SeeingTheta

ABSTRACT 5697

ISOQOL 2026 · PRAGUE · ORAL SESSION 109: Analytical Methods

rmcorr vs MLCorr

A head-to-head contest for the within-person correlation

Performance of repeated-measures (rmcorr) versus multilevel (MLCorr) correlations in longitudinal psychometric data: a Monte Carlo simulation study

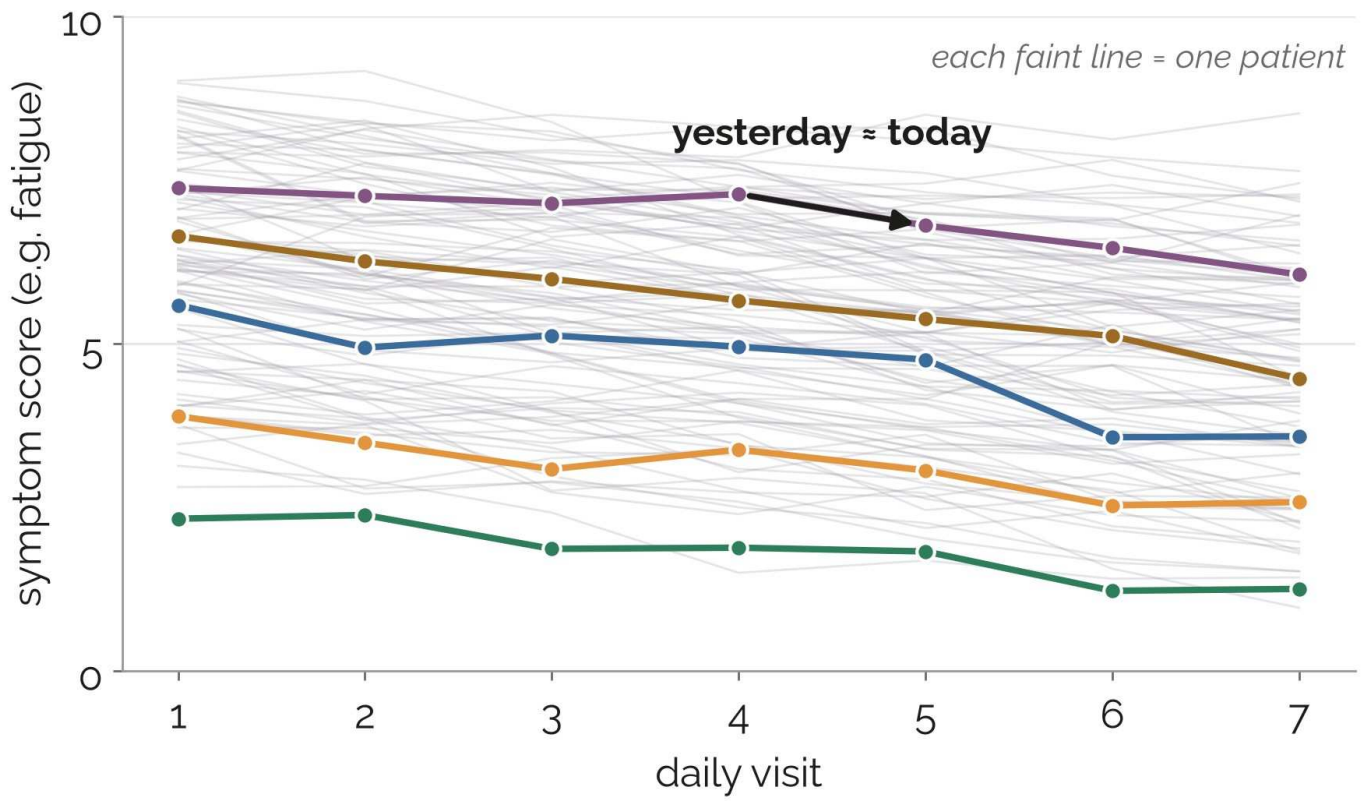
PRESENTED BY Marie-Jeanne Holveck, SeeingTheta

WITH Piper Fromy, SeeingTheta

DATE Monday 19 October 2026, 4:05 PM



The arena: what repeated PRO data look like



about 7
daily visits per patient

100 to 1,000
patients per study

autocorrelation
today's score echoes yesterday's

Illustrative simulated data: 100 patients, 7 daily visits each, Colour = 5 highlighted patients.

Each patient gives several scores, and neighbouring visits are alike.

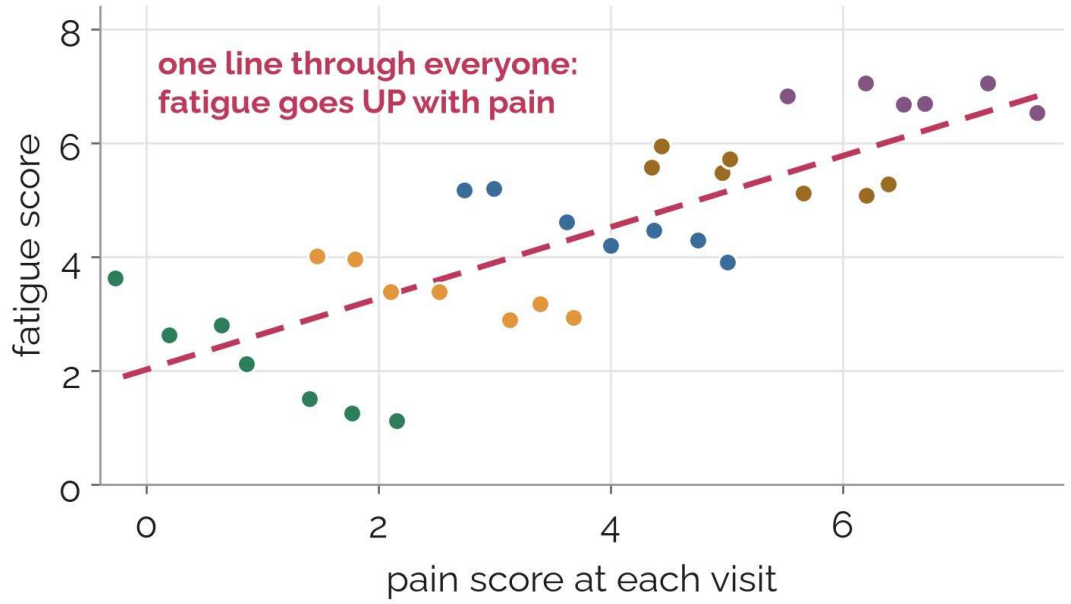
Why the within-person correlation matters

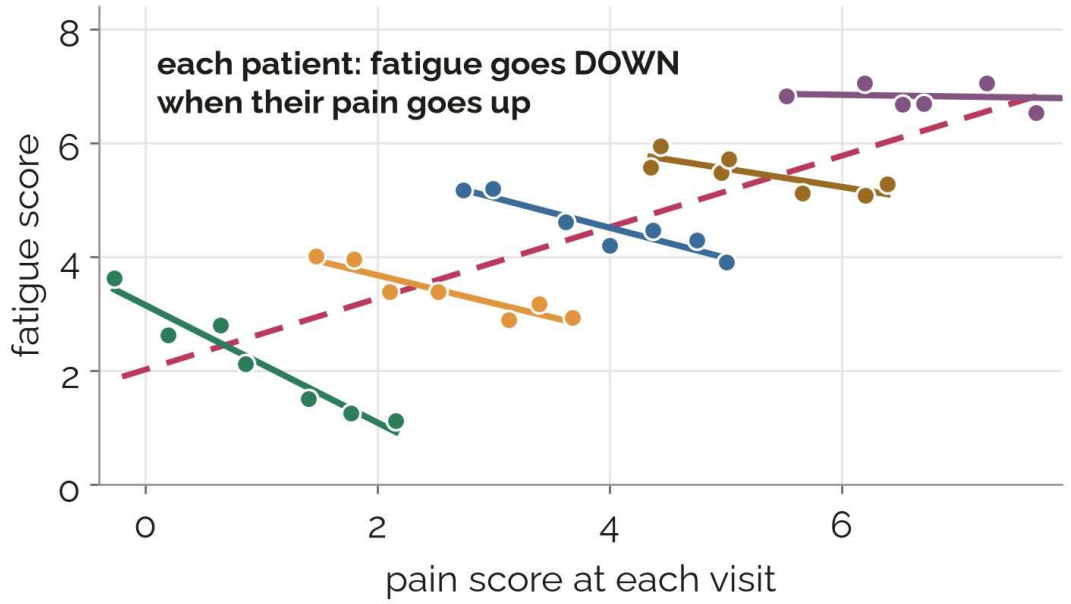
THE BETWEEN-PERSON QUESTION

Do people with more pain also have more fatigue? This compares one person with another.

THE WITHIN-PERSON QUESTION

When my pain is above my own usual level, is my fatigue above my usual level too?

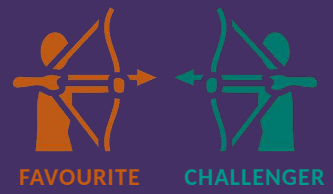




Illustrative simulated data: 5 patients, 7 visits each (35 measurements). Colour = patient.

Ignore the people and you get the wrong answer. Two methods claim the right one.

Meet the contenders: the tale of the tape

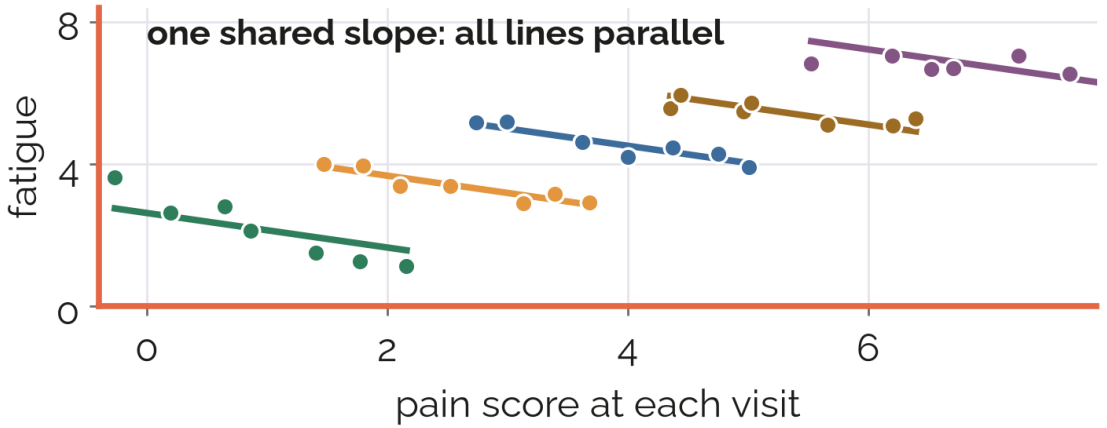


BEFORE THE MATCH · THE CONTENDERS

rmcorr vs MLCorr

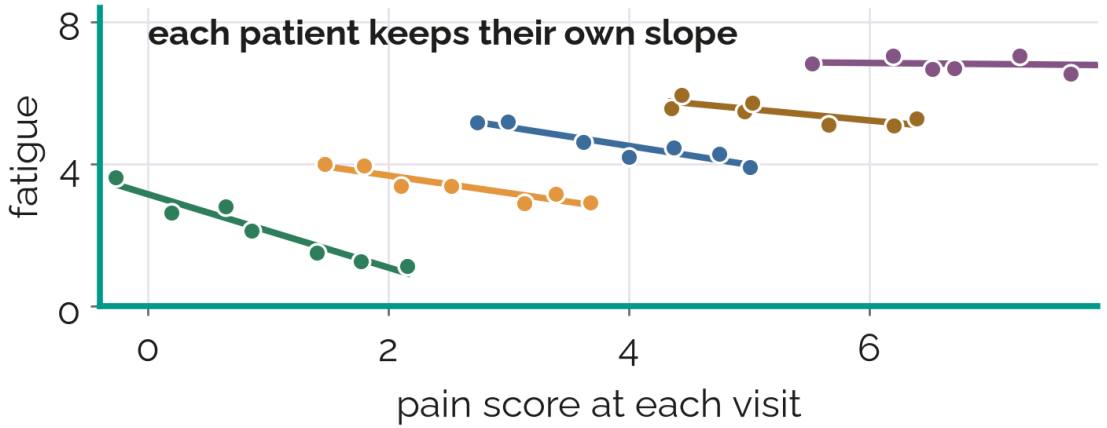
THE ORANGE CORNER · rmcorr repeated-measures correlation

"If a typical patient's pain goes up, what happens to their fatigue? Assumes every patient responds the same way."



THE GREEN CORNER · MLCorr multilevel correlation

"The same question, and by the way: how much do patients differ in how they respond?"



BOTH compare each patient with their own usual level, then estimate the average within-person slope

SLOPES one shared slope for everyone

AUTOCORRELATION ignored: visits treated as independent

SLOPES average slope, plus how much patients differ

AUTOCORRELATION modelled: AR(1) or CAR(1) structure

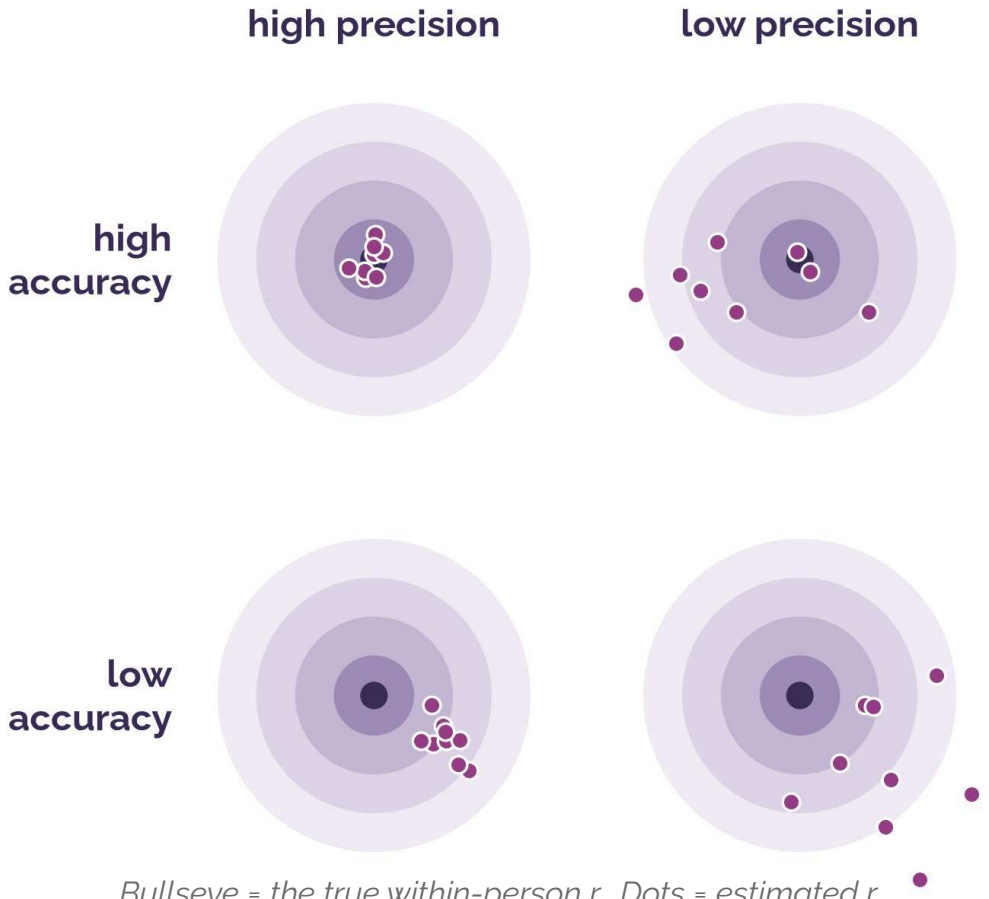
Same target for both: the average within-person r. The key difference is autocorrelation.

How the judges score: accuracy and precision

FAVOURITE CHALLENGER

BEFORE THE MATCH · THE ROUNDS

rmcorr vs MLCorr



Bullseye = the true within-person r. Dots = estimated r.

ROUND 1. COVERAGE · THE MAIN EVENT

Does the 95% CI contain the true r 95% of the time? This needs accuracy and precision together.

ROUND 2. BIAS · ACCURACY

Is the estimated r centred on the true r, on average?

ROUND 3. CI WIDTH · PRECISION

How tight is the 95% Confidence Interval (CI)?

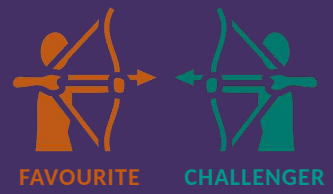
OUR PREDICTIONS

rmcorr is less accurate (misses more often) because ignoring autocorrelation inflates its estimated r.

The gap attenuates when patients respond alike, with same slope.

A good method hits the true r, and knows how sure it is.

The obstacle course: conditions from the literature



Obstacle	Levels	Grounded in
Patients respond differently (slope heterogeneity)	SD = **0**, **0.25** or **0.75** times the average slope	McNeish & Hamaker 2020; Fisher 2018; Bolger & Laurenceau 2013
Visits per person (T)	**7** or **14**	Ringwald 2022 (about 7 days); 14 brackets it
Missing data	**complete**, or **about 20% Missing At Random (MAR)** where a patient is more likely to skip after a high score or later in the study	Vachon 2019; Rintala 2019; Diggle & Kenward 1994; Rubin 1976
Residual autocorrelation (ϕ = 0.3)	first-order AutoRegressive **AR(1), regular visits**, or first-order Continuous-time AutoRegressive **CAR(1), irregular visits**	Kuppens 2010; de Haan-Rietdijk 2017
Sample size (N)	**100**, **300** or **1,000**	Ringwald 2022 (300, 1,000); N = 100 as a stress test
True within-person r	**0.1**, **0.2** or **0.3** (and **0, to count false positives**)	Gignac & Szodorai 2016

Held fixed: between-person r = 0.3. The outcome is continuous and normal; ordinal scores are future work.

240 conditions × 1,000 simulated datasets × 2 methods = 480,000 fits

Round 1, coverage: rmcorr loses ground as studies grow

AFTER ROUND 1 · COVERAGE

rmcorr 0 : 1 MLCorr

Coverage (accuracy and precision) = how often does the 95% CI around r contain the true r?

MLCorr holds close to 95% throughout. **rmcorr** gets worse with larger N, larger true r, 7 rather than 14 visits, and irregular visits (CAR(1), in backup).



AR(1) residuals, MAR about 20% missing, averaged over slope heterogeneity. Grey band = 95% ± simulation noise (1.96 Monte Carlo SE). All conditions in backup.

The surprise: the bigger the study, the worse rmcorr does.

Rounds 2 and 3, the replay: why rmcorr gets worse with more people

AFTER ROUNDS 2 and 3 · BIAS · 1 DRAW

rmcorr 0 : 2 MLCorr

COVERAGE · ACCURACY AND PRECISION

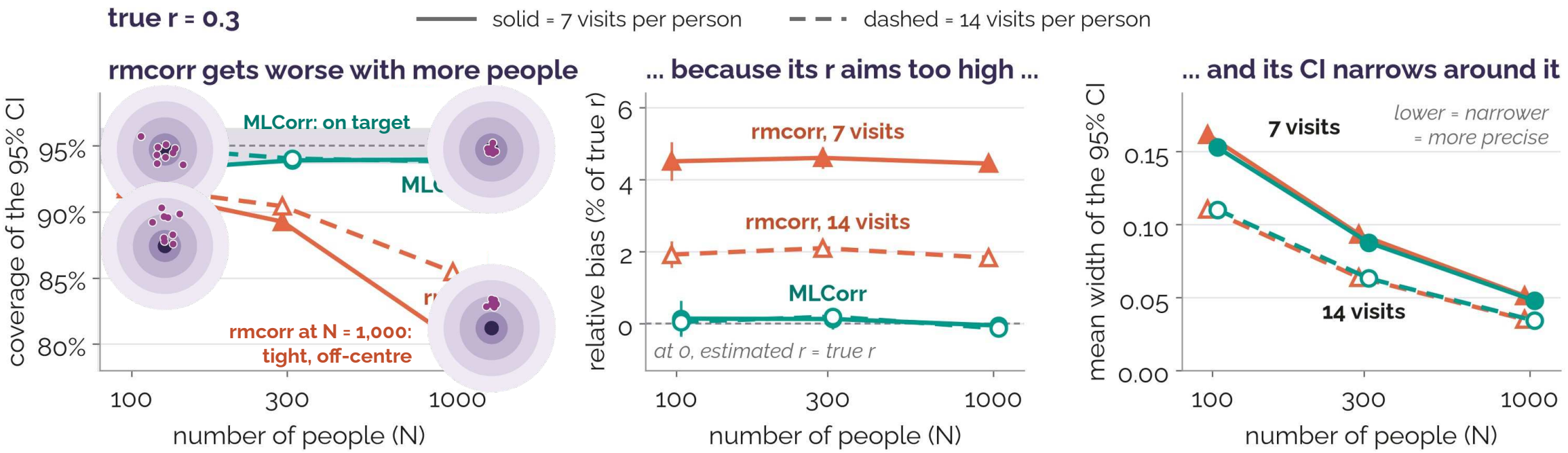
How often does the CI around r contain the true r?

BIAS · ACCURACY

On average, how far is the estimated r from the true r?

CI WIDTH · PRECISION

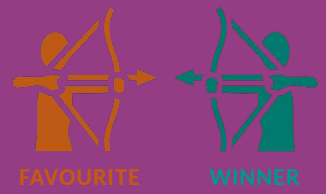
How wide is the 95% CI around the estimated r?



True within-person r = 0.3. AR(1) residuals, MAR about 20% missing, averaged over slope heterogeneity. Grey band = 0.95 ± 1.96 MCSE. Bars = ± 1.96 MCSE.

More data makes a biased method more confidently wrong. Doubling the visits halves the bias.

The verdict: MLCorr wins on points



FINAL SCORE · 2 DRAWS

rmcorr 0 : 3 MLCorr

Round	What the judges ask	Winner
1. Coverage	Does the 95% CI contain the true r?	**MLCorr**
2. Bias	Is the estimate centred on the true r?	**MLCorr**
3. CI width	How tight is the interval?	**draw**
4. False positives*	At true r = 0, is the error rate 5%? (yes)	**draw**
5. Power*	Does it find a real effect? (yes)	**MLCorr**

** Backup rounds. Power: MLCorr equal or slightly higher at N = 100 and r = 0.1.*

In typical PRO diaries (300 to 1,000 patients, 7 visits), **rmcorr misses the true r twice as often as MLCorr on average, and up to four times as often with irregular visits.**

Indeed, the 95% CI misses the true r:

rmcorr: 4 to **29%** (regular visits),
5 to **42%** (irregular visits)

MLCorr: 4 to **7%** (regular visits),
4 to **9%** (irregular visits)

The shortfall grows with study size, where researchers trust rmcorr most.



OUR ADVICE

Prefer **MLCorr** for within-person r in repeated PRO data.

None of the 240 simulated conditions favoured rmcorr. MLCorr ties or wins on every measure.

Thank you

THE SEEINGTHETA TEAM



Piper Fromy
Psychometrics, Methodology and Strategic Consulting



Elisabeth Berger
Statistics and Programming



Kacylia Pistoia
Psychometrics and Methodology



Maryam Bayram
Psychometrics and Methodology



Nikola Markiwicz
Psychometrics and Methodology



Marie-Jeanne Holveck
Statistics and Programming



Come and meet the team and Iris at the SeeingTheta booth.

We welcome collaborators to extend this work.

CONTACT

Piper Fromy and Marie-Jeanne Holveck
pip@seeingtheta.com
marie@seeingtheta.com

Questions welcome

BACKUP

The full scorecards

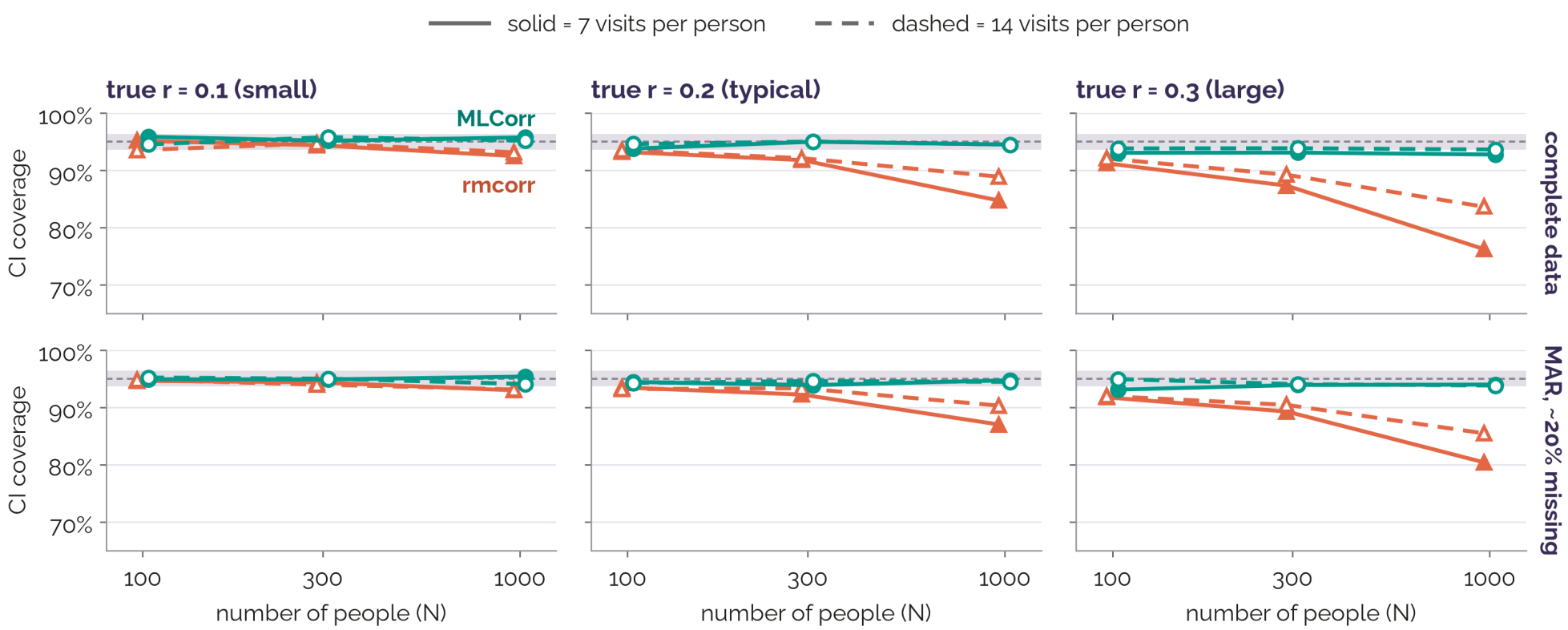
IN THIS SECTION

- Coverage in every condition
- Bias, theory check and CI width
- Paired coverage difference
- rmcorr's opposite biases cancelling vs honest MLCorr
- False positives and power
- Centring and autocorrelation
- References



Backup: coverage in every condition, AR(1) regular visits

Coverage = how often the 95% CI around r contains the true r. Grey band: 95% target ± simulation noise.

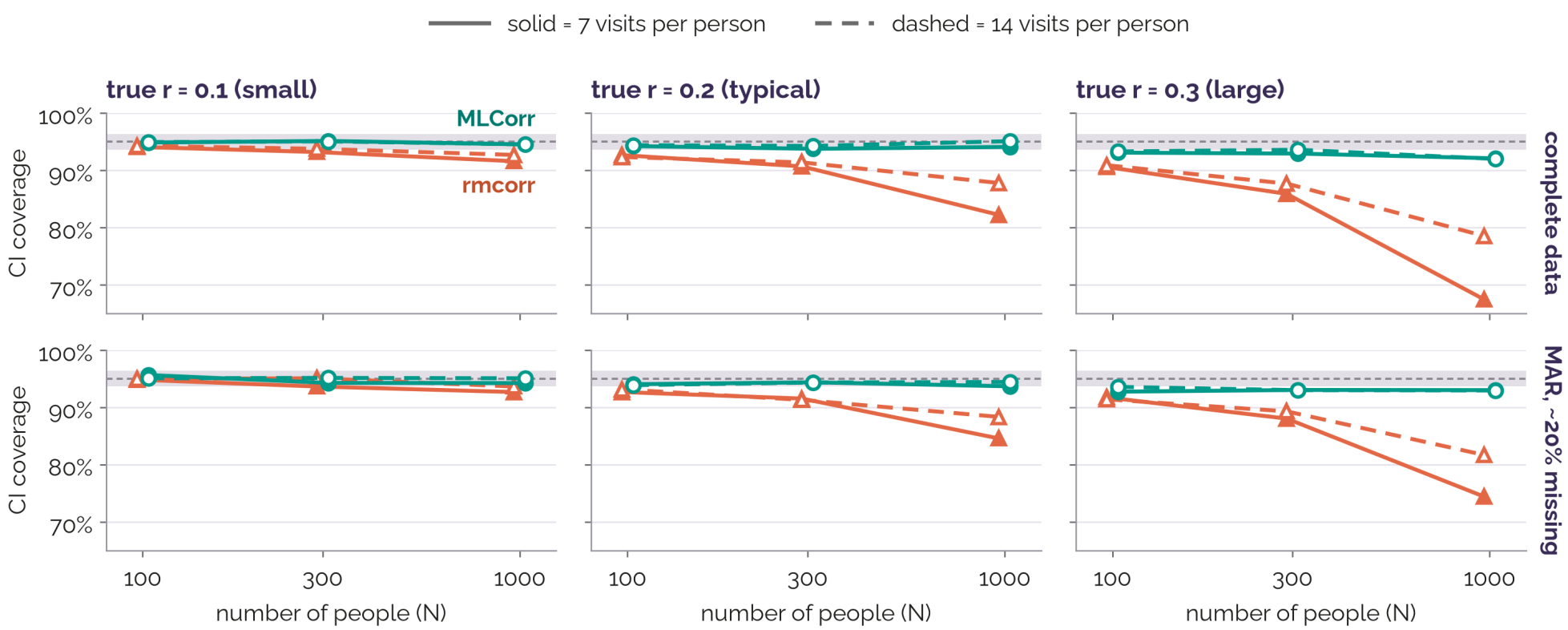


Averaged over slope heterogeneity. Solid = 7 visits, dashed = 14 visits.

MLCorr holds near 95% in every cell. rmcorr falls as N and r grow.

Backup: coverage in every condition, CAR(1) irregular visits

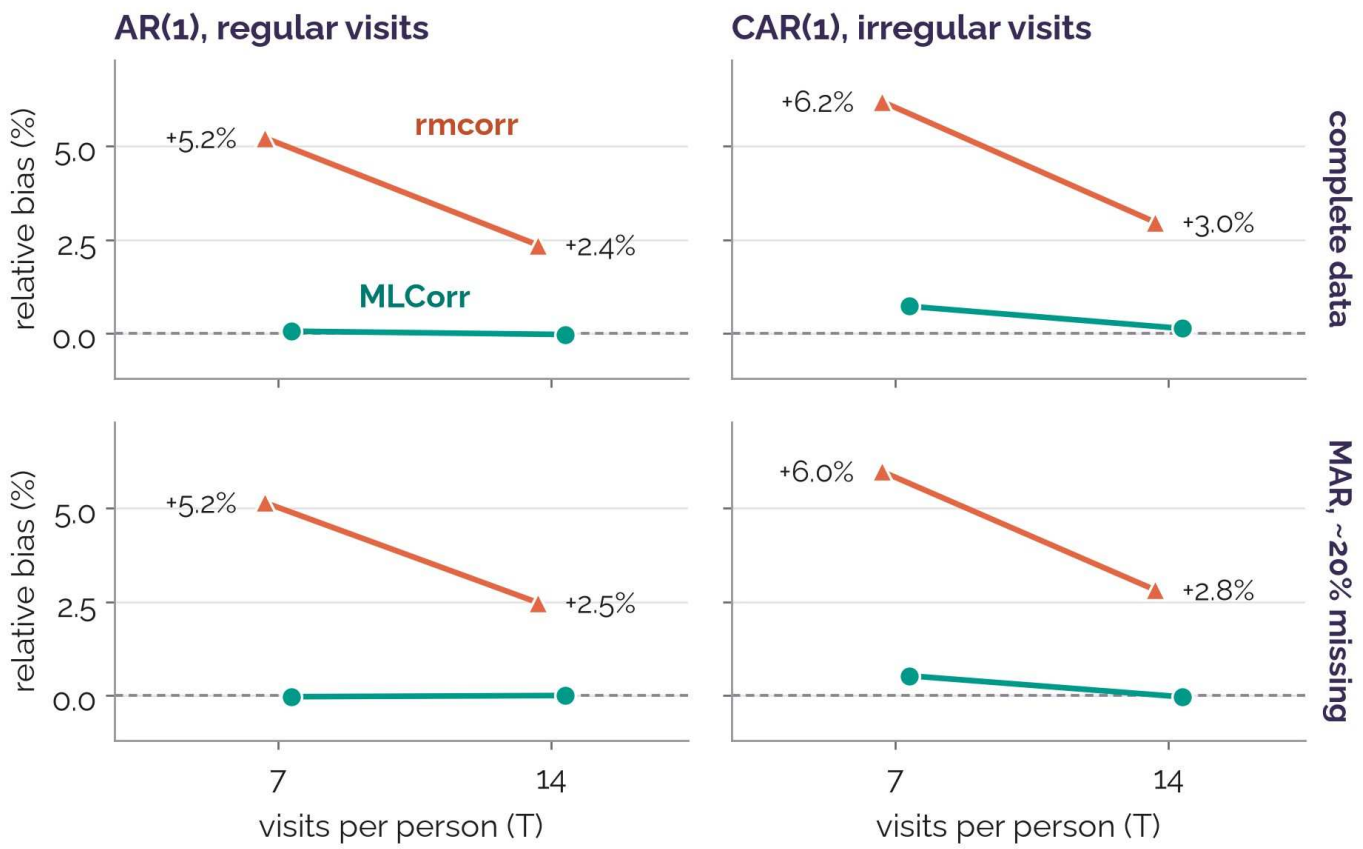
Coverage = how often the 95% CI around r contains the true r. Irregular visits hurt **rmcorr** more.



Averaged over slope heterogeneity. Solid = 7 visits, dashed = 14 visits.

With irregular visits, rmcorr's coverage falls to 68% on average at the hardest setting.

Backup: bias by number of visits



Averaged over N, true r and slope heterogeneity. Bars = ±1.96 Monte Carlo SE.

BIAS · ACCURACY

On average, how far is the estimated r from the true r? Shown as % of the true r. 0 = on target.

RESULT

MLCorr: close to 0 everywhere. **rmcorr**: about +5 to 6% at 7 visits, +2.4 to 3% at 14.

WHY

Both methods subtract each person's mean. rmcorr also assumes independent visits, so with autocorrelated residuals, the centring removes more noise than signal, so r comes out too high.

Doubling the visits halves rmcorr's bias. Short diaries are where it bites.

Backup: the simulation agrees with theory



WHAT THIS CHECKS

Crosses: bias predicted by a closed-form formula. Dots: bias in the simulation.

RESULT

They agree for every r and both visit counts. **MLCorr** sits at the predicted 0.

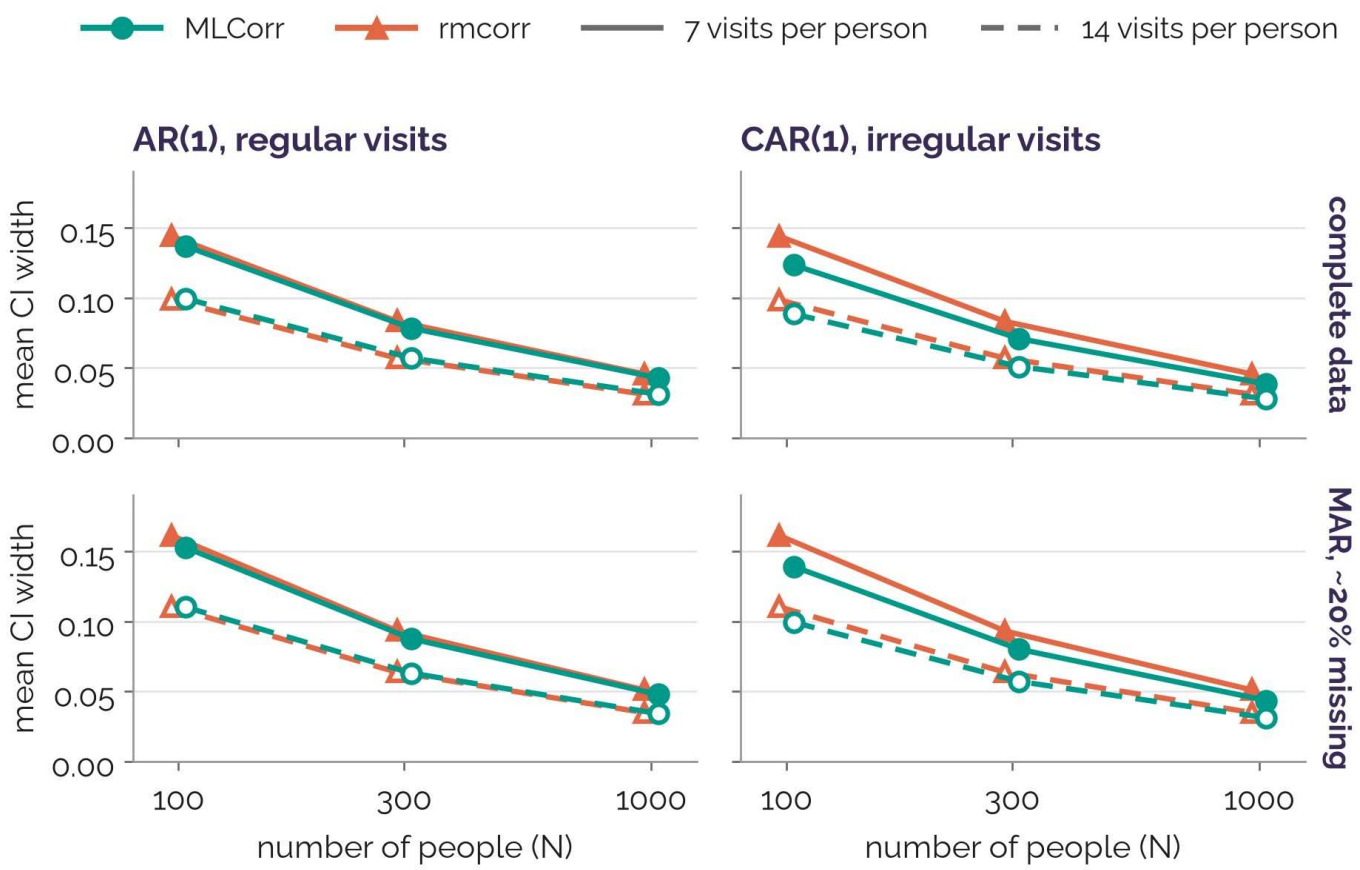
SO

The bias we report is the bias theory expects, not a coding artefact or a result due to chance.

Complete data, no slope heterogeneity, AR(1), averaged over N. Bars = ±1.96 MCSE.

The simulation is well calibrated.

Backup: MLCorr does not win by casting a wider net



True r = 0.3, averaged over slope heterogeneity.

CI WIDTH · PRECISION

How wide is the 95% CI around r? Lower = more precise.

RESULT

Both intervals narrow as N grows. **MLCorr**'s is as narrow as **rmcorr**'s, or narrower.

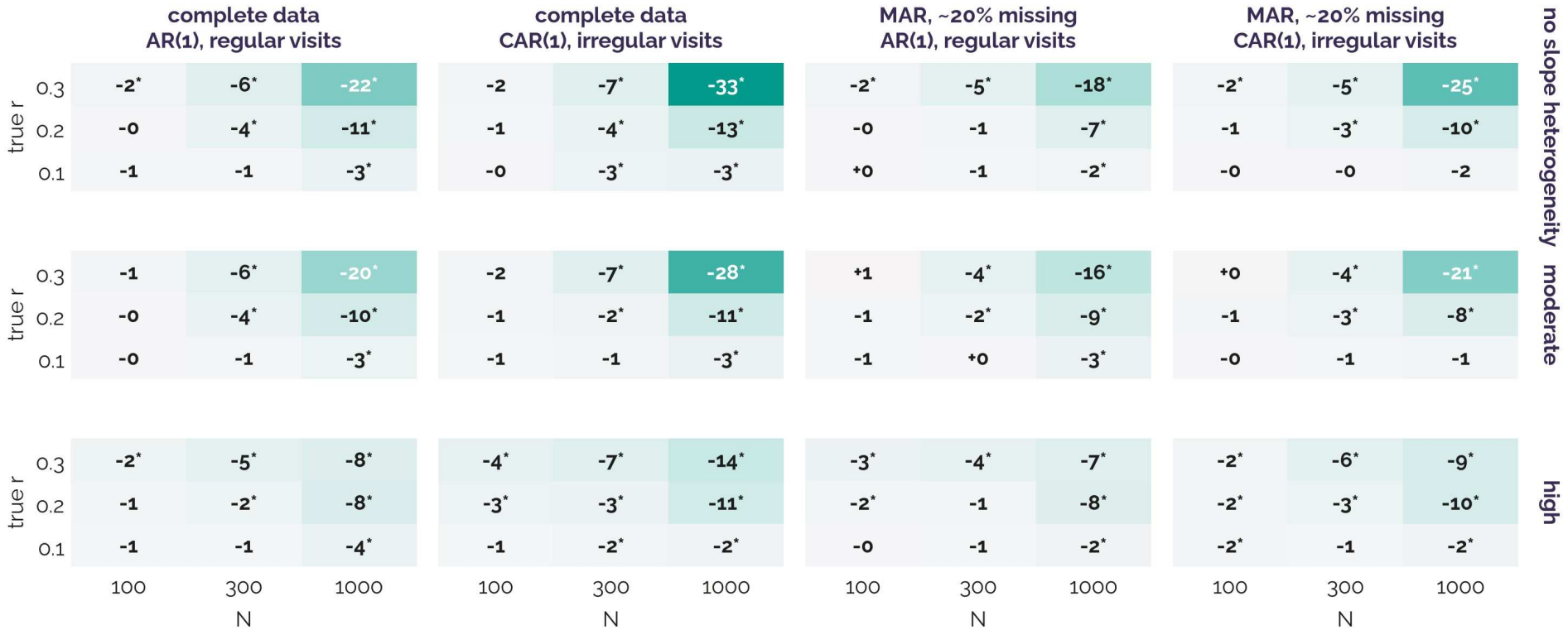
SO

MLCorr covers better because its interval sits on the true r. Coverage is about centring, not width.

Same width, better aim.

Backup: paired coverage difference, 7 visits per person

rmcorr minus **MLCorr** coverage, same datasets, in % points. Teal: **MLCorr** better. * = beyond noise.



MLCorr covers better almost everywhere. The gap is largest at high r and large N.

Backup: paired coverage difference, 14 visits per person

Same layout as before. With 14 visits the gaps are about half as large (worst cell: -16 points).

		complete data AR(1), regular visits			complete data CAR(1), irregular visits			MAR, ~20% missing AR(1), regular visits			MAR, ~20% missing CAR(1), irregular visits		
	true r	100	300	1000	100	300	1000	100	300	1000	100	300	1000
no slope heterogeneity	0.3	+1	-4*	-11*	-1	-4*	-16*	-1	-3*	-10*	-0	-2	-14*
	0.2	-0	-2*	-6*	-1	-1	-8*	-1	-0	-4*	+0	-2*	-6*
	0.1	+0	-1	-2*	-1	-2*	-2	+0	-1*	-1	-1	-1	-1
moderate	0.3	+1	-3*	-10*	+0	-3*	-14*	-2*	-3*	-8*	-0	-2	-13*
	0.2	-0	-3*	-5*	-0	-2*	-8*	-0	-1	-5*	+0	-2*	-6*
	0.1	-2*	-1	-2*	-1	-1	-1	-1	-2*	-0	-0	+0	-1
high	0.3	-7*	-7*	-8*	-6*	-10*	-10*	-6*	-5*	-6*	-6*	-8*	-7*
	0.2	-3*	-3*	-6*	-4*	-5*	-5*	-2*	-2*	-4*	-2*	-5*	-6*
	0.1	-1	-1*	-2*	+0	-1	-3*	-0	-0	-2*	+1	+0	-2*
			N			N			N			N	

Against our prediction, the gap narrows as slope heterogeneity grows.

Backup: Two opposite errors cancelling in rmcorr versus the honest MLCorr

Estimated r = signal / √ (signal² + noise²)

With increasing N

Method	Coverage	Bias	CI width	Conclusion
rmcorr	↓ poor	→ persistent	↓	centres each person (both methods do), then treat the visits as independent; with autocorrelated residuals, centring removes more noise than signal, and the independence assumption reads that as a tighter link, so r estimate is inflated. The bias is fixed; larger N just shrinks the CI around it, so coverage falls.
MLCorr	→ good	→ none	↓	models the autocorrelation and the slope spread, so it stays on target with an honest interval.

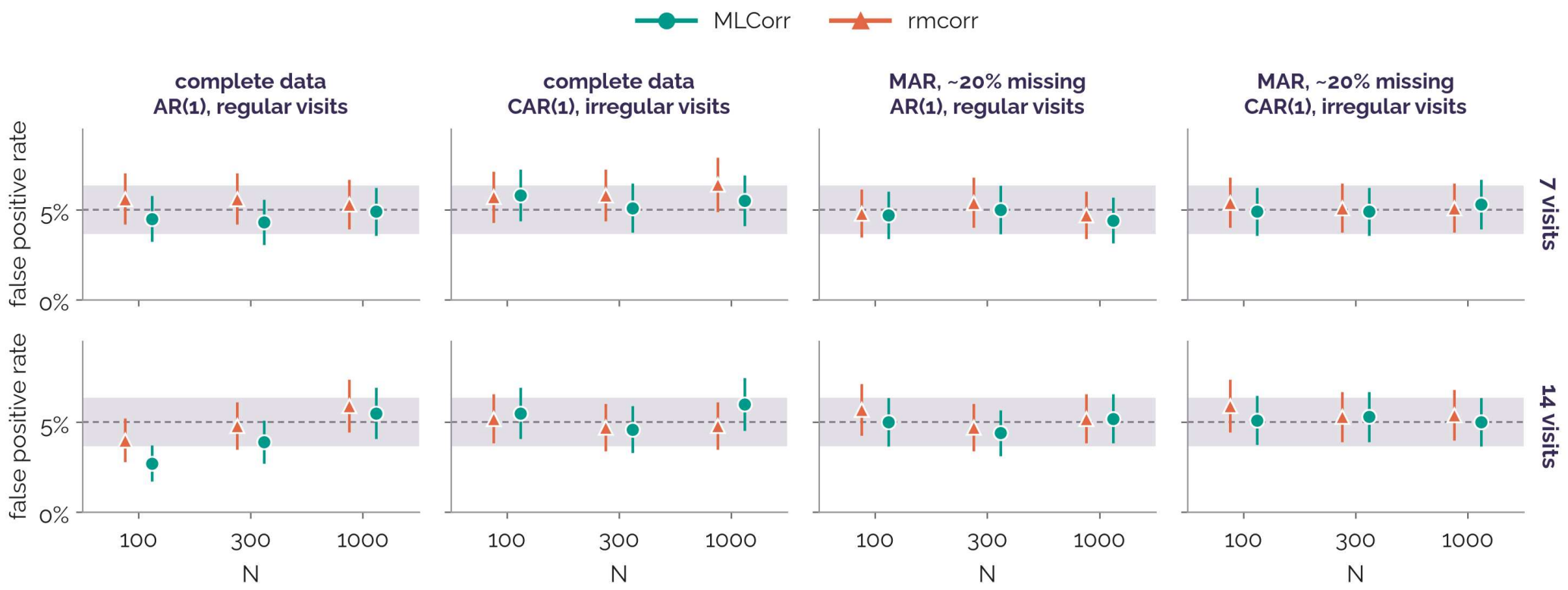
With increasing slope heterogeneity

Method	Coverage	Bias	CI width	Conclusion
rmcorr	↑less bad	↓ toward 0	→	varied slopes add extra between-person spread into the outcome (add noise) which dilutes r back down. So the two noise effects partly cancel: bias shrinks, coverage recovers. But it's cancellation, not correctness, and the CI ignores the slope spread.
MLCorr	→ good	→ none	↑widens	models the autocorrelation and the slope spread, so it stays on target with an honest interval.

Two biases cancel in rmcorr: correct by luck, not by design. MLCorr is right on purpose, and its interval is honest.

Backup: false positives at true r = 0

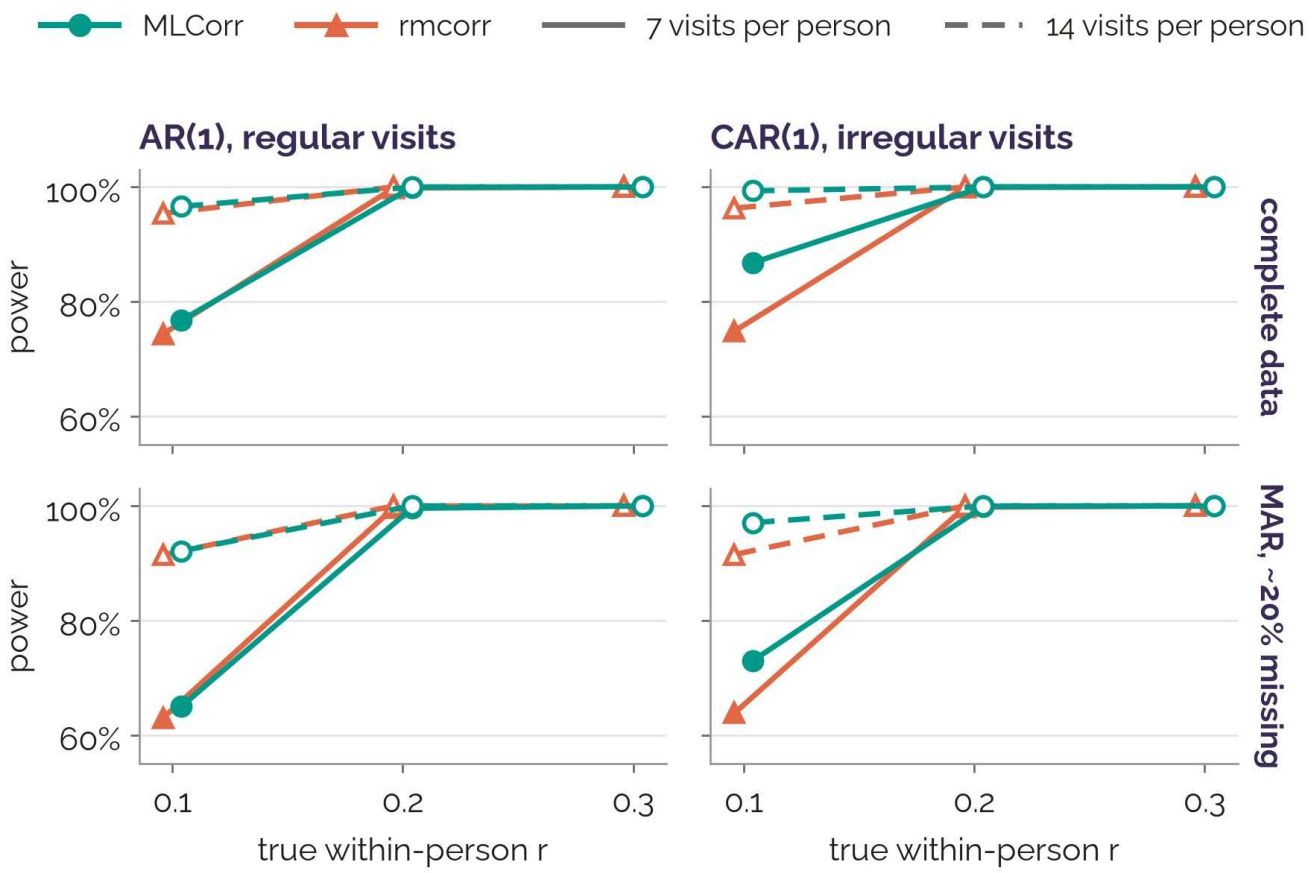
Type I error = at true r = 0, how often a method wrongly calls the correlation significant. Target: 5%.



Bars = ±1.96 Monte Carlo SE. Grey band = 5% ± simulation noise. No slope heterogeneity at r = 0.

No excess false positives. rmcorr's problem is estimating a real effect, not inventing one.

Backup: power at the smallest sample (N = 100)



Averaged over slope heterogeneity. At N = 300 and 1,000 both methods are near 100%.

POWER

At true r above 0, how often a method detects it. Power of 75% = the effect is missed 25% of the time.

RESULT

From r = 0.2 both are at 100%. At r = 0.1, **MLCorr** is equal or ahead: 87% vs 75% (CAR(1), complete, 7 visits).

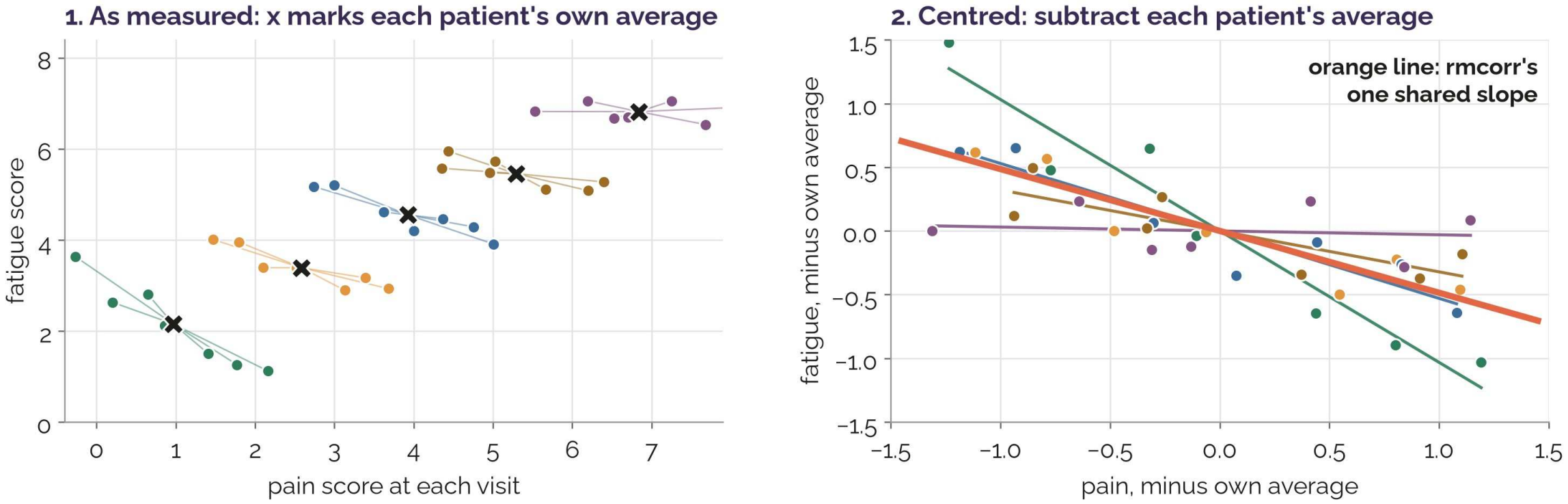
SO

MLCorr's honest intervals cost no power.

We are not trading sensitivity for validity.

Backup: what both methods do first, centre each person

Subtract each patient's own average. The five clouds land on top of each other at zero.

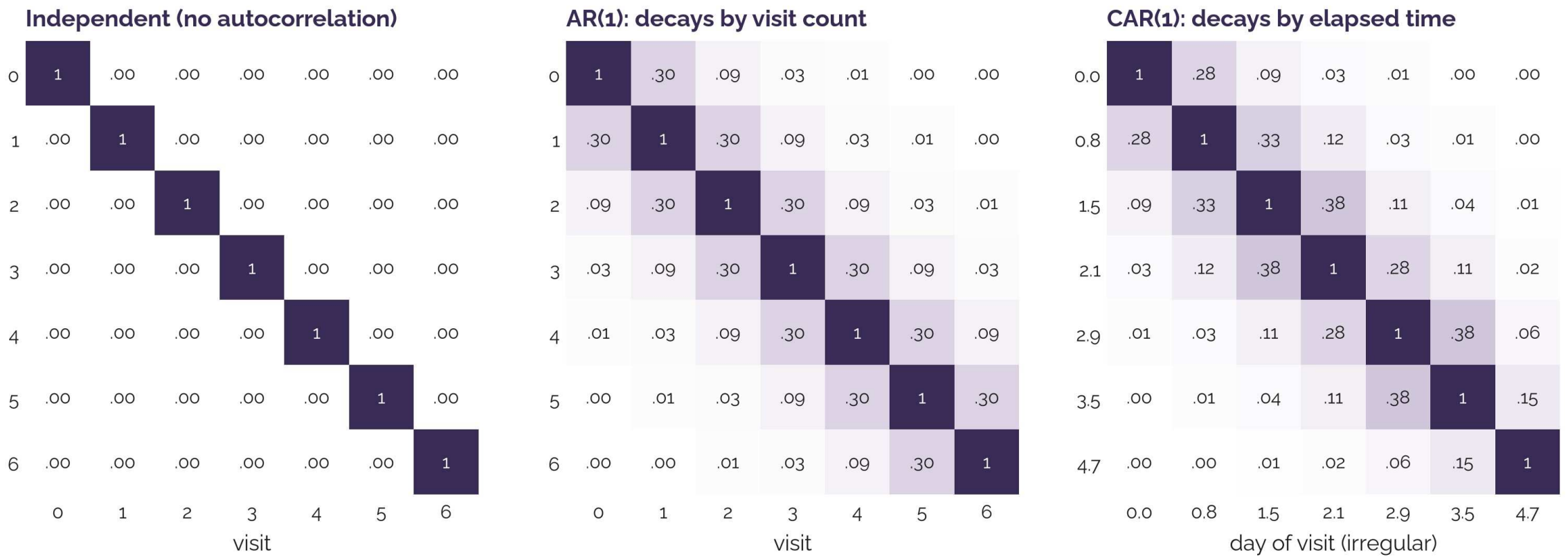


Illustrative simulated data: 5 patients, 7 visits each.

Centring removes between-person differences. Both methods start here, but differ in what comes next.

Backup: residual autocorrelation (ϕ = 0.3)

Correlation between the residuals of two visits, for 7 visits. Left: no autocorrelation, for comparison.



CAR(1) panel: one example person; days rounded to 0.1. Gaps between visits are random (mean 1 day).

AR(1) counts visits. CAR(1) uses the real time between visits.

References

Bolger, N., & Laurenceau, J.-P. (2013). Intensive longitudinal methods: An introduction to diary and experience sampling research. Guilford Press.

De Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. Frontiers in Psychology, 8, 1849. https://doi.org/10.3389/fpsyg.2017.01849

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. Applied Statistics, 43(1), 49. https://doi.org/10.2307/2986113

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. PNAS, 115(27). https://doi.org/10.1073/pnas.1711978115

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. Personality and Individual Differences, 102, 74-78. https://doi.org/10.1016/j.paid.2016.06.069

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. Psychological Science, 21(7), 984-991. https://doi.org/10.1177/0956797610372634

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. Psychological Methods, 25(5), 610-635. https://doi.org/10.1037/met0000250

Ringwald, W. R., Manuck, S. B., Marsland, A. L., & Wright, A. G. C. (2022). Psychometric evaluation of a Big Five personality state scale for intensive longitudinal studies. Assessment, 29(6), 1301-1319. https://doi.org/10.1177/10731911211008254

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. Psychological Assessment, 31(2), 226-235. https://doi.org/10.1037/pas0000662

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. JMIR, 21(12), e14475. https://doi.org/10.2196/14475

The sources behind each factor level on the obstacle course (slide 6)