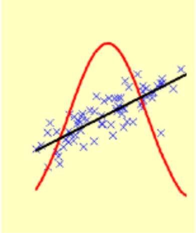


Stephen Senn

Edinburgh

stephen@senns.demon.co.uk

@stephensenn

Acknowledgements

My thanks to Pierre Verweij and the committee for the invitation

This work was partly supported by the European Union’s 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. “IDeAL”

A former captain of industry speaks

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

A leading researcher speaks

Not only will genetic tests predict responsiveness to drugs on the market today, but also genetic approaches to disease prevention and treatment will include an expanding array of gene products for use in developing tomorrow's drug therapies.

Francis S Collins, *NEJM*, 1999

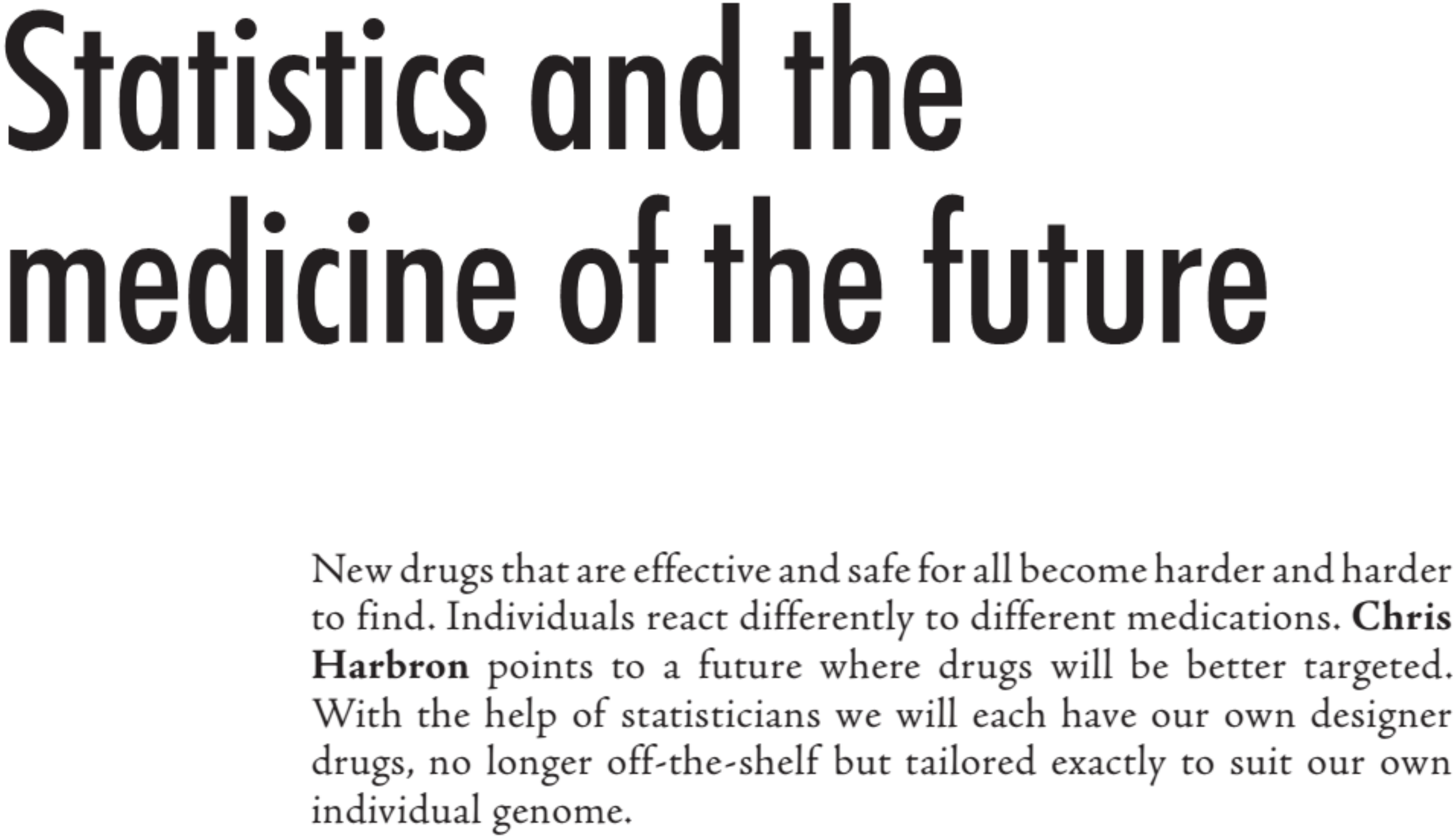
The editor of a leading journal speaks

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.)

Significance get’s in on the act



Statistics and the medicine of the future

New drugs that are effective and safe for all become harder and harder to find. Individuals react differently to different medications. **Chris Harbron** points to a future where drugs will be better targeted. With the help of statisticians we will each have our own designer drugs, no longer off-the-shelf but tailored exactly to suit our own individual genome.

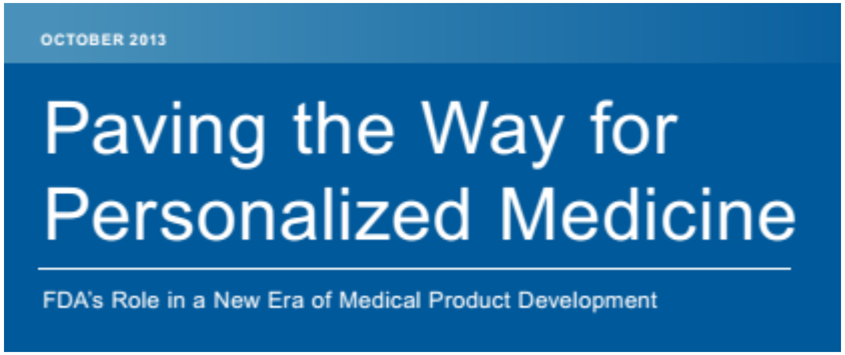
June 2006

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)

The world’s leading regulatory agency speaks



OCTOBER 2013

Paving the Way for Personalized Medicine

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

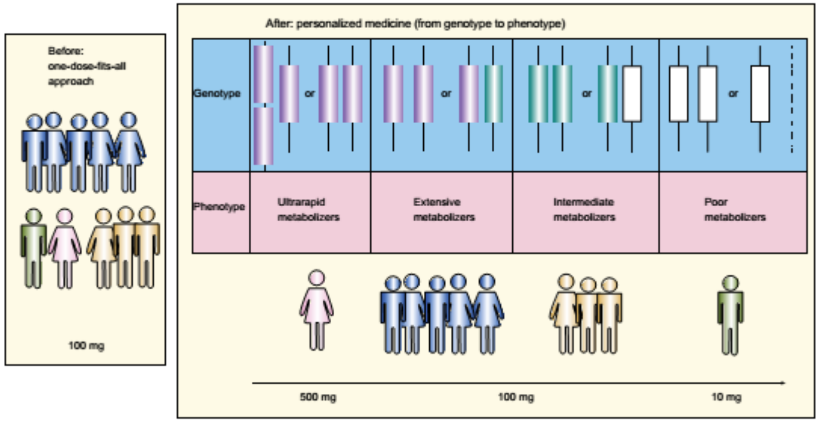
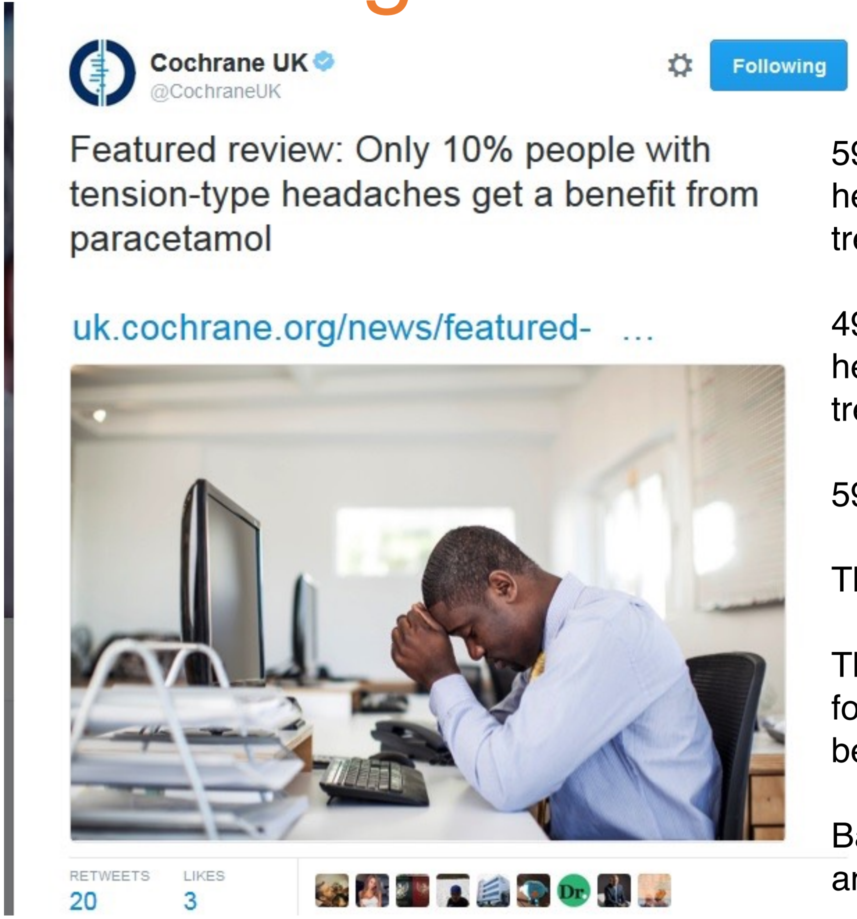

Before: one-dose-fits-all approach
100 mg
After: personalized medicine (from genotype to phenotype)
Genotype
or
or
or
or
Phenotype
Ultrarapid metabolizers
Extensive metabolizers
Intermediate metabolizers
Poor metabolizers
500 mg
100 mg
10 mg

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.)

The leading evidence based medicine organisation speaks



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

uk.cochrane.org/news/featured- …

59% had no or at worst mild headache **<u>after 2 hours</u>** when treated with paracetamol

49% had no or at worst mild headache **<u>after 2 hours</u>** when treated with placebo

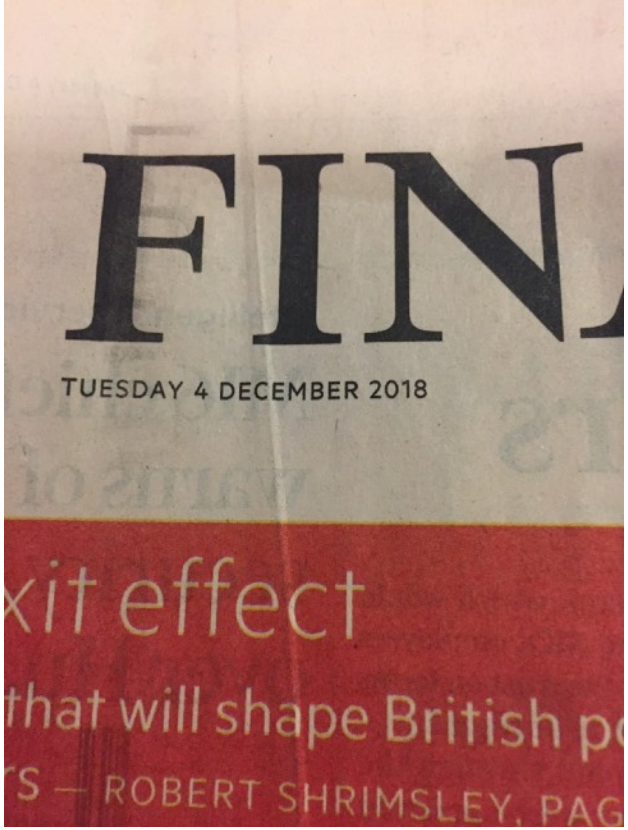
59%-49% = 10%

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

The Researchers are Enthusiastic
and the Financial Press



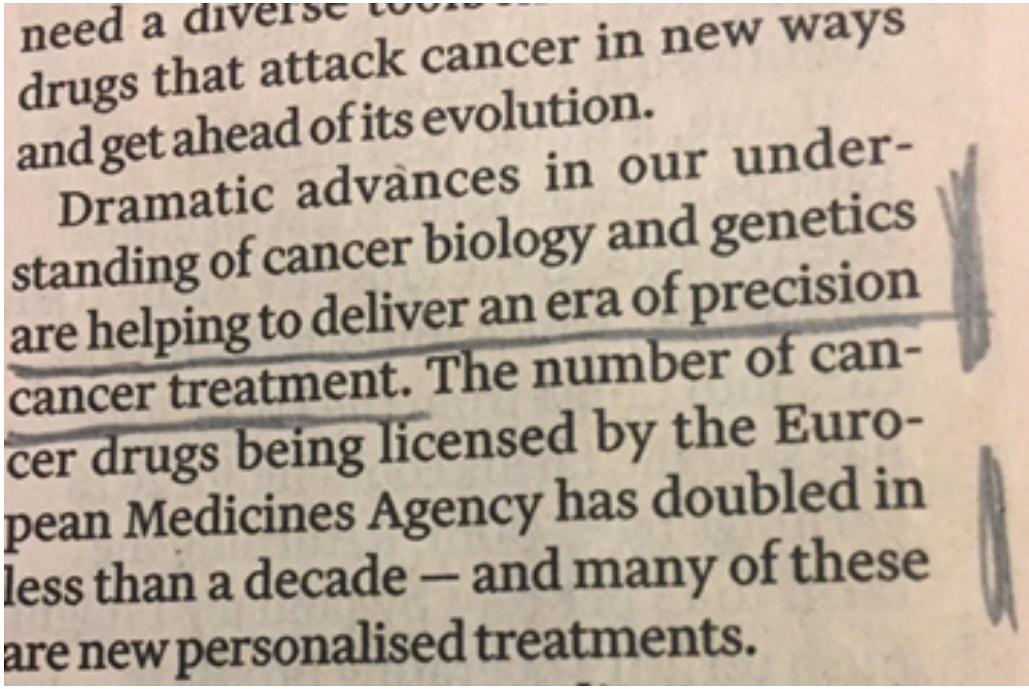
FIN

TUESDAY 4 DECEMBER 2018

xit effect

that will shape British p

S — ROBERT SHRIMSLEY, PAG



need a diverse

drugs that attack cancer in new ways and get ahead of its evolution.

Dramatic advances in our understanding of cancer biology and genetics are helping to deliver an era of precision cancer treatment. The number of cancer drugs being licensed by the European Medicines Agency has doubled in less than a decade — and many of these are new personalised treatments.

Olivia Rossanese, Institute of Cancer Research, December 2018

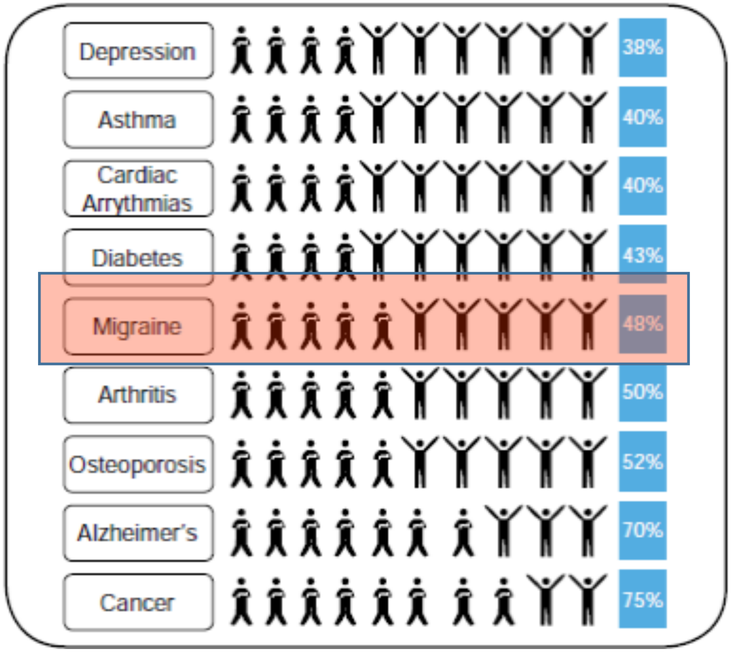
Warning

We tend to believe the truth is in there.

Sometimes it isn’t and the danger is that we will find it anyway

Statistics on non-responders

What the FDA says



Paving the way for personalized medicine, FDA Oct 2013

Where the FDA got it

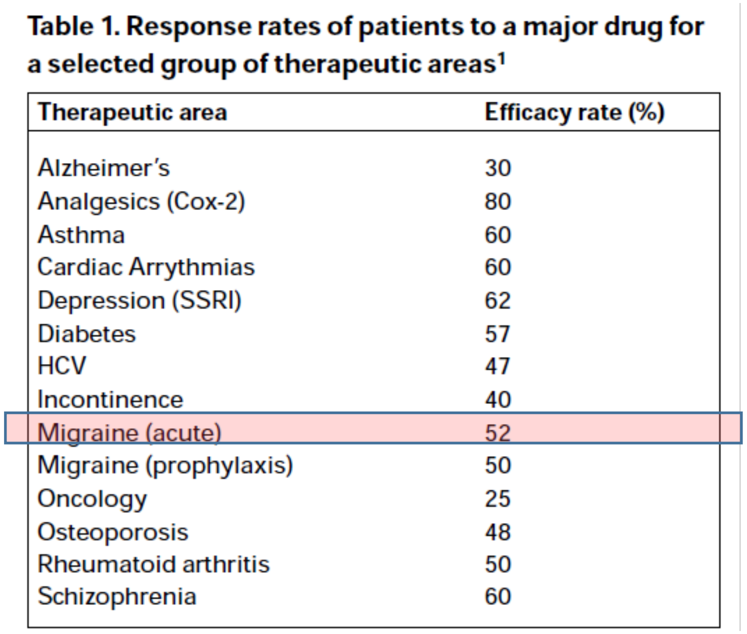


Table 1. Response rates of patients to a major drug for a selected group of therapeutic areas[1]

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

Back to the source of the source

Where the FDA got it

Where those who got it got it

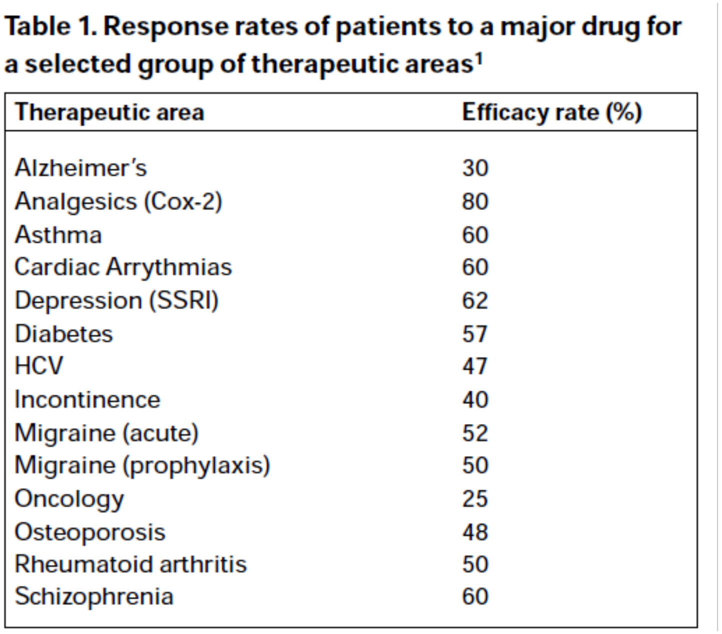
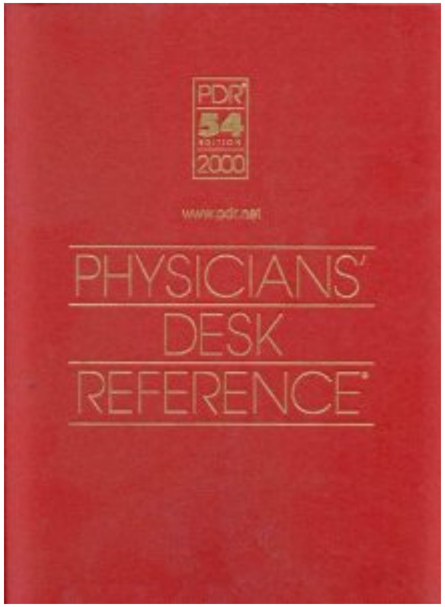
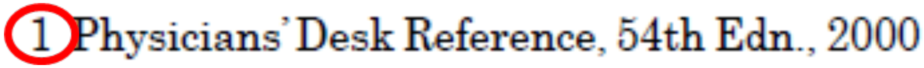


Table 1. Response rates of patients to a major drug for a selected group of therapeutic areas[1]

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





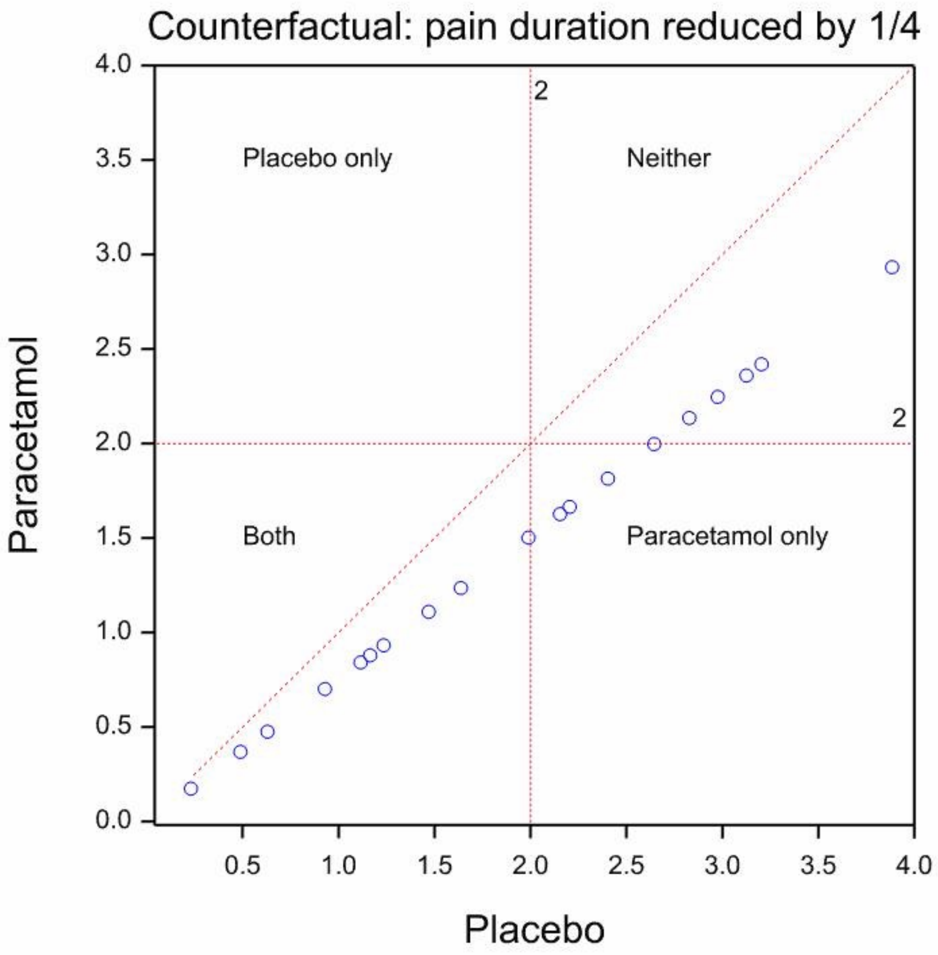
1 Physicians' Desk Reference, 54th Edn., 2000

My thesis

- We have a plague of zombie statistics
  - They are ugly
  - They are evil
  - They refuse to die
- Even if the figures were right in some numerical sense, they wouldn’t mean what they are assumed to mean
- We need to do something!

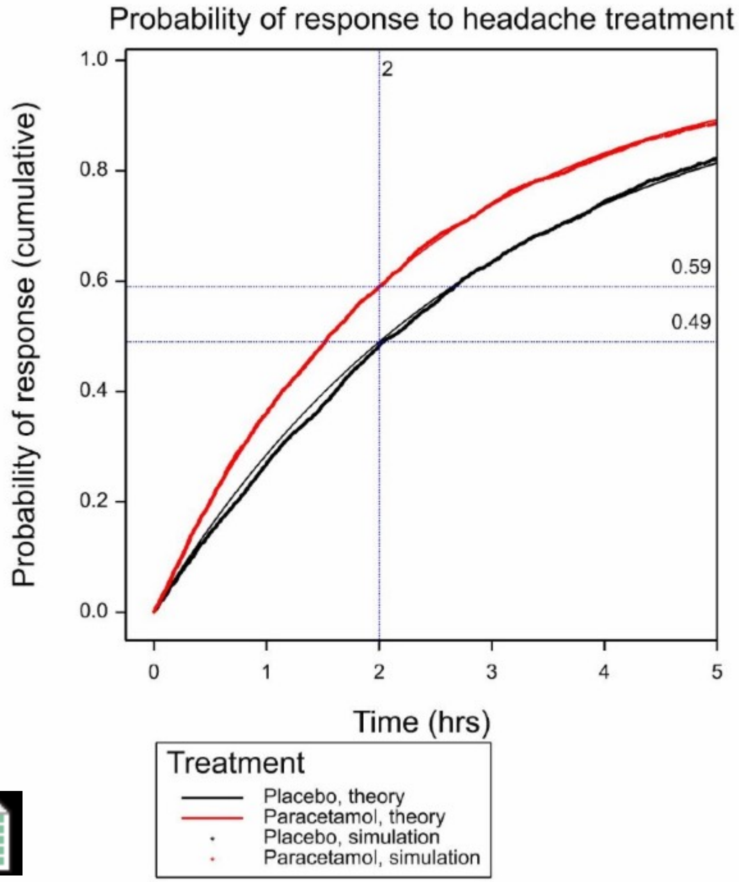
Back to the headache
A Recipe to Mimic the Cochrane Results

- Generate one random number, U_i, for each of 6000 headaches, $i = 1,2…6000$
- Calculate pairs of headache
  - 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 each headache can only receive one treatment
  - The other is *counterfactual*
- Draw the empirical cumulative distribution


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

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



Microsoft Excel Worksheet

Lessons

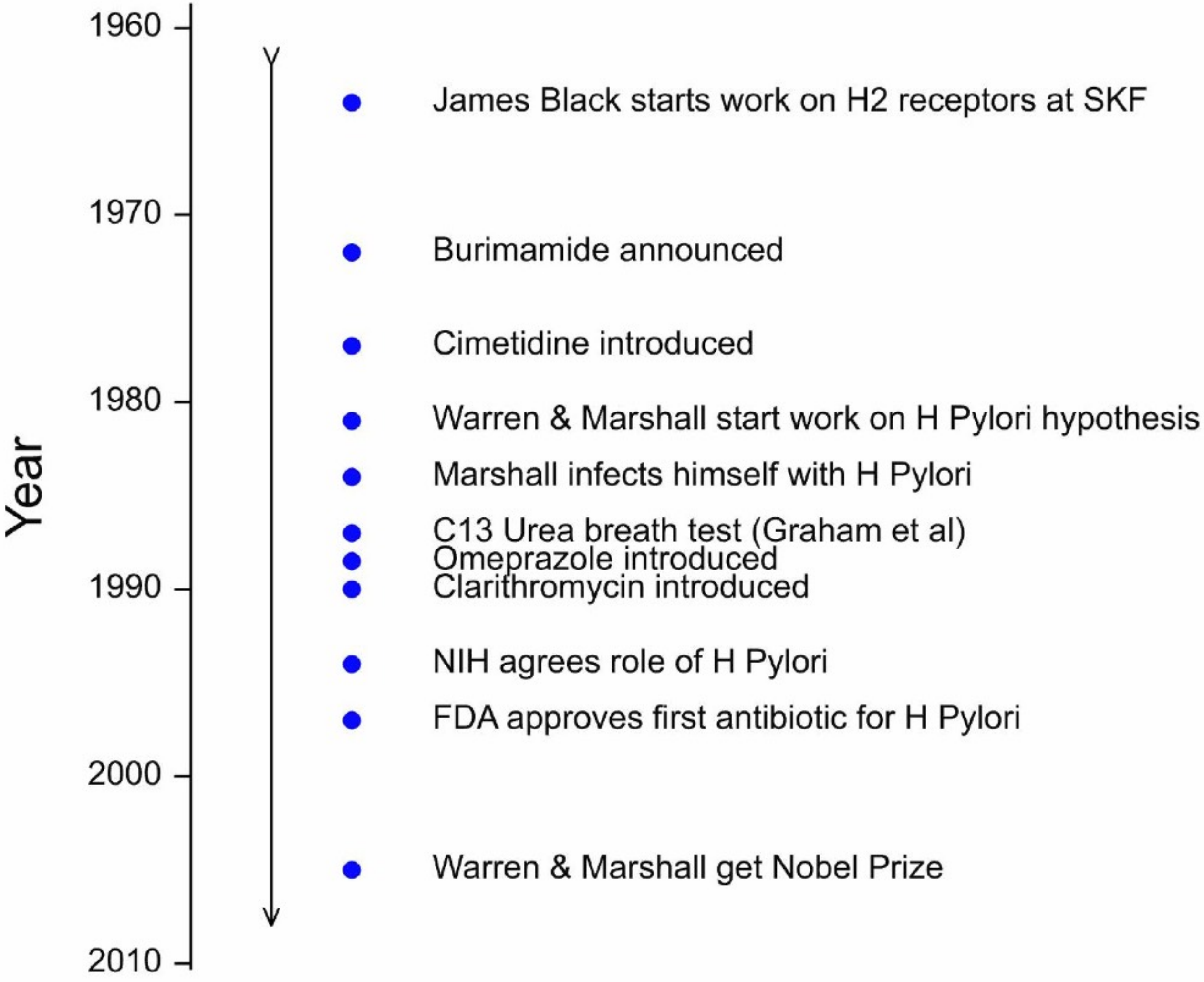
Particular

- 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

A time line of ulcer treatment

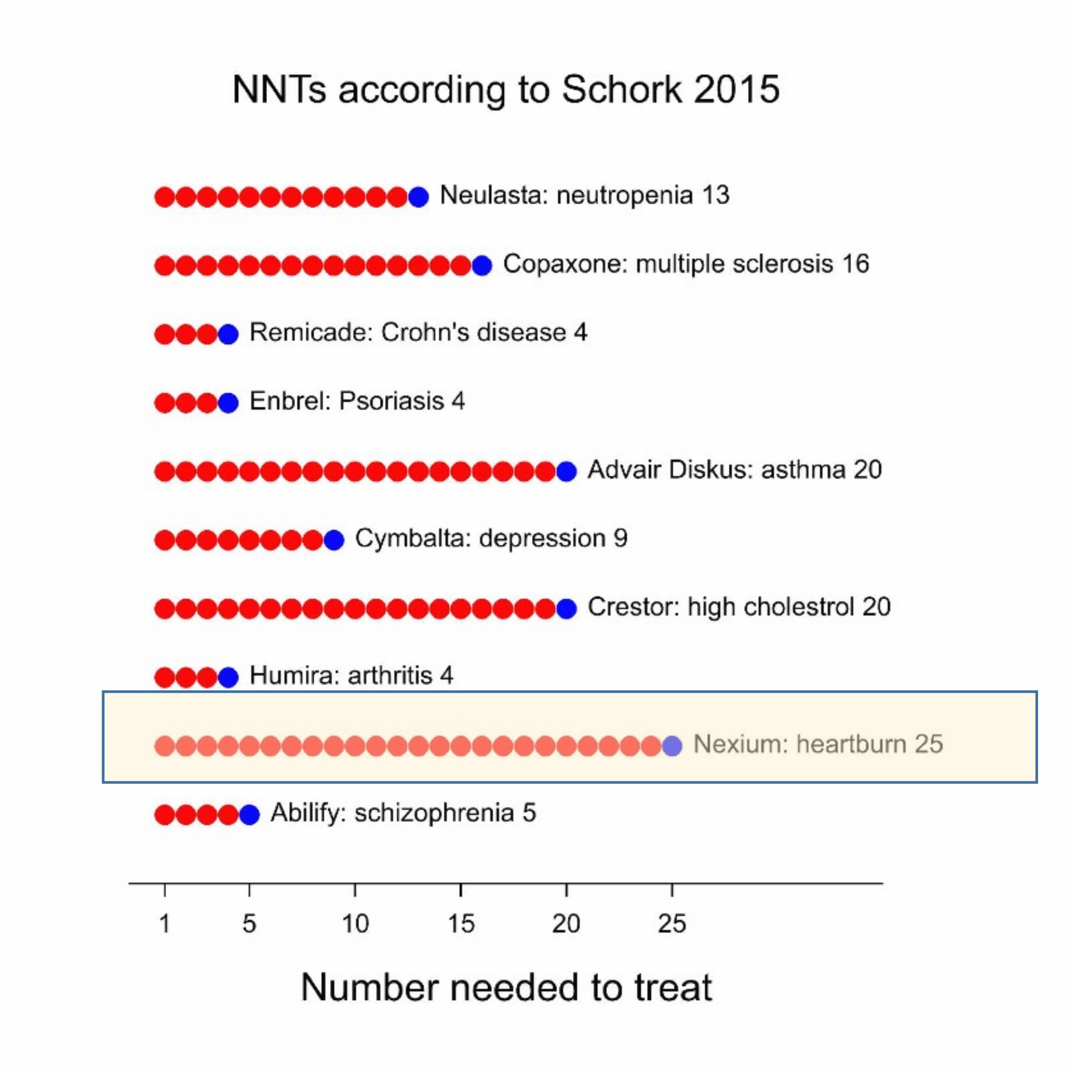


- James Black starts work on H2 receptors at SKF
- Burimamide announced
- Cimetidine introduced
- Warren & Marshall start work on H Pylori hypothesis
- Marshall infects himself with H Pylori
- C13 Urea breath test (Graham et al)
- Omeprazole introduced
- Clarithromycin introduced
- NIH agrees role of H Pylori
- FDA approves first antibiotic for H Pylori
- Warren & Marshall get Nobel Prize

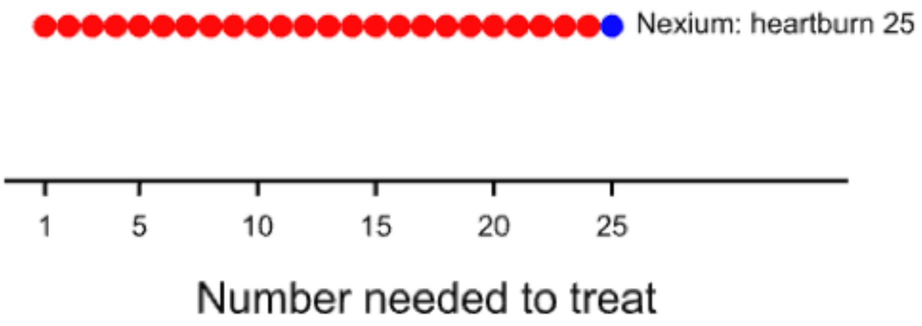
Wasteful medicine?

"Every day, millions of people are taking medications that will not help them. The top ten highest-grossing drugs in the United States help between 1 in 25 and 1 in 4 of the people who take them. "

Schork, N, *Nature* 2015



Nexium revisited



- This is for Nexium (*Esomeprazole*) compared to other proton pump inhibitors
- This is a highly successful class of treatment
- 'Response' is 92% in one case and 88% in the other *at 8 weeks* (Labenz et al, 2005)
- Response rises over time and would probably increase *beyond* 8 weeks
- The claim that only 1 in 25 benefit is nonsense

Two extreme cases
Illustrated with the EXPO study

		Esomeprazole		
		Not healed	Healed	Total
Pantoprazole	Not healed	7.9	4.8	12.7
	Healed	0.0	87.3	87.3
	Total	7.9	92.1	100.0

Case where no patient would respond on Pantoprazole who did not on Esomeprazole (Nexium)

		Esomeprazole		
		Not healed	Healed	Total
Pantoprazole	Not healed	0.0	12.7	12.7
	Healed	7.9	79.4	87.3
	Total	7.9	92.1	100.0

Case where all patients who did not respond on Esomeprazole (Nexium) would respond on Pantoprazole

Ulcer treatment in the last quarter of the 20th century

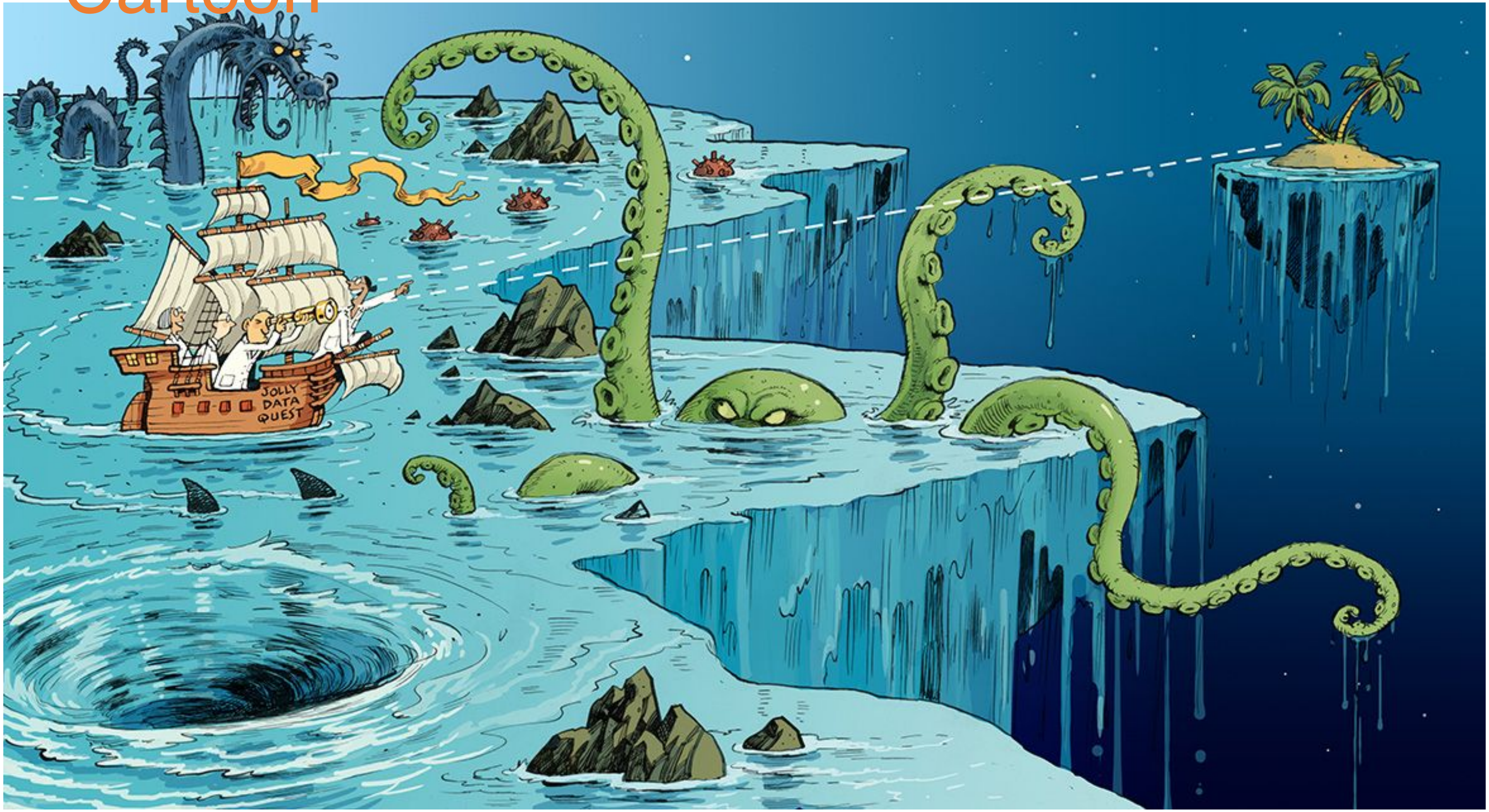
- A great success story
- Four to five major innovations/discovery
  - H2 antagonists
  - Proton pump inhibitors
  - The role of H Pylori
  - C13 Urea breath test
  - Antibiotics developed to treat H Pylori
- Together these transformed the treatment of ulcer treatment
- General surgery largely wiped out
- For the many not just the few
  - Admittedly the combination of the C13 test and antibiotic treatment involves a degree of personalisation
- But now look at how the personalised medicine bandwagon misrepresents it

We have moved from finding highly effective treatments for most patients to trying to find expensive ones for almost nobody at all

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 the standards we use 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

Slide with the Obligatory Purloined Cartoon



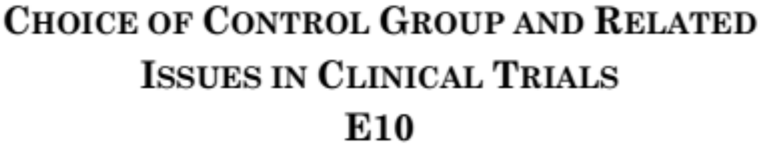
Senn, Nature, 29 November 2018

Baseline as a counterfactual? An example

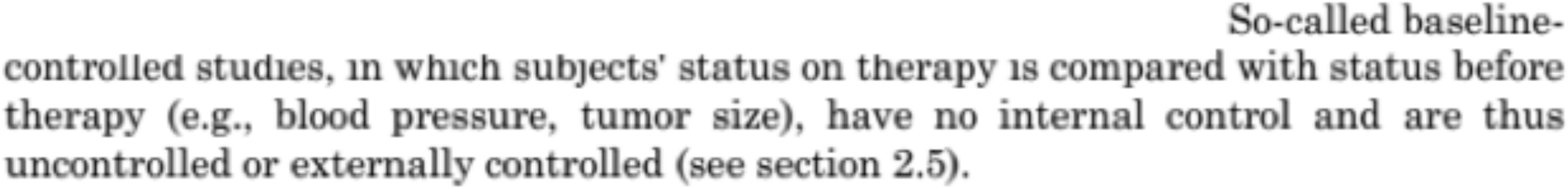
- Long term trial (20 years) of an anti-aging cream by hundreds of subjects
- Perfect compliance
  - Obviously a fictional example!
- Measure based on ‘wrinkle score’
  - Number of wrinkles at the end of the trial compared to the baseline value for every subject
- The average wrinkle score at the end of the trial was zero
- Clearly the cream did not work

What the example makes clear

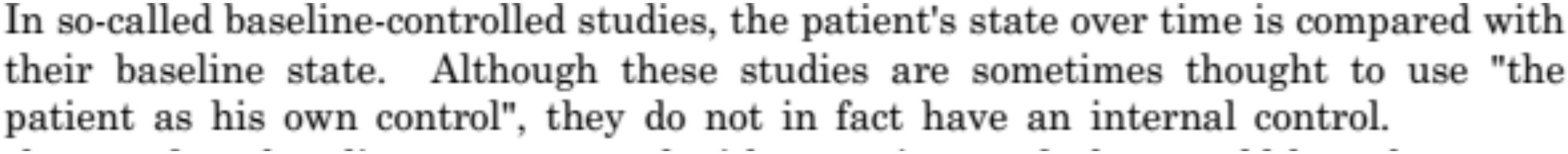
- Baselines are not in general suitable as counterfactuals
- Alas, the past is not on offer
  - I say this as someone who is now aged 66 !
- Decision making, including decision making as regards which treatment to take, is a choice between possible futures
- A baseline is only useful to the degree it helps us predict this future
- That is not the way clinical trials work or at least should work
- This is all very very obvious
- So why do we ignore at an individual level, what we all know at the level of a trial?



Choice of Control Group and Related Issues in Clinical Trials E10



So-called baseline-controlled studies, in which subjects' status on therapy is compared with status before therapy (e.g., blood pressure, tumor size), have no internal control and are thus uncontrolled or externally controlled (see section 2.5).



In so-called baseline-controlled studies, the patient's state over time is compared with their baseline state. Although these studies are sometimes thought to use "the patient as his own control", they do not in fact have an internal control.

15% increase in forced expiratory volume in one second (FEV_1)
Some regulatory magic

Some key efficacy PD responses were similar between xxx and yyy following 180 µg single dose inhalation. The percentage of responders (15% increase in FEV_1 from baseline) was 63% and 52% for xxx and yyy, respectively

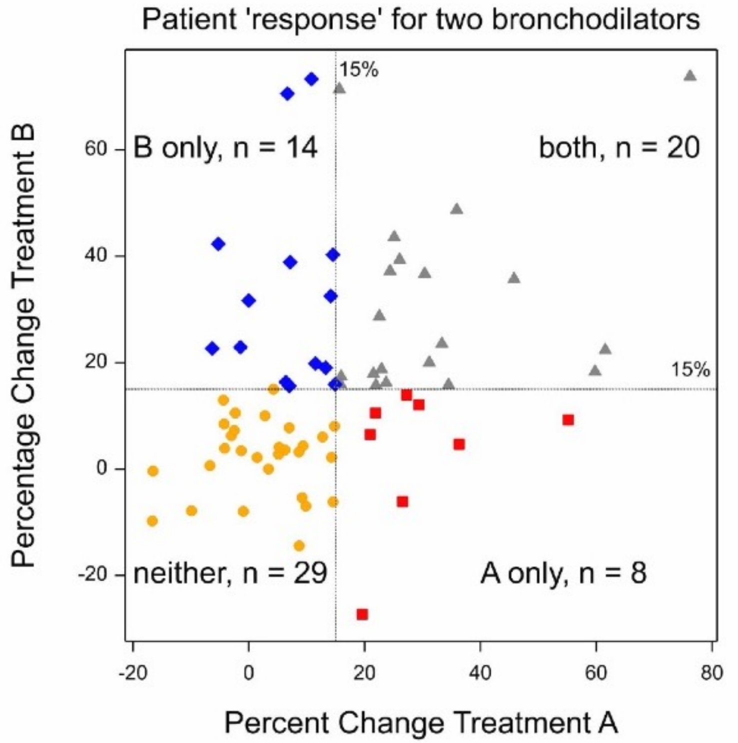
From a cross-over trial submitted to the FDA as part of a review

This is cited just to show that the standard of 15% bronchodilation is used

I shall now use an example of my own

So can we identify individual response with cross-overs?

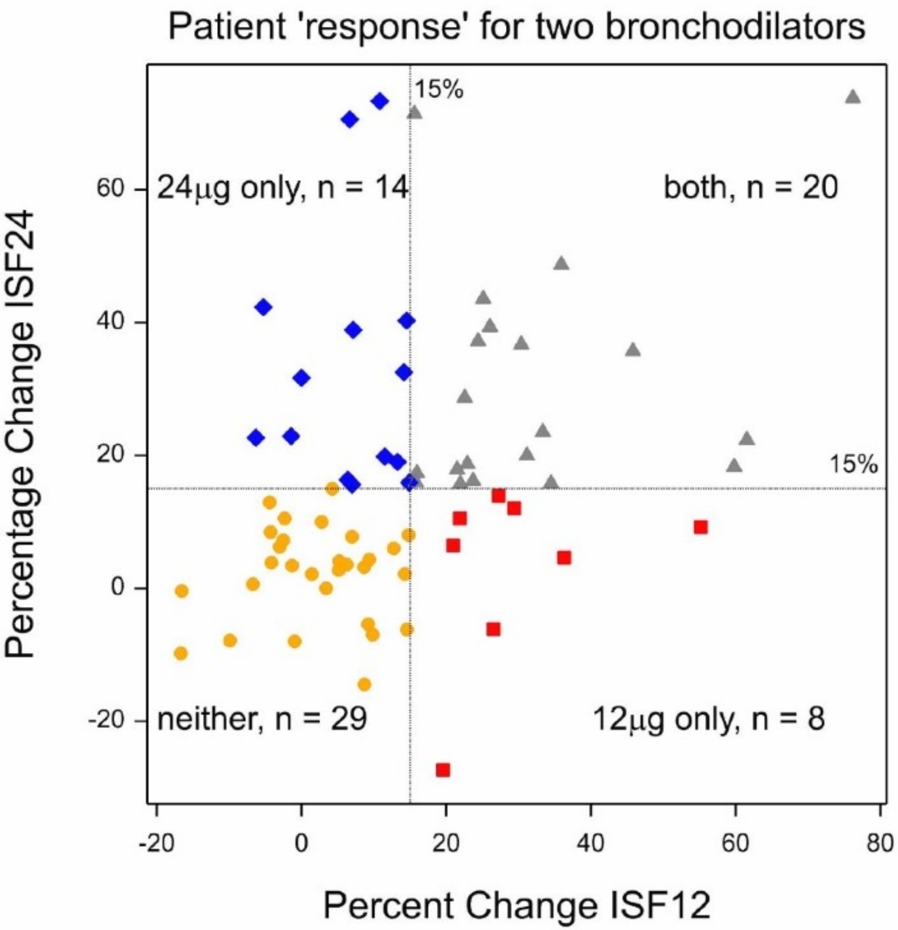
- Cross-over trial in asthma
  - 71 patients
  - Forced expiratory volume in one second (FEV_1) at 12 hours
- FDA definition of response is ≥ 15% increase compered to baseline
- There seem to be a number of patients who respond to B and not to A and vice versa
- Clearly if we can find predictive characteristics of them we can improve treatment
- Can't we?



(c) Stephen Senn 2019

Differential response? Not so fast

- A is ISF 12µg, formoterol
- B is ISF 24µg formoterol
- 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



The real lessons

- *Other things being equal* a high NNT is indicative of a poorer treatment but it is not a valid shortcut to studying variation
- We need to understand and master the variation in the system
- We need to not naively over-interpret differentiation in observed response
  - Some of it may be genuine treatment-by-patient interaction
  - Much of it may be within-patient variation
- In many (but not all) cases the task facing us will be to deliver better average medicine
- In this connection there is one big problem we continually overlook
- The main source of variation in the system is not patients
- It's doctors

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
- This is what is inspiring what Brent James is doing at Inter-Mountain health
- At the moment we are making a bad job of this
- NNTs and responder analysis are no substitute for serious study of components of variation
- There is no point publishing and developing yet more complicated fancy stuff involving mixed models if we fail at the first hurdle
- Once you have sold the post by permitting naïve causal definitions the battle is already lost
- We must do better in fighting the omic hype
    - and I would include me but I am retired

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		

The supply of truth always greatly exceeds its demand

John F Moffitt

Suggested further reading

Araujo, A., S. Julious and S. Senn (2016). "Understanding Variation in Sets of N-of-1 Trials." PLOS ONE **11**(12): e0167167.

Holland, P. W. (1986). "Statistics and Causal Inference." Journal of the American Statistical Association **81**(396): 945-960.

Robins, J. and S. Greenland (1989). "The probability of causation under a stochastic model for individual risk." Biometrics **45**(4): 1125-1138.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." Journal of educational Psychology **66**(5): 688.

Senn, S. J. (2001). "Individual Therapy: New Dawn or False Dawn." Drug Information Journal **35**(4): 1479-1494.

Senn, S. J. (2004). "Individual response to treatment: is it a valid assumption?" BMJ **329**(7472): 966-968.

Senn, S. J. (2016). "Mastering variation: variance components and personalised medicine." Statistics in Medicine **35**(7): 966-977.

Senn SJ. (2018) Statistical pitfalls of personalized medicine. *Nature. 2018;563(7733):619-621.*

Snapinn, S. and Q. Jiang (2011). "On the clinical meaningfulness of a treatment's effect on a time-to-event variable." Stat Med **30**(19): 2341-2348.