Cool regressions for the ‘Midlife Physical Activity and SuperAging: A Link to Better Cognitive Function, Functional Status, and Mental Health in Older Women’

Josef Mana

In this report, I present a set of analyses based propensity-scores matching followed by g-computation of the averagae treatment on controls (ATC) effect of midlife/lifelong physical activity on current cognitive superaging status (SA), current physical activity level (PA) mean cognitive performance (Cognitive CS), cognitive screening (MMSE), depressive symptoms (GDS-15), anxiety (GAI), and functional non-independence (FAQ) symptoms in elderly ladies. Check the repository https://github.com/josefmana/cosactiw-brief-report.git.

Introduction

This report is a buffet of results inspired by Table 1 of the main article. In the table, we provide in-sample comparisons of ‘distribution similarity’ (via Mann-Whitney U test) and ‘stochastical independence’ (via Pearson’s χ^2 test) between NANOK and COSACTIW samples putatively representing the difference between samples with ***a lot of midlife physical activity (COSACTIW)*** and ***much less midlife physical activity (NANOK)***. In other words, we evaluated whether midlife physical activity relates to different distributions (e.g., higher/lower scores) in variables such as composite cognitive score (Cognition CS), cognitive screening (MMSE), functional independence difficulties (FAQ), depressive (GDS-15) or anxiety symptoms (GAI), and whether physically active women during midlife in our sample were more or less likely to achieve superager status (SA) or to be more physically active currently (PA) than women not as much physically active during midlife. Unfortunately, these samples differed substsantially in education distribution and to a lesser degree, age variance, possible ruining our fun from in-sample comparisons if causal inference with practical implications for larger population is of interest. Consequently, in this report I derive estimates that should be more robust than in-sample comparisons under the assumption that age and education are sufficient for explaining away confounding between midlife physical activity and the outcomes of interest described above.

Warning

The assumption, that age and education are sufficient to close all the back-doors (Cinelli et al., 2024; Pearl, 1995; Perković et al., 2018; Shpitser et al., 2012), is both critical to ensuring validity of presented results as correct estimates of causal effects, and possibly easily disputable. However, having some assumption about causal relations, reporting it, and logically deriving results based on this assumption is superior to stating no assumption about causal relations and still interpreting results as causal, even if the assumption is incorrect.

Tip 1: Use of Tips in the text

Throughout the report, I use the green 'Tip' boxes such as this one to try re-phasing information from the main text to what I hope will be more accessible form of presentation. To start, the introduction section says:

- I looked into COSACTW-vs-NANOK differences in the same variables as those reported in Table 1,
- I adjusted/'controlled' for age and education level in some fancy statistical procedure,
- if we believe age and education are 'good controls' then adjusting for them leads to valid results.

Important

I also use this red box for sentences that could be used in a finished article's Methods or Results section (or used as inspiration for such sentences). For example, the preceding section could be summarised in the Methods (Statistical analyses) or Results (its own section such as 'Robustness checks' or 'Post-hoc comparisons') section as:

- *"Since sampling procedures used for COSACTIW and NANOK resulted in samples with non-equivalent age and education level distributions, and both age and education can confound the effect of midlife physical activity on outcomes of interest, we further present comparisons adjusted for age, education-level, and their interaction with study group (COSACTIW vs NANOK)."*

Methods

To adjust our estimates for age and education, I opted for propensity scores matching (to 'balance' the samples in terms of age and education), followed by a set of weighted regressions (logistic or linear depending on the outcome) of outcome on group (COSACTIW vs NANOK), education level, age, and interaction of each education and age with group, and finished by estimating marginalised effects of several flavours (marginal effect of group, effect of group at each education level, and group/education interaction) via g-computation. I check neither effect of education because it would imply a different set of statistical models, nor effect of age from the same reason as well as because age ain't no variable that could be really intervened on.

Note

The type of effect I estimated in this report is the so-called *average treatment effect on controls (ATC)*. If valid (i.e., if its assumptions hold) it says what is the effect of expanding a treatment (midlife physical activity in our case) to those not receiving it. In other words, it tries to estimate how much controls (NANOK, not really physically active during midlife) would be better off in whatever variable we investigate (such as probability of SA nebo functional non-independence), if they were instead part of 'treatment' (COSACTIW, really a lot physically active during midlife).
There are other estimands one could use instead, principally:

- *average treatment effect on treated (ATT)* says what if instead being physically active during midlife the COSACTIW women were not physically active during midlife, how much would it hurt?
- vanilla *average treatment effect (ATE)* says what difference would be between a sample where all participants would be physically active during midlife, and the same sample if all the participants were not physically active during midlife?

I did not use ATT or ATE because (i) ATC seemed to be of most practical interest, (ii) ATT ended up having way to low 'effective sample size.'

Tip 2: Methods used

In this bunch of the text I simply list procedures used. The main takeaways for you should be that in the report I present numbers reflecting:

- means scores/proportions (e.g., percent of SA) in COSACTIW and NANOK that are not skewed by education or age,
- the difference between COSACTIW and NANOK that are not skewed by education or age,

- means scores/proportions in COSACTIW and NANOK separately for low and high education levels,
- differences between COSACTIW and NANOK depending on education (do COSACTIW high education differ from COSACTIW low education more than NANOK high education differ from NANOK low education?)

❗ Important

In a full-blown article, this would go somewhere to Methods (it is quite a mouthful, it could be shortened for sure):

- *"To adjust for confounding due to age and education, we used propensity scores matching to estimate marginal effect of midlife physical activity (operationalised as group membership to either COSACTIW or NANOK samples) on a set of cognitive function, and mental health outcomes. We used full matching on propensity scores estimated by a logit regression of the group membership on age and education level, which yielded good balance, as indicated by Figure 1. Next, we estimated a set of logistic and linear regressions for nominal and continuous variables respectively, regressing each outcome on group membership, age, education level as well as group memebership interaction with both age and educaton. In all regressions, the full matching weights were included in the estimation as implemented in the glm() function of base R (R Core Team, 2024). Finally, we estimated group-specific average values of each outcome of interest adjusted for age and education level, and compared these quantities via g-computation with a cluster-robust variance estimate of its standard error as implemented in the 'marginaleffects' R package (*Arel-Bundock et al., *Forthcoming). Both, the matching and g-computation analyses were used to estimate the Average Treatment Effect on the Control (ATC) with COSACTIW representing treatment group and NANOK representing the control group."*

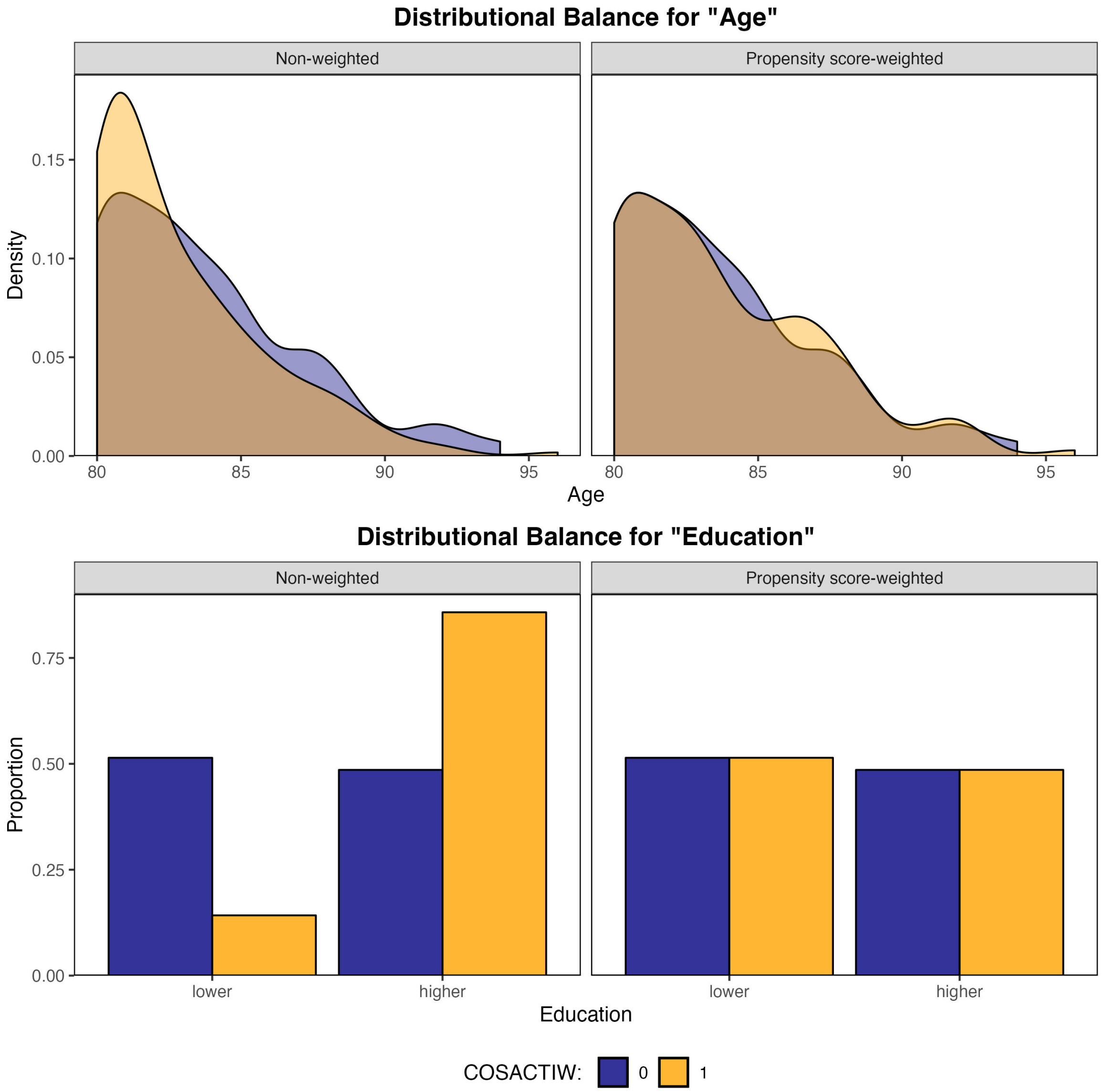

Figure 1: Disgtributional balance of age (top row) and education level (bottom row) in original sample (left column) and propensity scores weighted (right column) sample.

Results

⚠ Warning

In this section I present only the results as they relate to specified research questions. I do not show any model diagnostics or further technical discussion underlying reliability of the results (depending on whether statistical models 'behave well'). Although this is a very nerdy stuff, one point that could arise during review should be discussed nonetheless - I used classical Gaussian (normal) linear regression for outcome variables that are clearly not Gaussian (at least in the sample, maybe even never) such as MMSE, FAQ, GDS-15, and GAI. There are four main reasons why I do not consider the non-normality to matter much are:

1. for parameter estimates bias (i.e., 'validity'), the more important factor seems to be correct causal specification even when the data-generating process is very skewed and quite exotic, and classical linear regression still achieves correct frequentist properties (Type I and II errors, let us call them 'reliability' here) in such cases (Scholz & Bürkner, 2024); on the other hand, sure, we could do better if we wanted to dive deep into it, see https://users.aalto.fi/~ave/casestudies/Nabiximols/nabiximols.html if you are a masochist),
2. least squares linear regression is a maximum entropy solution for finding a mean of a distribution with homoscedasticity, and I tried posterior predictive checks on FAQ model - sample mean and SD were well reproduced by the model, distribution shape was not - the results will most likely still be valid for inference, not so for prediction though (cf. Zhang et al., 2023),
3. the continuous variables are sums of answers, they should converge to normal distribution in theory if well constructed,
4. the g-computation estimates of standard errors ought to be 'robust' to non-normality and (more importantly) heteroscedasticity from what I understood.

In this section, I present two kinds of estimates, (i) COSACTIW vs. NANOK comparisons adjusted for age and education (Table 1), and (ii) COSACTIW vs. NANOK comparisons within each education level as well as comparisons between COSACTIW-versus-NANOK differences between education levels (i.e., sample/education level interaction) (Table 2). Nominal variables (SA and PA) are presented as estimated per group percentages for description/prediction and odds ratios (ORs) for comparisons. Continuous variables (Cognitive CS, MMSE, GDS-15, GAI and FAQ) are presented as estimated means for description/prediction and differences between means for comparisons. For each outcome variable, I present the estimate and its 95% confidence interval (CI), and for comparisons I further present p-values and s-values for statistical testing. S-value (i.e., the Shannon information) is a cognitive tool to help researchers intuitively evaluate strength of evidence against a null hypothesis contained in the results as equivalent to the number consecutive 'heads' tosses that would provide the same amount of evidence against the null hypothesis that the coin is fair (Cole et al., 2021; Greenland, 2019). All null hypothesis take the form of 'no difference' implying OR = 1 or difference-in-means = 0.

Table 1: Model-based predictions and comparisons between COSACTIW and NANOK samples marginalised over age and education level

	Predictions[a]		Comparisons[b]		
	COSACTIW[c]	NANOK[c]	Estimate[c]	p-value	s-value[d]
SA	37.25% [31.02, 43.93]	32.54% [22.43, 44.58]	1.15 [0.73, 1.80]	0.548	0.87
PA	90.12% [84.75, 93.74]	70.38% [57.73, 80.53]	1.27 [1.07, 1.50]	0.006	7.37
Cognitive CS	0.24 [0.16, 0.32]	0.20 [0.05, 0.36]	0.04 [-0.20, 0.28]	0.739	0.44
MMSE	28.21 [28.00, 28.41]	27.28 [26.91, 27.65]	0.92 [0.26, 1.58]	0.006	7.29
GDS-15	1.85 [1.53, 2.17]	3.01 [2.43, 3.58]	-1.16 [-2.01, -0.31]	0.007	7.11
GAI	2.68 [2.22, 3.14]	4.00 [3.18, 4.82]	-1.33 [-2.76, 0.09]	0.067	3.90
FAQ	0.37 [0.24, 0.51]	0.96 [0.72, 1.20]	-0.59 [-1.08, -0.10]	0.018	5.80

[a] Values represent model average prediction of proportions (SA and PA) or means (continuous variables) marginalised over levels of education and age.
[b] Values represent odds ratios (SA and PA) or difference between means (continuous variables) of each outcome (rows) comparing COSACTIW vs NANOK marginalised over levels of education and age derived via g-computation after propensity scores matching for average treatment effect on controls (NANOK).
[c] Values are presented as estimate [95% confidence interval].
[d] Shannon information transform of p-values. It can be interpreted as an answer to the question of how many consecutive 'heads' tosses would provide the same amount of evidence (or 'surprise') against the null hypothesis that the coin is fair?

Table 2: Model-based education-wise predictions and comparisons between COSACTIW and NANOK samples marginalised over age

	Predictions[a]		Comparisons[b]		
Education	COSACTIW[c]	NANOK[c]	Estimate[c]	p-value	s-value[d]
SA					
higher	42.62% [33.49, 52.27]	37.71% [22.96, 55.16]	1.14 [0.76, 1.73]	0.526	0.93
lower	32.45% [24.41, 41.67]	28.01% [15.66, 44.93]	1.15 [0.48, 2.79]	0.751	0.41
higher / lower	-	-	0.99 [0.37, 2.63]	0.986	0.02
PA					
higher	92.48% [85.11, 96.36]	74.41% [55.88, 86.97]	1.25 [1.01, 1.55]	0.036	4.79
lower	87.31% [79.22, 92.55]	66.26% [48.29, 80.51]	1.28 [0.98, 1.68]	0.069	3.87
higher / lower	-	-	0.98 [0.69, 1.38]	0.895	0.16
Cognitive CS					
higher	0.27 [0.15, 0.39]	0.27 [0.05, 0.48]	0.04 [-0.22, 0.29]	0.782	0.35
lower	0.21 [0.10, 0.33]	0.14 [-0.07, 0.36]	0.04 [-0.35, 0.44]	0.824	0.28
higher - lower	-	-	-0.01 [-0.48, 0.46]	0.970	0.04
MMSE					
higher	28.71 [28.42, 29.01]	27.62 [27.08, 28.15]	1.04 [0.40, 1.67]	0.001	9.44
lower	27.73 [27.44, 28.02]	26.96 [26.44, 27.48]	0.81 [-0.32, 1.95]	0.161	2.64
higher - lower	-	-	0.22 [-1.07, 1.52]	0.736	0.44
GDS-15					
higher	1.33 [0.87, 1.79]	3.16 [2.33, 3.99]	-1.87 [-3.10, -0.64]	0.003	8.47
lower	2.34 [1.89, 2.79]	2.86 [2.05, 3.66]	-0.49 [-1.66, 0.68]	0.412	1.28
higher - lower	-	-	-1.38 [-3.08, 0.32]	0.112	3.16
GAI					
higher	2.28 [1.62, 2.94]	4.06 [2.87, 5.24]	-1.90 [-3.72, -0.08]	0.040	4.63
lower	3.06 [2.42, 3.70]	3.94 [2.80, 5.09]	-0.79 [-2.97, 1.39]	0.477	1.07
higher - lower	-	-	-1.11 [-3.96, 1.74]	0.445	1.17
FAQ					
higher	0.28 [0.09, 0.48]	0.79 [0.45, 1.14]	-0.54 [-1.06, -0.03]	0.037	4.74
lower	0.46 [0.27, 0.64]	1.12 [0.78, 1.45]	-0.64 [-1.47, 0.20]	0.135	2.89
higher - lower	-	-	0.09 [-0.90, 1.08]	0.857	0.22

[a] Values represent model average prediction of proportions (SA and PA) or means (continuous variables) marginalised over levels of education and age.

[b] Values represent odds ratios (SA and PA) or difference between means (continuous variables) of each outcome (rows) comparing COSACTIW vs NANOK marginalised over levels of education and age derived via g-computation after propensity scores matching for average treatment effect on controls (NANOK).

[c] Values are presented as estimate [95% confidence interval].

[d] Shannon information transform of p-values. It can be interpreted as an answer to the question of how many consecutive 'heads' tosses would provide the same amount of evidence (or 'surprise') against the null hypothesis that the coin is fair?

Appendix

References

Arel-Bundock, V., Greifer, N., & Heiss, A. (Forthcoming). How to interpret statistical models using marginaleffects in R and Python. *Journal of Statistical Software*.

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, *53*(3), 1071–1104. https://doi.org/10.1177/00491241221099552

Cole, S. R., Edwards, J. K., & Greenland, S. (2021). Surprise! *American Journal of Epidemiology*, *190*(2), 191–193.

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, *73*(sup1), 106–114.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Perković, E., Textor, J., Kalisch, M., & Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, *18*(220), 1–62. http://jmlr.org/papers/v18/16-319.html

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Scholz, M., & Bürkner, P.-C. (2024). *Prediction can be safely used as a proxy for explanation in causally consistent bayesian generalized linear models*. https://arxiv.org/abs/2210.06927

Shpitser, I., VanderWeele, T., & Robins, J. M. (2012). *On the validity of covariate adjustment for estimating causal effects*. https://arxiv.org/abs/1203.3515

Zhang, S., Heck, P. R., Meyer, M. N., Chabris, C. F., Goldstein, D. G., & Hofman, J. M. (2023). An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, *120*(33), e2302491120. https://doi.org/10.1073/pnas.2302491120