The importance of appropriate subgroup evaluation for regulatory decision making

Hannover Medical School

Armin Koch, DE SAWP and BSWP member
Institut für Biometrie
Medizinische Hochschule Hannover
Carl-Neuberg-Str. 1
30625 Hannover

Did the importance of subgroups change?

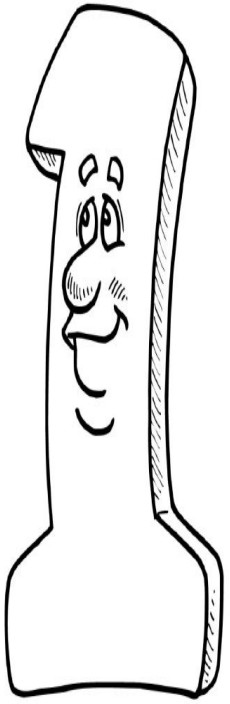
Medics will say no, because they were always interested in subgroups.

However, standards of evidence have changed:

- in former times (two-trials rule of the FDA) we had 2 (usually PBO controlled) studies in the US ***and*** 2 (usually active controlled) studies in the EU.
- nowadays assessment of efficacy and benefit/risk is based on one world-wide pivotal study planned with an adaptive design intended to justify licensing in all the ICH-regions.

☞ If consistency / replication is considered important, nowadays assessment needs to be done *within* instead of *between* studies.

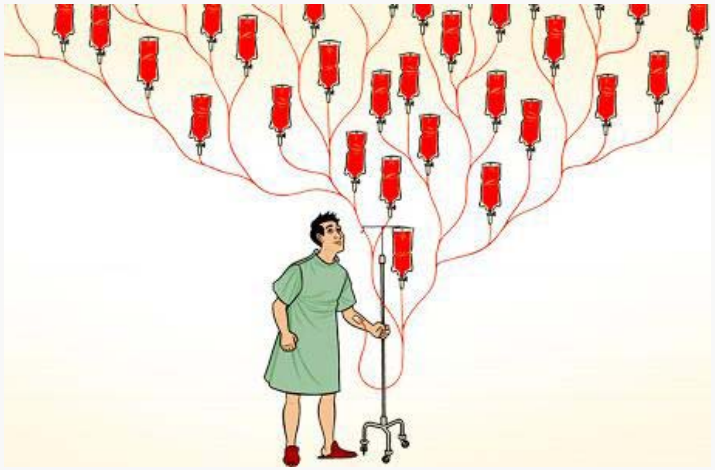
☞ Biomarkers become increasingly important in drug research and challenge all concepts of subgroup assessment



The future, maybe:
Personalized Medicine:

Definition:

use of genetic or other molecular biomarker information to improve the safety, effectiveness and health outcomes of patients via more efficiently targeted risk stratification, prevention and tailored treatment management approaches.



from M. Papaluca-Amati

Subgroups in Phase III clinical trials

Paradigm of phase III clinical research:

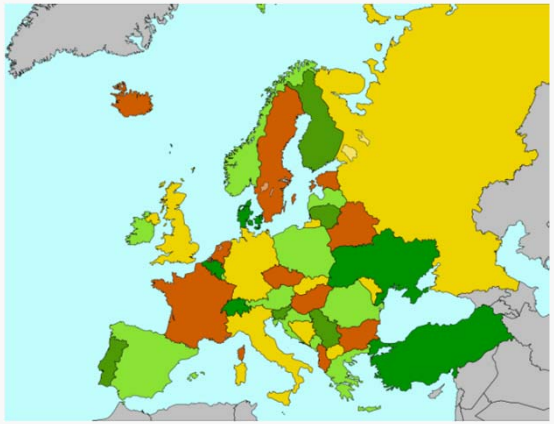
Trials should not fine-tune the patient population.

Flip-side of the coin:

Consistency of the treatment-effect in relevant subgroups of the patient population is non-trivial and needs to be verified.

Relevant subgroups:

Something that needs to be defined, but demography, gender, disease characteristics, co-medication, center, region and country are plausible candidates.

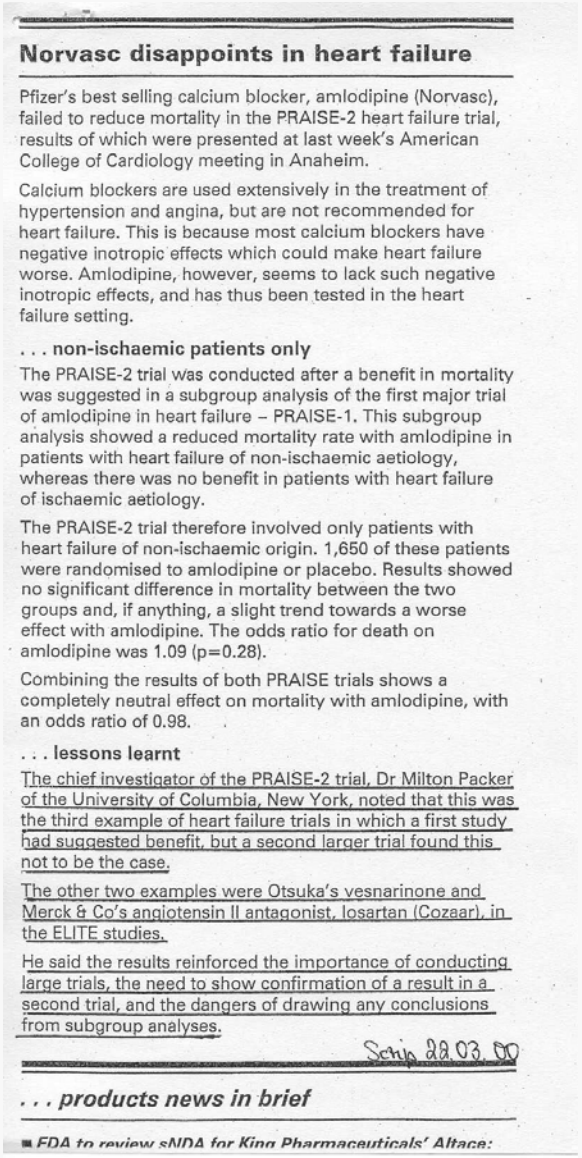


Empirical evidence exists, that looking into subgroups for significance may be dangerous

HF trial PRAISE 1 suggested efficacy of Amlodipine in subgroup of non-ischemic patients,

but PRAISE 2 didn't replicate benefit on mortality (P=0.28).

Luckily no treatment recommendation has been based on subgroups, but replication has been attempted.



Norvasc disappoints in heart failure

Pfizer's best selling calcium blocker, amlodipine (Norvasc), failed to reduce mortality in the PRAISE-2 heart failure trial, results of which were presented at last week's American College of Cardiology meeting in Anaheim.

Calcium blockers are used extensively in the treatment of hypertension and angina, but are not recommended for heart failure. This is because most calcium blockers have negative inotropic effects which could make heart failure worse. Amlodipine, however, seems to lack such negative inotropic effects, and has thus been tested in the heart failure setting.

. . . non-ischaemic patients only

The PRAISE-2 trial was conducted after a benefit in mortality was suggested in a subgroup analysis of the first major trial of amlodipine in heart failure – PRAISE-1. This subgroup analysis showed a reduced mortality rate with amlodipine in patients with heart failure of non-ischaemic aetiology, whereas there was no benefit in patients with heart failure of ischaemic aetiology.

The PRAISE-2 trial therefore involved only patients with heart failure of non-ischaemic origin. 1,650 of these patients were randomised to amlodipine or placebo. Results showed no significant difference in mortality between the two groups and, if anything, a slight trend towards a worse effect with amlodipine. The odds ratio for death on amlodipine was 1.09 (p=0.28).

Combining the results of both PRAISE trials shows a completely neutral effect on mortality with amlodipine, with an odds ratio of 0.98.

. . . lessons learnt

The chief investigator of the PRAISE-2 trial, Dr Milton Packer of the University of Columbia, New York, noted that this was the third example of heart failure trials in which a first study had suggested benefit, but a second larger trial found this not to be the case.

The other two examples were Otsuka's vesnarinone and Merck & Co's angiotensin II antagonist, losartan (Cozaar), in the ELITE studies.

He said the results reinforced the importance of conducting large trials, the need to show confirmation of a result in a second trial, and the dangers of drawing any conclusions from subgroup analyses.

Scrip 22.03.00

. . . products news in brief

■ *FDA to review sNDA for King Pharmaceuticals' Altace;*

Consequence:
Positive conclusions require pre-specification

Issue has been discussed within and outside CHMP:

"When exploratory, these [subgroup analyses] should be interpreted cautiously. Market approval of a compound is based on the overall trial results, and, importantly no drug has so far been approved or not approved either in the US or in the EU on the basis of subgroup analysis."

(Maggioni, Darne, Atar, Abadie, Pitt, Zannad
Cardiology (107), 97 2007)

European guidance on multiplicity in clinical trials states that:

A specific claim of a beneficial effect in a particular subgroup requires pre-specification of the corresponding null hypothesis and an appropriate confirmatory analysis strategy. It is highly unlikely that claims based on subgroup analyses would be accepted in the absence of a significant effect for the overall study population.

(PtC on Multiplicity issues in clinical trials, Sec. 4)

Sometimes overall results do not tell the truth:

Primary endpoint is ESRD and more severe events:

Table: primary composite endpoint in a study comparing test, reference and Placebo in patients with diabetic nephropathy on background of anti-hypertensive therapy:

	Test	***Reference***	***RR 95%-CI P-value***
all patients	189/579 (32.6%)	233/567 (41.1%)	0.794 (0.682; 0.926) 0.0032
male	104/378 (27.5%)	145/359 (40.4%)	0.681 (0.554; 0.837) 0.0002
female	85/201 (42.3%)	88/208 (42.3%)	1.000 (0.797; 1.254) 0.9980
adjusted analysis			0.811 (0.696; 0.944) 0.0070*

* Breslow&Day-Test for heterogeneity P-value is 0.0141

Sometimes overall results do not tell the truth:

CV safety endpoint:

Table: secondary cardiovascular composite endpoint1 in a study comparing Test, Reference and Placebo in patients with diabetic nephropathy on background of anti-hypertensive therapy:

	Test	***Reference***	***RR*** ***95%-CI*** ***P-value***
all patients	141/579 (24.4%)	129/579 (22.8%)	1.093 (0.887; 1.347) 0.4046
male	86/378 (22.8%)	90/359 (25.1%)	0.908 (0.701; 1.174) 0.4608
female	55/201 (27.4%)	39/208 (18.8%)	1.459 (1.017; 2.095) 0.0405
adjusted analysis			1.065 (0.863; 1.314) 0.5559*

1endpoint is a composite of cardiovascular death, nonfatal MI, hospitalisation for HF, stroke, above-ankle amputation

Benefit/risk assessment

An overall positive treatment effect may be put into perspective in subgroups by:

- no effect in a relevant subgroups of the patient population
- indication of harm
- negative benefit/risk in subgroups
- substantial heterogeneity

Assessment of subgroups

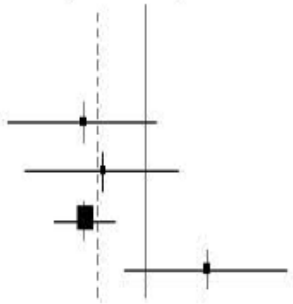
- is an essential part of benefit/risk assessment
- reflects, how physicians decide, who should be treated

The dangerous impression of being in the "comfort zone"

The Plato trial, comparing Ticagrelor to Clopidogrel in 18,000 patients with ACS demonstrated superiority, but regional differences became obvious from the results (p_{Het}~0,05).

Is it wise to pretend that this is an American problem?

Figure 20 Forest Plot: Results by Region (K-M)



Characteristic	Total Patients	KM % at Month 12 Ti.	KM % at Month 12 Cl.	HR (95% CI)
Region				
Asia/Australia	1714	11.4	14.8	0.80 (0.61, 1.04)
Cent/Sth America	1237	15.2	17.9	0.86 (0.65, 1.13)
Euro./Mid E./Afr.	13859	8.8	11.0	0.80 (0.72, 0.90)
North America	1814	11.9	9.6	1.25 (0.93, 1.67)

Torn between two extremes

The subtle balance between:

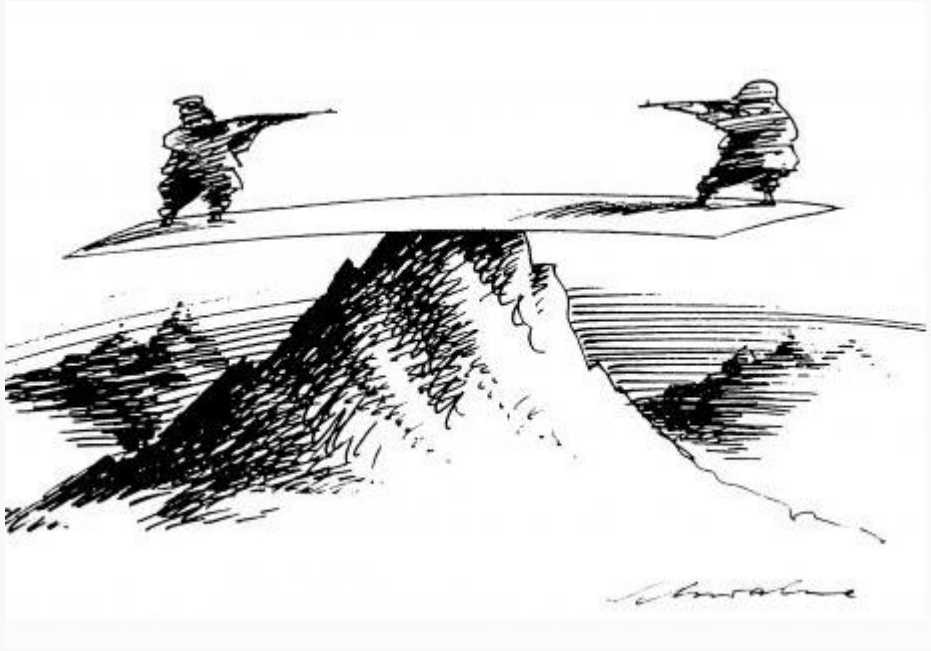
increasing the type-1-error by means of multiple testing in subgroups

and

overlooking important untoward effects in subgroups

can only be ameliorated by means of pre-planning and specification of what is a relevant subgroup and a relevant difference between subgroup-effects at the planning stage.

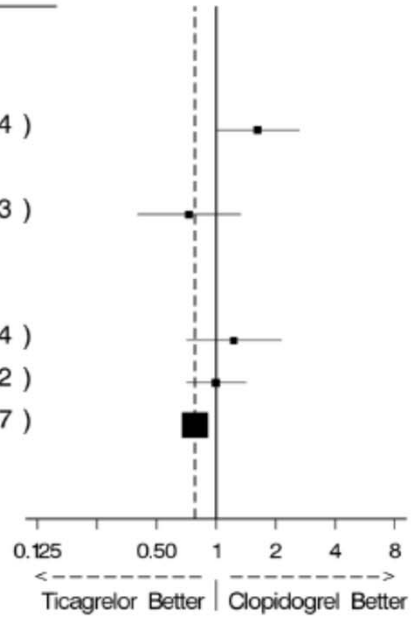
In this, statisticians fear eventually too much to be mislead by (good quality data).



Subgroup assessment is informative

One of the many additional analyses for PLATO:

Region	ASA Dose (mg)	Ticagrelor N	Ticagrelor E	Clopidogrel N	Clopidogrel E	HR (95% CI)
US						
	>=300	324	40	352	27	1.62 (0.99, 2.64)
	>100 – <300	22	2	16	2	
	<=100	284	19	263	24	0.73 (0.40, 1.33)
Non–US						
	>=300	140	28	140	23	1.23 (0.71, 2.14)
	>100 – <300	503	62	511	63	1.00 (0.71, 1.42)
	<=100	7449	546	7443	699	0.78 (0.69, 0.87)



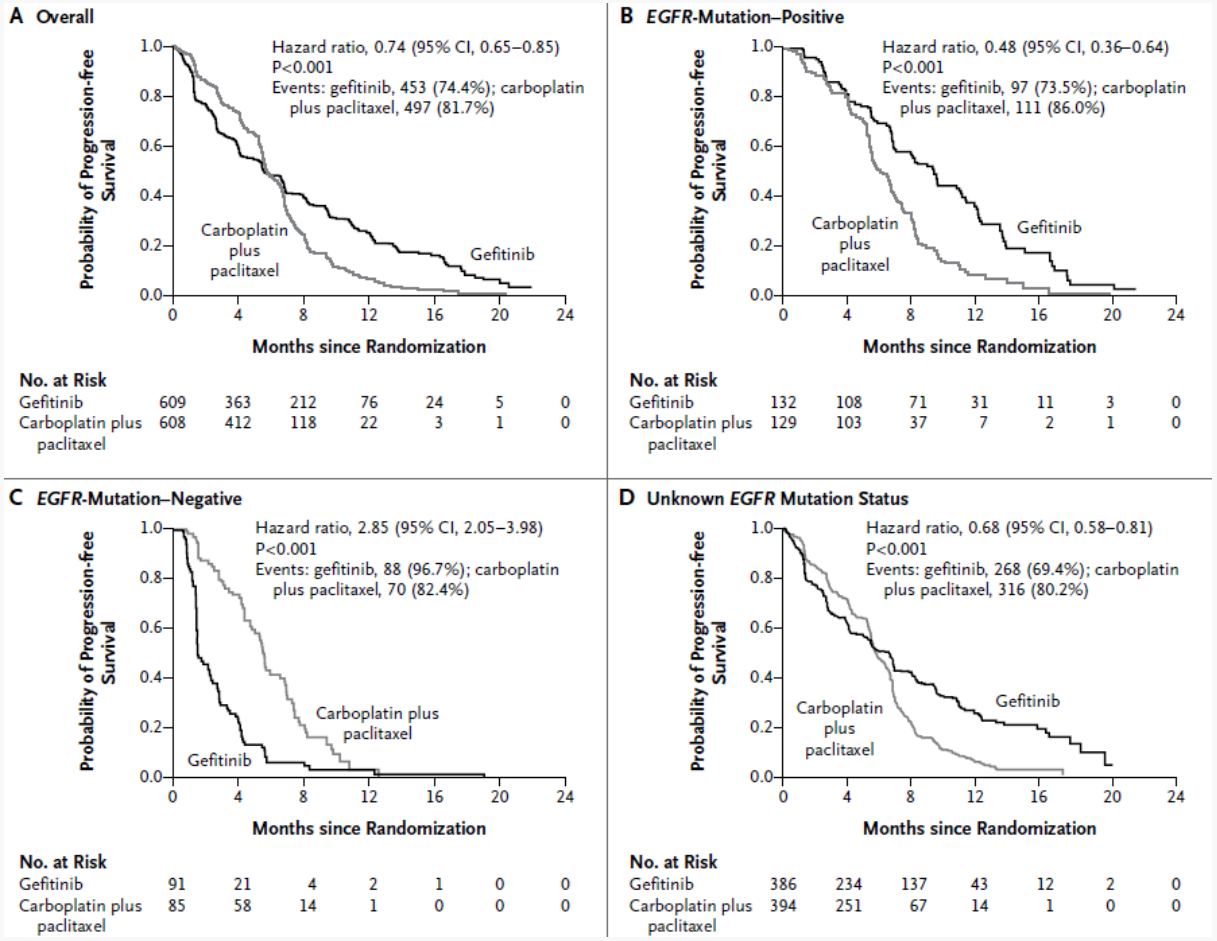
(from the Astra Zeneca preparatory material)

Subgroup-GL: Under which conditions could a subgroup finding be convincing?

Crossing survival as an example for non-conclusive overall outcome

EGFR mutation status in the IPASS-trial comparing Gefitinib to Carboplatin + Paclitaxel in patients with NSCLC.

(Mok et al. (2009)



Subgroup Guideline:

A discussion about what is needed to challenge the main outcome of a trial from the perspective of:

Heterogeneity extent of differences in the target patient population regarding prognostic or predictive factors. The more heterogeneous the population, the more important are subgroup investigations.

Consistency extent to which estimated treatment effects in relevant subgroups assures that the overall treatment effect applies to the breadth of the trial population.

Credibility describes the extent to which subgroup findings can be concluded as being well substantiated and hence relied on for decision making. Credibility depends on the degree of well-founded, a priori definition, the **biological plausibility** (mainly a clinical or pharmacological judgement) for a particular finding and **replication**.

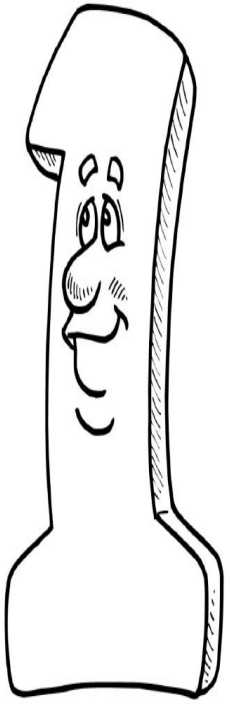
The ICH-E17 discussion:

Needs and standards of evidence have changed:

- in former times (two-trials rule of the FDA) we had 2 (usually PBO controlled) studies in the US ***and*** 2 (usually active controlled) studies in the EU.
- nowadays assessment of efficacy and benefit/risk is based on one world-wide pivotal study planned with an adaptive design intended to justify licensing in all the ICH-regions.

☞ If one study is supposed to provide the required evidence for many regions, this trial will need a lot of diligence at the planning stage.

☞ Nobody can have her/his “own significance” (so we need to understand the overall treatment effect), but there are enormous opportunities to learn (Japanese are also living in the US ☺).



Under which conditions could a subgroup finding be convincing?

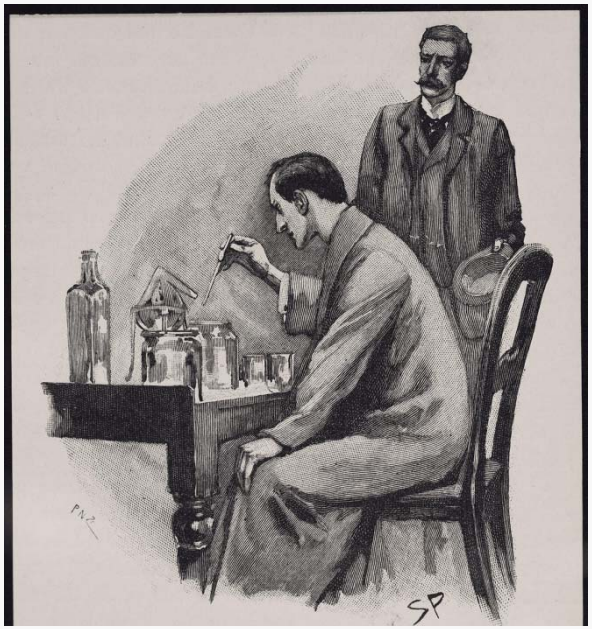
Guiding principles for this case-by-case decision include:

- a pharmacological rational, or a mechanistically plausible explanation, should at best exist for differential treatment effects in subgroups,
- a priori, or external evidence should exist that the subgroup is a well defined entity ("well known"),
- stratification of the randomisation as an indicator,
- convincing P-value (not borderline in a borderline trial)
- the overall outcome of the trial should at a minimum substantiate the claim that no harm is introduced by the experimental treatment,
- good overall safety and subgroup safety, or convincing benefit/risk assessment from subgroup is possible
- Replication (from other trials, from phase II trials, from other trials in the same indication)

Biomarkers challenge the concept

Compare to: Sung et al: BMJ 2012 (344), published 15 March 2012.

… and at the assessment stage:
a signal is a signal is a signal…



Methodological problems exist with repeated testing even if we restrict ourselves to relevant effects, but

In first place a signal should be taken seriously!

- ...when you have eliminated the impossible, whatever remains, however improbable, must be the truth.
- Eliminate all other factors, and the one which remains must be the truth.
- It is an old maxim of mine that when you have excluded the impossible, whatever remains, however improbable, must be the truth.

(Sherlock Holmes, various occasions)

… probably a chance finding
… should be concluded after careful assessment, only

Finis:

- Subgroups add “credibility” to the overall outcome of the trial,
- subgroup analyses are an integral part of benefit/risk assessment,
- T1E of different importance in proof of efficacy and B/R assessment,
- improvements in flagging procedures and likelihood of chance findings are highly welcome and PSI working group’s contribution on methodology is highly appreciated,
- statisticians can help beyond that!