What Randomisation Can and Cannot do for You

Stephen Senn

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Acknowledgements

Thank you for the kind invitation

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"

THE ANALYSIS OF GROUPS OF EXPERIMENTS

By F. YATES and W. G. COCHRAN

Statistical Department, Rothamsted Experimental Station, Harpenden

Lack of randomness is then only harmful in so far as it results in the omission of sites of certain types and in the consequent arbitrary restriction of the range of conditions. In this respect scientific research is easier than technical research.

P558

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. Journal of Agricultural Science, 28(4), 556-580

Outline

- A game of chance to explain randomisation
- The value of randomisation (scientific research)
  - A false criticism
  - The importance of ratios
  - An example: the TARGET study
- Limitations of randomisation and how to deal with them (technical research)
  - Additive scales
  - Prediction
- Conclusions

Game of Chance

- Two dice are rolled
  - Red die
  - Black die
- You have to call correctly the odds of a total score of 10
- Three variants
  - Game 1 You call the odds and the dice are rolled together
  - Game 2 the red die is rolled first, you are shown the score and then must call the odds
  - Game 3 the red die is rolled first, you are not shown the score and then must call the odds

Total Score when Rolling Two Dice

Black Die Score	Red Die Score					
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Variant 1. Three of 36 equally likely results give a 10. The probability is 3/36=1/12.

Total Score when Rolling Two Dice

Black Die Score	Red Die Score: 1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Variant 2: If the red die score is 1,2 or 3, probability of a total of10 is 0. If the red die score is 4,5 or 6 the probability of a total of10 is 1/6.

Variant 3: The probability = (½ x 0) + (½ x 1/6) = 1/12

The Morals

- You can't treat game 2 like game 1.
  - You must condition on the information you receive in order to act wisely
  - You must use the actual data from the red die
- You can treat game 3 like game 1.
  - You can use the *distribution in probability* that the red die has
- You can't ignore an observed prognostic covariate in analysing a clinical trial just because you randomised
  - That would be to treat game 2 like game 1
- You can ignore an unobserved covariate precisely because you did randomise
  - Because you are entitled to treat game 3 like game 1

A Red Herring

"Even if there is only a small probability that an individual factor is unbalanced, given that there are indefinitely many possible confounding factors, then it would seem to follow that the probability that there is some factor on which the two groups are unbalanced (when remember randomly constructed) might for all we know be high. Prima facie those frequentist statisticians who argue that randomization "tends" to balance the groups in all factors commit a simple quantificational fallacy." John Worrall 2002

- One sometimes hears that the fact that there are indefinitely many covariates means that randomisation is useless
- This is quite wrong
- It is based on a misunderstanding that variant 3 of our game should ***not*** be analysed like variant 1
- I showed you that it ***should***

You are not free to imagine anything at all

- Imagine that you are in control of all the thousands and thousands of covariates that patients will have
- You are now going to allocate the covariates and their effects to patients
  - As in a **simulation**
- If you respect the actual variation in human health that there can be, you will find that the net total effect of these covariates is bounded

$$Y = \beta_0 + \tau Z + \beta_1 X_1 + \cdots \beta_k X_k + \cdots$$

Where *Z* (which is equal to either 0 or 1) is a treatment indicator, τ is the treatment effect, and the *Xs* are covariates. You are not free to arbitrarily assume any values you like for the *Xs* and the βs because the variance of *Y* must be respected.

What happens if you don’t pay attention

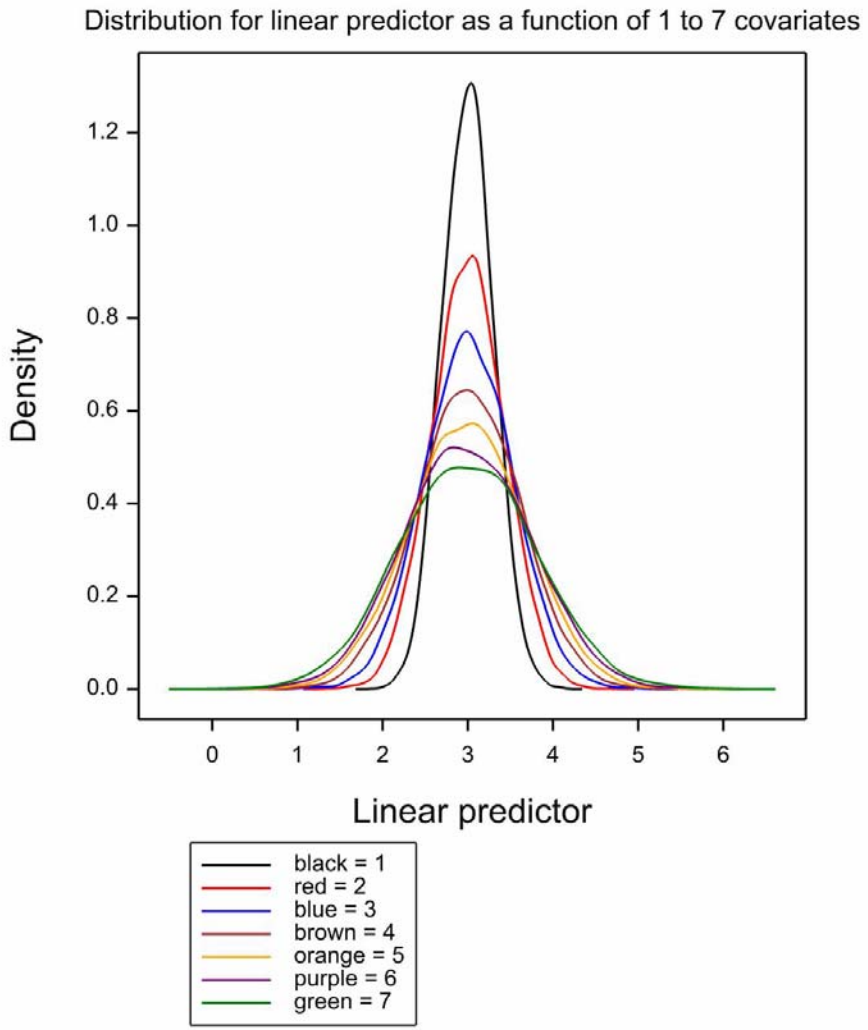
Simulation of the linear predictor as the number of covariates increases from 1 to 7

However, the variance of each covariate is the same and the coefficient is the same and the covariates are assumed orthogonal

We can see that the variance of the predictor keeps on increasing

The values soon become impossible

But in reality the total contribution that the covariates can make is bounded



In fact this is pointless

Look at the equation again

$$Y = \beta_0 + \tau Z + \beta_1 X_1 + \cdots \beta_k X_k + \cdots$$

We have to take care how we choose the parameters of the $X_1, .. X_k \; and \; \beta_1 \ldots \beta_k$ and what we have to guide us are the possible values of *Y*. But suppose we re-write the equation

$$Y = Y^* + \tau Z$$

Where

$$Y^* = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \cdots$$

Now there is only one unknown, Y^* not indefinitely many, and this is all that we need to consider

So Worrall's argument is wrong

Worrall's argument boils down to saying that if a series is infinite its sum can't be bounded.

But how about the sum

$$S = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} \ldots \ ?$$

The importance of ratios

- So from one point of view **there is only one covariate** that matters
  - potential outcome
    - If you know this, all other covariates are irrelevant
- And just as this can vary between groups in can vary within
- The t-statistic is based on the ***ratio*** of differences *between* to variation *within*
- Randomisation guarantees (to a good approximation) the unconditional behaviour of this ratio and that is all that matters for what you can't see (game 3)
- An example follows

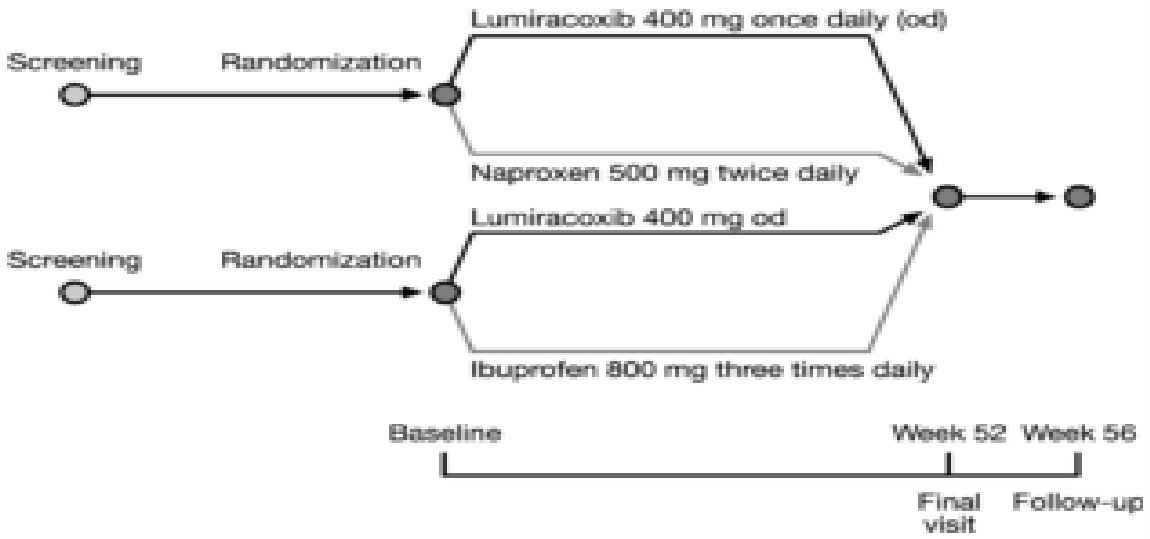


Figure 1. Therapeutic Arthritis Research and Gastrointestinal Event Trial – study design.

Better non-CONSORT diagram in the design paper: Hawkey et al
Aliment Pharmacol Ther 2004; 20: 51–63

.

Why this complicated plan?

- The treatments have different schedules
  - Lumiracoxib once daily
  - Naproxen twice daily
  - Ibuprofen 3 times daily
- To blind this effectively would require very complicated double dummy loading schemes
- So centres were recruited into
  - either lumiracoxib versus naproxen
  - or lumiracoxib versus ibuprofen

Baseline Demographics

	Sub-Study 1		Sub Study 2	
Demographic Characteristic	**Lumiracoxib n = 4376**	**Ibuprofen n = 4397**	**Lumiracoxib n = 4741**	**Naproxen n = 4730**
Use of low-dose aspirin	975 (22.3)	966 (22.0)	1195 (25.1)	1193 (25.2)
History of vascular disease	393 (9.0)	340 (7.7)	588 (12.4)	559 (11.8)
Cerebro-vascular disease	69 (1.6)	65 (1.5)	108 (2.3)	107 (2.3)
Dyslipidaemias	1030 (23.5)	1025 (23.3)	799 (16.9)	809 (17.1)
Nitrate use	105 (2.4)	79 (1.8)	181 (3.8)	165 (3.5)

Formal statistical analysis of baseline comparability

- Usually I do not recommend doing this
- If we have randomised we know that differences must be random
  - Testing could be used to examine cheating
- However here there was randomisation within sub-studies and *not* between
- It thus becomes interesting to see if the tests can detect the difference between the two

Baseline Chi-square P-values

	Model Term		
Demographic Characteristic	**Sub-study (DF=1)**	**Treatment given Sub-study (DF=2)**	**Treatment (DF=2)**
Use of low-dose aspirin	< 0.0001	0.94	0.0012
History of vascular disease	< 0.0001	0.07	<0.0001
Cerebro-vascular disease	0.0002	0.93	0.0208
Dyslipidaemias	<0.0001	0.92	<0.0001
Nitrate use	< 0.0001	0.10	<0.0001

Outcome Variables

All four groups

	Sub-Study 1		Sub Study 2	
Outcome Variables	**Lumiracoxib n = 4376**	**Ibuprofen n = 4397**	**Lumiracoxib n = 4741**	**Naproxen n = 4730**
Total of discontinuations	1751 (40.01)	1941 (44.14)	1719 (36.26)	1790 (37.84)
CV events	33 (0.75)	32 (0.73)	52 (1.10)	43 (0.91)
At least one AE	699 (15.97)	789 (17.94)	710 (14.98)	846 (17.89)
Any GI	1855 (42.39)	1851 (42.10)	1785 (37.65)	1988 (21.87)
Dyspepsia	1230 (28.11)	1205 (27.41)	1037 (21.87)	1119 (23.66)

Outcome Variables

Lumiracoxib only

	Sub-Study 1	Sub Study 2
Outcome Variables	**Lumiracoxib n = 4376**	**Lumiracoxib n = 4741**
Total of discontinuations	1751 (40.01)	1719 (36.26)
CV events	33 (0.75)	52 (1.10)
At least one AE	699 (15.97)	710 (14.98)
Any GI	1855 (42.39)	1785 (37.65)
Dyspepsia	1230 (28.11)	1037 (21.87)

Deviances and P-Values

Lumiracoxib only fitting Sub-study

	Statistic	
Outcome Variables	**Deviance**	**P-Value**
Total of discontinuations	13.61	0.0002
CV events	2.92	0.09
At least one AE	1.73	0.19
Any GI	21.31	<0.0001
Dyspepsia	47.34	< 0.0001

How we already use modelling, data and additive scales

- Interspecies scaling
- Bioequivalence
  - log relative bioavailability is additive but difference in absolute bioavailability is not
- Dose proportionality
- Use of additive scales in phase III
  - Log hazard
  - Log-odds ratio

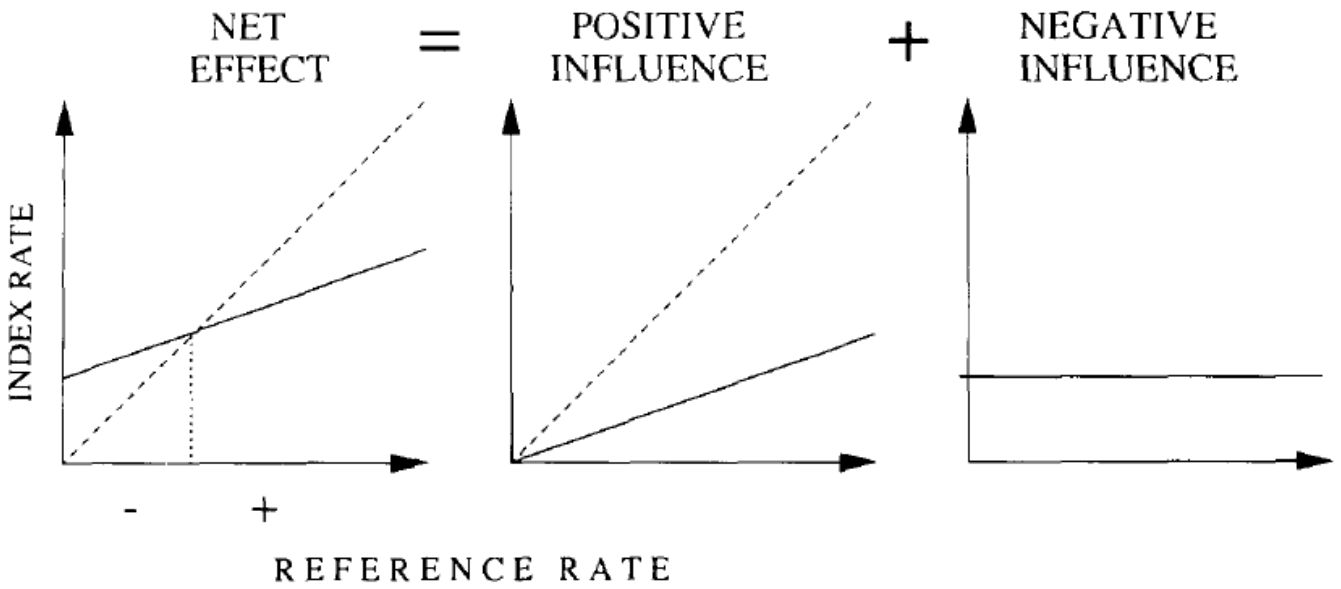


Figure 1 Direction of net effects. If a treatment does not affect the course of disease in any way, the index and reference rates of an outcome in an RCT will, when plotted on an *X*–*Y* graph across subgroups of patients at different levels of risk, fall (on the average) on the identity line (---). If there is a positive influence that reduces the reference rate with a constant proportion (middle panel) and a negative influence that induces a risk that is uniform over subgroups (right panel), their net effect will sum up as shown in the left panel.

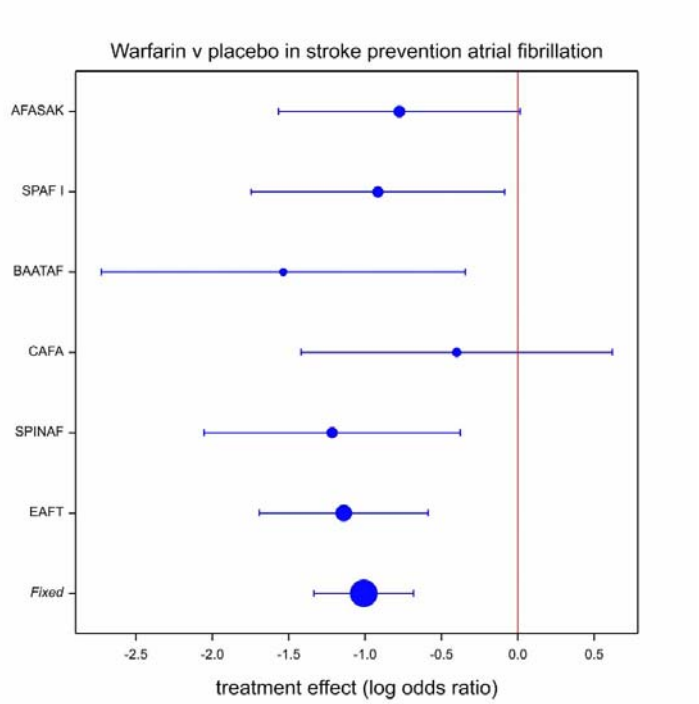
Controlled Clinical Trials, 1989

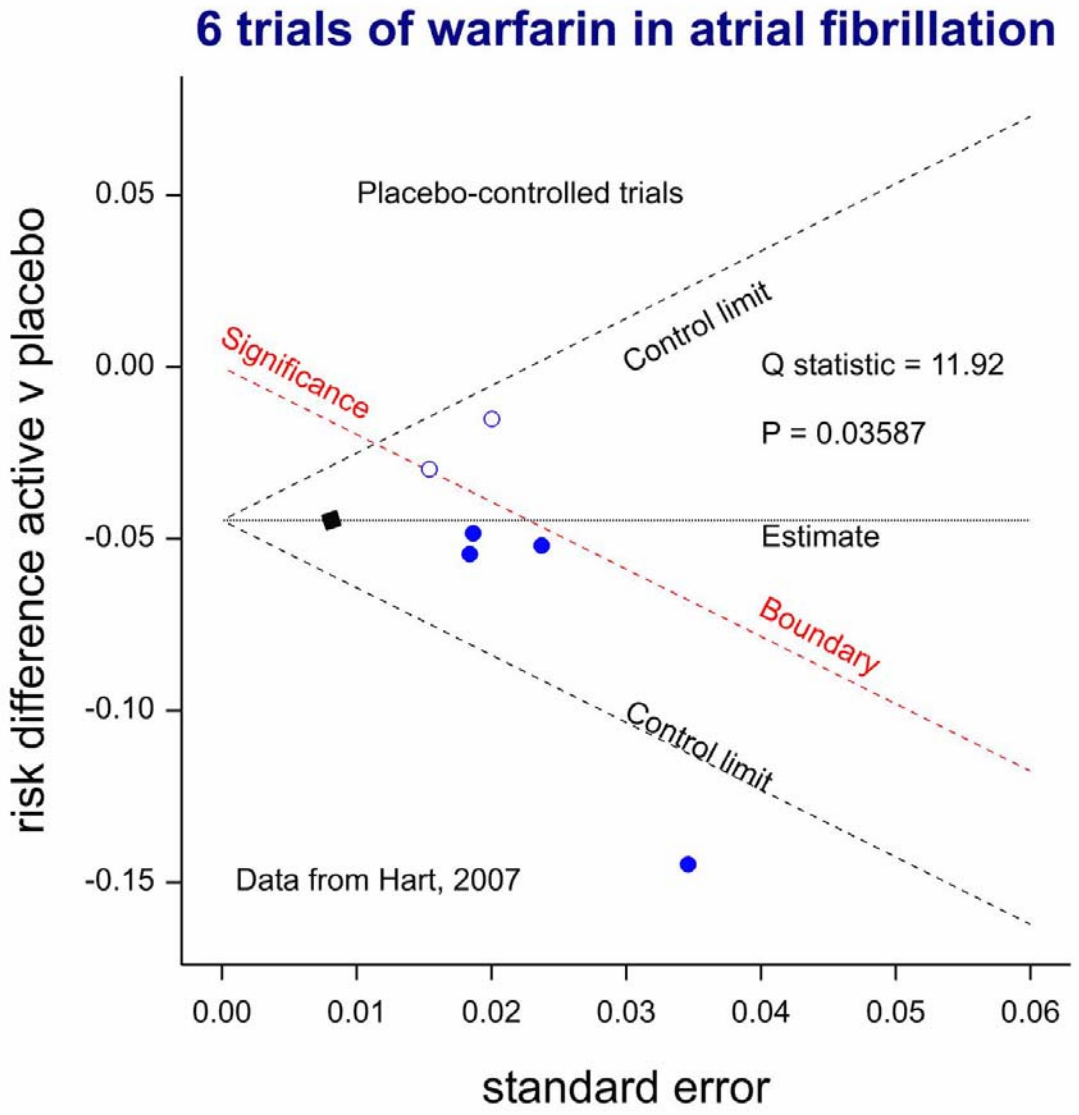
Implications of the Lubsen-Tijssen Model

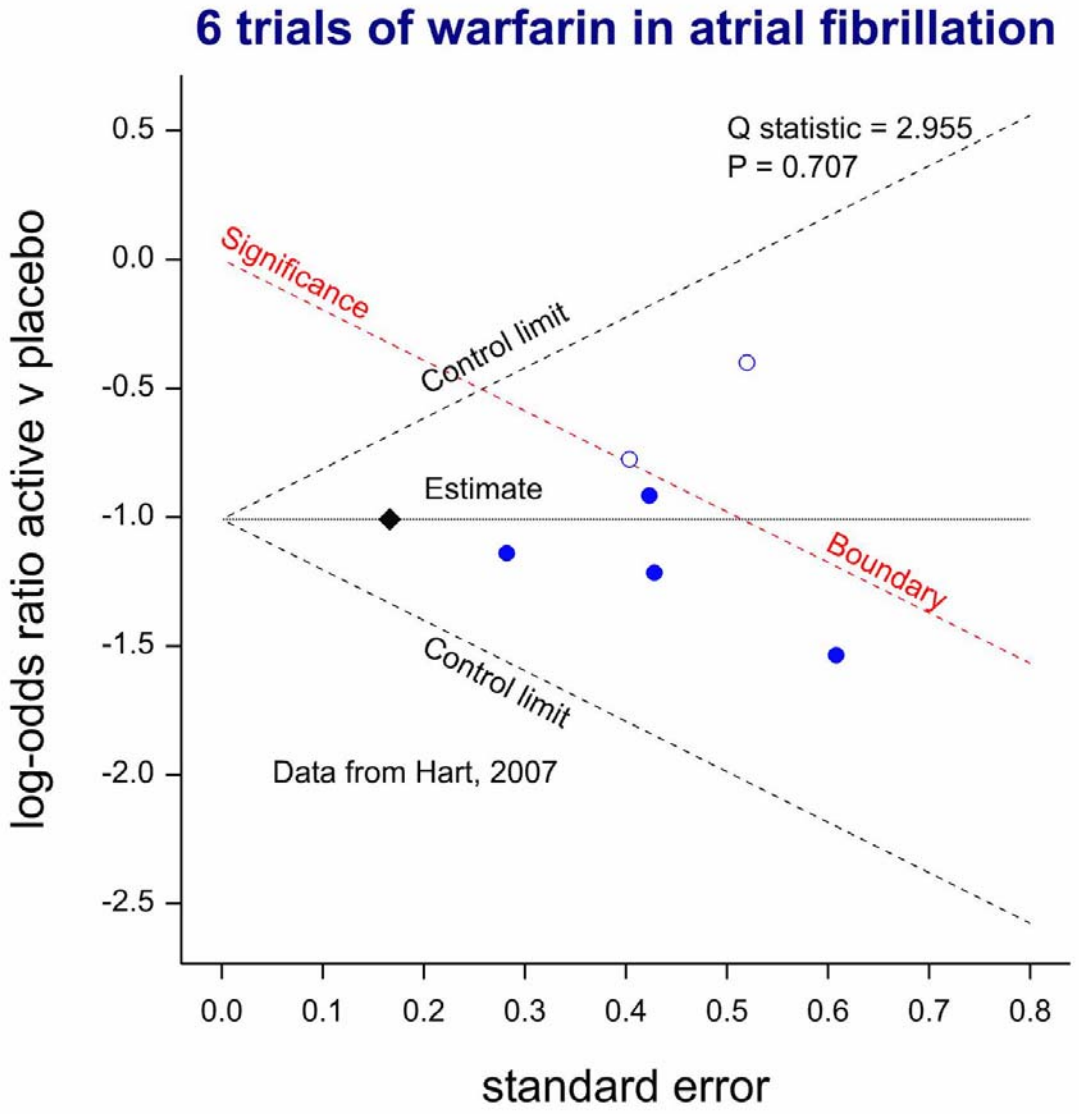
- We need to study treatment benefit on disaggregated (of harm) additive scale
- We will need real world data on harms
- We will need real world data on background risk
- We will need models
- We will need cooperation between
  - Medics and statisticians working on clinical trials
  - Statisticians, epidemiologists, health economists, medics and others working in real world data

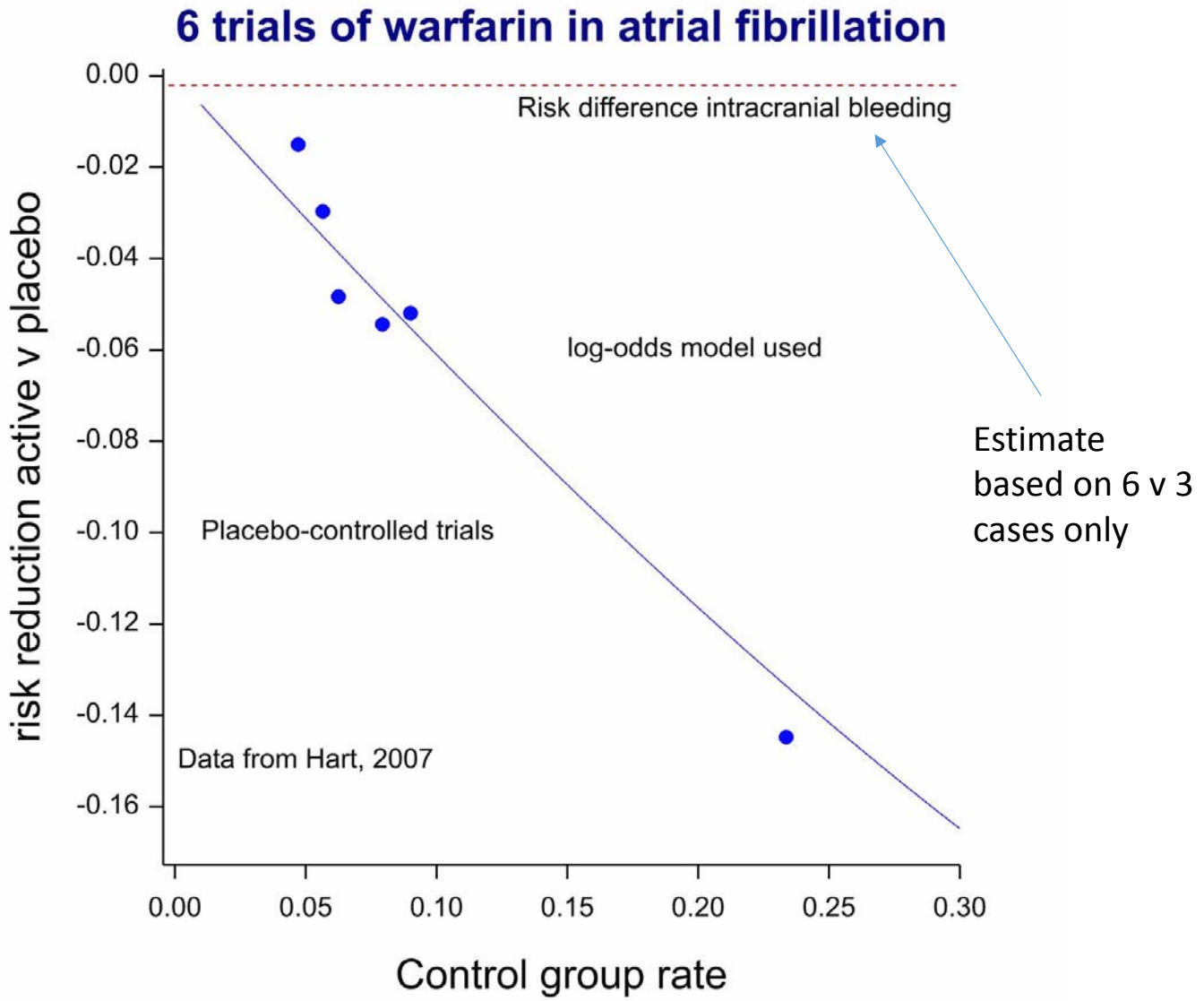
Example of Atrial Fibrillation

- 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
Placebo-controlled trials
Control limit
Significance
Q statistic = 11.92
P = 0.03587
Estimate
Boundary
Control limit
Data from Hart, 2007
risk difference active v placebo
0.05
0.00
-0.05
-0.10
-0.15
0.00
0.01
0.02
0.03
0.04
0.05
0.06
standard error


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
standard error
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



Estimate based on 6 v 3 cases only

Conclusion

- Randomisation is valuable
- Randomisation is not enough
  - Not all questions we need to ask can be answered using RCTs
  - Even when we can use RCTs we need to translate the results so they can be used in the clinic
- Modelling is important
  - Additive scales
  - Back transformation
- There is a ton of technical statistical theory on randomisation
  - If you want to talk about it, it's a good idea to be familiar with it