Randomisation isn’t perfect but doing better is harder than you think

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”

SEVENTH FRAMEWORK PROGRAMME

IDEAL

My thanks to BFF4 for further support and to Xiao-Ling for the invitation

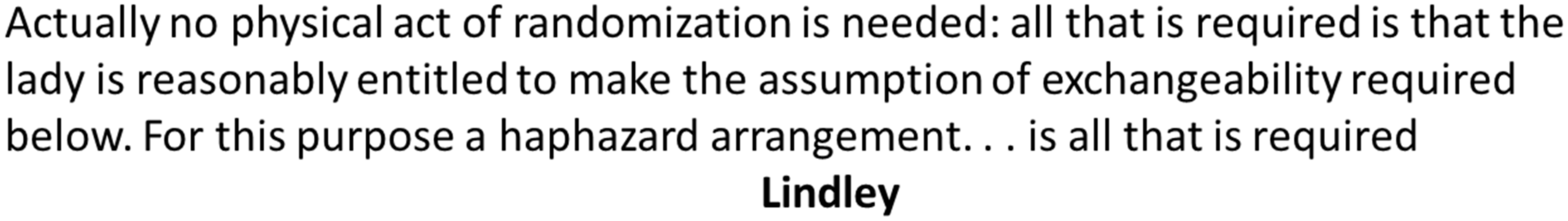
Outline

- Three criticisms
- A game of chance
- Answering the criticisms
  - – Lindley (briefly)
  - – Urbach (briefly)
  - – Worrall (in detail)
- A practical example
- My philosophy of randomisation and analysis

Lindley improves on Fisher?

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. The subject has been told in advance of what the test will consist, namely she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance

Fisher



Actually no physical act of randomization is needed: all that is required is that the lady is reasonably entitled to make the assumption of exchangeability required below. For this purpose a haphazard arrangement. . . is all that is required

Lindley

Urbach (1985) improves clinical trials?

1. . . . . For example, one could arrange for the matching to be performed by a panel of doctors representing a spectrum of opinion on the likely value of the drugs and whose criteria of selection have been made explicit. (p. 272)

2. …or one could simply permit the subjects to choose their own groups, always ensuring of course that they have not been informed of which treatment is to be applied to which group

Worrall’s influential criticism

“Even if there is only a small probability that an individual factor is un- balanced, 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 anyone knows be high.”

Worrall, 2002

A 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

		Red Die Score					
		1	2	3	4	5	6
Black Die Score	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

		Red Die Score					
		1	2	3	4	5	6
Black Die Score	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
- Whatever your philosophy, *it is still valuable to know that the game has been played fairly*

The Reality

Trialists continue to use their randomisation as an excuse for ignoring prognostic information and they continue to worry about the effect of factors they have not measured. Neither practice is logical.

Lindley

- What is ‘haphazard’?
  - Lindley is silent on the point
- Random is clearly defined by Fisher
  - Every one of the 70 sequences is equally probable
- Any departure from this requires modelling the correlation between subject and experimenter’s thought processes
- Good luck!

Urbach

Panel of doctors

- You treat patients when they fall ill
  - As anybody who has run clinical trials will know
- This scheme is almost always impossible
- But to the extent something is possible he is setting up a man of straw
- Blocking plus randomisation is (fairly) common

Patients choose themselves

- This has the variance of at least a randomised trial
- Even if we assume patients choose independently we still as Bayesians need a prior distribution on P(choose A)
- Eg $Beta(\alpha, \beta)$
- But best is $\alpha = \beta \to \infty$, which is what randomisation provides

The wisdom of Fisher

. . . if I want to test the capacity of the human race for telepathically perceiving a playing card, I might choose the Queen of Diamonds, and get thousands of radio listeners to send in guesses. I should then find that considerably more than one in 52 guessed the card right. Experimentally this sort of thing arises because we are in the
habit of making tacit hypotheses, e.g. ‘Good guesses are at random except for a possible telepathic influence.’

But in reality it appears that red cards are always guessed more frequently than black

Fisher writing to Harold Jeffrey (Bennett, 1990 p 268-269)

Summing up Worrall

*Randomization controls for all confounders…..*if there are many possible factors affecting the outcome it is actually very likely that some of them are unbalanced. Thus, in practice if it is noticed after randomization that the two groups are unbalanced with respect to a variable that is thought to affect the outcome, then the groups are re-randomized or adjusted (Worrall 2002)

Reiss and Ankeny, 2016, *Stanford Encyclopedia of Philosophy*

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

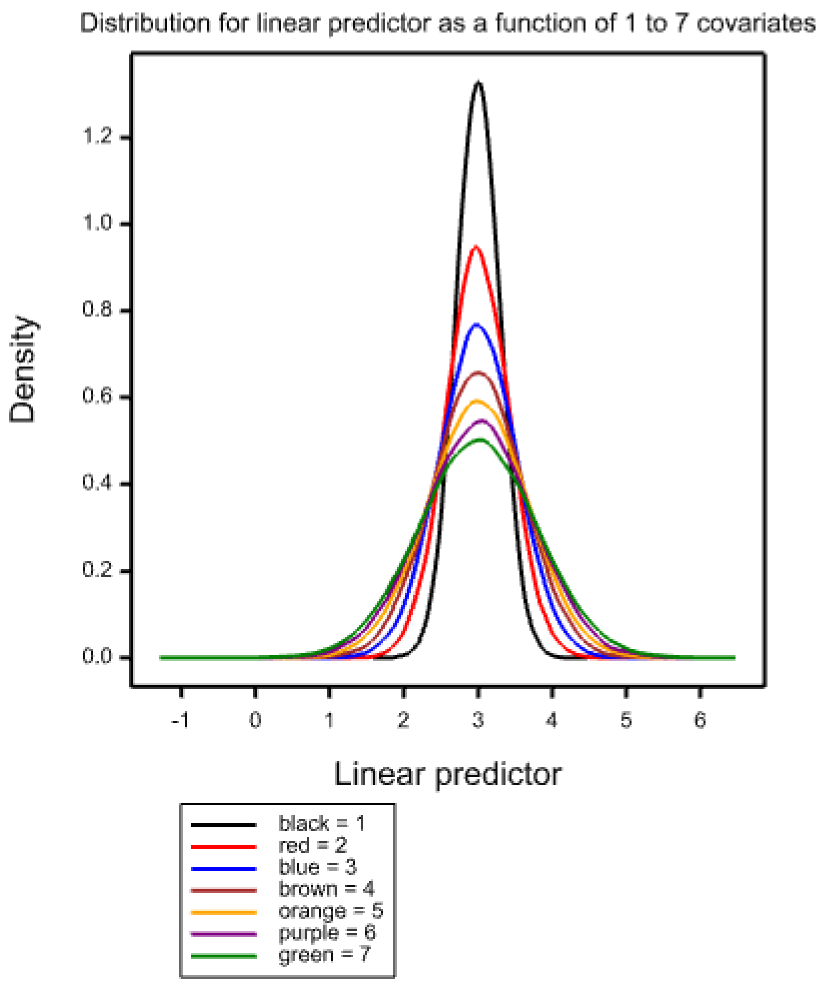
Simulation of the linear predictor as the number of covariates increases from 1 to 7

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

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

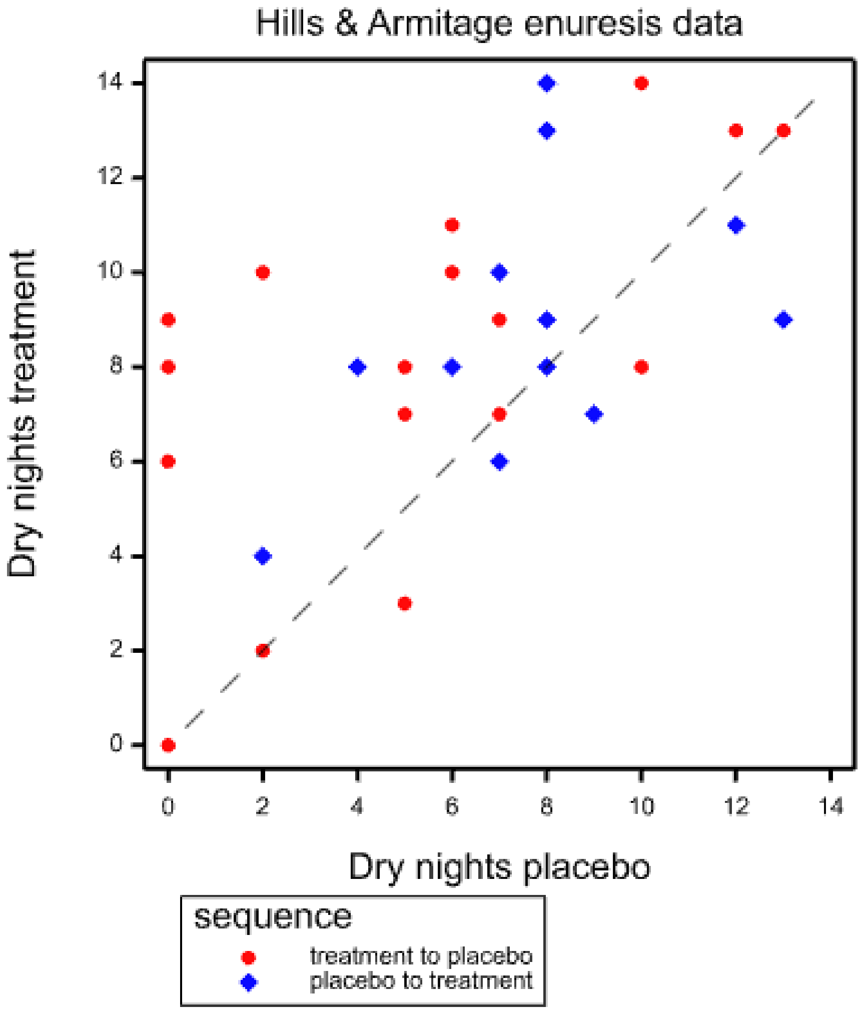
It is astonishing that so many have been taken in by this but if they had actually tried to calculate, their noses would have been rubbed in the problem

The importance of ratios

- In fact 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

Hills and Armitage 1979

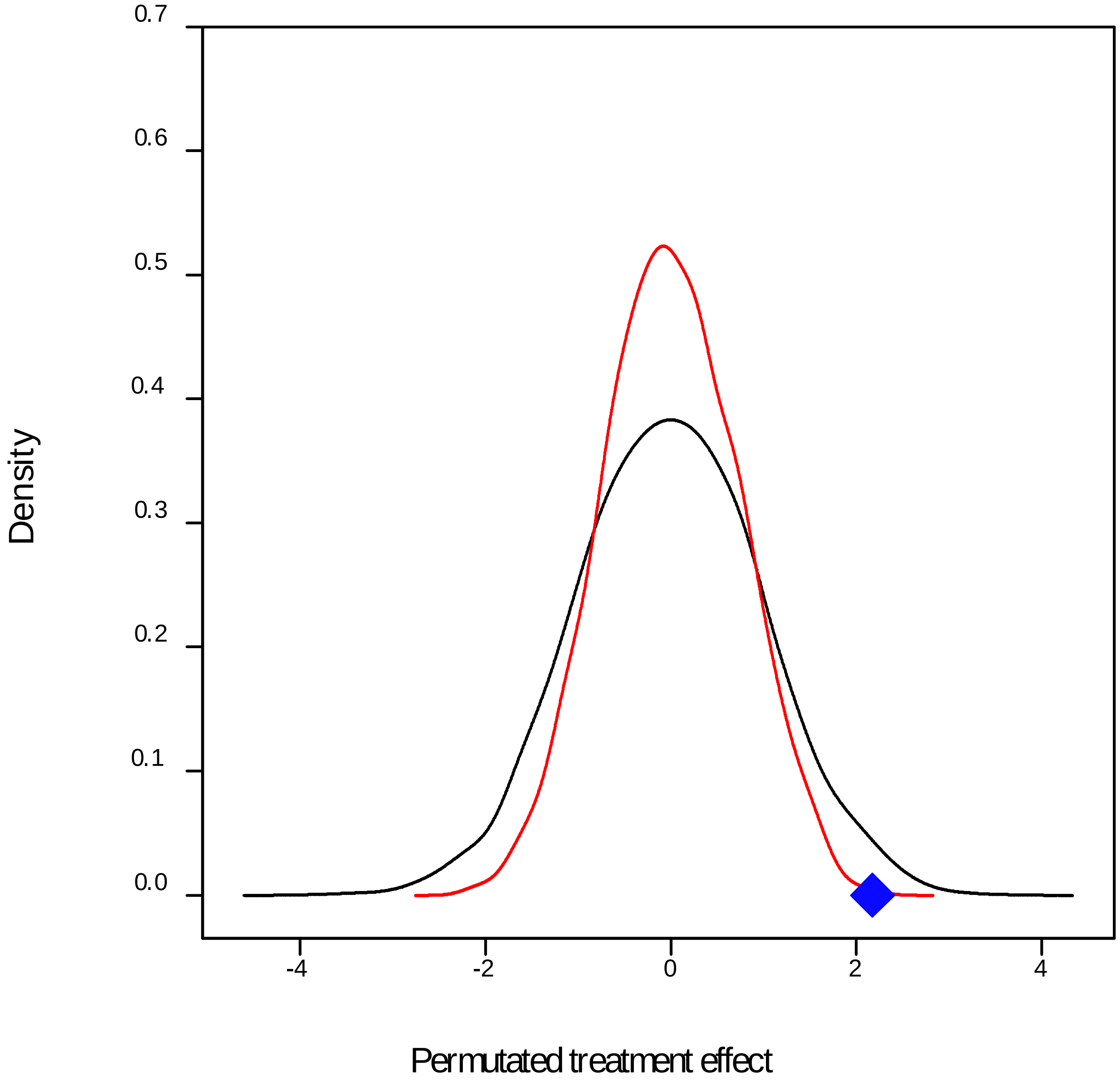
- Trial of enuresis
- Patients randomised to one of two sequences
  - Active treatment in period 1 followed by placebo in period 2
  - Placebo in period 1 followed by active treatment in period 2
- Treatment periods were 14 days long
- Number of dry nights measured



Cross-over trial in Eneuresis

Two treatment periods of 14 days each

1. Hills, M, Armitage, P. The two-period cross-over clinical trial, *British Journal of Clinical Pharmacology* 1979; **8**: 7-20.



Blue diamond shows treatment effect whether or not we condition on patient as a factor.

It is identical because the trial is balanced by patient.

However the permutation distribution is quite different and our inferences are different whether we condition (red) or not (black) and clearly balancing the randomisation by patient and not conditioning the analysis by patient is wrong

The two permutation* distributions summarised

Summary statistics for Permuted difference no blocking

Number of observations =	10000
Mean =	0.00561
Median =	0.0345
Minimum =	-3.828
Maximum =	3.621
Lower quartile =	-0.655
Upper quartile =	0.655
P-value for observed difference	0.0340

Summary statistics for Permuted difference blocking

Number of observations =	10000
Mean =	0.00330
Median =	0.0345
Minimum =	-2.379
Maximum =	2.517
Lower quartile =	-0.517
Upper quartile =	0.517
P-value for observed difference	0.0014

*Strictly speaking randomisation distributions

Two Parametric Approaches

Not fitting patient effect

Estimate	s.e.	t(56)	t pr.
2.172	0.964	2.25	0.0282

(P-value for permutation is 0.034)

Fitting patient effect

Estimate	s.e.	t(28)	t pr
.			
2.172	0.616	3.53	0.00147

(P-value for Permutation is 0.0014)

What happens if you balance but don't condition?

That is to say, permute values respecting the fact that they come from a cross-over but analysing them as if they came from a parallel group trial

Approach	Variance of estimated treatment effect over all randomisations*	Mean of variance of estimated treatment effect over all randomisations*
Completely randomised Analysed as such	0.987	0.996
Randomised within-patient Analysed as such	0.534	0.529
Randomised within-patient Analysed as completely randomised	0.534	1.005
*Based on 10000 random permutations		

In terms of t-statistics

Approach	Observed variance of t-statistic over all randomisations*	Predicted theoretical variance
Completely randomised Analysed as such	1.027	1.037
Randomised within-patient Analysed as such	1.085	1.077
Randomised within-patient Analysed as completely randomised	0.534	1.037@

*Based on 10000 random permutations

@ Using the common falsely assumed theory

The Shocking Truth

- The validity of conventional analysis of randomised trials does not depend on covariate balance
- It is valid because *they are not* perfectly balanced
- If they were balanced the standard analysis would be *wrong*
- The cross-over trial balances for *30,000 genes and all history to date for each patient*
- The parallel group trial does not
- *Because* it does not, it posts a higher variance
- If we have taken care to balance all these tens of thousands of covariates, *analysing as if we hadn't is wrong*

Being a statistician means never having to say you are certain

What the armchair critics have overlooked is that it is not enough to say that randomisation will not guarantee a perfect point estimate. No statistician ever said it would.

The *probability* statement has to be attacked

But does it really matter?

- Does all this obsession with concurrent control and randomisation really matter
- Couldn’t we just get by with historical controls?
- The TARGET study shows the problem

The TARGET study

- One of the largest studies ever run in osteoarthritis
- 18,000 patients
- Randomisation took place in two sub-studies of (nearly) equal size
    - Lumiracoxib versus ibuprofen
    - Lumiracoxib versus naproxen
- Purpose to investigate cardiovascular and gastric tolerability of lumiracoxib
    - That is to say side-effects on the heart and the stomach

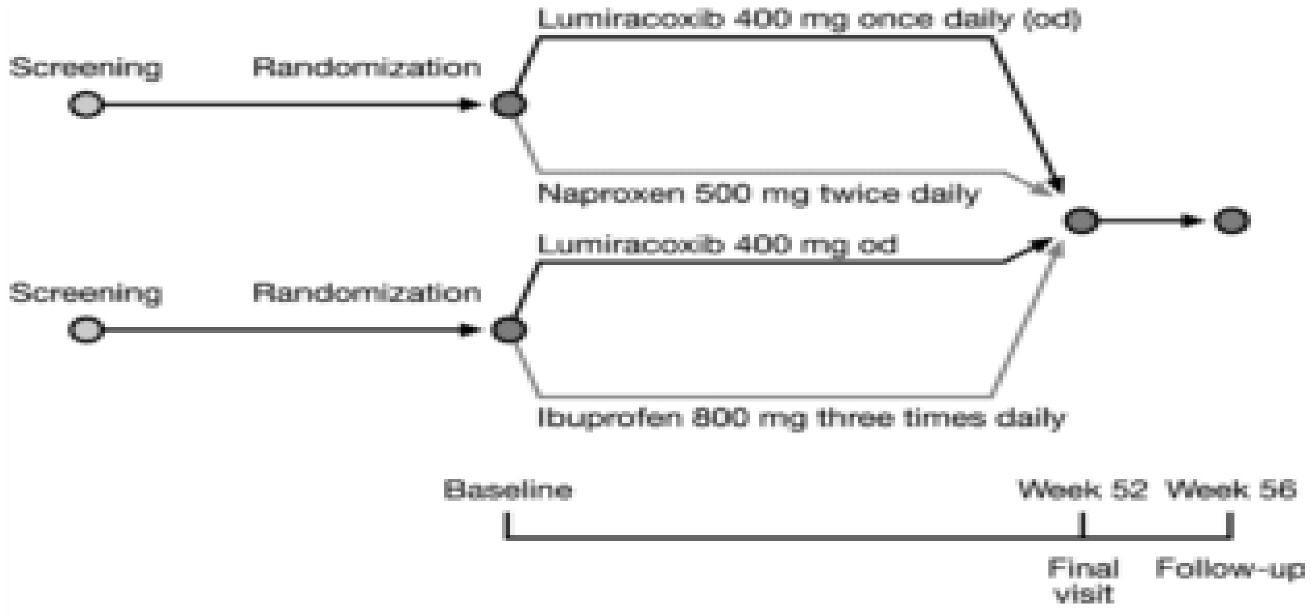


Figure 1. Therapeutic Arthritis Research and Gastrointestinal Event Trial – study design.

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 Deviances

	Model Term		
Demographic Characteristic	**Sub-study (DF=1)**	**Treatment given Sub-study (DF=2)**	**Treatment (DF=2)**
Use of low-dose aspirin	23.57	0.13	13.40
History of vascular disease	70.14	5.23	47.41
Cerebro-vascular disease	13.54	0.14	7.75
Dyslipidaemias	117.98	0.17	54.72
Nitrate use	39.83	4.62	29.17

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

To sum up

- There are important differences between the sub-studies at the outset which would be extremely unlikely to occur by chance
- On the other hand the sort of difference that we see within sub-studies at baseline is the sort that could arise very easily by chance
- So it seems at least that not randomising can be very dangerous
- In this trial provided we compare treatments within sub-studies there is no problem

Lessons from TARGET

- If you want to use historical controls you will have to work ***very*** hard
- You need at least two components of variation in your model
  - Between centre
  - Between trial
- And possibly a third
  - Between eras
- What seems like a lot of information may not be much
- Concurrent control and randomisation seems to work well

My Philosophy of Clinical Trials

- Your (reasonable) beliefs dictate the model
- You should try measure what you think is important
- You should try fit what you have measured
  - Caveat : random regressors and the Gauss-Markov theorem
- If you can balance what is important so much the better
  - But fitting is more important than balancing
- Randomisation deals with unmeasured covariates
  - You can use the distribution *in probability* of *unmeasured* covariates
  - For *measured* covariates you must use the actual *observed* distribution
- Claiming to do ‘conservative inference’ is just a convenient way of hiding bad practice
  - Who thinks that analysing a matched pairs t as a two sample t is acceptable?

What’s out and What’s in

Out	In
Log-rank test	Proportional hazards
T-test on change scores	Analysis of covariance fitting baseline
Chi-square tests on 2 x 2 tables	Logistic regression fitting covariates
Responder analysis and dichotomies	Analysis of original values
Balancing as an excuse for not conditioning	Modelling as a guide for designs

Unresolved Issue

- In principle you should never be worse off by having more information
- The ordinary least squares approach has two potential losses in fitting covariates
  - Loss of orthogonality
  - Losses of degrees of freedom
- This means that eventually we lose by fitting more covariates

Resolution?

- The Gauss-Markov theorem does not apply to stochastic regressors
- In theory we can do better by having random effect models
- However there are severe practical difficulties
- Possible Bayesian resolution in theory
- A pragmatic compromise of a limited number of prognostic factors may be reasonable

To sum up

- There are a lot of people out there who fail to understand what randomisation can and cannot do for you
- Statisticians need to tell them firmly and clearly what they need to understand
- Getting dirty and wading in are great aids to thinking

Finally

I leave you with this thought

Statisticians are always tossing coins but do not own many

ADDITIONAL TARGET STUDY SLIDES

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