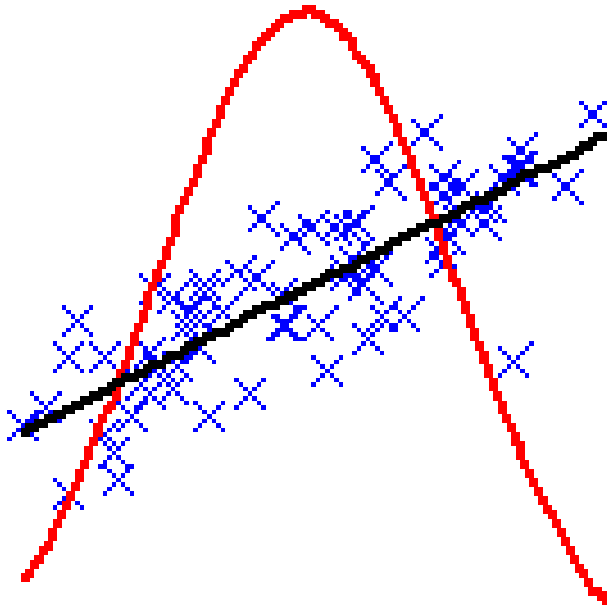
Clinical trials are about comparability, not generalisability



Stephen Senn

@stephensenn

stephen@senns.uk

Outline

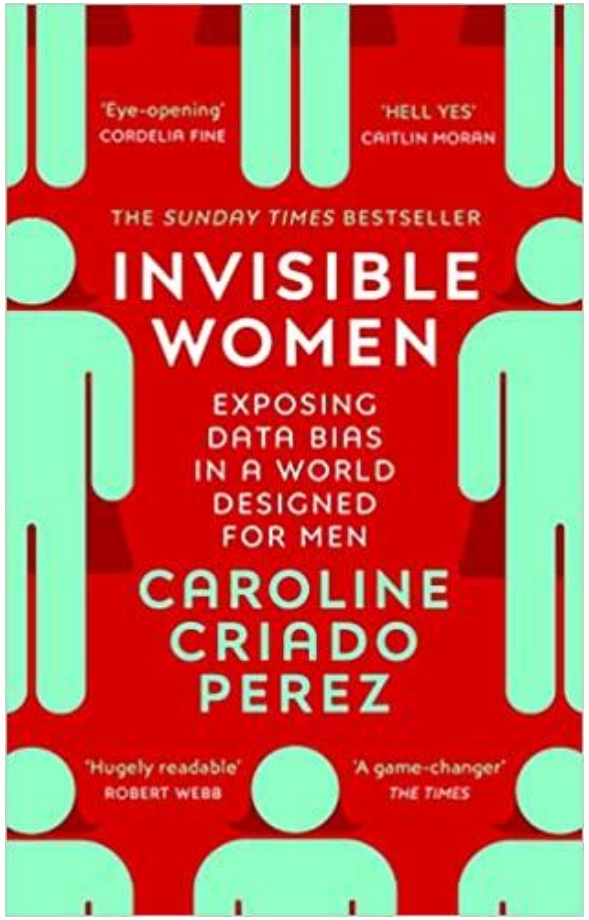
- Three examples
  - Bioequivalence and sex
  - Rats
  - The Lanarkshire Milk Experiment
- Lessons
- Two sources of confusion
  - Statistical frameworks
  - Objectives of analysis
    - Why the sampling paradigm is irrelevant
    - Implications for analysis
- Translating/transporting effects
- Conclusions

Three Examples

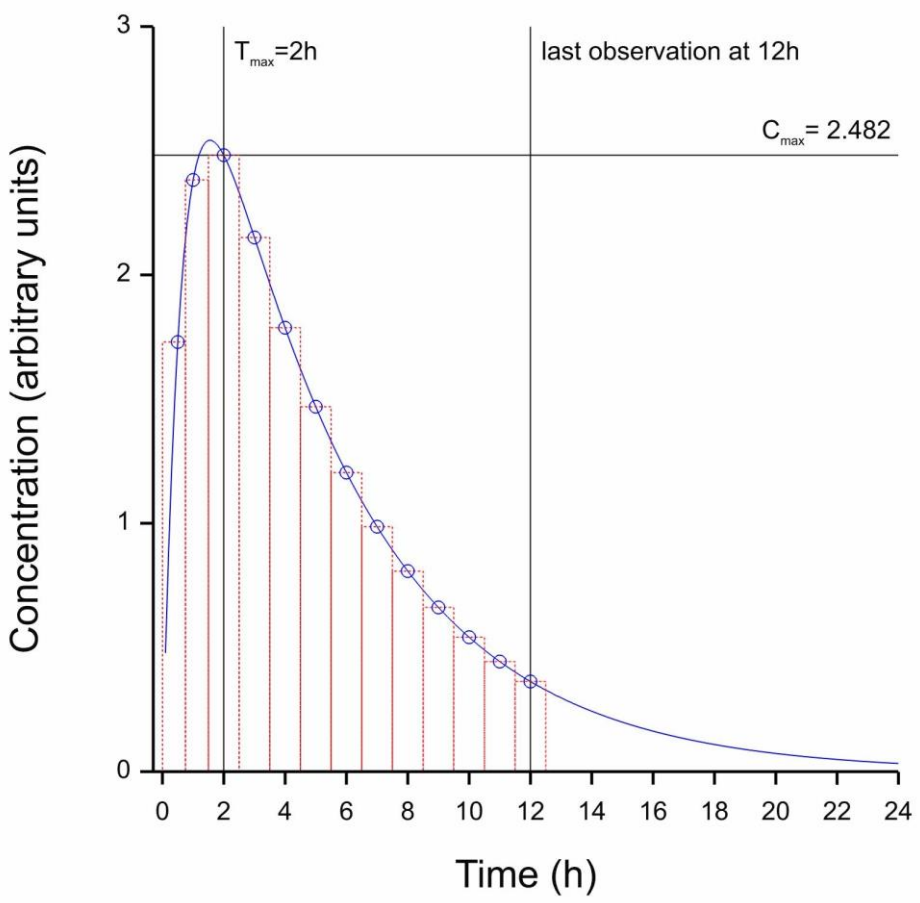
Volunteers, Rats, Schoolchildren

Invisible women?

- Interesting account (2019) of many ways that women are badly served by data systems
  - For example, housework is (still) mainly carried out by women but(ironically) not counted towards gross domestic product
- Makes many important points
- I am just going to criticise one



Vive la différence?

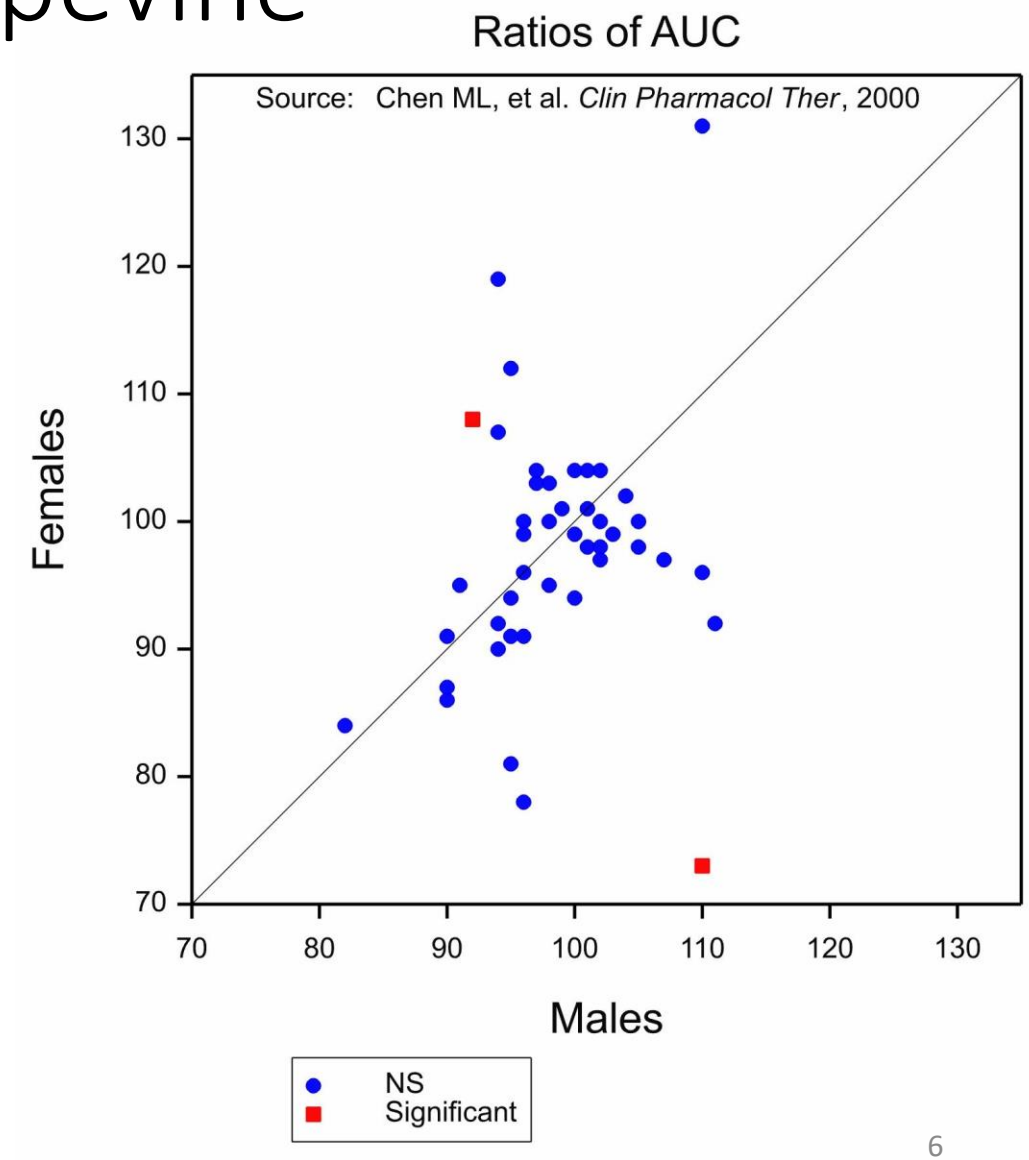


Illustrative example of a concentration time curve that might be used to calculate AUC

"Drug trials for generic drugs are much less rigorous than original trials…and they are conducted 'almost exclusively' in young adult males…sure enough, in 2002 the FDA's Center for Drug Evaluation showed 'statistically significant differences in men and women in bioequivalence for most generic drugs compared with reference drugs"[56] (PP203-204)

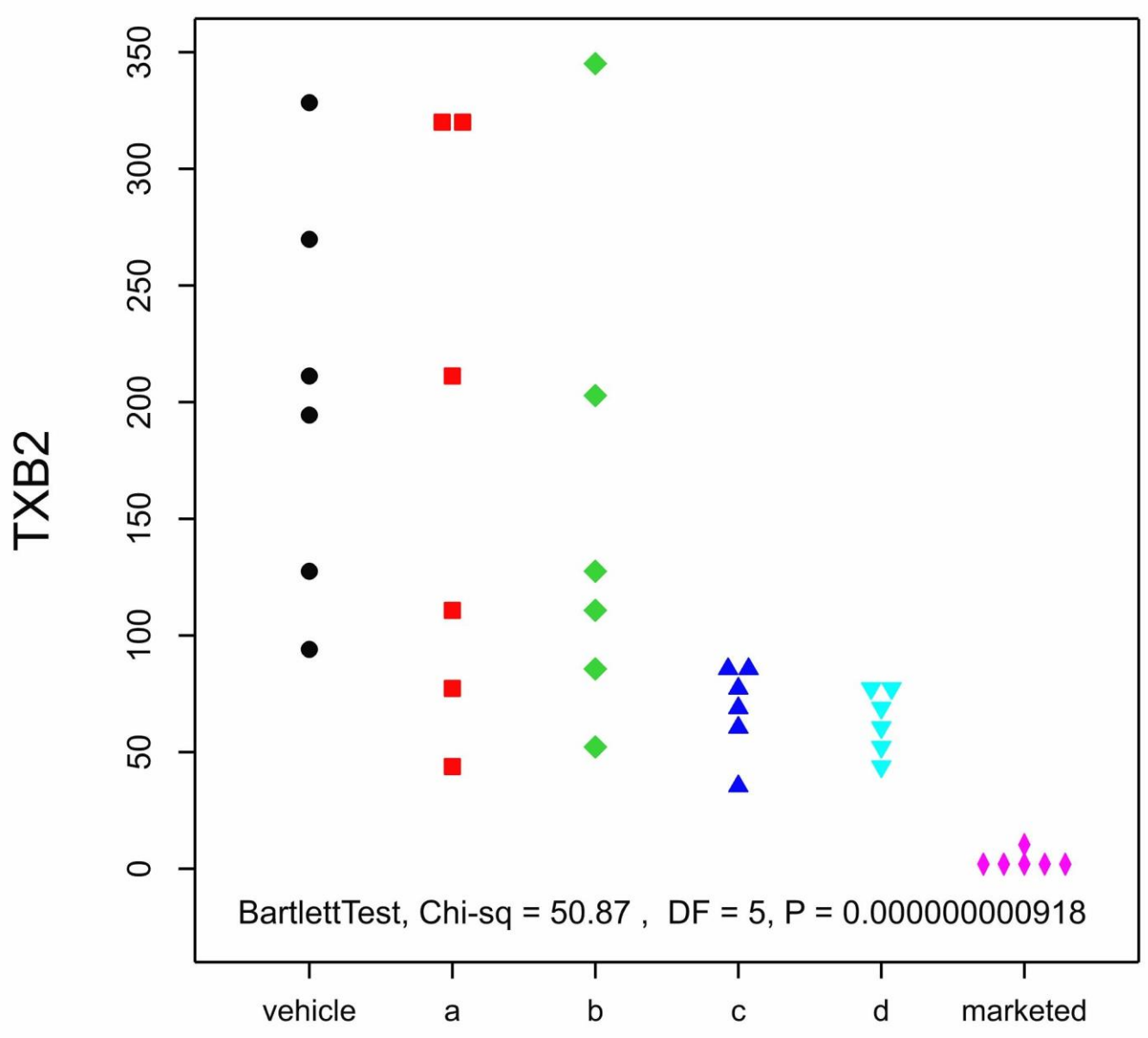
Hearing it through the grapevine

Citer	Citation
Criado-Perez 2019	McGregor 2017
McGregor, 2017	Koren et al 2013
Koren et al, 2013	Chen et al 2000
Chen et al 2000	“A statistically significant (P < .05) sex-by-formulation interaction was observed in two data sets for AUC and five for Cmax”



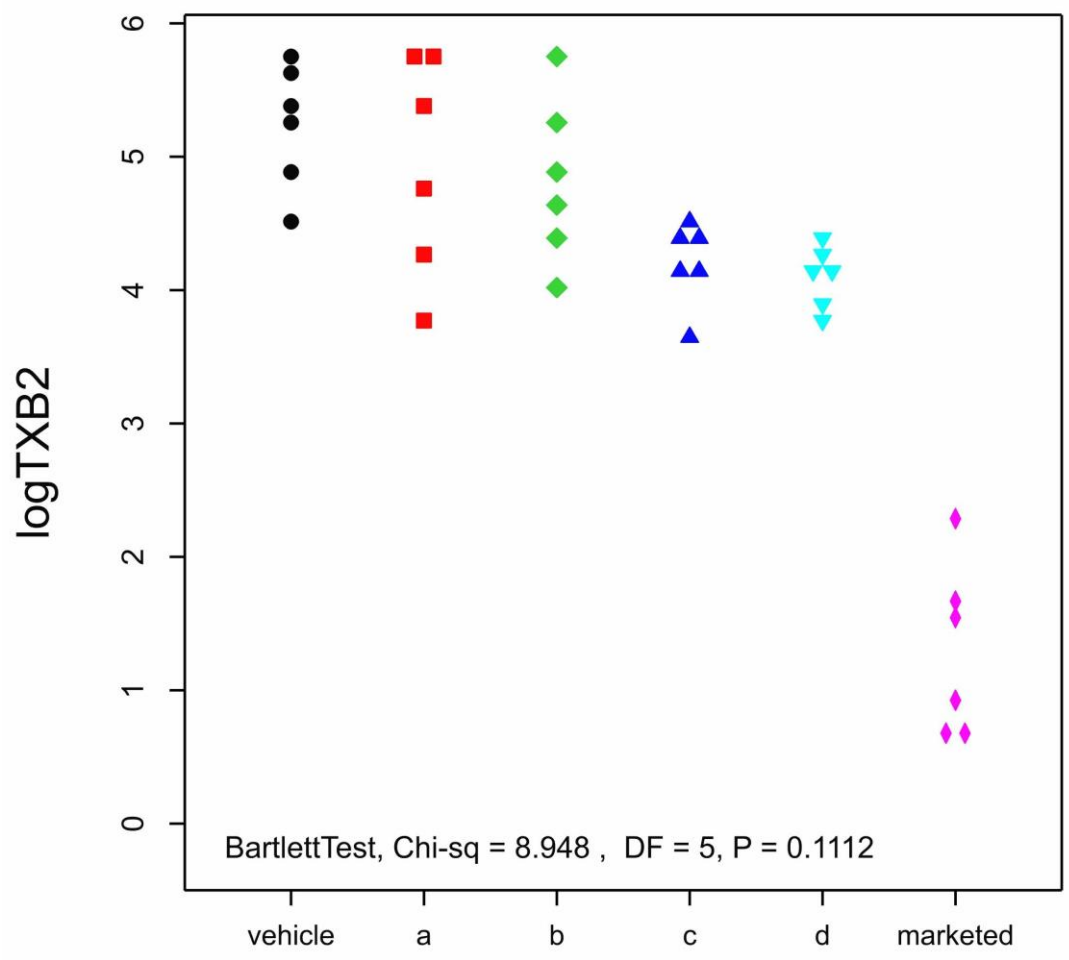
Benchmarks

- 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 were sacrificed and TXB2 is measured.



A little bit of magic

- There is clearly a difference between treatments
- However, it is not just means that are affected but variances
- Bartlett's test highly significant
- Variances seem to increase with the mean
- Suggests a log-transformation
- This seems to restore homoscedasticity



The Lanarkshire Milk Experiment (LME)

- More than 18,000 pupils in 67 schools in Lanarkshire were enrolled in an experiment on nutrition in 1930
- Effect of raw and pasteurised milk given over four months on height and weight studied
- Authors pooled all controls together
- Experiment attracted interest of Fisher, Student, Karl Pearson and others

1:1 allocation within schools
Approximately ¼ of schoolchildren received raw ¼ pasteurised and ½ acted as controls.

Raw milk schools

- Control
- Raw milk

Pasteurised milk schools

- Control
- Pasteurised milk

Lessons

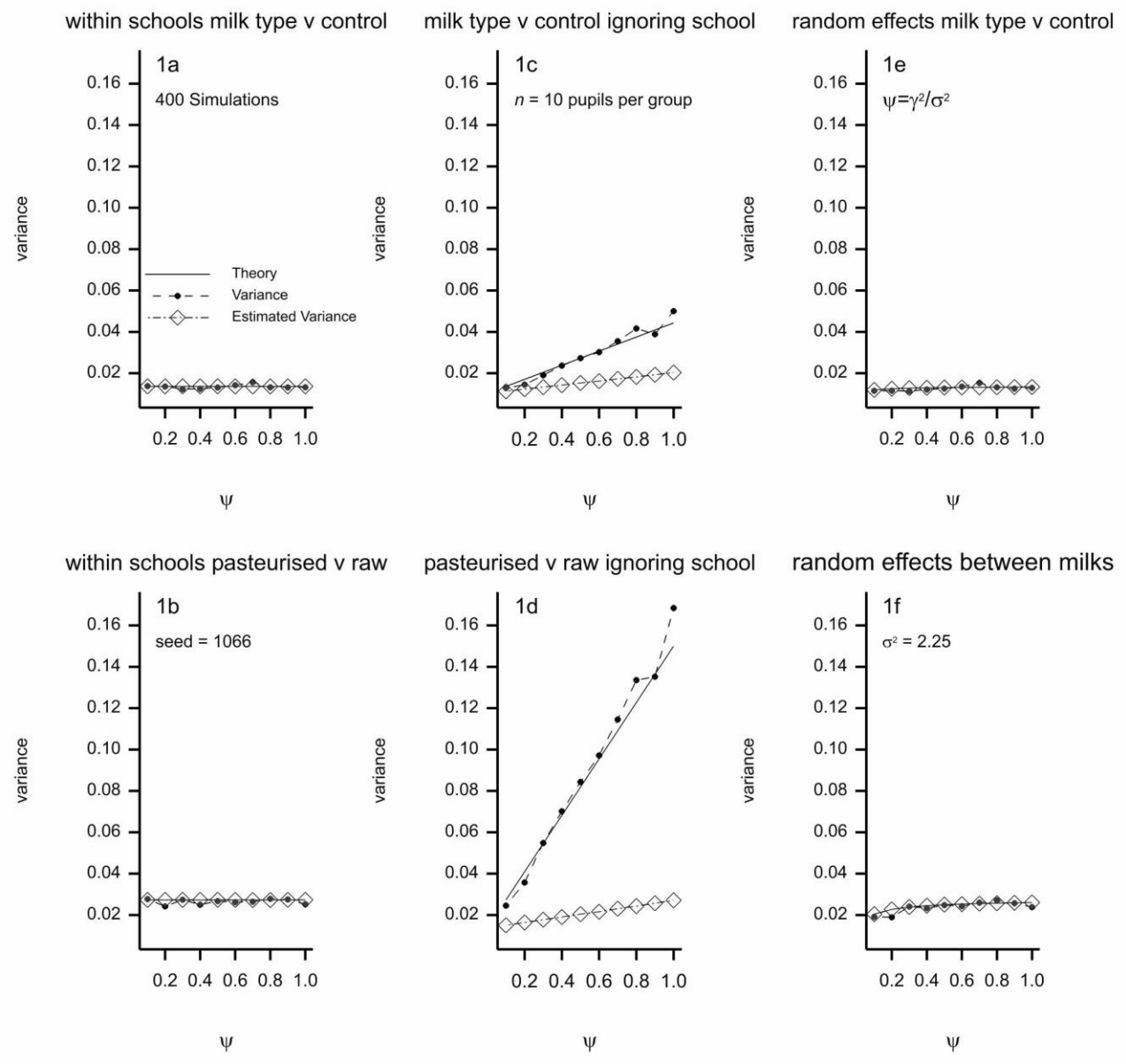
The good, the bad and the useful

Experimental inference

Example	What I don't worry about	What I worry about	What I conclude
Invisible women	That bioequivalence trials don't use a random sample of patients.	Getting the scale right.	Take care in translating results but don't fuss about sex differences in bioequivalence.
Rats	Whether the rats are representative (of what?)	Was the design appropriate? Was there only one cage per treatment? Was the standard error calculated appropriately?	Make sure you get the analysis right and then think carefully about how to use the results.
Lanarkshire Milk Experiment	Were the schoolchildren representative?	Was the comparison fair? Was the allocation subverted?	Take care that the analysis reflects the design.

Tricky statistics

- The LME is a cluster allocated incomplete blocks design,
- This requires careful analysis that reflects variation between and within schools.
- Pooling all controls together a) is inefficient and b) will lead to incorrect calculation of standard errors.
- This has nothing to do with sampling theory. It's an allocation issue.



Sources of confusion

Frameworks and questions

Statistical frameworks

Frameworks

Framework	Notes
Random sampling	Data are generated from a population of interest using probability sampling
Randomisation	Data are generated by random assignment of treatment to units
Linear model	Conditional on predictors, treated as fixed, outcomes are assumed to have random errors
Multivariate analysis	Variates vary together and we study certain conditional distributions

Problems and issues

- Random sampling is most often invoked in elementary courses
- It has little if any relevance to clinical trials
  - Randomisation is used **not** random sampling
- In practice linear models are very often used
  - and sometimes (particularly in epidemiology) incorrectly described as being *multivariate*
- This is all rather confused
  - Sometimes it does not matter but sometimes it does

Possible questions a clinical trial might answer

Questions

- Q1. Was there an effect of treatment in this trial?
- Q2. What was the average effect of treatment in this trial?
- Q3. Was the treatment effect identical for all patients in the trial?
- Q4. What was the effect of treatment for different subgroups of patients?
- Q5. What will be the effect of treatment when used more generally (outside of the trial)?

Implications

- Clinical trials are best at answering Q1 and Q2
- Some trials can answer Q3
- Usually resources won't permit addressing Q4 adequately
- Of course Q5 is very important but any answer will be highly speculative and in particular if you can't answer Q1 and Q2 well, you are likely to get it wrong

Two extremes

Predictive (Q5)

- The population is taken to be ‘patients in general’
  - Of course this really means future patients
  - They are the ones to whom the treatment will be applied
- We treat the patients in the trial as an appropriate selection from this population
  - This does not require them to be typical but it does require additivity of the treatment effect
  - Additivity implies low heterogeneity of effect on the scale used

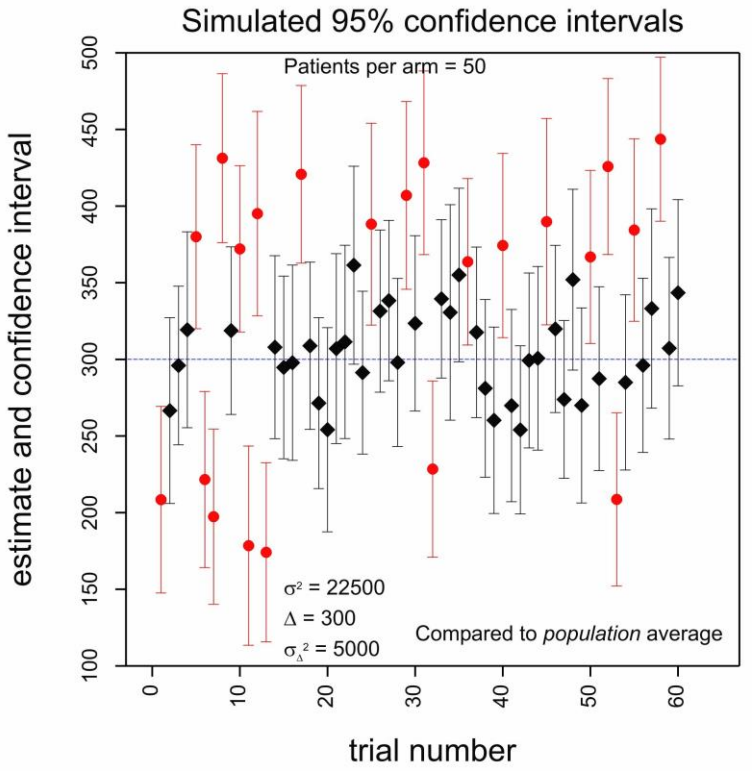
Causal (Q1 & Q2)

- We take the patients as fixed
- We want to know what the effect was for them
- Unfortunately there are missing counterfactuals
- What would have happened to control patients given intervention and vice-versa
- The population is the population of all possible allocations to the patients studied

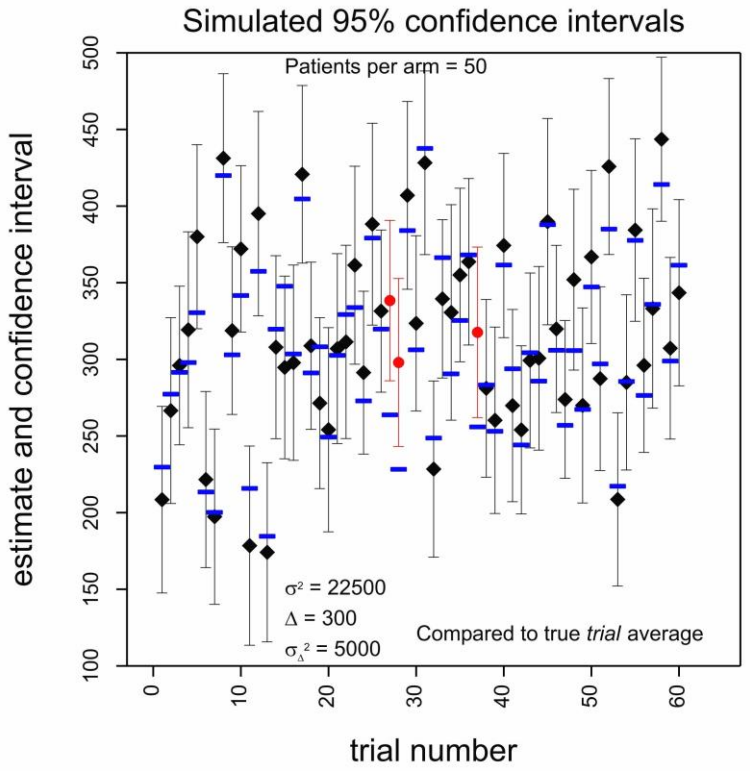
Coverage probabilities for two questions

Heterogeneity of treatment effect large

Predictive



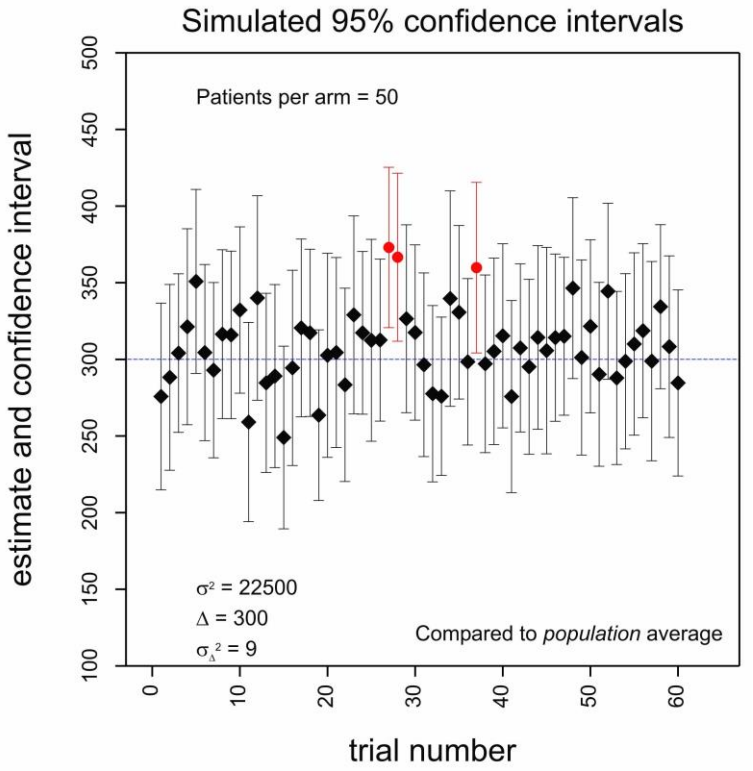
Causal



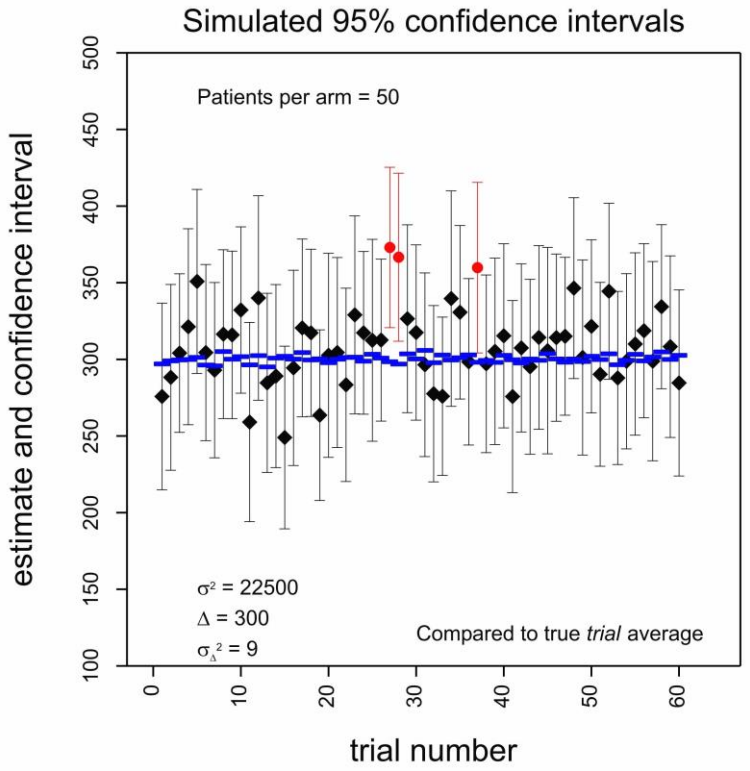
Coverage probabilities for two questions

Heterogeneity of treatment effect small

Predictive



Causal



Yates & Cochran

J Ag Science 1938

On representativeness

"..it is usually impossible to secure a set of sites selected entirely at random…the deliberate inclusion of sites representing extreme conditions may be of value. Lack of randomness is then only harmful in so far as it results in the omission of certain types and in the consequent arbitrary restriction of the range of conditions. In this respect scientific research is easier than technical research."

P558

On generalisation

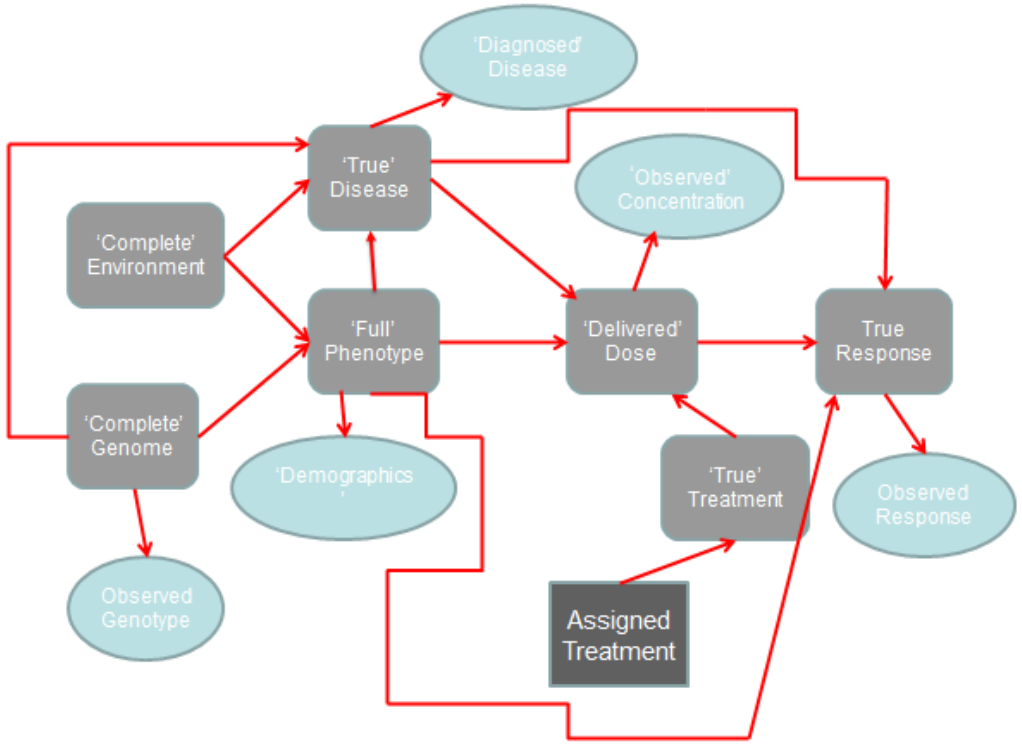
"If the mean square for varieties is significant, this indicates the significance of the average differences of the average difference of the varieties over the particular set of places chosen. If varieties ×place is also significant…it is clear that the choice of place must effect the magnitude of the average difference between varieties…even if varieties ×place is not significant, this fact cannot be taken as indicating no variation in the varietal differences." P560

Basic thesis

- At some level we all know that these questions are not the same and cannot be answered using one analysis
  - We tend not to recognise this formally
- As we get more ambitious our formal statements tend to underestimate the real uncertainty
- Nevertheless, it is worth answering the simple questions well
- The causal inference movement is increasingly ambitious in attempting to answer the more difficult questions
- This work is very interesting but in my opinion …
  - It rarely treats uncertainty seriously
  - It rarely deal with study effects
  - It incorrectly assume that representative sampling takes place
  - We are in danger of overstating the promise of ‘big data’ in particular from observational studies

Lost in translation?

From trials to practice



2

From trials to clinical practice

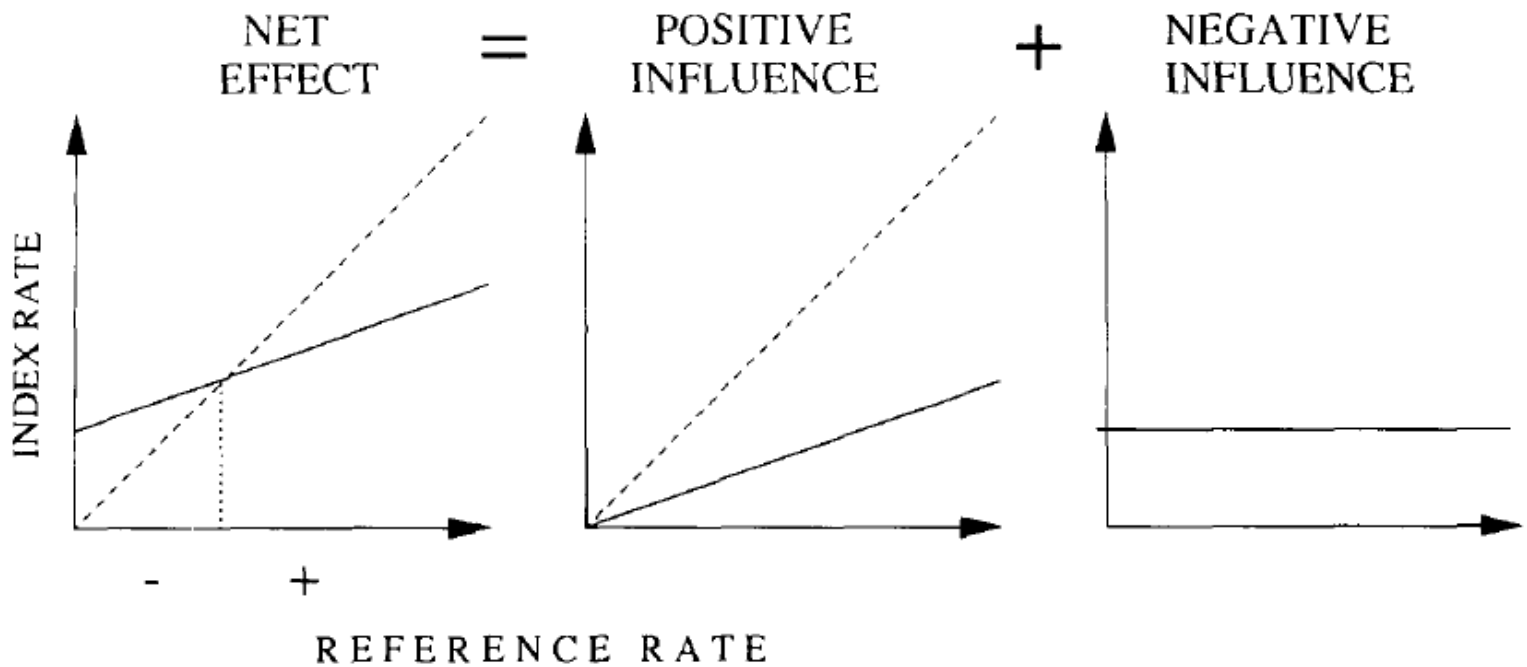


Figure 1 Direction of net effects. If a treatment does not affect the course of disease in any way, the index and reference rates of an outcome in an RCT will, when plotted on an *X–Y* graph across subgroups of patients at different levels of risk, fall (on the average) on the identity line (---). If there is a positive influence that reduces the reference rate with a constant proportion (middle panel) and a negative influence that induces a risk that is uniform over subgroups (right panel), their net effect will sum up as shown in the left panel.

Controlled Clinical Trials, 1989

Interspecies scaling

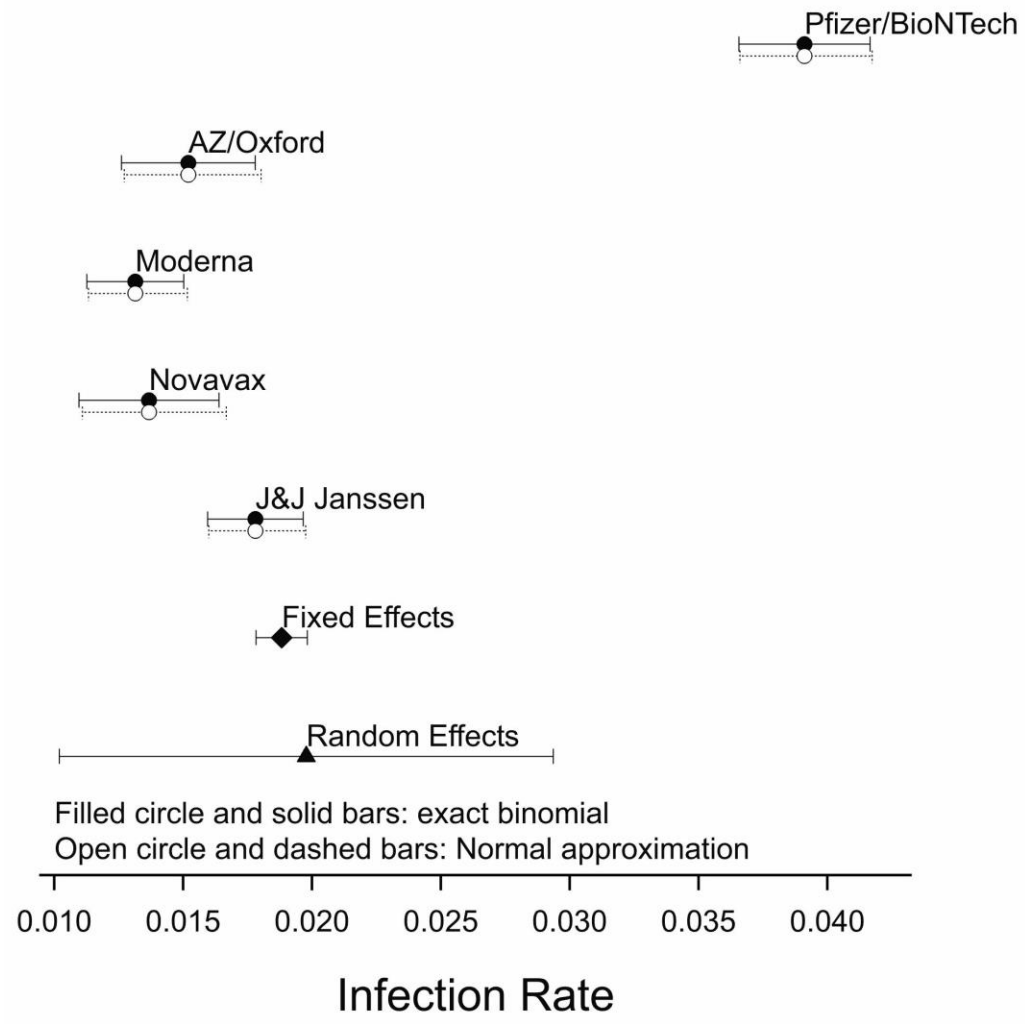
- Kleiber's "law" indicates that metabolic rate for a great number of species is proportional to body weight to the power of ¾
- Anderson and Holford have applied this to clearance also
  - Much of the sex difference between men and women can be explained by this
  - Also the difference between children >2 years old and adults
- Note that we rarely exploit this sort of thing in real world *treatment* , where it would translate into different doses for different individuals but prefer to chase subgroup interactions within studies.

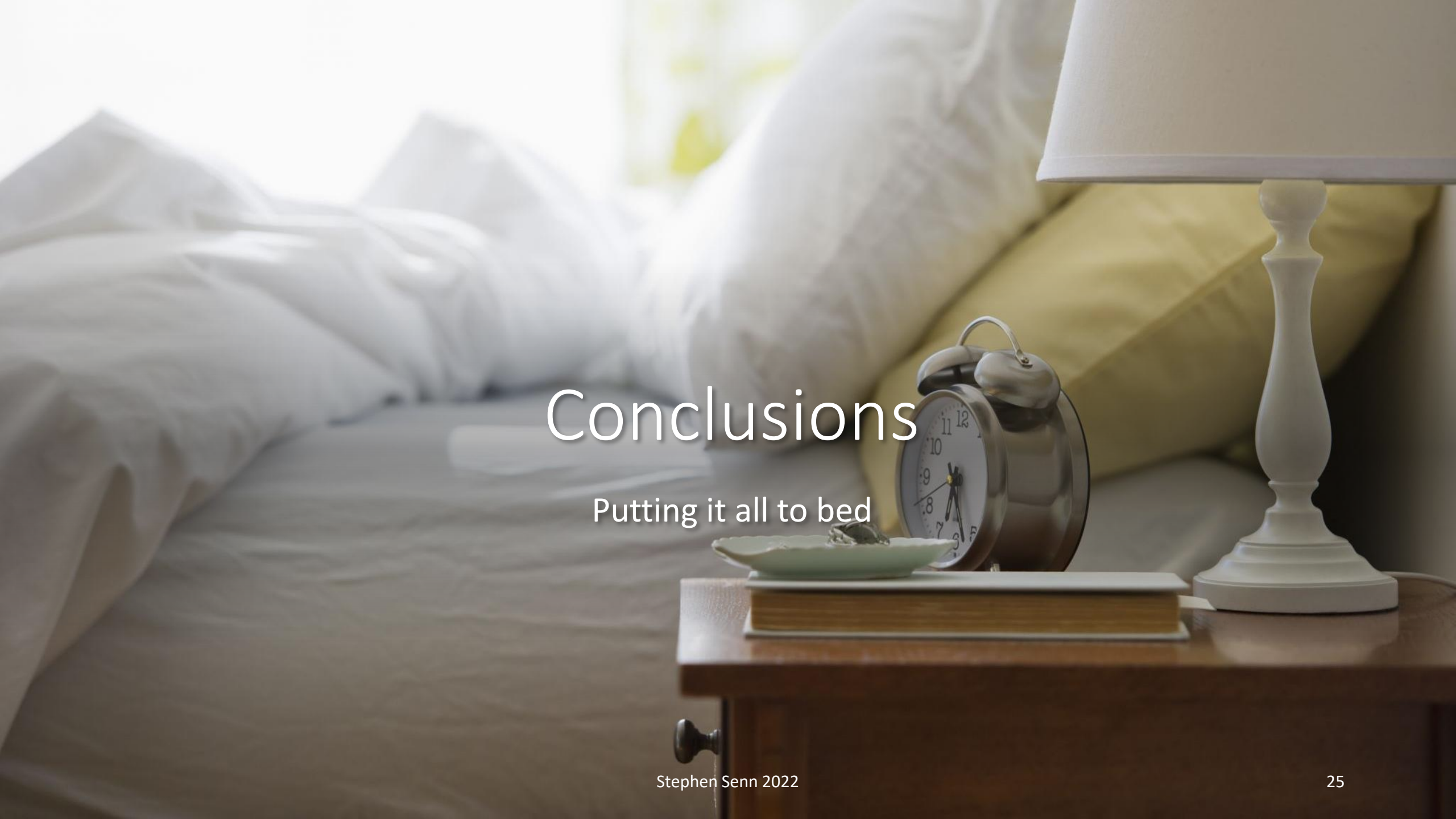
Max Kleiber 1893-1976

Vaccine trials

- Pfizer/BioNTech planned to recruit approximately 42,000 subjects to their big COVID-19 vaccine trial
- But most of these subject are completely irrelevant to inference
- What matters is cases
- You use cases under control, Y_c to estimate how many cases you would have had, had the vaccine not worked
- The number of cases the vaccine prevented is estimated (approximately) as Y_v-Y_c and the divisor is (approximately) Y_c, not the number of subjects
- VE=(Y_C-Y_V)/Y_C
- Case rates will vary dramatically over time as the pandemic waxes and wanes

Variation in control group rates between several large vaccine trials





Conclusions

Putting it all to bed

Lessons

- We have very poor if any control of the *presenting process*
  - It is a complete misconception that inclusion criteria determine this.
- But we can achieve (fairly) tight control of the *allocation algorithm*
- The allocation algorithm guides not only how we should estimate effects but also estimate standard errors of effects
  - And our intended model also guides out choice of design
- Study main effects can be very large and **must** be allowed for
  - Be very sceptical of any structural causal model that leverages observational data by invoking a *faithfulness* assumption
- Translation to practice is
  - Difficult
  - Requires using appropriate scale and models
  - Gives no guarantees
- We must not let the unattainable best become the enemy of the good