Simplicity, Complexity and Modelling in Clinical Trials

Stephen Senn

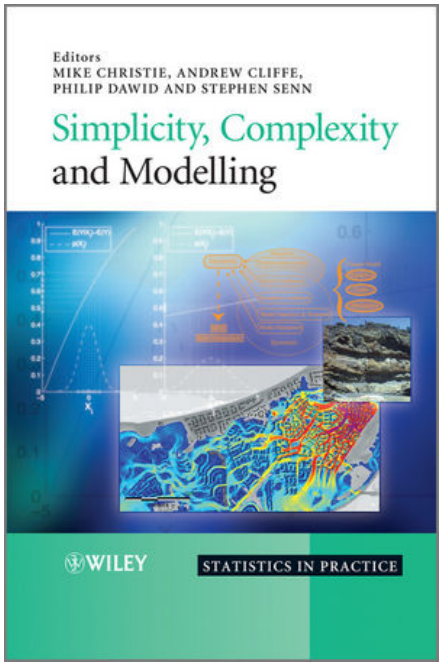
LUXEMBOURG INSTITUTE OF HEALTH

RESEARCH DEDICATED TO LIFE

Acknowledgements

This work is partly supported by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. "IDeAl"

However, some of the ideas grew out of work on the EPSRC Simplicity Complexity and Modelling project EP/E018173



An apology

Although I have worked for many years on clinical trials, my only involvement with cancer has been occasional membership of data safety monitoring boards

Thus, my examples are not taken from trials in cancer

The relevance (or not) to cancer of anything I have to say I leave for others to judge

Outline

- Examples where we have a considerable gain by increasing complexity
- Examples where we do better to be simple
- Examples where more complex designs and modelling can teach us to be simpler
- Recommendations and conclusions

General thesis:
as complex as necessary but no more

For complexity

- Complex allocation needs complex analysis
- Baseline covariates carry useful information
- Some apparently simple transformations mislead and destroy information and should be avoided

For simplicity

- Some complex models have unfortunate side-effects
- Overfitting can reduce capacity to predict
- Complex models can hide dangerous implicit assumptions

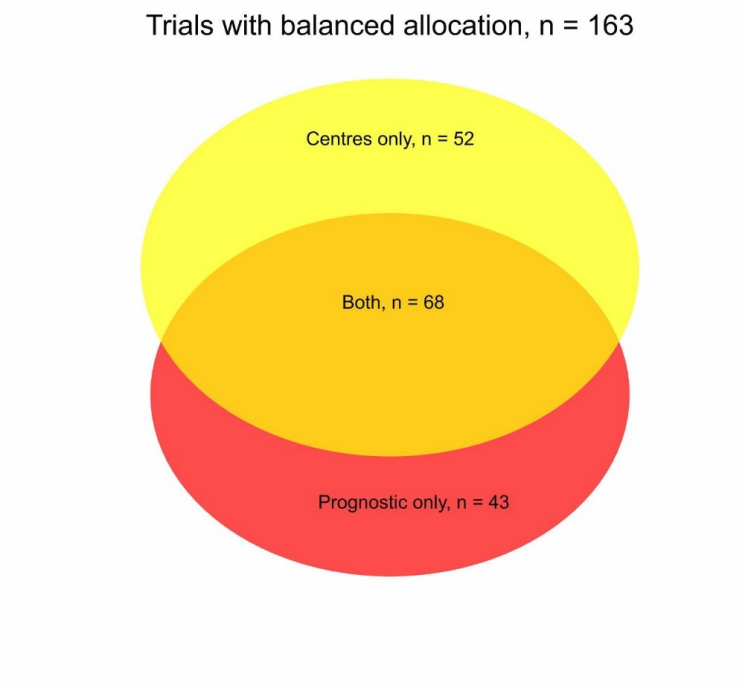
For complexity

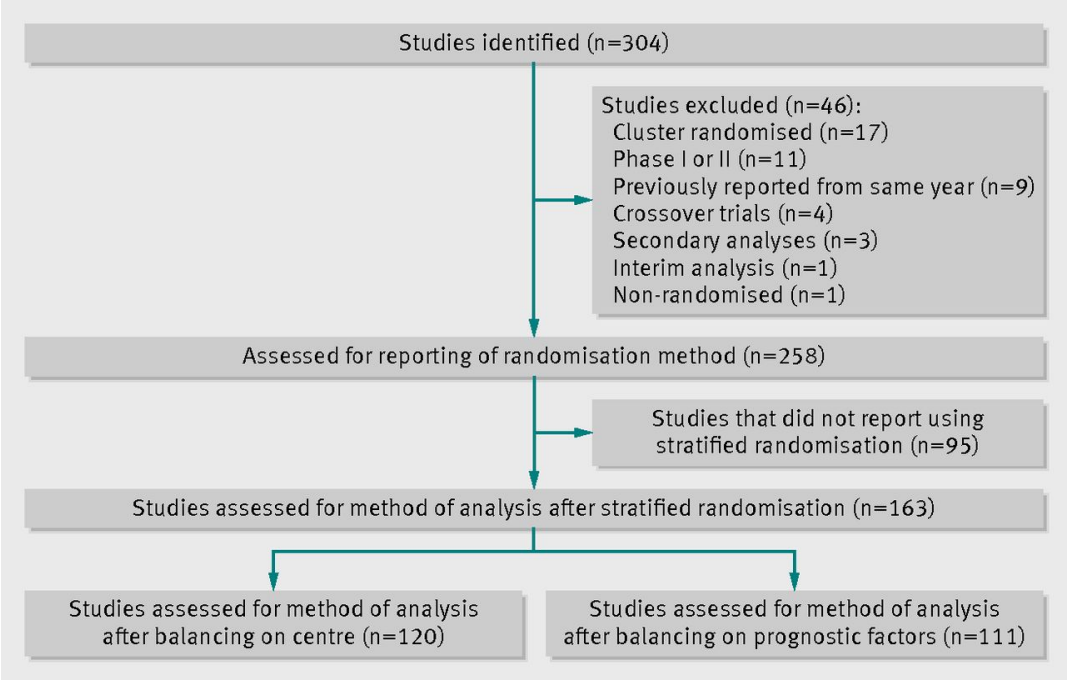
Where we lose information by being too simple

Failure to allow for design in analysis

Survey by Kahan and Morris 2012

Survey of trials published in *BMJ, JAMA, NEJM* & *Lancet* in 2010





© Kahan and Morris, BMJ, 2012

Findings as regard analysis

Balancing by centre

Strategy	Number	Percent
Adjusted in primary analysis	31	26%
Adjusted in secondary analysis	4	3%
Did not adjust	68	57%
Unclear	17	14%
Total	120	100%

Did not adjust + unclear = 71%

Balancing by prognostic factors

Strategy	Number	Percent
Adjusted for all in primary analysis	40	36%
Adjusted for some in primary analysis	4	4%
Adjusted in secondary analysis	10	9%
Did not adjust	45	41%
Unclear	12	11%
Total	111	100%

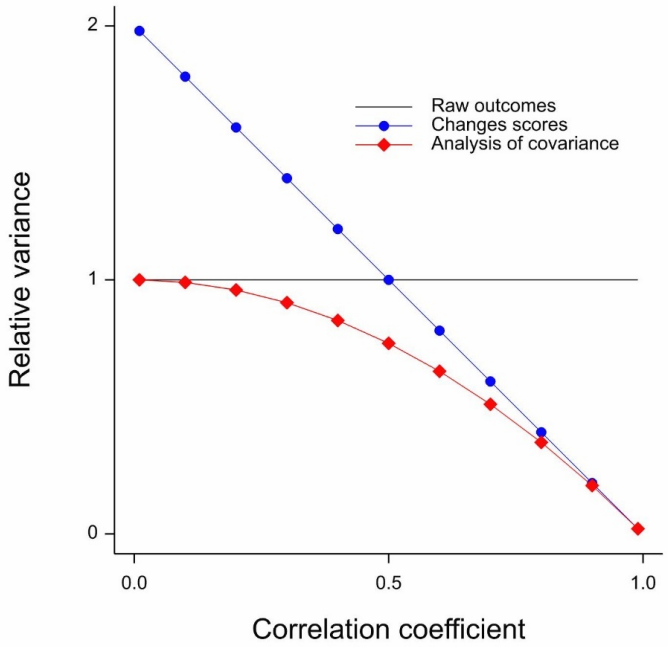
Did not adjust + unclear = 52%

In summary

- K & M Found fewer than 50% of clinical trials (in leading journals) that balanced by centre or prognostic factor declared that the main analysis took account of this
- An opportunity for increasing efficiency is being missed
- For linear models, the standard errors will be larger than they should be
- For non-linear models, effective treatments will have estimates biased towards the null
- This is something of a scandal

Change from baseline
Waste in the name of simplicity

- A common habit for ‘true’ baselines is to use them to construct a change-score by simple subtraction
- Assuming equal variances at baseline and outcome this increases the variance unless the correlation is greater than 0.5
- Analysis of covariance (invented 1931) is (asymptotically) better than either raw scores or change scores

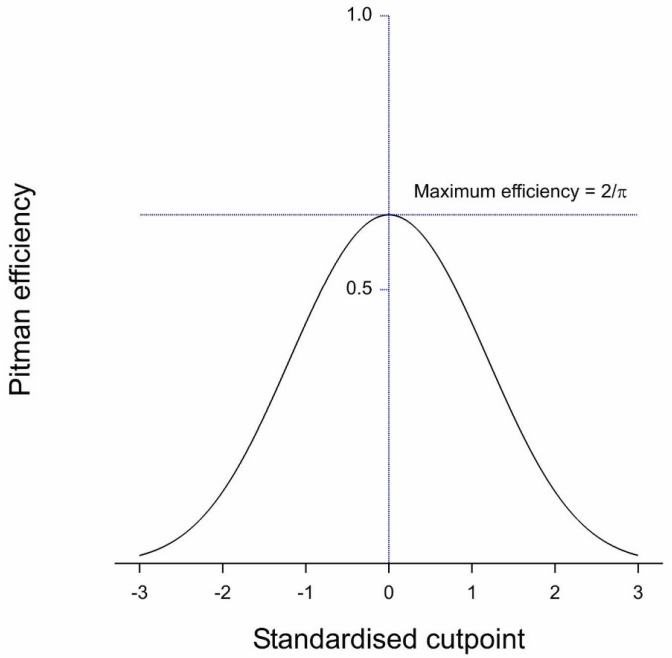


Responder dichotomies
Criminal waste in the name of simplicity

- Responder dichotomies compound the change score crime by replacing them by a binary score
- If the cut-point is at the median, the sample size must be increased by

$$100\times(\pi/2-1)\approx 57\%$$

- For other cut-points it is worse
- Encourages false belief in causal differences between ‘responders’ and non-responders
  - To be picked up later



For simplicity

Where we lose information by being too complex

Repeated measures analysis

- Much repeated measures analysis using mixed models could be simply replaced by a summary measures approach
- This will often be nearly as efficient
  - In fact for simple correlation structures and complete data will be fully efficient
- Helps understanding
- Furthermore some repeated measures approach implicitly violate intention to treat. See
  - Senn, Stevens and Chaturvedi, 2000
  - Bamia, White, Kenward, 2013

How analysis of repeated measures can violate ITT

What everybody agrees is unacceptable

- It is common advice that you should not correct for post randomisation covariates
- For example, measures of the form Y_3-Y_1 instead of Y_3-Y_0

What everybody assumes is fine

- Ordinary least squares estimates of slopes
- But suppose that you have three post randomisation measurements at equal intervals
- The ordinary least squares measure of the slope is

$$(Y_3-Y_1)/2$$

An example of the wrong sort of complexity: Thall and Vail, 1990

- Repeated seizures for 59 patients over 4 two week periods
  - Compares two treatments Placebo and progabide
- Has been cited 461 times by 2017 according to Google Scholar
- Cites Leppik et al (1985) for the data
- Seven different covariance models proposed by Thall & Vail
- Very many different models proposed since

…, the predicted mean seizure rate for the progabide group is either higher or lower than that for the placebo group, accordingly as the baseline count does or does not exceed a critical threshold…. This suggests that progabide may be contraindicated for patients with high seizure rates.

Thall and Vail p666

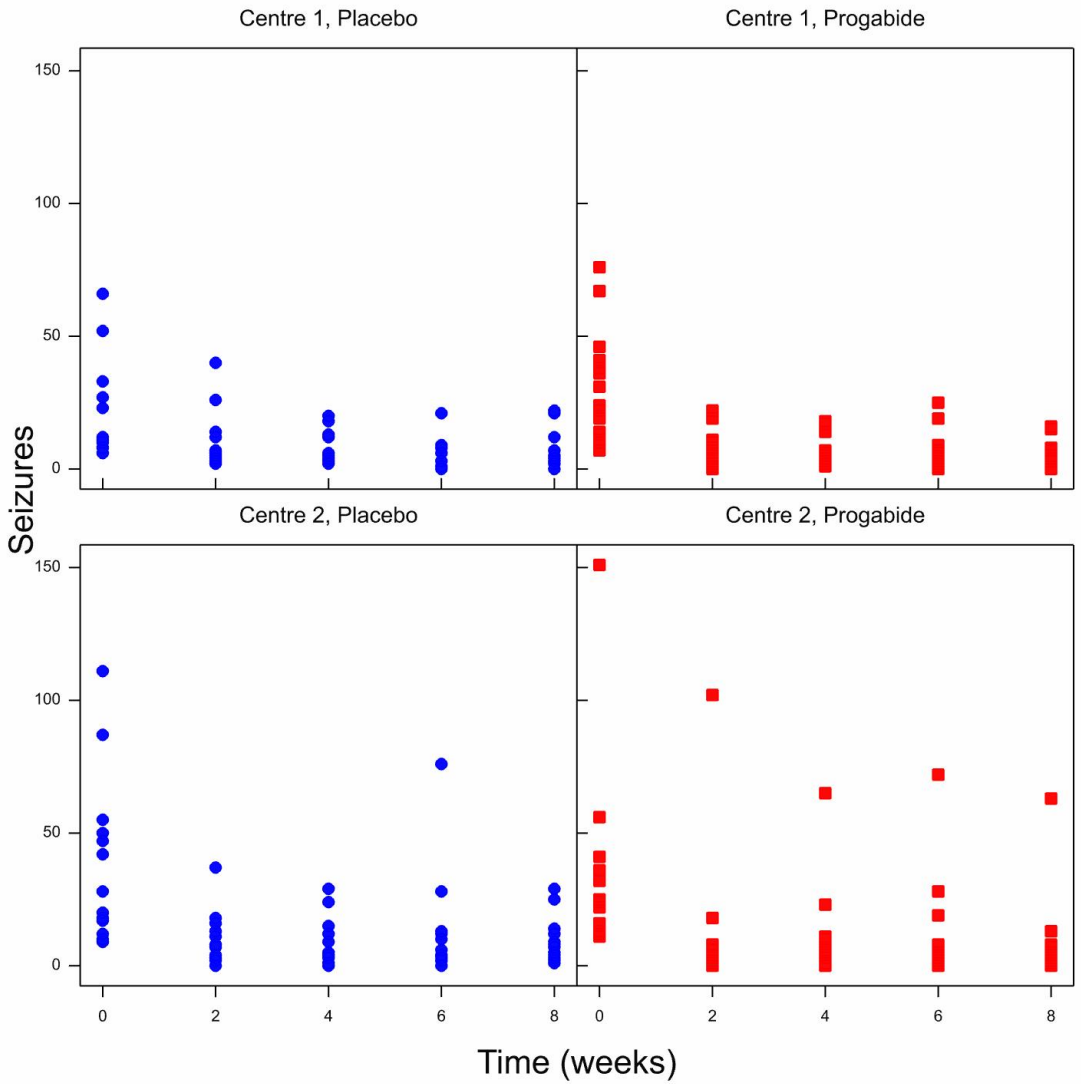
But data and what they mean are important

- The data are not from Leppik et al 1985 but 1987
- It should be obvious from the patient numbers that it is a two centre trial, but nobody appears to notice this
- The division into four two-week periods has *no clinical meaning whatsoever*
- Modelling this as a correlated series is pointless and possible misleading

Table 2

Successive two-week seizure counts for 59 epileptics. Covariates are adjuvant treatment (0 = placebo, 1 = progabide), eight-week baseline seizure counts, and age (in years).

ID	Y_1	Y_2	Y_3	Y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
.	.	.	.	.	.	.	.



In my view the analysis of data as four visits is pointless and we might as well analyse the totals. Leppik *et al.* (1987), using all the data of the original crossover trial, found no convincing evidence of a treatment effect and I am suspicious of any analyses of the first-period data only, including those of Lee and Nelder and Thall and Vail (1990), that do. Fitting total seizures as a function of centre, age and base-line seizure in addition to treatment using either Poisson regression and allowing for overdispersion or a negative binomial model, or using the square root of the number of seizures in a linear model, I find no convincing evidence of a treatment effect.

Increasing the number of covariates in a linear model

- Adding predictive covariates to a model makes the residual error smaller
- But it makes the design matrix somewhat less well-conditioned
- Second order efficiency is also affected
    - Fewer degrees of freedom for estimating the error variance
    - Less favourable t-distribution for confidence intervals
- Eventually as we add covariates we lose
- Problem in small trials

Complexity that may yield simplicity

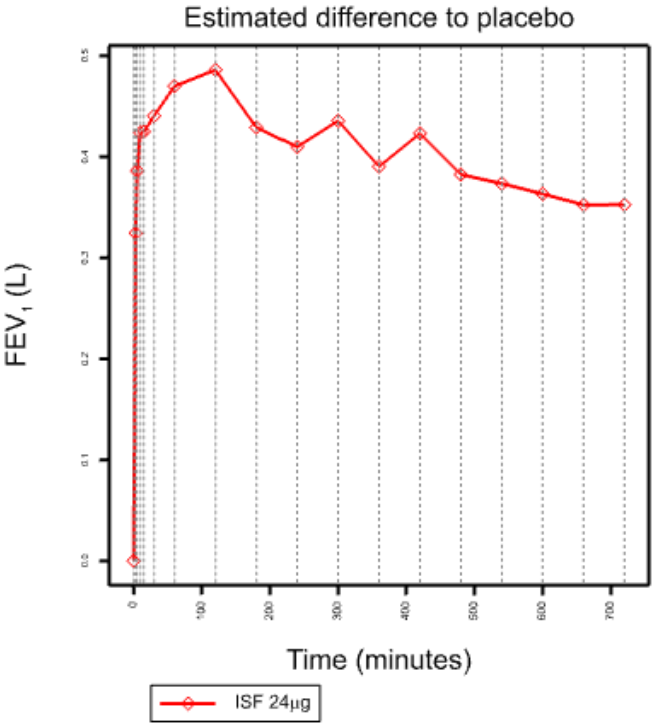
How being complex can sometimes yield insights that lead to simplicity

A complex design in asthma

		Formulation of Formoterol		
		ISF	**MTA**	**Nothing**
Dose	**0 µg**			Placebo
	6 µg	ISF6	MTA6	
	12 µg	1SF12	MTA12	
	24 µg	ISF24	MTA24	

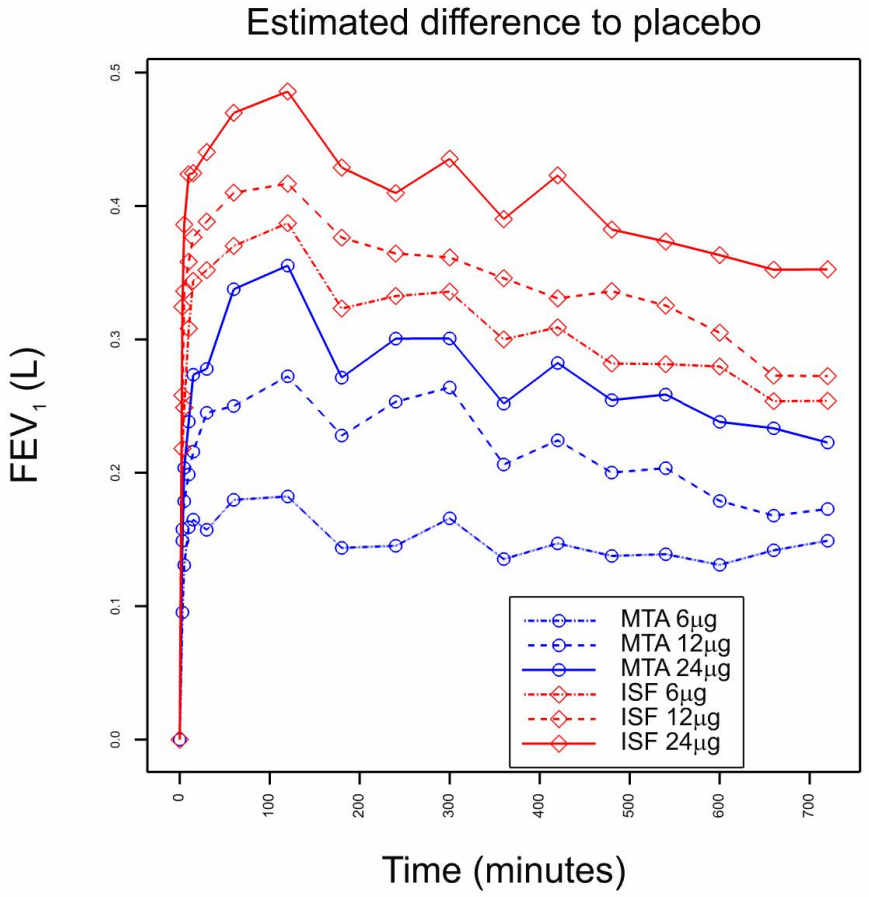
Senn, S. J., Lillienthal, J., Patalano, F., & Till, M. D. (1997). An incomplete blocks cross-over in asthma: a case study in collaboration. In J. Vollmar & L. A. Hothorn (Eds.), *Cross-over Clinical Trials (pp. 3-26). Stuttgart: Fischer.*

- Parallel assay
- Cross-over
- Incomplete blocks
- Seven treatments
- Five periods
- Twenty-one sequences
- Forced expiratory volume in one second (FEV_1)
- 18 time-points over 12 hours
- Log AUC of FEV_1 as main outcome



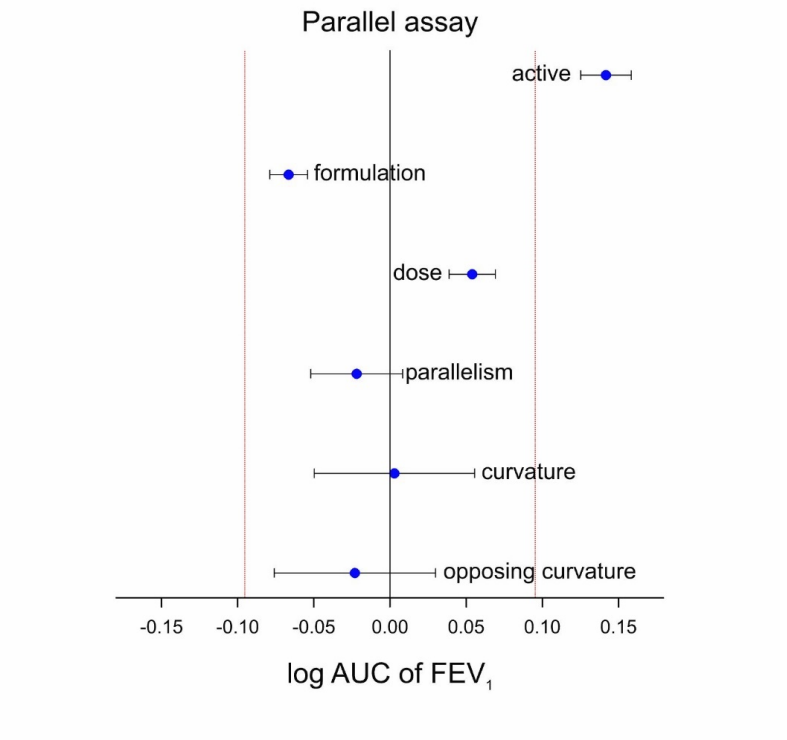
Results

- Perfect dose response 6μg, 12μg, 24μg within each formulation
- Big surprise is complete separation of formulations
- Formulations not at all equivalent
- MTA 24μg appears to be less potent than ISF 6μg



Analysis of Contrasts

	ISF (Reference)			MTA (Test)			Pla-cebo
Contrast	**6**	**12**	**24**	**6**	**12**	**24**	
Active	1/6	1/6	1/6	1/6	1/6	1/6	-1
Formulation	-1/3	-1/3	-1/3	1/3	1/3	1/3	0
Dose	-1/2	0	1/2	-1/2	0	1/2	0
Parallelism	-1	0	1	1	0	-1	0
Curvature	-1	2	-1	-1	2	-1	0
Opposing curvature	-1	2	-1	1	-2	1	0



Implications

As regards comparing formulations

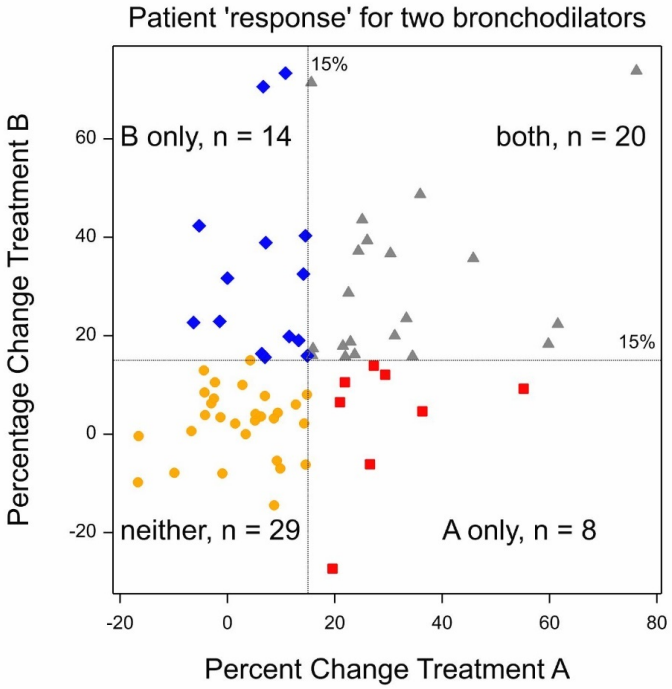
- The formulations are clearly not equipotent
- The difference between formulations is a great as the difference between doses
- But the model adequacy contrasts are all non-significant
- Linear (in the log dose) appears to work
- A careful complicated design killed the new formulation

But there is more

- The fact that patients have been measured many times enables us to say something about individual response
- Consider a common (very stupid) definition of response
  - 15% increase in FEV_1 above baseline
- Now look at 'responders' 12 hours after treatment for two of the formulations…

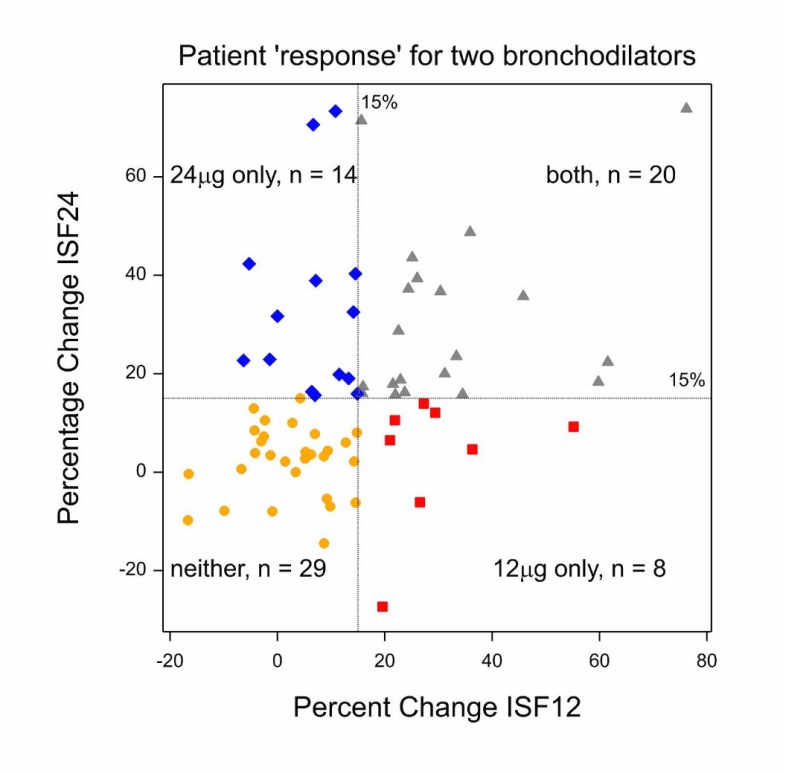
The case for personalised medicine

- There seem to be a number of patients who respond to B and not to A and vice versa
- Clearly if we can find predictive characteristics of them we can improve treatment
- Next stop, precision medicine



The case against personalised medicine

- A is ISF 12μg, the second most potent of the six formulations and doses tested
- B is ISF 24μg the most potent of the six formulations and doses tested
- It is biologically extremely implausible that patients could respond to 12μg and not to 24μg
- Yet apparently 8 out of 71 patients did
- Conclusion: naïve simple views of causality and response aren't good enough and more complex design and analysis is needed

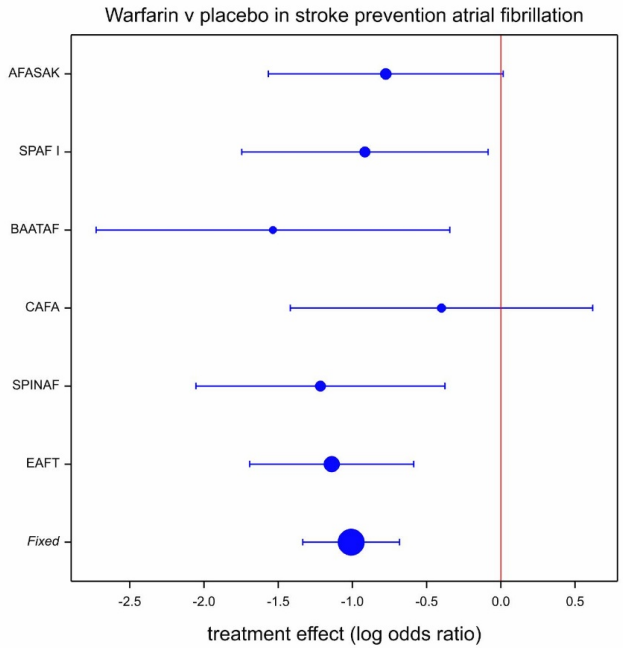


Responder analysis is the work of the devil

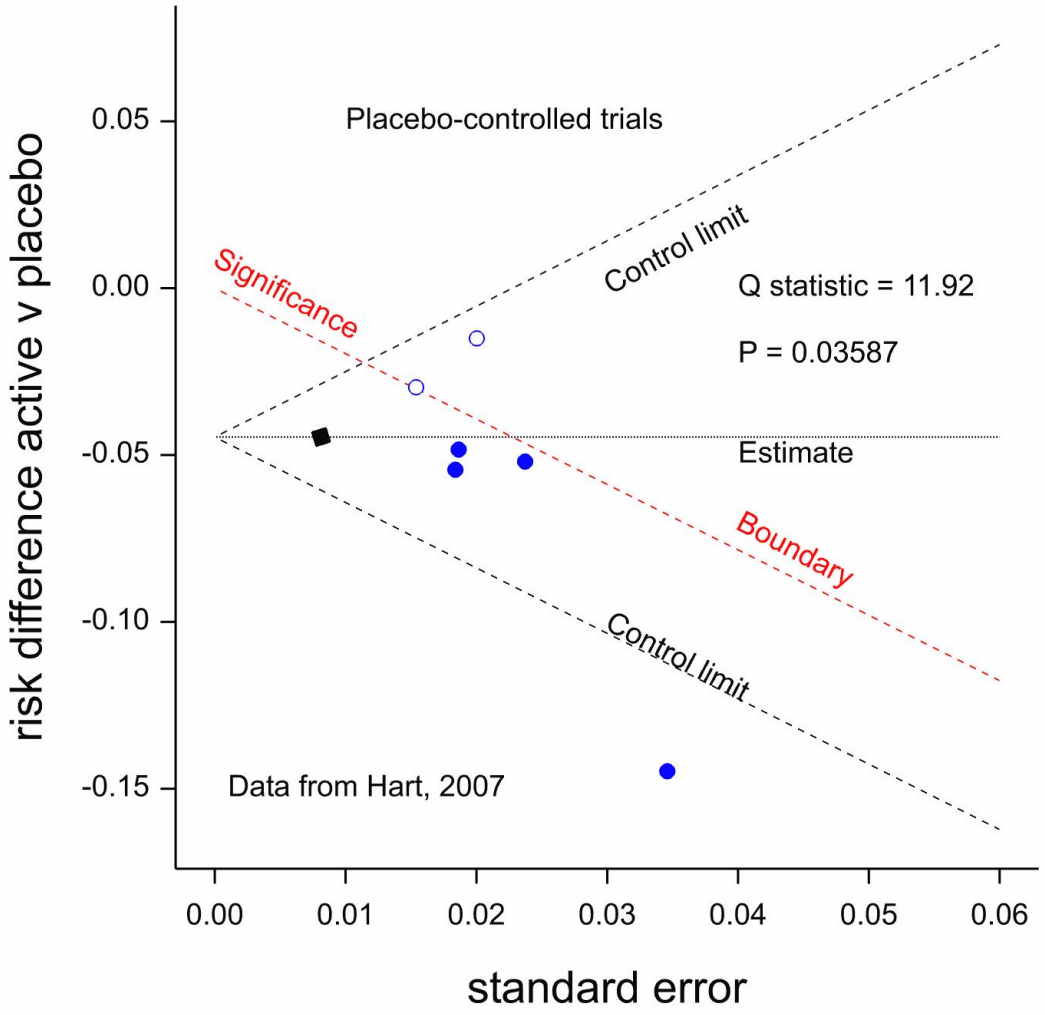
Any statistician who collaborates in this crime deserves to languish in data-analysis hell

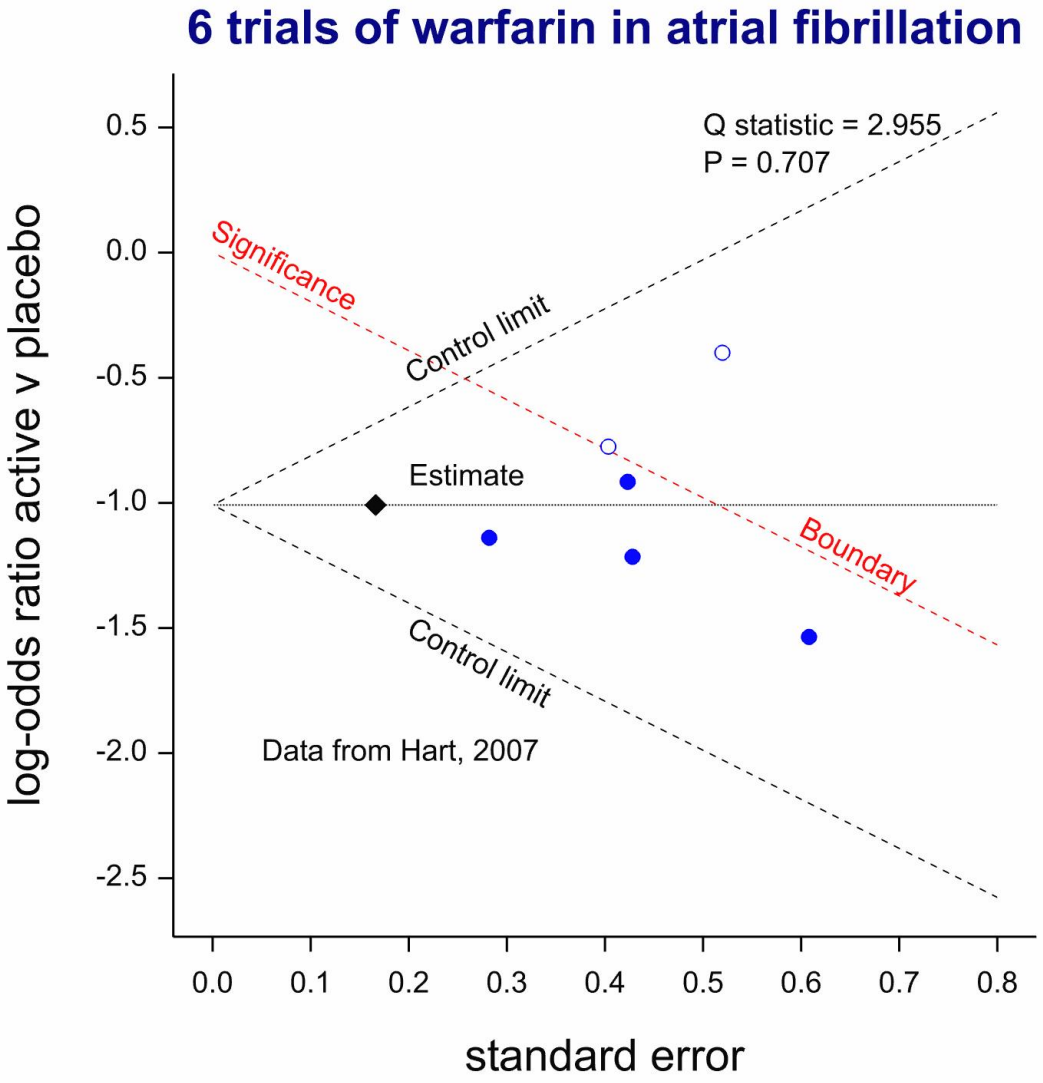
Example of Atrial Fibrillation

- Such patients are at higher risk of stroke
- Meta-analysis (reproduced in Hart et al 2007)concluded that warfarin has a beneficial protective effect
- But there is a risk of intracranial bleeding
- Who should get warfarin?

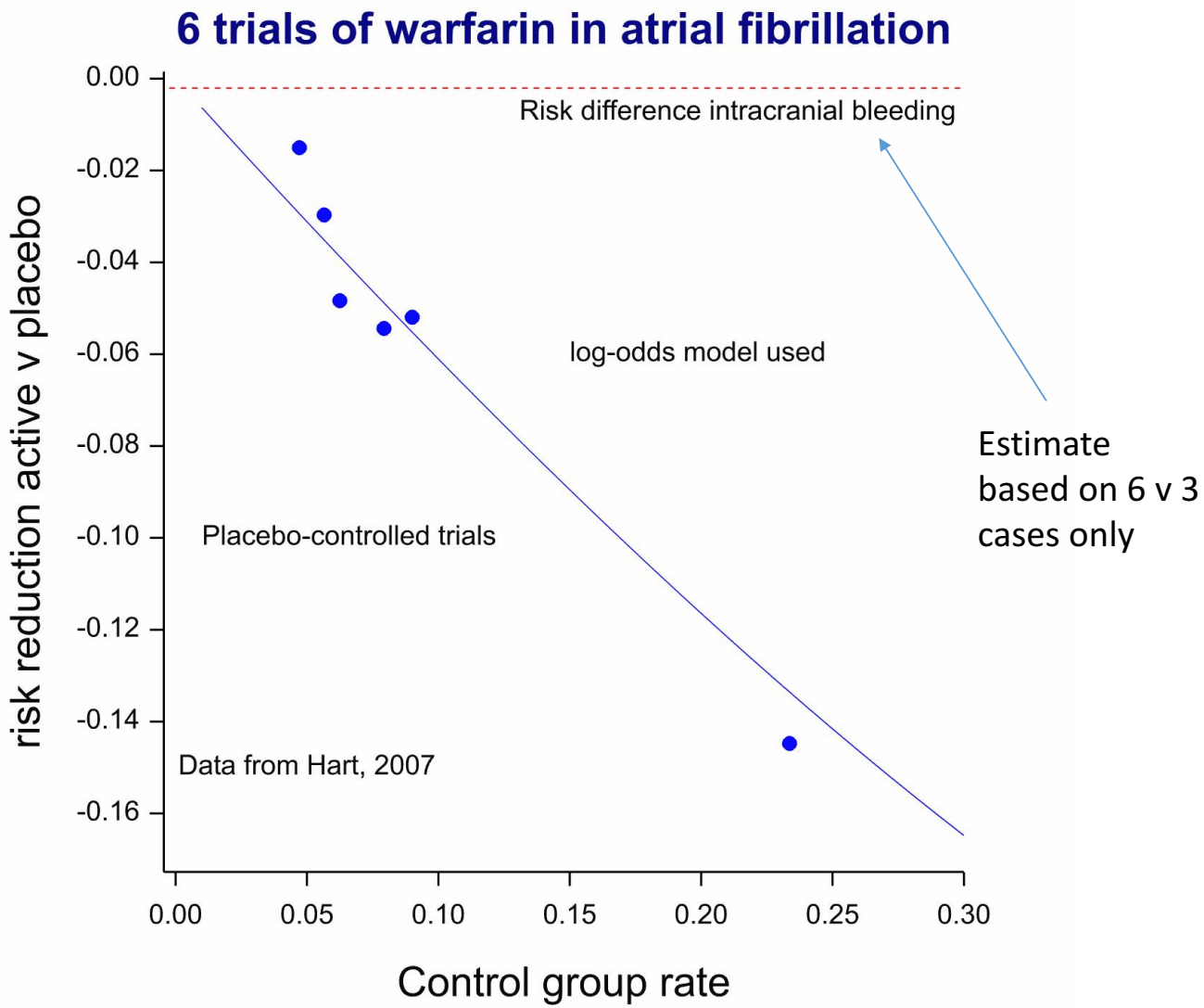


6 trials of warfarin in atrial fibrillation




6 trials of warfarin in atrial fibrillation
Q statistic = 2.955
P = 0.707
Significance
Control limit
Estimate
Boundary
Control limit
Data from Hart, 2007
log-odds ratio active v placebo
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
standard error

The line gives the prediction if the common log-odds ratio estimate is applied to the control group rate



Recommendations and conclusions

Reminding ourselves why we do this

Conclusions

- The appropriate degree of complexity is a matter of judgement
- The key to getting the right degree is maintaining a sense of purpose
  - Does the complexity reflect pharmacology etc to the degree needed?
  - Have we followed through: analysis that reflects design and design that serves the analysis needed?
  - Are we doing it to increase our understanding of the effects of treatment?
  - Are we being complex in the right places
    - Elaborate models of responder dichotomies are pointless
  - Measurement matters

Recommendation

Always ask yourself this:

Am I really interested in finding out about the effects of treatment?