Subgroups, clinical sense and statistical challenge

European Statistical Meeting on Subgroup Analyses

Kit Roes, 30 Nov 2012

University Medical Center
Utrecht

Overview

- Perspectives (again)
- (Example of) Subgroup analyses in “negative” trial
  - &Introduce suggested criteria
  - What do we/you believe (additivity)
- Inferential points
- Conclusions

Perspectives

Perspective of treating physician

Evidence based decision for the (next) patient to treat, selecting from the available treatment options.

Perspective of market authorisation of a new drug

Evidence based decision of allowing physicians to add a new drug to their treatment options.

Provide information to guide the prescribing physician.

Subgroups: Perspectives from regulatory*

- To *confirm* consistency across *(all)* subgroups of clinical importance.

- To *identify* safety problems limited to a subgroup.
- To *identify* subgroups with larger effect, in positive study.
- To *check* specific subgroups that a priori are suspected to show less or no treatment effect.

- To *identify* subgroup(s) that demonstrate relevant effect, in case the overall effect is not significant.

*Grouin, Coste, Lewis (2005), J. of Biopharm. Stat.

Regulatory environment moving towards

- Including relative efficacy and comparative effectiveness into drug development plans.*
- Information from patient and payer perspective available at market authorisation.
- Perspective of stratified prediction of treatment effects increasingly important (e.g. companion diagnostics).

*Eichler, Bloechl-Daum, Abadie, Barnett, König and Pearson (2010). Relative efficacy of drugs: an emerging issue between regulatory agencies and third-party payers. Nat Rev Drug Disc

And these perspectives are not the same….

In many areas the “best available” treatment (strategy) for a particular patient is not known.

Treatment decisions are less evidence based as we would like to see them.

Dodgy subgroup analyses

If your drug didn’t win overall in your trial, you can chop up the data in lots of different ways, to try and see if it won in a subgroup: maybe it works brilliantly in Chinese men between fifty-six and seventy-one. This is as stupid as playing ‘Best of three… ….Best of five…’ And yet it is commonplace.

Example: the DECS trial

Prophylactic corticosteroids in cardiac surgery to attenuate the inflammatory response to cardiopulmonary bypass and surgical trauma.

ORIGINAL CONTRIBUTION

JAMA, November 7, 2012—Vol 308, No. 17

Intraoperative High-Dose Dexamethasone for Cardiac Surgery
A Randomized Controlled Trial

- RCT in 4494 patients undergoing cardiac surgery with cardiopulmonary bypass

- Single high dose dexamethosone (1 mg/kg) versus placebo

- Primary outcome: The composite of death, myocardial infarction, stroke, renal failure, or respiratory failure, within 30 days of randomization.

Primary Results

Table 2. Primary Study End Point and Components of the Primary Study End Point in the Dexamethasone and Placebo Groups

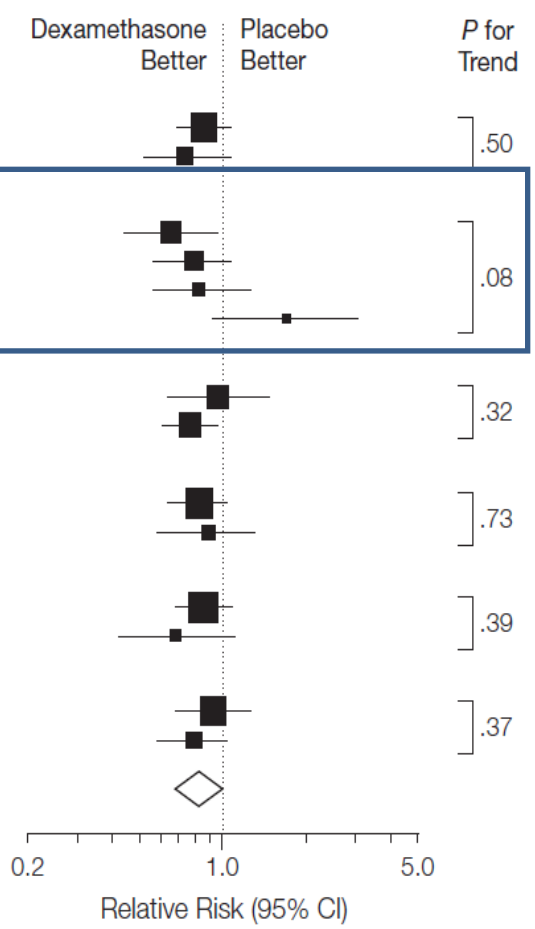
	No. (%) of Patients		
	Dexamethasone (n = 2235)	Placebo (n = 2247)	Relative Risk (95% CI)
Primary study end point[a]	157 (7.0)	191 (8.5)	0.83 (0.67-1.01)
Components of the primary study end point			
Death	31 (1.4)	34 (1.5)	0.92 (0.57-1.49)
Myocardial infarction	35 (1.6)	39 (1.7)	0.90 (0.57-1.42)
Stroke	29 (1.3)	32 (1.4)	0.91 (0.55-1.50)
Renal failure	28 (1.3)	40 (1.8)	0.70 (0.44-1.14)
Respiratory failure	67 (3.0)	97 (4.3)	0.69 (0.51-0.94)

P-value on primary composite outcome: 0.07

Subgroup analyses (*pre-planned*)

Figure 2. Forest Plot of Subgroup Analyses

Subgroup	No. of Patients		No. of Events		Relative Risk (95% CI)	*P* for Trend
	Dexamethasone	Placebo	Dexamethasone	Placebo		
Sex						
Male	1622	1628	112	130	0.86 (0.68-1.10)	.50
Female	613	619	45	61	0.74 (0.52-1.08)	
Age, y						
<65	913	962	38	62	0.65 (0.44-0.96)	.08
65-74	818	781	57	70	0.78 (0.56-1.09)	
75-79	344	336	38	44	0.84 (0.56-1.27)	
≥80	159	168	24	15	1.69 (0.92-3.10)	
EuroScore						
0-4	1025	1060	43	46	0.97 (0.64-1.45)	.32
≥5	1180	1147	113	142	0.77 (0.61-0.98)	
Diabetes						
No	1816	1805	115	142	0.81 (0.64-1.02)	.73
Yes	415	436	41	49	0.88 (0.59-1.30)	
COPD						
No	1987	1978	133	153	0.87 (0.69-1.08)	.39
Yes	243	266	24	38	0.69 (0.43-1.12)	
Prolonged CPB						
No	1674	1659	85	92	0.92 (0.69-1.22)	.37
Yes	536	563	71	96	0.78 (0.59-1.03)	
Total	2235	2247	157	191	0.83 (0.67-1.01)	



Subgroup analyses (pre-planned)

In patients younger than 65 years, the RR for mortality was 0.42 (95% CI, 0.13-1.34; *P* = .13), but it gradually increased with age to 3.87 (95% CI, 1.10-13.6; *P*=.02) in patients aged 80 years or older.

Secondary endpoints

No a priori multiplicity correction for (19) secondary endpoints.

Only exploratory conclusions, and only if p-value below 0.0025.

“In an exploratory analysis of secondary end points, a *reduced incidence of respiratory failure* was found, which was accompanied by an *overall reduced time to weaning from mechanical ventilation*, a *lower risk of pneumonia*, and a *reduction in ICU and hospital stay*.”

In conclusion: DECS

In conclusion, in our trial of adults undergoing cardiac surgery, the use of intraoperative dexamethasone did not reduce the 30-day incidence of major adverse events compared with placebo.

In the paper:

- suggested new trial on the pulmonary outcomes
- should also consider patient selection for the therapy
  - a larger beneficial effect in younger patients and no apparent benefit in those aged 80 years or older.

And what should cardiac surgeons do?

Subgroup Analysis of Trials Is Rarely Easy (SATIRE)

Sun et al. BMJ 2012;344:e1553 doi: 10.1136/bmj.e1553
(Published 15 March 2012)

1. *Is the subgroup variable a characteristic at randomization? ☑*
2. Was the subgroup variable a stratification factor at randomisation?
3. *Was the hypothesis specified a priori? ☑*
4. Was the subgroup effect one of a small number of hypothesized effects tested? ☑
5. *Does interaction test suggests a low likelihood that chance explains the apparent subgroup effect (**sign at 5%)?** ☑?*
6. Is the significant interaction effect independent of other potential subgroup effects? ☑
7. Was the correct direction of subgroup effect specified a priori?
8. Is the interaction consistent across studies?
9. Is the interaction consistent across closed related outcomes within the study? ☑
10. Is there indirect evidence that supports the hypothesized interaction?

What do we/you believe?

- If there is a treatment effect, it is equal for all subjects (strong additivity).
- If there is a treatment effect, it will likely vary between subjects.
- (treatment effect also visible in variance)

Senn, Statist. Med. 2004; 23:3729–3753

Subgroup analysis in a “negative” trial

Strategies & overall type I error levels

Simple subgroup testing at 2.5% one-sided

- Type I error increased:
  - Up to almost 5% for small subgroups
  - Up to 4.2% for subgroup 50%
  - Approaching 2.5% for large subgroups

Interaction test (5%) & subgroup testing (at 2.5% one-sided)

- Type I error increased
  - Up to 3.5% for small subgroups
  - Up to 3.0% for subgroup 50%
  - Approaching 2.5% for large subgroups

Subgroup analysis in a “negative” trial

Replication of the same trial (overall type I error 0.000625)

- Type I error increased
  - Up to almost 0.0012 for small subgroups
  - Up to 0.0009 for subgroup 50%
  - Approaching 0.000625 for large subgroups

Replication by trial in the subgroup

- Type I error (of observing it twice)
  - Up to about 2.55% for small subgroups
  - Approaching 2.5% for large subgroups

Reflection on the DECS example

- Overall result non-significant (p=0.07), interaction test (p=0.08) and young age group (RR, 0.65; 95% CI, 0.44-0.96; *P* = .03).

- Heterogeneity primarily driven by mortality.

- Overall results might justify action
  - (precautionary, safety in elderly)
  - (support for use in the younger, given exploratory evidence of other advantages – which need to hold in the younger)

Inferential points in single trial

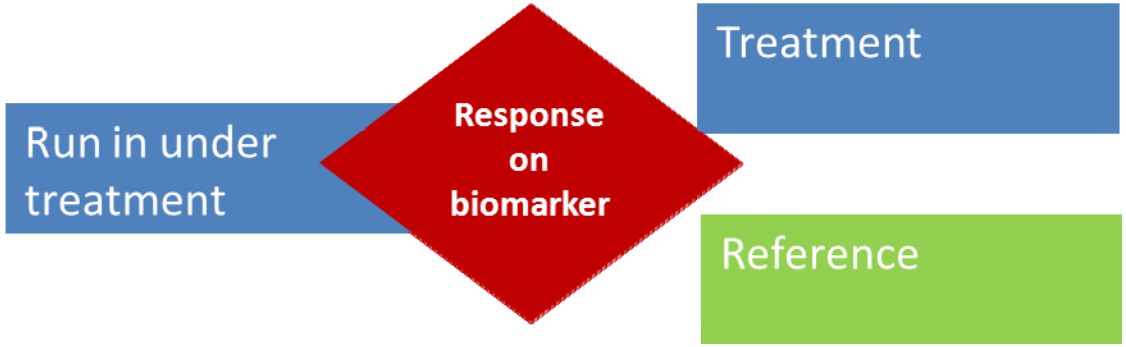
Subgroup analysis

- Standard of evidence to define:
  - For positive action (to treat, to license).
  - For negative action (not to treat, not to license).

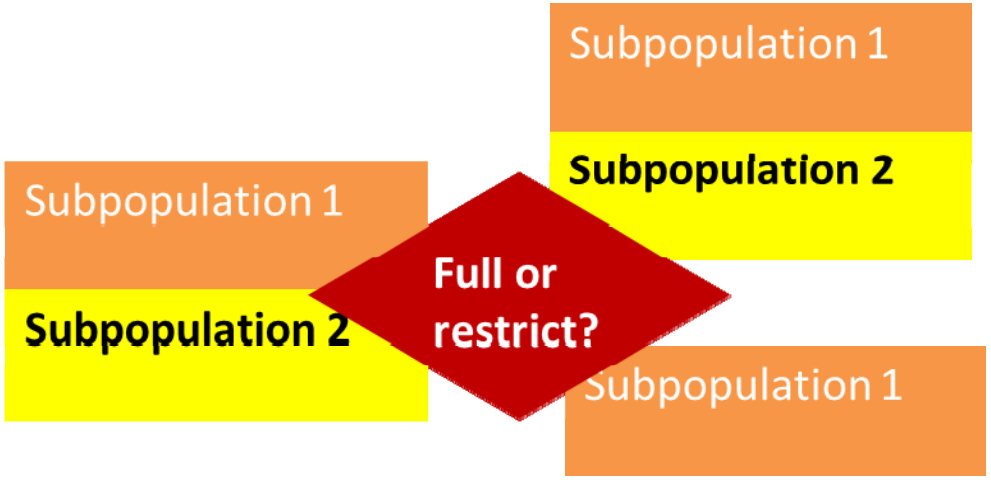
- Standard of evidence to define:
  - Justified replication through additional experimentation.
  - Are all those proposed seriously worthwhile?

Enrichment designs

Restriction randomisation to promising subgroup before randomisation.



Restriction randomisation to promising subgroup based on interim results.



Inferential points innovative designs

- Potentially increase efficiency.
- Create substantial “missing data”.
  - Subgroups / subpopulations may not be so black and white (not even if they are genetic).
  - Effective (and safe) treatment may be withheld from patients that could benefit (e.g. oncology).

Discussion

(1) Regulatory environment moving towards

- Including relative efficacy and comparative effectiveness into drug development plans.
- Information from patient and payer perspective available at market authorisation.
- Supported by research program (Innovative Medicines Inititiative)

(2) Treating physician perspective

- More complete and definite conclusions (from trials), including tailoring treatment to patients evidence based.

Should we focus on *DESIGN* of (Large? Pragmatic?) trials that allow more definite conclusions on multiple subgroups?

Conclusion

- Challenges of interpreting subgroups analyses, particularly if overall results are ambiguous.
- Also in such cases, interaction test and replication provide substantial protection.
- From the patient and treating physician perspective, *DESIGN* of trials that can provide fuller answers may be preferable to selective, but efficient designs.