Randomisation: Misunderstanding, myths and truth.

Stephen Senn

Acknowledgements

Acknowledgements

Thank you to the ULB for the 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"

Plan

I am unsure as to how much you actually know about randomisation in clinical trials, so I am going to give some practical background on this and give you some examples.

There is not much in the way of theory in these but this will give you some feel for the real world of clinical trials

However the practice is important because it is in ignoring the practice that critics have gone astray

The second part will be very different. I am going to cover some of the theory of randomisation

Part I

Basics and practical examples

My excuse for this elementary introduction

Some very clever people have made fools of themselves by telling clinical trialists how they would do things better without understanding the basics

Basics of clinical trials

The principle of concurrent control

- Clinical trials are comparative
- A new treatment is compared to a control *in the same trial*
  - Sometimes a standard treatment
  - Sometimes a placebo
- Why?
  - Because we know results vary strongly from trial to trial
- Typically a random choice is made as to which patient gets which treatment

Basics of clinical trials

Patient recruitment is sequential

- Patients are (nearly always) entered sequentially
  - You treat patients when they ‘present’ at the clinic
  - They don’t all present at the same time
  - They are often recruited over several months
  - Some of them may have finished the trial before others have started
- This reality must be faced by all allocation procedures
- Randomisation can be regarded as a game between trialist and physician.
- Object is to prevent the physician biasing the allocation

Basics of clinical trials

What is randomisation?

- Randomisation is the process of deciding which patient gets which treatment using an element of chance
- It can be ***unrestricted***
  - For example we toss a coin for every new patient
- It can be ***restricted*** in some way
  - Randomised blocks
    - For example in every set of 8 patients, 4 get the new treatment and 4 get the standard but the order is left to chance

Basics of clinical trials

The physical basis of blinding

- The physical basis of blinding involves placebos
- Dummy drugs must be available that match each treatment in terms of taste, colour etc.
- Every drug manufacturer is required to also make placebos to their won drugs for clinical research *and provide them to competitors when asked*

Basics of clinical trials

The use of dummies for blinding

- Drugs do not resemble each other
- Therefore in comparative trials we need placebos to each drug
- The patient will receive
  - Either active A and **placebo to B**
  - Or **placebo to A** and active B
- This is called *the double dummy technique*
- If the dosing schedules are not the same (e.g. once daily versus twice daily) then placebo occasions have to be organised
- This is called *double dummy loading*

Basics of clinical trials

Randomisation is necessary for blinding

- Humans are not good at choosing random sequences
- They tend to avoid repetitions
- If a particular sequence looks random to you it may also look random to another
- The other's ability to guess may be greater than you think
- This means that probability calculations become very speculative
- The solution is to randomise

Two approaches to allocation

Pre-randomisation

- Identical looking but numbered packs are sent out to clinics
- At random some of the packs are the experimental treatment and some are the control
- The investigator gives the *next* patient the pack with the *lowest* remaining number
- Only at the end of the trial will be it revealed who got what

Central-randomisation

- The investigator enters the patient remotely onto the system giving all patient details
- If entry onto the trial is approved the system tells the investigator what pack number to pick up
- Only once the patient is entered is the pack number revealed and then it is too late for the investigator to change his mind

Basics of clinical trials

Types of clinical trial

- Parallel group
  - Patients are individually randomised to treatment
    - for example they either get A or B
- Cross-over trial
  - Patients are randomised to sequences of treatment for the purpose of comparing individual treatments
    - For example they are randomised to A followed by B or to B followed by A
- Cluster randomised trial
  - Patients are randomised by centre to receive a treatment
    - For example, either all the patients in that centre receive A or all the patients receive B

What Does Randomisation not Do?

- Does not guarantee patients are representative
- Does not guarantee balance as regards prognostic factors
  - Better than many other systems
  - But only guarantees balance in expectation
  - For a limited number of factors stratification can be added
- Also not possible for all factors

An Example

- Most double blind trials are run using the double dummy technique
- Patients either receive A and placebo to B or B and placebo to A
- But we hardly ever randomise the order in which the two pills are taken
- Why not?

Is randomisation fully efficient?

No.
A simple example. We have a parallel group trial currently we have n_A patients on A and n_B on B. To minimise the variance of the treatment effect we should deliberately allocate the next patient to the treatment with the lower number of patients. (If the trial is currently balanced it makes no difference.)

Proof Assume without loss of generality that $n_A < n_B$
If we allocate to **group A**, the variance of the estimated treatment effect will be proportional to

$$\frac{1}{n_A+1}+\frac{1}{n_B}=\frac{n_A+n_B+1}{n_An_B+n_B} \text{ .....(1)}$$

If we allocate to **group B**, the variance of the estimated treatment effect will be proportional to

$$\frac{1}{n_A}+\frac{1}{n_B+1}=\frac{n_A+n_B+1}{n_An_B+n_A} \text{ .....(2)}$$

However, the numerator of (1) and (2) are identical and the denominators only differ in the last term, $+n_B$ (expression 1) or $+n_A$ (expression, 2), as the case may be. However since $n_A<n_B$ we have (1) < (2), therefore we should allocate to group A

But randomisation is often pretty good

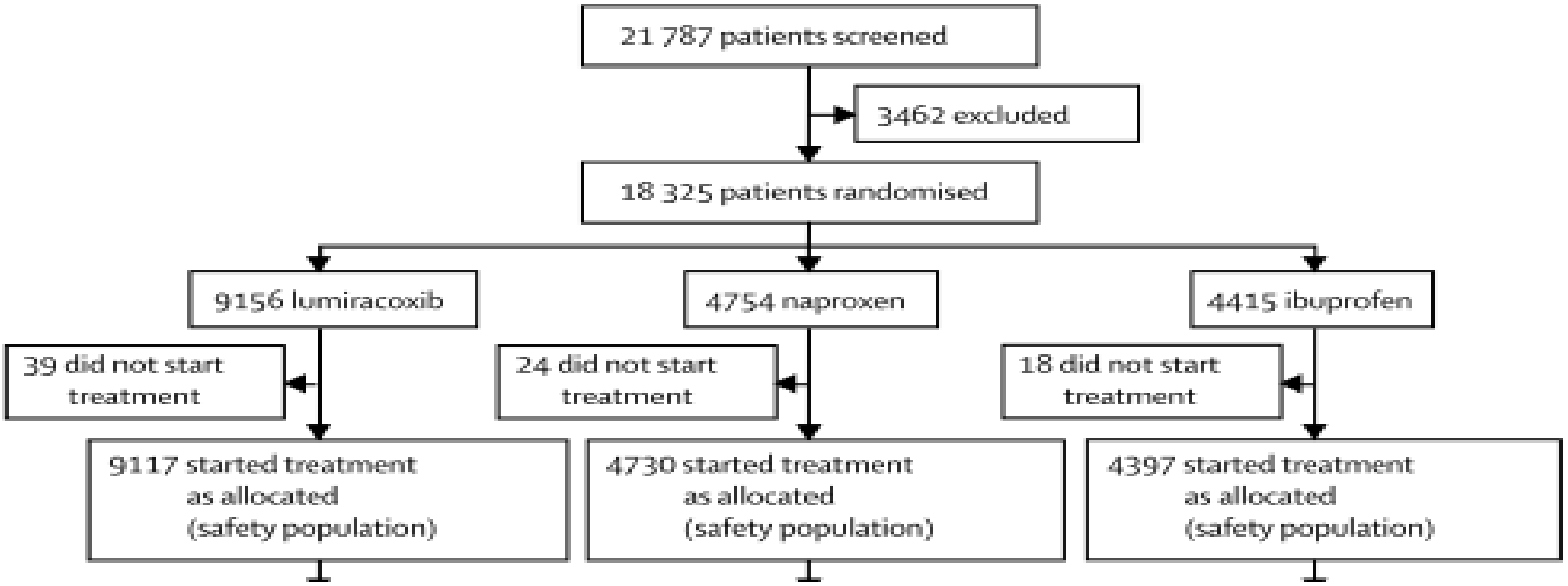
"…or by whether their hair is parted on the left or the right, or one could simply permit the subjects to choose their own groups, always ensuring of course that they have not been informed of which treatment is to be applied to which group…"

Urbach 1985, p271

- Suppose that we assume that the patients choose independently and let θ be the probability that a patient chooses A so $(1-\theta)$ is the probability the patients choses B
- Randomisation is like knowing $\theta = 0.5$
- But every value of θ other than 0.5 is worse

The TARGET study

- One of the largest studies ever run in osteoarthritis
- 18,000 patients
- Randomisation took place in two sub-studies of 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



Extract from **incorrect** CONSORT diagram in the Lancet

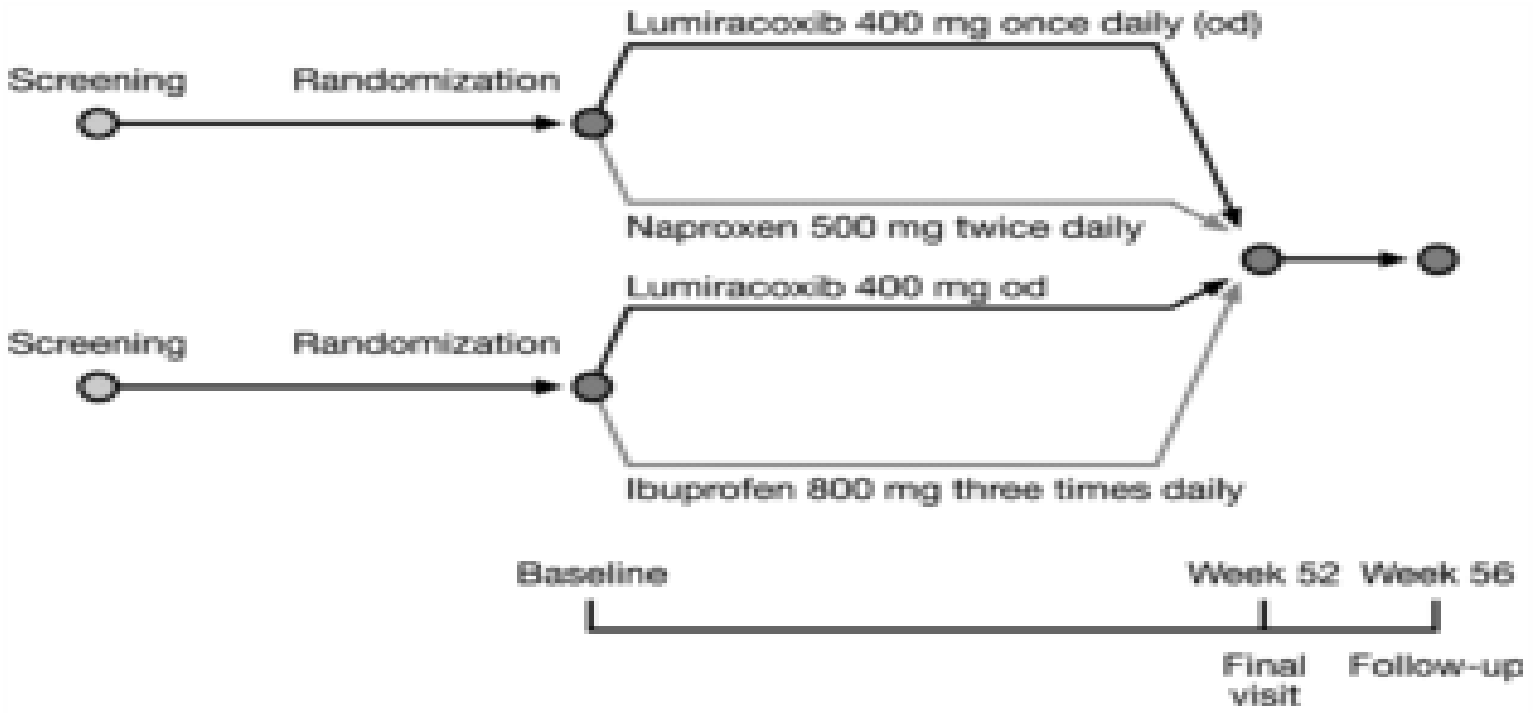


Figure 1. Therapeutic Arthritis Research and Gastrointestinal Event Trial – study design.

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 and 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

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

Is it a between-centre difference?

- TARGET had 18,224 patients in 849 centres
- Centre size varied from 12 to 167 patients with an average of 22
- However the deviance at outcome for sub-study amongst lumiracoxib for discontinuation is 37.4 and for dyspepsia is 16.8 and these are not easily explained as being due to *random differences* between centres

A Simple Model

An unrealistic balanced trial

n patients per arm, c centres in total with p patients per centre

$$2n = pc, \qquad n = \frac{pc}{2}$$

Between-centres variance is γ^2 within-centre variance is σ^2.

Design	Variance of Treatment Contrast
Completely randomised	$4\frac{(\gamma^2+\sigma^2)}{cp}$
Randomised blocks (centre blocks)	$4\frac{\sigma^2}{cp}$
Cluster randomised	$4\frac{\left(\gamma^2+\frac{\sigma^2}{p}\right)}{c}$

When using external controls we have *at least* the variability of a cluster randomised trial

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

Part 2

Some theory

Game of Chance

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

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, the 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

Dice games

- You can't treat game 2 like game 1
  - You must condition on the information received
  - 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

Clinical Trials

- You can't ignore an observed prognostic covariate 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
  - You are entitled to treat game 3 like game 1

The error

- The error is to assume that because you can't use randomisation as a justification for ignoring information it is useless
- It is useful for what you don't see
- Knowing that the two-dice game is fair is important even though the average probability is not relevant to game two
- ***Average probabilities are important for calibrating your inferences***
  - Your conditional probabilities must be coherent with your marginal ones
    - ➢ See the relationship between the games

A Red Herring

"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

- 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

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

Where *Z* is a treatment indicator and the X 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

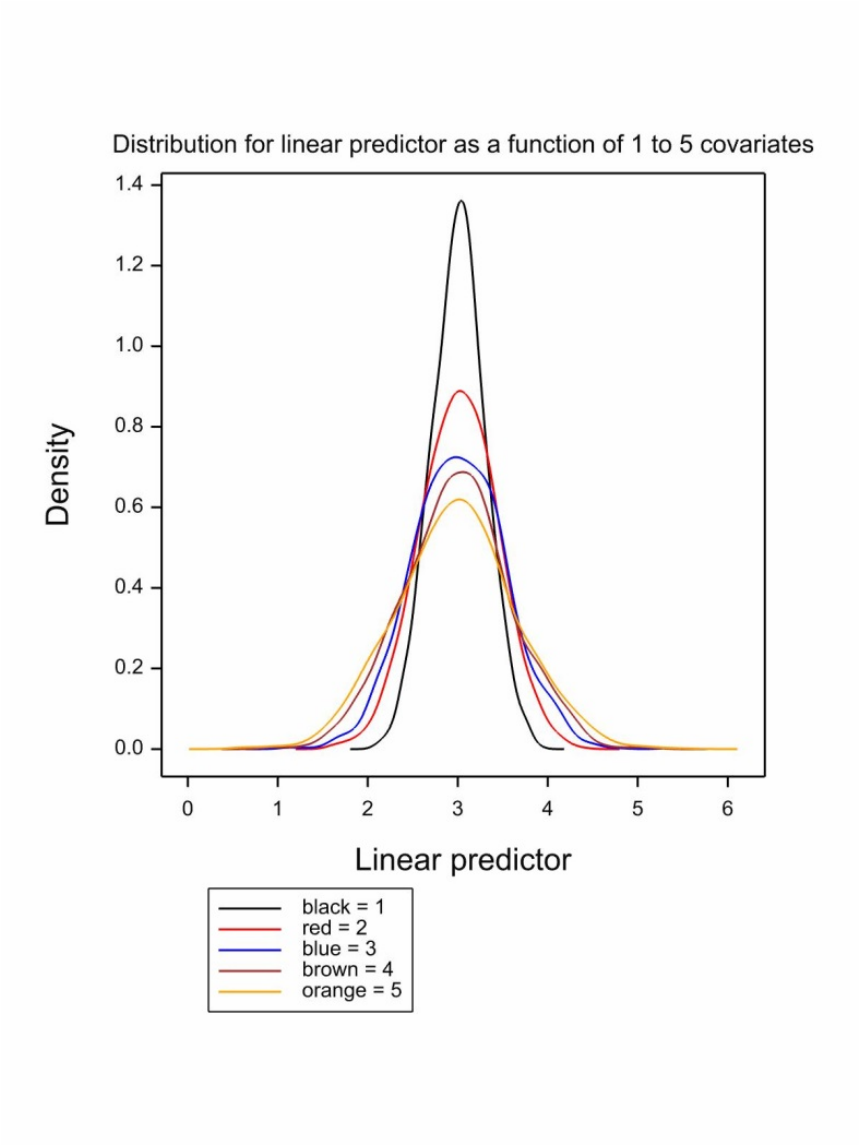
Simulation of the linear predictor as the number of covariates increases from 1 to 5

However, the variance of each predictor is the same and the coefficient is the same

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

The values soon become impossible

The total contribution that the predictors can make is bounded



In fact this is pointless

Look at the equation again

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

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

$$Y = Y^* + Z$$

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

The importance of ratios

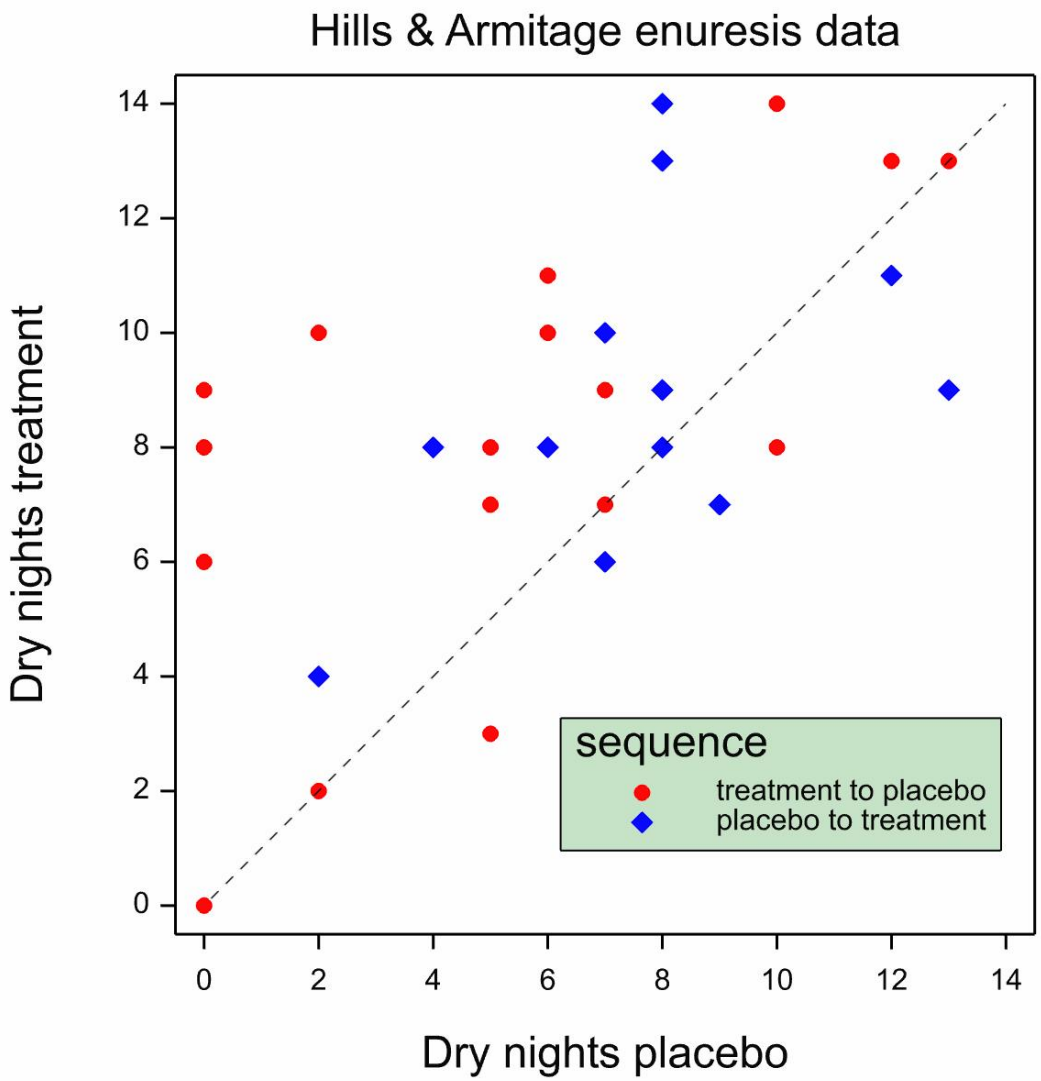
- So from one point of view there is only one covariate that matters
  - potential outcome
    - If you know this, all other covariates are irrelevant
- And just as this can vary between groups in can vary within
- The t-statistic is based on the ***ratio*** of differences *between* to variation *within*
- 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

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

Important points to note

- Because every patient acts as his own control all **patient level** covariates (of which there could be thousands and thousands) are perfectly balanced
- Differences in these covariates can have no effect on the difference between results under treatment and the results under placebo
- However, **period level** covariates (changes within the lives of patients) could have an effect
- My normal practice is to fit a period effect as well as patients effects, however, I shall omit doing so to simplify



Cross-over trial in Enuresis

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.

Two Parametric Approaches

Not fitting patient effect

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

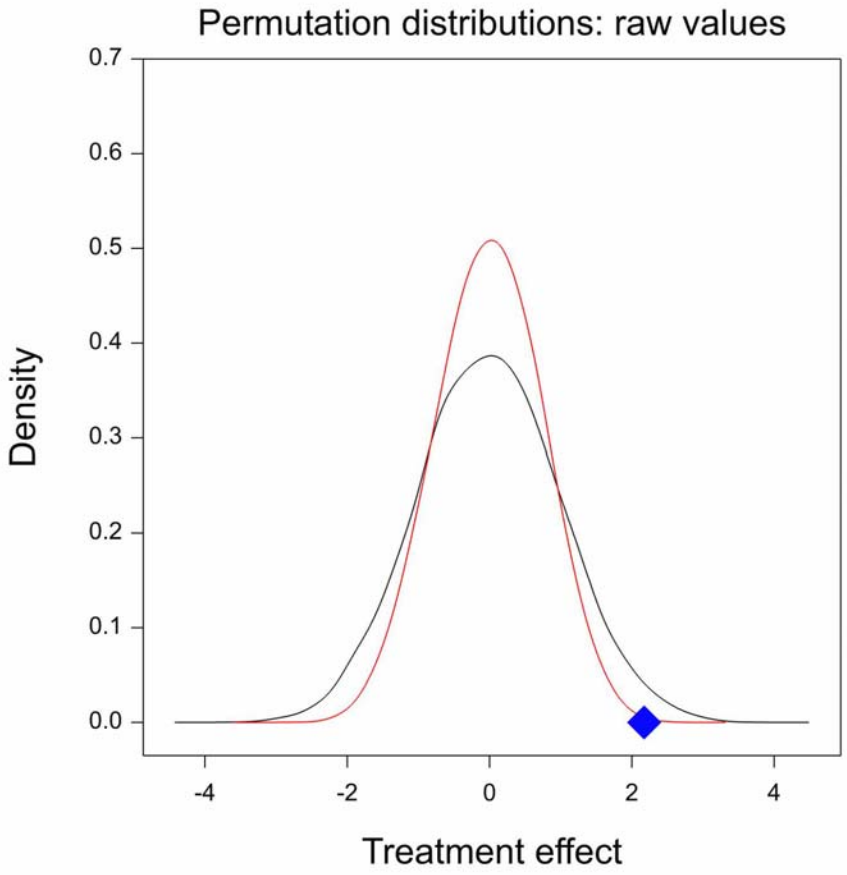
Fitting patient effect

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

Note that ignoring the patient effect, the P-value is less impressive and the standard error is larger

The method *posts higher uncertainty* because unlike the within-patient analysis it make no assumption that the patient level covariates are balanced.

Of course, in this case, since we know the patient level covariates are balanced, this analysis is invalid.



Blue diamond shows treatment effect whether we condition on patient or not 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 = -0.00319
- Median = -0.0345
- Minimum = -3.621
- Maximum = 3.690
- Lower quartile = -0.655
- Upper quartile = 0.655

P-value for observed difference 0.0344

(Parametric P-value 0.0282)

Summary statistics for Permuted difference blocking

Number of observations = 10000

- Mean = -0.00339
- Median = 0.0345
- Minimum = -2.793
- Maximum = 2.517
- Lower quartile = -0.517
- Upper quartile = 0.517

P-value for observed difference 0.001

(Parametric P-value 0.00147)

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*

Typical nonsense encountered in the medical press

‘The central telephone randomisation system used a minimisation algorithm to balance the treatment groups with respect to eligibility criteria and other major prognostic factors.’ (p24)

‘All comparisons involved logrank analyses of the first occurrence of particular events during the scheduled treatment period after randomisation among all those allocated the vitamins versus all those allocated matching placebo capsules (ie, they were “intention-to treat” analyses).’ (p24)

1. (2002) MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* **360**:7-22

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

- Log-rank test
- T-test on change scores
- Chi-square tests on 2 x 2 tables
- Responder analysis and dichotomies
- Balancing as an excuse for not conditioning

In

- Proportional hazards
- Analysis of covariance fitting baseline
- Logistic regression fitting covariates
- Analysis of original values
- 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

The Problem

- However, this seems to imply that in making inferences for randomised clinical trials we must condition on everything we observe
- All covariates must be in the model
- What is the effect on efficiency?
- Could it mean that more information is worse than less?

A Quote from Jack Good

The use of random sampling is a device for obtaining apparently precise objectivity but this precise objectivity is attainable, *as always*, only at the price of throwing away some information (by using a *Statistician's Stooge* who knows the random numbers but does not disclose them)…

…But the use of sampling without randomization involves the pure Bayesian in such difficult judgments that, at least if he is at all Doogian, he might decide by Type II rationality, to use random sampling to save time.

Resolution?

- In theory we can do better than ordinary least squares by having random effect models
  - Gauss-Markov theorem does not apply to stochastic regressors
- However there are severe practical difficulties
- Possible Bayesian resolution in theory
- A pragmatic compromise of a limited number of prognostic factors may be reasonable
- This is exactly what ICHE9 suggests
  - Awareness of the issues seems to be much greater amongst drug regulators than amongst journal editors

To sum up

- Randomisation makes a valuable contribution to handling unobserved variables
- Randomisation does not guarantee balance of unobserved variables
  - This balance is not needed
  - If it applied conventional analyses would be invalid
- Randomisation is not an excuse for ignoring prognostic covariates
- Some technical challenges remain but these are challenges of modelling not randomisation *per se*

Finally

I leave you with
this thought

Statisticians are always tossing coins but do not own many

RA Fisher at Mablethorpe