Variances as ends not means: designing to understand variation

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Stephen Senn

Acknowledgements

My thanks to Axel Krebs-Brown for organizing this and to the EFSPI for support

My thanks to Andreas Laupacis and Jennifer Deevy for providing me with a copy of the classic paper on NNTs by Laupacis et al (1988)

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"

Disclaimers

These are my personal opinions and should not be attributed to the ASA, the LIH or, indeed any colleagues or other parties

I shall pick up some examples of problems to illustrate the point.

Just as a careless use of P-values doe not indicate that all the science in the paper in question is bad, a careless use of NNTs (or other statistics claiming to indicate personal response) does not indicate that the paper it appeared in is bad.

In some cases the problems are actually with publicity given to a paper rather than the content of the papers

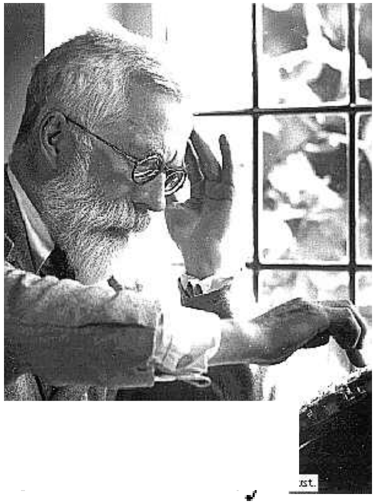
The difficulty of statistics is regularly underestimated by everybody, including me.

Outline

The statistician's motto

- Two anniversaries
  - Variance 100 years
  - NNT 30 years
- How we are misunderstanding variation
- How we can do it better
- What statisticians need to do

Fisher’s great paper of 1918



of the mean square error. When there are two independent causes 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

XV.—**The Correlation between Relatives on the Supposition** of Mendelian Inherit-**ance.** By **R. A. Fisher,** B.A. *Communicated by* Professor J. ARTHUR THOMSON. (With Four Figures in Text.)

Laupacis, Sackett, Roberts, 1988

- Considers the problems of physicians trying to prioritise
- Investigates various measures
- Proposes the reciprocal of the risk reduction
  - Calls this the Number Needed to be Treated
    - Now often called Number Needed to Treat (NNT)
- Discusses various caveats

The Caveats

- Combining baseline risk and treatment effect in one measure can be misleading
- RCTs are often expressed in disease specific terms but NNTs need overall benefit and harm
- Compliance is a practical issue for the physician (and patient!)
- Variation from trial to trial
- Different lengths of follow-up, different NNTs
- Some treatments do not begin to be effective until long after they are started

What is an NNT?

The number needed to treat (NNT) is an epidemiological measure used in communicating the effectiveness of a health-care intervention, typically a treatment with medication. The NNT is the average number of patients who need to be treated to prevent one additional bad outcome (e.g. the number of patients that need to be treated for one of them to benefit compared with a control in a clinical trial). It is defined as the inverse of the absolute risk reduction. It was described in 1988 by McMaster University's Laupacis, Sackett and Roberts. The ideal NNT is 1, where everyone improves with treatment and no one improves with control. The higher the NNT, the less effective is the treatment.

Wikipedia entry consulted 15 February 2018
https://en.wikipedia.org/wiki/Number_needed_to_treat

What Wikipedia should say is…

The NNT is the average number of patients who need to be treated to prevent one additional ‘bad’ outcome, where such an outcome is often an arbitrary dichotomy that may partly vary randomly from patient to patient and for a given patient from occasion to occasion

Anybody familiar with the notion of “number needed to treat” (NNT) knows that it's usually necessary to treat many patients in order for one to benefit. NNTs under 5 are unusual, whereas NNTs over 20 are common.

Richard Smith, *BMJ*, 13 December 2003

(Richard Smith was the editor of the *BMJ* for many years and remains a very interesting commentator of medicine and health.)

Cochrane UK @CochraneUK

Following

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

uk.cochrane.org/news/featured- …

RETWEETS 20 LIKES 3

59% had no or at worst mild headache after 2 hours when treated with paracetamol

49% had no or at worst mild headache after 2 hours when treated with placebo

59%-49% = 10%

Therefore 10% benefitted

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

Based on a review of 23 studies and 6000 patients

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?

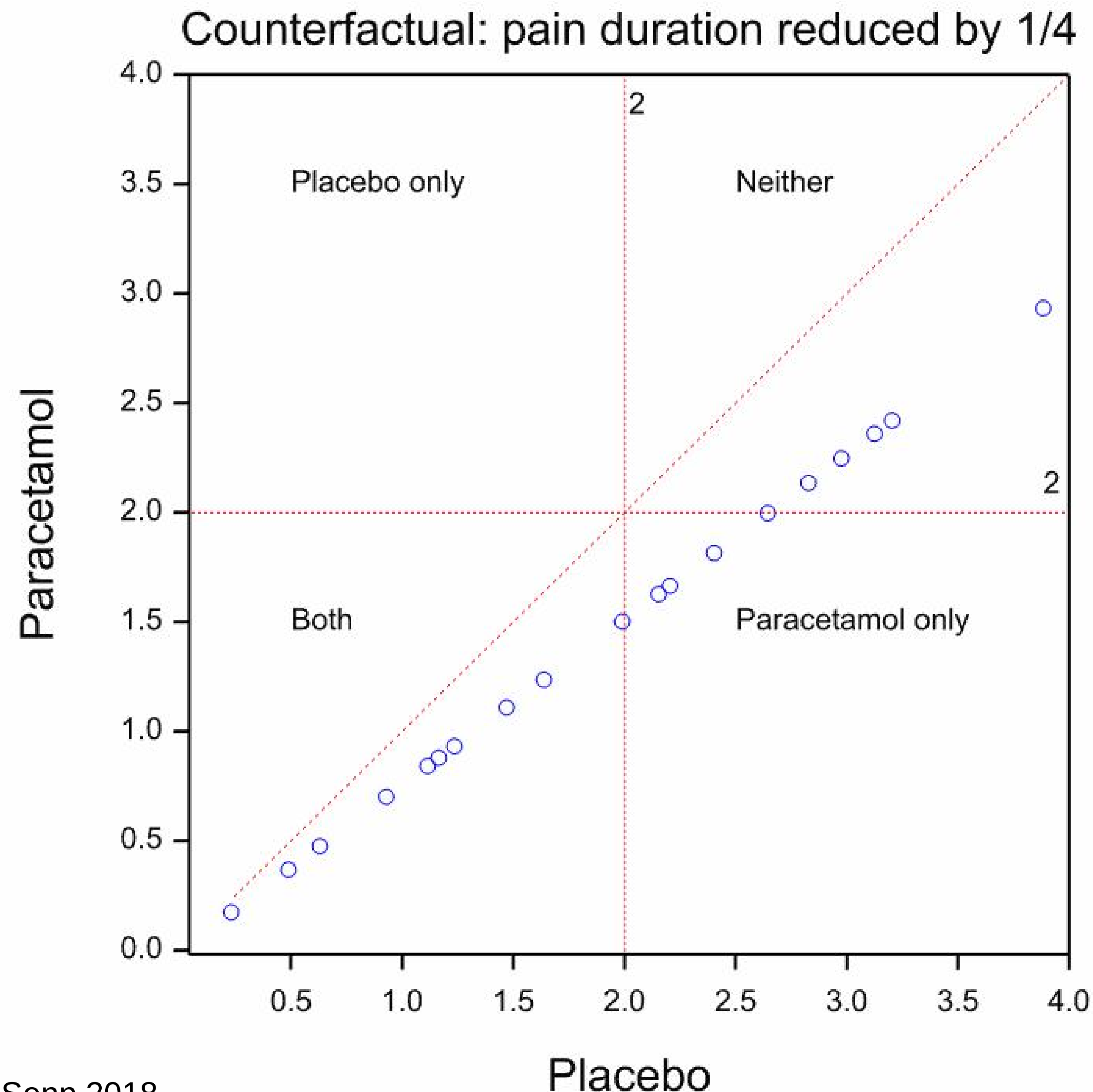
The International Headache Society recommends the outcome of being pain free two hours after taking a medicine.

(So does the FDA as regards migraine)

Mr Jones responded. Mrs Smith didn’t.

A Recipe to Mimic the Cochrane Results

- Generate one random number, U_i, for each of 6000 headaches, $i = 1,2\ldots6000$
- Calculate pairs of headaches
  - Y_{i1} =-log(U_i)2.97 (placebo headache duration)
  - Y_{i2} =-log(U_i)2.24 (paracetamol headache duration)
- Now randomly erase one member of each pair
  - – Because headache can only receive one treatment
  - – The other is counterfactual
- Draw the empirical cumulative distribution


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

Dichotomania

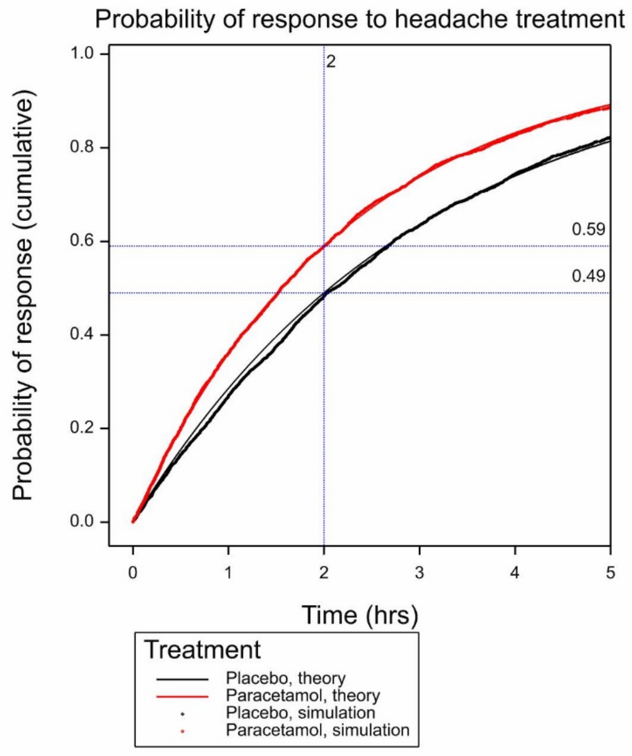
Some simulated pain headache durations

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 this

Why this recipe?

- The exponential distribution with mean 2.97 is chosen so that the probability of response in under two hours is 0.49
  - This is the placebo distribution
- The exponential distribution with mean 2.24 is chosen so that the probability of response in under two hours of 0.59
  - This is the paracetamol distribution
- This is what you would see if *every* headache were reduced to the same degree (about ¼)
- It is also mimics exactly the Cochrane result



Lessons

Particular

- The NNT of 10 is perfectly compatible with paracetamol having *exactly the same proportionate effect on every headache*
- Nothing in the data we are given says anything whatsoever about differential response

In general

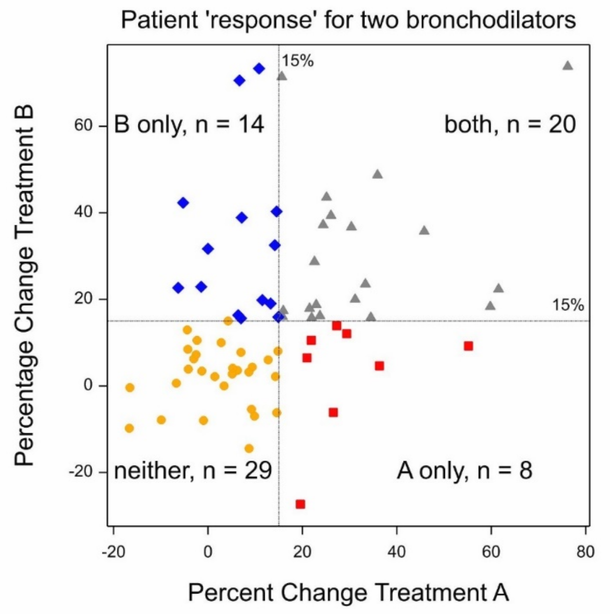
- An NNT cannot tell you what proportion of patients responded
- To think so is a straightforward conceptual mistake
- Claims regarding the proportion who respond based on NNTs are misleading

Putting my cards on the table

- Medicine has always involved differentiation and personalisation based on diagnosis
  - Type I v Type II diabetes
  - Pneumonia vs tuberculosis
- Progress will continue to involve this
- Nevertheless, methodological errors are being made in understanding variation
- To correct this misunderstanding will involve clear thinking, clever design, and good analysis
- The people to deliver this are statisticians

The case for personalised medicine

- Cross-over trial in asthma
  - 71 patients
  - Forced expiratory volume in one second (FEV_1) at 12 hours
- FDA definition of response is ≥ 15% increase compered to baseline
- 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
- A green light for personalised medicine

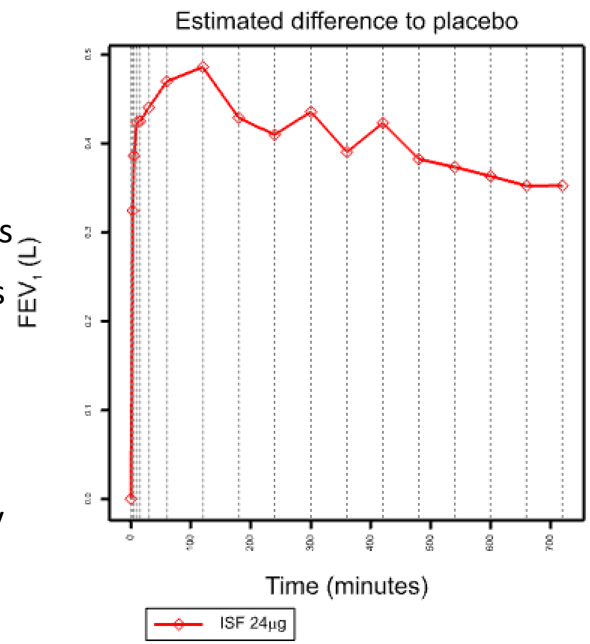


Some further details
A complex design in asthma comparing two formulations of formoterol

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

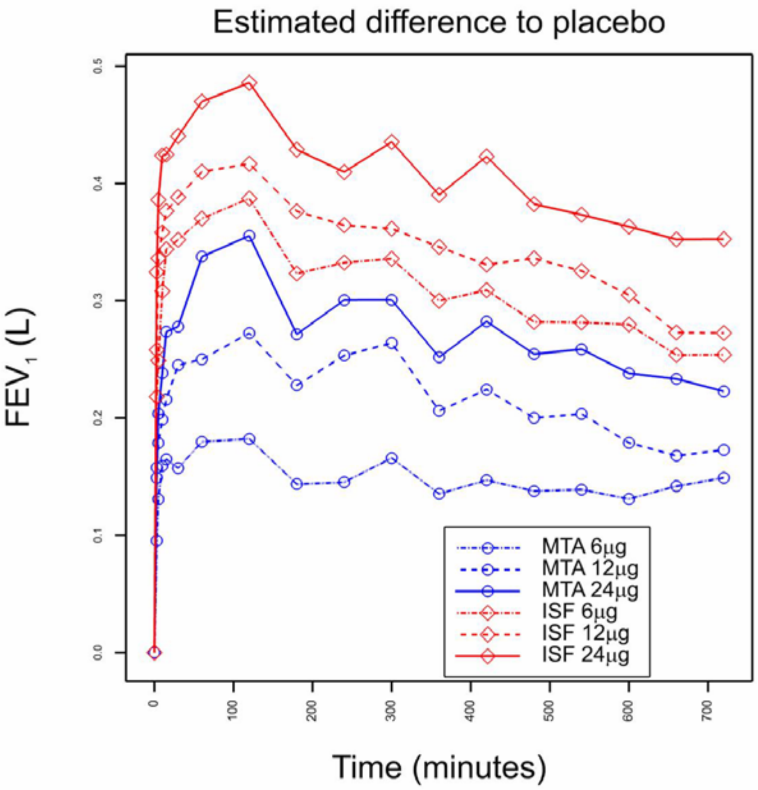
Senn, S. J., Lillienthal, J., Patalano, F., & Till, M. D. (1997). An incomplete blocks cross-over in asthma: a case study in collaboration. In J. Vollmar & L. A. Hothorn (Eds.), *Cross-over Clinical Trials (pp. 3-26). Stuttgart: Fischer.*

- 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



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



Implications

As regards comparing formulations

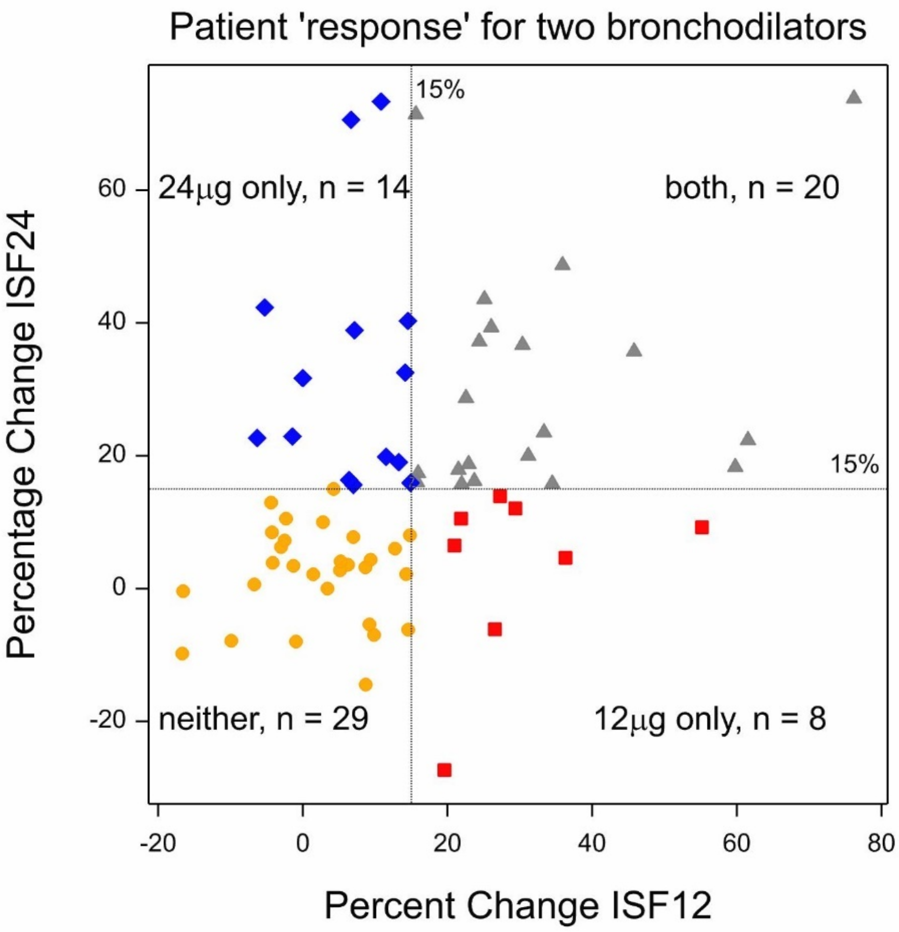
- The formulations are clearly not equipotent
- The difference between formulations is a great as the difference between doses
- A careful complicated design killed the new formulation

But there is more

- The fact that patients have been measured many times enables us to say something about individual response
- Let's go back to the FDA's (not very sensible) definition of response
  - 15% increase in FEV_1 above baseline
- Now look again at 'responders' 12 hours after treatment for two of the formulations…

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
- What can the explanation be?
- Large within patient variability
- Conclusion: naïve simple views of causality and response aren't good enough and more complex design and analysis is needed

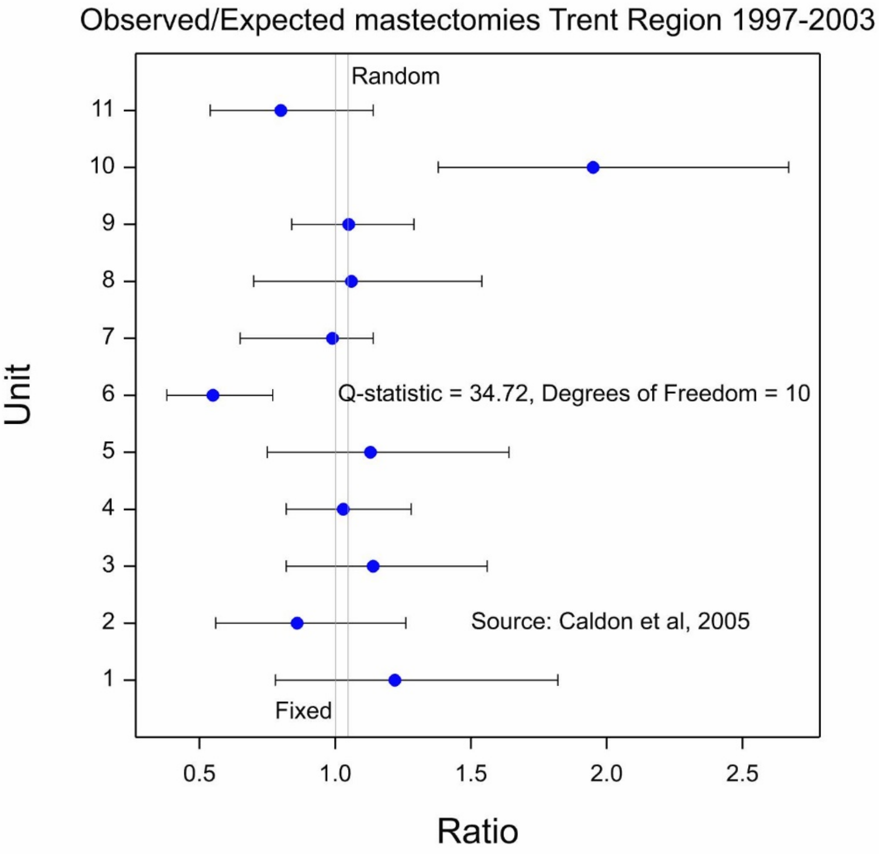


How responder analysis misleads us:
Six depressingly common sins

- Poor choice of counterfactual
  - Baseline does not necessarily predict what would happen in the absence of treatment
- Bad measures
  - Percent change from baseline is known to be a highly variable and badly behaved measure
- Arbitrary dichotomy
  - There is nothing magic about 15% and dichotomising loses information
- Linguistic confusion
  - *Responder* does not mean 'was *caused* to improve' it means 'was *observed* to improve'
- Causal naivety
  - Subsequence is not consequence
- Failure to replicate
  - If you want to exclude within-patient variability as an explanation you have to know how big it is. That involves measuring patients more than once

Case-mix fails to explain variation in mastectomy rates: management of screen-detected breast cancer in a UK region 1997–2003 (Caldon et al 2005)

- 1997-2003
- 11 units
- 5109 cases and 1829 mastectomies
- Significant variation in rates between units p<0.00001
- Not explained by case mix p<0.00001
- 2-fold variation ratio observed to expected and 4-fold for cancers < 15mm diameter



The central problem in management and leadership is failure to understand the information in variation. Lloyd S Nelson (quoted by WE Deming)

- As Deming, the guru of quality control taught us, it is the duty of every manager to understand the variation in the system
- At the moment we are making a bad job of this
- Our goal should be *superb* medicine
- Personalised medicine is a *means* to help us achieve this goal but it is not the *goal*
- We need to build personalised medicine on top of excellent average medicine
- This requires developing evidence-based guidelines and encouraging physicians to use them

Sources of Variation in Clinical Trials

Label	Source	Description
A	Between treatments	The average difference between patients over all treatments and all randomisations
B	Between patients	The difference between patients given the same treatment. (Averaged over both experimental and control treatments.)
C	Patient by treatment interaction	The extent to which the 'true' difference between treatments will vary from patient to patient. (Equivalently the extent to which the difference between patients will vary from treatment to treatment.)
D	Pure within patient error	The extent to which the 'response' would vary for a given patient when given the same treatment on different occasions. (Averaged over all patients and both treatments.)

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

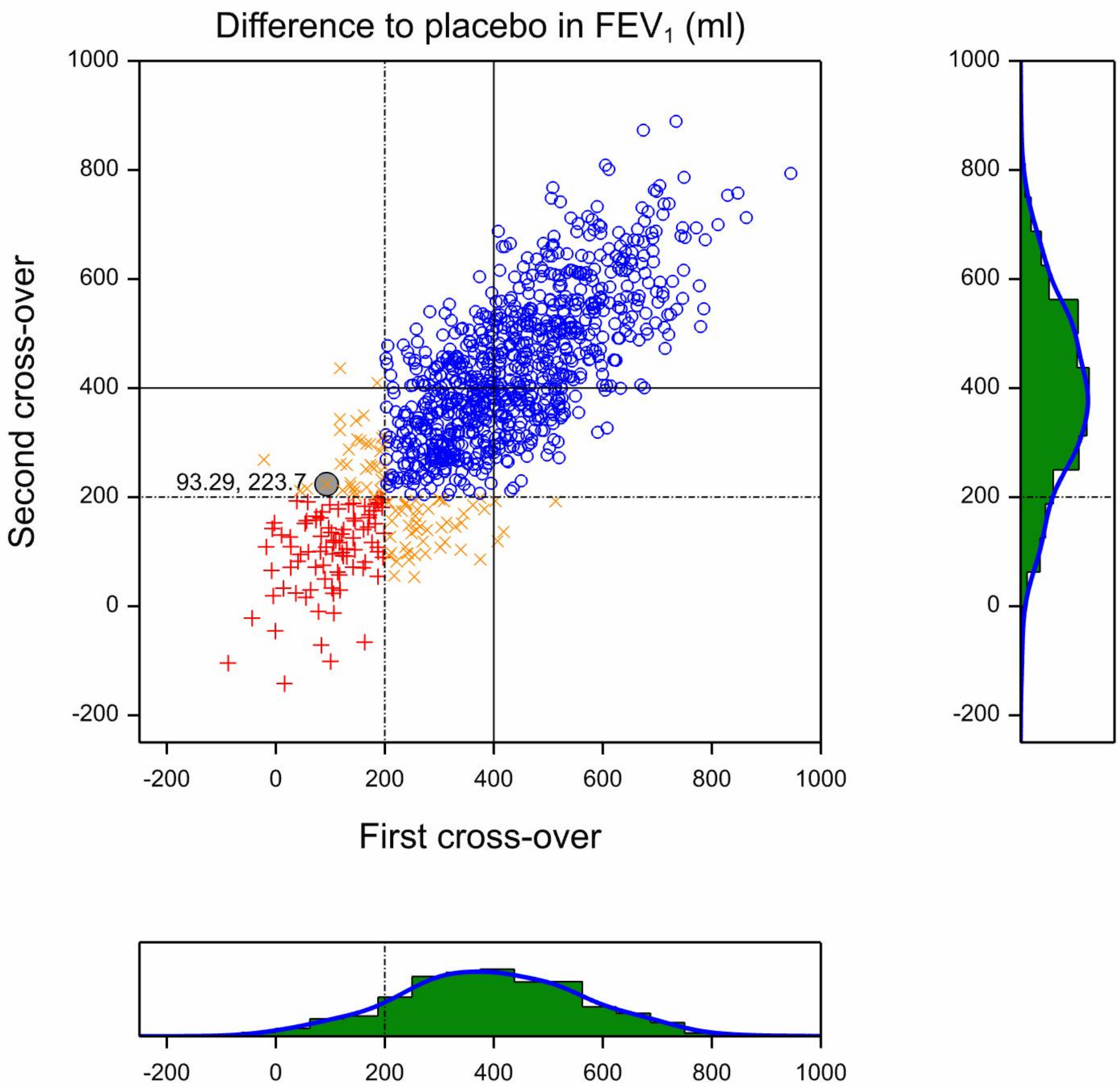
Identifiability and Clinical Trials

Type of Trial	Description	Identifiable effects	Error term
Parallel	Each patient receives one treatment	A	B+C+D
Cross-over	Each patient receives each treatments	A+B	C+D
Repeated cross-over	Each patient receives each treatment at least twice	A+B+C	D

A design to do better
Repeated cross-over
Also known as n-of-1 trials

- We allocate patients randomly to receive either A followed by B or vice versa
- Then we repeat this
- Patients will be allocated to one of the four sequences on the right

	First Cross-over		Second Cross-over	
	Period			
Sequence	1	2	3	4
I	A	B	A	B
II	B	A	B	A
III	A	B	B	A
IV	B	A	A	B



Difference to placebo in FEV1 (ml)
Second cross-over
First cross-over
93.29, 223.7
1000
800
600
400
200
0
-200
-200
0
200
400
600
800
1000

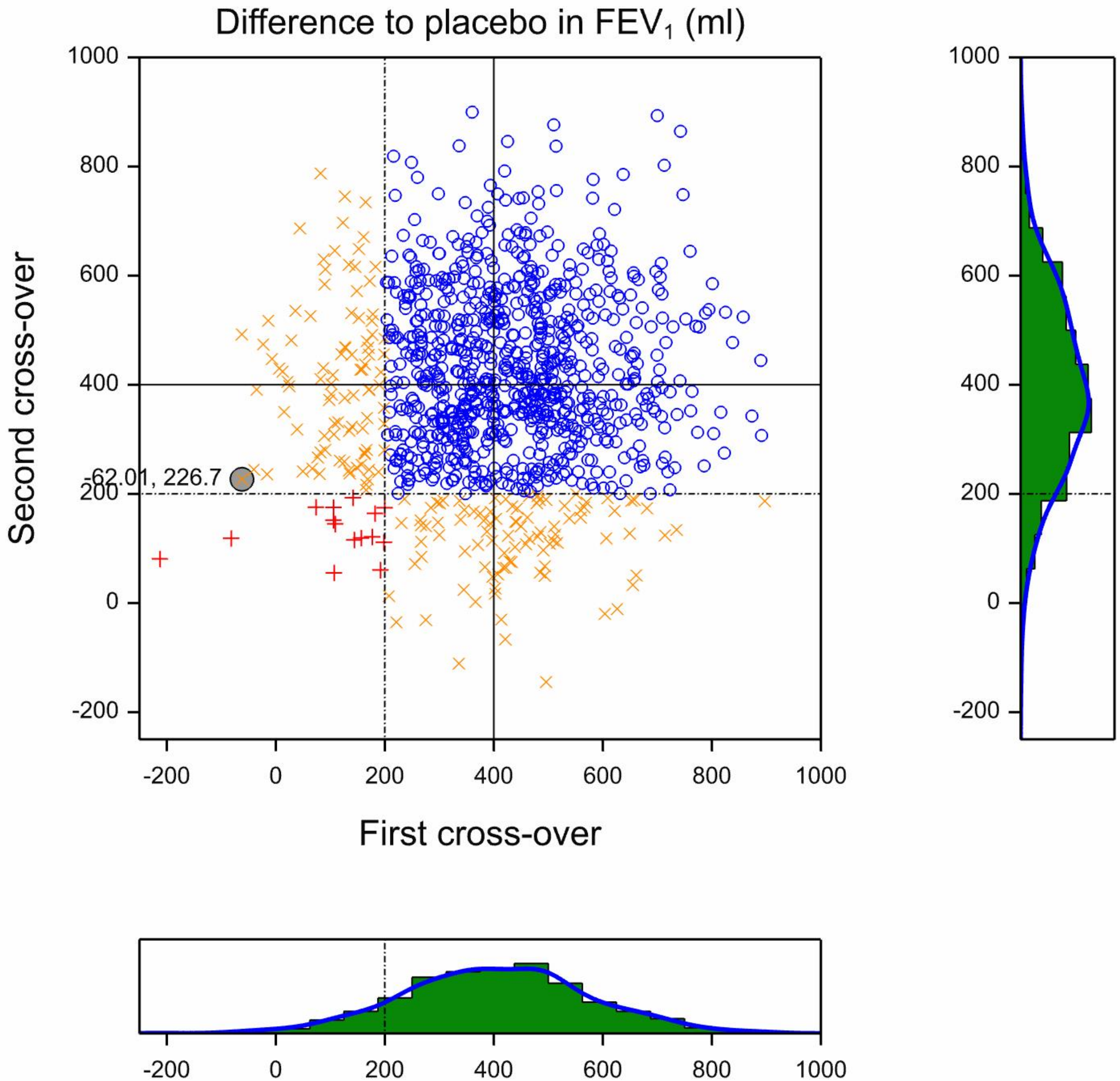
Results 1

		Second Cross-over		
		Responder	Non-responder	Total
First Cross-Over	Responder	827	49	876
	Non-responder	38	86	124
	Total	865	135	1000

correlation coefficient is 0.8

Conditional probability of "response" in 2^{nd} cross-over

827/876=0.94 if responded1st
38/124=0.31 if did not respond 1st



Difference to placebo in FEV_1 (ml)
Second cross-over
First cross-over
-62.01, 226.7
1000
800
600
400
200
0
-200
-200
0
200
400
600
800
1000

Results 2

		Second Cross-over		
		Responder	Non-responder	Total
First Cross-Over	Responder	797	95	892
	Non-responder	93	15	108
	Total	865	135	1000

correlation coefficient is 0.0.2

Conditional probability of "response" in 2nd cross-over

797/892=0.89 if responded1st
93/108=0.86 if did not respond 1st

Personal View

Improving average medicine

- We are not doing average medicine well
- It's healthcare delivery by doctors that is currently the biggest problem
  - Identify the most important EBM guidelines
  - Oversee their implementation in medicine
  - Monitor the individual results

Personalising for further improvement

- Designing clinical trials to identify components of variation
- Identification of those diseases where it will make the biggest difference
- Personalisation where it can

RA Fisher going on a random walk and hitting an absorbing barrier