Mastering variation: Variance components and personalised medicine

Stephen Senn

crp SANTÉ
CENTRE DE RECHERCHE PUBLIC
COMPETENCE CENTER FOR METHODOLOGY AND STATISTICS

Acknowledgements

Many thanks 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"

but not necessarily at University Challenge

Genes, Means and Screens

It will soon be possible for patients in clinical trials to undergo *genetic tests* to identify those *individuals who will respond favourably to the drug* candidate, *based on their genotype*…. This will translate into *smaller, more effective clinical trials* with corresponding *cost savings* and ultimately better treatment in general practice. … *individual patients will be targeted* with specific treatment and personalised dosing regimens to maximise efficacy and minimise pharmacokinetic problems and other side-effects.

Sir Richard Sykes, FRS, 1997

My emphasis

Basic thesis

- Both sides of the regulatory divide are convinced that there is a strong element of personal response to treatment
- The truth is that nobody knows
- This is because we statisticians have failed to teach others about components of variation
- And some of us have failed to learn about components of variation also

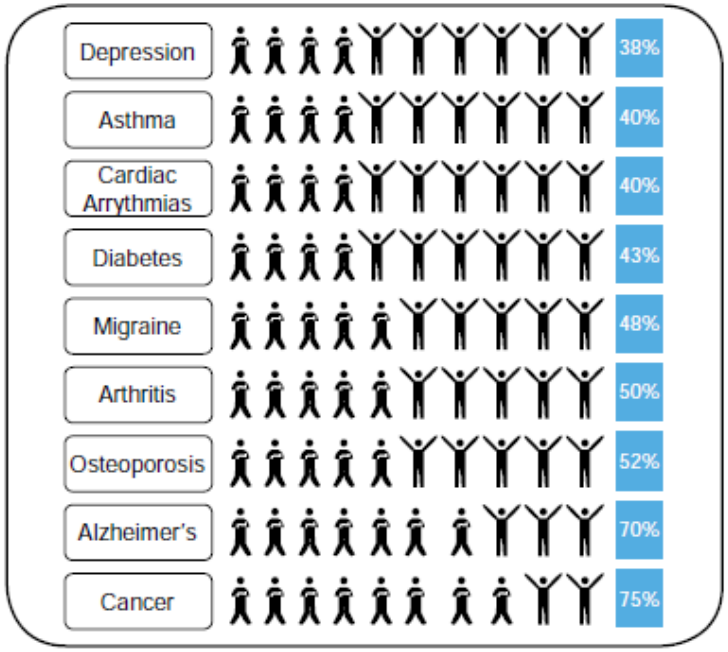
Seven key aspects

1. Better communication of the problems by statisticians to their colleagues
2. *Application of decision analysis to determine when personalisation is worth pursuing*
3. Appropriate design for teasing out components of variation
4. Application of random effect methodology for improving estimates
5. Translating from additive to relevant scales
6. Application of Deming´s ideas to understanding the system
7. *Realistic monitoring and feedback*

Zombie statistics 1

Percentage of non-responders

What the FDA says



Paving the way for personalized medicine, FDA Oct2013

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

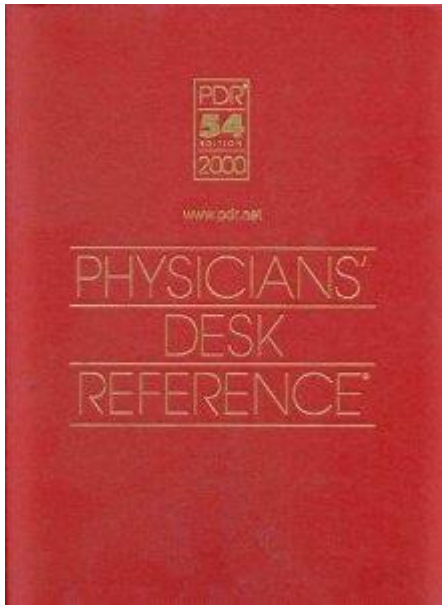
Zombie statistics 2

Where they got it

Where those who got it 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

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

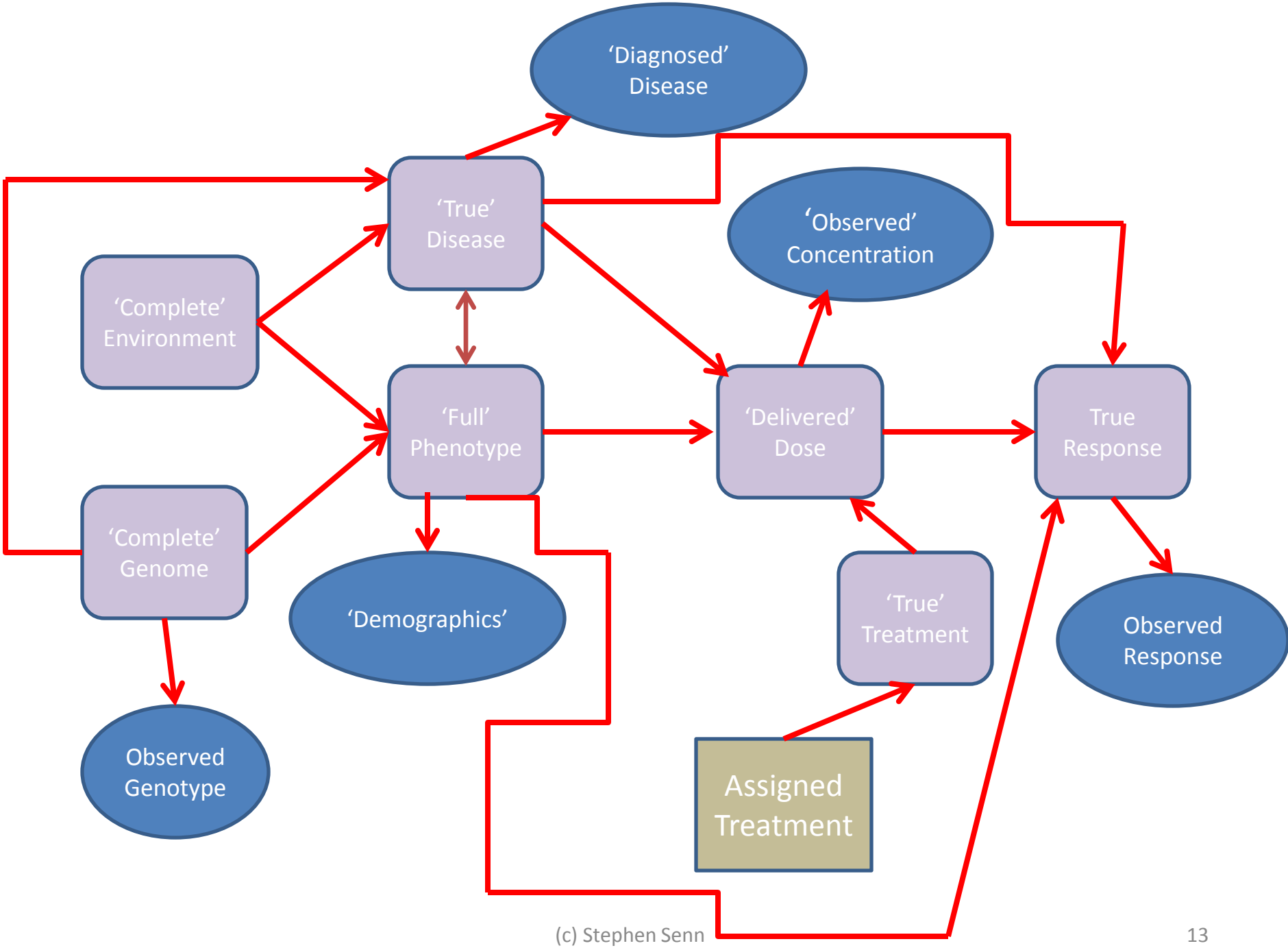
Vic Reeves

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

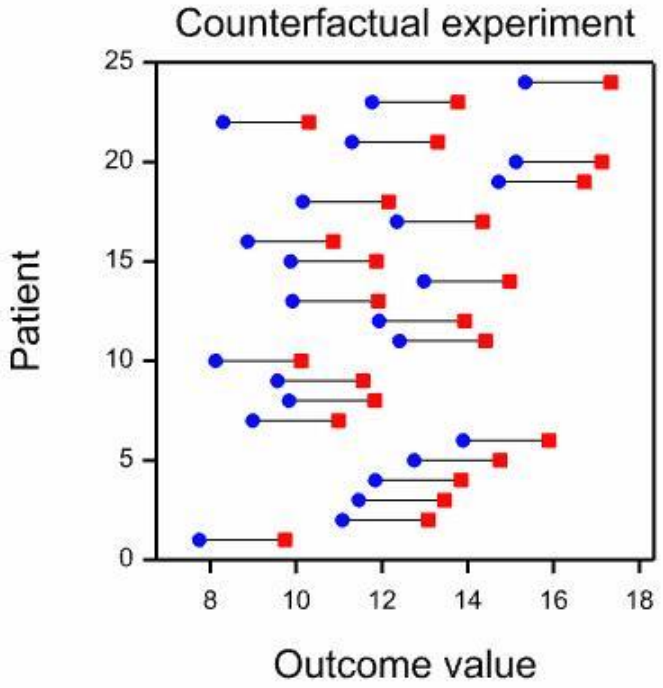
Sources of Variation in Clinical Trials

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



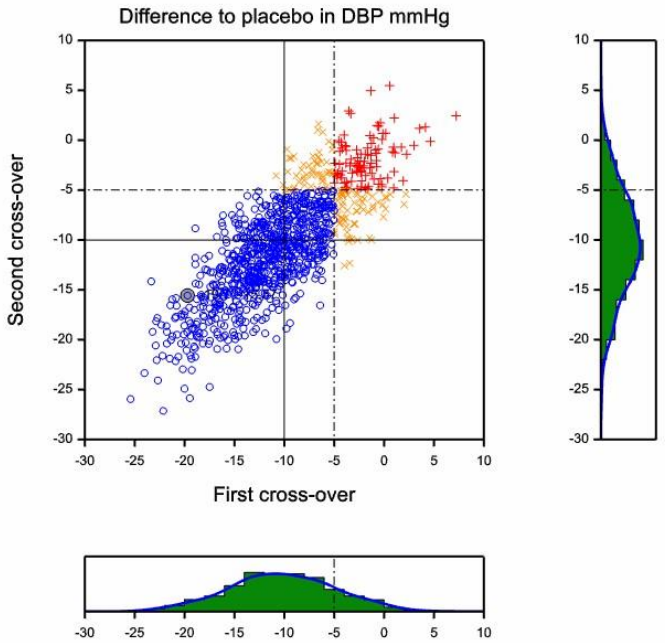
Left-hand panel: what you would see if patients could be treated both ways
Note how difference active-placebo is constant

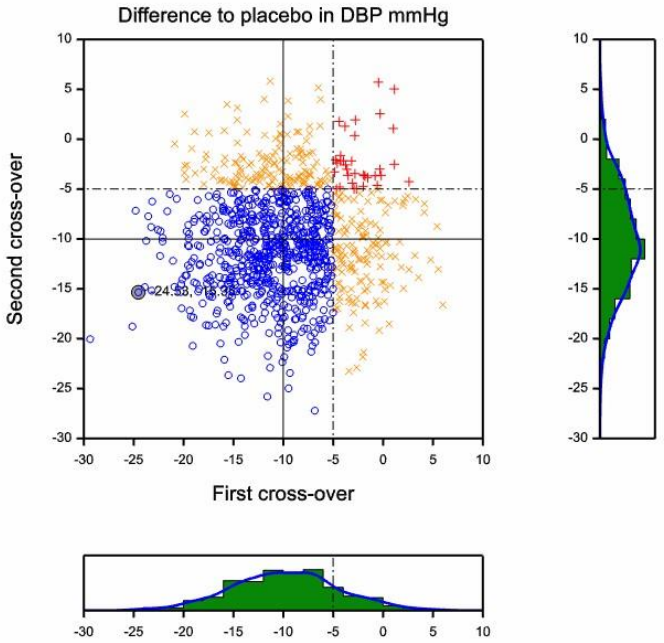
A Thought Experiment

- Imagine a cross-over trial in hypertension
- Patients randomised to receive ACE II inhibitor or placebo in random order
- Then we do it again
- Each patient does the cross-over twice
- We can compare each patient's response under ACE II to placebo twice

Design

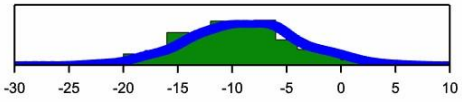
	First Cross-over		Second Cross-over	
	Period			
Sequence	1	2	3	4
I	A	B	A	B
II	B	A	B	A
III	A	B	B	A
IV	B	A	A	B

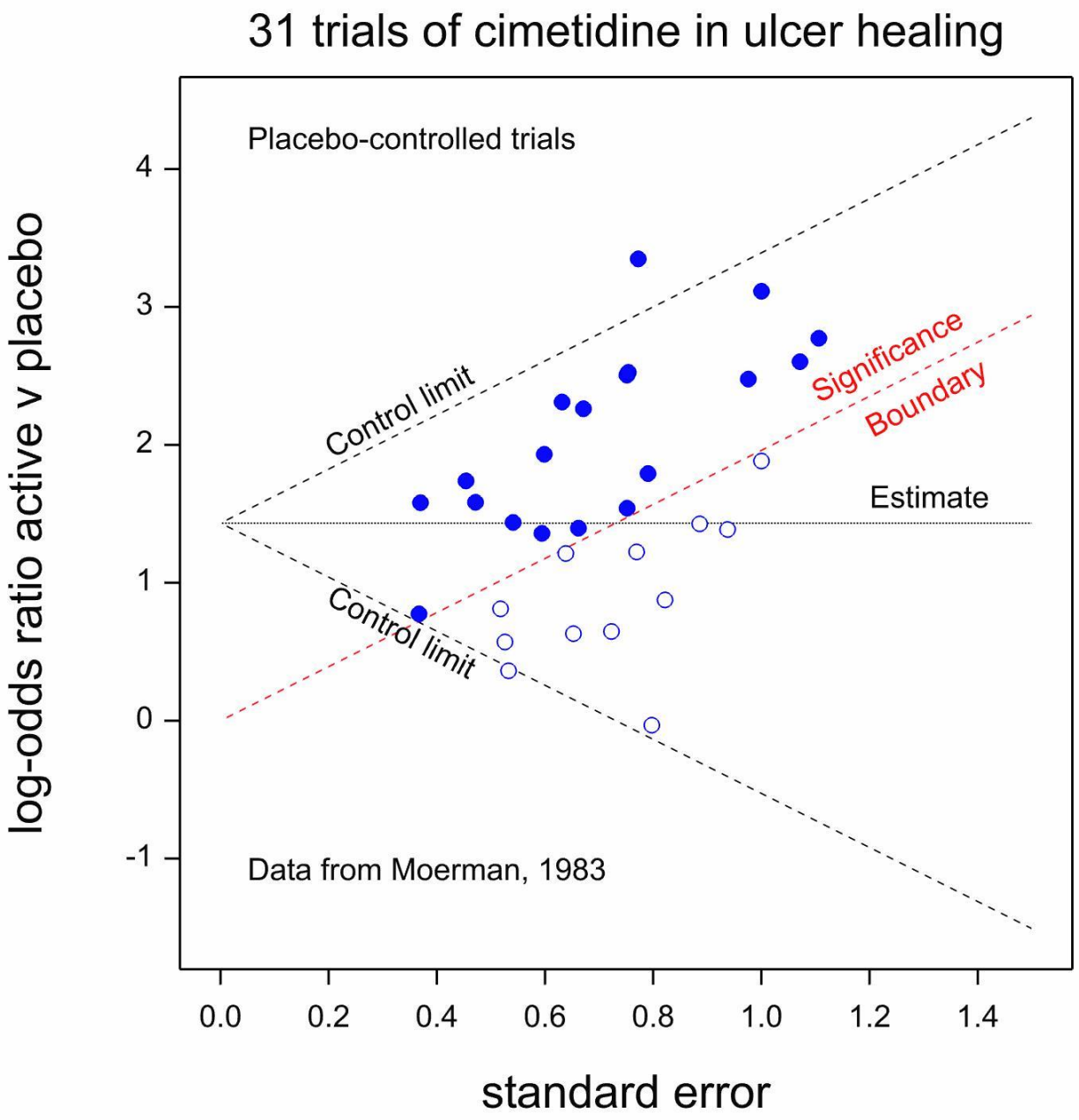




Patients are treated in two cross-over trials , thus permitting two estimates of the difference between active treatment and placebo. The difference on the second occasion is plotted against the first. Blue = response on both occasions, red = non-response on both occasions, orange = response on one occasion but not the other.

The marginal distributions are given as green histograms. LHS response on first occasion predicts response on second. RHS response on first occasion does **not** predict response on second. If you had only carried out one cross-over you would have the picture below. Which case does it apply to?



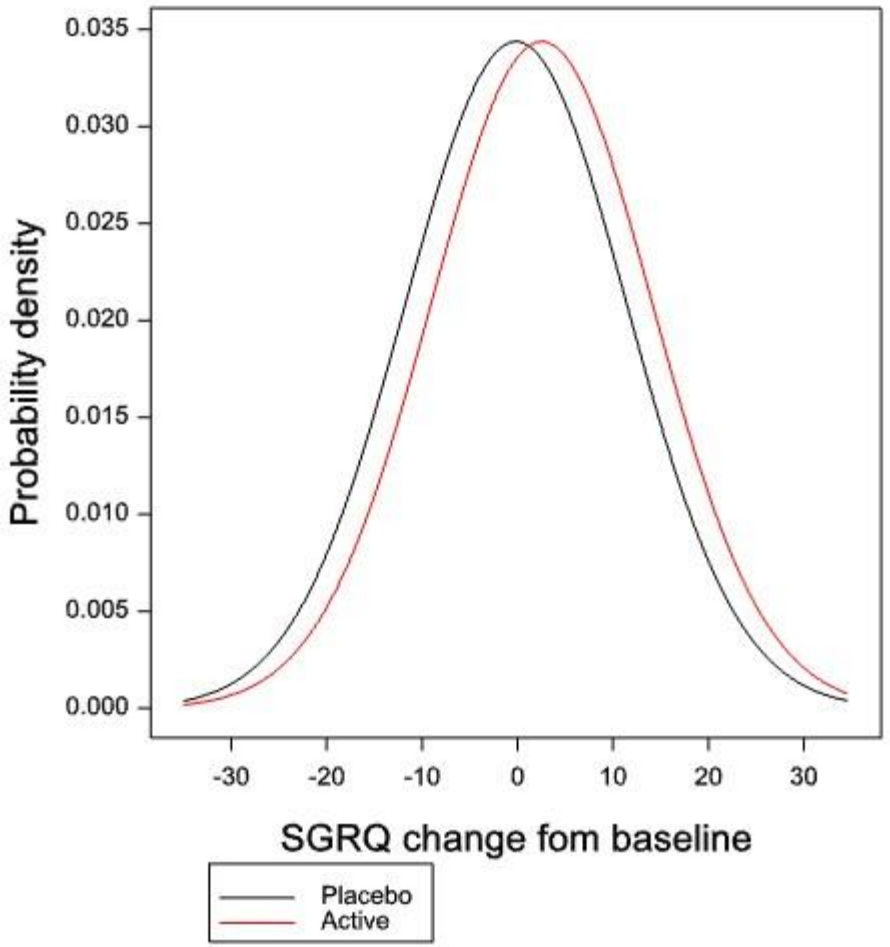

31 trials of cimetidine in ulcer healing
Placebo-controlled trials
Control limit
Significance
Boundary
Estimate
Control limit
Data from Moerman, 1983
log-odds ratio active v placebo
standard error
4
3
2
1
0
-1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Tiotropium v Placebo in Chronic Obstructive Pulmonary Disease

From the UPLIFT Study, *NEJM*, 2008

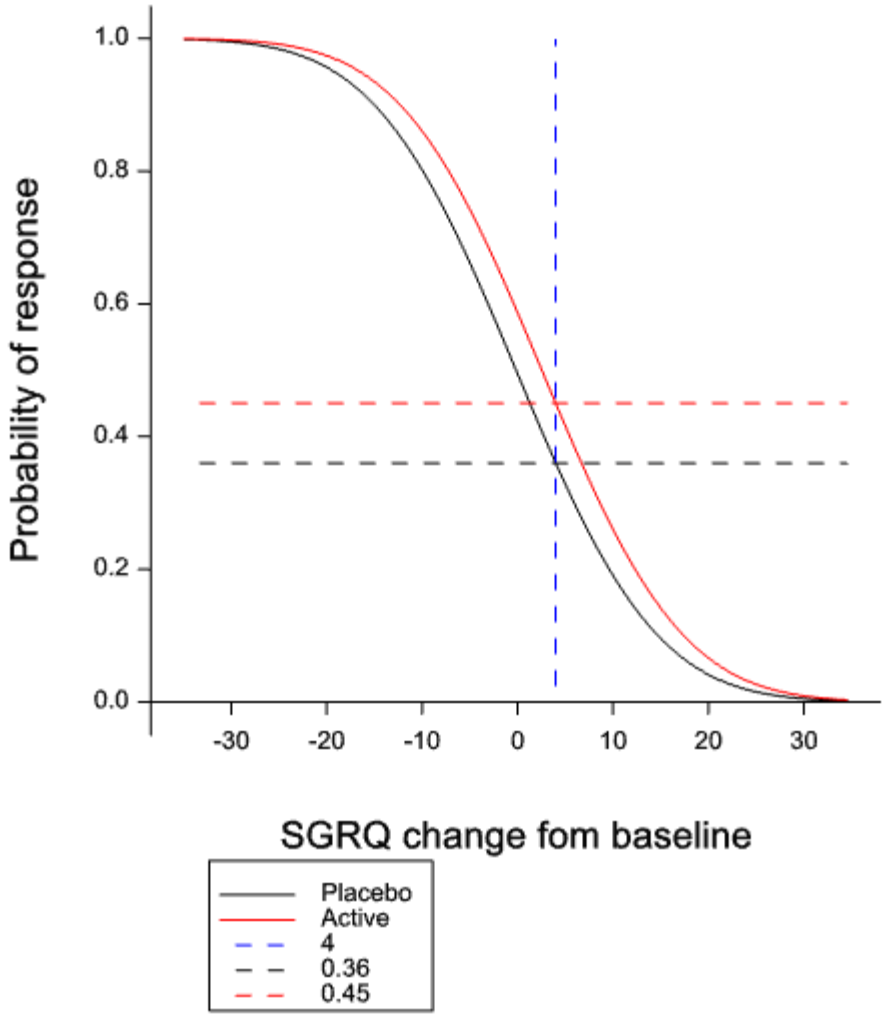
Significant differences in favor of tiotropium were observed at all time points for the mean absolute change in the SGRQ total score (ranging from 2.3 to 3.3 units, P<0.001), although the differences on average were below what is considered to have clinical significance (Fig. 2D). **The overall mean between-group difference in the SGRQ total score at any time point was 2.7** (95% confidence interval [CI], 2.0 to 3.3) in favor of tiotropium (P<0.001). **A higher proportion of patients in the tiotropium group than in the placebo group had an improvement of 4 units or more in the SGRQ total scores from baseline at** 1 year (49% vs. 41%), 2 years (48% vs. 39%), 3 years (46% vs. 37%), and **4 years (45% vs. 36%)** (P<0.001 for all comparisons).

(My emphasis)



Two Normal distributions with the same spread but the Active treatment has a mean 2.7 higher.

If this applies every patient under active can be matched to a corresponding patient under placebo who is 2.7 worse off



A cumulative plot corresponding to the previous diagram.

If 4 is the threshold, placebo response probability is 0.36, active response probability is 0.45.

In summary…this is rather silly

- If there is sufficient measurement error even if the true improvement is identically 2.7, some will show an 'improvement' of 4
- The conclusion that there is a higher proportion of *true* responders *by the standard of 4 points* under treatment than under placebo is quite unwarranted
- So what is the point of analysing 'responders'?

Who are the authors?

1. Tashkin, DP, Celli, B, Senn, S, Burkhart, D, Kesten, S, Menjoge, S, Decramer, M. A 4-Year Trial of Tiotropium in Chronic Obstructive Pulmonary Disease, *N Engl J Med* 2008.

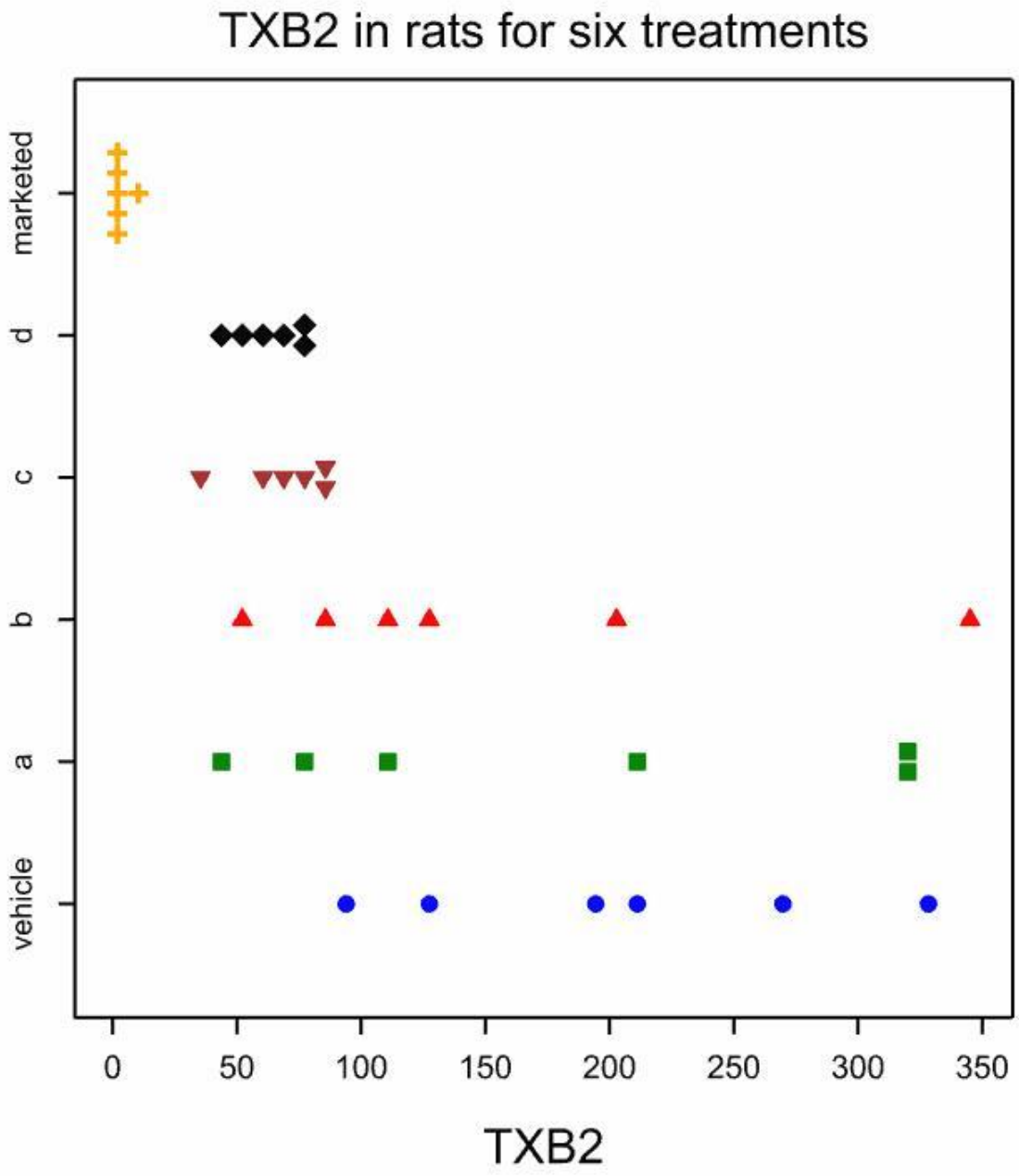
Personal note. I am proud to have been involved in this important study and have nothing but respect for my collaborators. The fact that, despite the fact that two of us are statisticians, we have ended up publishing something like this shows how deeply ingrained the practice of responder analysis is in medical research. We must do something to change this.

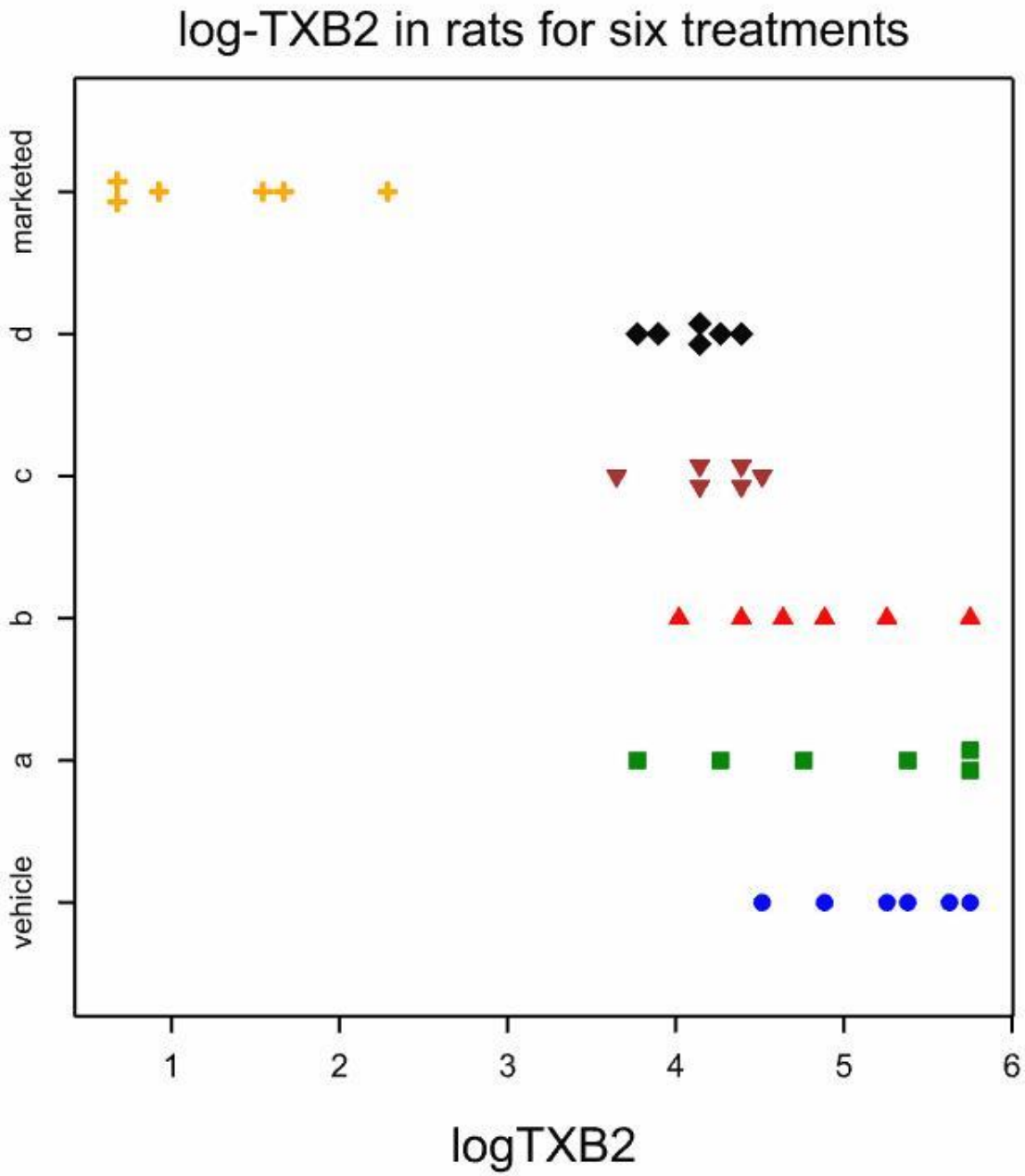
Beware of clinically relevant measures

- Such measures may not have good statistical properties
- There may be variation in terms of these measures that can be removed by transformation
- Analysis is better on the ‘additive’ scale
- Back transformation can be used for perdiction

An Example from Research

- Four experimental p38α kinase inhibitors
- Vehicle and marketed product as controls
- Thrombaxane B2 (TXB2) is used as a marker of COX-1 activity
- Six rats per group were treated for a total of 36 rats
- At the end of the study rats are sacrificed and TXB2 is measured.


TXB2 in rats for six treatments
marketed
d
c
b
a
vehicle
0
50
100
150
200
250
300
350
TXB2


log-TXB2 in rats for six treatments
marketed
d
c
b
a
vehicle
1
2
3
4
5
6
logTXB2

Analyses of Original and Transformed Data

Original

Source	DF	SS	MS	F	P
Treat	5	184596	36919	6.31	< 0.001
Error	30	175439	5848		
Total	35	360035			

Transformed

Source	DF	SS	MS	F	P
Treat	5	62.7	12.54	40.1	< 0.001
Error	30	9.4	0.31		
Total	35	72.1			

Untransformed

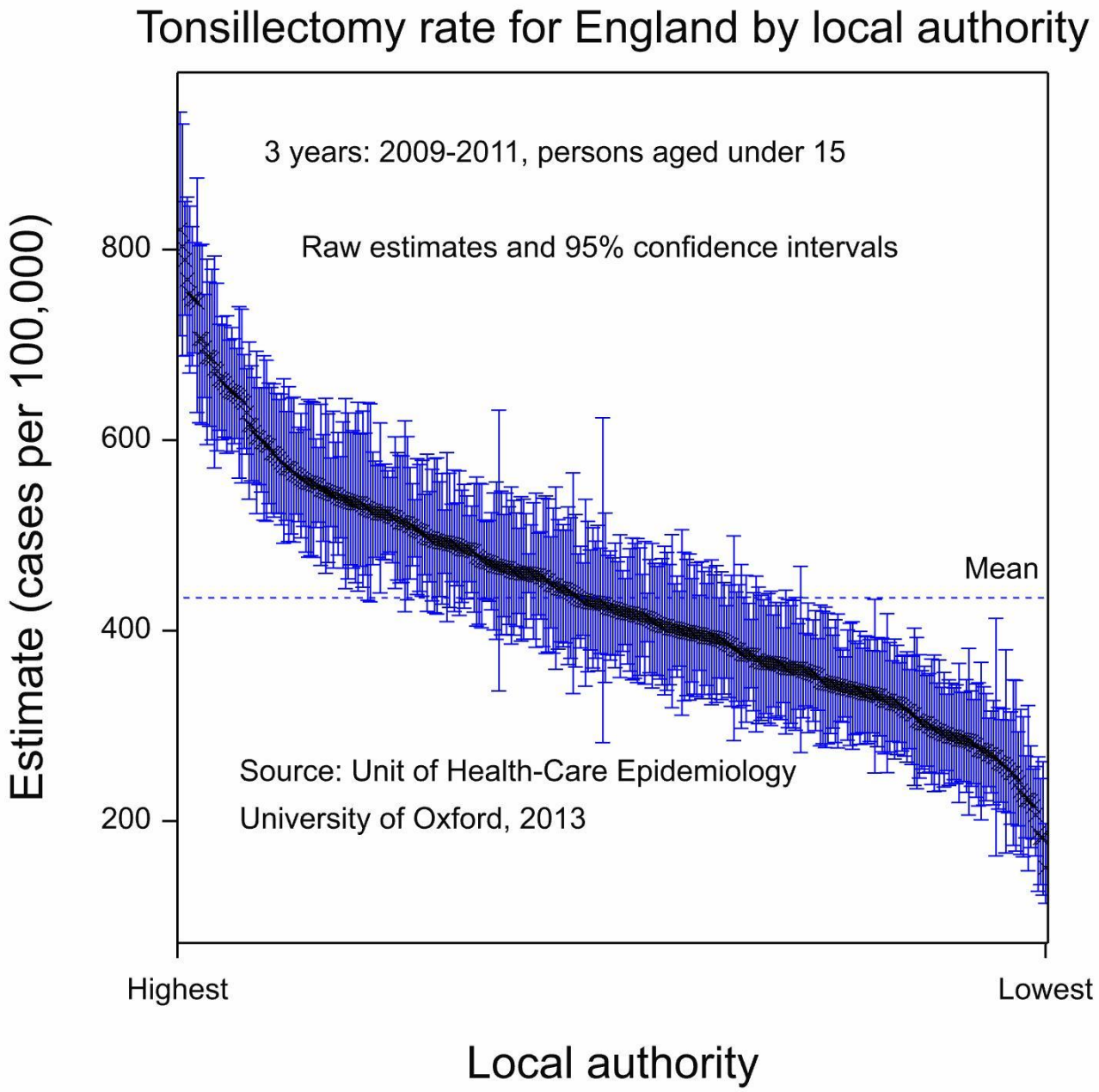
*** Bartlett's Test for homogeneity of variances ***
Chi-square 50.87 on 5 degrees of freedom: probability < 0.001

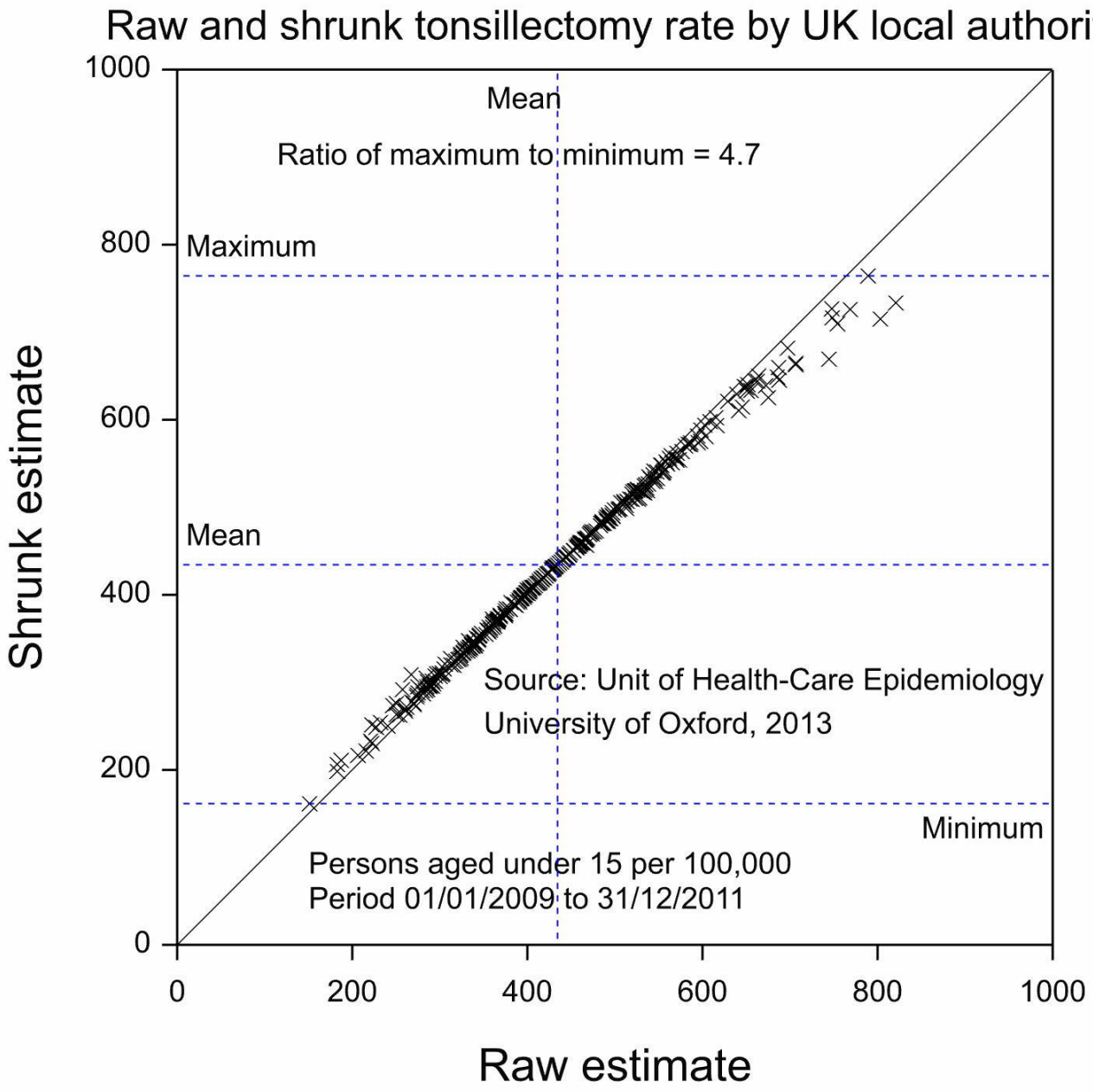
Log-transformed

*** Bartlett's Test for homogeneity of variances ***
Chi-square 8.95 on 5 degrees of freedom: probability 0.111

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
Estimate (cases per 100,000)
800
600
400
200
Mean
Source: Unit of Health-Care Epidemiology
University of Oxford, 2013
Highest
Lowest
Local authority


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

“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		

Who's to blame?

- Statisticians
  - (Including me)
- Our life scientist colleagues don't understand variation
- We do
- We should tell them the truth

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

The supply of truth always greatly exceeds its demand

John F Moffitt