Reliably basing conclusions on subgroups of clinical trials

Hannover Medical School

Armin Koch
Institut für Biometrie
Medizinische Hochschule Hannover
Carl-Neuberg-Str. 1
30625 Hannover

Views expressed in this presentation are those of the author and not necessarily those of the SAWP or the BSWP at the ema

Planning, conduct and assessment of studies

I believe:

- everything that can be pre-planned, should be pre-planned: assessment of findings in this context follows established rules,
- this is sometimes not enough: during assessment not-pre-planned, but plausible subgroups are of importance,
- statistics can be helpful to tackle this problem better

Situations to be distinguished:

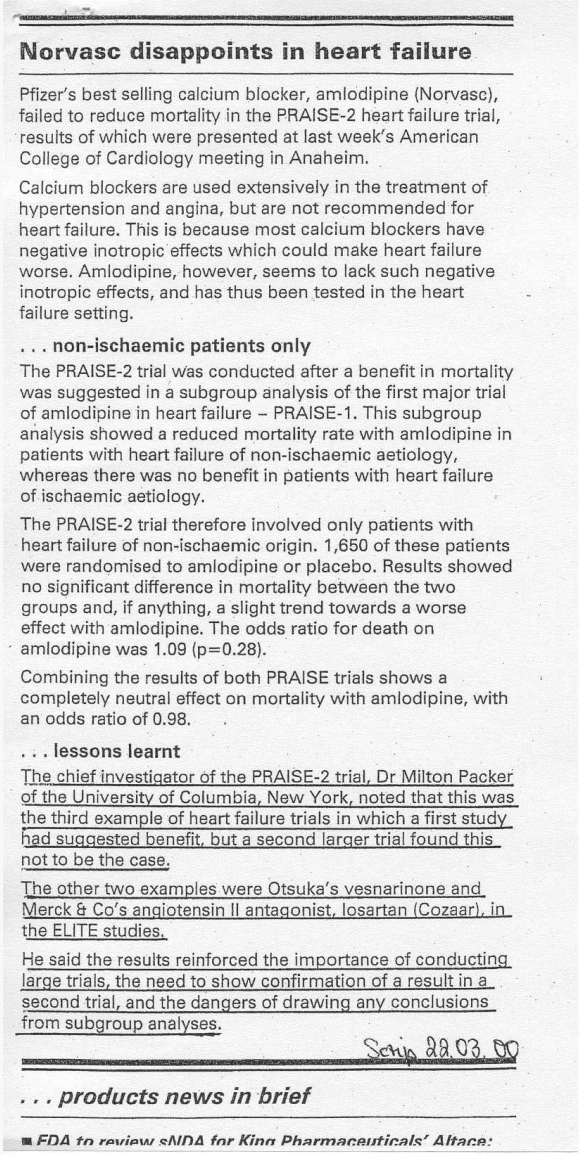
- pre-specification of a subgroup (and confirmatory strategy)
- excluding a subgroup from a significant trial
- "refinement" of treatment effect in a significant trial
- positive subgroup in a negative trial

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)

Perception of regulatory work with respect to assessment strategies

$P < 0.05$ (two-sided)	$P \geq 0.05$ (two-sided)
Cartoon: man signing license	Cartoon: man trashing application

In case significance was the only thing to assess in drug licensing, regulatory agencies would not be needed.

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

The precautionary principle:

PtC on multiplicity issues in clinical trials:

Sec. 4 Summary: “... A license may be restricted if unexplained strong heterogeneity is found in important sub-populations, or if heterogeneity of the treatment effect can reasonably be assumed but cannot be sufficiently evaluated for important sub-populations”.

ICH-E9:

It is even more important to understand the basis of any heterogeneity characterized by marked qualitative interactions, and failure to find an explanation may necessitate further clinical trials before the treatment effect can be reliably predicted.

The precautionary principle

In some instances 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

Please note:

P-values can guide assessment, but significance can not be the criterion to decide.

Under which conditions could a subgroup finding be convincing?

In case the overall trial is not significant usually from statistical grounds no further confirmatory testing is possible (type I error is exhausted).

Any step further can only be based on a case by case decision.

Most important:

- a generally acceptable argument should exist, why straightforward replication is not possible,

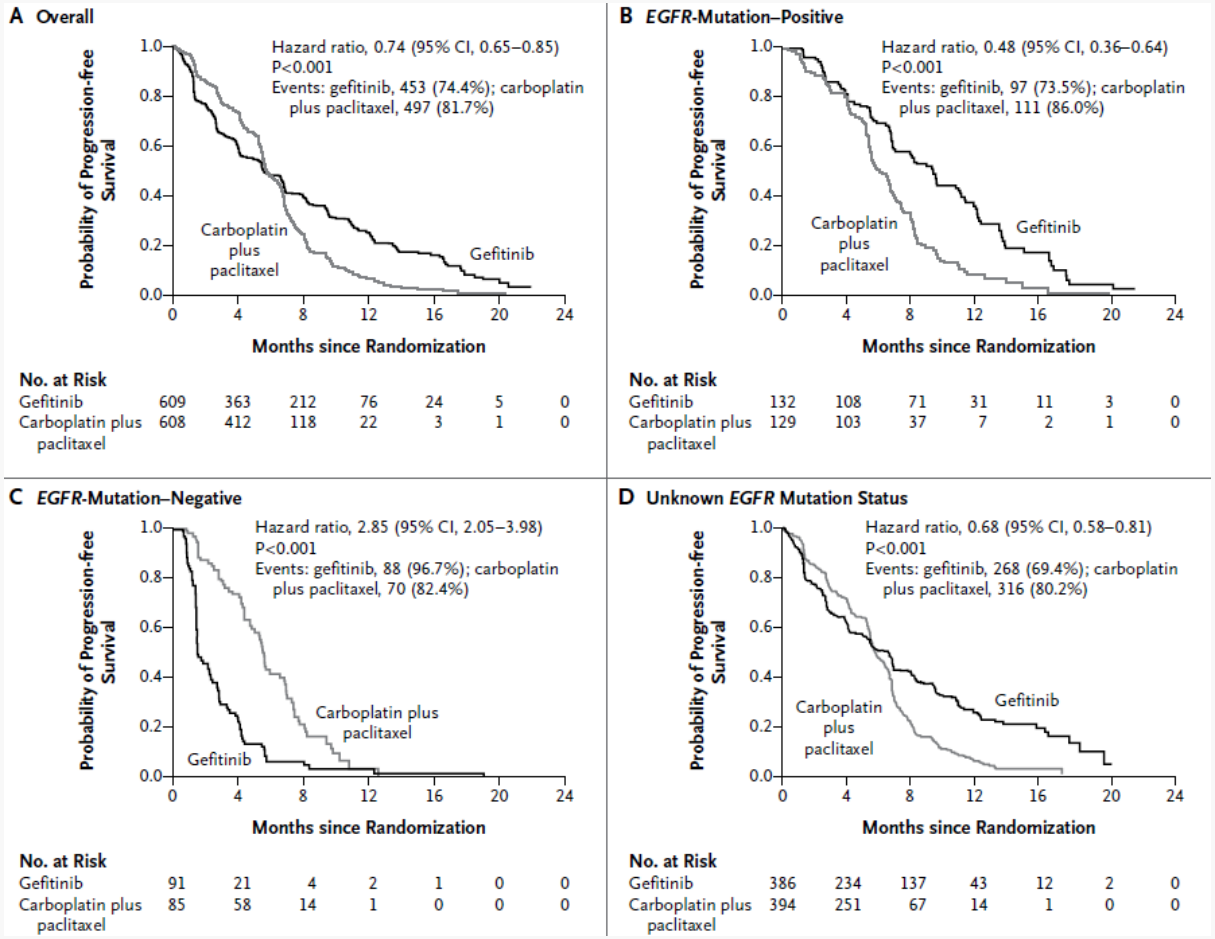
because,

- replication of promising subgroup-findings in an independent trial

is standard if trial is overall not significant, but subgroup findings suggest efficacy at least in parts of the patient population.

Under which conditions could a subgroup finding be convincing?

Crossing survival as an example for non-conclusive overall outcome



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. It is not possible to claim treatment benefit in one subgroup, if another subgroup suggests that also harm may be introduced,
- good overall safety and subgroup safety, or convincing benefit/risk assessment from subgroup is possible
- substantial heterogeneity?

Under which conditions could a subgroup finding be convincing?

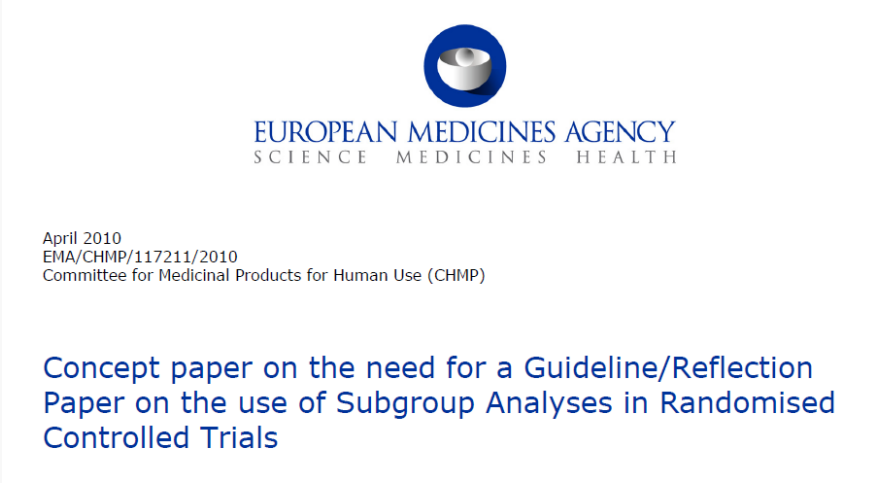
Replication:

- applications often include more than one trial,
- similar trials may be inspected for similar findings in subgroups (yet undetected class effects)

that may be used to add credibility to a positive conclusion.

European guidance is underway,

will discuss assessment of not-preplanned subgroups in "significant" and non-significant trials.



Finis:

- the end of the journey is still unclear
- because you need to be very brave if you wish to leave the arena of pre-planned decision making with full control of the type-1-error
- P<0.05 is no longer the criterion, epidemiology-style decision making needed
- the discussion about biomarkers (well known and not so well known) favorably adds to what needs to be considered in this context.

Cartoon: Lucky Eddy testing bridge