Epib 605

A few random thoughts

Jay Brophy MD PhD
McGill University — Departments of Medicine, Epidemiology & Biostatistics

2026-02-17

Week 6

- Breaking Up Prolonged Sitting to Improve Cardiometabolic Risk: Dose–Response Analysis of a Randomized Crossover Trial
- Conclusion: The present study provides important information concerning efficacious sedentary break doses. Higher-frequency and longer-duration breaks (every 30 min for 5 min) should be considered when targeting glycemic responses, whereas lower doses may be sufficient for BP lowering.

What is GPT-5's opinion?

According to M365 Copilot, powered by a GPT-5-based chat model

> "This study makes a meaningful contribution to the sedentary-behavior and cardiometabolic-risk literature by empirically comparing multiple frequency × duration combinations of sedentary breaks within a single crossover design. Despite some limitations, the study effectively advances the field."

Major strength:

- Randomized crossover design reduces inter-individual variability and allows within-subject comparisons, enhancing statistical power and precision in effect estimation

Minor limitations:

- Small sample size and underpowered design
- Short-term, acute effects only; no data on long-term outcomes or sustainability
- Highly controlled laboratory conditions - not reflective real-world

Efficacy of crossover design

Parallel vs. Crossover Sample-Size Efficiency

Total parallel N required to match precision of a crossover, using the classical ratio 2 / (1 – ρ)

Within-subject correlation (ρ)	Parallel : Crossover sample-size ratio = 2 / (1 - ρ)	Parallel total N if crossover uses n = 17
0.5	4.00	68
0.6	5.00	85
0.7	6.67	113
0.8	10.00	170

The more similar a person's responses are across the two periods (higher correlation), the more a crossover "leverages" that similarity to reduce noise —
so a parallel trial must grow in size to keep up with the crossover's rising efficiency

Figure 1

> "...30 min for 5-min dose (mean, 11.8; SE, 4.7; P = 0.017). Attenuations were also observed for the other sedentary break doses, most notably for the every 30 min for 1-min dose (mean, 6.7; SE, 4.6; P = 0.159), but were not statistically"significant"

But aren't we more interested in the 30/5 vs 30/1 comparison? Are those differences significantly significant?

```r
# test for interaction
m1 <- 11.8; m2 <- 6.7; se1 <- 4.7; se2 <- 4.6
# Risk difference: non-immigrant minus immigrant
rd  <- m1 - m2; se_rd <- sqrt(se1^2 + se2^2)   # as
ci_rd <- rd + c(-1,1) * 1.96 * se_rd; z <- rd / se_
p <- 2 * pnorm(-abs(z))
```

```
$RD
[1] 5.1

$CI95
[1] -7.789888 17.989888

$z
[1] 0.7754916

$p_value
[1] 0.4380492
```

Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician. 2006

McGill

Am improved visualization

A more serious problem

- What was analyzed? Mean change from baseline i.e Time 1 vs time 2
- What was randomized? Intervention 1 vs intervention 2?
- Due to the randomization scheme, comparisons must be between treatment strategies with baseline values as covariates — not across time periods via change scores
- Represents major methodological error as regression to the mean (RTM) can lead to a bias overestimation
- Correct approach - mixed-effects ANCOVA with fixed effects for treatment, and period-specific baseline as a covariate

References

1. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. BMJ. Nov 10 2001;323(7321):1123-4.
2. Senn S. Change from baseline and analysis of covariance revisited. Stat Med. Dec 30 2006;25(24):4334-44. doi:10.1002/sim.2682

It doesn't finish there!

In small-N designs with modest signal-to-noise, "statistically significant" effects tend to be

(i) more likely than you expect to have the wrong sign (Type S error), and

(ii) when the sign is right, the magnitude is often exaggerated relative to the true effect (Type M error, the "exaggeration ratio").

(iii) The lower the true power for a realistic effect, the higher the expected exaggeration among the subset of significant results.

Reference: Gelman A, Carlin JB. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci. Nov 2014;9(6):641-651.

ChatGPT a second opinion

> Given the baseline/RTM flaw and related issues that you raise, I withdraw my earlier positive characterization. The manuscript addresses a worthwhile question with a strong design, but as currently analyzed it does not provide reliable evidence.

A few slides

Speaker notes