MRC Hubs for Trials Methodology Research

ConDuCT Hub

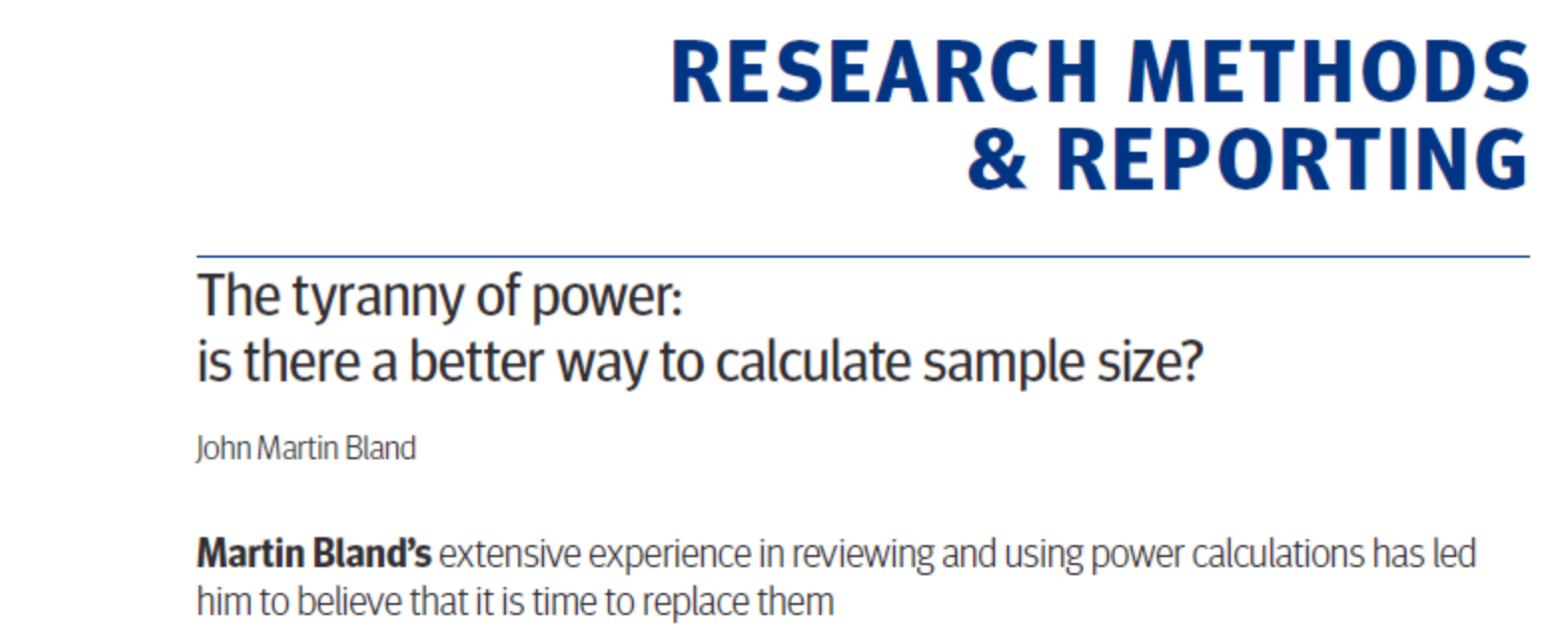
Holding on to power:

why confidence intervals are not (usually) the best basis for sample size calculations

Chris Metcalfe

University of Bristol &

MRC ConDuCT Hub for Trials Methodology


RESEARCH METHODS
& REPORTING

The tyranny of power:
is there a better way to calculate sample size?

John Martin Bland

Martin Bland's extensive experience in reviewing and using power calculations has led him to believe that it is time to replace them

- Power / sample size calculations have been useful in increasing the typical size of studies
- Calculate the sample size needed to give a good chance (= power) that a true treatment effect is demonstrated in a study

Problems of power?

- Need knowledge of the research area (at least some idea of variability / control group response)
- Difficulties in deciding “How big a difference do you want to detect?” Researchers often state the difference they can detect with their expected sample size
- Need to choose a primary outcome measure

- *“If we ask researchers to present their results as confidence intervals rather than significance tests, I think we should also ask them to base sample size calculations on confidence intervals.”*

- "The width of the confidence interval for the difference between two similar percentages is given by:
- **$\pm 1.96\sqrt{(2p(100-p)/n)}$**
- where *n* is the number in each group and *p* is the percentage expected to experience the event."
- With *d* as the confidence interval width and re-arranging …

- $n = (1 / d)^2 \times 2 \times p \times (100 - p) \times 1.96^2$

- Compare this with the sample size formula for the long established use of confidence interval based analysis in **equivalence studies**:

- $n = (1 / d)^2 \times 2 \times p \times (100 - p) \times (1.96 + 0.84)^2$

- Two treatments are in truth equivalent, with p = 15% of patients in both arms developing the disease under study
- Sample size for 95% confidence interval with width $\pm d$ excluding more than a 5% absolute difference between the treatments
  - Bland's formula: 392 patients per arm.
    - 50:50 chance of demonstrating equivalence
  - Standard formula: 800 patients per arm.
    - 80% chance of demonstrating equivalence

Problems of power?

- Need knowledge of the research area (at least some idea of variability / control group response)
- Difficulties in deciding “How big a difference do you want to detect?” Researchers often state the difference they can detect with their expected sample size
- Need to choose a primary outcome measure

How big a difference?

- for **equivalence studies**, a true treatment difference of zero is assumed and $\pm d$ is the maximum difference which would still allow the two treatments to be considered as clinically equivalent

Detecting a difference in effect

- At first a 95% confidence interval that excludes unimportant differences appears desirable
- But this introduces the difficulty of assuming a true difference between the treatments being compared ...
- ... with *d* then the true difference minus the minimum important difference

Conclusions

- Power is integral to study planning, giving a good chance that a true treatment effect will be detected even if underestimated
- Sample size calculations based on confidence-intervals:
  - still need to accommodate power
  - still need to be based on a particular outcome
  - still need the minimum important difference to be defined, **and also the true difference**