Response, quality and variation:

what drug development may be missing

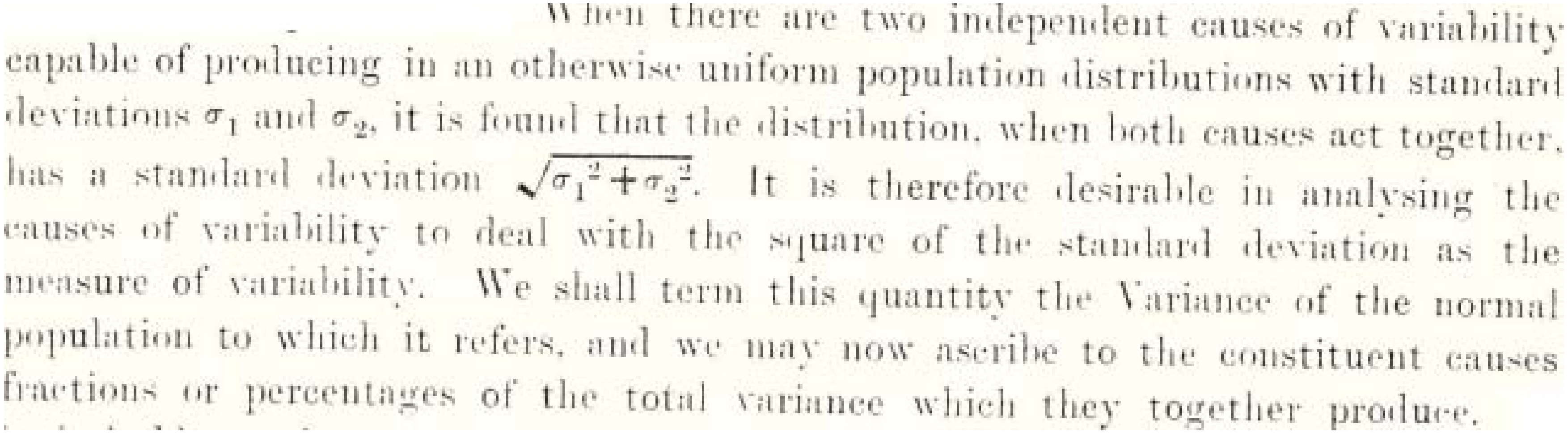
Stephen Senn

LUXEMBOURG
INSTITUTE
OF **HEALTH**
RESEARCH DEDICATED TO LIFE

Acknowledgements

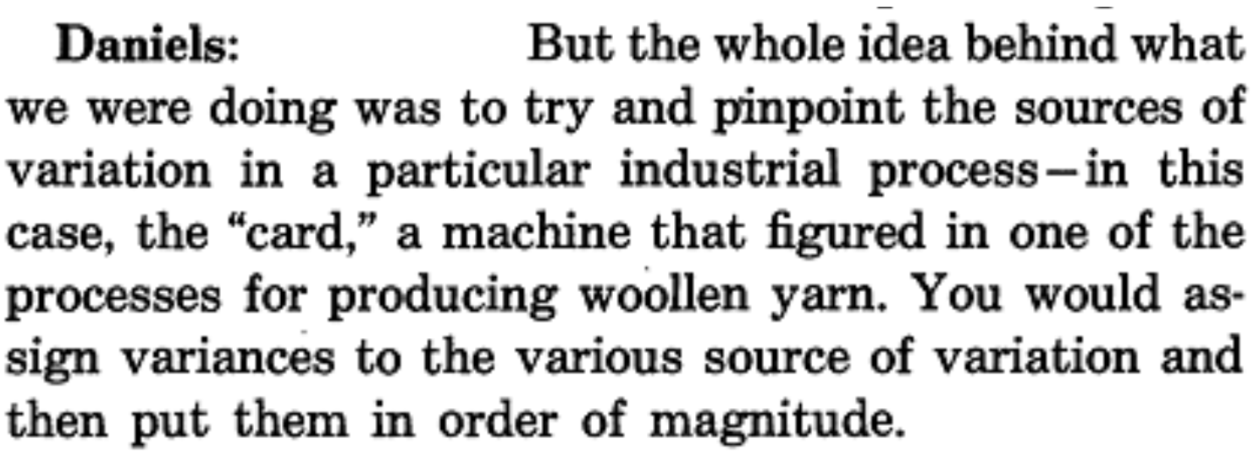
I am grateful for the invitation from Statistical Solutions to address you on a topic dear to me heart

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"



When there are two independent causes of variability capable of producing in an otherwise uniform population distributions with standard deviations σ_1 and σ_2, it is found that the distribution, when both causes act together, has a standard deviation $\sqrt{\sigma_1^2+\sigma_2^2}$. It is therefore desirable in analysing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance of the normal population to which it refers, and we may now ascribe to the constituent causes fractions or percentages of the total variance which they together produce.

Fisher 1918



Daniels: **But the whole idea behind what we were doing was to try and pinpoint the sources of variation in a particular industrial process—in this case, the "card," a machine that figured in one of the processes for producing woollen yarn. You would assign variances to the various source of variation and then put them in order of magnitude.**

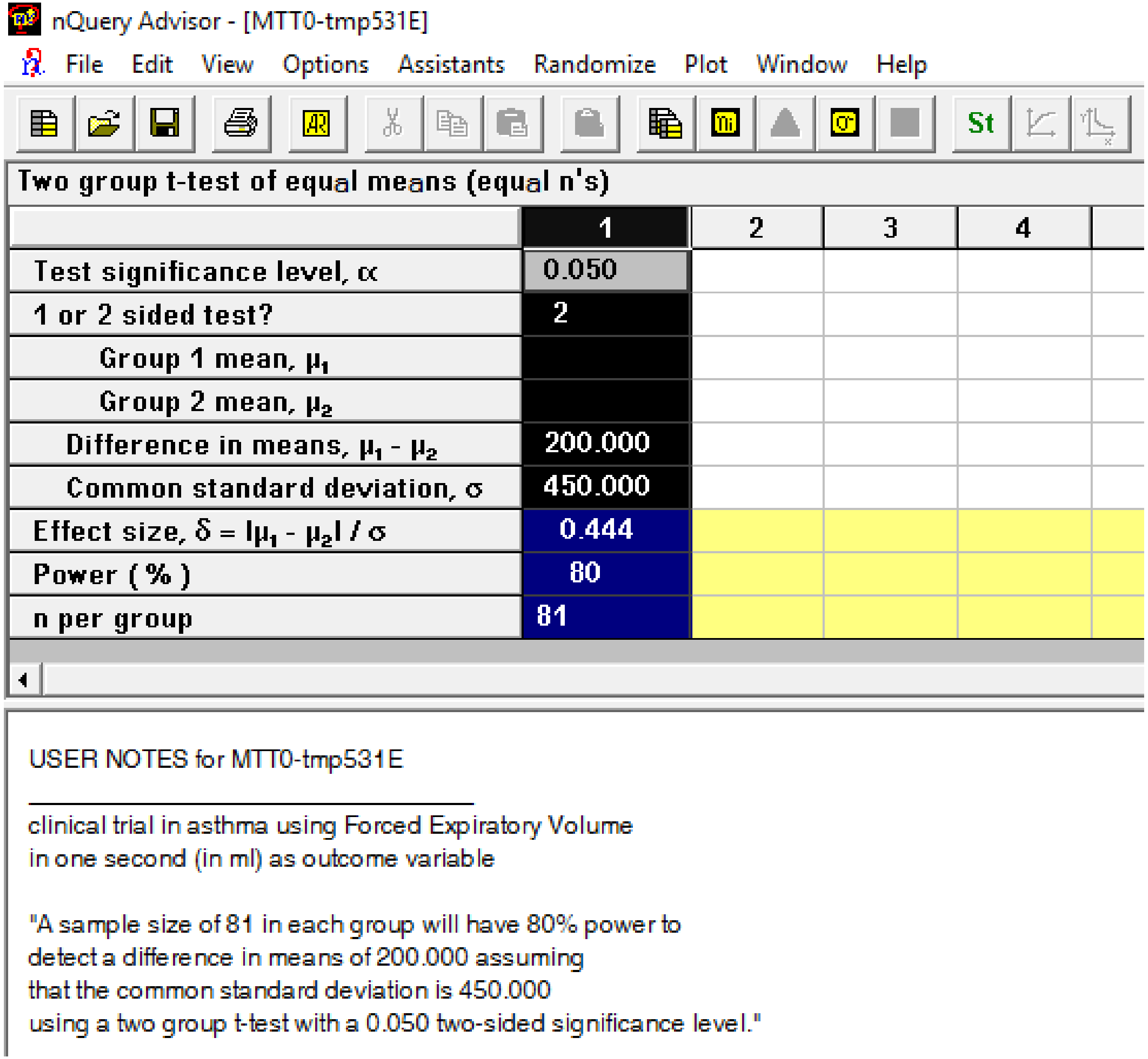
Henry Daniels to Whittle, 1993 but describing work at Wool Research 1935-1942

Variances

Good pharma

Allowing for variation

- The pharmaceutical industry has regularly used formal sample size calculations on clinical trials
- The effect of sample sizes on (sought for) signal to noise ratios has been understood and planned for
- There has been a lot of practical and good work in this direction
  - (But we probably need to move beyond power)



Two group t-test of equal means (equal n's)

	1	2	3	4
Test significance level, α	0.050			
1 or 2 sided test?	2			
Group 1 mean, μ_1				
Group 2 mean, μ_2				
Difference in means, $\mu_1 - \mu_2$	200.000			
Common standard deviation, σ	450.000			
Effect size, $\delta = \vert\mu_1 - \mu_2\vert / \sigma$	0.444			
Power (%)	80			
n per group	81			

USER NOTES for MTT0-tmp531E

clinical trial in asthma using Forced Expiratory Volume
in one second (in ml) as outcome variable

"A sample size of 81 in each group will have 80% power to
detect a difference in means of 200.000 assuming
that the common standard deviation is 450.000
using a two group t-test with a 0.050 two-sided significance level."

Variances

Fair Pharma

Reducing variances

Good

- Cross-over trials in early phases
- Blocking by centre
- Use of covariates
- Clever dose-finding
- Some good work on timing of observations
- Generally done better in micro-array design than academic sponsors

Bad

- Silly models for carry-over
- More use could be made of covariates
- Phase I healthy volunteer trials are a design desert
- Communication between theoreticians and practitioners has not be great

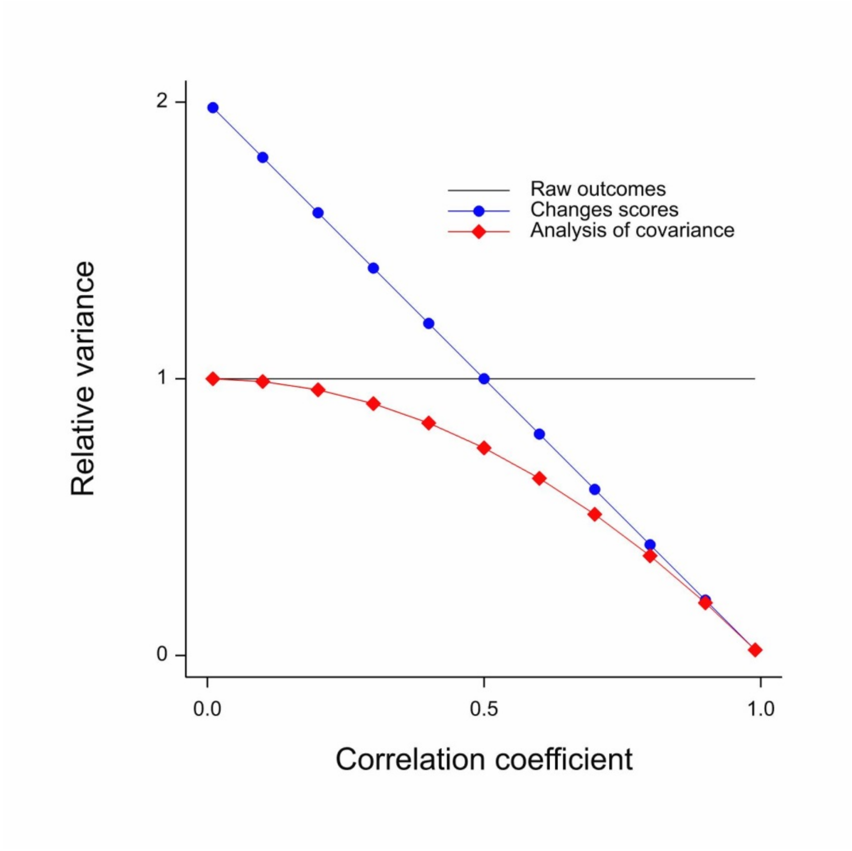
Variances

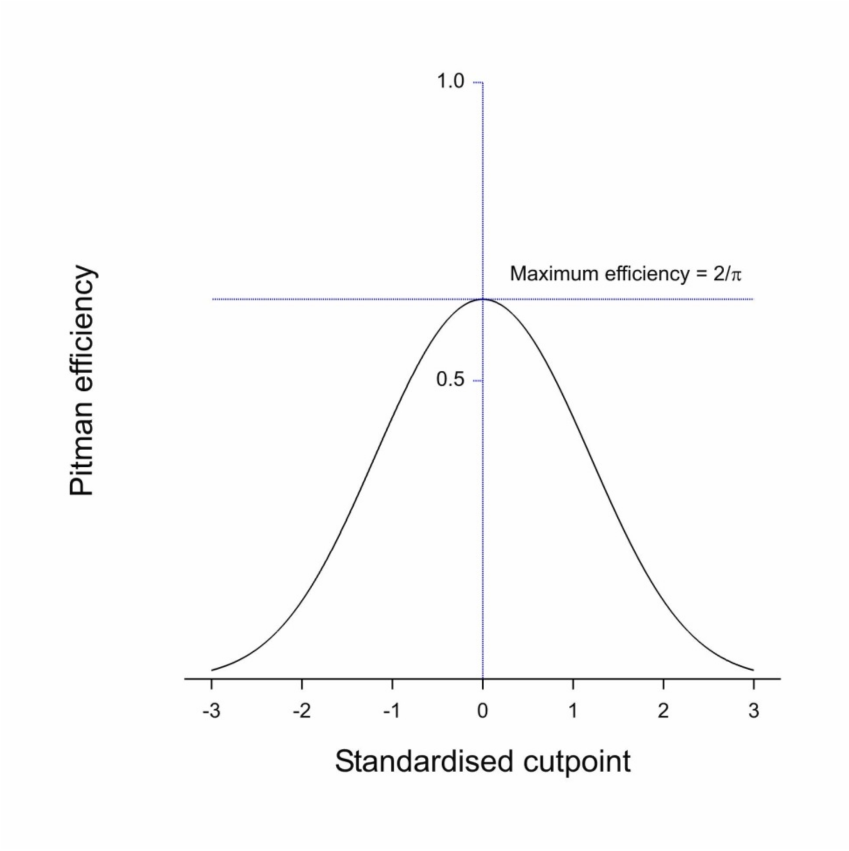
Bad Pharma

Increasing variances

- Change from baseline instead of analysis of covariance
- Refusal to model
  - Heard at the FDA ‘we don’t do modelling’
- Dichotomania

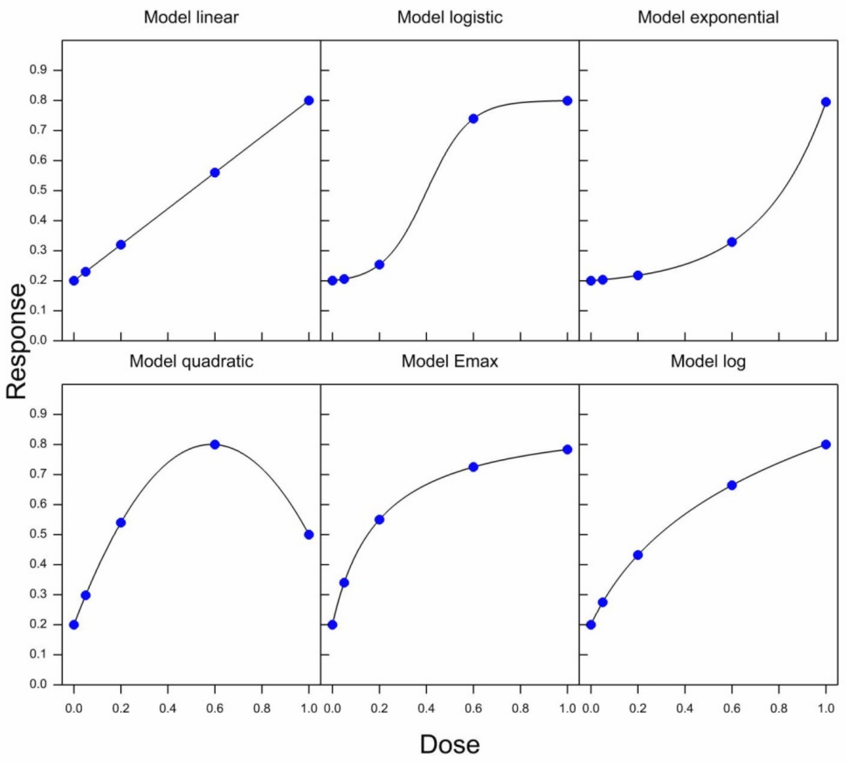
Increasing the variance





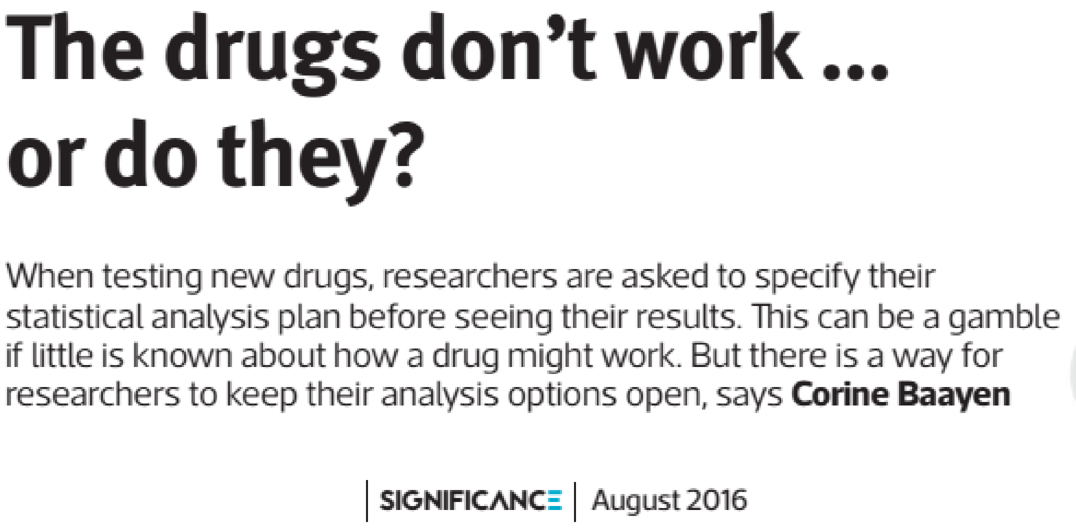
Merck goes dose-finding for migraine

- Complicated design using three stages, 8 doses and 517 patients
- Subsequent analysis by sophisticated MCP-Mod methodology developed at Novartis illustrated by Corine Baayen in *Significance*
- This is all very clever with the exception of one incredibly stupid thing



How they threw information away

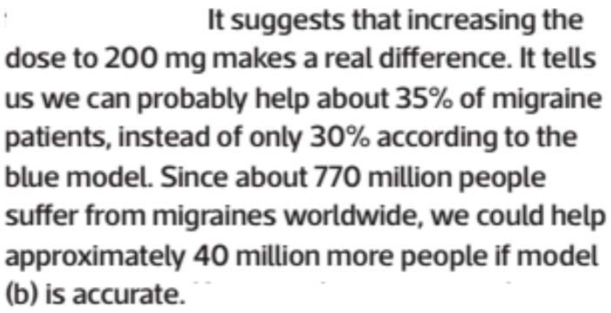
“In each group they measured how many patients were free of pain after two hours”



The drugs don’t work … or do they?

When testing new drugs, researchers are asked to specify their statistical analysis plan before seeing their results. This can be a gamble if little is known about how a drug might work. But there is a way for researchers to keep their analysis options open, says **Corine Baayen**

SIGNIFICANCE | August 2016



It suggests that increasing the dose to 200 mg makes a real difference. It tells us we can probably help about 35% of migraine patients, instead of only 30% according to the blue model. Since about 770 million people suffer from migraines worldwide, we could help approximately 40 million more people if model (b) is accurate.

A question for you

Alas Smith and Jones

Ms Smith had her headache reduced from 8 hours duration to 6 (reduced by 2 hrs or 25%)

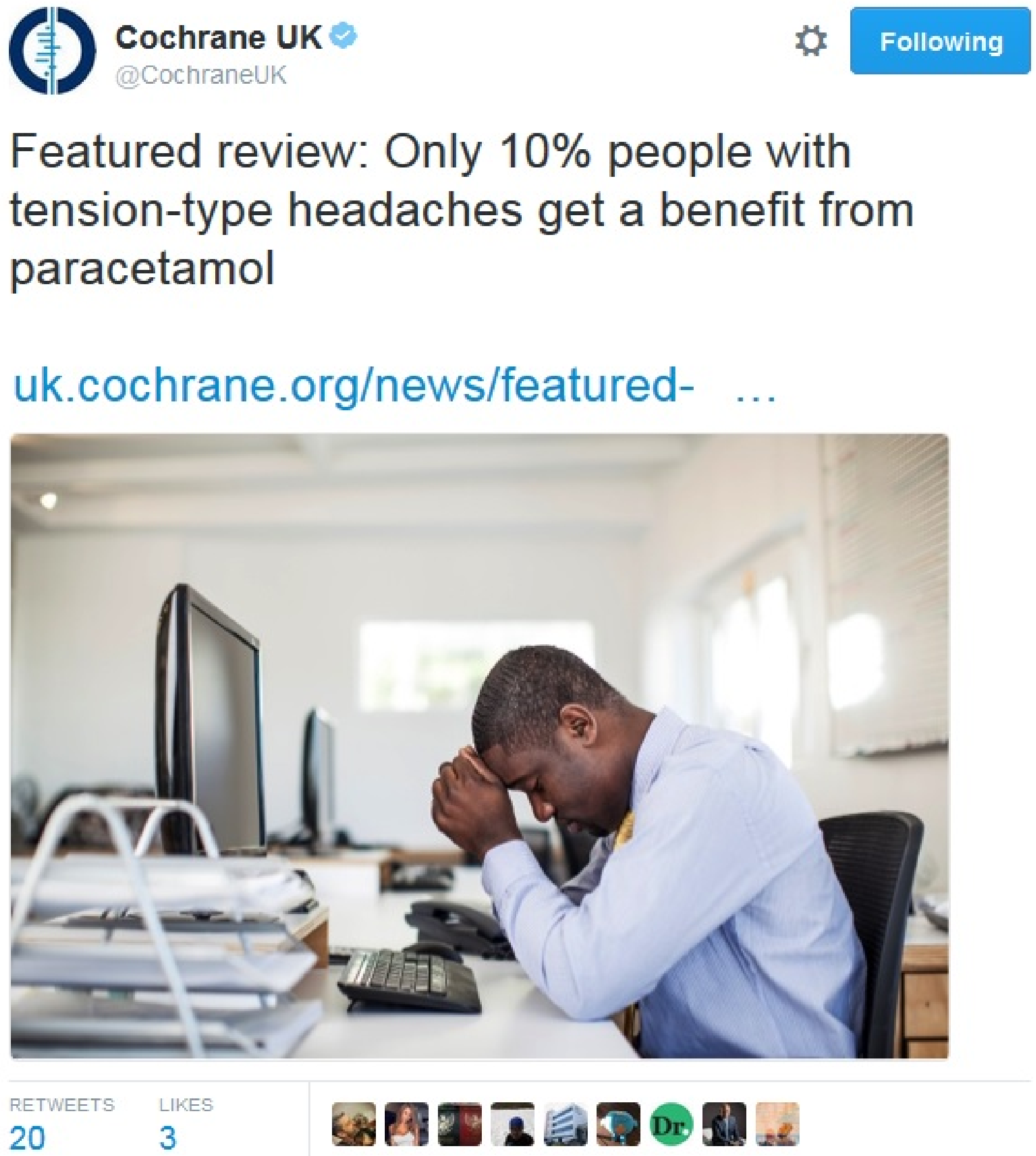
Mr Jones had his headache duration reduced from 2hr05’ to 1hr55’ (reduced by10 minutes or 8%)

Who had the greater benefit?

The International Headache Society recommends the outcome of being pain free two hours after taking a medicine.

So does the FDA

Mr Jones responded. Mrs Smith didn’t.



Featured review: Only 10% people with tension-type headaches get a benefit from paracetamol

uk.cochrane.org/news/featured- …

59% had no headache after 2 hours when treated with paracetamol

49% had no headache after 2 hours when treated with placebo

59%-49% = 10%

Therefore 10% benefitted

The number needed to treat for one extra patient to have a benefit is 10

Painful comparison

Cochrane Collaboration meta-analysis

- Meta-analysis of placebo-controlled trials of paracetamol in tension headache
- 23 studies
- 6000 patients in total
- Outcome measure:
  - Pain free by 2 hours

Baayen *Significance* article

- Explanation of Novartis's MCP-Mod dose-finding approach using a trial run by Merck
- 7 doses + placebo
- 517 patients in total
- Outcome measure
  - Pain free by 2 hours

In both cases

- The patients were only studied once
- A dichotomy of a continuous measure was made
- Patients were labelled as responders and non-responders
- A causal conclusion was drawn that went beyond simply comparing proportions
  - Baayen talked about the proportion of patients who would respond
  - Cochrane talked about the proportion of patients to whom it would make a difference in terms of response

Headaches or patients?

- In fact conclusions about the proportion of patients who will regularly have a response to treatment cannot be drawn from such studies
- You cannot separate headaches and patients
- Furthermore the dichotomy causes causal confusion

We tend to believe “the truth is in there”, but sometimes it isn’t and the danger is we will find it anyway

What I propose to do

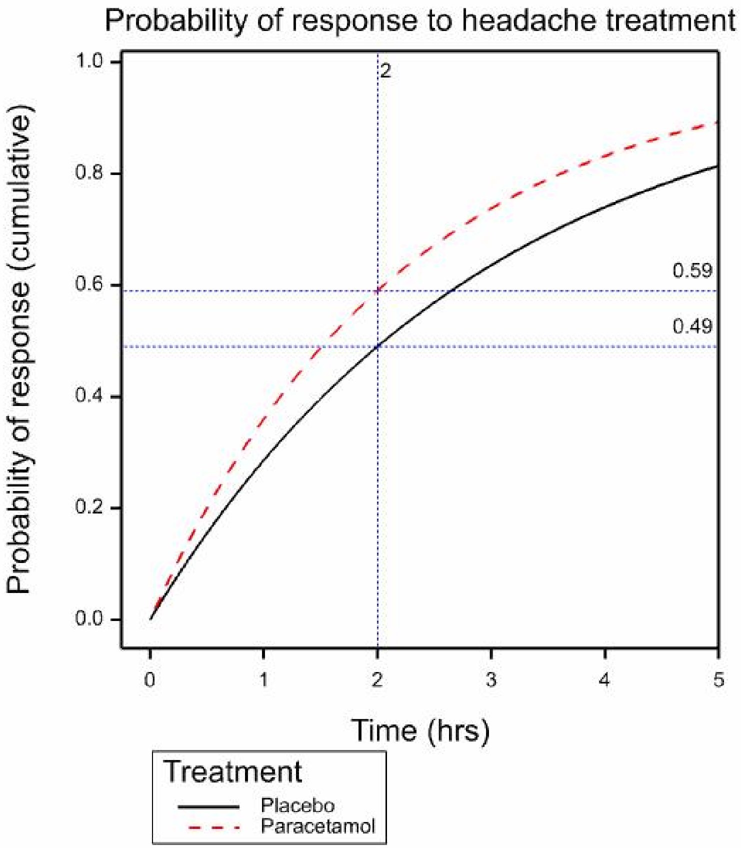
- Create a simple statistical model to mimic the Cochrane result
  - – In terms of time to pain resolution every patient will have the same proportional benefit
    - In fact I shall be using a form of *proportional hazards model*
  - – The dichotomy will classify patients as responders as non-responders
  - – We will be tempted to conclude that some don't benefit and some do and that this is a permanent feature of each patient

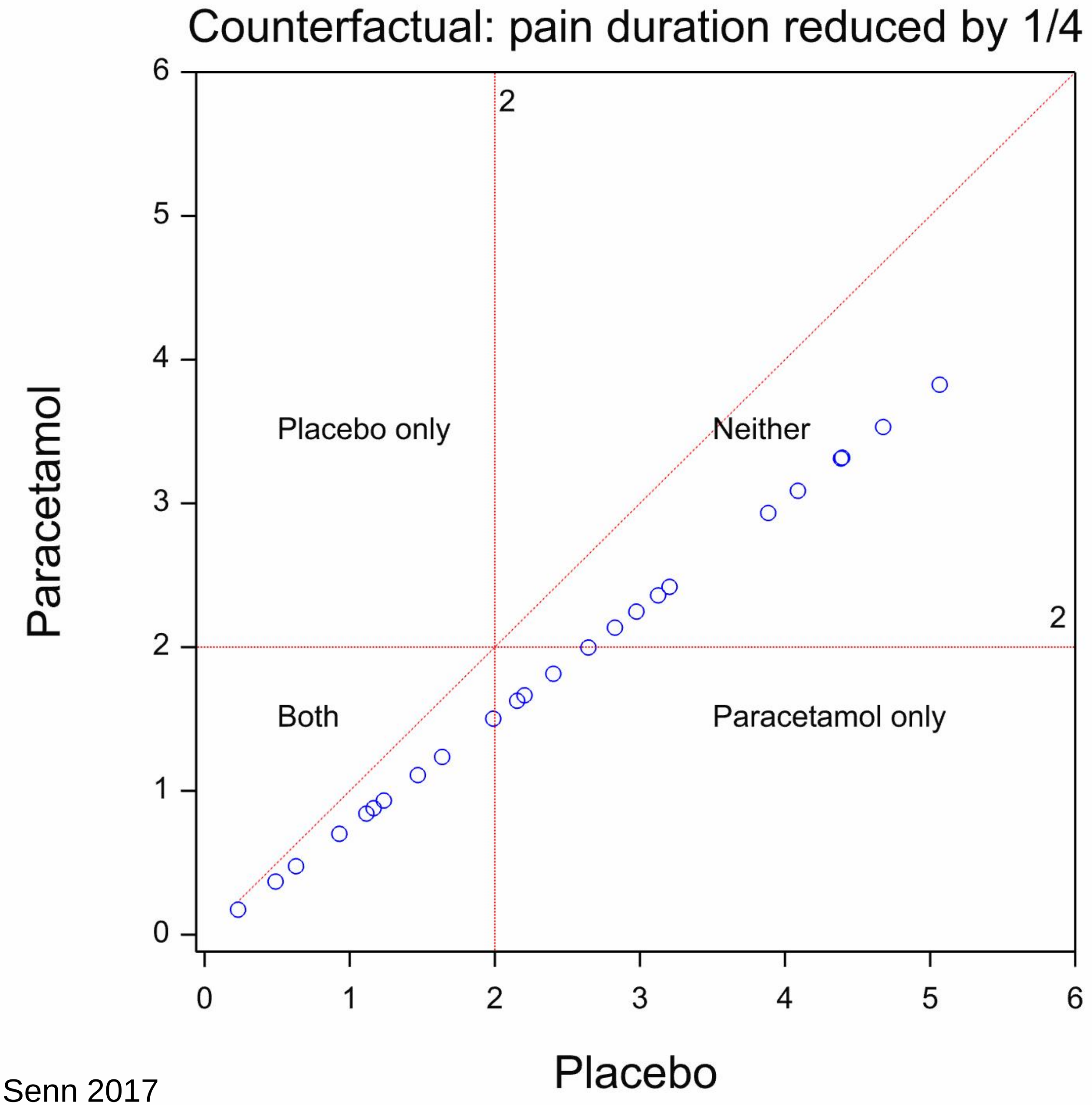
The Numerical Recipe

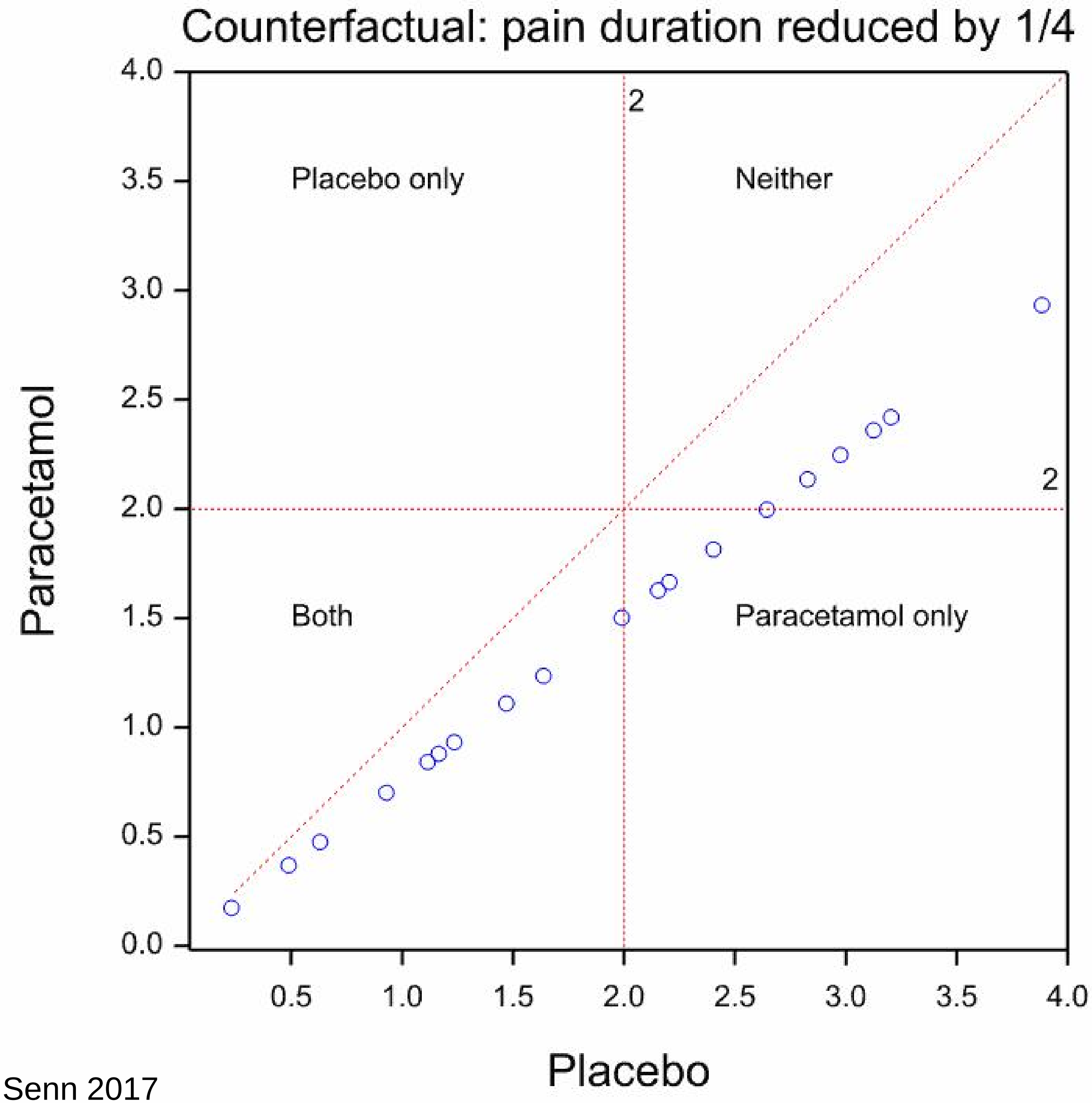
- I shall generate pain duration times for 6000 headaches treated with placebo
  - This will be done using an exponential distribution with a mean of just under 3 hours (2.97 hrs to be exact)
  - Each such duration will then be multiplied by just over ¾ (0.755 to be exact) to create 6000 durations under paracetamol
- I shall then take the 6000 pairs and randomly erase one member of the pair to leave 3000 unpaired placebo values and 3000 unpaired paracetamol values
- I shall then analyse the data

Why this recipe?

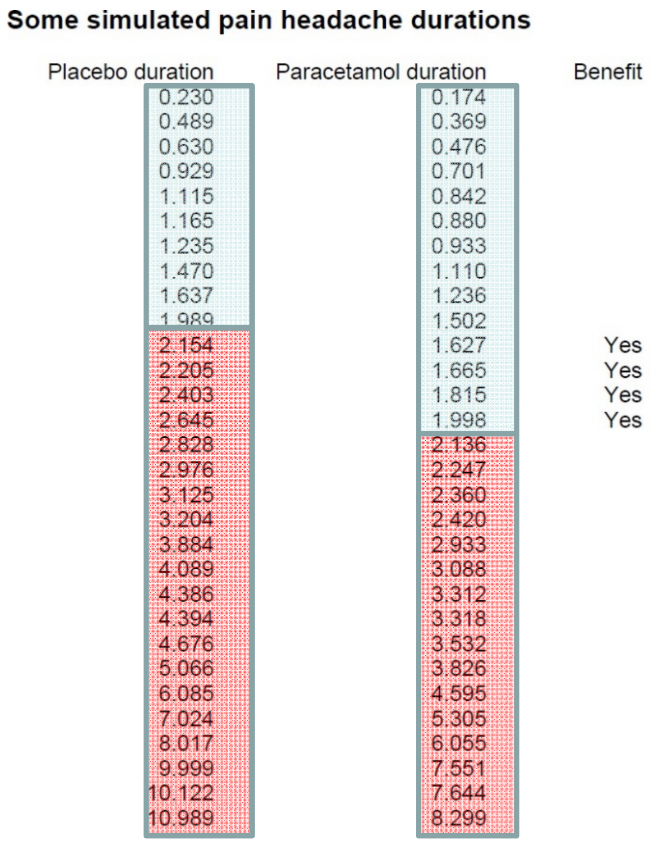
- The exponential distribution with mean 2.970 is chosen so that the probability of response in less than two hours is 0.49
  - This is the placebo distribution
- Rescaling these figures by 0.755 produces another exponential distribution with a probability of response in under two hours of 0.59
  - This is the paracetamol distribution




Counterfactual: pain duration reduced by 1/4
Paracetamol
Placebo
0
1
2
3
4
5
6
2
2
Placebo only
Neither
Both
Paracetamol only


Counterfactual: pain duration reduced by 1/4
Placebo only
Neither
Both
Paracetamol only
2
2
Paracetamol
Placebo
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Dichotomania



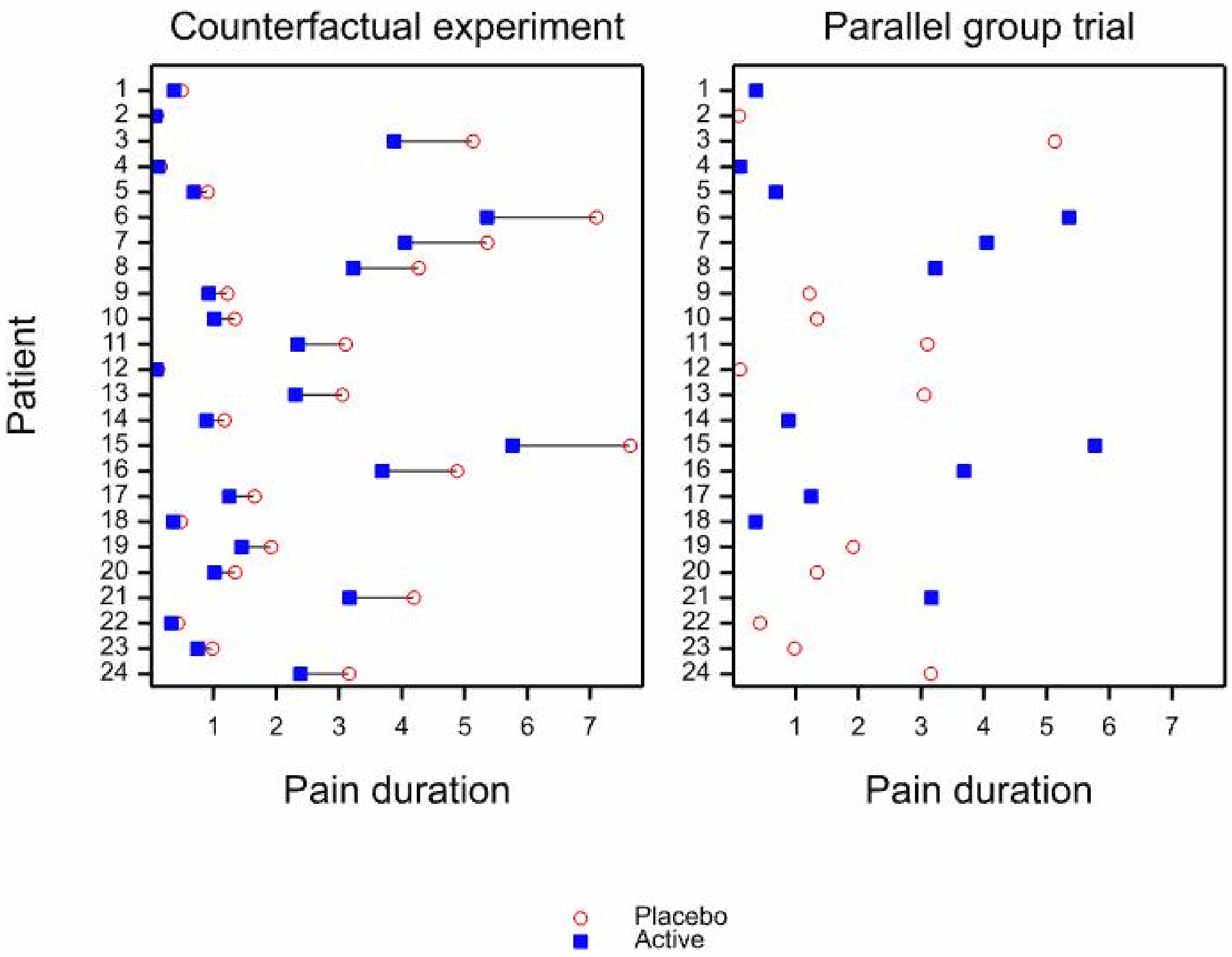
Some simulated pain headache durations

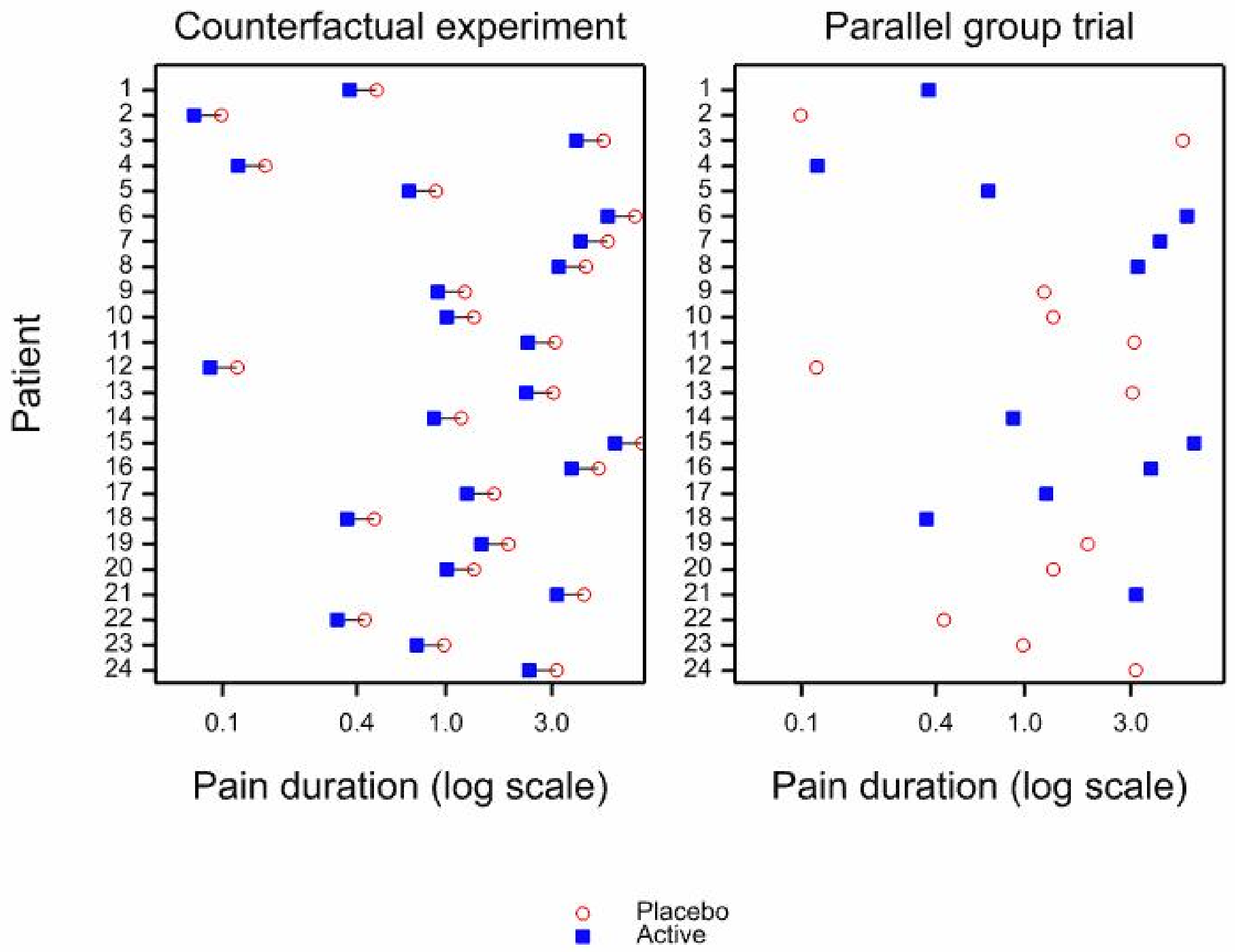
Placebo duration	Paracetamol duration	Benefit
0.230	0.174	
0.489	0.369	
0.630	0.476	
0.929	0.701	
1.115	0.842	
1.165	0.880	
1.235	0.933	
1.470	1.110	
1.637	1.236	
1.989	1.502	
2.154	1.627	Yes
2.205	1.665	Yes
2.403	1.815	Yes
2.645	1.998	Yes
2.828	2.136	
2.976	2.247	
3.125	2.360	
3.204	2.420	
3.884	2.933	
4.089	3.088	
4.386	3.312	
4.394	3.318	
4.676	3.532	
5.066	3.826	
6.085	4.595	
7.024	5.305	
8.017	6.055	
9.999	7.551	
10.122	7.644	
10.989	8.299	

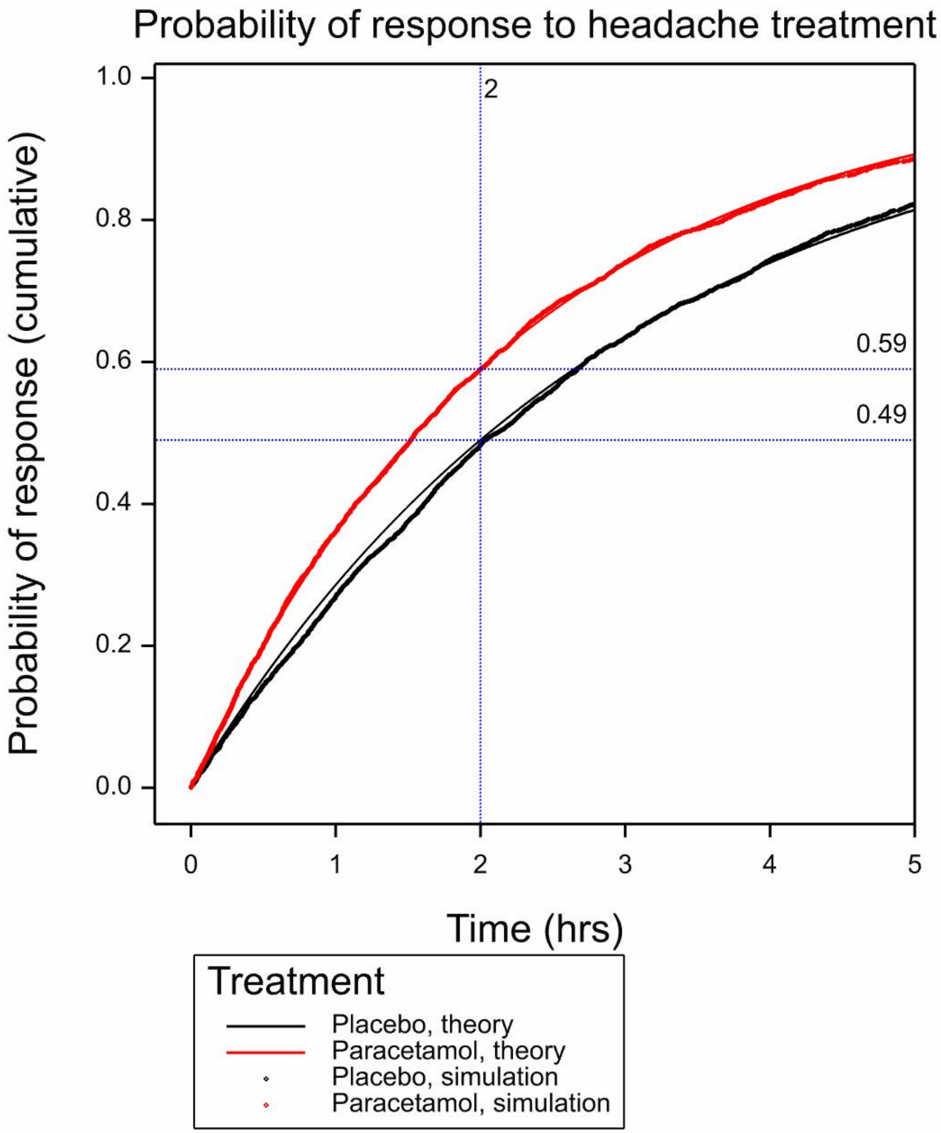
- We lose information through such dichotomies
- We tend to believe our own nonsense labels
  - Response
  - Non-response
- We then delude ourselves that Nature also believes our nonsense
- Next stop: *personalised medicine*

However

- So far I have only gone half way in my simulation recipe
- I have simulated a placebo headache and a corresponding paracetamol headache
- However I can’t treat the same headache twice
- One of the two is *counterfactual*
- I now need to get rid of one member of each factual/counterfactual pair


Counterfactual experiment
Parallel group trial
Patient
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
1
2
3
4
5
6
7
Pain duration
Pain duration
○ Placebo
■ Active


Counterfactual experiment
Parallel group trial
Patient
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
0.1
0.4
1.0
3.0
Pain duration (log scale)
○ Placebo
■ Active


Probability of response to headache treatment
Probability of response (cumulative)
1.0
0.8
0.6
0.4
0.2
0.0
2
0.59
0.49
0
1
2
3
4
5
Time (hrs)
Treatment
Placebo, theory
Paracetamol, theory
Placebo, simulation
Paracetamol, simulation

Summary statistics for Responder: Treatment Placebo

```
Number of observations =   3000
Mean =   0.482
Median =              0
```

Summary statistics for Responder: Treatment Paracetamol

```
Number of observations =   3000
Mean =   0.589
Median =              1
```

To sum up

- The results reported are perfectly consistent with paracetamol having the same effect *on every single headache*
- This does not have to be the case but we don’t know that it isn’t
- The combination of dichotomies and responder analysis has great potential to mislead
- Researchers are assuming that because some patients ‘responded’ in terms of arbitrary dichotomy there is scope for personalised medicine

A previous Prime Minister of the UK speaks

This agreement will see the UK lead the world in genetic research within years. I am determined to do all I can to support the health and scientific sector to unlock the power of DNA, turning an important scientific breakthrough into something that will help deliver better tests, better drugs and above all better care for patients....

David Cameron, August 2014 (my emphasis)

OCTOBER 2013

Paving the Way for Personalized Medicine

FDA's Role in a New Era of Medical Product Development

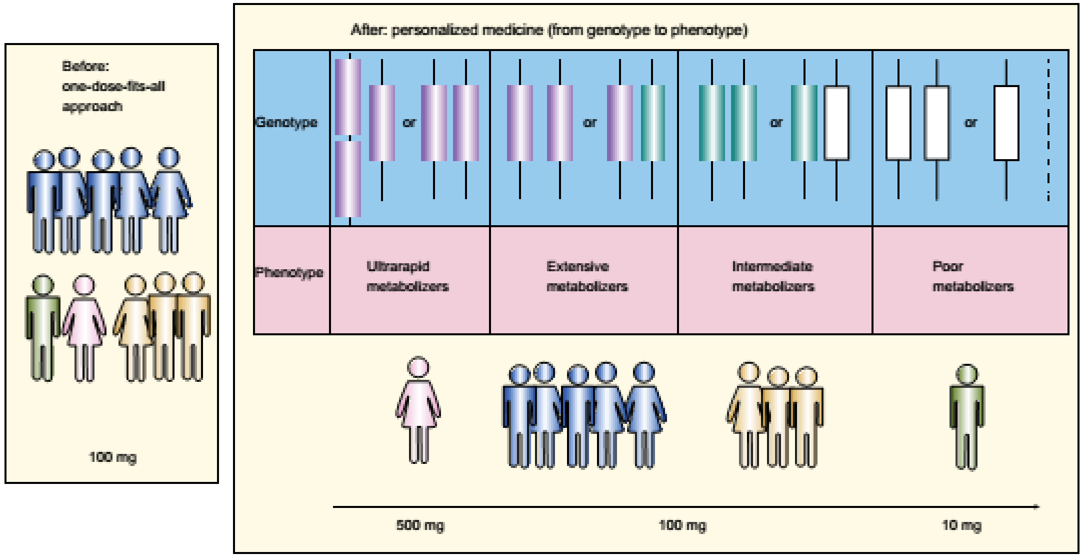
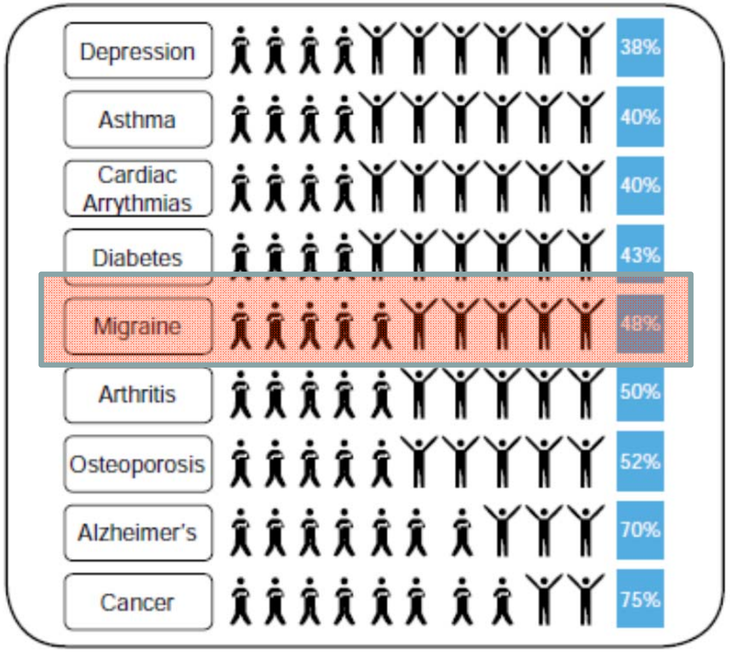


Figure 1. Representation of the trial-and-error or one-dose-fits-all approach versus personalized medicine. The left panel shows a situation in which everyone gets the same dose of a drug, regardless of genotype. The right panel shows a personalized medicine approach in which the dose of the drug is selected based upon genotypical, and therefore phenotypical, variability of the metabolizing enzyme. (Source: Xie, H., Frueh, F.W., (2005). Pharmacogenomics steps toward personalized medicine. *Personalized Medicine*, 2(4), 333.)

Zombie statistics 1

Percentage of non-responders

What the FDA says



Paving the way for personalized medicine, FDA Oct 2013

Where they got it

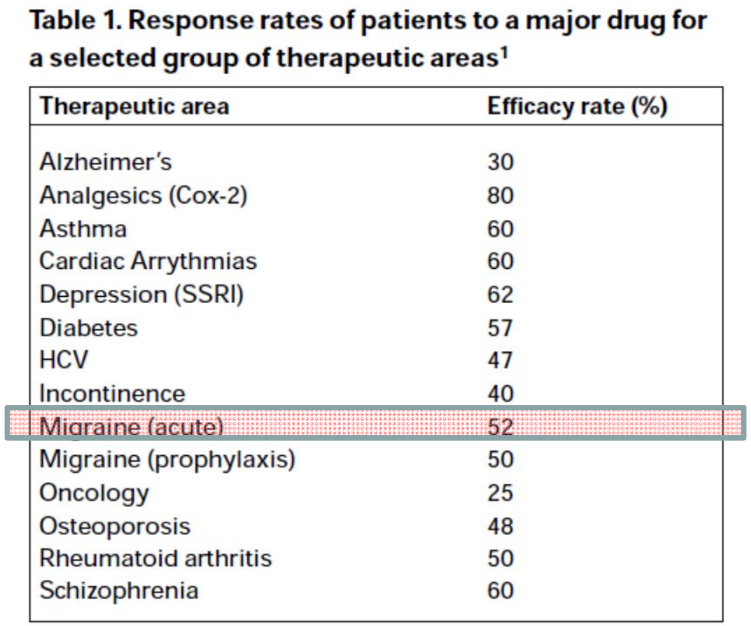


Table 1. Response rates of patients to a major drug for a selected group of therapeutic areas[1]

Therapeutic area	Efficacy rate (%)
Alzheimer's	30
Analgesics (Cox-2)	80
Asthma	60
Cardiac Arrythmias	60
Depression (SSRI)	62
Diabetes	57
HCV	47
Incontinence	40
Migraine (acute)	52
Migraine (prophylaxis)	50
Oncology	25
Osteoporosis	48
Rheumatoid arthritis	50
Schizophrenia	60

Spear, Heath-Chiozzi & Huff, *Trends in Molecular Medicine*, May 2001

Zombie statistics 2

Where they got it

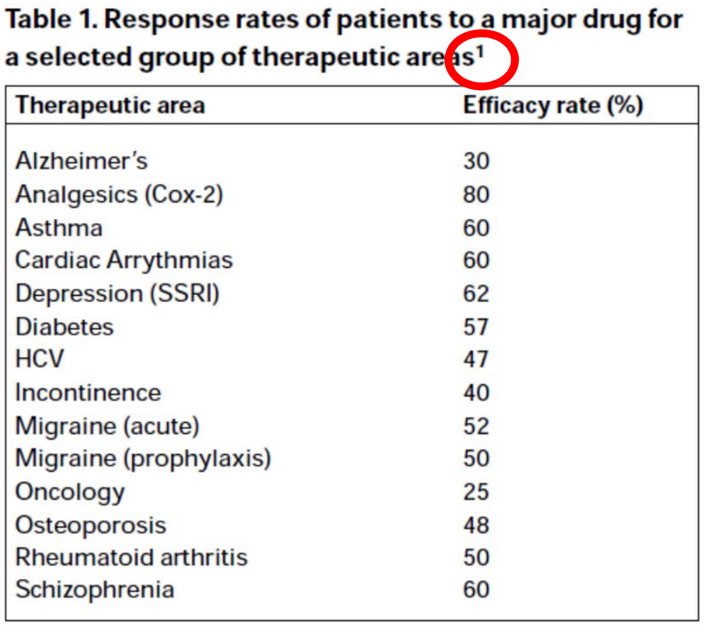
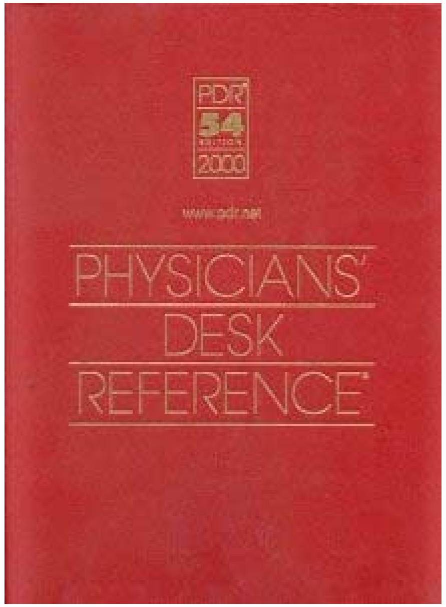


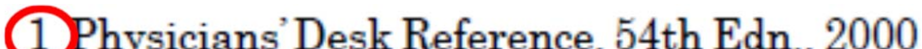
Table 1. Response rates of patients to a major drug for a selected group of therapeutic areas[1]

Therapeutic area	Efficacy rate (%)
Alzheimer's	30
Analgesics (Cox-2)	80
Asthma	60
Cardiac Arrythmias	60
Depression (SSRI)	62
Diabetes	57
HCV	47
Incontinence	40
Migraine (acute)	52
Migraine (prophylaxis)	50
Oncology	25
Osteoporosis	48
Rheumatoid arthritis	50
Schizophrenia	60

Spear, Heath-Chiozzi & Huff, *Trends in Molecular Medicine*, May 2001

Where those who got it got it





1 Physicians' Desk Reference, 54th Edn., 2000

The Real Truth

- These are zombie statistics
- They refuse to die
- Not only is the FDA’s claim not right, it’s not even wrong
- It’s impossible to establish what it might mean even if it were true

88.2% of all statistics are made up on the spot

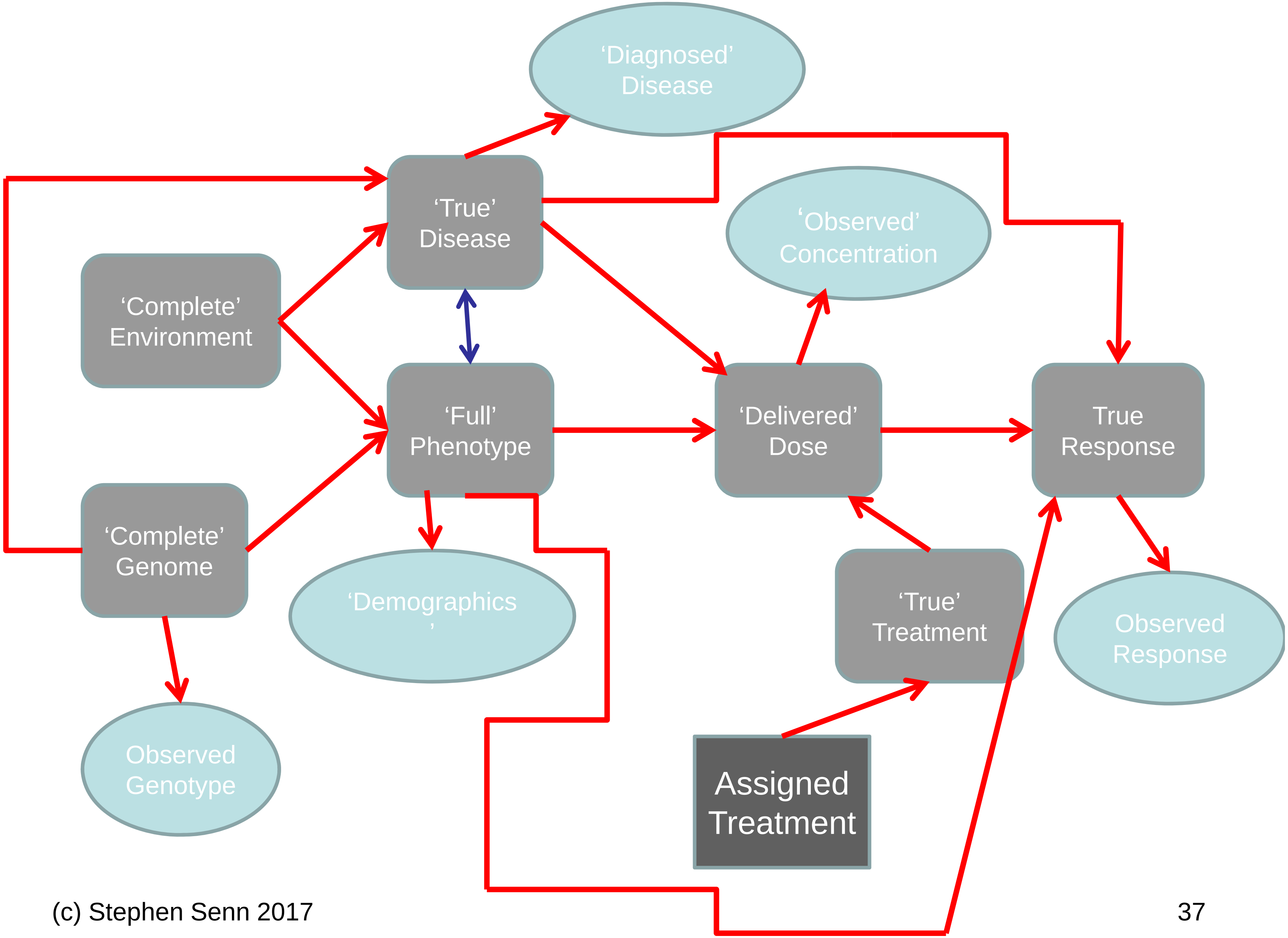
Vic Reeves

The Pharmacogenomic Revolution?

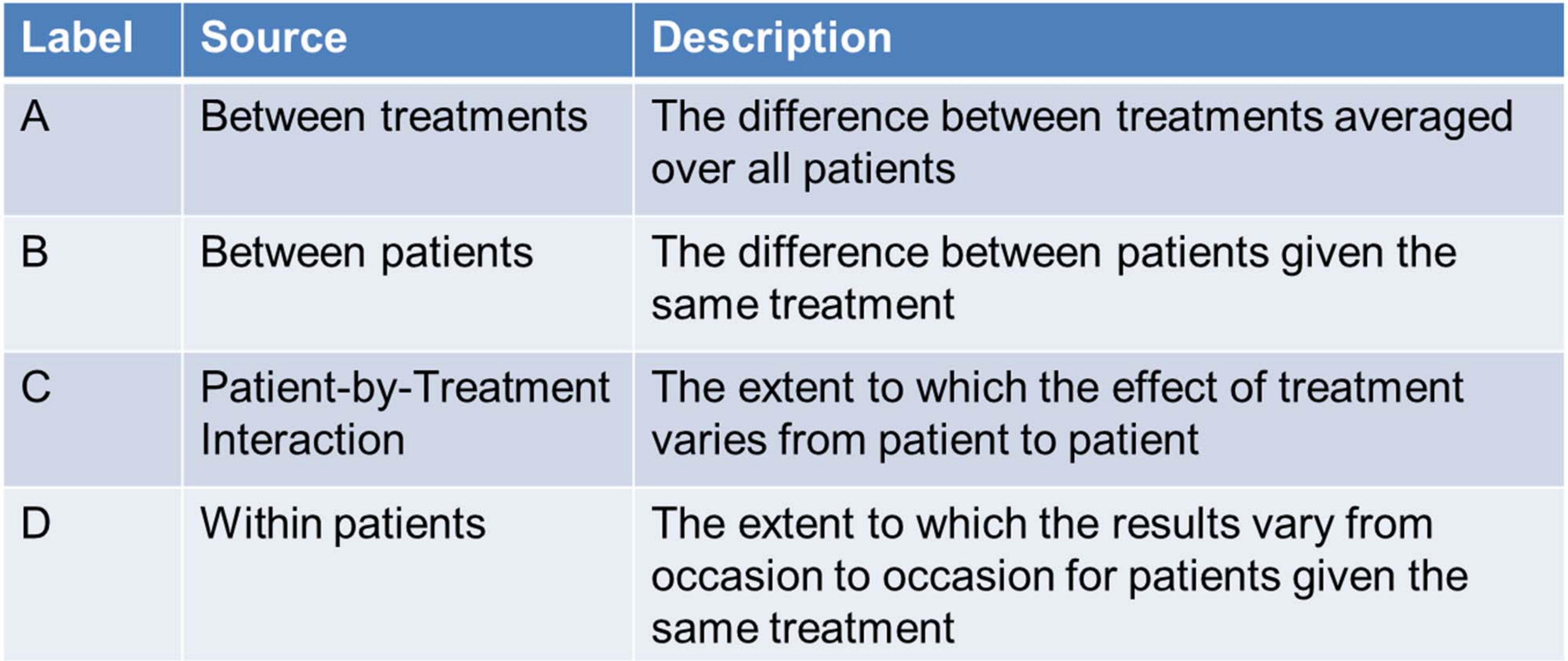
- Clinical trials
  - Cleaner signal
  - Non-responders eliminated
- Treatment strategies
  - “Theranostics”
- Markets
  - Lower volume
  - Higher price per patient day

Implicit Assumptions

- Most variability seen in clinical trials is genetic
  - Furthermore it is not revealed in obvious phenotypes
    - Example: height and forced expiratory volume (FEV_1) in one second
    - Height predicts FEV_1 and height is partly genetically determined but you don’t need pharmacogenetics to measure height
- We are going to be able to find it
  - Small number of genes responsible
  - Low (or no) interactive effects (genes act singly)
  - We will know where to look
- We are going to be able to do something about it
  - May require high degree of dose flexibility
- In fact we simply don’t know if most variation in clinical trials is due to individual response let alone genetic variability


‘Diagnosed’ Disease
‘True’ Disease
‘Observed’ Concentration
‘Complete’ Environment
‘Full’ Phenotype
‘Delivered’ Dose
True Response
‘Complete’ Genome
‘Demographics’
‘True’ Treatment
Observed Response
Observed Genotype
Assigned Treatment

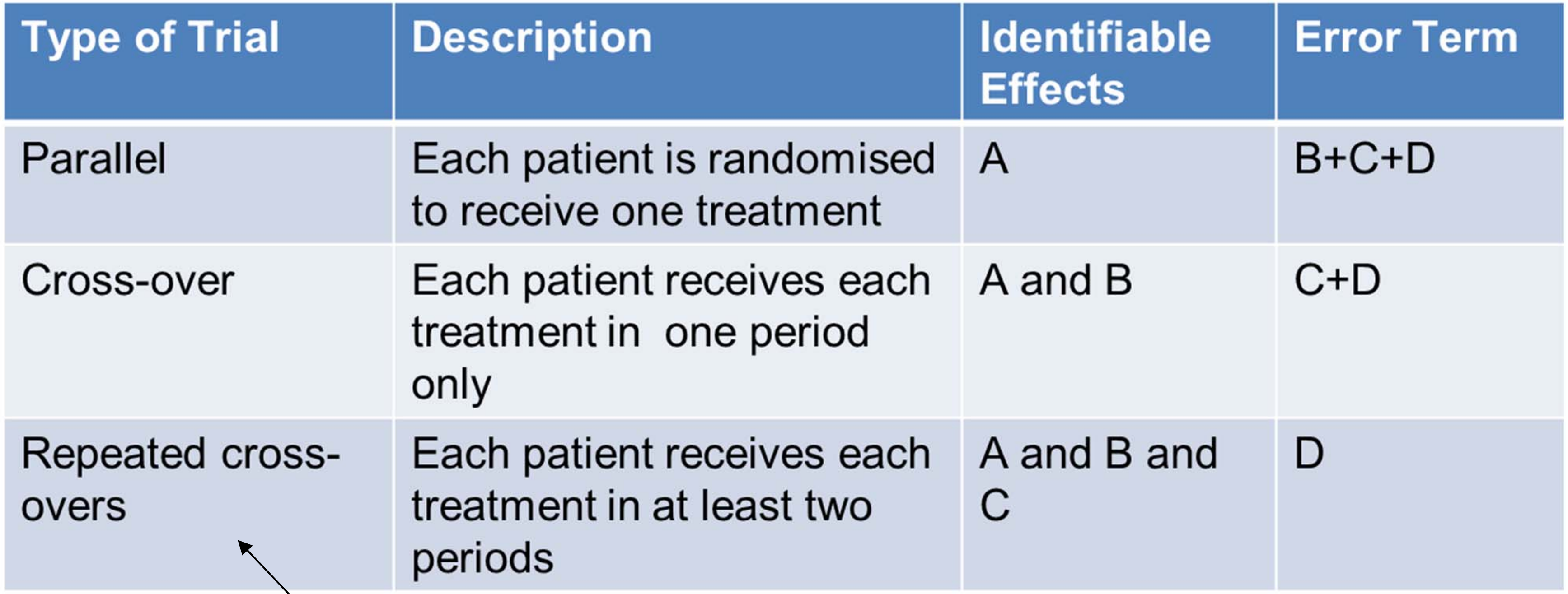
Sources of Variation in Clinical Trials



Label	Source	Description
A	Between treatments	The difference between treatments averaged over all patients
B	Between patients	The difference between patients given the same treatment
C	Patient-by-Treatment Interaction	The extent to which the effect of treatment varies from patient to patient
D	Within patients	The extent to which the results vary from occasion to occasion for patients given the same treatment

Senn SJ. Individual Therapy: New Dawn or False Dawn. *Drug Information Journal* 2001;35(4):1479-1494.

Identifiability and Clinical Trials

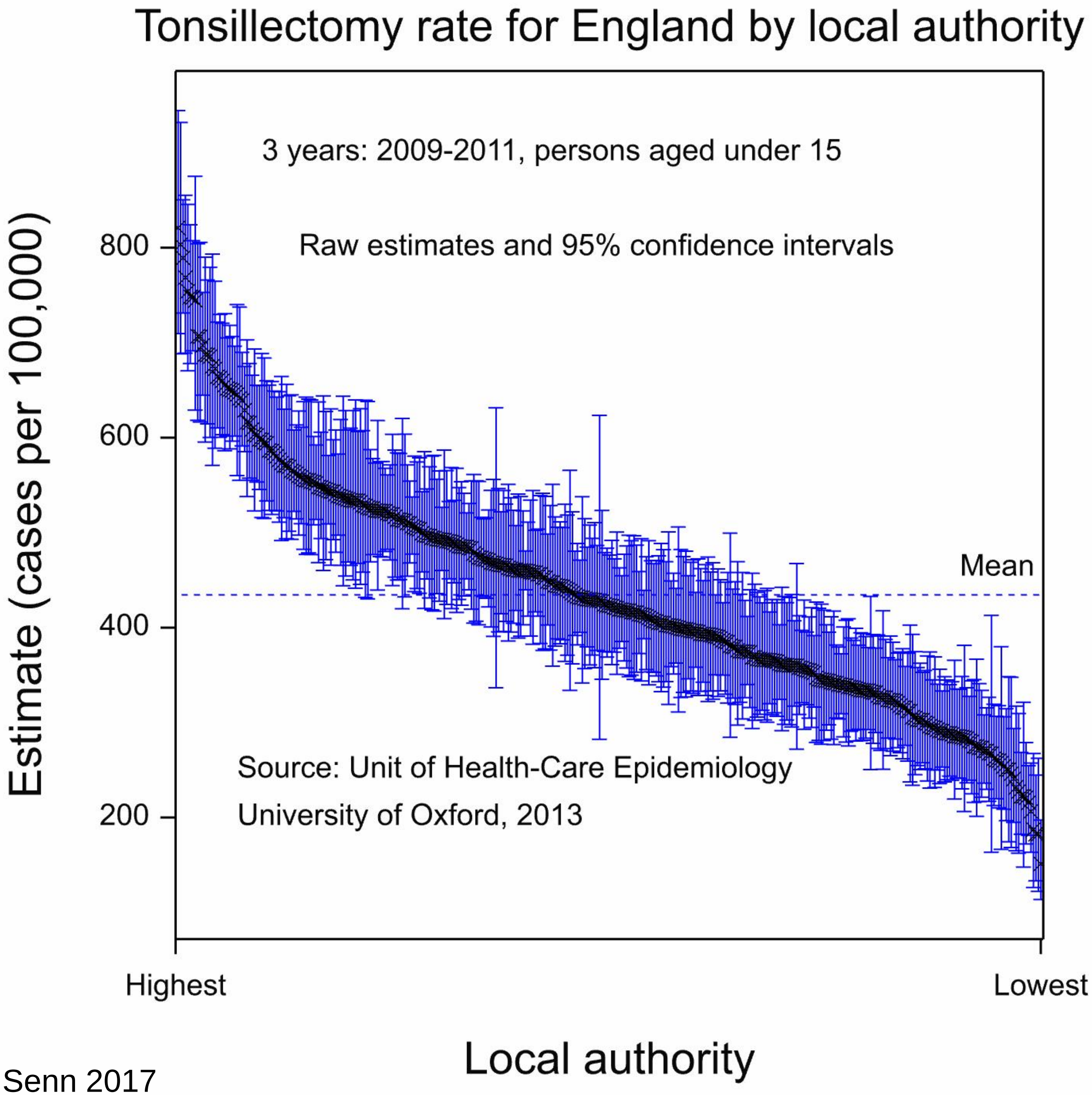


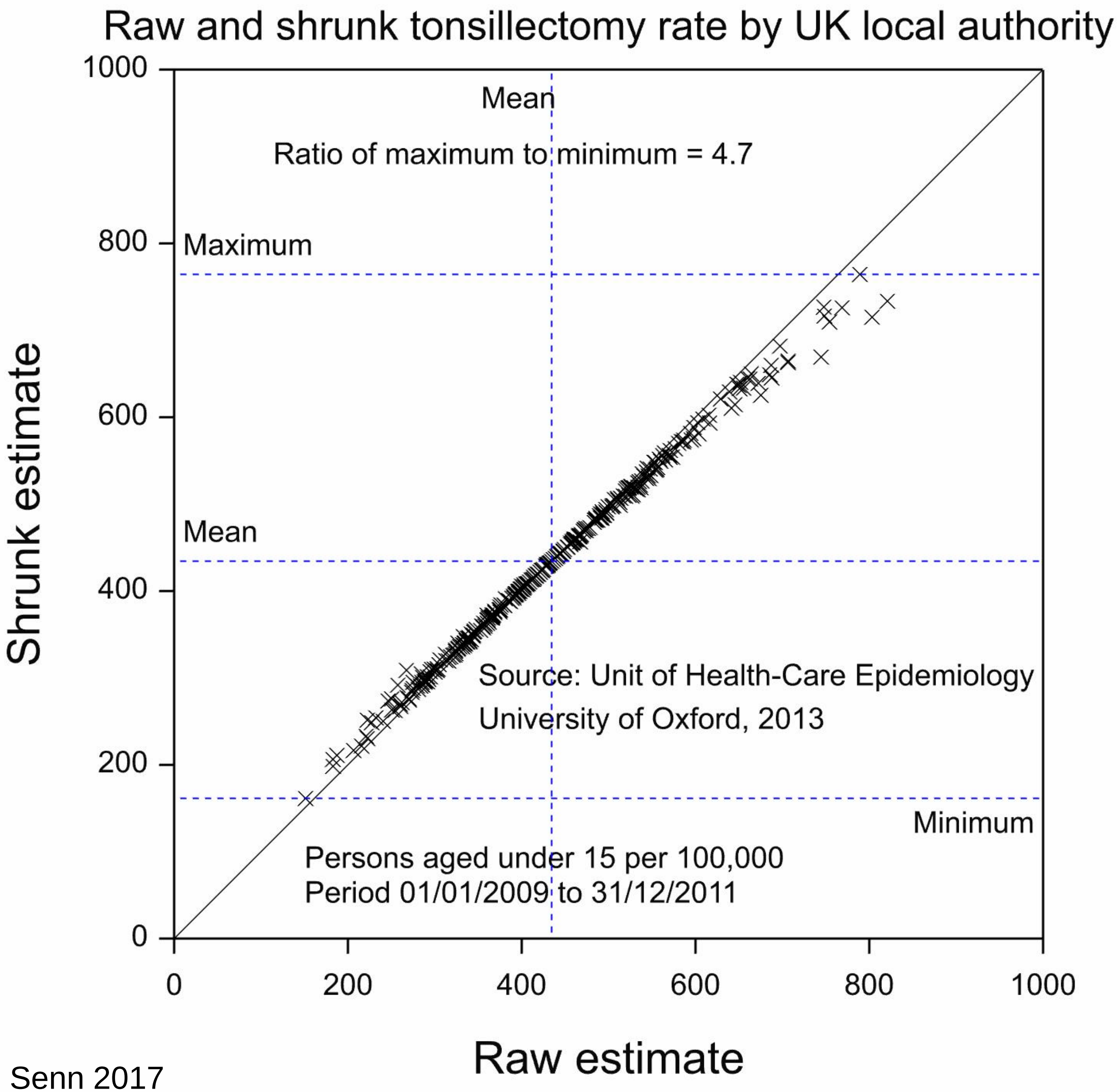
Type of Trial	Description	Identifiable Effects	Error Term
Parallel	Each patient is randomised to receive one treatment	A	B+C+D
Cross-over	Each patient receives each treatment in one period only	A and B	C+D
Repeated cross-overs	Each patient receives each treatment in at least two periods	A and B and C	D

Also known as *n of 1 trials*
See StatSols blog
http://blog.statsols.com/making-it-personal-n-of-1-trials-allowing-for-individuality-but-not-overdoing-it

In the Meantime

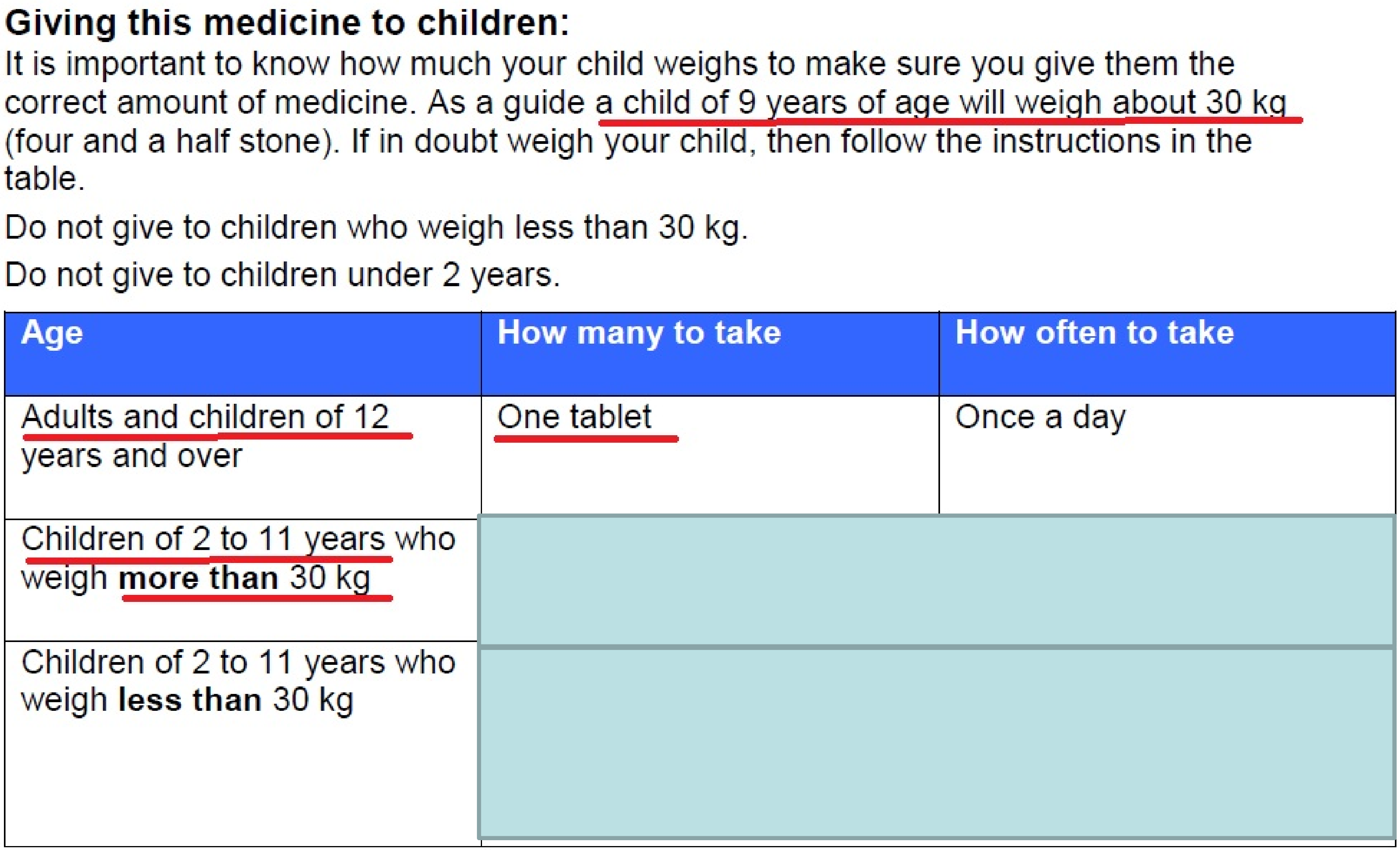
- There is a massive source of unwanted variation
- Doctors
- Variation in practice is so large that it cannot be justified by variation in patients
- This is the basic idea behind the way that Intermountain Health under the leadership of Brent James has been applying Deming’s principles to health care


Tonsillectomy rate for England by local authority
3 years: 2009-2011, persons aged under 15
Raw estimates and 95% confidence intervals
Mean
Source: Unit of Health-Care Epidemiology
University of Oxford, 2013
Estimate (cases per 100,000)
800
600
400
200
Highest
Lowest
Local authority


Raw and shrunk tonsillectomy rate by UK local authority
Mean
Ratio of maximum to minimum = 4.7
Maximum
Mean
Source: Unit of Health-Care Epidemiology
University of Oxford, 2013
Minimum
Persons aged under 15 per 100,000
Period 01/01/2009 to 31/12/2011
Shrunk estimate
0
200
400
600
800
1000
Raw estimate
0
200
400
600
800
1000

“Guys, it’s more important that you do it the same way than what you think is the right way.”

Brent James, Advice to doctors



Giving this medicine to children:

It is important to know how much your child weighs to make sure you give them the correct amount of medicine. As a guide a child of 9 years of age will weigh about 30 kg (four and a half stone). If in doubt weigh your child, then follow the instructions in the table.

Do not give to children who weigh less than 30 kg.

Do not give to children under 2 years.

Age	How many to take	How often to take
Adults and children of 12 years and over	One tablet	Once a day
Children of 2 to 11 years who weigh **more than** 30 kg		
Children of 2 to 11 years who weigh **less than** 30 kg		

Advice

- Don’t let the label ‘responder’ infect your brain
- A ‘responder’ is a patient who was *observed* to get better by some arbitrary standard
- A ‘responder’ is not a patient who was *caused* to get better by the drug
- Subsequence is not consequence
- To establish who really responds and who does not you need to work very hard

Conclusion

- We have done very well in handling some aspects of variation in clinical trials
- However, it is high time we did better in investigating the sources of variation
- Mastering variation is the key to high quality medicine

The supply of truth always greatly exceeds its demand

John F Moffitt