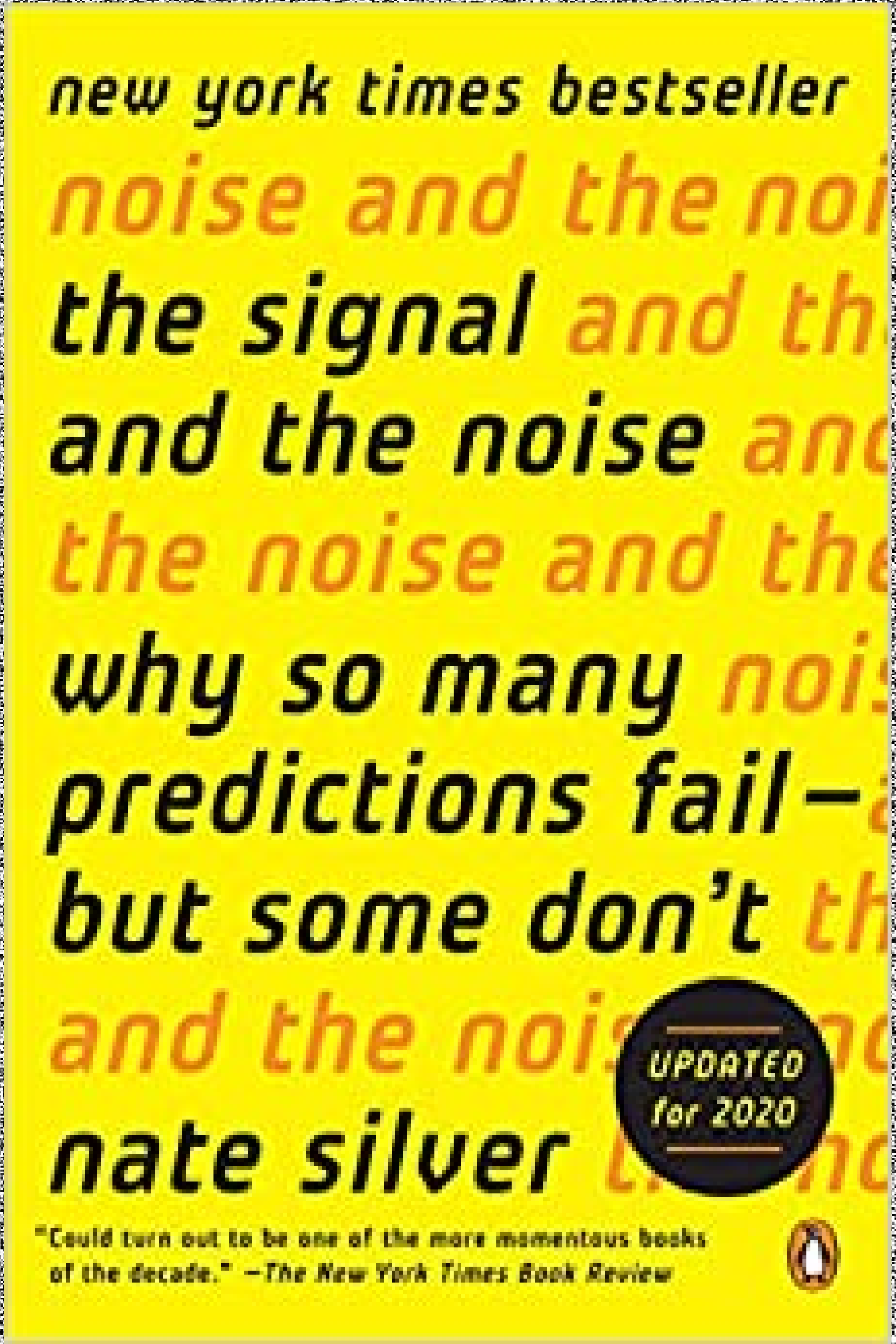
The scientific reproducibility crisis - winner's curse and other causes

Matthias von Herrath, Novonordisk and

Ondrej Libiger, J&J

(Matthias@lji.org)



new york times bestseller

the signal and the noise

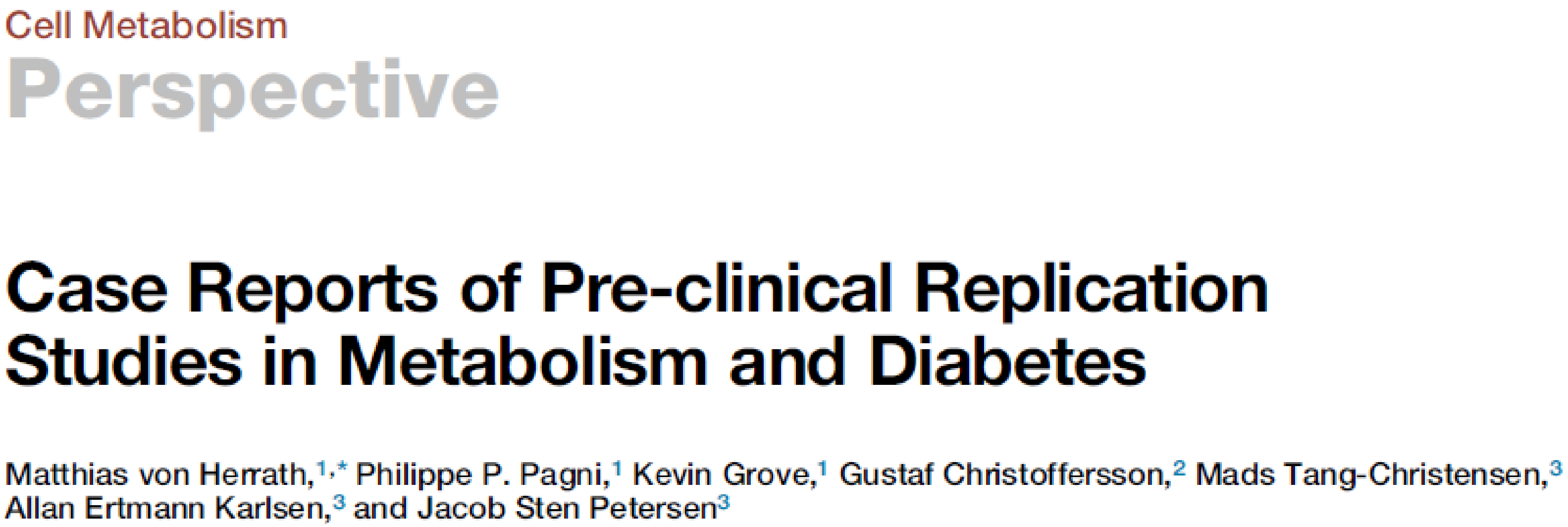
why so many predictions fail—but some don't

nate silver

UPDATED for 2020

"Could turn out to be one of the more momentous books of the decade." —*The New York Times Book Review*

Majority of published research does not replicate, lacks robustness



Cell Metabolism

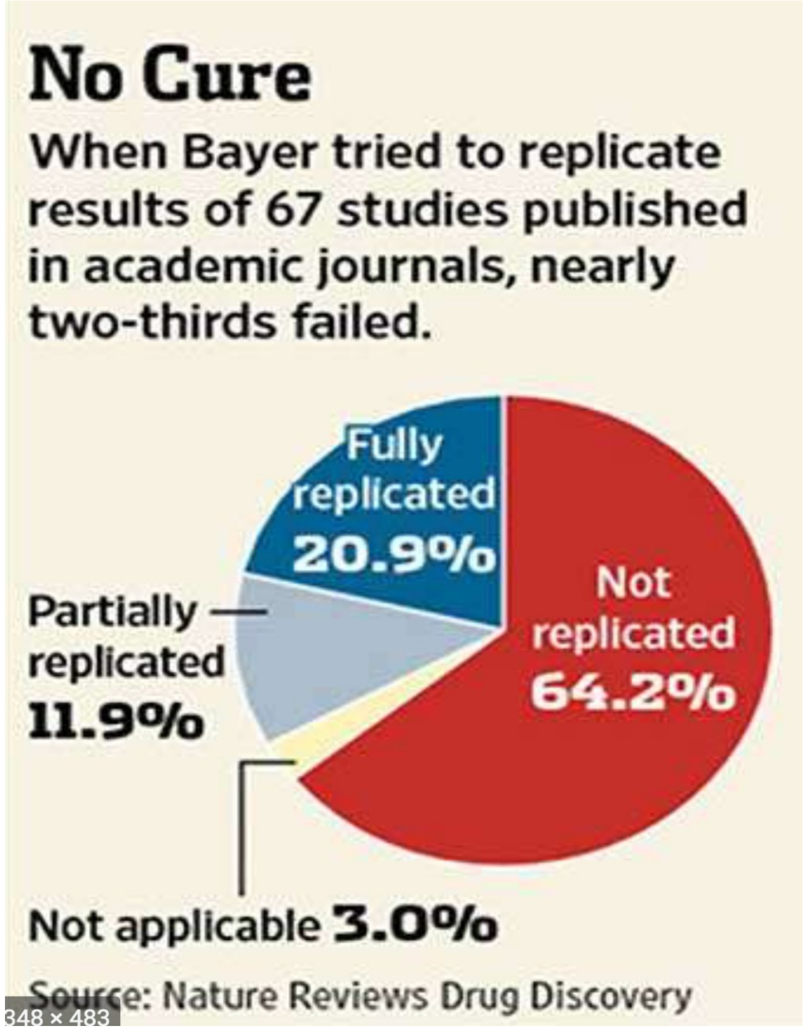
Perspective

Case Reports of Pre-clinical Replication Studies in Metabolism and Diabetes

Matthias von Herrath,[1,*] Philippe P. Pagni,[1] Kevin Grove,[1] Gustaf Christoffersson,[2] Mads Tang-Christensen,[3]
Allan Ertmann Karlsen,[3] and Jacob Sten Petersen[3]

janssen | PHARMACEUTICAL COMPANIES OF Johnson&Johnson

novo nordisk®



No Cure

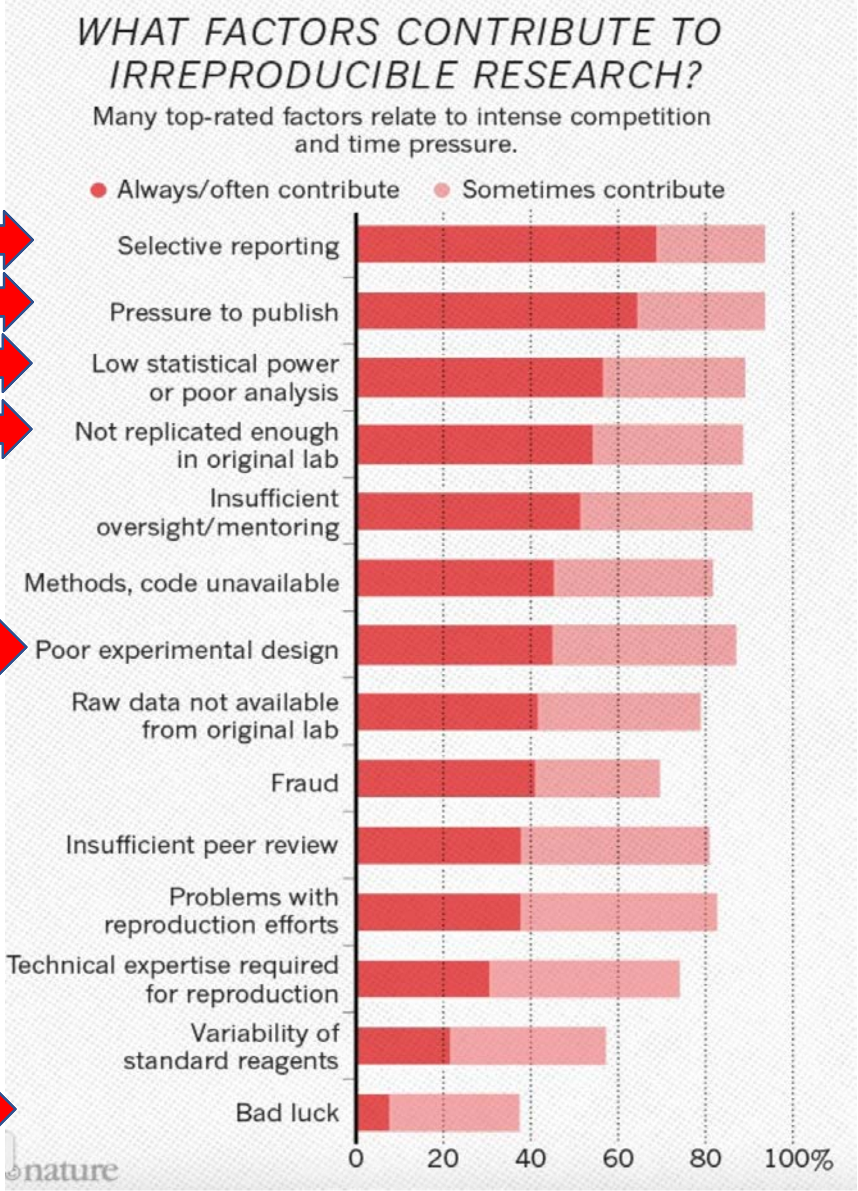
When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Source: Nature Reviews Drug Discovery

Prinz et al., 2011; doi.org/10.1038/nrd3439-c1

Robustness

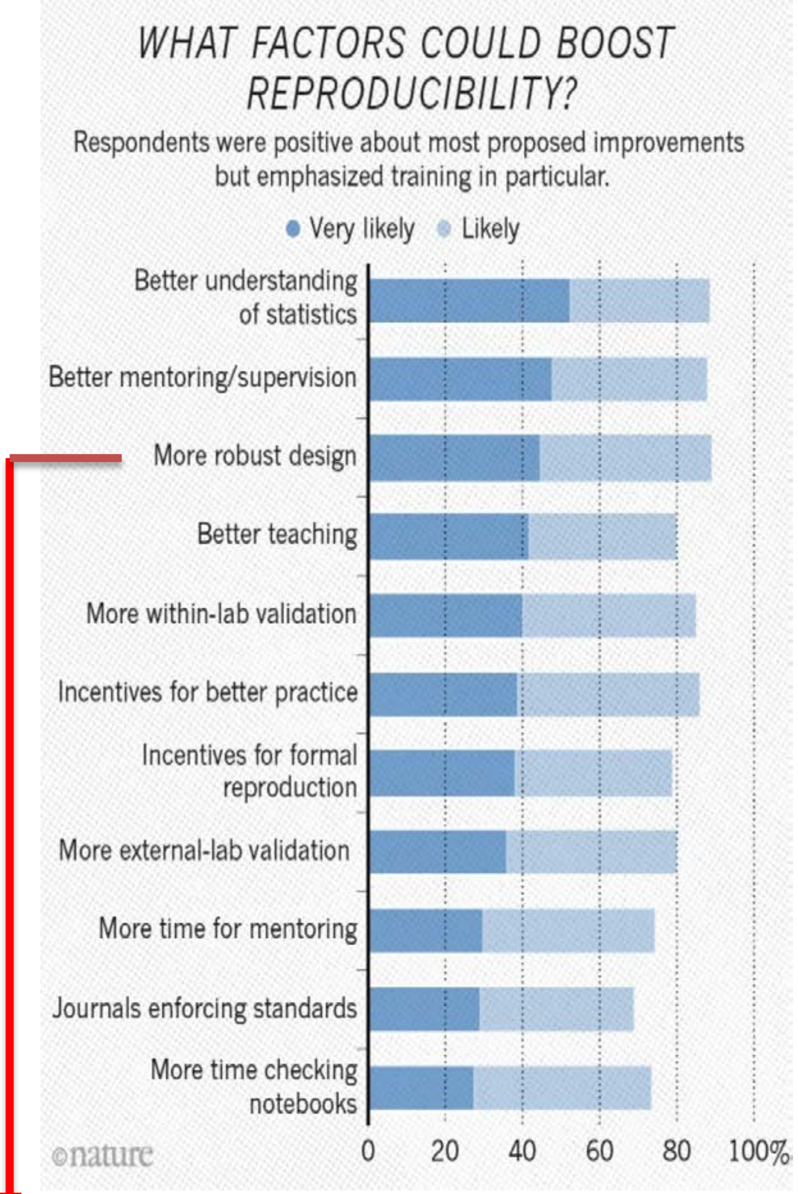
Winner's curse



WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

Baker & Penny, Nature 2016



WHAT FACTORS COULD BOOST REPRODUCIBILITY?

Respondents were positive about most proposed improvements but emphasized training in particular.

Higher statistical power!!

Winner’s curse

- The **exaggeration of effect sizes** (e.g., differences) in published reports and the low probability of study replication

- If 100 labs conduct the same experiment, and only the labs with statistically significant results publish, the reported effect sizes will be exaggerated (and may even be wrong).

- This is especially true when the experiments have **low statistical power** (inadequate sample size given effect size and measurement variability).

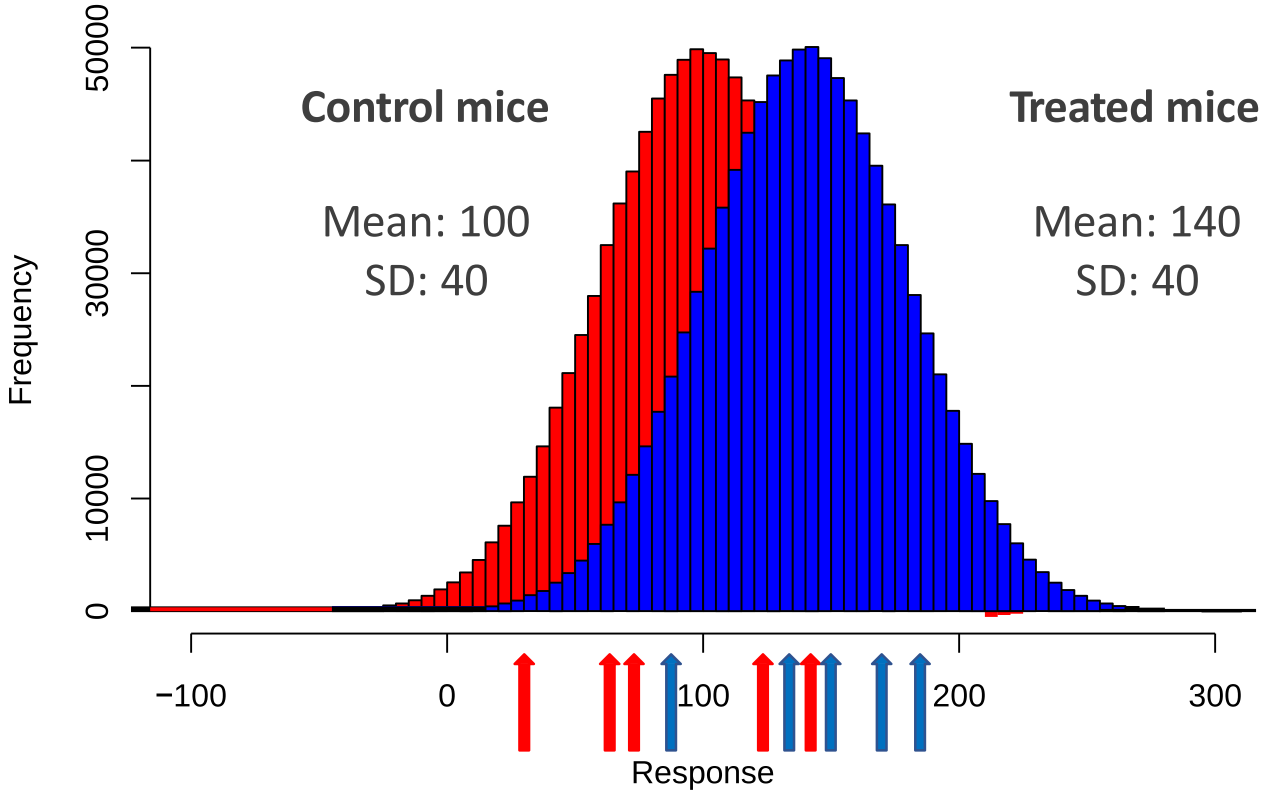
Statistical power

- Definition: ability to reject null hypothesis when it is false ("Ability to detect an effect of a given magnitude or larger, if such an effect is present")

- Power of 80%: out of a 100 experiments, 80 will correctly reject the null hypothesis (20 won't – false negatives)

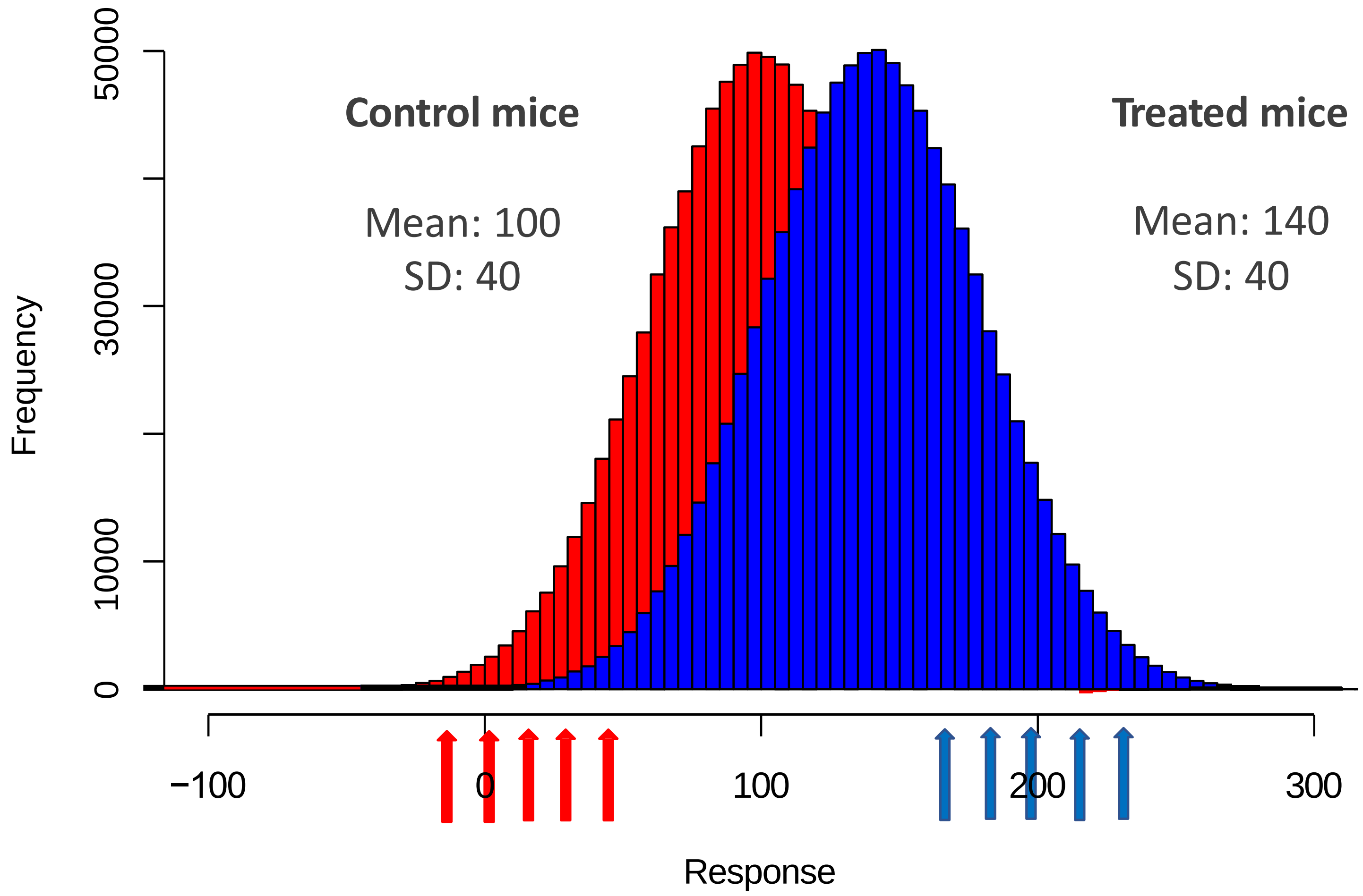
- Statistical power <- sample size + magnitude of effect + measurement variability + alpha (significance) level + statistical test + experimental design

Example: Simulated data allows us to compare results with the ground truth

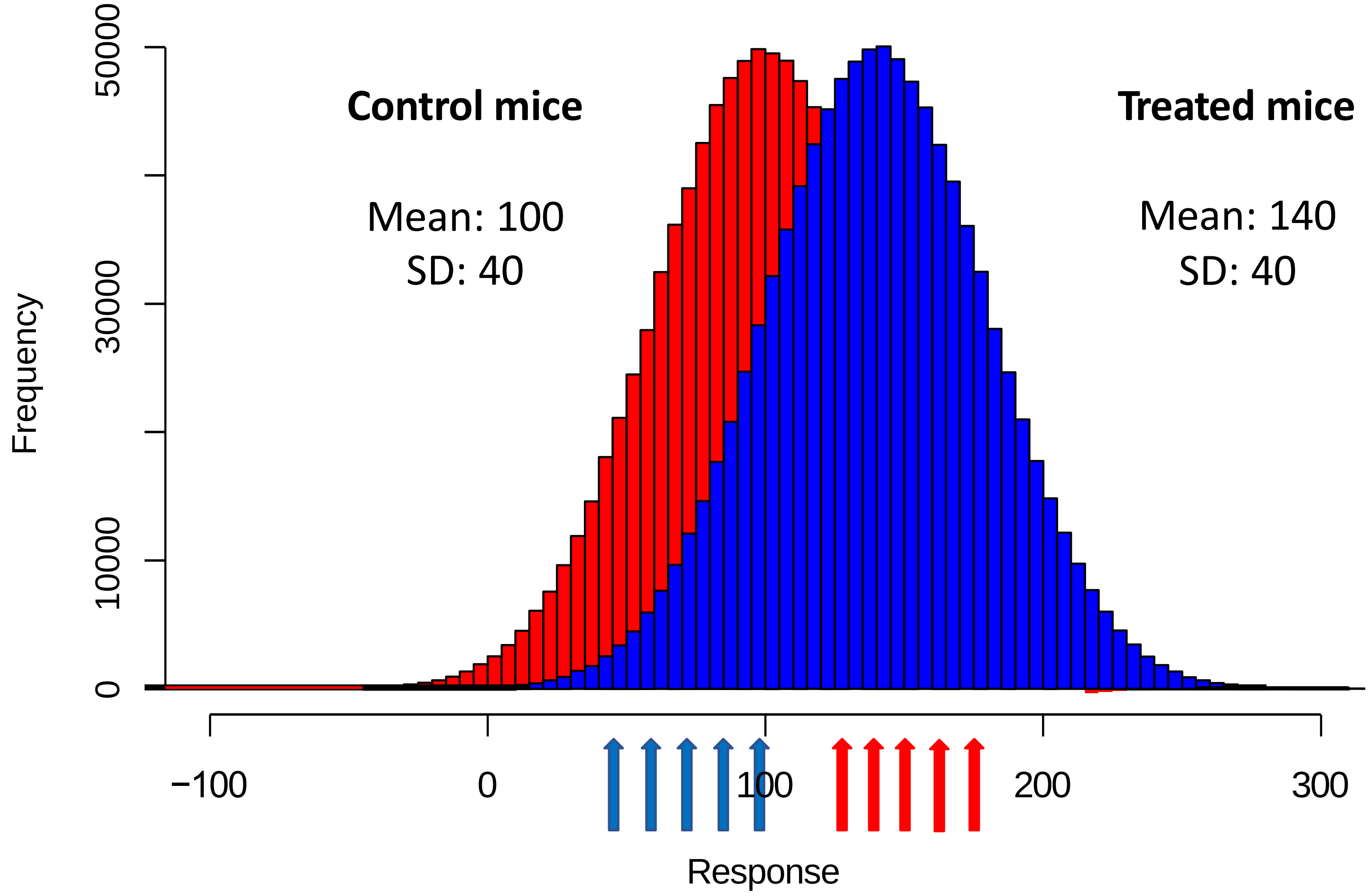


1. **Sample N control mice** (draw N random numbers from the distribution for Vehicle group)
2. **Sample N treated mice** (draw N random numbers from the distribution for Treated group
3. **Determine the effect size**, i.e., difference (95% CI) of means between control and treated mice
4. **Test the null hypothesis** of no difference between groups
5. **Repeat 100 times**

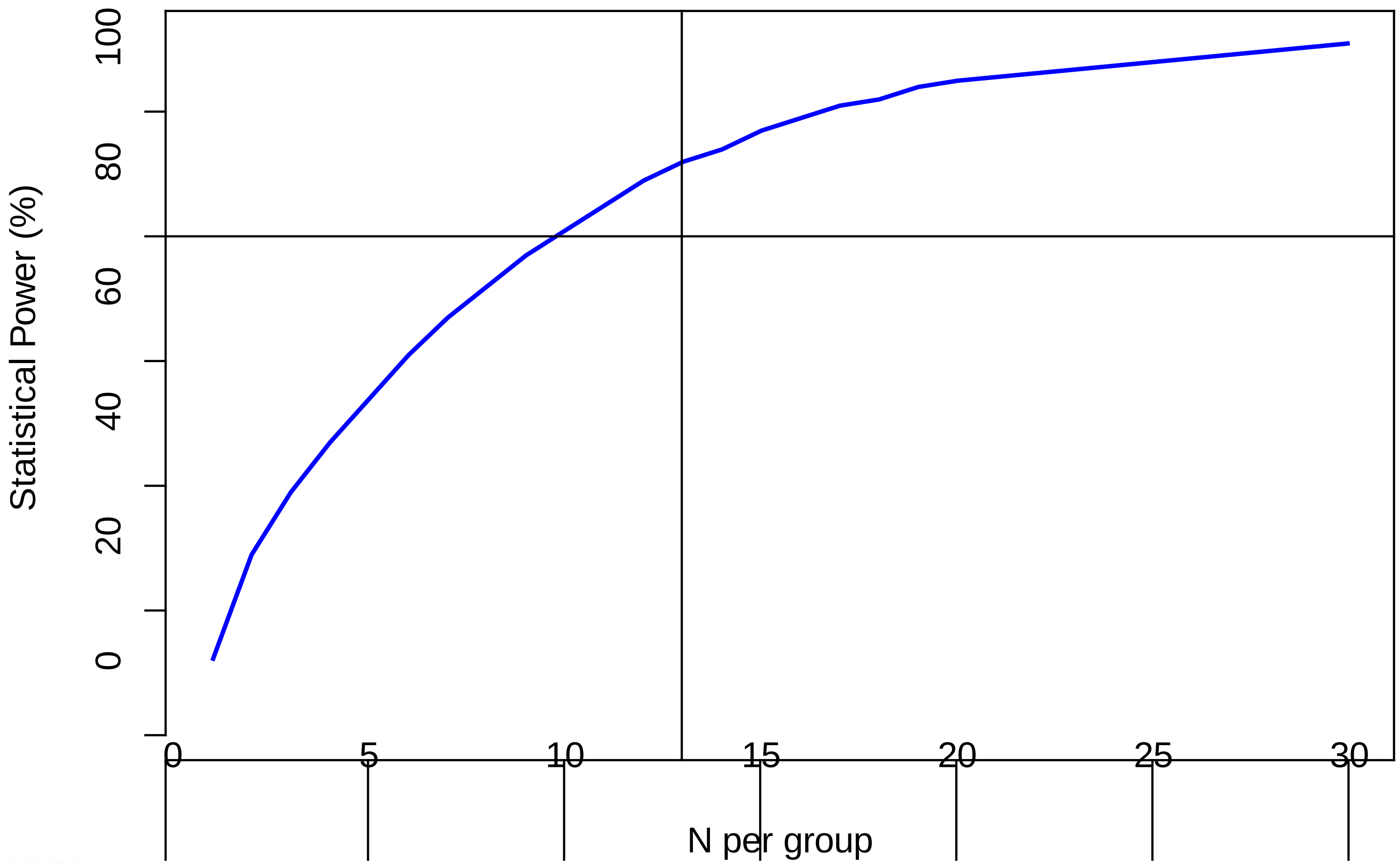
Example sampling – estimates a large effect



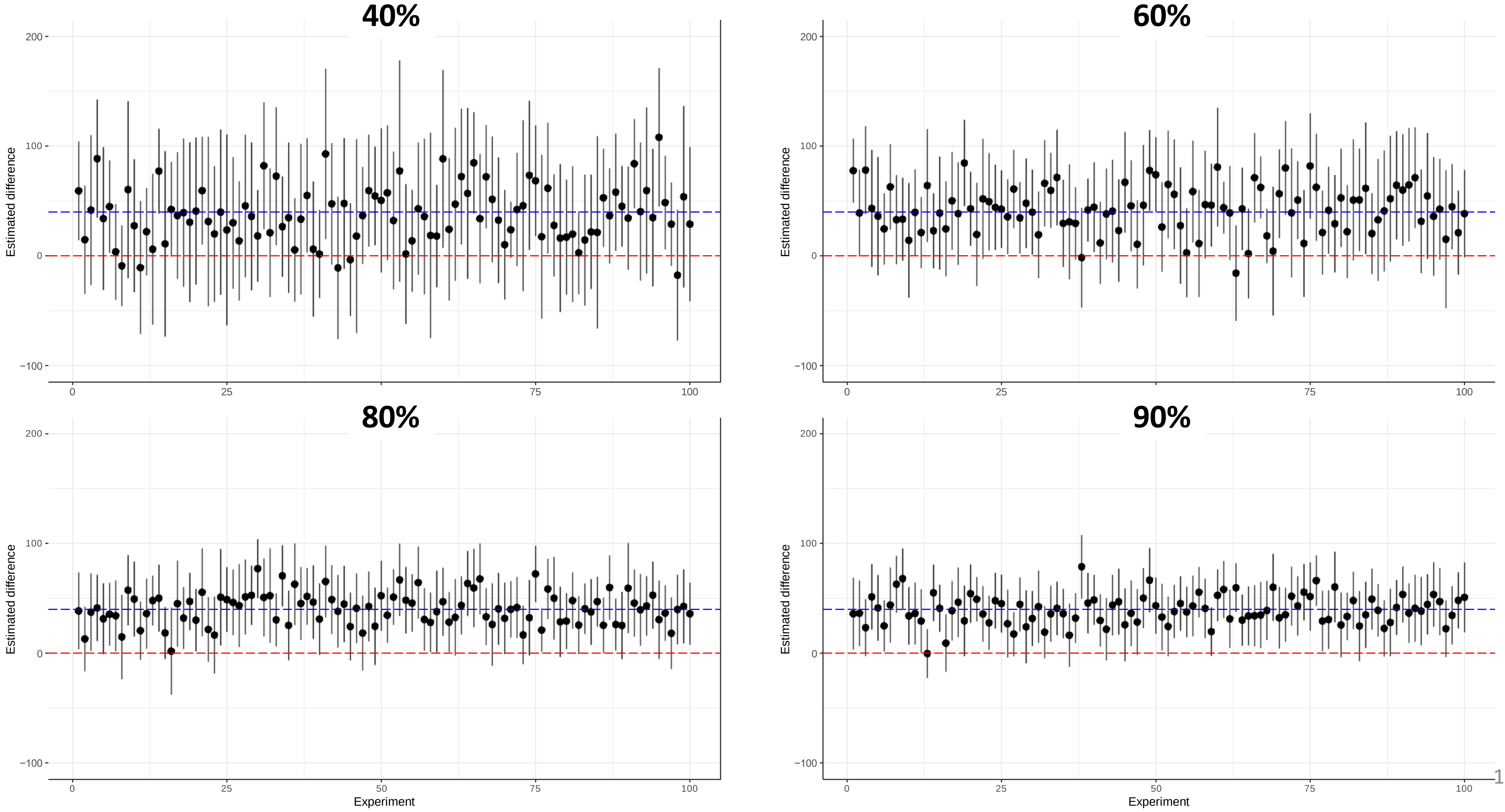
Example sampling – estimates no effect



Choice of N per group determines the statistical power of the hypothesis test



Compare results for different power



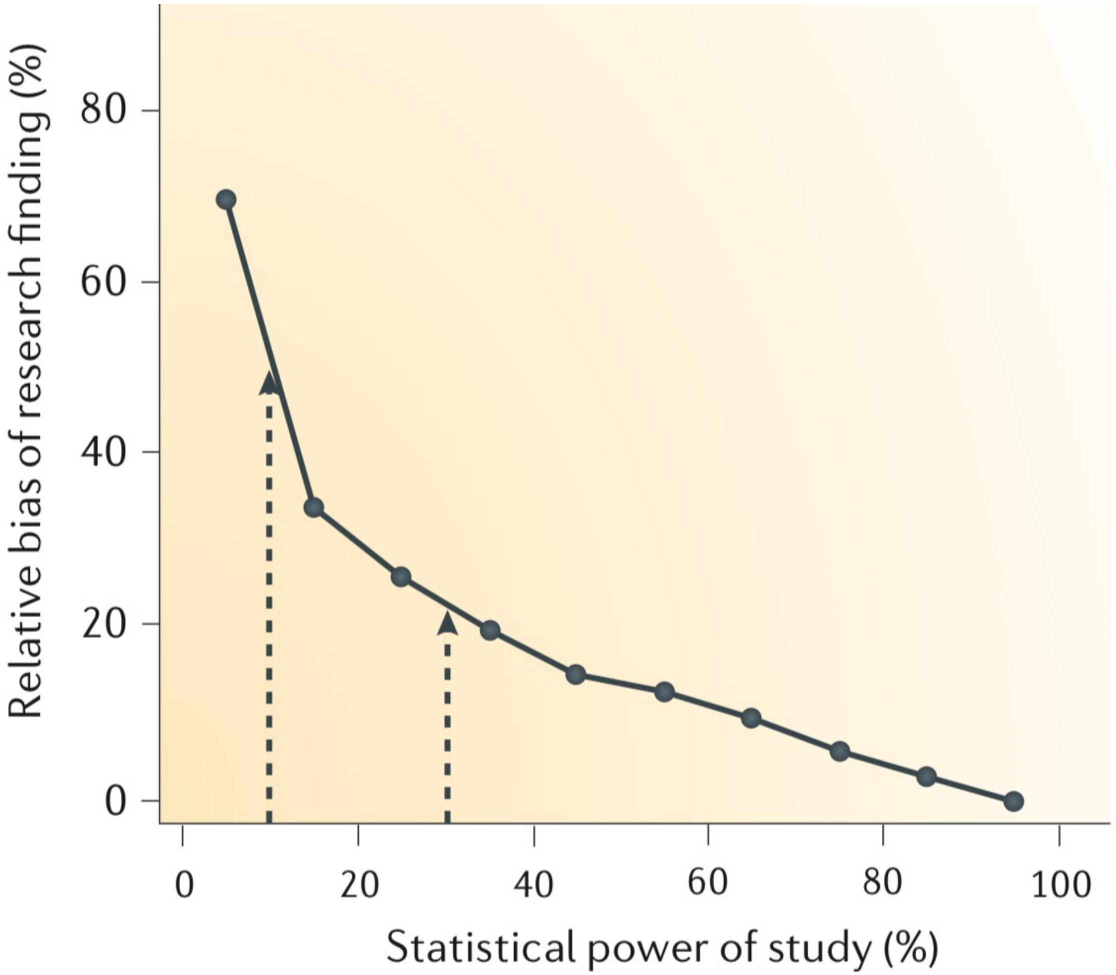
40%

60%

80%

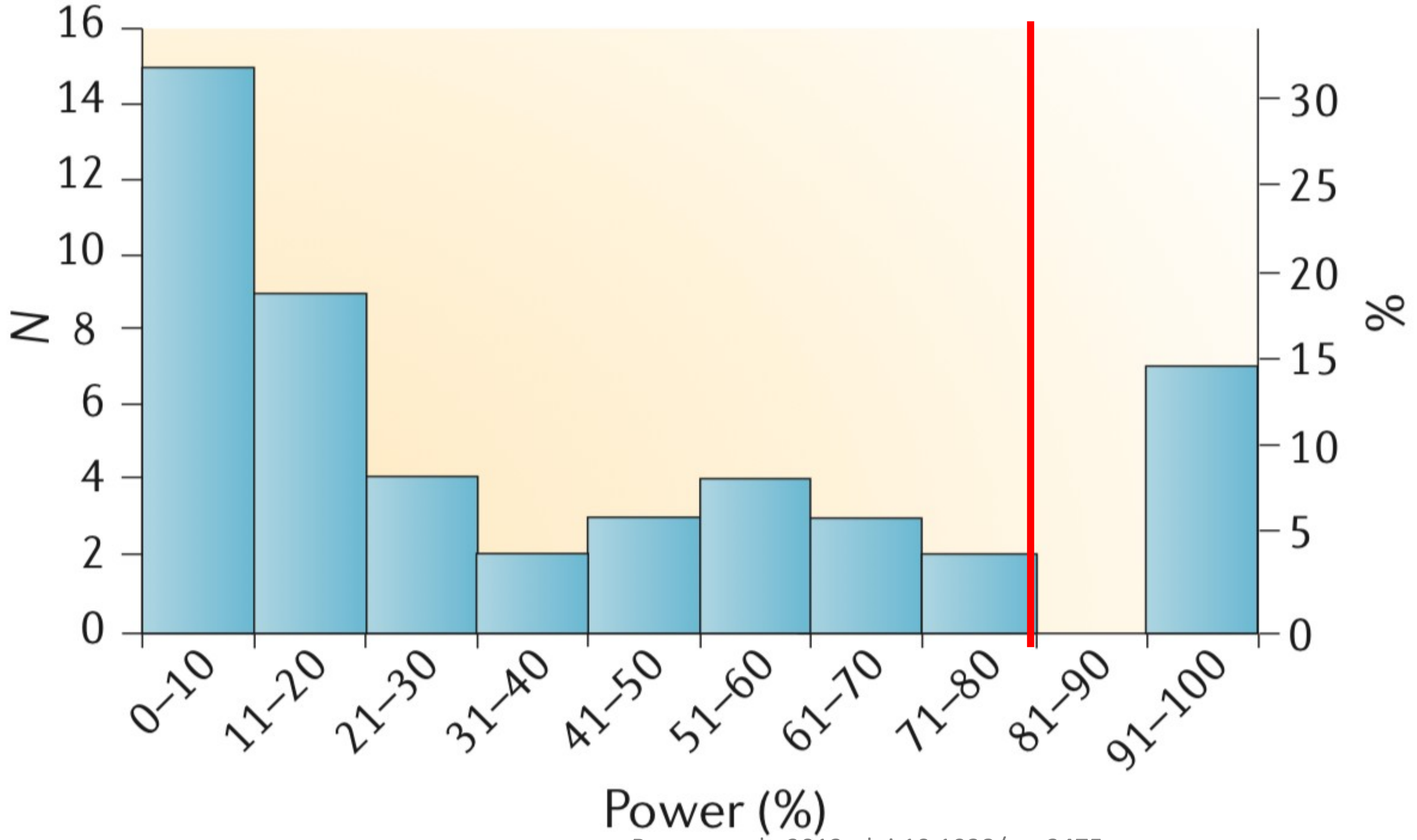
90%

Winner's curse: low power results in exaggerated estimates of effect size



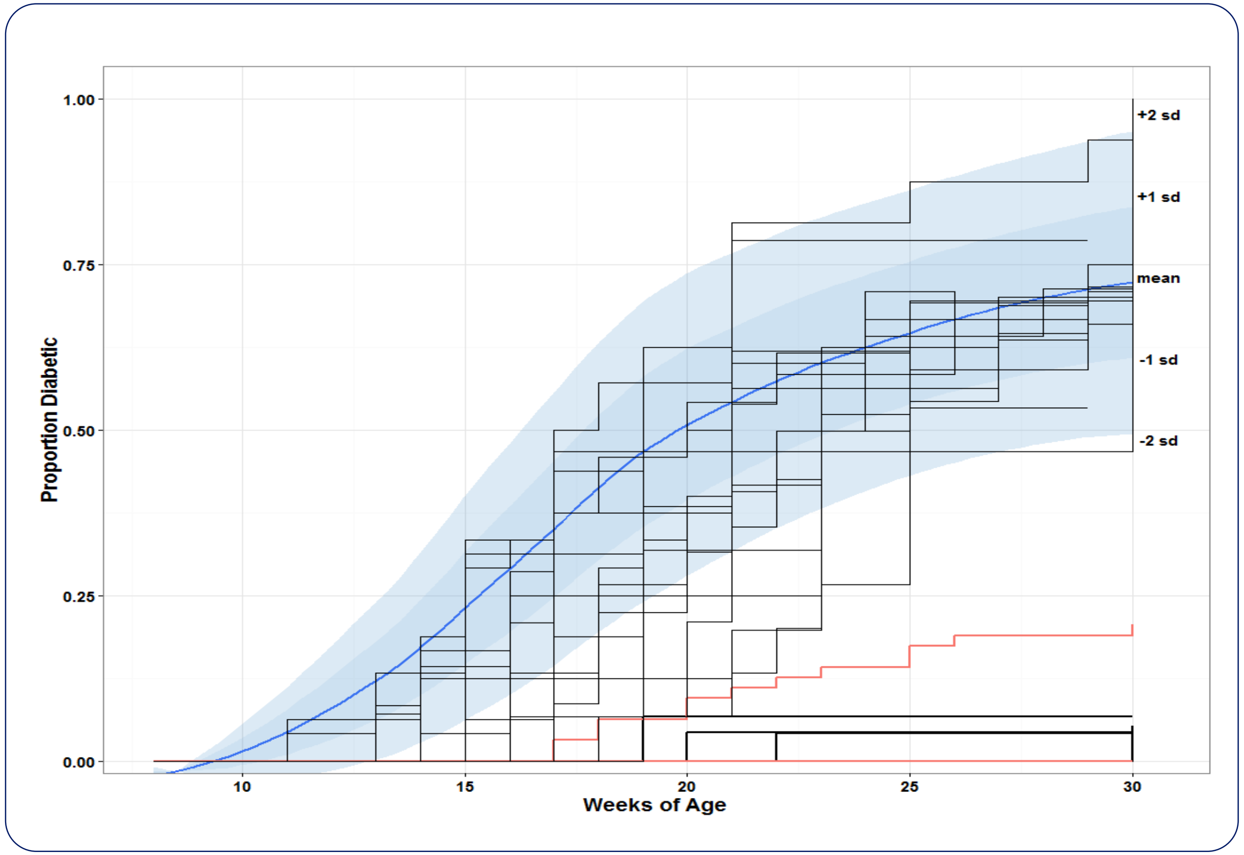
Most published studies have low power

Studies with lower statistical power are at best pilots, but often portrayed (when published) as conclusive. The need to overreach conclusions contributes to distortion of scientific reality



Button et al., 2013; doi:10.1038/nrn3475

Case study – Variability of the NOD mouse model for type 1 diabetes



Many published targets – inherent model ‘noise’

Immunotherapy with translational potential

Mouse numbers are often underestimated, we recommend > 16/group in prevention, >35/group after onset

Studies should be randomized, ideally blinded, and repeated at different sites (cf. Gill et al. Diabetes. 2016 May;65:1310)

A comprehensive matrix of antigens did give none or no robust protection from diabetes in the NOD model (NovoNordisk studies)

NN hormonally inactive insulins #1, 2 and 3 are Novo Nordisk's proprietary insulins with varying degrees of reduced affinity for the insulin receptor

	ORAL TRACK	SUBCUTANEOUS TRACK
Early Prevention	Mouse insulin[a] Porcine insulin [*after Zhang et al. PNAS 1991;88:10252*][a] NN hormonally inactive insulin#1[a]	
Late Prevention	Mouse insulin[a] Porcine insulin[a] NN hormonally inactive insulin#1[a] NN hormonally inactive insulin#2[b] NN hormonally inactive insulin#3	NN hormonally inactive insulin#1 in protamine-acetate NN hormonally inactive insulin#1 in protamine-sulphate NN hormonally inactive insulin#1 in IFA NN hormonally inactive insulin#1 in Intralipid NN hormonally inactive insulin#2 [*after Karounos et al. J. Clin. Invest. 1997; 100:1344*][b] NN hormonally inactive insulin#2 in protamine-sulphate Human proinsulin peptide Insulin mimotope 3 [*after Daniel et al. J Exp Med 2011;208:1501*]
Recent Onset	Mouse insulin + liraglutide Porcine insulin + liraglutide NN hormonally inactive insulin#1 + liraglutide	NN hormonally inactive insulin#1 in protamine-sulphate + liraglutide

Publications from in-house work:

[a]Pham et al *Clin Immunol* 2016;164:28

[b]Grönholm et al. *Diabetologia* 2017;60:1475

Sobering conclusions – antigenic therapy NOD

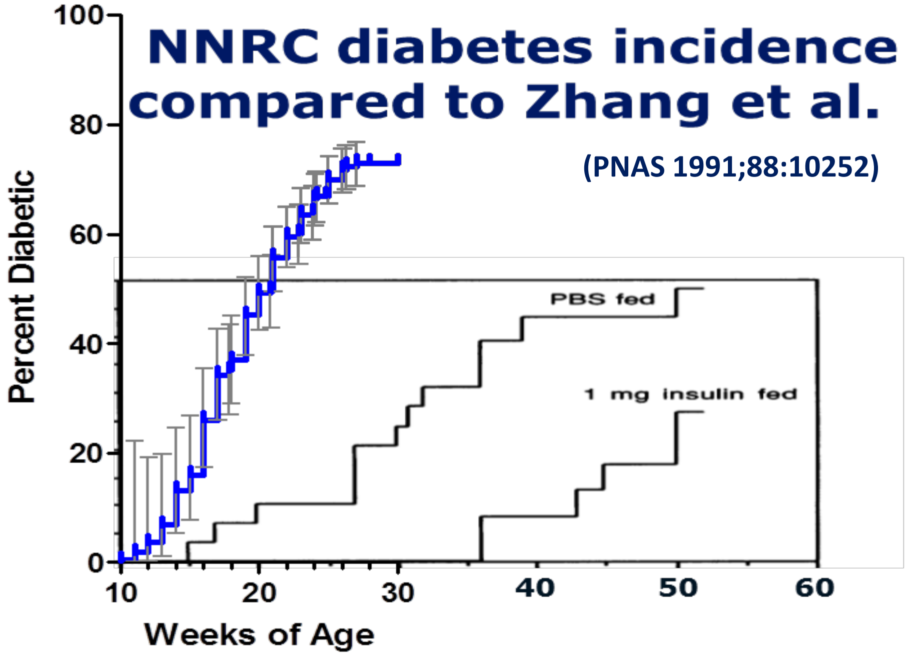
Reproducibly worked:

- InsB9:23 in IFA
- In house DNA immuno-therapy (proinsulin)

Lack of robustness:

- Oral insulins in various formulations
- All peripheral peptides in adjuvants or with acylation to prolong half-lives or via pumps

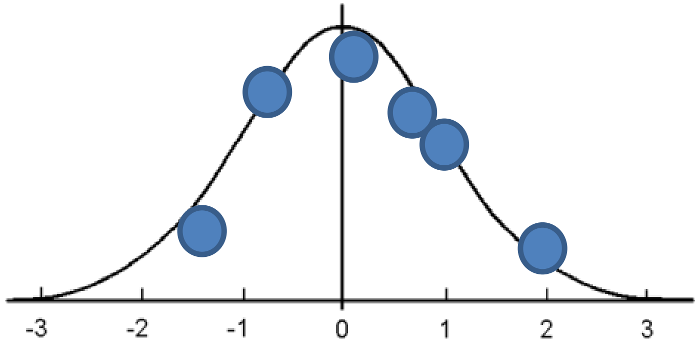
Variation of T1D incidence in the NOD model:



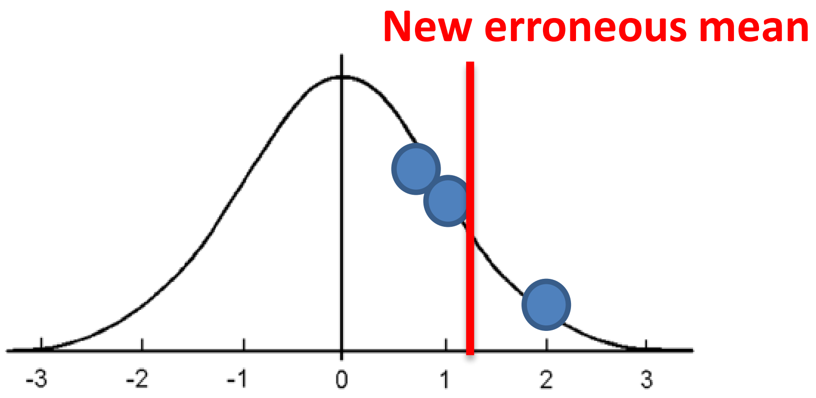
Effect of oral administration of porcine insulin on T1D in female NOD mice. Life table analysis of the control group and the group fed 1 mg of insulin (P = 0.02, Log rank test). Porcine insulin was administered twice for 5 weeks then once weekly thereafter until one year of age with treatment beginning at 5 weeks of age (n=27-30).

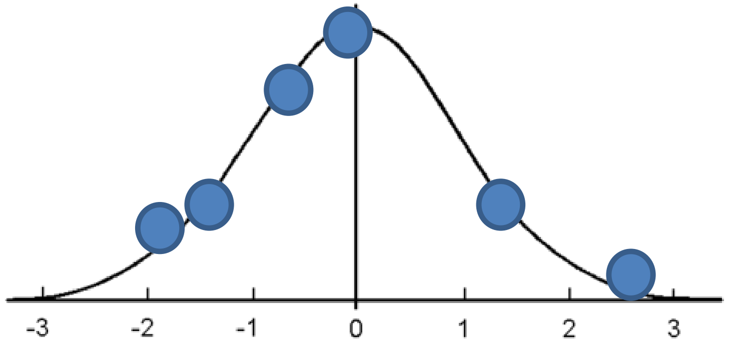
Displayed in blue is the combined incidence in untreated and PBS controls from NNRC-Seattle demonstrating the difference in rate of disease onset and incidence (n=176).

Particularly problematic scenarios assuming positivity bias in publications

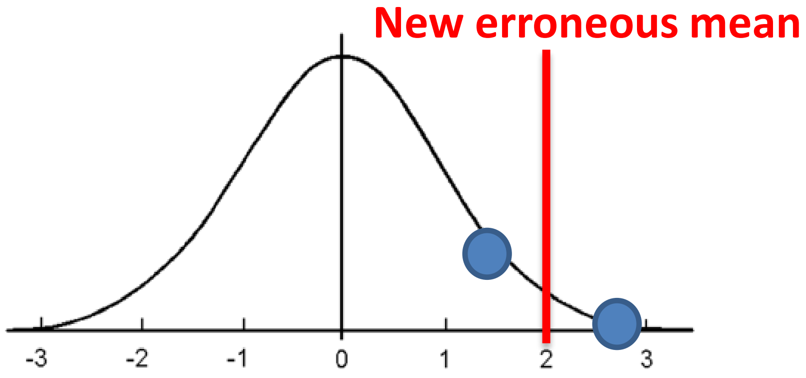


Low Number of experiments or replications per study





High SE



Reality

Reality

Conclusions – changes we should embrace to make the scientific method robust again

More expensive

- Requiring ≥80% power results in more reliable results and replicable experiments; blinded studies, determine minimal detectable difference and biological relevance in advance
- Use pure reagents and optimal technology, share resources – collaborate for this, science has evolved and become too complex to yield meaningful results in single laboratories only

Change in mindset

- Eliminate positivity bias – negative results need to be published and such studies/papers need to be career relevant
- Embrace a more collaborative scientific model, this will become more relevant as science and underlying technology become increasingly complex as well as for human research
- The current system is in 'over-drive', publish fewer, but better studies