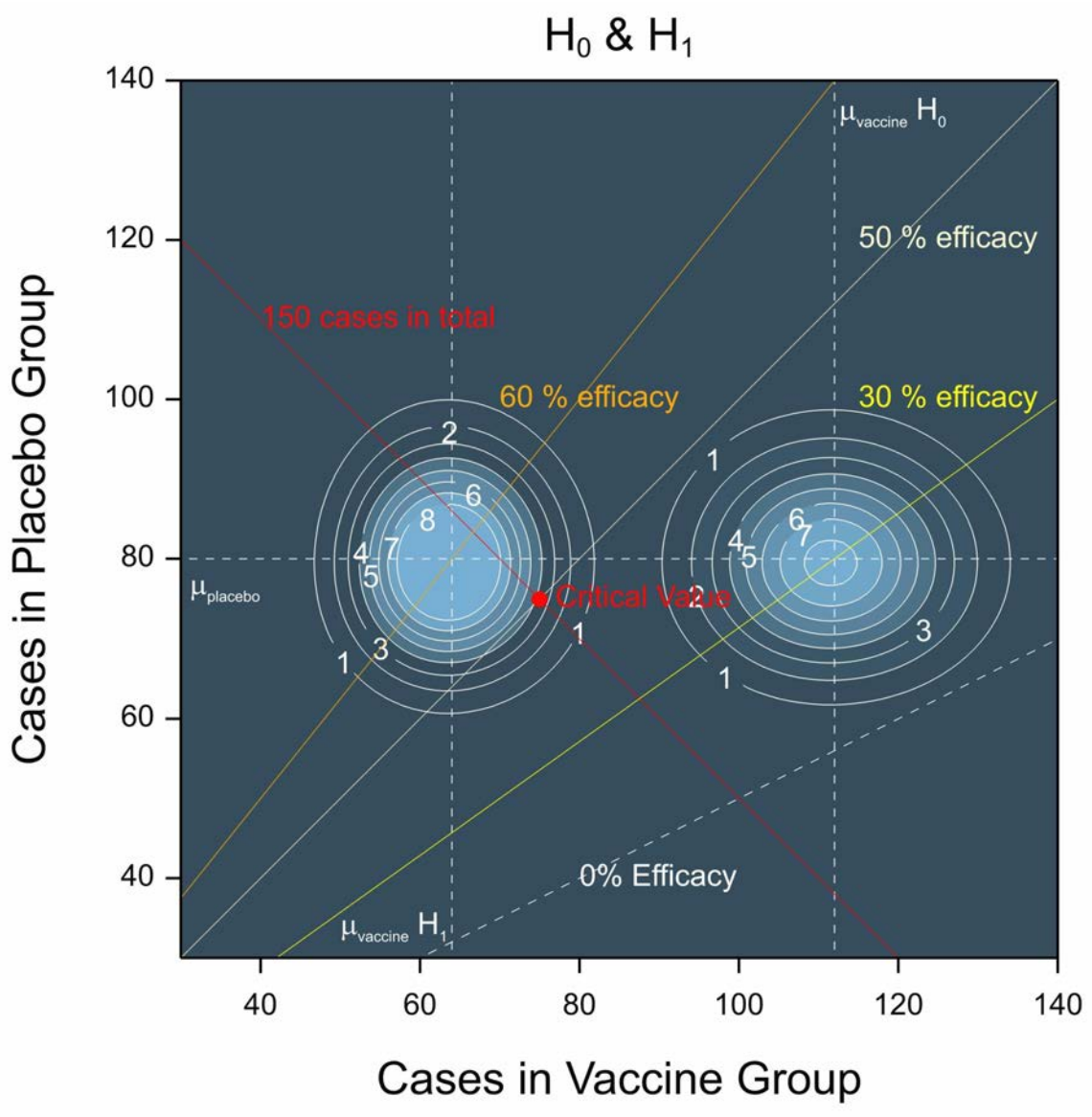
Randomisation and Control in the Age of Coronavirus

Stephen Senn

Edinburgh



Acknowledgment

I thank Deborah Mayo for the invitation to give this lecture and also for frequently hosting my blogs, many of which are reflected in it.

Outline

Part I Randomisation in general

- Some common criticisms
- Randomisation explained by a game of chance
- The criticisms addressed

Part II Randomisation & coronavirus

- Poor practice in therapeutic trials
- The limitations of historical controls
- Ongoing vaccine trials

Part I

Randomisation in general

It’s all about balance

A foundational and strong assumption of RCTs (once the sample is chosen) is the achieving-good randomisation assumption…

…As long as there are important imbalances we cannot interpret the different outcomes between the treatment and control groups as simply reflecting the treatment’s effectiveness. Researchers thus need to better reduce the degree of known imbalances – and thus biased results – by better using, for example, larger samples and stratified randomisation.

Alexander Krauss, 2018.

Indefinitely many confounders?

Even if there is only a small probability that an individual factor is un-balanced, given that there are indefinitely many possible confounding factors, then it would seem to follow that the probability that there is some factor on which the two groups are unbalanced (when remember randomly constructed) might for all anyone knows be high.

Worrall, 2002

Balance requires an infinity of randomisations

In order to begin to address this problem of confounding factors, the randomization would have to be repeated an indefinite number of times But in RCTs, randomization is usually done only once Thus, defenders of the special causal ability of RCTs make claims about the epistemic powers of actual RCTs based on what would happen in ideal RCTs (The presence of the phrase "in the long run" betrays the slide to theoretical claims) If we were to randomize forever, the limiting-average effect of the treatment would yield information of the sort desired by RC'T enthusiasts . “

Borgerson, 2009

It’s all about homogeneity

…any difference in outcome between the test group and the control group should be caused by the tested interventions, since all other differences should be homogenously distributed between the two groups
E Rocca E and RL Anjum 2020

The common mistake

- All of these commentaries start from the false assumption that error elimination is the necessary goal of experimentation
- However, error is inevitable *and must be accepted as such*
- Statistical methods calculate not just a point estimate but an associated measure of uncertainty
  - Standard error
  - Confidence interval
- What the statistical argument is designed to produce is a valid probabilistic statement
- The critics need to show that balance is imperfect but that the allowance that analysis makes for imbalance is inadequate

A Game of Chance

- Two dice are rolled
  - Red die
  - Black die
- You have to call correctly the odds of a total score of 10
- Three variants
  - Game 1 You call the odds and the dice are rolled together
  - Game 2 The red die is rolled first, you are shown the score and then must call the odds
  - Game 3 The red die is rolled first, you are not shown the score and then must call the odds

Total Score when Rolling Two Dice

		Red Die Score					
		1	2	3	4	5	6
Black Die Score	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Variant 1. Three of 36 equally likely results give a 10. The probability is 3/36=1/12.

Total Score when Rolling Two Dice

		Red Die Score					
		1	2	3	4	5	6
Black Die Score	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Variant 2: If the red die score is 1,2 or 3, probability of a total of 10 is 0.

If the red die score is 4,5 or 6 the probability of a total of10 is 1/6.

Variant 3: The probability = (½ x 0) + (½ x 1/6) = 1/12

The Morals

Specifically

- All three games are uncertain
- Perfect prediction is impossible
- Valid probabilistic prediction *is* possible
- You can't treat game 2 like game 1.
  - You must condition on the information you receive in order to act wisely
  - You must use the actual data from the red die
- You can treat game 3 like game 1.
  - You can use the *distribution in probability* that the red die has

More generally

- You can make valid statements about the outcomes of clinical trials provided that you embrace uncertainty
- You can't ignore an observed prognostic covariate in analysing a clinical trial just because you randomised
  - Analysis of covariance may be employed
- You can ignore an unobserved covariate precisely because you did randomise

You are not free to imagine anything at all

- Imagine that you are in control of all the thousands and thousands of covariates that patients will have
- You are now going to allocate the covariates and their effects to patients
  - As in a **simulation**
- If you respect the actual variation in human health that there can be, you will find that the net total effect of these covariates is bounded

$$Y = \beta_0 + \tau Z + \beta_1 X_1 + \cdots \beta_k X_k + \cdots$$

Where *Z* (which is equal to either 0 or 1) is a treatment indicator, τ is the treatment effect, and the *Xs* are covariates. You are not free to arbitrarily assume any values you like for the *Xs* and the βs because the variance of *Y* must be respected.

What happens if you don’t pay attention

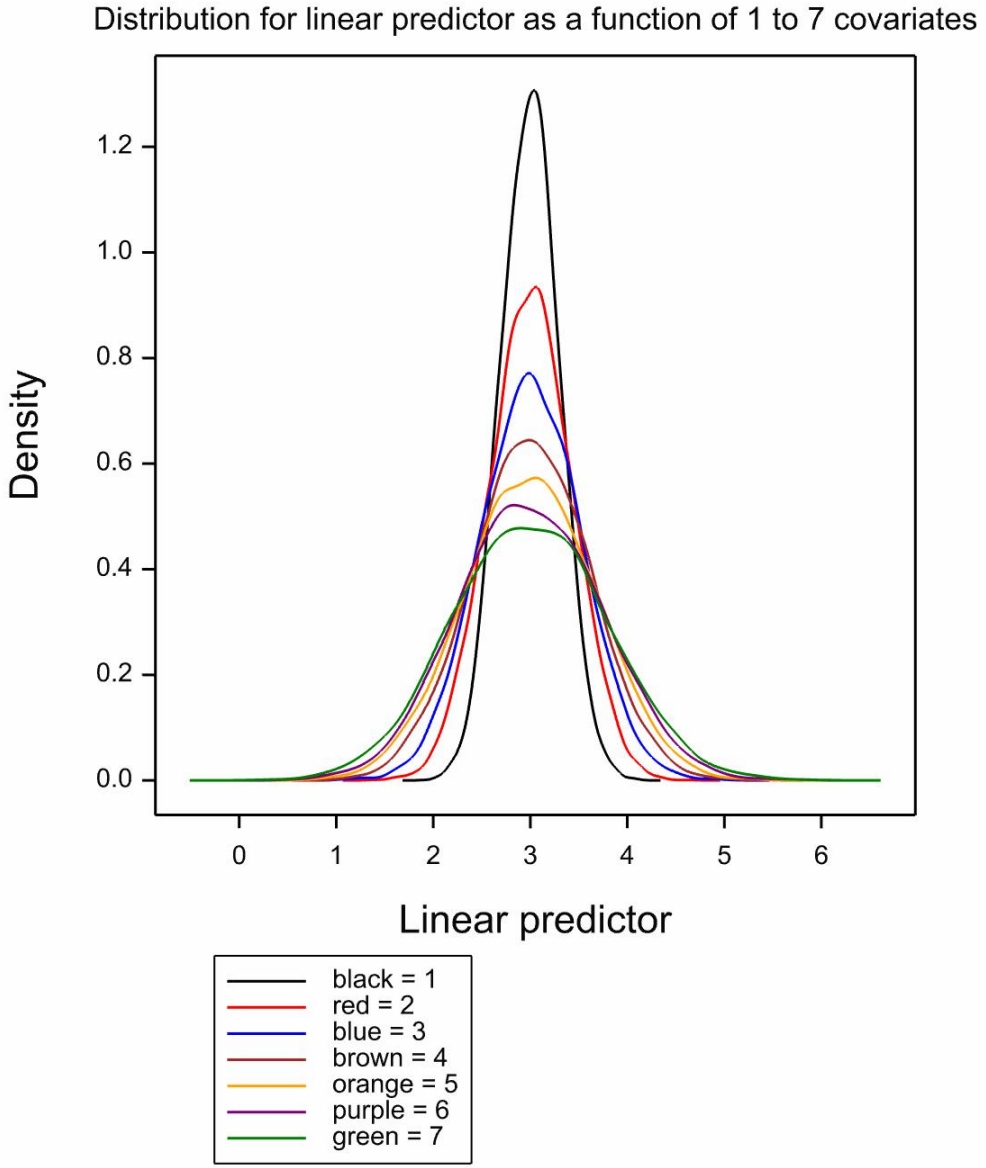
Simulation of the linear predictor as the number of covariates increases from 1 to 7

However, the variance of each covariate is the same and the coefficient is the same and the covariates are assumed orthogonal

We can see that the variance of the predictor keeps on increasing

The values soon become impossible

But in reality the total contribution that the covariates can make is bounded



In fact this is pointless

Look at the equation again

$$Y = \beta_0 + \tau Z + \beta_1 X_1 + \cdots \beta_k X_k + \cdots$$

We have to take care how we choose the parameters of the $X_1, .. X_k \; and \; \beta_1 \ldots \beta_k$ and what we have to guide us are the possible values of *Y*. But suppose we re-write the equation

$$Y = Y^* + \tau Z$$

Where

$$Y^* = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \cdots$$

Now there is only one unknown, Y^* not indefinitely many, and this is all that we need to consider

So Worrall's Argument is Wrong

Worrall's argument boils down to saying that if a series is infinite its sum can't be bounded.

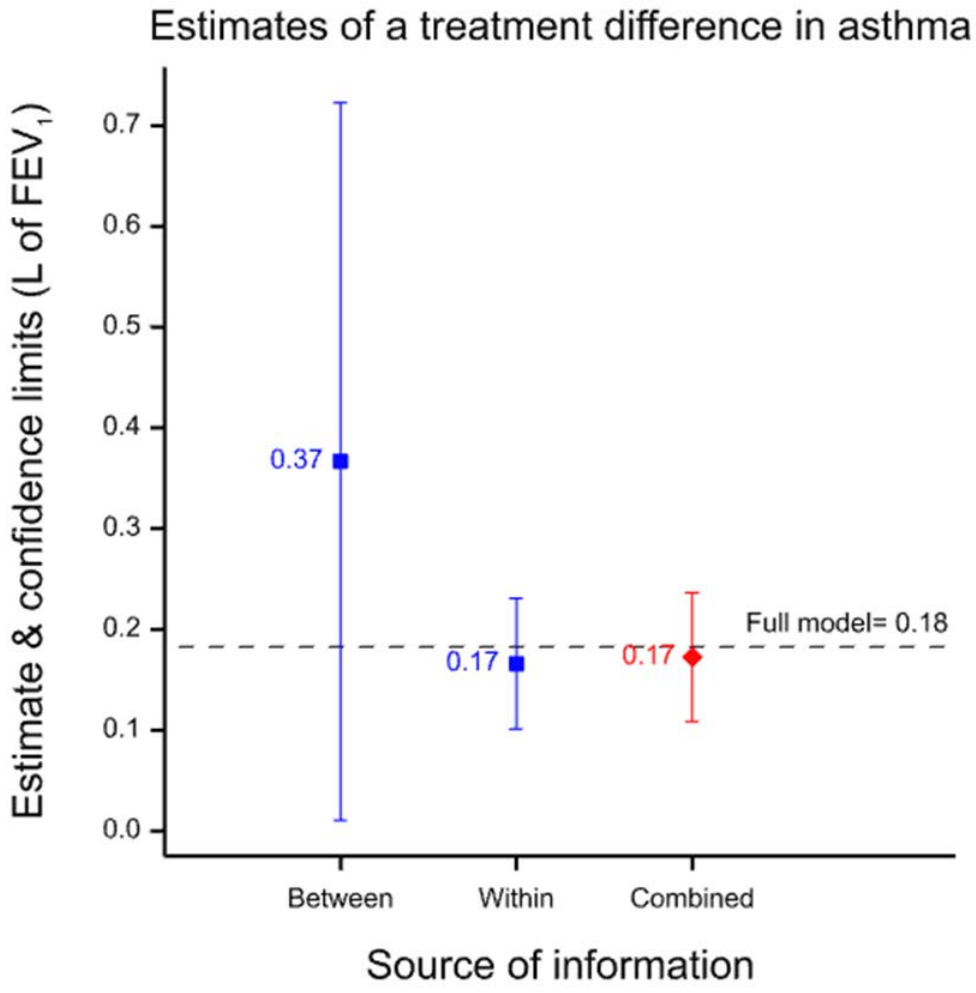
But how about the sum

$$S = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} \ldots \ ?$$

An incomplete block trial in asthma

- Seven treatments were compared in a five period cross-over trial
- Outcome was forced expiratory volume in one second (FEV_1)
- Each patient received five of the seven treatment in a given sequence to which they were randomised
- I shall compare two of the treatments
  - ISF24 and MTA6
- I can do this in two different ways
  - Using 71 patients who were given both of these treatments
  - Using 37 patients who were given ISF24 but not MTA6 and 37 who were given MTA6 and ISF24
- The first *within*-patient comparison balances for tens of thousands of genetic and environmental factors and the second *between*-patient doesn't

Stop obsessing about balance



- The within-patient estimate is much more precise than the between-patient estimate
- However, it is not more valid
- The extra ‘balance’ is reflected in a narrower confidence interval
- The analysis of the between-patient data has made an allowance for imbalance

The Shocking Truth

- The validity of conventional analysis of randomised trials does not depend on covariate balance
- It is valid because *they are not* perfectly balanced
- If they were balanced the standard analysis would be *wrong*
- The cross-over trial balances for *30,000 genes and all history to date for each patient*
- The parallel group trial does not
- *Because* it does not, it posts a higher variance
- If we have taken care to balance all these tens of thousands of covariates, analysing the result as if we hadn't is wrong

An analogy

To criticise randomisation because it does not guarantee balance is like criticising a bridge-builder because thermal expansion means the bridge will stretch, while overlooking that the design explicitly allows for this fact.

The analysis of randomised trials makes an allowance for imbalance.

Part II

Randomisation in the age of coronavirus

The pressure to drop standards

- The urgency of dealing with the pandemic has led some to propose or even employ trials that do not use concurrent control
- Some external standard is employed instead
  - Historical data
  - Patients who refused to be randomised
  - Patients in other centres
- This is generally a bad idea
- If it is to be done it requires careful analysis as to how exactly the data should be used

One armed trials
Two common strategies using historical controls

One sample test approach

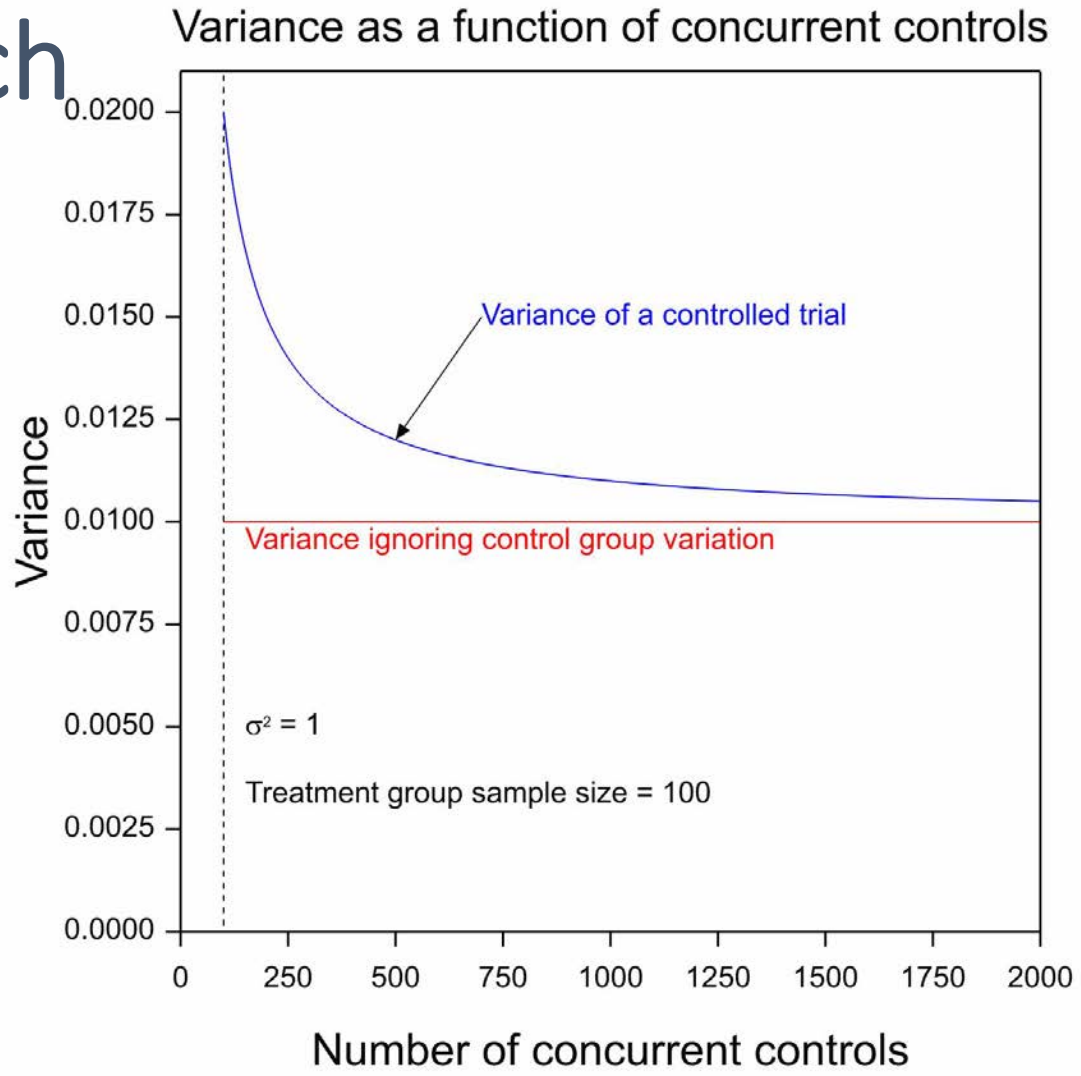
- Nominate a target ‘response’ based on historical data
  - For example: 20% complete remission by 6 months
- Carry out test to show that response for new treatment is ‘statistically significantly better’ than historical standard.

Two sample test approach

- Find a set of historical controls regarded as being comparable
  - based either on a single study or a number of studies
- Treat these as if they were the control arm in a standard (parallel group) clinical trial

One sample test approach

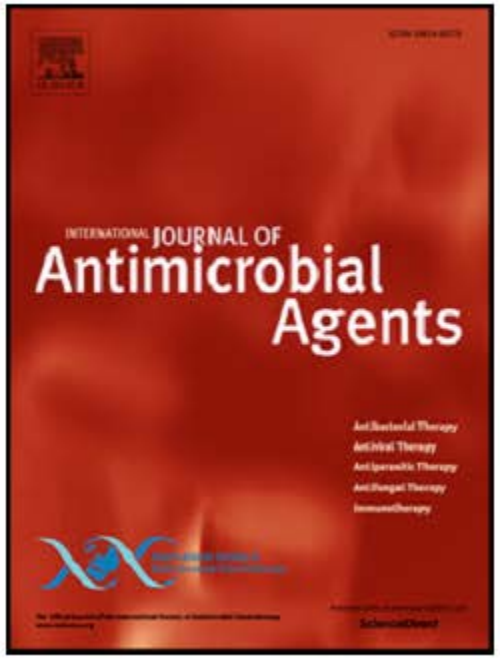
- Except for within-patient studies we basically never carry out one sample tests
  - Two values are reduced to a single difference
- We always allow not only for uncertainty in the test group but also in the control group
- So the only way that this would be approximately right is if the historical rate could be regarded as coming from an extremely large concurrent control group
  - As (say) in a trial in which at least 20 times as many patients had been randomised to control as to treatment
- Clearly these conditions cannot be satisfied
- So there is only one way to describe such an analysis
- **Cheating**



Journal Pre-proof

Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial

Philippe Gautret , Jean-Christophe Lagier , Philippe Parola , Van Thuan Hoang , Line Meddeb , Morgane Mailhe , Barbara Doudier , Johan Courjon , Valérie Giordanengo , Vera Esteves Vieira , Hervé Tissot Dupont , Stéphane Honoré , Philippe Colson , Eric Chabrière , Bernard La Scola , Jean-Marc Rolain , Philippe Brouqui , Didier Raoult



French Confirmed COVID-19 patients were included in a single arm protocol from early March to March 16th, to receive 600mg of hydroxychloroquine daily and their viral load in nasopharyngeal swabs was tested daily in a hospital setting. Depending on their clinical presentation, azithromycin was added to the treatment. Untreated patients from another center and cases refusing the protocol were included as negative controls.

Statistical differences were evaluated by Pearson's chi-square or Fisher's exact tests as categorical variables, as appropriate. Means of quantitative data were compared using Student's t-test. Analyses were performed in Stata version 14.2.

Using historical controls

- When one naively uses historical controls as is they were concurrent one implicitly makes strong assumption
- That historical results are comparable with current ones despite differences in
  - Time
  - Place
  - Equipment
  - Personnel
- The following example shows that this is dangerous

The TARGET study

- One of the largest studies ever run in osteoarthritis
- 18,000 patients
- Randomisation took place in two sub-studies of equal size
  - Lumiracoxib versus ibuprofen
  - Lumiracoxib versus naproxen
- Purpose to investigate CV and GI tolerability of lumiracoxib

Baseline Demographics

	Sub-Study 1		Sub Study 2	
Demographic Characteristic	**Lumiracoxib n = 4376**	**Ibuprofen n = 4397**	**Lumiracoxib n = 4741**	**Naproxen n = 4730**
Use of low-dose aspirin	975 (22.3)	966 (22.0)	1195 (25.1)	1193 (25.2)
History of vascular disease	393 (9.0)	340 (7.7)	588 (12.4)	559 (11.8)
Cerebro-vascular disease	69 (1.6)	65 (1.5)	108 (2.3)	107 (2.3)
Dyslipidaemias	1030 (23.5)	1025 (23.3)	799 (16.9)	809 (17.1)
Nitrate use	105 (2.4)	79 (1.8)	181 (3.8)	165 (3.5)

Baseline Chi-square P-values

	Model Term		
Demographic Characteristic	**Sub-study (DF=1)**	**Treatment given Sub-study (DF=2)**	**Treatment (DF=2)**
Use of low-dose aspirin	< 0.0001	0.94	0.0012
History of vascular disease	< 0.0001	0.07	<0.0001
Cerebro-vascular disease	0.0002	0.93	0.0208
Dyslipidaemias	<0.0001	0.92	<0.0001
Nitrate use	< 0.0001	0.10	<0.0001

Lessons from TARGET

- If you want to use historical controls you will have to work *very* hard
- You need at least two components of variation in your model
  - Between centre
  - Between trial
- And possibly a third
  - Between eras
- What seems like a lot of information may not be much

Toy example of historical controls

The data

- Acute Myeloid Leukaemia
- 18 studies available with a total number of 1232 patients
- Number of responders available

The strategy

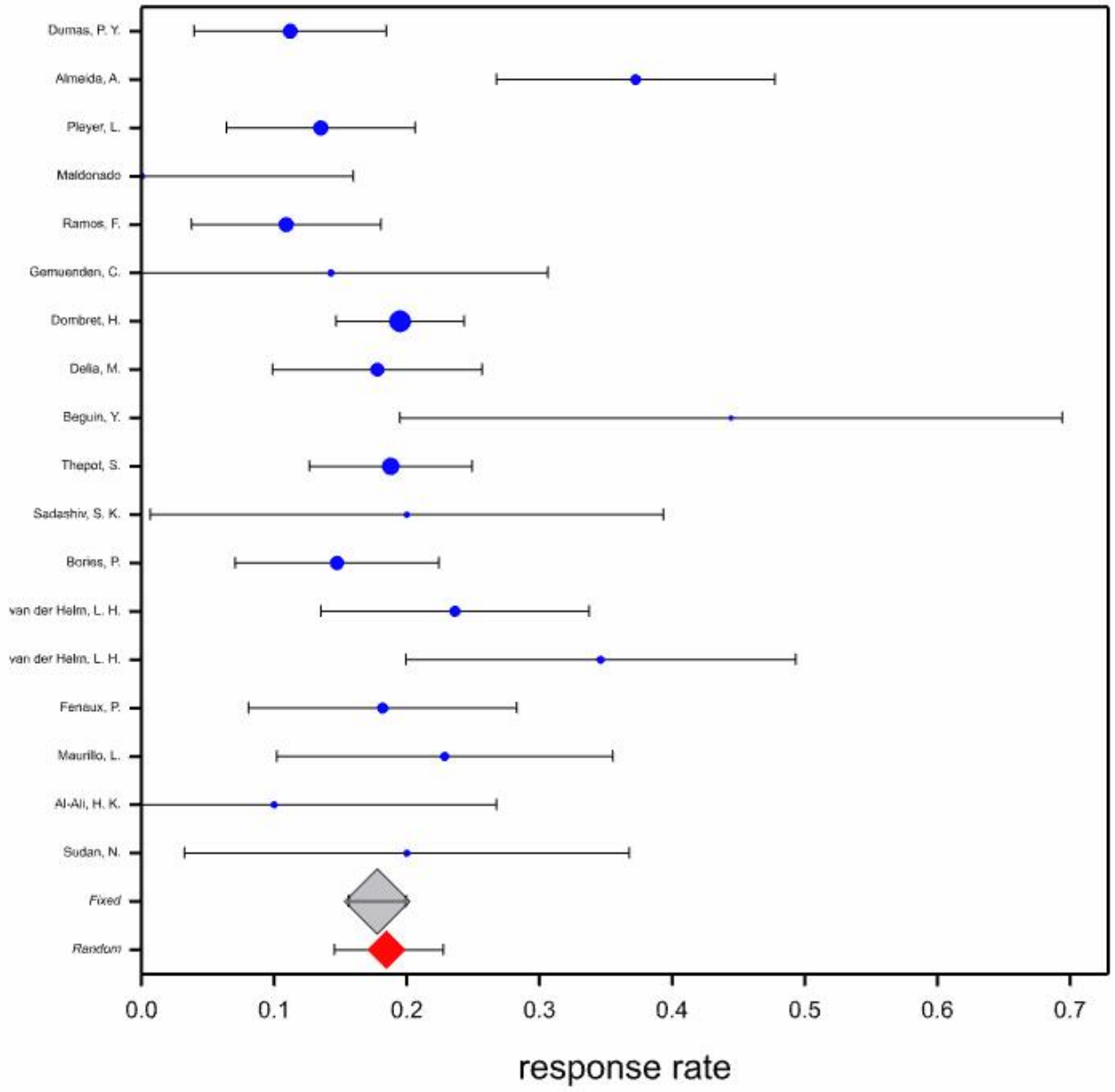
- Calculate response rate for each trial
- Calculate per trial standard error using naive overall response rate but per trial number of patients
- Put the results inside a simple random effects meta-analysis

Toy example

Warning: Clearly since these are response rates negative values are impossible

CIs for some trials would include negative probabilities

Nevertheless, the example is instructive of the dangers of assuming you have lots of data



Number of patients = 1232

Number of concurrent controls that would beat any number of historical ones = 49

Recommendations when planning the selection and integration of historical controls

- Identification, pre-specification & agreement on a suitable historical data-set
  - Because otherwise you could pick and choose your historical controls
- An agreed, enforceable and checkable plan for recruiting the experimental arm in advance of doing so
  - Because otherwise you could selectively recruit to your advantage
- A finalised analysis plan prior to beginning the trial
  - Because blinding is impossible
- Use of a hierarchical model with sufficient complexity
  - Because many components of variation are involved
- Emphasis on number of historical trials rather than patients
  - Because otherwise components of variation cannot be estimated

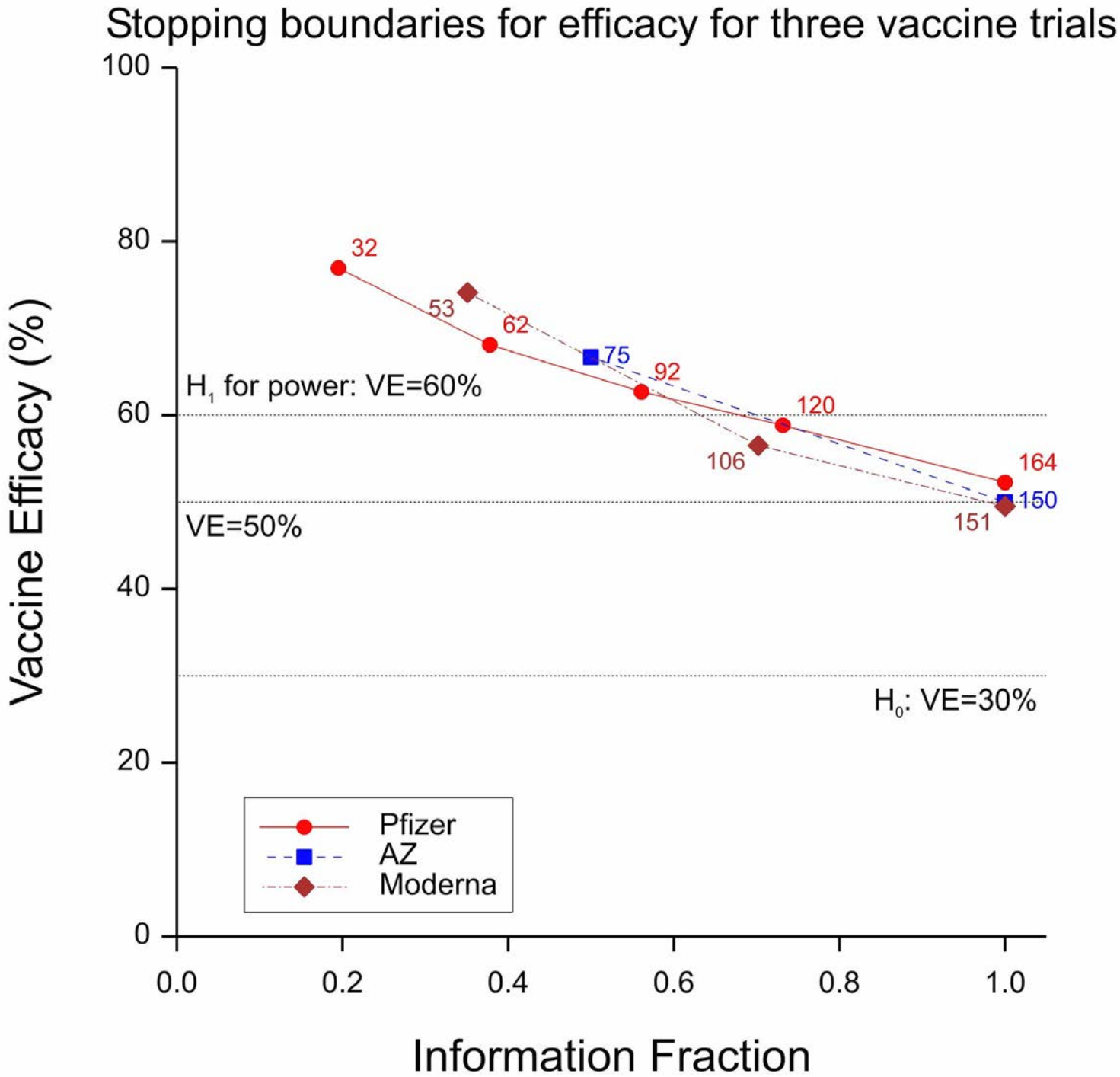
Conclusion
Historical data? Think cluster not parallel

- Some good work has been done recently that recognises that variation between studies is the key
- For example work by Heinz Schmidli and colleagues in a Bayesian framework
- But whether you model the variance components in a Bayesian or a frequentist way is less important than recognising what they are
- The care you have to take with various matters of organisation is as important as the analysis

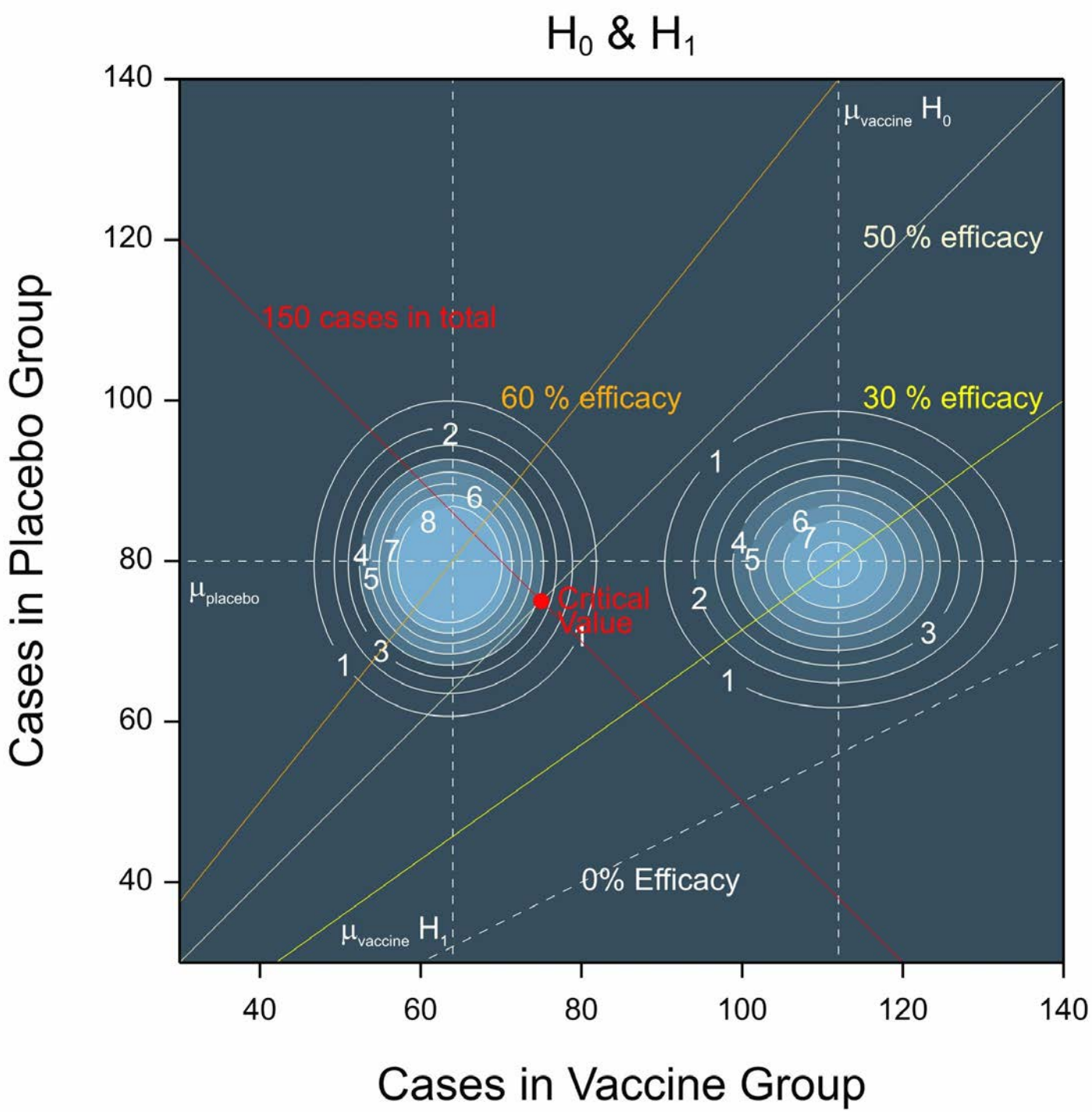
You can be rich in data but poor in information

Three large vaccine trials

Feature	Pfizer/ BioNTech	AstraZeneca	Moderna
Number on vaccine	21,999	20,000	15,000
Number on control	21,999	10,000	15,000
Number of events targeted	164	150	151
Assumed control rate	0.65	0.80	0.75
Null hypothesis efficacy %	30	30	30
Efficacy for power %	60	60	60
Planned looks	5	2	3
Inferential approach	Bayesian	Frequentist	Frequentist


Stopping boundaries for efficacy for three vaccine trials
Vaccine Efficacy (%)
Information Fraction
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
H1 for power: VE=60%
VE=50%
H0: VE=30%
32
53
62
75
92
106
120
151
164
150
Pfizer
AZ
Moderna

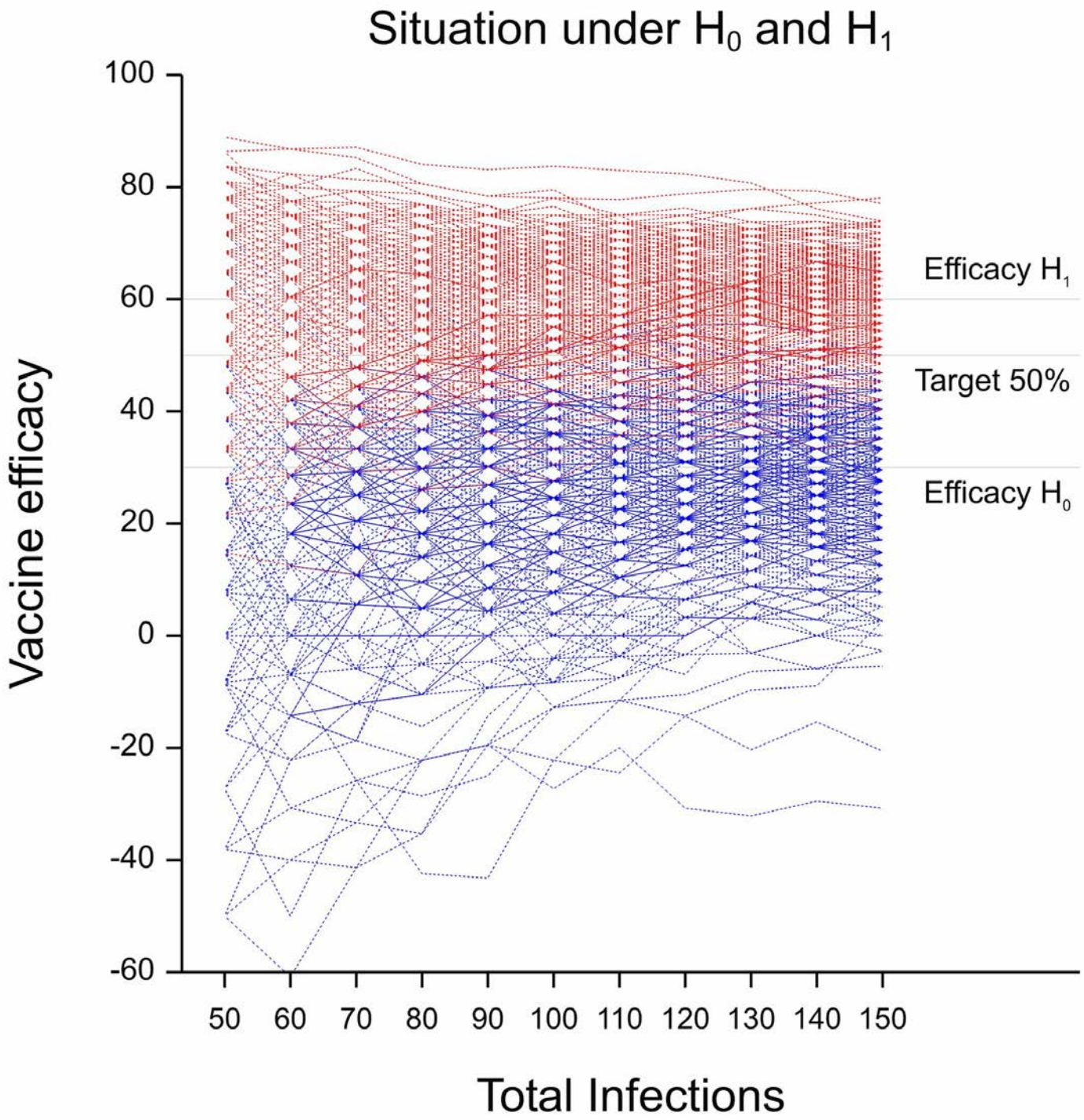
Representation of the AstraZeneca trial (ignoring complications of an interim look)



Some points from the AZ trial

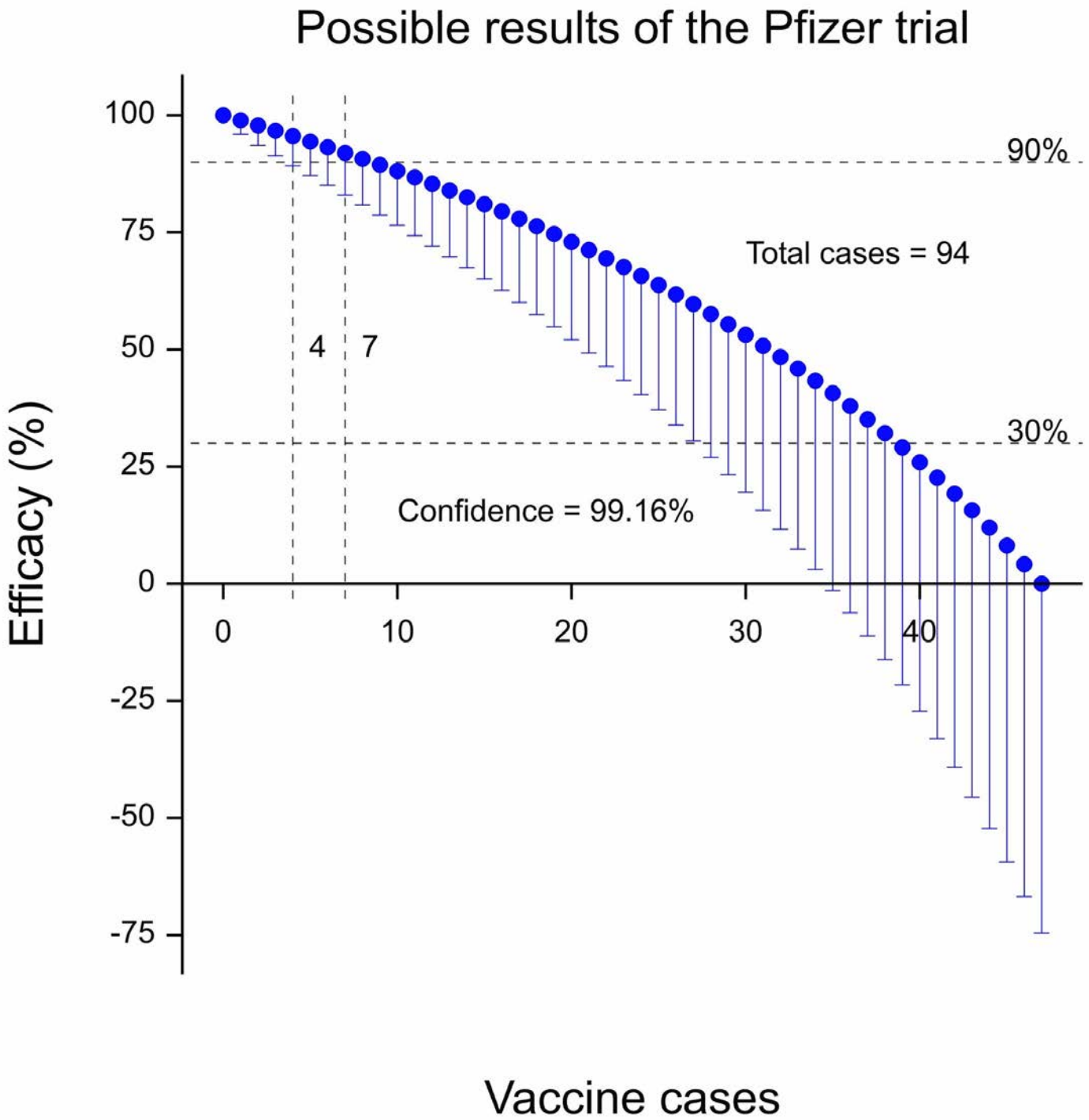
- We need to have assumptions about typical rates of infection to plan the trial
  - Because we need to estimate how many subject to recruit in order to have the requisite number of infections
- These rates are very much a matter of guesswork
- Since we have concurrent controls our (possibly very poor) guesses are not part of the inference
- If we did not have concurrent control, our (possibly very poor) guesses would have to provide the yardstick by which efficacy was judged

Simulation of the Moderna trial (ignoring complications of interim looks)



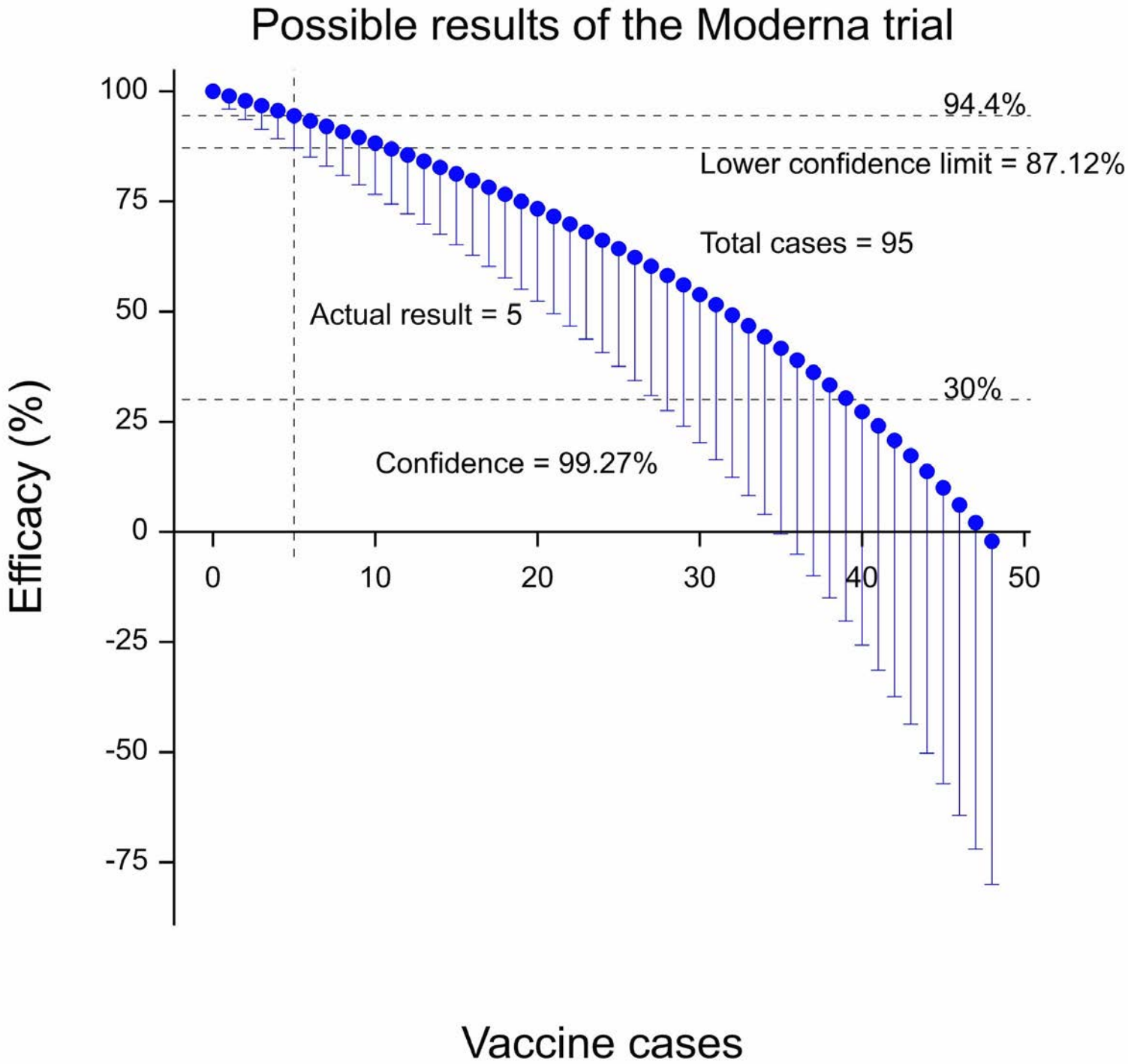
On 9 November 2020 Pfizer claimed 90% efficacy based on 94 cases

Confidence interval allows for the sequential nature of the trial



On 16 November 2020 Moderna claimed 94.5% efficacy based on 94 cases

Confidence interval allows for the sequential nature of the trial



Conclusion

- The obsession with balance has caused many to mistake how randomised trials work
- We need to know how well we know what we think we know and this is what statistical analysis is meant to provide
- Good design is key to making good conclusions
- In the age of coronavirus we have had a mixed record
  - Doing poorly in some therapeutic trials
  - Doing well in some large vaccine trials

A word from the master

Nevertheless, though its logical necessity is easily apprehended, the question of the validity of the estimates of error used in tests of significance was for long ignored and is still often overlooked in practice.

RA Fisher, *The Design of Experiments,* p42