Does the choice of Sleep Regularity Index software change the conclusions? A re-analysis of the NHANES 2011–2014 sleep regularity–obesity association

Jürgen Degenfellner[1]

[1]Institute of Physiotherapy, ZHAW School of Health Sciences, Winterthur, Switzerland

July 28, 2026

Abstract

Background. The Sleep Regularity Index (SRI) can be computed by different open-source packages. Czeisler et al. (2025) showed that the package alone changes SRI values and their apparent associations with health outcomes, prompting a reporting standard (RIRI). I test whether a published SRI–obesity finding changes under the standard calculators, and why.
Methods. NHANES 2011–2014 provides a ready-made minute-level sleep/wake classification (`PAXPREDM`); feeding both packages this same series holds sleep detection fixed and isolates the SRI computation. On the byte-identical series of my published sample ($n = 7085$) I recomputed the SRI with `GGIR` and `sleepreg` at defaults and compared them to the published SRI (computed by hand with the Phillips formula). I then re-estimated the published survey-weighted log-BMI model under each calculator, reporting multiplication factors with 95% bootstrap confidence intervals, and refitted it split by recording completeness.
Results. On identical input the calculators disagreed: package SRIs were about 7 median points lower, correlated 0.81 (`GGIR`) and 0.67 (`sleepreg`) with the published SRI, reshuffled about half of participants across quintiles, and `sleepreg`'s default 5-day minimum discarded 30% of the sample. Every calculator preserved the inverse direction, but not the magnitude: the fully adjusted top-versus-bottom-quintile effect on BMI shrank from about −5% to −1.4% (`sleepreg`). The divergence was structured, driven by how each package pairs days across a recording gap, and confined to incompletely recorded participants: among the 60% with seven consecutive days the effect was calculator-invariant, whereas the entire calculator dependence arose in the 40% with a day gap, an instance of regression dilution from exposure measurement error.
Conclusions. The qualitative finding is robust to the calculator, but its effect size and sex/ethnic effect modification are calculator-dependent: a measurement-error phenomenon, not a property of the association. Authors should report the calculator, version and minimum-day/missing-data options, and present effect sizes with intervals.

Keywords: Sleep Regularity Index; measurement error; regression dilution; reproducibility; secondary data analysis; NHANES; software implementation

1 Introduction

Sleep regularity, operationalised by the Sleep Regularity Index (SRI) (Phillips et al., 2017), is the percentage-scaled probability that a person is in the same state (asleep/awake) at two time points exactly 24 h apart, ranging from −100 (perfectly inverted) through 0 (random) to 100 (identical days).

Czeisler et al. (2025) compared the two widely used open-source SRI calculators, `sleepreg` (Windred et al., 2021) and `GGIR` (Migueles et al., 2019; van Hees et al., 2015), in 73,794 adults. They found that SRI scores differed markedly between packages, in both absolute and relative terms. Applied to prospective models of mortality, incident type 2 diabetes, and incident atrial fibrillation, the calculation method alone changed results and interpretation. They proposed a 14-item reporting standard (Reporting Items for Regularity Indices, RIRI).

In earlier work (Degenfellner et al., 2026) my colleagues and I reported an inverse cross-sectional association between the SRI and obesity in two nationally representative NHANES cycles (2011–2014, $n = 7085$): more regular sleep was associated with lower BMI, more so in women and most clearly in non-Hispanic White and other/multi-racial women. That study used *neither* package; the SRI was computed by hand, applying the original Phillips et al. (2017) formula to NHANES's own minute-level wake/sleep classification (the variable `PAXPREDM`). This raises a direct, answerable question prompted by RIRI: **would those conclusions change if the SRI had been computed with the standard packages instead of by hand?**

This is not merely a theoretical concern; it is already visible *within this cohort*. Reported SRIs for the NHANES 2011–2014 accelerometry span a wide range: a median of 84.5 (Wang et al., 2023, Suppl. Table 8), a mean of 63.0 (Yang et al., 2024) and a median of 63.9 (unweighted 61.5) in the published sample (Degenfellner et al., 2026), roughly 20 SRI points apart. These analyses differ in their analytic samples and processing pipelines, not only in how the SRI is computed, so this particular spread reflects more than the calculator. The point is simply that no single agreed SRI exists even for one dataset. The present study removes those confounds, fixing the participants and the sleep/wake series (below), to isolate the calculator alone: the specific divergence RIRI quantifies. For reference, Windred et al. (2021) reported a median of 81 in the UK Biobank, a different cohort and device.

This setting takes Czeisler et al. (2025)'s comparison one step further. In their study the base sleep/wake classification was already held fixed: both packages received the *same* epoch-by-epoch sleep/wake states produced by `GGIR`, yet the SRIs still diverged, which Czeisler et al. (2025) attributed to the packages' different algorithms and to `sleepreg`'s additional re-handling of those shared states (its own nap and wake-after-sleep-onset rules, based on sustained inactivity). That leaves one further source of divergence still in play: the sleep/wake series itself, which the two packages did not treat identically. NHANES 2011–2014 closes it. It distributes a ready-made minute-level sleep/wake classification (`PAXPREDM`); I give this one series to all three calculators and run `sleepreg`'s SRI engine directly on it, so its nap and wake-after-sleep-onset reclassification, which needs the raw accelerometry, never runs. The sleep/wake series is then identical for every calculator, and any remaining divergence must come from the **SRI computation itself**: epoch resolution, day-pairing logic, missing-data handling, and minimum-day thresholds. Whether these implementation differences alone move a published association is what I test here, making NHANES an informative special case that complements RIRI.

2 Methods

2.1 Identical sleep/wake input

I reconstructed the minute-level binary sleep/wake series of the earlier study. From the NHANES 2011–2012 and 2013–2014 Physical Activity Monitor minute files (PAXMIN), `PAXPREDM` was recoded to binary sleep (sleep= 1, wake= 0, non-wear/unknown=missing). I applied the same inclusion filters: full-minute observations, no quality flags, ≥ 3 days with < 120 min non-wear, days with exactly 1440 minutes, ≥ 3 consecutive days including a weekend day, and $< 30\%$ missing binary sleep per participant. NHANES does not store the clock time of each minute directly; it stores the monitor's start time and a running count of 80 Hz samples. I used these to give every minute its real time of day (start time plus the sample number divided by 80). This matters because the SRI, and both packages, compare each day with the next at the same time of day, which only works with correct timestamps. The result is one shared minute-level sleep/wake series for all 7085 participants of the published sample, so `GGIR` and `sleepreg` run on identical input and any difference between them, or from the published SRI, reflects only the calculator.

2.2 The published SRI and the two package calculators

The baseline is the **published** SRI (Degenfellner et al., 2026), computed by hand with the Phillips formula at 1-min epochs over weekday-matched consecutive days ($\text{SRI} = -100 + \frac{200}{N_v} \sum 1[s_i = s_{i+1440}]$ over valid epoch pairs). On the reconstructed series I recomputed the SRI with the two standard packages, each at its default settings, and compared them to these published values.

1. **`sleepreg`** (`SRI_from_binary`, v1.3.5). The series is converted to the package's transition format and re-expressed at *10-second* epochs (`sleepreg`'s fixed internal resolution); the SRI is then computed over the *continuous* recording at a fixed 24 h lag (not weekday-matched); days with > 6 h missing are dropped; participants with < 5 valid comparison days return missing (default `minSRIdays`$= 5$).

2. **`GGIR`** (`CalcSleepRegularityIndex`, v3.3.6). `GGIR` computes a Phillips SRI for each pair of consecutive *calendar* days, matched by time of day at 1-min epochs and using the valid (non-missing) epochs in each pair, then averages these day-pair values into one SRI per person.

`sleepreg`'s published pipeline targets `GGIR` 2.0-0 and is incompatible with current `GGIR` (3.3.6), itself an instance of the reproducibility problem RIRI flags. I therefore ran `sleepreg`'s SRI engine directly from its (`GGIR`-independent) source functions on the binary series.

2.3 Agreement and association

Agreement between methods was summarised by Pearson and Spearman correlation, Bland–Altman limits of agreement, and cross-tabulation of SRI-quintile membership. For each SRI version I re-ran the published models on the same 7085-participant frame (survey design with 4-year MEC weights, strata and PSUs): the fully adjusted continuous model (log BMI on SRI plus 13 covariates), the quintile model (top vs bottom quintile), and the sex- and ethnicity-interaction models. I report effect sizes as multiplication factors $\text{MF} = \exp(\beta_{\text{SRI}}\,\Delta)$, where Δ is the fixed difference in median SRI between the top and bottom quintile, with 95% bootstrap confidence intervals (500 resamples of participants, refitting the survey-weighted model each time), matching the original study's procedure so that the "Original" row is directly comparable.[1] To compare calculators directly rather than by their separate intervals, I also bootstrapped the *difference* in multiplication factor between two calculators, resampling the same participants for both in each of 500 replicates (a paired bootstrap) and taking the 95% interval of the difference.

[1] As a robustness check, the design-based Taylor-linearisation intervals returned by `survey::svyglm` (which account for the strata and PSUs analytically rather than by resampling) were near-identical: for the published SRI under Model 2, 0.932 to 0.969 design-based versus 0.931 to 0.969 bootstrap.

Table 1: Agreement of the package SRIs with the published SRI, on the same NHANES minute-level sleep/wake classification (`PAXPREDM`, $n = 7085$).

	Median SRI	Pearson r vs. published	Same quintile as published	n with SRI (of 7085)
Original (published)	61.5	1.00	—	7085
`GGIR`	54.6	0.81	49.3%	7085
`sleepreg`	54.6	0.67	44.1%	4964

3 Results

3.1 SRI values disagree across packages on identical input

Median SRI was 61.5 (published), 54.6 (`GGIR`) and 54.6 (`sleepreg`), so both packages sat about 7 points lower. Against the published SRI, `GGIR` correlated 0.81 and `sleepreg` 0.67 (Spearman 0.79 and 0.69; Figure 1). Only 49.3% of participants fell in the same SRI quintile under `GGIR` as in the published values, and 44.1% under `sleepreg` (12% moved by ≥ 2 quintiles), so roughly half changed quintile depending on the calculator. `sleepreg`'s default 5-day minimum returned missing for 2121 of 7085 participants (30%), whereas `GGIR` retained all (Table 1).

3.2 The SRI–BMI effect size attenuates across calculators

Re-running the method of Table 3 in Degenfellner et al. (2026), the "Original" row reproduces the published values (Table 2, Figure 2), so every other row is directly comparable. The inverse direction was preserved under every calculator: every Model 1, Model 2 and women's Model 3 factor stayed at or below 1 (in men the association was already near the null), but the magnitude was not. The fully adjusted (Model 2) top-vs-bottom-quintile factor moved from 0.950 (95% CI 0.931, 0.969) (original) to 0.970 (95% CI 0.953, 0.988) (`GGIR`) to 0.986 (95% CI 0.967, 1.006) (`sleepreg`): an effect on BMI of about −5.0% (original) attenuating to −1.4% (`sleepreg`), with the `sleepreg` bootstrap interval spanning roughly −3.3% to +0.6%. The fully adjusted estimate *in women* (Model 3) moved likewise, from 0.918 (95% CI 0.893, 0.945) (original) to 0.941 (95% CI 0.919, 0.966) (`GGIR`) to 0.966 (95% CI 0.937, 0.994) (`sleepreg`). Comparing calculators directly (paired bootstrap of the difference), the fully adjusted top-vs-bottom-quintile effect was weaker under `GGIR` than under the published SRI by 0.020 (95% CI 0.008, 0.031) and under `sleepreg` by 0.045 (95% CI 0.027, 0.061); in women the corresponding differences were 0.024 (95% CI 0.008, 0.040) and 0.061 (95% CI 0.040, 0.084) (Figure 2B).

3.3 Effect modification weakens too

The sex×SRI and ethnicity×SRI interactions were clearest under the published SRI and `GGIR` and noticeably attenuated under `sleepreg`, where the interaction estimates shrank and their intervals widened (with 30% fewer participants). The "stronger in women" pattern is thus itself calculator-dependent in strength, though not in direction. This is most visible when the published study's main figure, Figure 2 of Degenfellner et al. (2026) (the fully adjusted model-4 predictions of BMI across the SRI range by sex and race/ethnicity), is reproduced under each calculator (Figure 3): the female slopes that carry the association, clearest for non-Hispanic White and other/multi-racial women, are steepest under the published SRI, shallower under `GGIR` and flattest under `sleepreg`,

Table 2: SRI–BMI multiplication factor (top vs bottom SRI quintile, Q1→Q5) under each SRI calculator, computed with the *identical method and models as Table 3 of Degenfellner et al. (2026)* (survey-weighted predictions at quintile-median SRI; 95% bootstrap CIs). Model 1: age + sex; Model 2: fully adjusted; Model 3: fully adjusted with sex×SRI interaction (sex-specific). The "Original" row reproduces the published values (shown in the *published* row beneath it), so every other row is directly comparable to the published paper. Only the SRI computation differs across rows.

SRI calculator	n	Model 1 (age+sex)	Model 2 (full adj.)	Model 3 (women)	Model 3 (men)
Original	7085	0.923 (0.907, 0.939)	0.950 (0.931, 0.969)	0.918 (0.893, 0.945)	0.982 (0.957, 1.002)
published		*0.923 (0.906, 0.939)*	*0.950 (0.932, 0.970)*	*0.920 (0.898, 0.945)*	*0.981 (0.961, 1.002)*
`GGIR`	7085	0.942 (0.927, 0.959)	0.970 (0.953, 0.988)	0.941 (0.919, 0.966)	1.000 (0.978, 1.023)
`sleepreg` (min 5 d)	4964	0.959 (0.938, 0.978)	0.986 (0.967, 1.006)	0.966 (0.937, 0.994)	1.009 (0.985, 1.038)
`sleepreg` (min 3 d)	6796	0.958 (0.940, 0.976)	0.981 (0.963, 0.999)	0.961 (0.933, 0.987)	1.002 (0.983, 1.024)

while the near-flat male slopes are unchanged. The effect modification is real under every calculator but its *strength* tracks the SRI's measurement quality.

3.4 The `sleepreg` attrition is a default effect, not a data limit

Almost all participants recorded 6–7 calendar days (< 5 raw days for only 1.4%). The 30% loss under `sleepreg` arises from two defaults acting together on NHANES's non-wear load: days with > 6 h missing are discarded, and only epoch pairs observed on *both* days count, so the number of valid comparison days falls below the `minSRIdays`= 5 default for 29.9% of participants. Lowering this single default to 3 (the original study's rule) recovers n from 4964 to 6796, but barely moves the effect estimate: the fully adjusted top-vs-bottom-quintile factor goes from 0.986 (95% CI 0.967, 1.006) at `minSRIdays`= 5 to 0.981 (95% CI 0.963, 0.999) at `minSRIdays`= 3 (Table 2). The default therefore explains the analytic sample loss but only a small fraction of the attenuation: even at `minSRIdays`= 3 the `sleepreg` effect (0.981) remains appreciably weaker than the published SRI (0.950), and the residual gap is intrinsic to its 10-second, continuous-lag arithmetic.

3.5 The disagreement is structured, not random: it concentrates in gapped recordings

The calculator disagreement is not uniform scatter around the identity line. The published-versus-package plots show several distinct point clouds with different slopes (Figures 1 and 5), which a single Pearson correlation cannot capture because it averages over a mixture. Colouring the points by candidate features identifies the latent variable: the number of *consecutive* recording days. Participants with seven consecutive valid days (about 60% of the sample) lie on the identity line, where the packages and the published SRI agree almost exactly. Participants with a *day gap* (non-consecutive days; about 40%) split off into separate clouds in which the package SRI is systematically lower, on average 22 SRI points lower at six valid days and 32 points lower at five. The gap between `sleepreg` and the published SRI correlates $r = 0.70$ with the number of non-consecutive day boundaries, but hardly at all with missingness or survey wave. The mechanism is the day-pairing convention: the hand Phillips calculation of the published study pairs days by *day of week* (and skips the gap), `GGIR` pairs consecutive calendar dates, and `sleepreg` uses a continuous real-time 24 h lag, so a gap in the recording makes the calculators compare different day pairs. The clouds are discrete because the number of valid days is. Fitting a separate line per valid-days group makes this explicit (Figure 4): regressing the `sleepreg` SRI on the published SRI gives a slope of 0.97 ($r = 0.98$) at seven consecutive days, 0.49 ($r = 0.54$) at six days, and 0.19 ($r = 0.14$) at five days, so the agreement collapses as days are lost and at five days the two are almost unrelated. The corresponding `GGIR` slopes are 0.94, 0.81, 0.60. The calculator one chooses therefore matters most for the participants with the least complete data.

3.6 The calculator dependence is confined to gapped recordings

The day-pairing account makes a clear prediction: where a recording has *no* gap the calculators return the same SRI, so the SRI–BMI effect should be *calculator-invariant* there, and all calculator dependence should live in the gapped recordings. Refitting the fully adjusted model separately by recording completeness confirms this. Among the 60% with seven consecutive valid days ($n = 4249$), the top-vs-bottom-quintile factor is essentially identical whichever calculator is used: 0.944 (published), 0.951 (`GGIR`) and 0.943 (`sleepreg`), a spread of just 0.008. Among the 40% with a day gap the same factor fans out: 0.959 (published), 0.981 (`GGIR`) and 0.998 (`sleepreg`), a spread of 0.039, with `sleepreg` now at essentially no effect. Every calculator therefore reproduces the published association in the well-recorded majority; they diverge only where a missing day forces their day-pairing rules apart. A paired bootstrap of the difference from the published SRI confirms the same split: among the seven-day recordings the differences are near zero (`GGIR` +0.007, (95% CI -0.003, 0.017); `sleepreg` +0.003, (95% CI -0.002, 0.009)), whereas among the gapped recordings they are larger (`GGIR` +0.022, (95% CI 0.001, 0.042); `sleepreg` +0.047, (95% CI 0.009, 0.092); Figure 6B).

4 Discussion

On the same fixed NHANES sleep/wake series (`PAXPREDM`), held identical for every calculator, the three SRI calculators still disagreed: ∼7-point median shifts, correlations as low as 0.67, half of participants reshuffled across quintiles, and a 30% swing in analytic sample size from a single default. This is a step beyond Czeisler et al. (2025), whose two packages shared `GGIR`'s base states but still differed in how `sleepreg` re-handled naps and wake-after-sleep-onset. An appreciable part of the divergence is therefore intrinsic to the SRI *computation* (epoch resolution, day-pairing, missing-data rules, minimum-day thresholds), not only to sleep classification.

4.1 Why the calculators disagree

With the sleep/wake series fixed for all three calculators, the divergence I observe must arise in the SRI computation itself. Reading the package source identifies four mechanisms, which my analyses then rank.

1. **Day pairing (dominant).** The original implements the Phillips formula by pairing epochs with the *same day-of-week index* one step apart (skipping non-consecutive days, wrapping Saturday→Sunday); `GGIR` (`CalcSleepRegularityIndex`) pairs each *consecutive calendar date* with the next, matched on time-of-day; `sleepreg` (`SRI_from_binary`) compares every epoch with the one exactly 24 h earlier along the *continuous* recording. When a recording is seven consecutive days the three rules pair the same epochs and agree almost exactly (Figure 4, left); when a day is missing they pair *different* epochs. The resulting bias is directional. With a gap, the calendar- and continuous-lag rules compare days that are further apart in real time, or offset from the true 24 h period. Sleep–wake patterns decorrelate as that interval

grows, so the overlap, and hence the SRI, is pulled *downward* (by about 22 points at six valid days and 32 at five). The published hand-Phillips rule instead skips the gap and compares only genuinely adjacent days, preserving the overlap; this is why the gapped recordings form the *lower* clouds rather than symmetric scatter. This mechanism accounts for most of the structured disagreement (the $r = 0.70$ between the `sleepreg`–published gap and the number of non-consecutive day boundaries).

2. **Minimum-day default.** `sleepreg` returns no value below `minSRIdays`= 5 valid comparison days; combined with its rule of dropping any day with > 6 h missing and counting only epoch-pairs observed on both days, NHANES's non-wear load pushes 30% of participants below the threshold. The published calculation and `GGIR` impose no such minimum. Relaxing the default to three days recovers most of the sample but changes the estimate only marginally.

3. **Epoch resolution.** The published calculation and `GGIR` compare 1-min epochs; `sleepreg` re-expresses the series at 10-s resolution. On an already-minute-level binary series this contributes little, but it is a further source of small, systematic differences.

4. **Missing-data / denominator convention.** Whether missing epochs are excluded (reducing the denominator) or imputed, and how, differs across implementations; in a companion multiverse analysis these imputation choices move the effect far less than the calculator choice, but they are not identical.

I verified that each calculator computes its intended quantity correctly: on synthetic recordings with a known SRI, the Phillips formula, `GGIR` and `sleepreg` all returned +100 for seven identical days, −100 for perfectly alternating days and 0 for days agreeing on half their epochs, and agreed on a pseudo-random pattern (Supplement). The divergence on real data is therefore a consequence of these algorithmic conventions, not of an implementation error. Nor is it an artefact of worse-fitting models: the fully adjusted model re-fitted with each calculator satisfied the regression assumptions equally well (near-identical residual skewness, kurtosis, variance inflation and homoscedasticity/linearity diagnostics; Supplement), so the attenuation under the packages is not a model-misspecification effect. I did, however, encounter one practical fragility: `sleepreg` (v1.3.5) aborts with an error when a recording begins exactly at local midnight, returning a missing value. Its pinned dependency on `GGIR` 2.0-0 also prevents installation alongside current `GGIR`, so I ran its SRI engine from source.

The implications for the substantive conclusion are mixed. The *direction* (more regular sleep, lower BMI, modified by sex and ethnicity) was reproduced by every calculator and is therefore robust. The *magnitude*, and the strength of the effect modification, were not: moving from the published SRI to `sleepreg` shrank the effect to under a third of its size (about −5% to −1.4%) and discarded a third of the sample. The results are better read as estimates: the family of plausible calculators maps to a range of effect sizes from about −5% to −1.4% (top vs bottom quintile), all in the same direction.

These results should *not* be read as significance verdicts. A reader applying a fixed $p < 0.05$ threshold might treat the calculators whose intervals include 1 as showing "no association" and the rest as showing one, and so draw opposite conclusions from the same data. That dichotomy has no principled basis: the 0.05 cutoff is arbitrary (Wasserstein et al., 2019). The point across calculators is not that some cross a line and others do not, but that the whole family of estimates points the same way with a magnitude that depends on the calculator; the effect sizes and their intervals are better read as a continuum than as significant/non-significant verdicts.

Why does a downward-biased, noisier SRI *weaken* the association rather than merely shift it? This is regression dilution: non-differential error in a continuous exposure attenuates its coefficient toward the null, the more so the larger the error. The packages inject such error, and they inject it almost entirely into the gapped 40% of participants. Splitting the fully adjusted model by recording completeness makes the mechanism explicit (Section 3.6): among the well-recorded 60%, where the calculators agree, the SRI–BMI factor is calculator-invariant (≈ 0.94 throughout). The entire calculator dependence of the overall estimate arises in the gapped subgroup, where the package SRIs are most degraded and their slopes approach the null (`sleepreg` 0.998). The attenuation is thus a property not of the association but of the exposure *measurement*: a quantitative example of the reproducibility problem RIRI describes.

4.2 Why published NHANES SRIs disagree so widely

The spread that motivates this study (a median SRI of 84.5 (Wang et al., 2023, Suppl. Table 8), a mean of 63.0 (Yang et al., 2024) and a median of 63.9 in the published sample (Degenfellner et al., 2026), all from the same NHANES 2011–2014 accelerometry) is wider than any single calculator produces here, and its origin cannot be recovered from the papers themselves. Beyond the SRI computation this study isolates, such analyses can differ in the sleep/wake source, the inclusion and non-wear filters, the epoch handling and the survey weighting. But because none release the code that produced the SRI, the reported values cannot be reconciled even when the stated definition is nominally the same. That unrecoverability, not any one calculator, is the reproducibility failure RIRI targets, and it is why reporting the full pipeline and sharing the code, as this reanalysis does, is the appropriate remedy.

These results support the RIRI recommendations even for datasets, like NHANES, where the sleep/wake classification is fixed. Authors should report the exact calculator and version, every non-default and default option (especially the minimum-day and missing-data rules), and, because the choice maps to a range of effect sizes, should present effect sizes with compatibility intervals rather than a single thresholded verdict. A limitation is that NHANES fixes the detection step, so my analysis isolates the computation component of the divergence; the raw-data detection component (not estimable here) would only add to it.

Recommendations for NHANES SRI users. (i) State the calculator and version and every minimum-day and missing-data option, default or not; these move both the estimate and the analytic sample. (ii) Prefer reporting effect sizes with compatibility intervals over a single thresholded verdict, since the plausible calculators map to a *range* of effect sizes. (iii) Treat the minimum-day default as an analytic choice: in NHANES, `sleepreg`'s `minSRIdays`$= 5$ silently removes a third of participants, and the loss is non-random with respect to wear time. (iv) Where feasible, report results under more than one calculator as a sensitivity analysis: a step beyond RIRI's disclosure checklist itself, but in keeping with its goal of reproducible SRI research (Czeisler et al., 2025).

Acknowledgements

The author thanks the NHANES participants and the National Center for Health Statistics for collecting and releasing the data.

Author contributions

Jürgen Degenfellner is the sole author: he conceived the study, performed all analyses, and wrote and approved the manuscript.

Statements and declarations

Ethical considerations. This is a secondary analysis of de-identified, publicly available NHANES 2011–2014 data. The NHANES protocol was approved by the National Center for Health Statistics (NCHS) Research Ethics Review Board. No further ethics approval was required for this reanalysis.

Consent to participate. All NHANES participants provided written informed consent at the time of the original survey.

Consent for publication. Not applicable; the study uses de-identified public data and reports no individually identifiable information.

Declaration of conflicting interests. The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding. The author received no financial support for the research, authorship, and/or publication of this article.

Data availability. All analyses use publicly available NHANES 2011–2014 data (`https://www.cdc.gov/nchs/nhanes/`). Code to reconstruct the sleep/wake series and reproduce every table and figure is openly available at `https://github.com/jdegenfellner/SRI_BMI_NHANES_2026_reanalysis_public`.

ORCID iD

Jürgen Degenfellner `https://orcid.org/0000-0001-6886-4583`

References

Czeisler MÉ, et al. (2025) Comparison of Sleep Regularity Index scores calculated by open-source packages and implications for outcomes research: rationale and design of the RIRI statement. *Sleep* 49(4): zsaf299.

Degenfellner J, Schernhammer ES, Strohmaier S (2026) Sex- and ethnic differences in the cross-sectional association between sleep regularity and obesity among US adults, NHANES 2011–2014. *J Biol Rhythms* 41(2): 205–218.

Migueles JH, Rowlands AV, Huber F, Sabia S, van Hees VT (2019) GGIR: a research community-driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. *J Meas Phys Behav* 2(3): 188–196.

Phillips AJK, Clerx WM, O'Brien CS, Sano A, Barger LK, Picard RW, et al. (2017) Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Sci Rep* 7: 3216.

van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. (2015) A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS One* 10(11): e0142533.

Wang X, Xu Y, Li X, Mansuri A, McCall WV, Liu Y, et al. (2023) Day-to-day deviations in sleep parameters and biological aging: findings from the NHANES 2011–2014. *Sleep Health* 9: 940–946.

Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond "$p < 0.05$". *Am Stat* 73(sup1): 1–19.

Windred DP, Jones SE, Russell A, Burns AC, Chan P, Weedon MN, et al. (2021) Objective assessment of sleep regularity in 60,000 UK Biobank participants using an open-source package. *Sleep* 44(12): zsab254.

Yang L, Wallen G, Barb J, Maki K, Farmer N (2024) 0863 Sleep regularity and cardiovascular health among U.S. adults. *Sleep* 47: A370–A371.

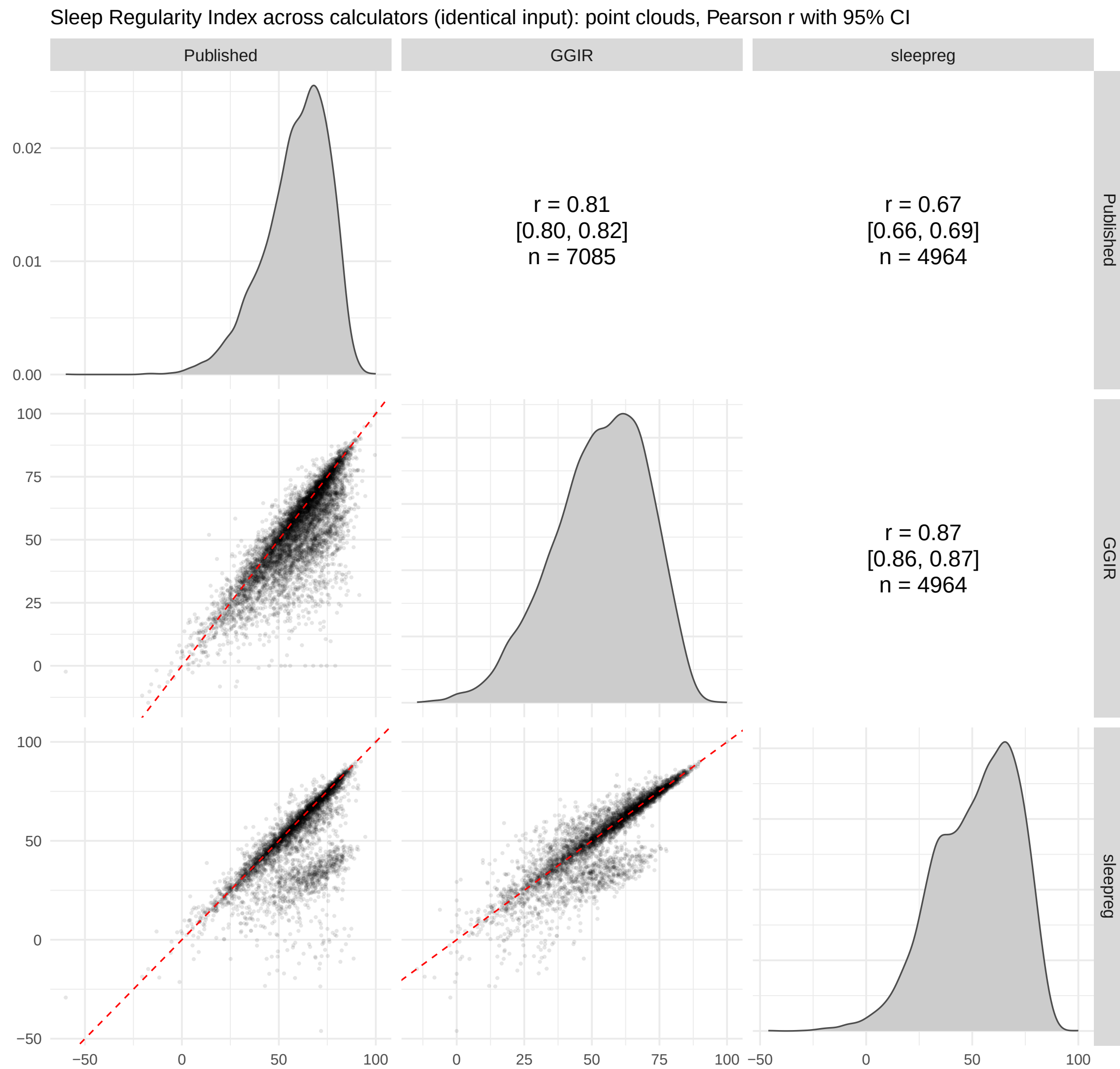

Figure 1: Pairwise agreement of the published SRI and the two packages on the same NHANES minute-level sleep/wake classification (`PAXPREDM`). Diagonal: density of each calculator's SRI. Lower triangle: all-versus-all point clouds (red dashed line is identity). Upper triangle: Pearson correlation with 95% confidence interval and the pairwise sample size. Note the secondary point cloud in the published-versus-package panels: a subgroup for whom the packages return a much lower SRI than the published values.

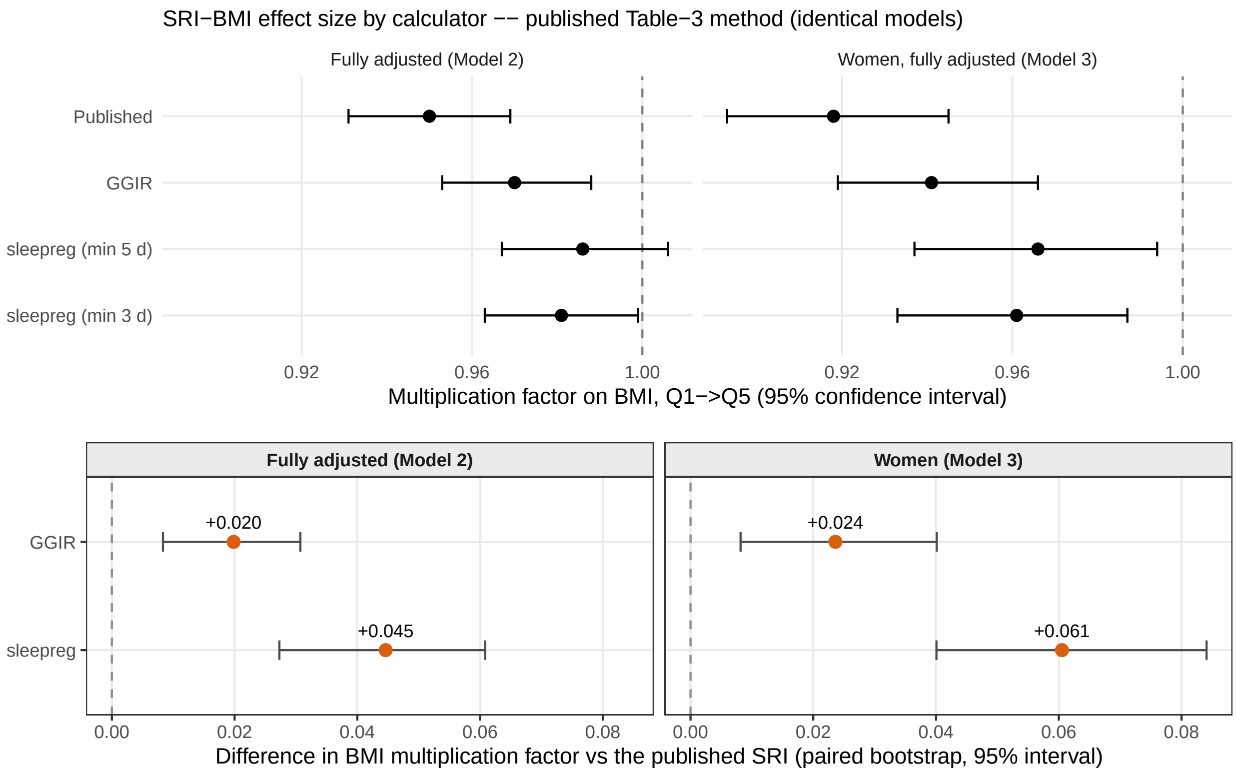

Figure 2: **(A, top)** SRI–BMI effect size (top vs bottom SRI quintile, multiplication factor on BMI with 95% bootstrap CI) by calculator, computed with the method and models of Table 3 in Degenfellner et al. (2026); left, fully adjusted (Model 2); right, women (Model 3). The inverse direction is preserved, but the magnitude attenuates from the published SRI to `GGIR` to `sleepreg`. **(B, bottom)** The *difference* in multiplication factor from the published SRI, estimated by a paired bootstrap (the same participants resampled for both calculators); point and 95% interval, with the dashed line marking no difference. Both packages give a weaker effect than the published SRI, more so `sleepreg`.

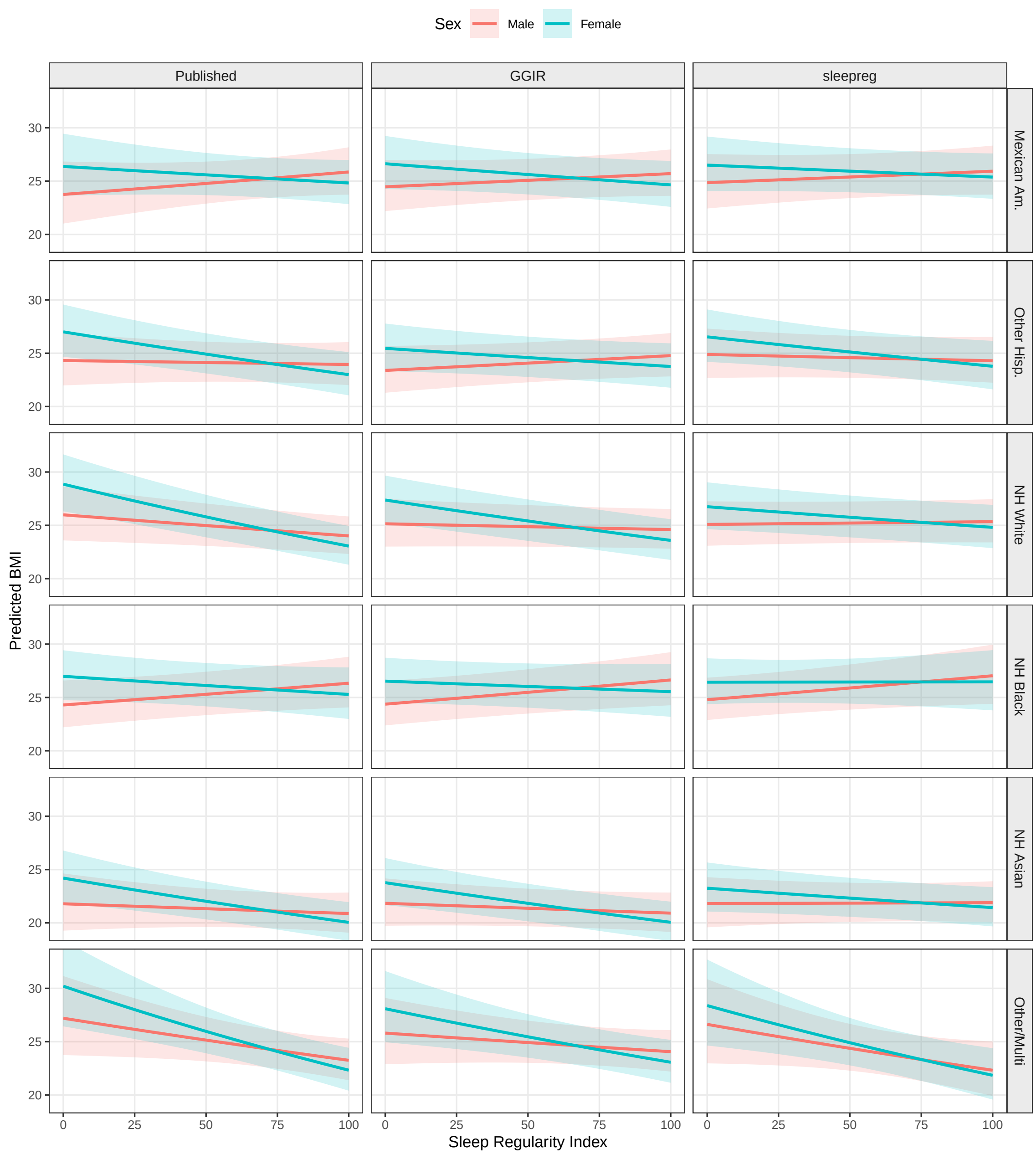

Figure 3: Figure 2 of Degenfellner et al. (2026) reproduced under each SRI calculator. Predicted BMI across the SRI range (0–100) by sex (Female teal, Male salmon) and race/ethnicity, from the fully adjusted interaction model (the published study's model 4: log-BMI on SRI in interaction with sex and with race/ethnicity, plus all covariates; survey-weighted), with design-based 95% bands. Rows are race/ethnicity, columns are the three calculators. The inverse female slopes attenuate from the published SRI (left) through `GGIR` to `sleepreg` (right), most clearly for non-Hispanic White and other/multi-racial women; the male slopes stay near flat throughout. The absolute BMI level reflects covariates held at reference, so only the slopes and their change across columns are to be compared. Bands are design-based and match a participant bootstrap to within about 5% in width.

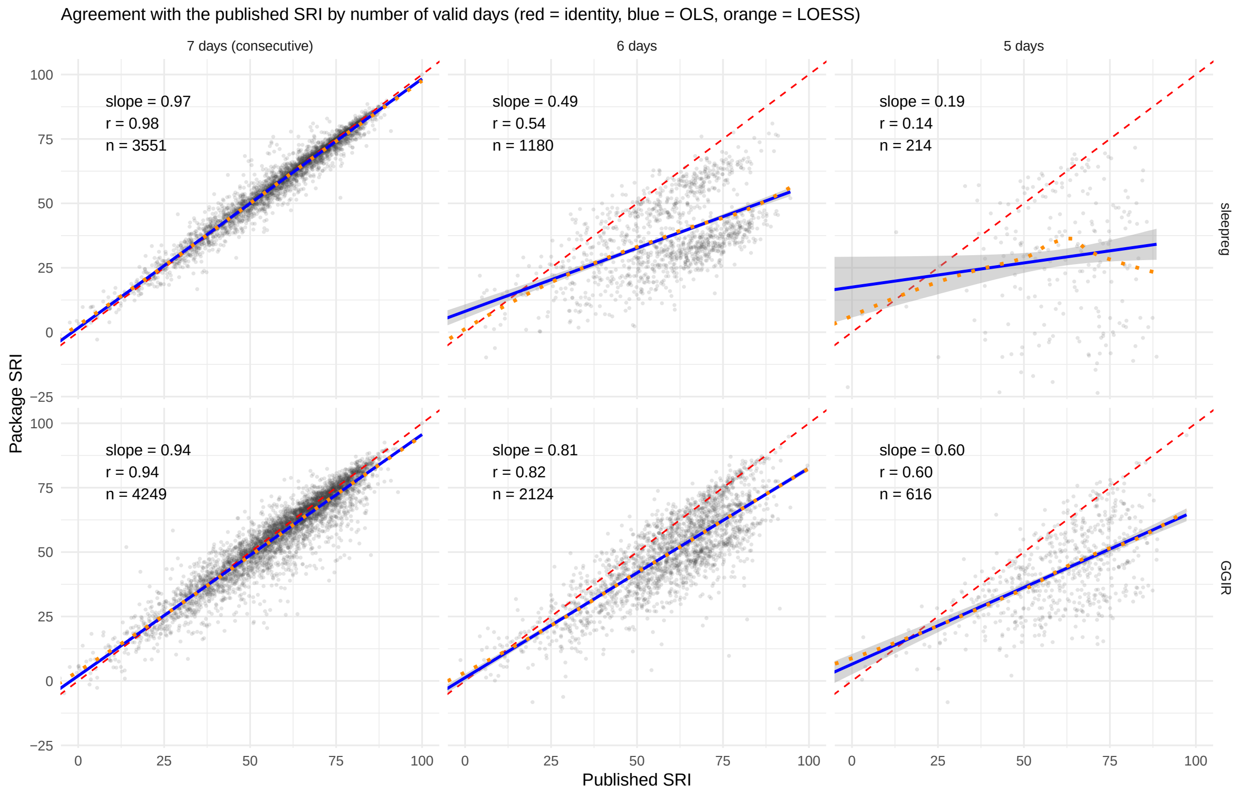

Figure 4: Package SRI versus published SRI, faceted by number of valid (consecutive) recording days (red dashed: identity; blue: OLS with 95% band; orange dotted: LOESS). At seven consecutive days agreement is essentially perfect (`sleepreg` slope 0.97, $r = 0.98$). With each missing day it degrades (slope 0.49 then 0.19 for `sleepreg`; 0.81 then 0.60 for `GGIR`). The fit is poor for the gapped groups not by accident: at six days the `sleepreg` cloud is visibly split into sub-clusters, and at five days there is barely any relationship ($r = 0.14$), so no single line, OLS or LOESS, describes it. The disagreement is a breakdown of agreement, not a uniform shift.

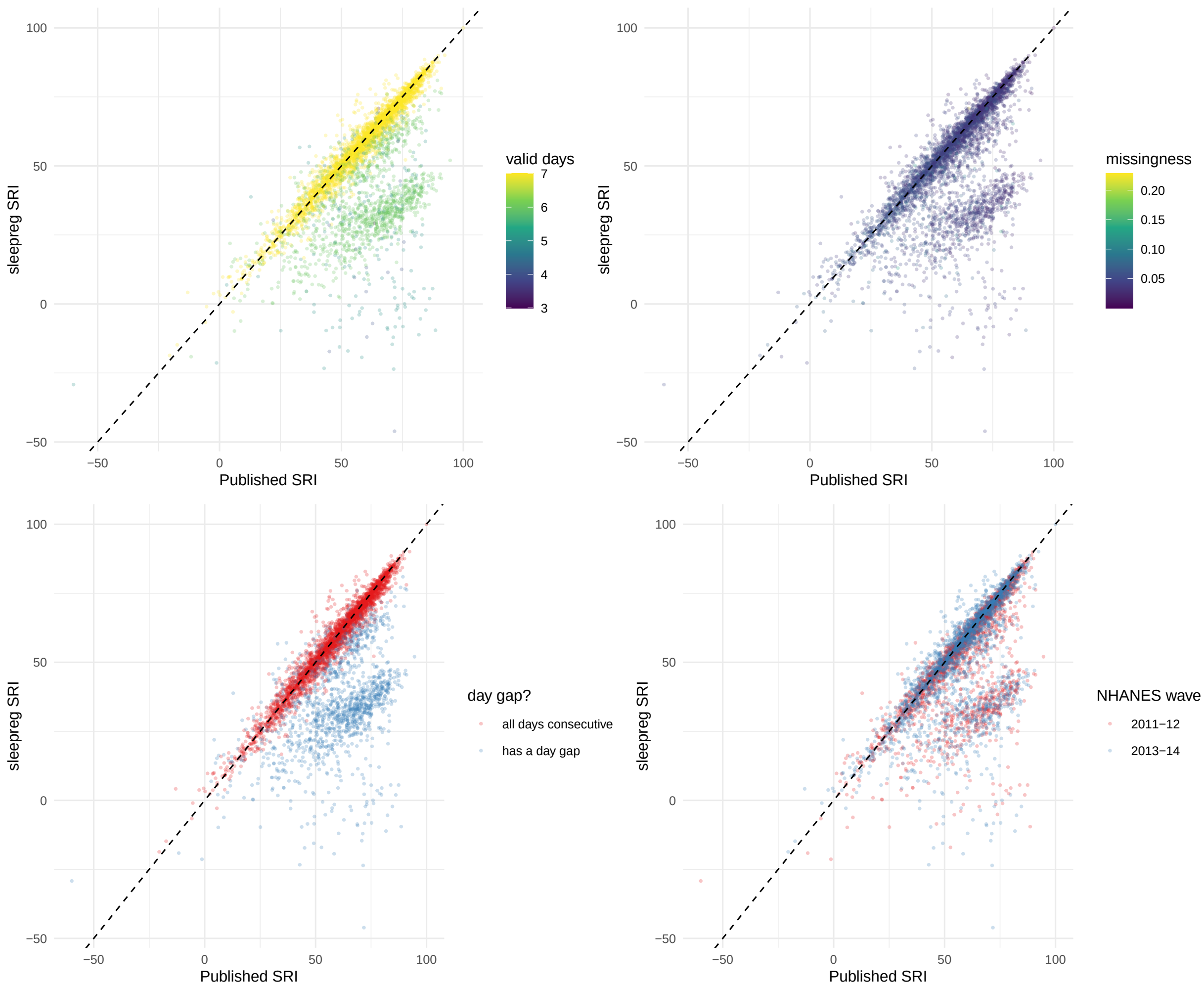

Figure 5: The `sleepreg`-versus-published SRI scatter, coloured by candidate per-participant features. The off-diagonal clouds are explained by the number of valid (consecutive) days / the presence of a day gap (top-left, bottom-left), not by missingness (top-right) or NHANES wave (bottom-right). Participants with all seven days consecutive lie on the identity line; those with a gap form the lower clouds.

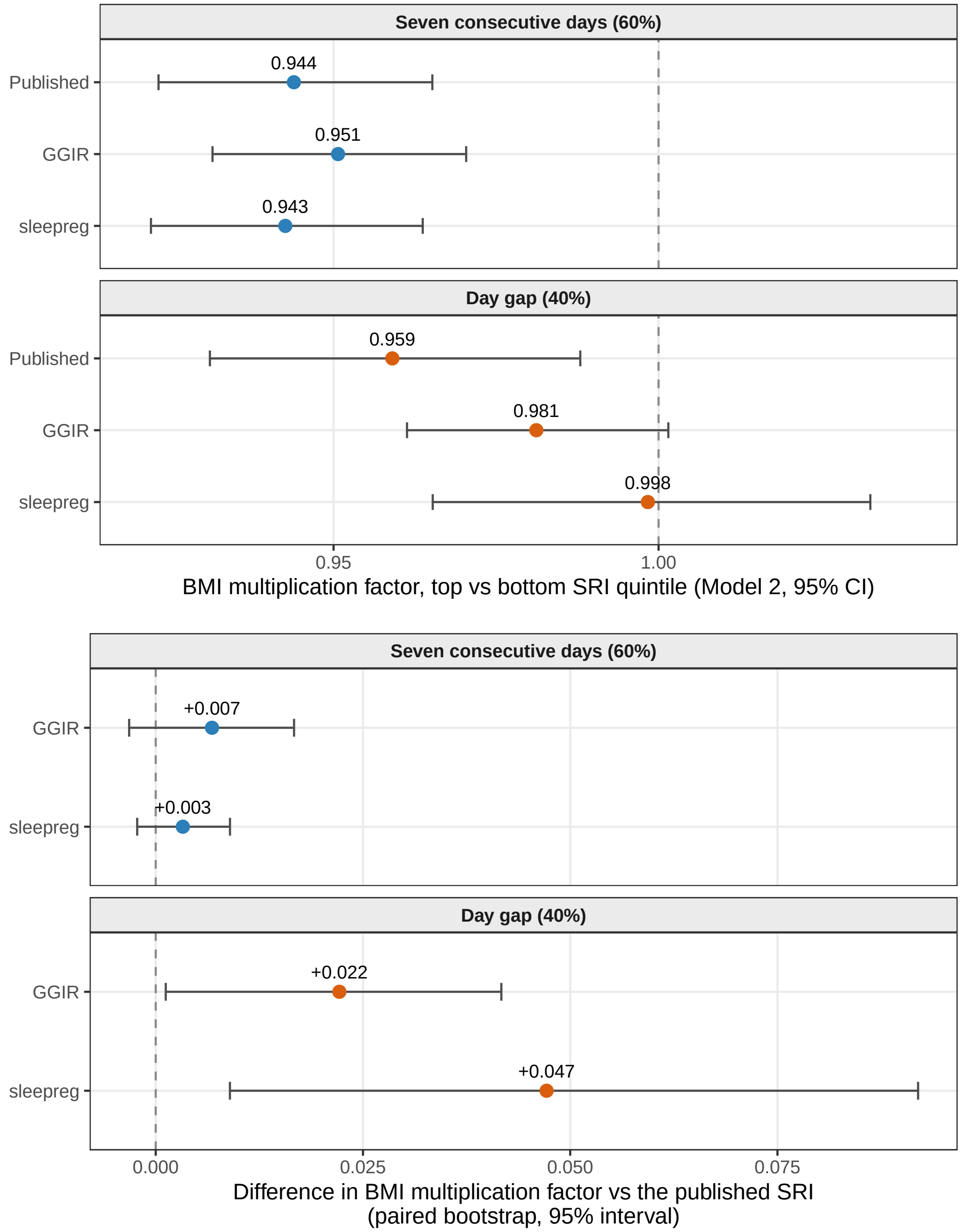

Figure 6: The SRI–BMI effect (Model 2, top vs bottom SRI quintile) by calculator, split by recording completeness. **(A, top)** Multiplication factor per calculator with 95% CI. Among participants with seven consecutive valid days the three calculators are essentially interchangeable (spread 0.008); all calculator dependence appears among those with a day gap, where the package estimates flatten toward the null. **(B, bottom)** The difference in multiplication factor from the published SRI (paired bootstrap; point and 95% interval, the dashed line marking no difference): near zero in the complete-recording subgroup (`GGIR` +0.007, `sleepreg` +0.003) and larger in the gapped subgroup (`GGIR` +0.022, `sleepreg` +0.047). The calculator dependence of the published association is thus a measurement-error effect of incompletely recorded participants.