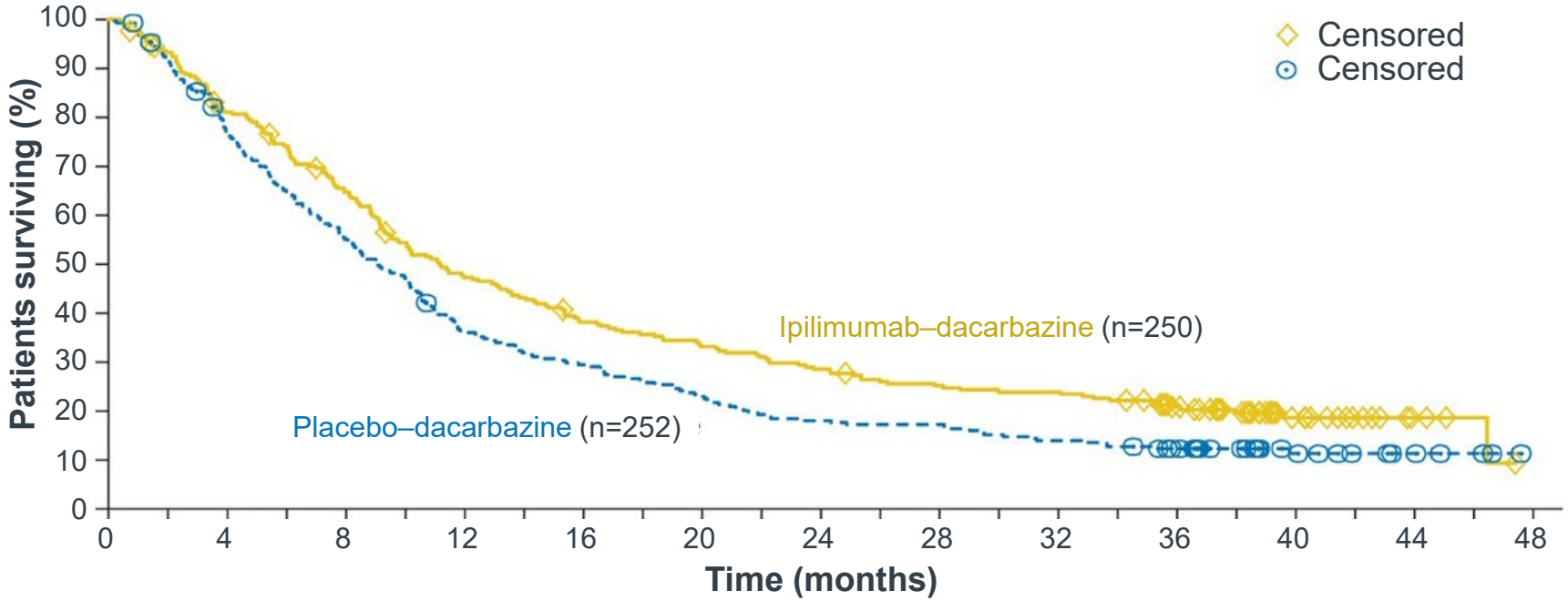
HR, hazard ratio; KM, Kaplan–Meier.
Slide courtesy of Paolo Bruzzi.

Different proportion of long-term survivors



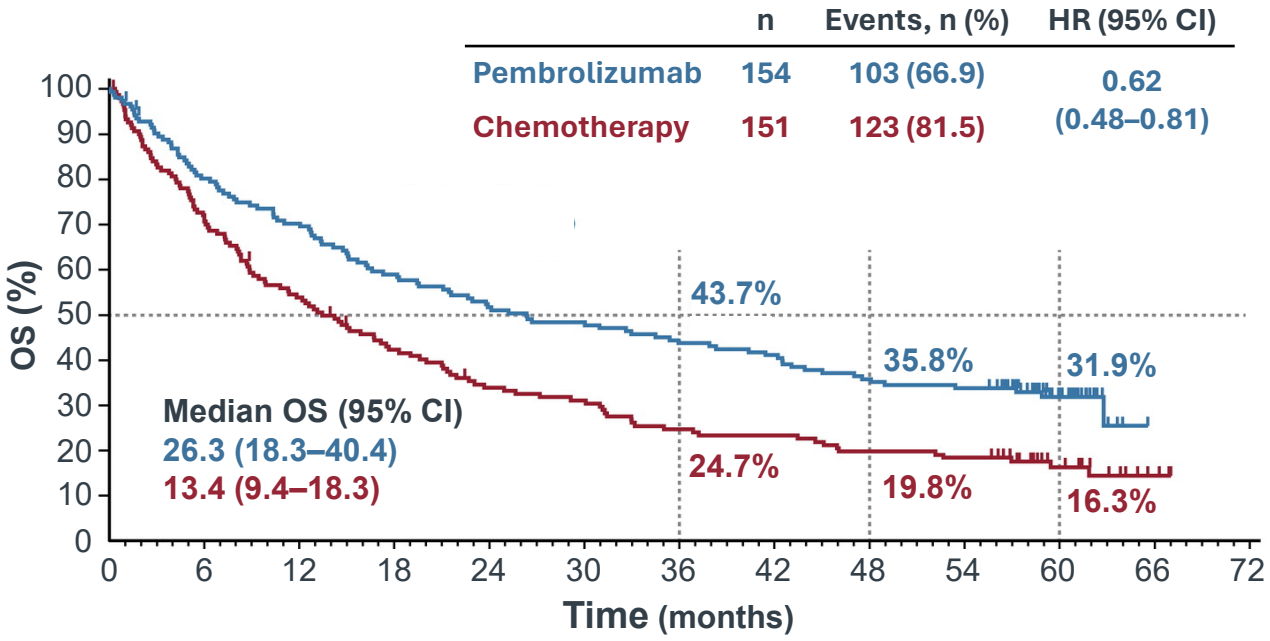
HR, hazard ratio; OS, overall survival.
Figure included for illustrative purposes only. Slide courtesy of Paolo Bruzzi.

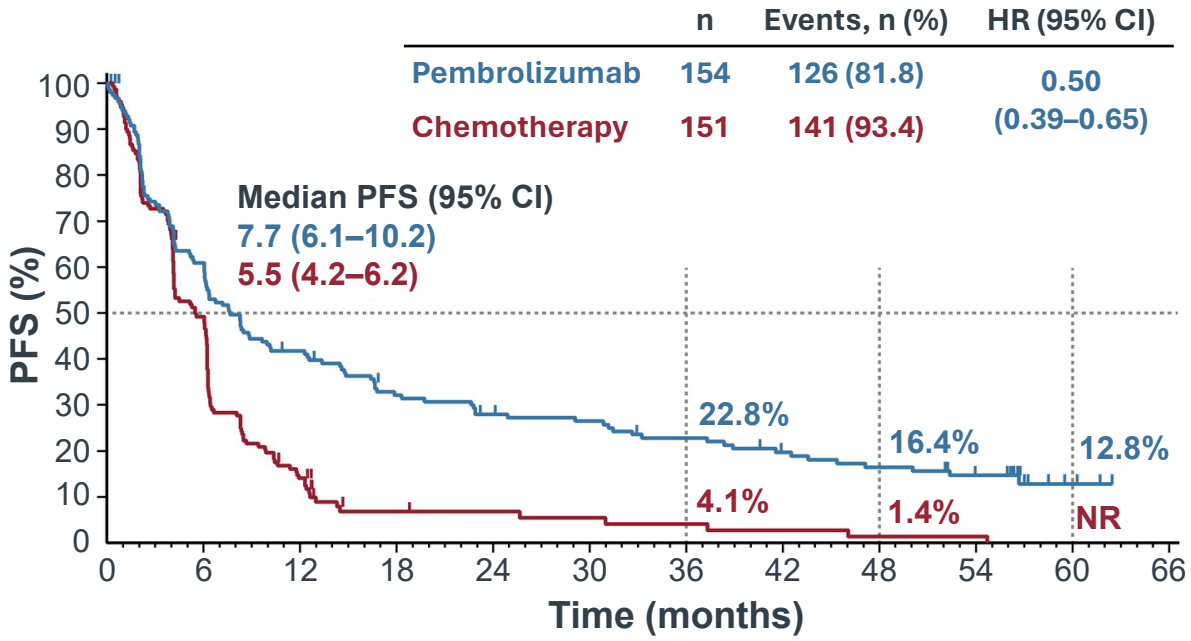
Example: Ipilimumab plus dacarbazine for previously untreated metastatic melanoma



Robert C *et al*. *N Engl J Med* 2011; 364 (26): 2517–2526.

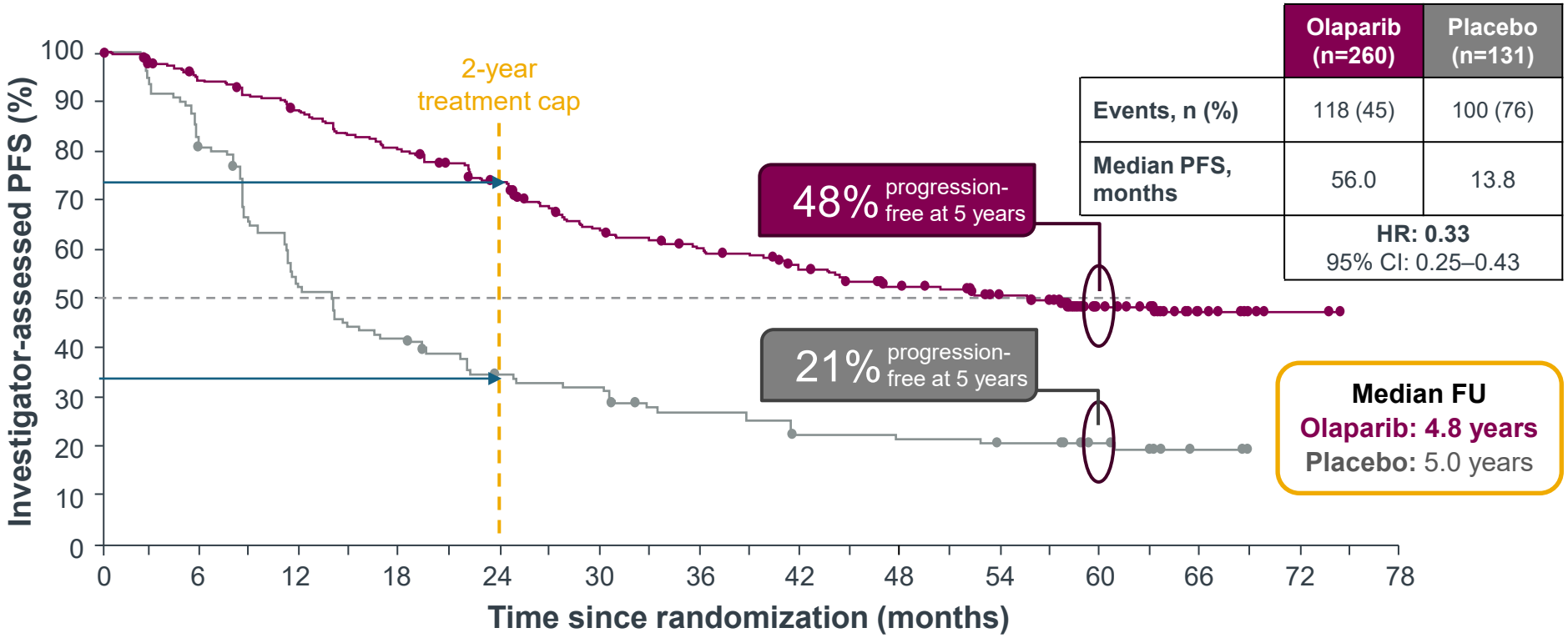
Example: 5-year outcomes with pembrolizumab vs. chemotherapy for metastatic non–small cell lung cancer





CI, confidence interval; HR, hazard ratio; NR, not reached; OS, overall survival; PFS, progression-free survival.
Reck M *et al*. *J Clin Oncol* 2021; 39 (21): 2339–2349.

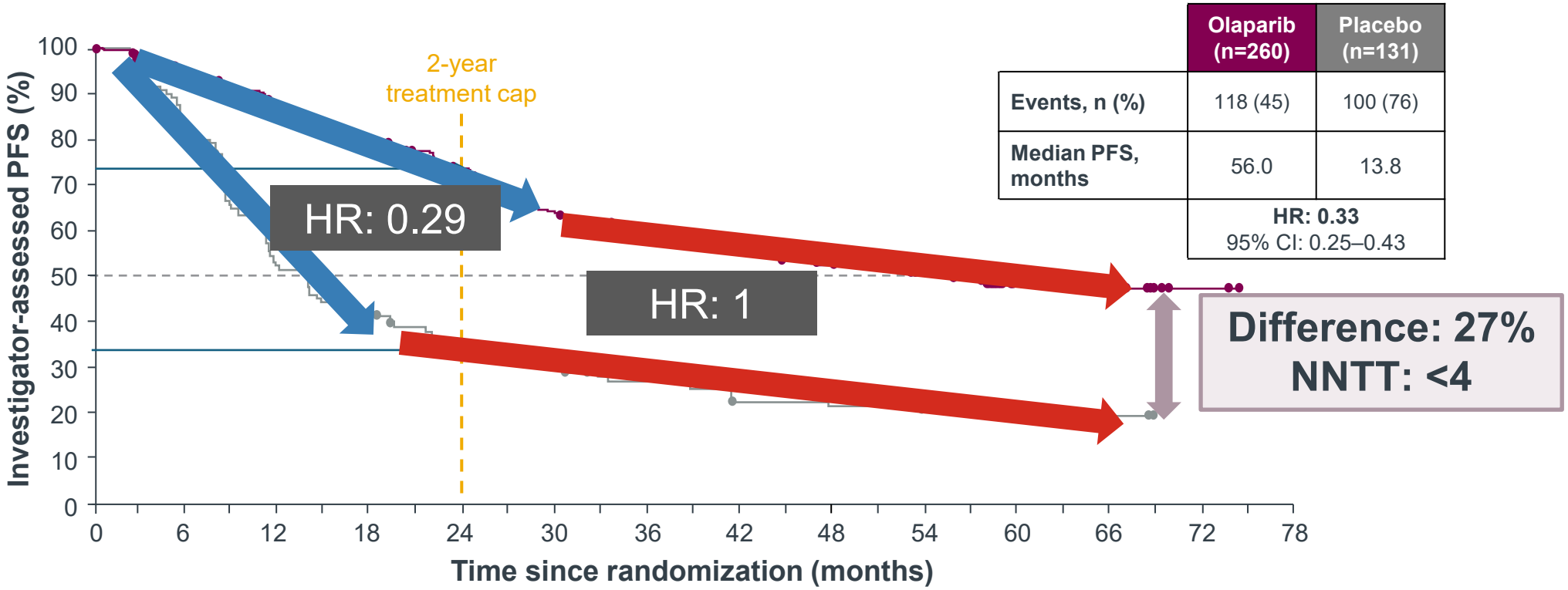
Example: Patients with OC and a *BRCAm* in response after first-line platinum-based chemotherapy derived significant PFS benefit from maintenance olaparib



	Olaparib (n=260)	Placebo (n=131)
Events, n (%)	118 (45)	100 (76)
Median PFS, months	56.0	13.8
HR: 0.33 95% CI: 0.25–0.43		

Data cut-off: March 2020. Investigator-assessed PFS. Median follow-up: olaparib, 4.8 years; placebo, 5.0 years.
CI, confidence interval; FU, follow-up; HR, hazard ratio; OC, ovarian cancer; PFS, progression-free survival.
Banerjee S *et al*. *Lancet Oncol* 2021; 22 (12): 1721–1731.

Example: After 5 years' FU, the PFS benefit derived with maintenance olaparib was sustained substantially beyond the end of treatment

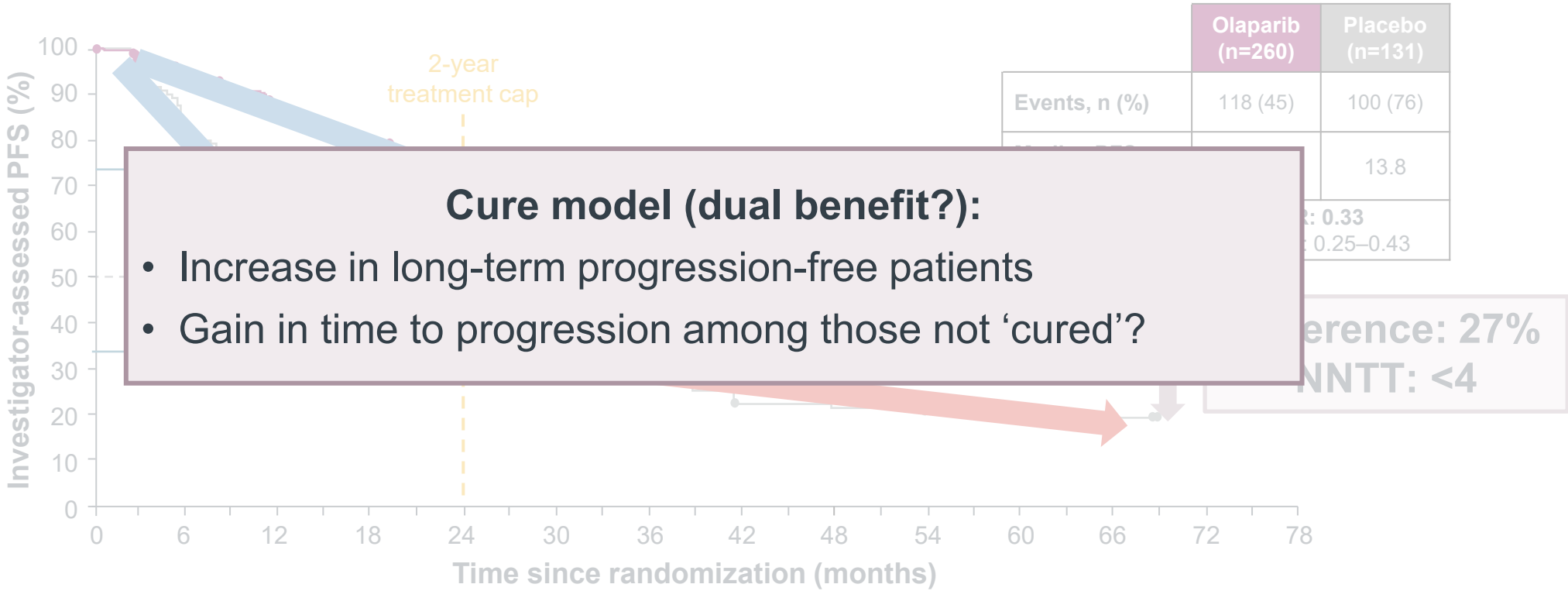


	Olaparib (n=260)	Placebo (n=131)
Events, n (%)	118 (45)	100 (76)
Median PFS, months	56.0	13.8
	HR: 0.33 95% CI: 0.25–0.43	

Difference: 27%
NNTT: <4

Data cut-off: March 2020. Investigator-assessed PFS. Median follow-up: olaparib, 4.8 years; placebo, 5.0 years.
CI, confidence interval; FU, follow-up; HR, hazard ratio; NNTT, number needed to treat; PFS, progression-free survival.
Banerjee S *et al*. *Lancet Oncol* 2021; 22 (12): 1721–1731.

Example: After 5 years' FU, the PFS benefit derived with maintenance olaparib was sustained substantially beyond the end of treatment



Cure model (dual benefit?):

- Increase in long-term progression-free patients
- Gain in time to progression among those not 'cured'?

Data cut-off: March 2020. Investigator-assessed PFS. Median follow-up: olaparib, 4.8 years; placebo, 5.0 years.
CI, confidence interval; FU, follow-up; HR, hazard ratio; NNTT, number needed to treat; PFS, progression-free survival.
Banerjee S *et al*. *Lancet Oncol* 2021; 22 (12): 1721–1731.

Conclusions

The summary indicators commonly used to describe the effect of anticancer treatments have different meanings and are not interchangeable

Proportion of long-term surivors

HR

Mean time to event

Understanding the type(s) of effect(s) of a specific treatment has important clinical implications *(e.g. small-for-many vs. large-for-few)*

Summary indicator	Meaning
Small treatment benefit for many patients	
Increase in **median time to event** (in restricted mean survival time)	**All** patients have the same (≈) **absolute** gain (e.g. 3 years)
HR	**All** patients have the same (≈) **proportional gain** (e.g. +50%)
Large treatment benefit for few patients	
Increase in the **proportion of long-term survivors** (e.g. OS, PFS)	**Few** patients have **large benefit** (become long-term survivors); most do not have any benefit

DCR, disease control rate; HR, hazard ratio; OS, overall survival; PFS, progression-free survival; QoL, quality of life.
Slide courtesy of Paolo Bruzzi.

2. Statistical plan, statistical significance, and multiplicity

Multiplicity

The general issue

Multiple endpoints

Solutions

Interim analyses
(subgroup analyses)

Slide courtesy of Paolo Bruzzi.

Statistical plan

A detailed statistical plan should be prepared before the start of the trial

(If the trial is double-blind, the plan can be prepared after the start of the study but before treatment codes are opened)

Slide courtesy of Paolo Bruzzi.

Statistical plan

- Contents:
  - Primary and secondary endpoints
  - Statistical analysis plan (contrasts, tests, populations, subgroup analysis)
  - Interim analyses
  - Significance levels
  - Power – sample size
  - Summary indicators

WHY?

Slide courtesy of Paolo Bruzzi.

Why is a statistical plan needed?

α, alpha.
Slide courtesy of Paolo Bruzzi.

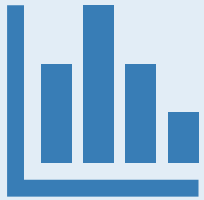
Multiplicity

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

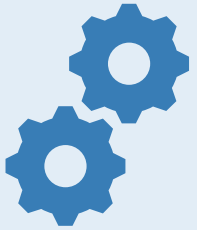
- Five consecutive reds at the roulette wheel
- Two cases with the same inherited mutation
- Three long-term survivors with advanced NSCLC

NSCLC, non–small cell lung cancer.
Slide courtesy of Paolo Bruzzi.

Analysis of RCTs: Reference criteria



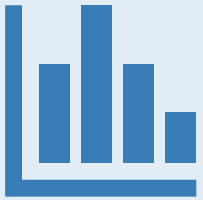
Statistical



Methodological

RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Analysis of RCTs: Reference criteria



Statistical

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

MULTIPLICITY

RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

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.)

AE, adverse event; QoL, quality of life.
Slide courtesy of Paolo Bruzzi.

The problem with multiplicity

- Statistical *P*: Probability to observe a difference as large as (or larger than) the one observed by chance if the null hypothesis (H0) is true (the two therapies have the same efficacy)

Example:

Arm A Responses: 30/40 (75%)

Arm B Responses: 20/40 (50%); *P*=0.02

- The probability of observing a difference of this size **because of chance alone** is 2%
- Conventional significance level to reject H0: 5%

H0, null hypothesis.
Slide courtesy of Paolo Bruzzi.

Therefore…

- We consider a difference as ‘real’ (i.e. not due to chance) when the probability that it occurred by chance alone is <5%

- Significance level = frequency of false-positive analyses among all those in which H0 is true

5%

Out of 100 studies (analyses) comparing treatments with the same efficacy, 5 studies will show a benefit of one treatment over the other just by chance (sampling error)

H0, null hypothesis.
Slide courtesy of Paolo Bruzzi.

Consequences

- If more than one test is conducted in a study, the probability that at least one of them shows a statistically significant difference is >5%

The probability of a false-positive test increases with an increasing number of tests

Slide courtesy of Paolo Bruzzi.

How much does it increase?

If independent tests → Rules of probability → Union

or

If correlated tests → It depends on the extent of the correlation

Slide courtesy of Paolo Bruzzi.

Risk of ≥1 false-positive test if independent

Number of tests	Probability $P<0.05$
1	5%
2	9.75%
3	14.3%
5	23%
10	40%
20	64%
40	87%

Slide courtesy of Paolo Bruzzi.

Example

- Drug A vs. Drug B in a cancer
- Three factors (e.g. sex, <50 or >50 years old, Stage I/II)
- Probability of a significant difference despite A = B:
  - 5% in primary analysis
  - 30% with 7 analyses (primary + 3×2 subgroups)
  - **75% with all 27 possible combinations of the three factors**
- How many factors can be examined in any disease?
  - Patient/disease stage/biology, clinical history, etc.

Slide courtesy of Paolo Bruzzi.

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

Slide courtesy of Paolo Bruzzi.

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**

Slide courtesy of Paolo Bruzzi.

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%)

α, alpha.
Slide courtesy of Paolo Bruzzi.

Analysis of results: Strategy

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

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.)

Slide courtesy of Paolo Bruzzi.

Question

In ALPINE, the sample size was adjusted because the anticipated effect size of ibrutinib was changed based on new data.

Does this change introduce a bias?

Brown JR *et al. N Engl J Med* 2023; 388 (4): 319–332 – supplementary appendix.

Statistical plan (1)

- **The statistical analysis plan should be prepared blinded to the study results**

It should include all planned analyses with:

- ❑ Endpoints and summary indicators
- ❑ Contrasts analyzed for statistical significance
- ❑ Statistical tests with *P*-values used for rejecting the null hypothesis in each analysis
- ❑ Timing (number of events) of each analysis

Slide courtesy of Paolo Bruzzi.

Statistical plan (2)

- Changes in the statistical analysis plan (adaptations) are permitted as long as they do not introduce any bias into the trial

- Examples:
  - Adaptations of eligibility criteria[1]
  - **Adaptations to maintain study power**[1] (ALPINE[2])
    - Based on **blinded interim analyses** of aggregate data

No bias

No need to be (but better if) planned in advance

1. Adaptive design clinical trials for drugs and biologics guidance for industry. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry. Accessed February 2024. 2. Brown JR *et al. N Engl J Med* 2023; 388 (4): 319–332 – supplementary appendix.

Bias-free adaptations to maintain study power

A. Based on blinded interim analyses

- Low event rate (increase sample size)
- Greater than expected proportion of patients with a good prognosis
- Lower than expected proportion of patients in a critical subgroup

B. Based on external information from other trials (ALPINE)

Slide courtesy of Paolo Bruzzi.

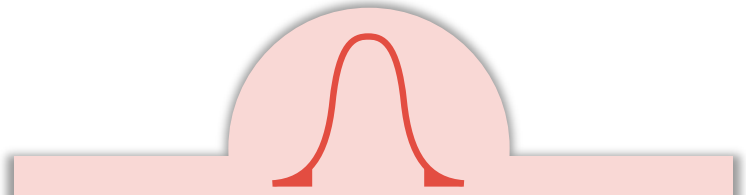
Statistical plan (3)

- Adaptations based on **unblinded analyses**:

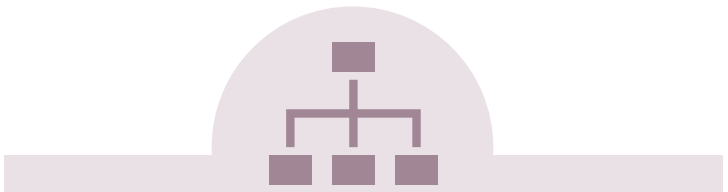
- For stopping early
- For dose selection studies
- Of patient subgroups based on treatment-effect estimates
- For endpoint selection based on interim estimates of treatment effect
- **Bias if not addressed in study design**

Slide courtesy of Paolo Bruzzi.

Methods to control multiplicity



1
Adjusted *P*-values

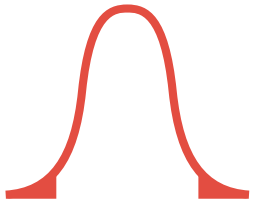


2
Hierarchical test procedures

3
Closed test procedures

Slide courtesy of Paolo Bruzzi.

1. Adjusted *P*-values

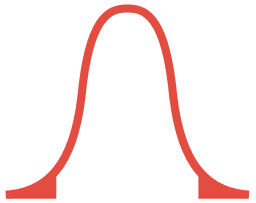


- Note:
  - Adjusted *P*-values
  - Alpha spending function
  - Alpha split
  - Others

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%

Slide courtesy of Paolo Bruzzi.

1. 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)

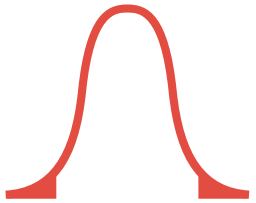
NB the correct formula is:

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

α, alpha.
Slide courtesy of Paolo Bruzzi.

1. Adjusted *P*-values



Single-step methods

- Examples: Bonferroni, Simes, Dunnett, …

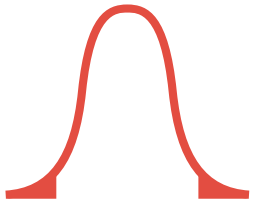
Stepwise methods

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

- Examples: Holm, Hochberg, step-down Dunnett…

Slide courtesy of Paolo Bruzzi.

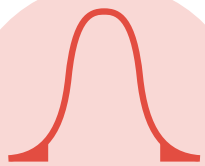
1. Adjusted *P*-values



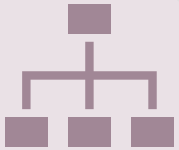
- **All methods involve a loss of statistical power**
  - Limit the number of analyses
  - Increase the sample size
  - Use hierarchical procedures
  - The Bonferroni method is the most conservative (least powerful)
    - Other more complex methods generally used

Slide courtesy of Paolo Bruzzi.

Methods to control multiplicity



1
Adjusted *P*-values

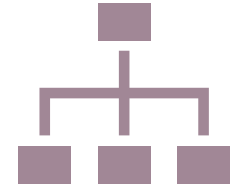


2
Hierarchical test procedures

3
Closed test procedures

Slide courtesy of Paolo Bruzzi.

2. Hierarchical test procedures



- Hypotheses are ordered in sequence and tested at level α until the first non-rejection

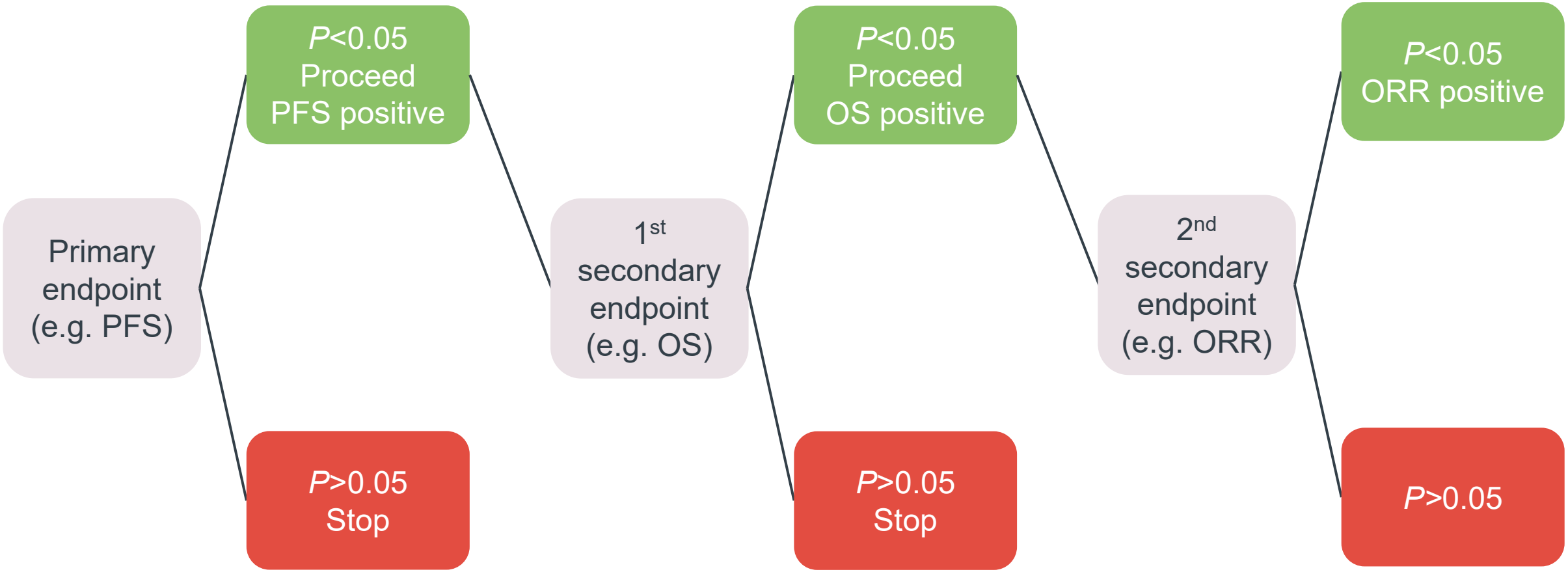
- In practice, first test:
  - If *P*>0.05 → stop (negative study)
  - If positive → second test

- **NO CORRECTION OF THE SIGNIFICANCE LEVEL IS REQUIRED**

- Sequence based on relevance, power, plausibility, etc.

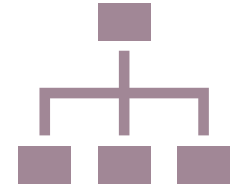
α, alpha.
Slide courtesy of Paolo Bruzzi.

2. Hierarchical test procedures (example)



ORR, overall response rate; OS, overall survival; PFS, progression-free survival.
Slide courtesy of Paolo Bruzzi.

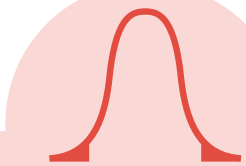
2. Hierarchical test procedures



- 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)

OS, overall survival; PFS, progression-free survival.
Slide courtesy of Paolo Bruzzi.

Methods to control multiplicity



1
Adjusted *P*-values

2
Hierarchical test procedures

3
Closed test procedures

Slide courtesy of Paolo Bruzzi.

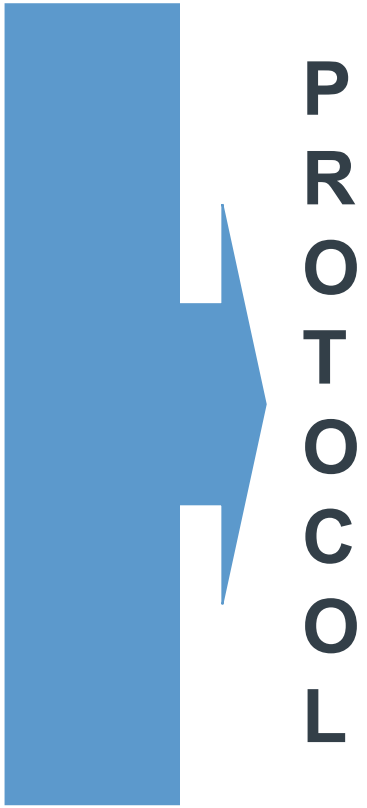
3. Closed test procedures

- 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

α, alpha.
Slide courtesy of Paolo Bruzzi.

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**



Slide courtesy of Paolo Bruzzi.

Question

ELEVATE-TN

Acalabrutinib ± obinutuzumab vs obinutuzumab + chlorambucil in treatment-naive chronic lymphocytic leukemia: 6-year follow-up of Elevate-TN[1]

What do the *P*-values for PFS and OS mean for A vs. A + O when the trial was powered to compare A arms vs. chemo + O?

A, acalabrutinib; chemo, chemotherapy; O, obinutuzumab; OS, overall survival; PFS, progression-free survival.

1. Sharman JP *et al*. Oral presentation at ASH 2023; San Diego, CA, USA, December 9–12, 2023 (Abstract 636).

ELEVATE-TN

Background

"Previous reports of ELEVATE-TN … demonstrated superior efficacy of acalabrutinib (A) ± obinutuzumab (O) compared with O + chlorambucil (Clb) in patients (pts) with treatment-naive (TN) chronic lymphocytic leukemia (CLL). Herein, updated results at 74.5 mo of follow-up are reported."

Conclusions

"With a median follow-up of 74.5 mo, the efficacy and safety of A+O and A monotherapy were maintained in pts with TN CLL, including in pts with high-risk genetic features. At 6 years of follow-up, PFS was significantly longer in pts treated with A+O vs A. Median OS was NR in any treatment arm and was significantly longer in pts treated with A+O vs O+Clb."

Methods

"All analyses are ad-hoc and P-values are descriptive."

mo, months; NR, not reached; OS, overall survival; PFS, progression-free survival.

Sharman JP *et al*. Oral presentation at ASH 2023; San Diego, CA, USA, December 9–12, 2023 (Abstract 636).

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

Slide courtesy of Paolo Bruzzi.

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

Slide courtesy of Paolo Bruzzi.

Multiplicity: Cases considered

Co-primary endpoints

Interim analyses

Subgroup analyses

Slide courtesy of Paolo Bruzzi.

Illustrative trial

	Anturane		Placebo		*P*-value
	Number of patients	Number of events	Number of patients	Number of events	
Total, n	813	74	816	89	0.28
Ineligible patients, n	38	10	33	4	–
Eligible patients, n	775	64	783	85	0.10
Unanalyzable deaths, n	–	20	–	23	–
Analyzable deaths, n	–	44	–	62	0.08
Analyzable cardiac deaths, n	–	43	–	62	0.06
Analyzable sudden cardiac deaths, n	–	22	–	37	0.04
Analyzable sudden cardiac deaths in first 6 months, n	–	6	–	24	0.003

N.B. Prevention of 75% of events

Slide courtesy of Paolo Bruzzi.

Multiple endpoints

Two strategies:

A. Co-primary endpoints

- The study is positive if **either endpoint** (or both) shows a significant difference
- Correction for multiplicity
  - E.g. if two analyses, the study is positive if either or both *P*-values <0.05/2 = 0.025
- Loss of power

B. Hierarchical procedure

- Endpoints are ordered
- First endpoint tested:
  - If *P*>0.05 → stop (negative study)
  - If *P*<0.05, study is positive → proceed with subsequent endpoints until *P*>0.05
- **No correction for multiplicity required**

Using both hierarchical tests and co-primary endpoints is conceptionally wrong and self-damaging.
It results in a loss of power without any advantage in return.

Slide courtesy of Paolo Bruzzi.

Example: LUX-Lung 7

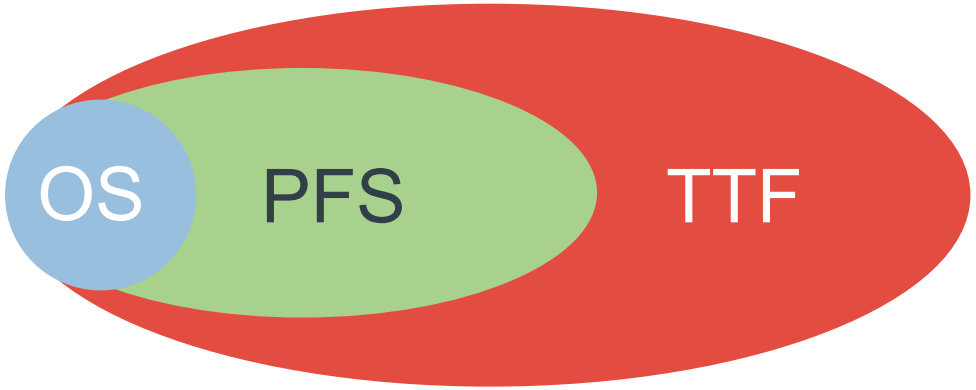
Park K *et al*. *Lancet Oncol* 2016; 17 (5): 577–589.

LUX-Lung 7: Three co-primary endpoints[1]

Two solutions:

A **At least one must be significant**

- Correction for multiplicity (e.g. $P<0.05/3$)
- Loss of power



OS, overall survival; PFS, progression-free survival; TTF, time to treatment failure.

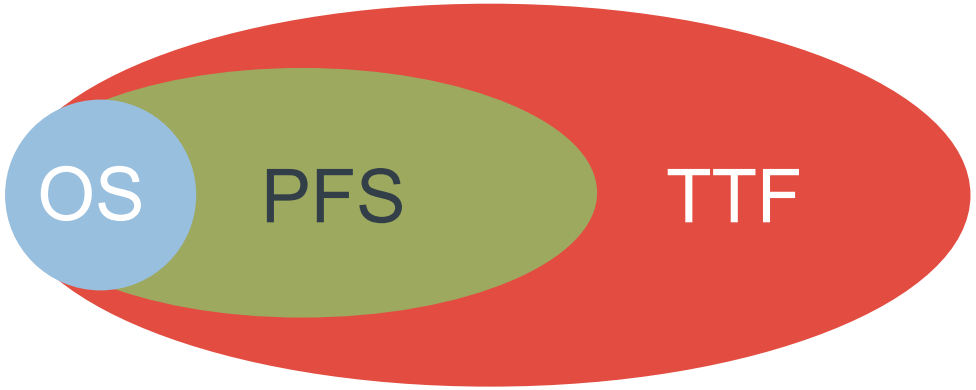
1. Park K *et al*. *Lancet Oncol* 2016; 17 (5): 577–589.

LUX-Lung 7: Three co-primary endpoints[1]

Two solutions:

B **Hierarchical approach**

- Order endpoints and stop when *P*>0.05
- No need to correct for multiplicity



OS, overall survival; PFS, progression-free survival; TTF, time to treatment failure.

1. Park K *et al*. *Lancet Oncol* 2016; 17 (5): 577–589.

LUX-Lung 7: Three co-primary endpoints[1]

Third ‘solution’ (the silliest one)

C All three endpoints must achieve significance

OS, overall survival; PFS, progression-free survival; TTF, time to treatment failure.

1. Park K *et al*. *Lancet Oncol* 2016; 17 (5): 577–589.

LUX-Lung 7: Final results

- Positive or negative study?

PFS: HR = 0.73 (95% CI: 0.57–0.95); *P*=0.017

TTF: HR = 0.73 (95% CI: 0.58–0.92); *P*=0.0073

OS: HR = 0.87 (95% CI: 0.66–1.15); ***P*=0.33**

Formally a negative study!

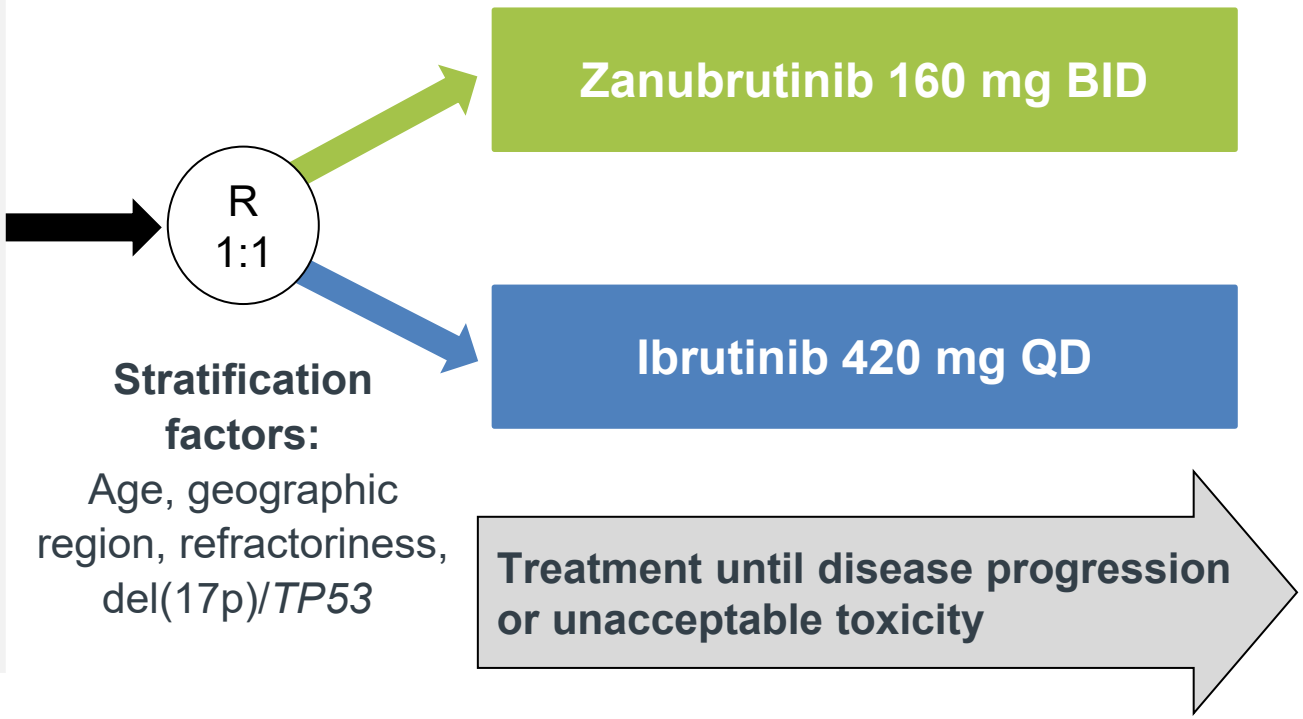
CI, confidence interval; HR, hazard ratio; OS, overall survival; PFS, progression-free survival; TTF, time to treatment failure.
Park K *et al*. *Lancet Oncol* 2016; 17 (5): 577–589.

ALPINE: Study design

R/R CLL/SLL with ≥1 prior treatment
(Planned N=600; actual N=652)

Key inclusion criteria
- R/R to ≥1 prior systemic therapy for CLL/SLL
- Measurable lymphadenopathy by CT or MRI
- Requires treatment per iwCLL

Key exclusion criteria
- Prior BTKi therapy
- Treatment with warfarin or other vitamin K antagonists



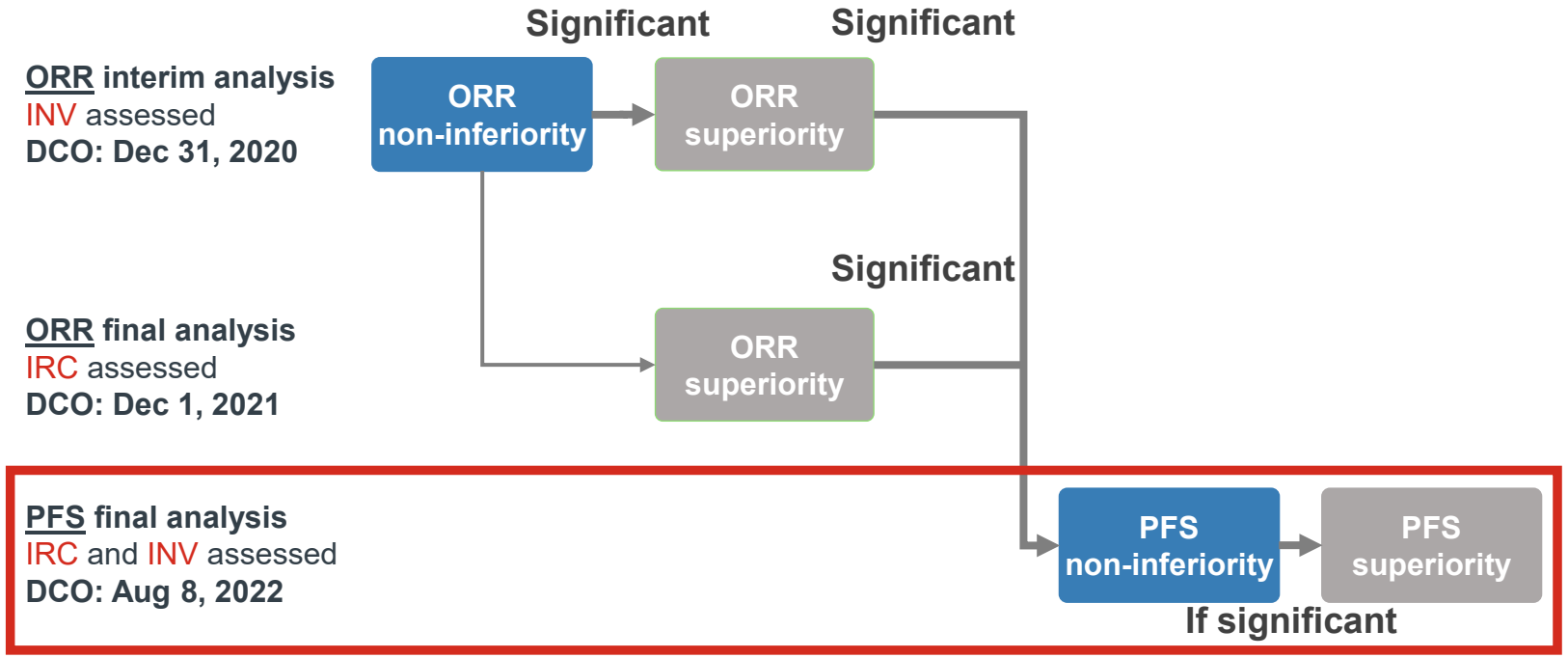
BID, twice a day; BTKi, BTK inhibitor; CLL, chronic lymphocytic leukemia; CT, computed tomography; del, deletion; iwCLL, International Workshop on Chronic Lymphocytic Leukemia; MRI, magnetic resonance imaging; QD, once daily; R, randomization; R/R, relapsed/refractory; SLL, small lymphocytic lymphoma.
Brown JR *et al*. Oral presentation at ASH 2022; New Orleans, LA, USA, December 10–13, 2022 (Abstract LBA-6).

ALPINE: Endpoints and statistical design

Primary endpoint
- ORR (PR + CR) non-inferiority and superiority (by investigator)

Key secondary endpoints
- PFS
- Incidence of atrial fibrillation

Other secondary endpoints
- DoR, OS
- TTF
- PR-L or higher
- Patient-reported outcomes
- Safety



- **ORR non-inferiority and superiority were demonstrated in the ORR interim and final analyses**
- **PFS was tested for non-inferiority under hierarchical testing when 205 events had occurred**

CR, complete response; DCO, data cut-off; DoR, duration of response; INV, investigator; IRC, independent review committee; ORR, overall response rate; OS, overall survival; PFS, progression-free survival; PR, partial response; PR-L, partial response with lymphocytosis; TTF, time to treatment failure.

Brown JR *et al*. Oral presentation at ASH 2022; New Orleans, LA, USA, December 10–13, 2022 (Abstract LBA-6).

Classical interim analysis

Comparative analyses of the primary endpoint to evaluate if there is a ‘significant’ difference requiring/allowing **the study to be stopped**

Slide courtesy of Paolo Bruzzi.

Interim analyses

Appropriate methodology:

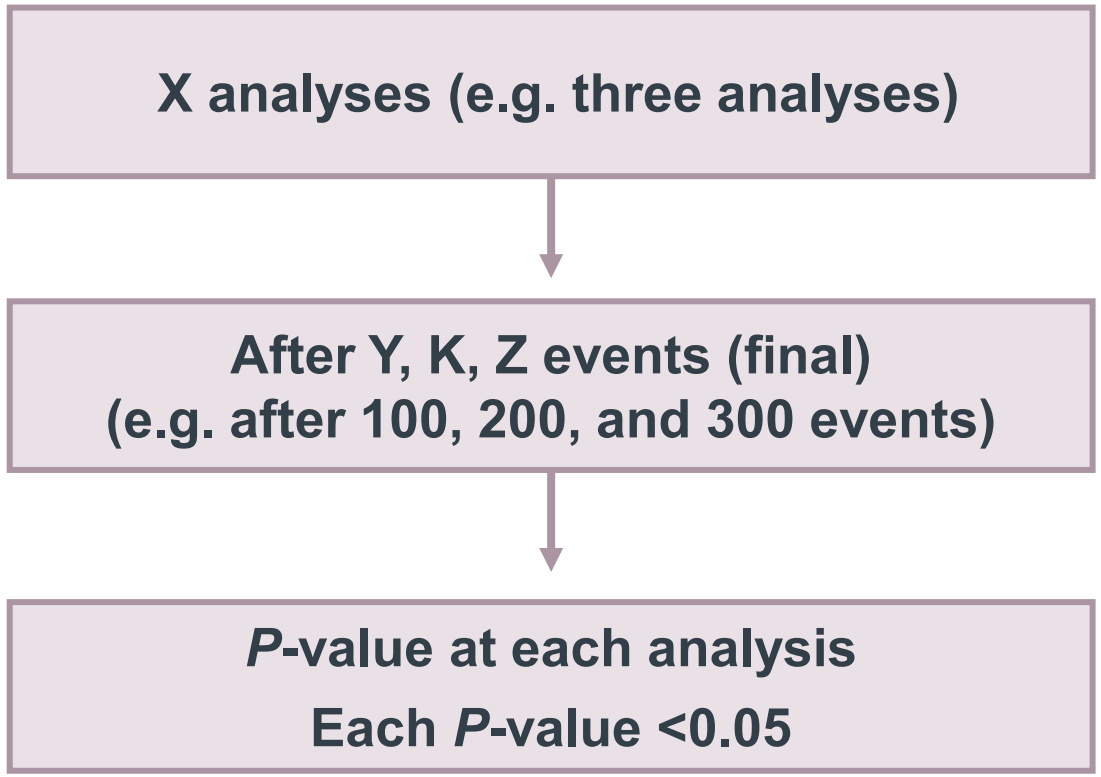
It is decided in advance when the analyses will be conducted

Adjusted *P*-values
The significance level required to stop the study is set at $P<0.05$

Strict levels at interim analyses to safeguard the power of the final analysis

Slide courtesy of Paolo Bruzzi.

Most widely used method: Group sequential analyses



Slide courtesy of Paolo Bruzzi.

Spending functions

- In practice, a spending function of the α error is established

Example: Five analyses

‘Constant’ spending (rarely used)

	Analysis 1	Analysis 2	Analysis 3	Analysis 4	Analysis 5
The study is stopped if the *P*-value is less than:	0.0158	0.0158	0.0158	0.0158	0.0158

‘Variable’ spending

	Analysis 1	Analysis 2	Analysis 3	Analysis 4	Analysis 5
The study is stopped if the *P*-value is less than:	0.0051 *or* 0.0005	0.0061 *or* 0.0005	0.0073 *or* 0.0005	0.0089 *or* 0.0005	0.0402 *or* 0.05

α, alpha.
Slide courtesy of Paolo Bruzzi.

Spending functions: Implications

- The ‘spending functions’ are computed to preserve the overall α error
- They involve a (usually moderate) loss of power
- With variable spending functions, the study is only stopped early if very strong effects are observed

Note:

- The ‘spending functions’ only take care of the α error
- Estimates of treatment effect are overestimated because of:
  - Regression to the mean
  - Only early events being considered

α, alpha.
Slide courtesy of Paolo Bruzzi.

Futility-based interim analyses

Trials may be halted after an interim analysis, based on futility

Why?

- Experimental therapy / limited prior data
- Anticipated therapeutic toxicity
- Anticipated low efficacy
  (based on the results of prior studies)
- Anticipated changes to the therapeutic landscape
  (e.g. emerging therapies with greater potential)
- Slow accrual (e.g. rare diseases) / low rate of events
- Anticipated costs

How?

Futility-based interim analyses are based on the probability…

- … of finding a significant difference at the end of the study (conditional power)
- … that the experimental therapy has the desired efficacy (Bayesian monitoring)

Interim analyses provide the opportunity to prematurely halt a trial without any 'statistical penalties'

This can be a very useful tool to stop a study when enrollment is languishing, or the therapy is no longer of interest

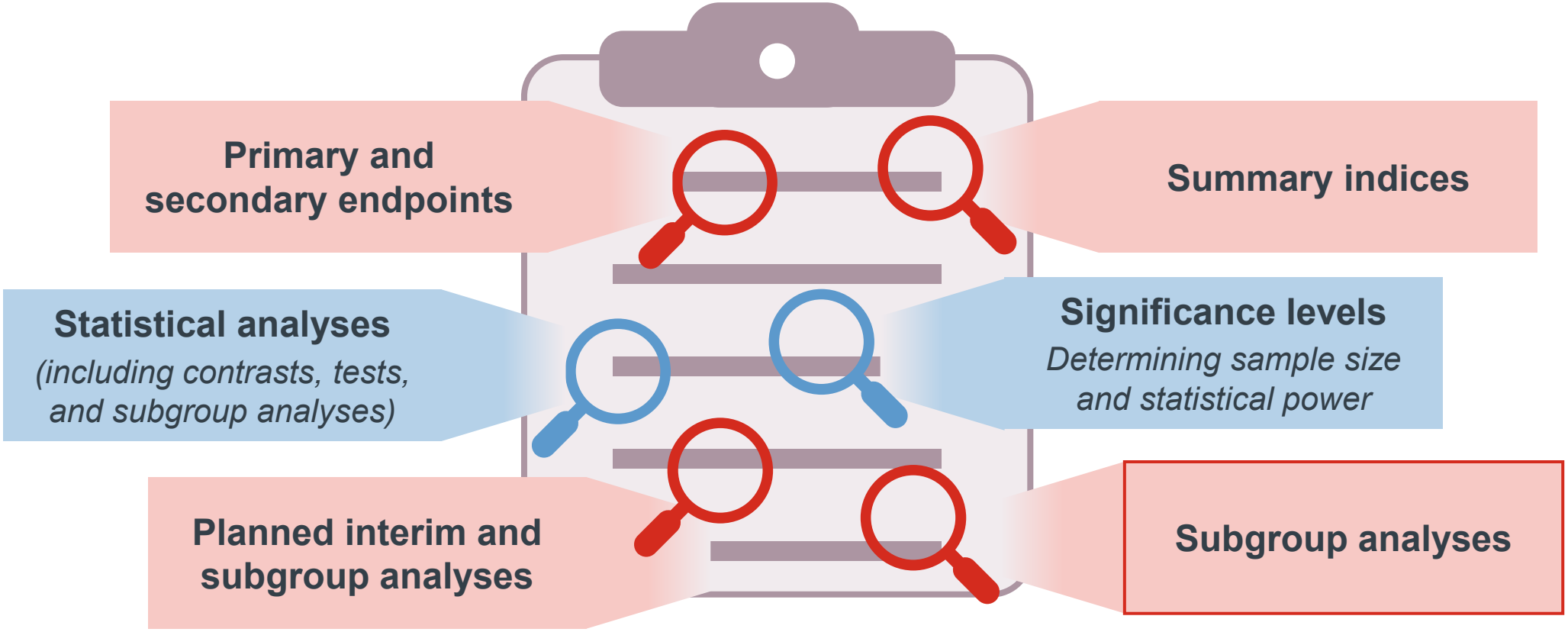
Slide courtesy of Paolo Bruzzi.

Interim analyses: Conclusions

- Interim analyses:
  - Are a useful tool for long-term studies
  - Must be carefully planned
  - Require specialized statistical support
- Unplanned interim analyses may compromise the validity of the study
- Slow enrolment/low rate of events: futility

Slide courtesy of Paolo Bruzzi.

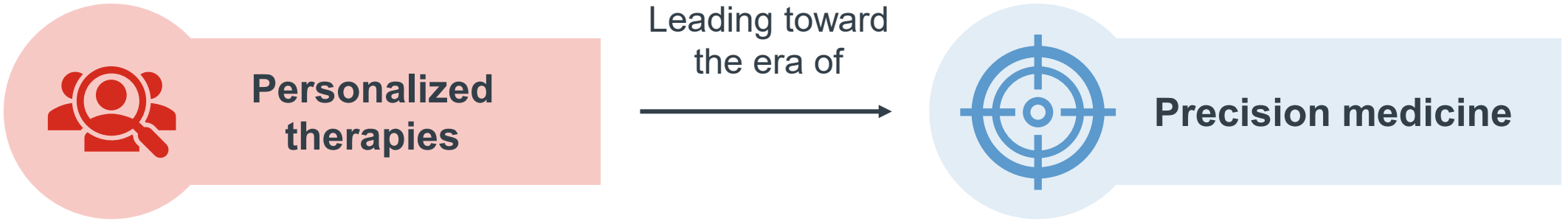
Statistical plan



Slide courtesy of Paolo Bruzzi.

Subgroup analyses

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



Slide courtesy of Paolo Bruzzi.

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

HR, hazard ratio; NHL, non-Hodgkin lymphoma; PD-L1, programmed death-ligand 1.
Slide courtesy of Paolo Bruzzi.

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 include a 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

HR, hazard ratio.
Slide courtesy of Paolo Bruzzi.

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 include a 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

These should be defined in the study protocol

HR, hazard ratio.
Slide courtesy of Paolo Bruzzi.

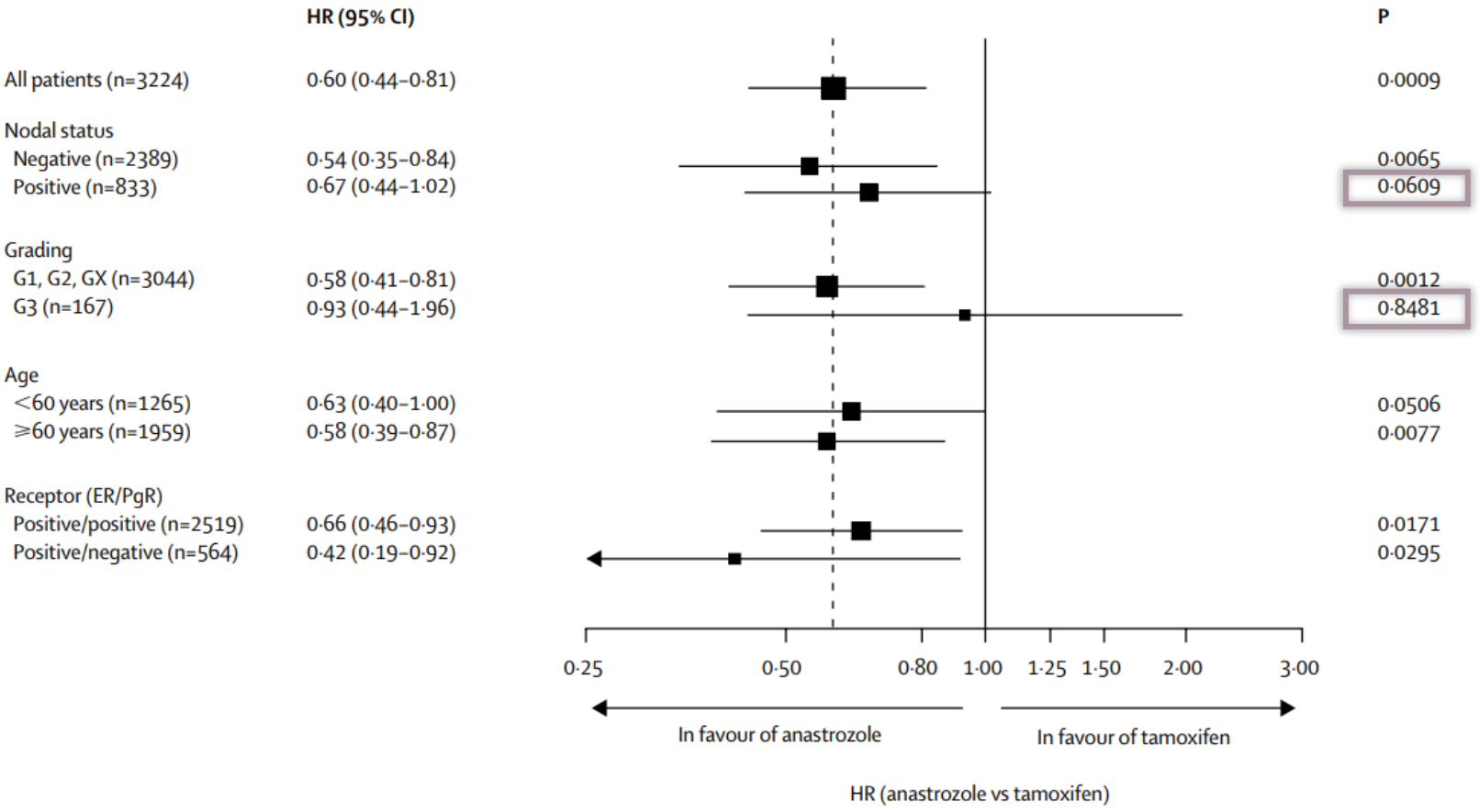
Issues with subgroup analyses: Statistical

The common statistical issues with subgroup analyses are:

- α_P **Improper significance testing**
-  **Multiplicity**

Slide courtesy of Paolo Bruzzi.

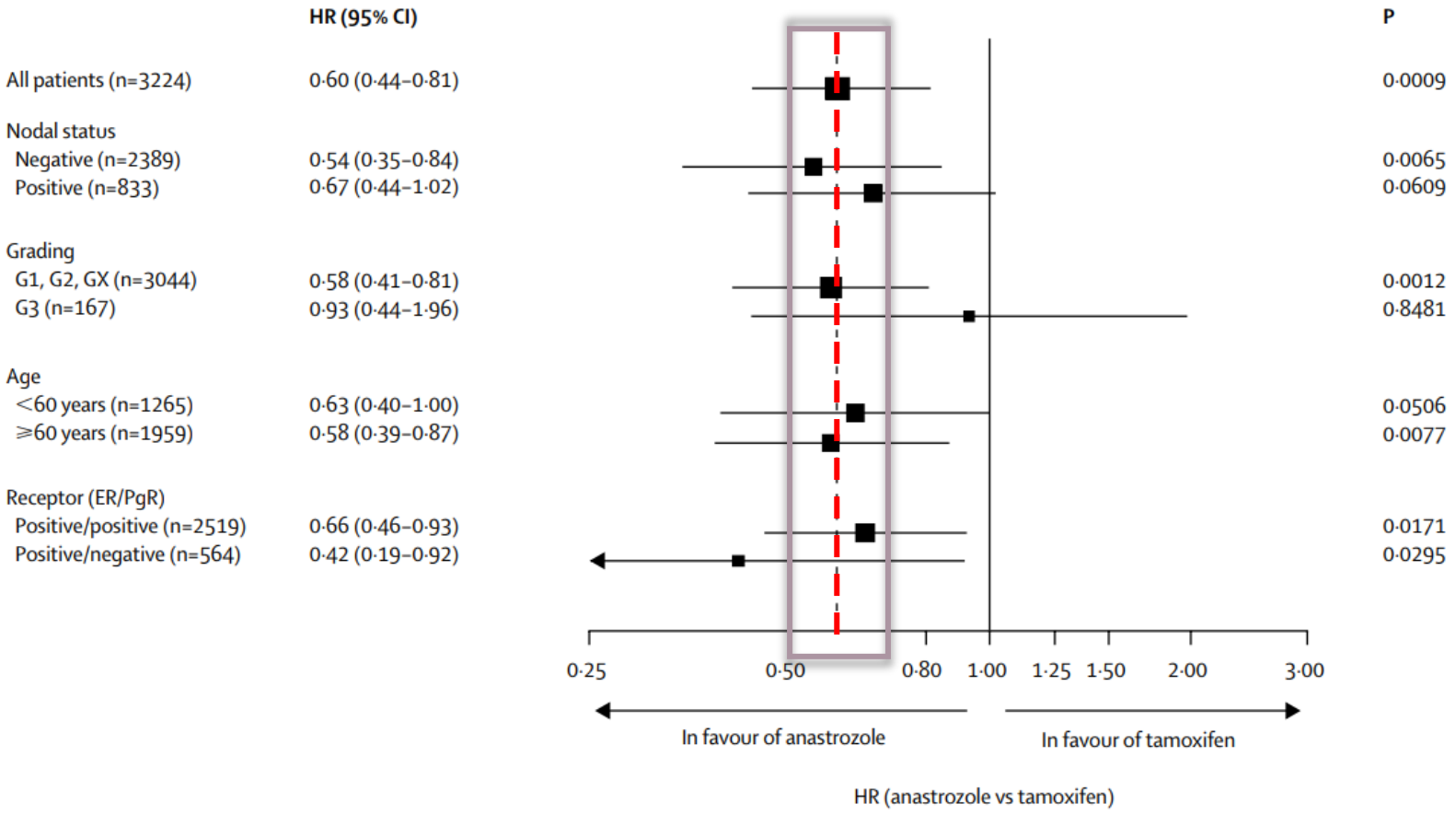
Improper significance testing



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

Figure included for illustrative purposes only.
CI, confidence interval; ER, estrogen receptor; G, Grade; HR, hazard ratio; PgR, progesterone receptor.
Jakesz R *et al*. *Lancet* 2005; 366 (9484): 455–462.

Proper subgroup analysis



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

Figure included for illustrative purposes only.
CI, confidence interval; ER, estrogen receptor; G, Grade; HR, hazard ratio; PgR, progesterone receptor.
Jakesz R *et al*. *Lancet* 2005; 366 (9484): 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

Slide courtesy of Paolo Bruzzi.

Improper 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)	

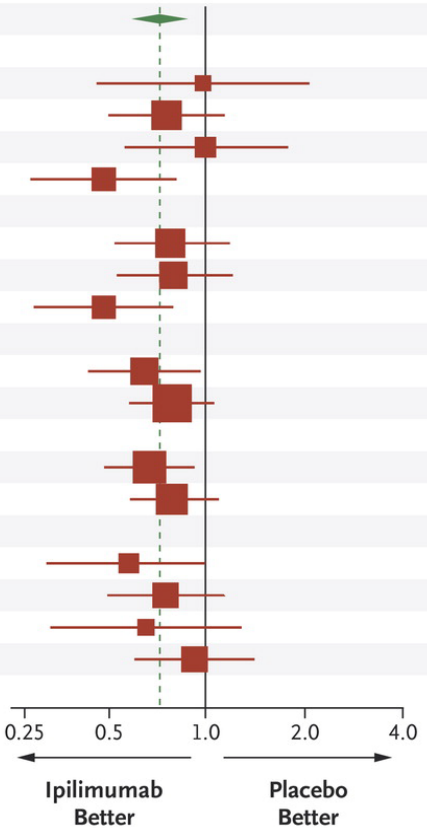


Figure included for illustrative purposes only.
CI, confidence interval.
Eggermont AMM *et al*. *N Engl J Med* 2016; 375 (19): 1845–1855.

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

H0, null hypothesis.
Slide courtesy of Paolo Bruzzi.

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

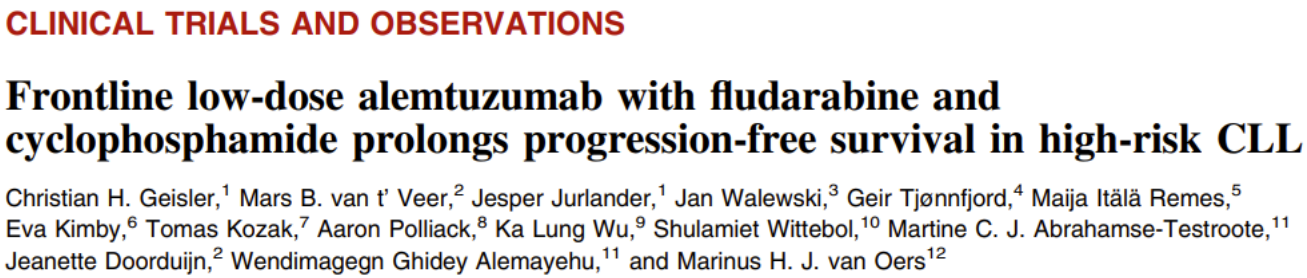
α split required

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

Statistical analysis plan!

α, alpha.
Slide courtesy of Paolo Bruzzi.

Predefined vs. post hoc endpoint analysis: Example



CLINICAL TRIALS AND OBSERVATIONS

Frontline low-dose alemtuzumab with fludarabine and cyclophosphamide prolongs progression-free survival in high-risk CLL

Christian H. Geisler,[1] Mars B. van t' Veer,[2] Jesper Jurlander,[1] Jan Walewski,[3] Geir Tjønnfjord,[4] Maija Itälä Remes,[5] Eva Kimby,[6] Tomas Kozak,[7] Aaron Polliack,[8] Ka Lung Wu,[9] Shulamiet Wittebol,[10] Martine C. J. Abrahamse-Testroote,[11] Jeanette Doorduijn,[2] Wendimagegn Ghidey Alemayehu,[11] and Marinus H. J. van Oers[12]

The following claim was made after the study:

"FCA prolonged the primary end point, progression-free survival (3-year progression-free survival 53 vs 37%, P = .01), but not the secondary end point, overall survival (OS).

However, a post hoc analysis showed that FCA increased OS in patients younger than 65 years (3-year OS 85% vs 76%, P = .035)."

But was this claim valid?

CLL, chronic lymphocytic leukemia; FCA, alemtuzumab with fludarabine and cyclophosphamide.
Geisler CH *et al*. *Blood* 2014; 123 (21): 3255–3262.

Predefined vs. post hoc endpoint analysis: Example

In addition to the primary endpoint (PFS) and secondary endpoint (OS), the following was defined in the statistical analysis plan:

- *"The treatment effects in subgroups were explored **in post hoc analyses by comparing the subgroup PFS and OS probabilities at 3 years**, … performing tests, including Cox regression analyses, **for interactions with treatment arm**."*
- *"The subgroups included genomic aberrations, IGHV mutational status, β_2-microglobulin, clinical stage, sex, and age. In view of recent important reports of substantial differences in response to immunotherapy and PFS between younger and older patients, … **the unplanned post hoc analysis included the outcomes of patients <65 and ≥65 years of age**, respectively."*

→ Planned post hoc analyses were designed to compare 3-year subgroup PFS and OS data with the overall treatment effect observed

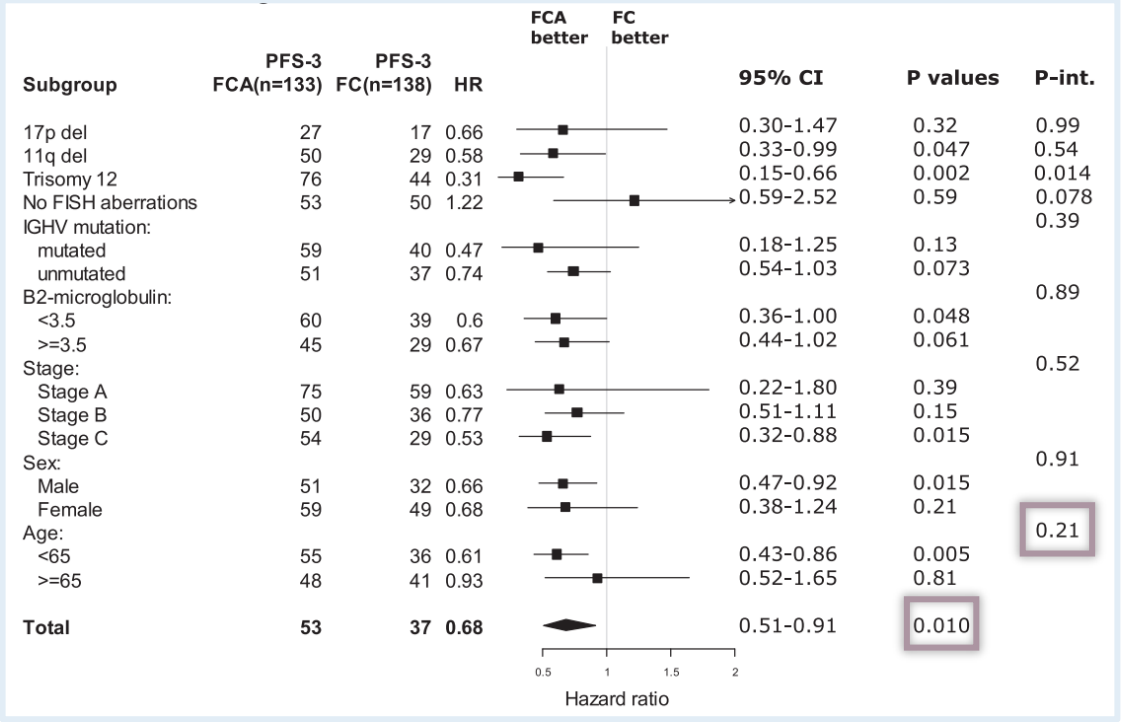
→ Age-based subgroup analyses were unplanned

OS, overall survival; PFS, progression-free survival.
Geisler CH *et al*. *Blood* 2014; 123 (21): 3255–3262.

Predefined vs. post hoc endpoint analysis: Example

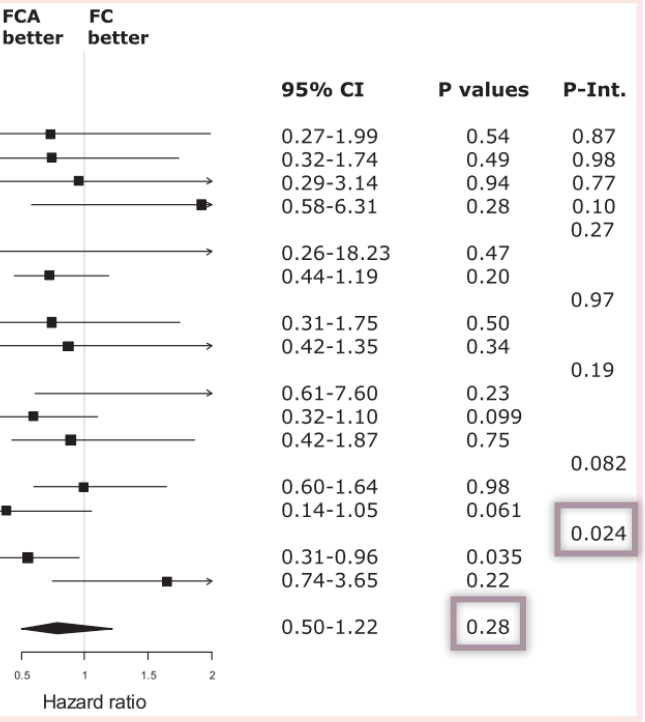
Is there sufficient evidence to claim an age-related treatment effect?

PFS



Subgroup	PFS-3 FCA(n=133)	PFS-3 FC(n=138)	HR	95% CI	P values	P-int.
17p del	27	17	0.66	0.30-1.47	0.32	0.99
11q del	50	29	0.58	0.33-0.99	0.047	0.54
Trisomy 12	76	44	0.31	0.15-0.66	0.002	0.014
No FISH aberrations	53	50	1.22	0.59-2.52	0.59	0.078
IGHV mutation:						0.39
mutated	59	40	0.47	0.18-1.25	0.13	
unmutated	51	37	0.74	0.54-1.03	0.073	
B2-microglobulin:						0.89
<3.5	60	39	0.6	0.36-1.00	0.048	
>=3.5	45	29	0.67	0.44-1.02	0.061	
Stage:						0.52
Stage A	75	59	0.63	0.22-1.80	0.39	
Stage B	50	36	0.77	0.51-1.11	0.15	
Stage C	54	29	0.53	0.32-0.88	0.015	
Sex:						0.91
Male	51	32	0.66	0.47-0.92	0.015	
Female	59	49	0.68	0.38-1.24	0.21	
Age:						0.21
<65	55	36	0.61	0.43-0.86	0.005	
>=65	48	41	0.93	0.52-1.65	0.81	
Total	**53**	**37**	**0.68**	0.51-0.91	0.010	

OS



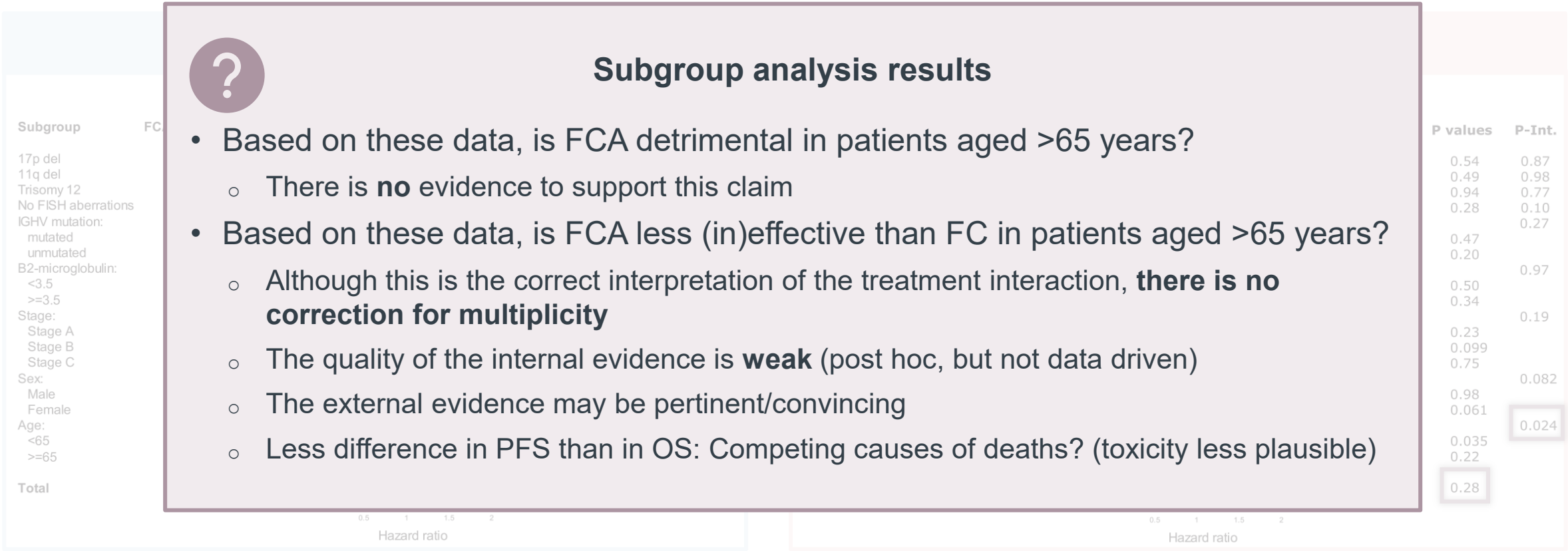
Subgroup	OS-3 FCA(n=133)	OS-3 FC(n=139)	HR	95% CI	P values	P-Int.
17p del	65	44	0.73	0.27-1.99	0.54	0.87
11q del	83	76	0.74	0.32-1.74	0.49	0.98
Trisomy 12	91	86	0.95	0.29-3.14	0.94	0.77
No FISH aberrations	89	87	1.92	0.58-6.31	0.28	0.10
IGHV mutation:						0.27
mutated	82	86	2.19	0.26-18.23	0.47	
unmutated	86	77	0.72	0.44-1.19	0.20	
B2-microglobulin:						0.97
<3.5	89	77	0.74	0.31-1.75	0.50	
>=3.5	80	73	0.87	0.42-1.35	0.34	
Stage:						0.19
Stage A	79	82	2.16	0.61-7.60	0.23	
Stage B	87	75	0.59	0.32-1.10	0.099	
Stage C	84	74	0.89	0.42-1.87	0.75	
Sex:						0.082
Male	83	77	0.99	0.60-1.64	0.98	
Female	93	72	0.38	0.14-1.05	0.061	
Age:						0.024
<65	88	76	0.55	0.31-0.96	0.035	
>=65	78	76	1.65	0.74-3.65	0.22	
Total	**85**	**76**	**0.78**	0.50-1.22	0.28	

CI, confidence interval; del, deletion; FC(A), (alemtuzumab with) fludarabine and cyclophosphamide; FISH, fluorescence *in situ* hybridization; HR, hazard ratio; int., interaction; OS(-3), (3-year) overall survival; PFS(-3), (3-year) progression-free survival.

Geisler CH *et al*. *Blood* 2014; 123 (21): 3255–3262.

Predefined vs. post hoc endpoint analysis: Example

Is there sufficient evidence to claim an age-related treatment effect?



Subgroup analysis results

- Based on these data, is FCA detrimental in patients aged >65 years?
  - There is **no** evidence to support this claim
- Based on these data, is FCA less (in)effective than FC in patients aged >65 years?
  - Although this is the correct interpretation of the treatment interaction, **there is no correction for multiplicity**
  - The quality of the internal evidence is **weak** (post hoc, but not data driven)
  - The external evidence may be pertinent/convincing
  - Less difference in PFS than in OS: Competing causes of deaths? (toxicity less plausible)

FC(A), (alemtuzumab with) fludarabine and cyclophosphamide; OS, overall survival; PFS, progression-free survival.
Geisler CH *et al*. *Blood* 2014; 123 (21): 3255–3262.

Conclusions

α_P

Modern trials are sufficiently designed to avoid multiplicity arising from statistical analysis

However, available statistical techniques do not correct for overestimates of treatment efficacy arising from multiple analyses

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

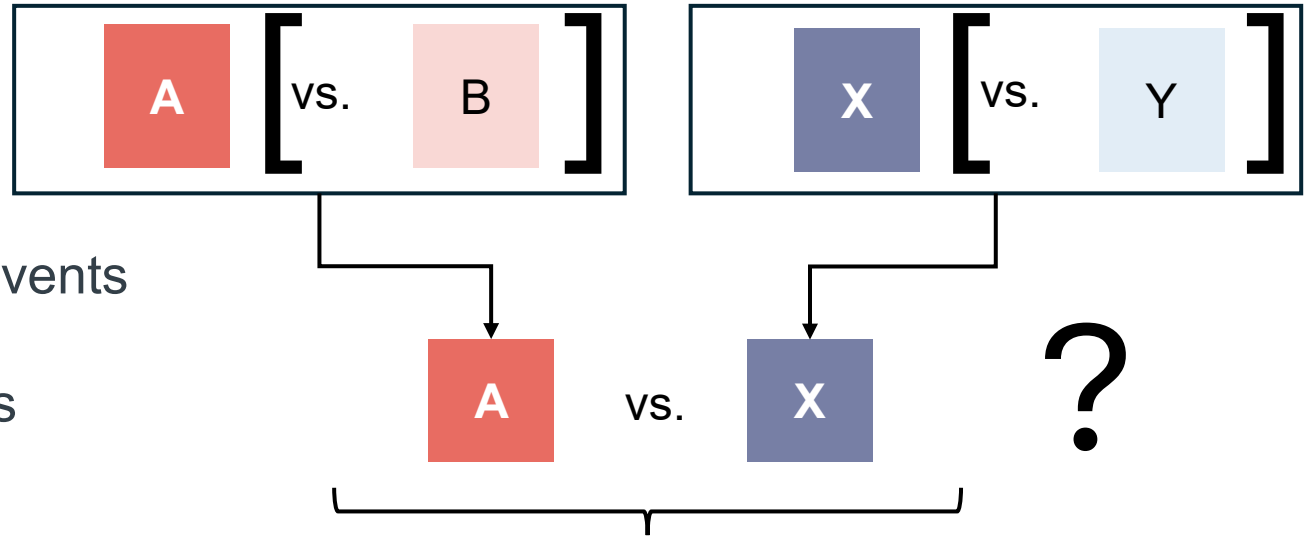
Slide courtesy of Paolo Bruzzi.

3. Indirect comparisons

Indirect comparisons

Studies in which…

a) Outcomes/incidences of clinical events
and
b) Effects of treatments/interventions

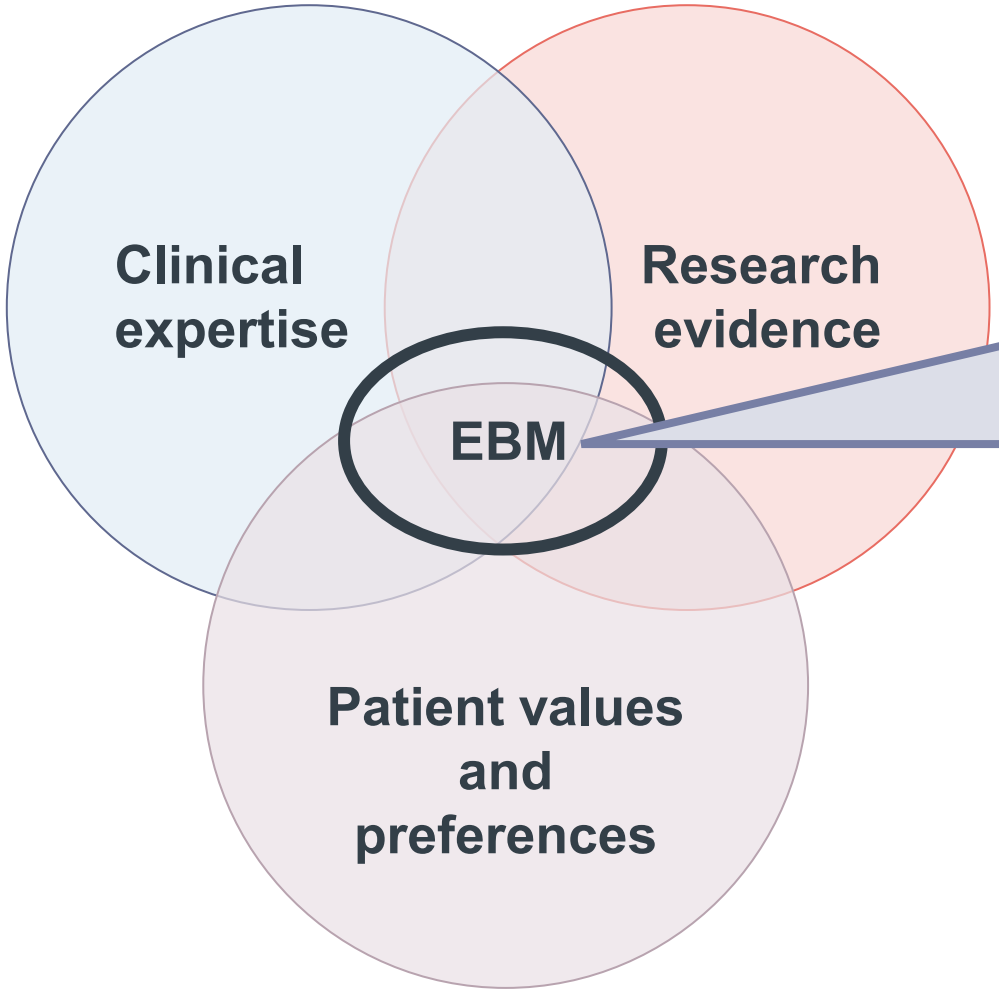


…are compared in groups of patients included in different studies

The aim is usually the comparative evaluation of the effectiveness and/or toxicity of different treatments or intervention strategies

Slide courtesy of Paolo Bruzzi.

Evidence-based medicine



Evidence-based medicine (EBM)

The integration of best research evidence with clinical expertise and patient values

Randomized controlled trials (RCTs) or systematic reviews of RCTs provide **evidence of 'efficacy'**

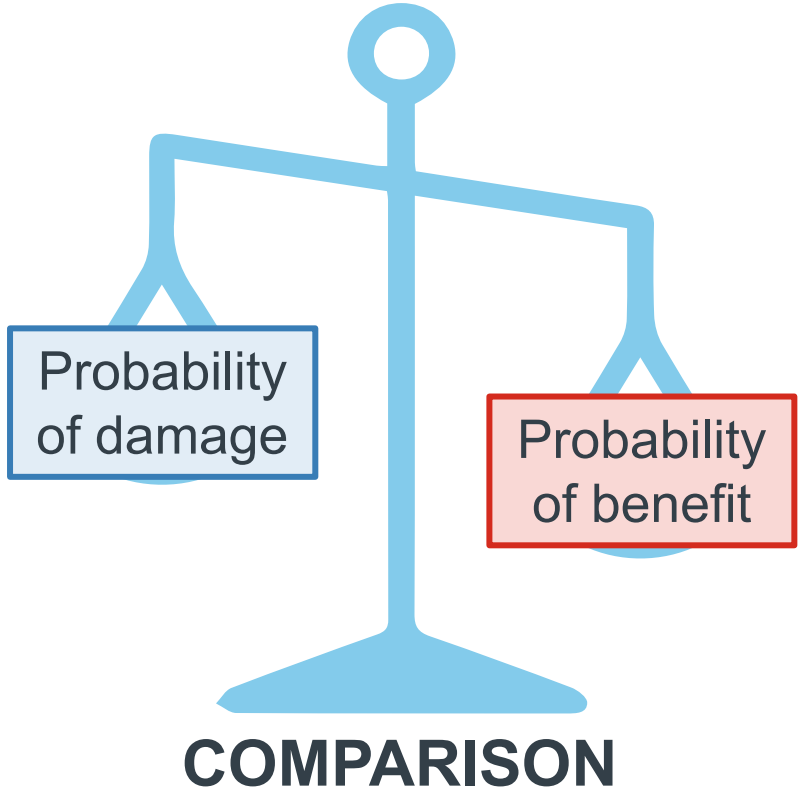
Slide courtesy of Paolo Bruzzi.

Evidence (proof of efficacy)

Results of one or more studies

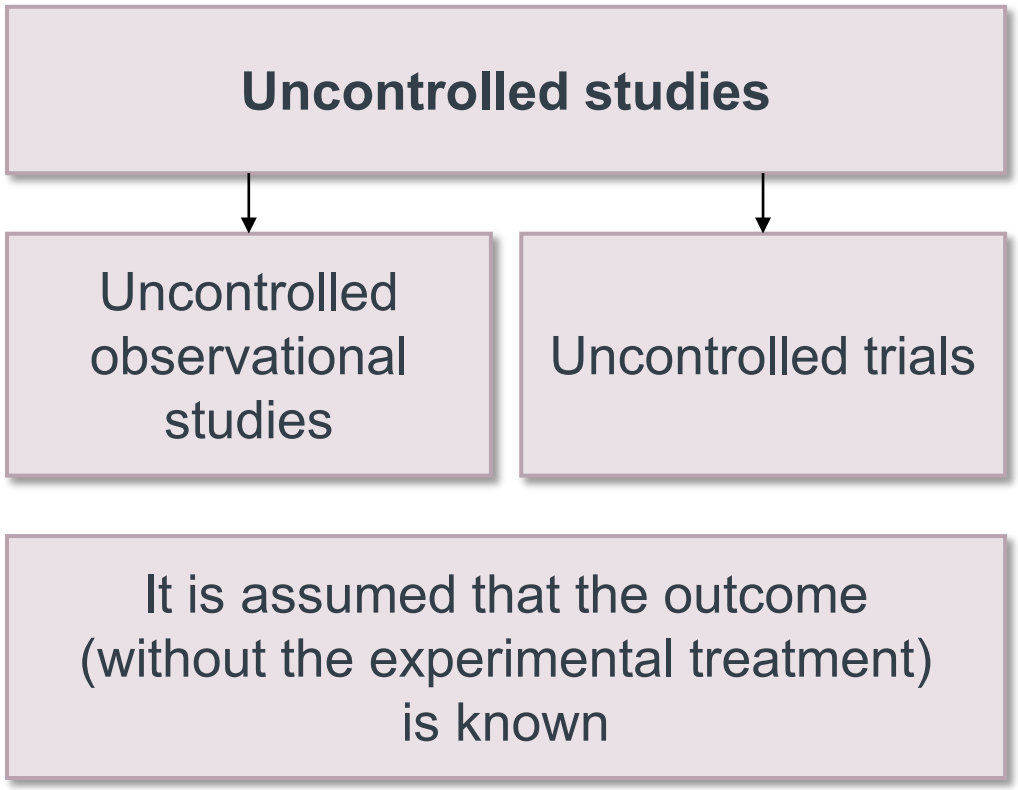
Evaluation

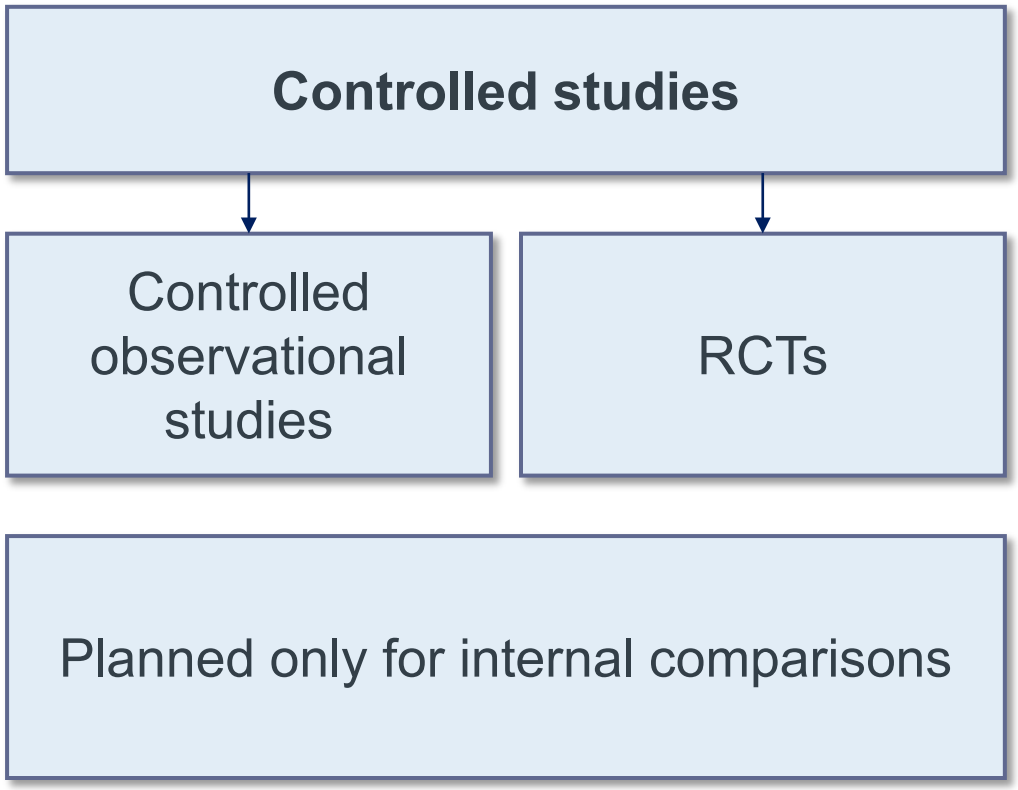
To what extent does an **intervention** (e.g., treatment, diagnosis action, etc.) represent a **valid option** for a patient?



Slide courtesy of Paolo Bruzzi.

Types of studies





RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Statistical analyses for comparison

Study designs

- Uncontrolled observational studies
- Controlled observational studies
- Uncontrolled trials
- RCTs

Statistical tools

- Univariate analysis
- Multivariate analysis
- Meta-analysis
- Meta-regression
- Network meta-analysis
- MAIC
- More complex analyses

MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Statistical analyses for comparison

Study designs

- Uncontrolled observational studies
- Controlled observational studies
- Uncontrolled trials
- **RCTs**

Statistical tools

- **Univariate analysis**
- **Multivariate analysis**
- **Meta-analysis**
- Meta-regression
- Network meta-analysis
- MAIC
- More complex analyses

MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

What if there is no RCT addressing the question of interest?

Study designs

- Uncontrolled observational studies
- Controlled observational studies
- Uncontrolled trials
- ~~**RCTs**~~

Statistical tools

- **Univariate analysis**
- **Multivariate analysis**
- **Meta-analysis**
- Meta-regression
- Network meta-analysis
- MAIC
- More complex analyses

?

Indirect comparisons!

MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Indirect comparisons

Studies in which…

a) Outcomes/incidences of clinical events

b) Effects of tre

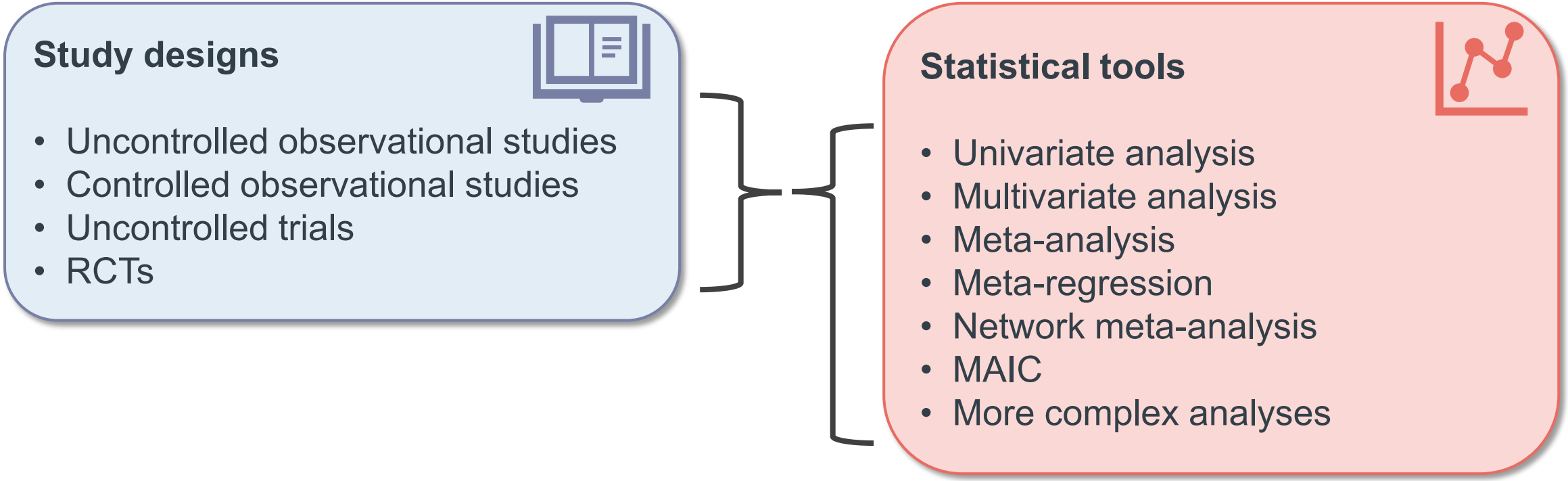
Can these be trusted?

…are compared in groups of patients included in different studies

The aim is usually the comparative evaluation of the effectiveness and/or toxicity of different treatments or intervention strategies

Slide courtesy of Paolo Bruzzi.

And if there is no RCT addressing the question of interest?



MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

And if there is no RCT addressing the question of interest?

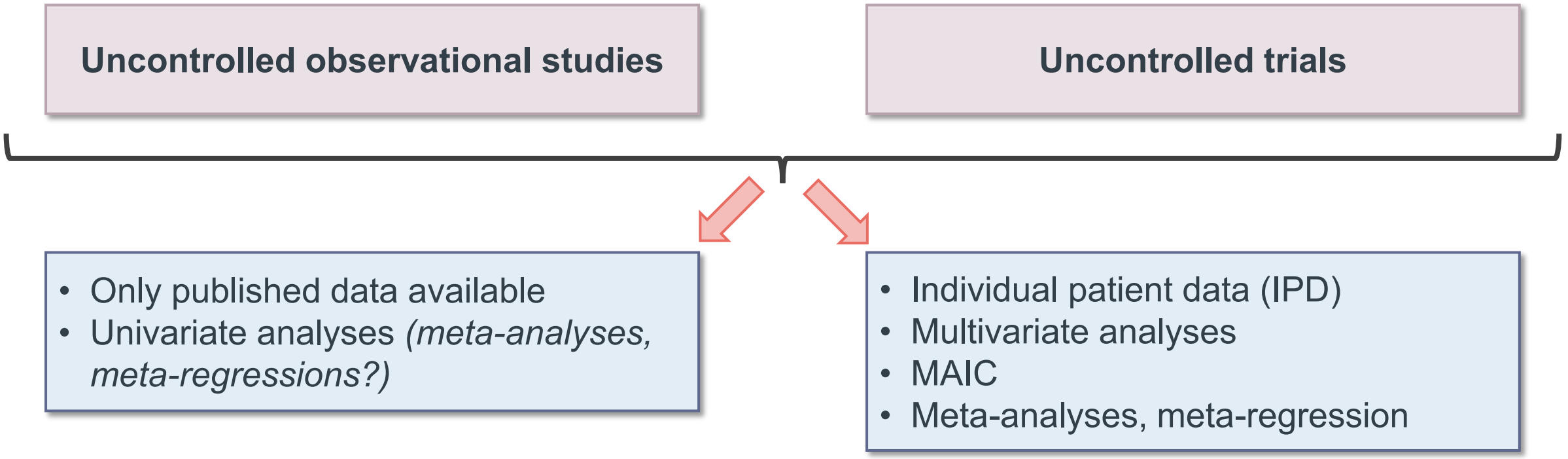
Uncontrolled observational studies

Uncontrolled trials

- In these analyses, it is only possible to compare outcomes/incidences of events
- Appear similar to RCTs, but patients were observed/treated in different studies

RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

And if there is no RCT addressing the question of interest?



MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection			
Attrition			
Assessment			
Analysis			

Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	**Randomization (intention-to-treat [ITT])**	
Attrition			
Assessment			
Analysis			

Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	**Randomization (ITT)**	
Attrition	Lost to follow-up Not evaluated		
Assessment			
Analysis			

ITT, intention-to-treat.
Slide courtesy of Paolo Bruzzi.

Attrition bias

"All cancer patients who continue to receive my treatment are still alive, after several years" – Luigi Di Bella
Immortal time bias

Are patients who drop out, lost to follow-up, or not evaluable, a random sample? Are they comparable in the two groups? **(ATTRITION BIAS)**

Treatment A: 200 treated → 40 cured = 20%

Treatment B: 200 treated → 100 evaluated → 40 cured = ?

Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	Randomization (ITT)	
Attrition	Lost to follow-up Not evaluated	**ITT** **Blinding**	
Assessment			
Analysis			

ITT, intention-to-treat.
Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	Randomization (ITT)	
Attrition	Lost to follow-up Not evaluated	ITT Blinding	
Assessment	Methods Bias	**Blinding** **Hard endpoints**	
Analysis			

ITT, intention-to-treat.
Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	Randomization (ITT)	
Attrition	Lost to follow-up Not evaluated	ITT Blinding	
Assessment	Methods Bias	Blinding Hard endpoints	
Analysis	Multiplicity	**Predefined statistical plan**	

ITT, intention-to-treat.
Slide courtesy of Paolo Bruzzi.

Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	Randomization (ITT)	**Randomized controlled double-blind trial**
Attrition	Lost to follow-up Not evaluated	ITT Blinding	
Assessment	Methods Bias	Blinding Hard endpoints	
Analysis	Multiplicity	Predefined statistical plan	

ITT, intention-to-treat.
Slide courtesy of Paolo Bruzzi.

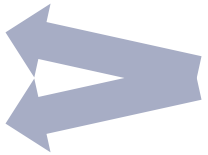
Treatment comparisons: Possible biases

Bias	Difference in	Solution	
Selection	Prognostic factors Predictive factors Will Rogers	~~**Randomization**~~ **(ITT)**	Statistical tools are not able to remove **ALL** these biases: **Accurate methodological revision**
Attrition	Lost to follow-up Not evaluated	ITT Blinding	
Assessment	Methods Bias	Blinding Hard endpoints	
Analysis	Multiplicity	Predefined statistical plan	

ITT, intention-to-treat.
Slide courtesy of Paolo Bruzzi.

Statistical methods for indirect comparisons

- ‘Raw comparison’
- Multivariate analysis
  - Direct
  - Through propensity score
- MAIC
- Simulated treatment comparison (STC)
- Network meta-analysis/meta-regression
  - Aggregate data
  - IPD



Uncontrolled studies

IPD, individual patient data; MAIC, matching-adjusted indirect comparison.
Slide courtesy of Paolo Bruzzi.

‘Raw comparison’ uncontrolled studies

In two groups of patients with COVID-19, one treated with Drug A and the other with Drug B, we observed 40/200 (20%) and 40/100 (40%) ‘recoveries’, respectively

What do we conclude?

Little or nothing!

Slide courtesy of Paolo Bruzzi.

Multivariate analysis: Uncontrolled studies

- Direct / via propensity score? It makes little difference

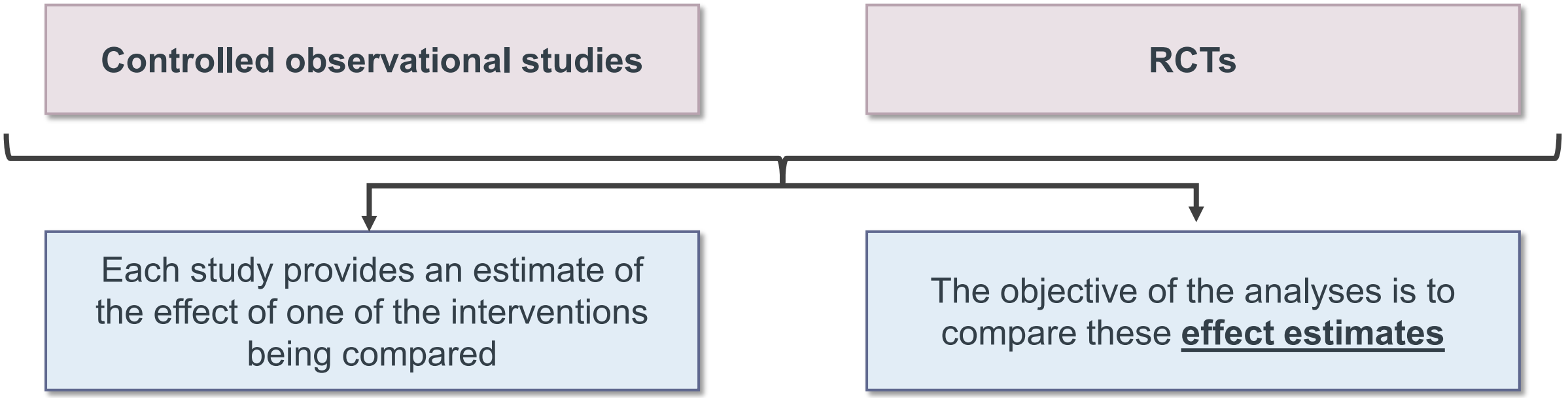
Requirement: Availability of IPD on all patients from all studies

Advantages
- Removal of differences in all **known and recorded** prognostic and predictive factors
- IPD from uncontrolled trials (observational studies) can be used

It represents an observational study

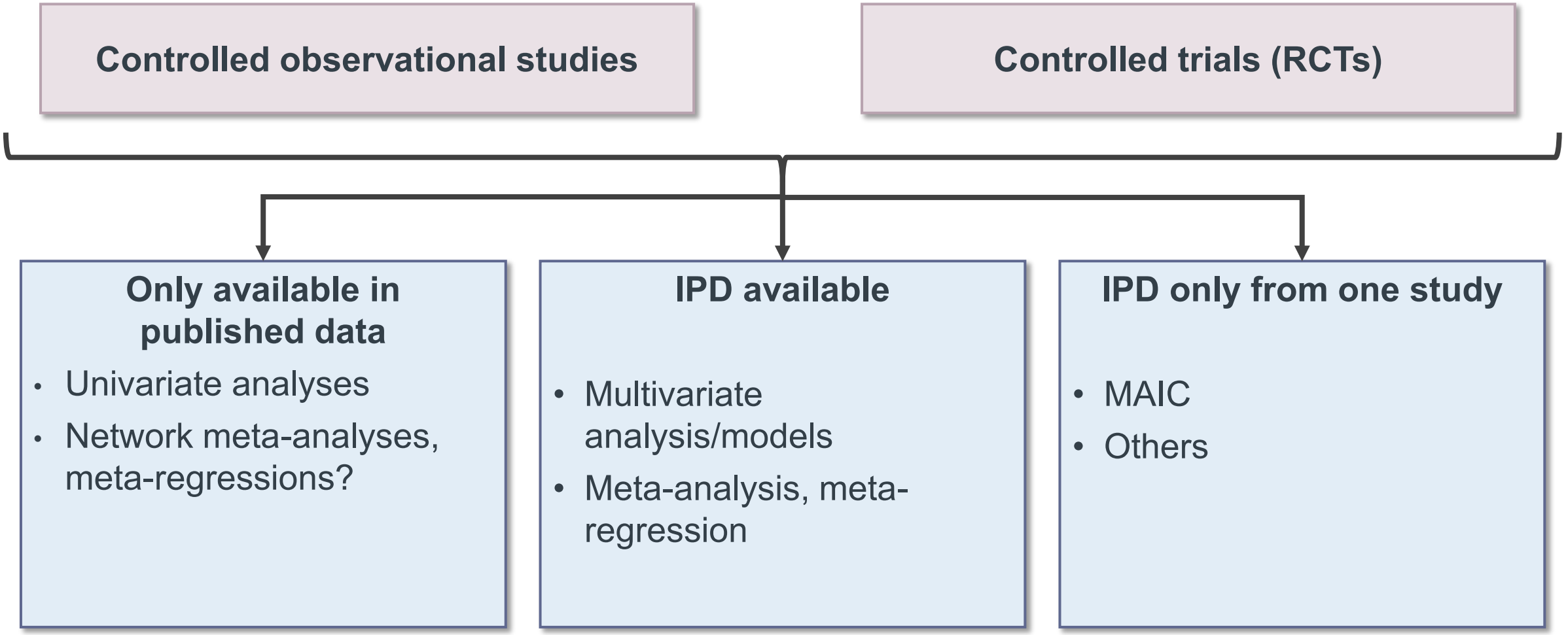
IPD, individual patient data.
Slide courtesy of Paolo Bruzzi.

And if there is no RCT addressing the question of interest?



RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

And if there is no RCT addressing the question of interest?



IPD, individual patient data; MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

1. ‘Raw comparison’: RCT

In two randomized trials on patients with COVID-19, one evaluating drug A vs. placebo and the other evaluating drug B vs. placebo, we observed:

RCT 1: Drug A deaths

40/200 (20%) vs. 60/200 (30%)
RR = 0.66

RCT 2: Drug B deaths

30/100 (30%) vs. 60/100 (60%)
RR = 0.5

What should I conclude?

Little or nothing

RCT, randomized controlled trial; RR, risk ratio.
Slide courtesy of Paolo Bruzzi.

2. Multivariate analysis: Comparison of the RCT results

Requirements:

- ☑ IPD of all patients from all RCTs being compared
- ☑ Trials being analyzed should have the same control arm
- ☑ A multivariate (usually proportional hazards) model is fitted with all relevant covariates and trial and treatment as strata

Direct adjustment or through a propensity score?

Little difference!

IPD, individual patient data; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

2. Multivariate analysis: Comparison of the RCTs results

Bias	Difference in	Problems
Selection	Prognostic factors Predictive factors Will Rogers	**Unknown/unmeasured factors?**
Attrition	Lost to follow-up Not evaluated	**Trial quality?**
Assessment	Methods Bias	**?**
Analysis	Multiplicity	**Can be addressed**

RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Unanchored MAIC

Requirement: Availability of IPD from at least one of several RCTs

If RCTs have different control therapies = unanchored MAIC analysis

Only experimental arms are considered	=	Comparison of **uncontrolled** observational studies

With a propensity score, groups of patients comparable with those of each of the ‘comparison’ trials are extracted from the experimental arm with IPD (approximate adjustment)

IPD, individual patient data; MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Unanchored MAIC: Possible biases

Bias	Difference in	Problems
Selection	Prognostic factors Predictive factors Will Rogers	**????**
Attrition	Lost to follow-up Not evaluated	**??**
Assessment	Methods Bias	**??**
Analysis	Multiplicity	**?**

MAIC, matching-adjusted indirect comparison.
Slide courtesy of Paolo Bruzzi.

Anchored MAIC

Requirement: Availability of IPD from at least one trial

RCT with same control treatment = anchored MAIC analysis

- Both arms are considered while maintaining randomization
- With a propensity score, treated and control groups comparable with those of each of the trials are extracted from the IPD trial (approximate adjustment)
- Same methodology as network meta-analysis ([NMA]; comparisons of HR, not patients), with the advantage of more precise adjustments
- Separate comparison with each trial

HR, hazard ratio; IPD, individual patient data, MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Anchored MAIC: Possible biases

Bias	Difference in	Problems
Selection	Prognostic factors Predictive factors Will Rogers	**OK** **Unknown/unmeasured predictive factors?** **OK**
Attrition	Lost to follow-up Not evaluated	**Trial quality?**
Assessment	Methods Bias	**?**
Analysis	Multiplicity	**Can be addressed**

MAIC, matching-adjusted indirect comparison.
Slide courtesy of Paolo Bruzzi.

Simulated treatment comparison

- Comparable to an **anchored** MAIC but using multiple regression techniques
- Similar requirements, advantages, and disadvantages

MAIC, matching-adjusted indirect comparison.
Slide courtesy of Paolo Bruzzi.

Network meta-analysis

Note:

The statistical methodologies of meta-analyses, NMAs, and meta-regression can be applied to any type of:

- **Study:** Observational/experimental NCT/RCT
- **Endpoint:** OS, PFS, ORR, QoL score, glycemia, etc.
- **Summary estimator:**
  - Absolute: % survival at 3 years, ORR, QoL after X years, average blood glucose, etc.
  - Relative: HR, delta, odds ratio, difference between averages, etc.

MA, NMA of RCTs: More immune from biases

- **Requirements:** ≥2 RCTs in the same disease, with links between trials for common control or experimental groups
- Summary indicators can be used (e.g. HR) or IPD from all studies
- An NMA analyzes all trials at once (network)

HR, hazard ratio; IPD, individual patient data; MA, meta-analysis; NCT, non-controlled trial; NMA, network meta-analysis; ORR, overall response rate; OS, overall survival; PFS, progression-free survival; QoL, quality of life; RCT, randomized controlled trial.
Slide courtesy of Paolo Bruzzi.

Direct and indirect evidence

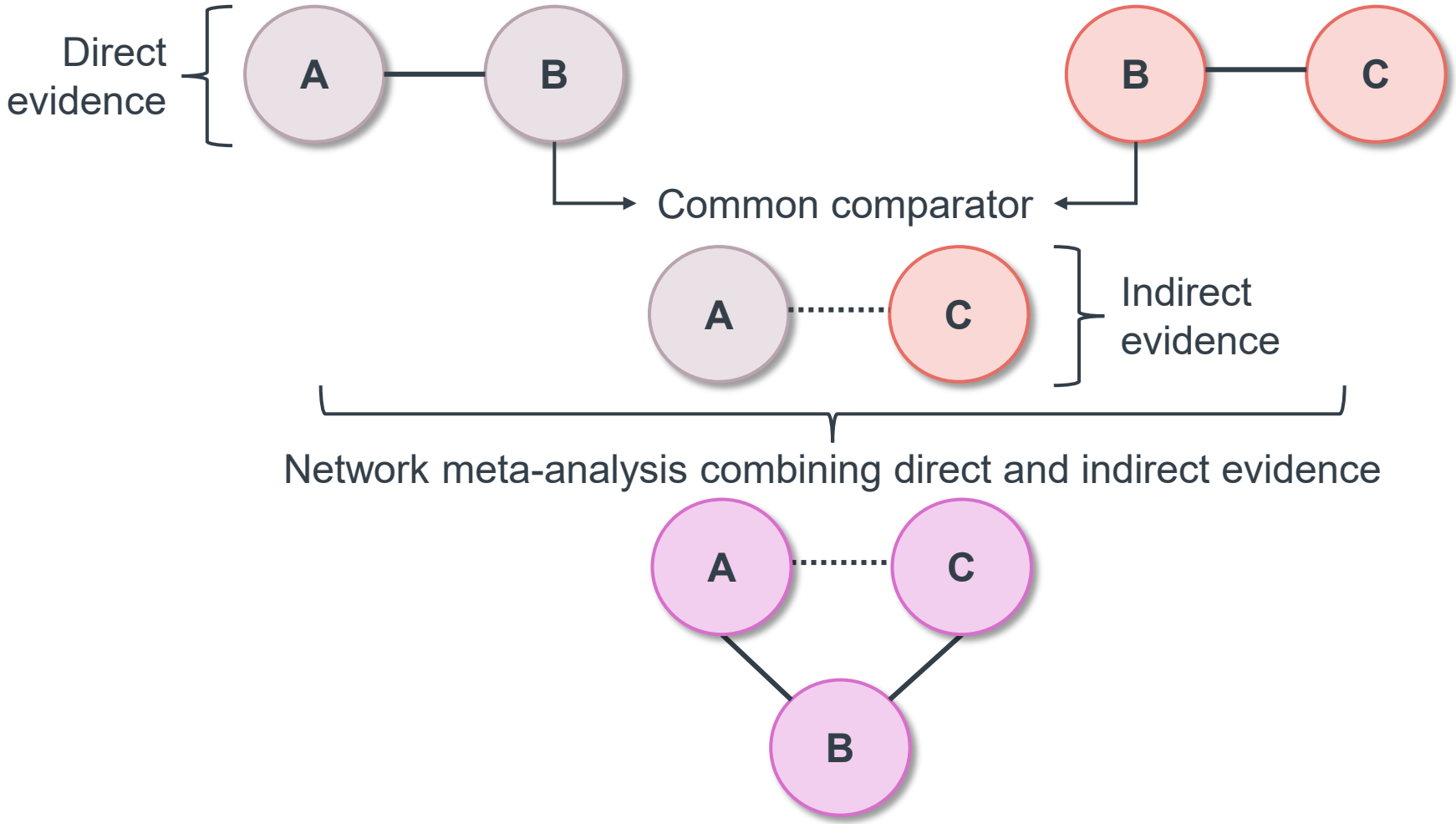
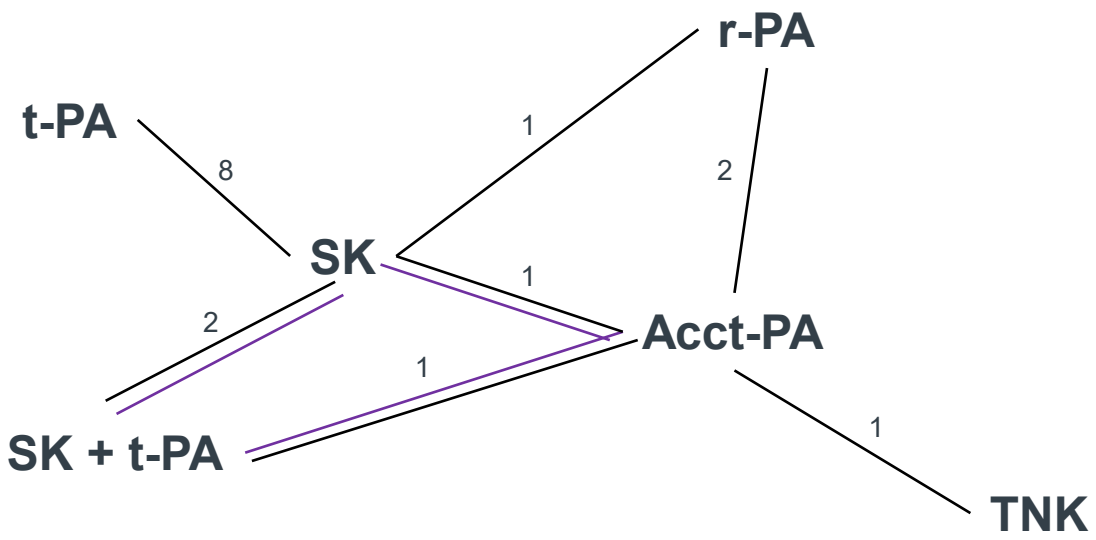


Figure adapted from Tonin FS *et al. Pharm Pract (Granada)* 2017; 15 (1): 943.

Example: Thrombolysis

Six treatments for acute myocardial infarction:

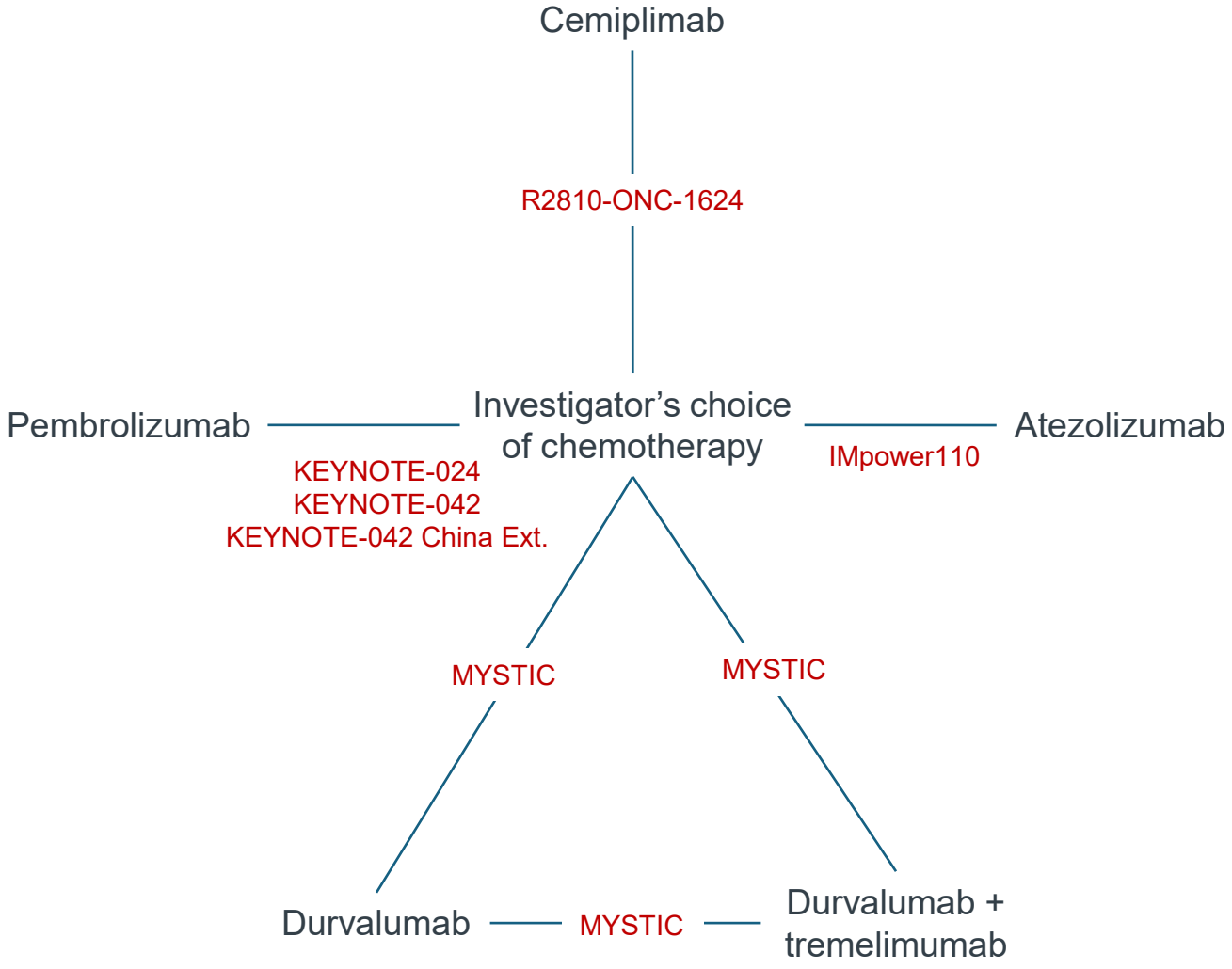
1. Streptokinase (SK)
2. Tissue plasminogen activator (t-PA)
3. Accelerated alteplase (Acct-PA)
4. Tenecteplase (TNK)
5. Reteplase (r-PA)
6. SK + t-PA



14 studies, 15 possible pairwise comparisons

Figure adapted from Boland A *et al. Health Technol Assess* 2003; 7 (15): 1–136.

Network of trials based on clinical systematic literature review focusing on monotherapy regimens in high PD-L1 population (PD-L1 ≥50% and TC3/IC3)



- Based on the systematic literature review focusing on monotherapy regimens, **six trials are included** in this network
- Studies are tiered based on the PD-L1 scoring method/assay used to determine patient eligibility (see table below)

Tier	Assay antibody	PD-1/PD-L1 inhibitors	Cell type	Concordance with 22C3 assay
Tier 1 PD-L1 TPS ≥50%, measured using the PD-L1 IHC **22C3 pharmDX** assay	**Dako 22C3**	Pembrolizumab (α-PD-1)	TC	N/A
		Cemiplimab (α-PD-1)	TC	N/A
Tier 2 High PD-L1 expression measured using an IHC assay that is <u>not</u> 22C3 pharmDX ***→Trials considered in sensitivity analyses only***	Ventana SP142	Atezolizumab (α-PD-L1)	TC/IC	Low (less sensitivity)
	Ventana SP263	Durvalumab (α-PD-L1)	TC	Moderate to high

IC, immune cell; IHC, immunohistochemistry; N/A, not applicable; PD-1, programmed cell death protein-1; PD-L1, programmed death-ligand 1; TC, tumor cell; TPS, tumor proportion score.
Figure adapted from Freemantle N *et al. Ther Adv Med Oncol* 2022; 14: 17588359221105024.

NMA: Possible biases

Bias	Difference in	Problems
Selection	Prognostic factors Predictive factors Will Rogers	**OK** **Unknown/unmeasured Predictive factors?**
Attrition	Lost to follow-up Not evaluated	**Differences across studies?**
Assessment	Methods Bias	**Differences across studies?**
Analysis	Multiplicity	**?**

MAIC, matching-adjusted indirect comparison.
Slide courtesy of Paolo Bruzzi.

(Network) meta-regression

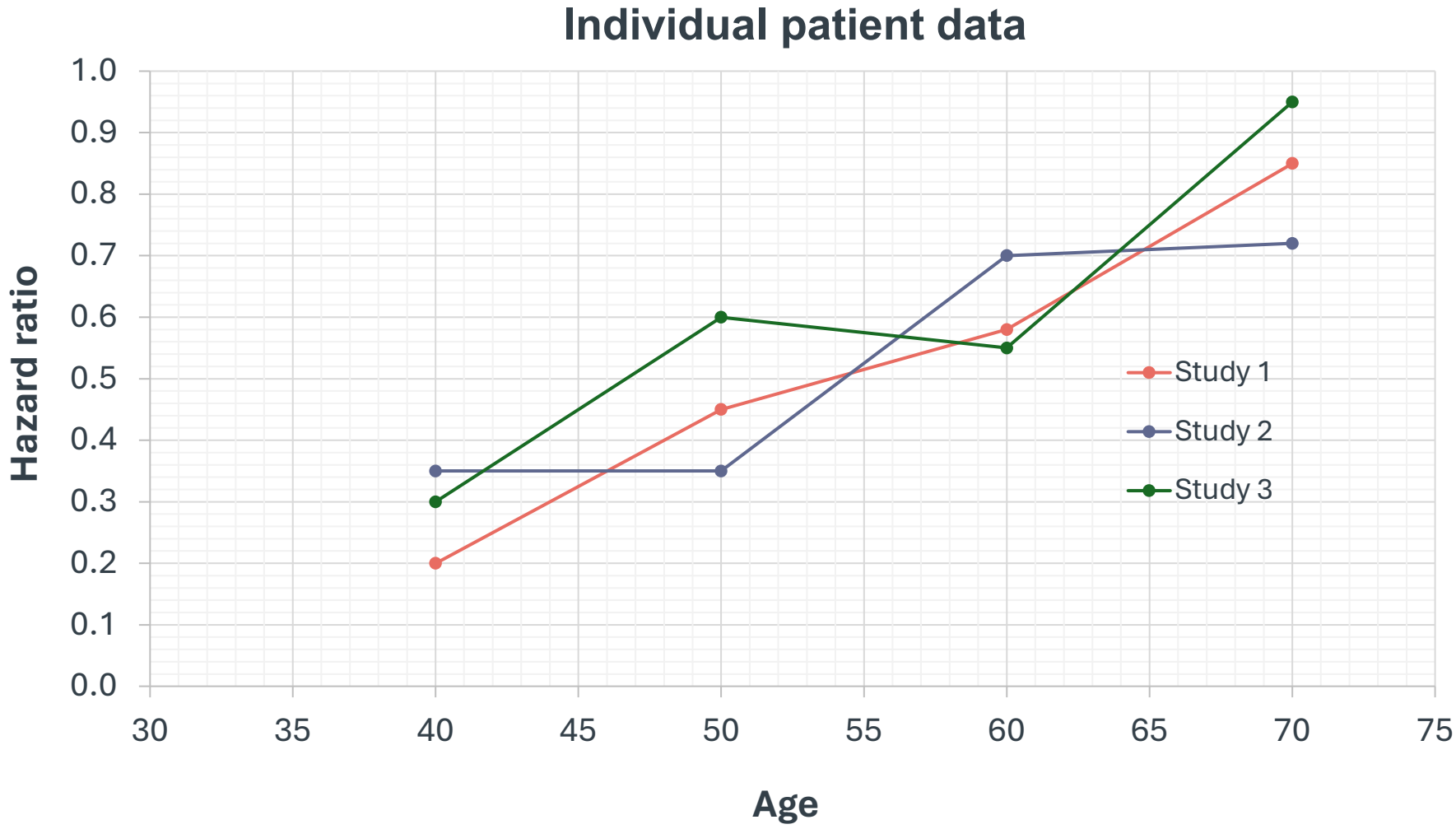
Used to evaluate whether the effect of an intervention is correlated with a continuous or ordered variable

Effectiveness of a therapy and age:

Published data	Mean age, years	HR
Trial 1	48	0.3
Trial 2	60	0.6
Trial 3	71	0.9

HR, hazard ratio.
Slide courtesy of Paolo Bruzzi.

(Network) meta-regression



Data provided by the speaker.
Slide courtesy of Paolo Bruzzi.

Requirements for indirect comparisons

	RCT	Same control Tx	IPD	Reliability
'Raw comparison'	No	No	No	-----
Multivariate analysis	No	No	Yes	-
Unanchored MAIC	Yes	No	1 trial	--
Anchored MAIC / STC	Yes	Yes	1 trial	+++
Network MA	Yes	Some	No	++
Meta-regression • Aggregate data • IPD	 Yes Yes	 Some Some	 No Yes	 ++ ++++

IPD, individual patient data; MA, meta-analysis; MAIC, matching-adjusted indirect comparison; RCT, randomized controlled trial; STC, simulated treatment comparison; Tx, treatment.
Slide courtesy of Paolo Bruzzi.

Indirect comparisons

In our enthusiasm for the statistical techniques to be used for indirect comparisons, we must not forget a simple truth:

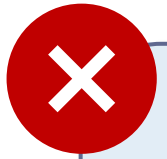
If these were effective, we would no longer need randomized trials

Slide courtesy of Paolo Bruzzi.

Conclusions

Indirect comparisons are useful tools for:

- Exploratory purposes (to generate hypotheses)
- Confirmatory purposes (to confirm opinions, formalize comparisons already made, etc.)



Indirect comparisons should not be used to:

- Demonstrate unknown effects (e.g. in subgroups)
- Reject consolidated theories/beliefs

Slide courtesy of Paolo Bruzzi.