Sleep Medicine Reviews 54 (2020) 101353

Contents lists available at ScienceDirect

Sleep Medicine Reviews

journal homepage: www.elsevier.com/locate/smrv

TECHNICAL REVIEW

Analyzing human sleep EEG: A methodological primer with code implementation

Roy Cox*, Juergen Fell

Department of Epileptology, University of Bonn, 53127 Bonn, Germany

ARTICLE INFO

Article history:
Received 16 March 2020
Received in revised form
30 April 2020
Accepted 30 April 2020
Available online 9 July 2020

Keywords:
Sleep
EEG
Slow oscillations
Sleep spindles
Cross-frequency coupling
Phase synchrony

SUMMARY

Recent years have witnessed a surge in human sleep electroencephalography (EEG) studies, employing increasingly sophisticated analysis strategies to relate electrophysiological activity to cognition and disease. However, properly calculating and interpreting metrics used in contemporary sleep EEG requires attention to numerous theoretical and practical signal-processing details that are not always obvious. Moreover, the vast number of outcome measures that can be derived from a single dataset inflates the risk of false positives and threatens replicability. We review several methodological issues related to 1) spectral analysis, 2) montage choice, 3) extraction of phase and amplitude information, 4) surrogate construction, and 5) minimizing false positives, illustrating both the impact of methodological choices on downstream results, and the importance of checking processing steps through visualization and simplified examples. By presenting these issues in non-mathematical form, with sleep-specific examples, and with code implementation, this paper aims to instill a deeper appreciation of methodological considerations in novice and non-technical audiences, and thereby help improve the quality of future sleep EEG studies.

Introduction

Accumulating evidence of sleep's involvement in cognition and neuropsychiatric disorders has propelled this unique brain state to the forefront of the fundamental and clinical neurosciences [1–3]. While the sleeping brain can be measured using a plethora of techniques, electroencephalography (EEG) remains the method of choice for human sleep research. As advances in EEG hardware (e.g., high-density amplifiers and caps) and analytic routines (e.g., time-frequency, functional connectivity, cross-frequency coupling analyses) have become commonplace in the wider neuroscience community, they have also begun permeating the sleep field. Indeed, by simultaneously capturing temporal, spectral, and spatial aspects of electrophysiological activity, modern EEG techniques (coupled with other neuroscience tools) hold great promise for understanding the complex organization and function of both non-rapid eye movement (NREM) and rapid eye movement (REM) sleep. For example, aspects of NREM activity (most notably, slow oscillations (SOs), sleep spindles, and their coupling), have been linked to aging and cognitive decline [4–6], schizophrenia [7–9], post-traumatic stress disorder [10,11], autism [12,13], as well as facets of cognitive, memory, and emotional functioning in healthy individuals (for reviews see [2,3,14]).

While increased interest in sleep EEG is to be welcomed, considerable theoretical and practical signal-processing expertise is needed to properly analyze sleep EEG and to assess the validity of published results. While the required level of expertise may be considered basic in certain engineering, physics, and mathematical communities, signal-processing training is less commonplace in the fields most likely to conduct or evaluate sleep EEG studies (e.g., psychology, neuroscience, medicine), at times negatively affecting the quality of published results. Although making errors is both inevitable and necessary to improve analytic skills, insufficient awareness of methodological issues may prevent error detection altogether. Combined with insufficient statistical rigor, improper application of analytic tools will yield erroneous and non-replicable results. By reviewing several methodological issues with relevant examples and code implementation, this paper is hoped to provide non-technical audiences an accessible entry point to contemporary sleep EEG analyses.

Our aim in this paper is twofold. First, we emphasize how small analysis choices can have large effects on outcome measures, and

* Corresponding author.
E-mail address: roycox.roycox@gmail.com (R. Cox).

https://doi.org/10.1016/j.smrv.2020.101353

Abbreviations

CA	common average
(d)PAC	(debiased) phase-amplitude coupling
EEG	electroencephalography
LM	linked mastoids
MEG	magnetoencephalography
NREM	non-rapid eye movement (sleep)
PLV	phase locking value
PSD	power spectral density
REM	rapid eye movement (sleep)
SL	surface Laplacian
SO	slow oscillation
SWA	slow wave activity
wPLI	weighted phase lag index

thereby interpretation of fundamental and clinical studies. While it is trivial that analysis choices will have some downstream effects, we have often been surprised by the degree of variability in outcomes. At the same time, repeating analyses in different ways while examining a multitude of potentially interesting outcome variables will increase the risk of false positives. Second, we stress the critical importance of checking that processing steps produce the desired output. Analysis scripts may run smoothly, generate reasonable outcome values, and yield compelling graphical plots, while not at all having performed the calculations the researcher envisioned, potentially compromising study findings. Such issues can be difficult to trace, but are greatly facilitated by visualizing intermediate results and applying analyses to stripped down or simulated data. To achieve these two aims, we discuss a set of data analysis procedures commonly encountered in sleep EEG studies, drawing attention to often overlooked issues and pointing out ambiguities in definitions. Accompanying Matlab scripts further illustrate these issues and may serve as starting points for custom analyses tackling fundamental and translational sleep EEG research questions.

In terms of scope, first, this paper does not provide an exhaustive overview of all relevant methodological issues, nor do we assume to cover the most important ones. While the chosen problems are certainly hoped to be of practical use, their main purpose is to illustrate the aims listed above. Second, many of the covered problems have been long and widely known in other signal processing and EEG communities, and indeed by sleep experts, and no novel insights are claimed. Rather, we hope that presenting these issues in non-mathematical form, from a dedicated sleep-EEG angle, and with code implementation, will make them more accessible, and their relevance more apparent, to non-technical audiences. Third, while the included topics and examples derive from our experience with human non-invasive scalp-level sleep EEG, many of the issues are equally applicable to source-reconstructed and invasive EEG recordings, magnetoencephalography (MEG), other species, and wakefulness. Fourth, since our focus is on data analysis, we will ignore electrode impedance[1] issues and the many preprocessing[2] steps that typically precede the generation of clean data. However, choices regarding filtering, artifact correction, channel interpolation, and so on, already constitute important decision points in the analysis chain.

Code and data

All code accompanying this paper was written in Matlab 2018a (Mathworks, Natick), and require Matlab's Signal Processing, Statistics and Machine Learning, and Curve Fitting Toolboxes, as well as EEGlab [16] functionality. Included data (58-channel NREM data separated into stages N2 and N3, sampled at 400 Hz, bandpass-filtered between 0.3 and 35 Hz) derives from three healthy participants undergoing whole-night polysomnography (EEG, electrooculography, and electromyography), sleep scored according to guidelines of the American Academy of Sleep Medicine [17], as reported previously [18,19]. Combined code and data (in EEGlab format) may be downloaded as a zip file from http://doi.org/10.5281/zenodo.3712074.

Spectral analysis

Spectral analysis is one of the best-known methods for describing signals with rhythmic components, and has a long history in sleep EEG [20]. It offers a broad and immediate overview of signal properties, including spectral components present, their variability, and overall signal quality, and it is often used to contrast experimental or clinical groups. While highly informative, variability in analytic strategies hampers study comparisons. Indeed, the term “power spectrum” itself is rather ambiguous, as it is often used rather loosely to refer to either power or power spectral density (PSD), two closely related but distinct concepts with different units (PSD: $\mu V^2/Hz$; power: μV^2). While PSD is considered the more natural representation, PSD and power spectra of discrete signals, like EEG, are related by a simple scaling factor, and for most practical purposes (comparing groups, correlating PSD/power to cognitive measures) they may be used interchangeably. The conceptual basics of the Fourier theorem, and details of the interrelations between time- and frequency-domain signal representations, and between power and PSD, can be found in Supplementary Material A and Fig. A-1.

Raw and relative/normalized PSD

PSD/power estimates for sleep EEG are commonly obtained using Welch's method (or Welch's periodogram; Supplementary Material A, [21]). These estimates are always positive, with many orders of magnitude difference between the slowest and fastest frequencies (Fig. 1A-i). For EEG signals, the relation between frequency f and PSD/power approximates a $1/f^{\,a}$ distribution (where the exponent a indicates how quickly the spectrum drops), also known as power-law scaling.[3] The presence of rhythmic signal components is then expressed as a positive deviation from this background “$1/f$” activity. Note that while every frequency (band) always has non-zero power, this does not imply the presence of oscillatory activity. Conversely, true oscillations (as assessed from time-domain analyses) are not always reflected by a noticeable peak in the PSD/power spectrum, especially if they are relatively infrequent and/or of low amplitude (e.g., sporadic SOs during N2

[1] Counter-intuitively, higher electrode impedance typically leads to larger rather than smaller signal amplitude [15].

[2] High-pass filter cutoffs often employed for waking EEG (0.5 or 1 Hz) tend to filter out most SO activity, so lower cutoffs are needed for (NREM) sleep data (e.g., 0.1 or 0.3 Hz).

[3] Here, “power” refers to the mathematical operation of exponentiation (“raising to a power”), not spectral power.

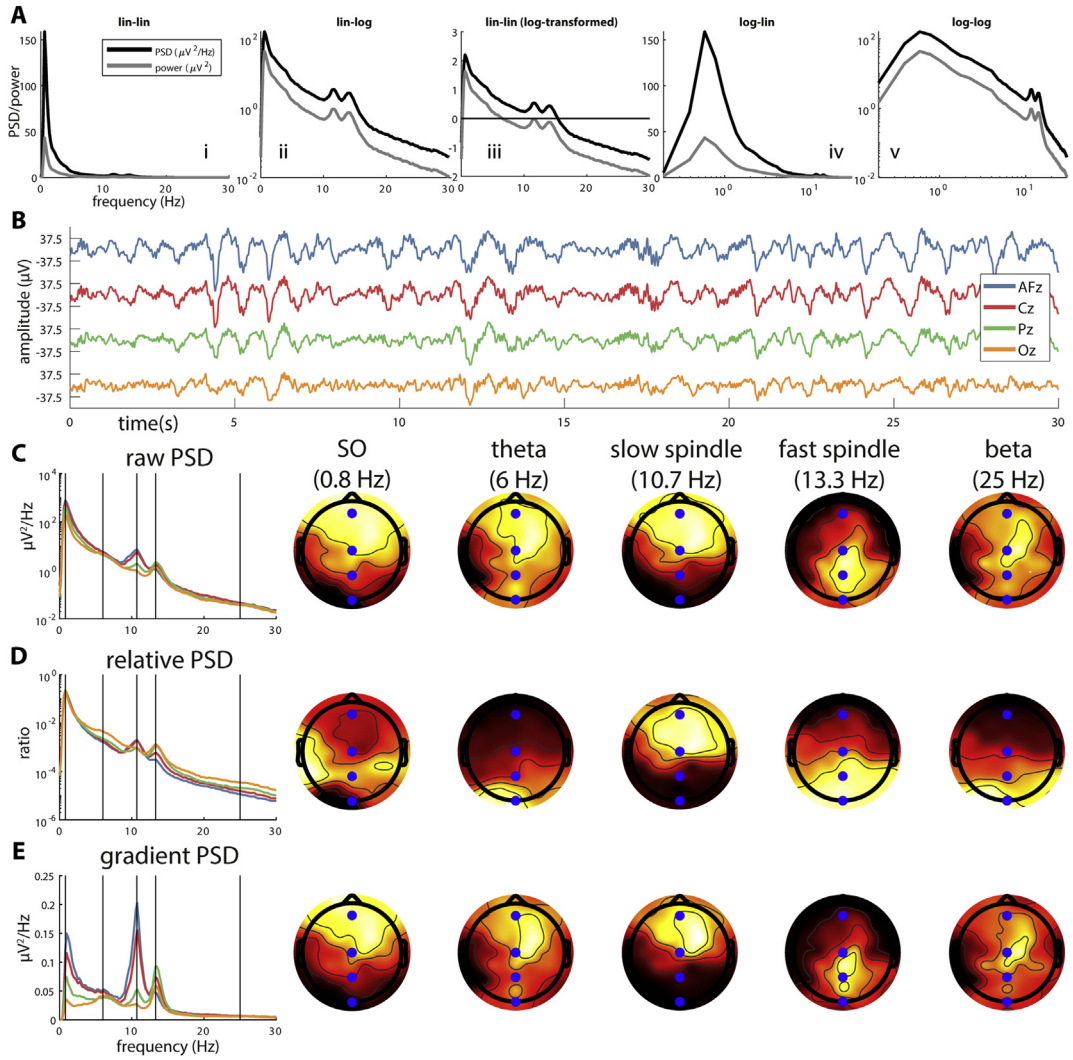

Fig. 1. Power spectral density, power, and normalizations. (A) Different representations of power spectral density (PSD) and power (N3 spectra of channel Cz using Welch's method). Raw PSD and power are related by a scaling factor (or an offset in logarithmic space). Note the virtual absence of power at the lowest frequencies due to high-pass (0.3 Hz) filtering. (B) 30 s excerpt of N3 sleep for four midline channels, showing largest-amplitude slow oscillations (SOs) on frontal channel (blue). (C) Raw PSD spectra for same example channels (left) show differential SO power consistent with (B), as well as channel differences in slow and fast spindle power. Vertical lines indicate frequencies of PSD topographies in right panels (log-transformed for visualization, scaled to data range), showing typical SO, slow spindle, and fast spindle topographies. (D) As C, but for relative power (raw PSD divided by sum of raw 0–30 Hz PSD). (E) As C, but using temporal gradient of time-domain signal for PSD estimation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

sleep).[4] It is also possible that rhythmic activity, when not of sinusoidal shape, is reflected by multiple spectral peaks (harmonics). Moreover, spectral power is affected by various kinds of artifacts (e.g., muscle activity, ocular activity, sweat artifacts). Hence, while spectral shape provides a rough indication of the presence or absence of oscillatory rhythms, it is not conclusive on its own.

Because of 1/*f* scaling, raw PSD/power is usually either plotted on a logarithmic y scale (Fig. 1A-ii), or log-transformed[5] and plotted on a linear scale (Fig. 1A-iii). While the shapes of panels ii and iii are identical, log-transformed PSD/power involves negative values at higher frequencies where untransformed PSD/power <1, which can lead to unintuitive behavior (e.g., greater summed PSD/power for the narrower 0.5–20 Hz range than for the wider 0.5–30 Hz range, due to differential inclusion of negative values). For this reason, untransformed PSD values should ideally be used for further analysis (with the possible exception of statistical testing, for which log or other transformations may be required). When covering a wider frequency range (e.g., intracranial EEG or MEG), it can be helpful to space the x axis logarithmically as well (panels iv and v in Fig. 1A; [22]).

Additional complexities arise when determining "power" within a certain frequency band (e.g., 0.5–4 Hz slow wave activity, or SWA). For continuous signals, power is defined as the integral over the relevant portion of the PSD function. For discretely sampled signals, like EEG, this can be approximated by summing the PSD values from the relevant frequency bins and multiplying by the appropriate bin size (or equivalently, summing power across bins). But because PSD and power are just scaled versions of each other, directly summing PSD values is usually of little consequence. Moreover, many researchers average, rather than sum, values

[4] Other techniques, such as time-frequency approaches via the short-time Fourier transform or wavelets (see section "Instantaneous phase and amplitude"), may be more appropriate for detecting transient oscillations.

[5] While any logarithmic transform is appropriate, base-10 is most often used.

across bins. While it again makes little practical difference whether this is done for PSD or power, summing and averaging yield different patterns when widths of frequency bands are not equal. In addition, some studies sum or average log-transformed instead of untransformed values. These details are not always documented, leaving it unclear whether band-limited activity concerns PSD or power, raw or log-transformed values, and averaged, summed, or integrated values.

Raw PSD/power has a relatively straightforward connection to signal amplitude, with channels expressing larger signal amplitudes typically showing larger power, as shown for raw PSD in Fig. 1 B, C. As such, raw spectra are useful when absolute differences in signal amplitude are deemed meaningful (e.g., topographical analyses). On the other hand, distance to reference affects signal amplitude [23], and individuals often exhibit sizable differences in signal amplitude that are at least partly due to factors of no interest (gyral folding, cap positioning, skull shape and thickness, etc.). For this reason, spectra are often normalized with respect to total power, thereby indicating the relative contribution of each spectral component to the signal. Here, many options exist for defining both the nominator (raw or log-transformed PSD) and denominator (raw or log-transformed PSD, sum or average across frequency bins, considering all bins or a limited range (e.g., 0–30 Hz)). In all cases, however, the denominator is heavily influenced by low-frequency activity (similar to band-limited power), such that normalizing by total power tends to equalize SO activity between channels or individuals. Again, the literature often leaves unspecified exactly how "relative power" has been defined. In addition, band-limited power (e.g., SWA) is sometimes derived from an already normalized spectrum, adding even more options to those from the previous paragraph.

Regardless of the exact definition of relative power, the consequences of power normalization can be significant. Comparing PSD spectral shapes of Fig. 1C, D (left), it becomes apparent that, compared to raw PSD, the use of relative PSD shifts spectra upward or downward. This not only reduces channel differences in the SO band, but also affects which of these example channels shows greatest fast spindle power (raw: Pz; relative: Oz). Considering PSD topographies for several frequencies (Fig. 1C, D, right), relative PSD removes the frontal dominance of 0.8 Hz SO activity, and shifts the hotspot of 13.3 Hz fast spindle activity posteriorly. Strikingly, raw and relative PSD yield entirely different topographies for 6 Hz theta and 25 Hz beta frequencies, with either anterior or posterior scalp regions showing greatest power. Code to reproduce Fig. 1 can be found in *powerDefined.m* and *powerNormalization.m*.

A similar story holds when comparing spectral profiles between N2 and N3 using either raw or relative PSD (Supplementary Material A and Fig. A-2). Hence, where, or in which sleep stage, one considers power to be greatest depends on the chosen normalization. Therefore, normalization should be tailored to the analysis goal. In general, the baseline (denominator) for normalization should not vary with factors of interest. For example, if one wishes to 1) account for individual differences in signal amplitude, but also retain 2) within-subject topographical patterns and 3) within-subject sleep stage differences, one might construct a within-subject normalization factor by averaging total power across channels and sleep stages, and normalize all channel-stage spectra relative to this single baseline.

Due to the 1/*f* distribution, oscillatory components may sometimes manifest as "shoulders" rather than clear spectral peaks. A simple way to enhance peaks is by taking the time-domain signal gradient (as a rough approximation of the temporal derivative), and estimating PSD from this signal [24]. Although this does not affect relative ordering of channel spectra and topographies compared to raw PSD (Fig. 1E), it may facilitate the isolation of spectral components [18]. More sophisticated methods to counteract 1/*f* activity involve fitting this component and subtracting it from the raw PSD [25,26]. However, as of yet no guidelines exist on how to handle different fits across channels, individuals, and sleep stages, while it is all but guaranteed that these choices, too, will affect results.

Finally, amplitude topographies derived from event detectors (e.g., SOs or spindles) typically correspond more closely to raw than relative PSD topographies. This means, for example, that spindle amplitude topographies may be more reflective of broadband power differences than spindle-specific variability, as shown in Supplementary Material B, Fig. B-1, and *powerAndSpindles.m*.

Montage choice

The issue of montage or reference choice has been contentious for almost as long as EEG has been around. Although the basic notion that reference affects signal properties is well understood, it is not always appreciated how far-reaching the consequences may be. Importantly, all montages are instantaneous transformations that do not consider temporal dynamics (i.e., computations are performed independently at each sample). Nonetheless, montage choice has a sizable effect on signals' temporal properties. We consider how three commonly used referencing schemes affect various aspects of the sleep EEG.

Montage affects signal amplitude and polarity

Fig. 2 shows a 30 s excerpt of N3 signal for four midline channels, employing linked mastoids (LM; A), common average (CA; B), and surface Laplacian (SL; C) montages. Briefly, LM can be considered the "default" montage for sleep EEG, with the reference locations over relatively (but certainly not fully) inactive sites. With CA, each electrode is referenced to the average of all electrodes, under the assumption that all potentials should sum to zero. Finally, the SL montage highlights local activity underlying each recording site, while reducing the effects of a common reference and volume conduction (i.e., near-instantaneous transmission of electric fields from neural sources through the brain tissue). For additional background, see Supplementary Material C.

As shown in Fig. 2A, the LM reference shows the typical frontal expression of large-amplitude SOs, which becomes weaker towards posterior channels. At the same time, signal shape and polarity across channels are relatively similar, as when an SO is visible on all channels (first vertical line). In contrast, while the CA reference produces frontal SOs that look similar to those with the LM reference, a polarity inversion occurs between Cz and Pz, resulting in anti-phase relations between anterior and posterior SOs (second and third vertical lines). Moreover, posterior SOs are of much greater amplitude compared to the LM reference, approaching amplitudes of the frontal signal, while Cz SOs are much smaller. This behavior results from the CA computation: Because activity across all electrodes by definition must sum to zero at each sample, large negative amplitudes in frontal areas must be matched by similarly large positive amplitudes elsewhere. In line with its focus on local activity, the SL montage generates more variability between channels, but also generates some mild polarity reversals relative to the LM montage. Code to reproduce Fig. 2 can be found in *referencePolarity.m*. Of note, these montage-dependent variations in relative signal amplitudes have repercussions for PSD topographies, similar to those seen for power normalizations. This is shown in detail in Supplementary Material C, Fig. C-1, and *referencePSDTopography.m*.

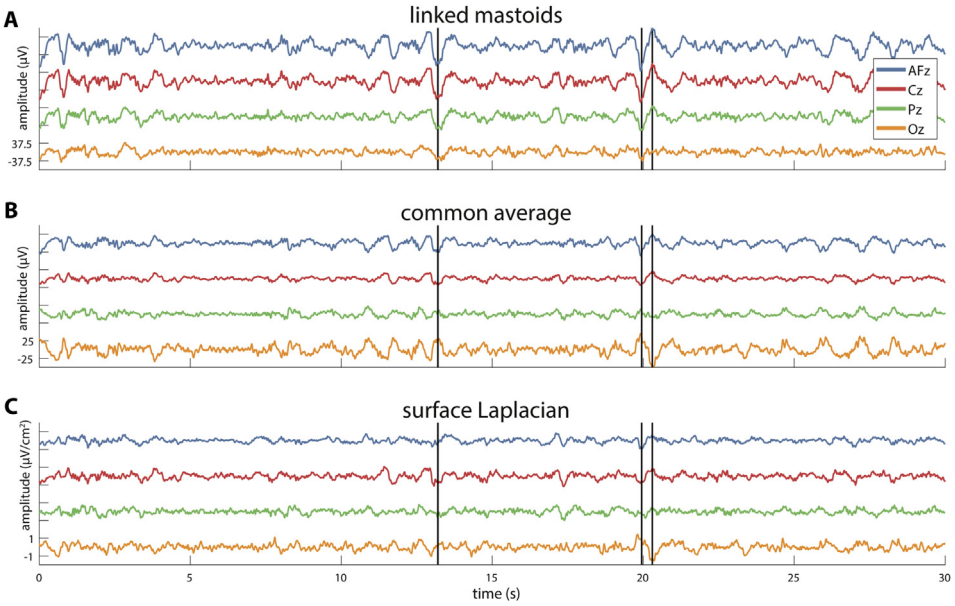

Fig. 2. Comparison of electroencephalography traces with different montages. 30 s of N3 sleep for four midline channels. Vertical lines indicate troughs/peaks of slow oscillations. (A) Common linked mastoids reference. (B) Common average reference. (C) Surface Laplacian estimated from spherical splines.

Montage affects functional connectivity

Reference choice also has a large influence on functional connectivity (see [27] for an overview of metrics). We consider the phase locking value (PLV, [28]) as a measure of phase synchrony. Here, consistent relations of oscillatory phase between brain regions at the same frequency are thought to enable effective communication between the underlying neuronal groups [29]. Topographies of Fig. 3A show PLV for the LM montage between a seed channel and the rest of the scalp for SO (left), slow spindle (middle), and fast spindle activity (right), with the seed channel chosen in correspondence with PSD topographies of Fig. C-1A. At each frequency, connectivity is largest with nearby channels and decreases with distance to seed. Topographies of average phase difference between the seed and the rest of the scalp indicate virtually no phase differences (0°), matching the between-channel similarity of Fig. 2A. In stark contrast, the CA reference yields strong connectivity from the seed to both nearby and distant channels (Fig. 3B), with a pattern roughly following that of the PSD topographies of Fig. C-1B. Moreover, connectivity to distant regions shows a prominent anti-phase relation at each frequency, consistent with the polarity reversals described earlier. Finally, SL maps

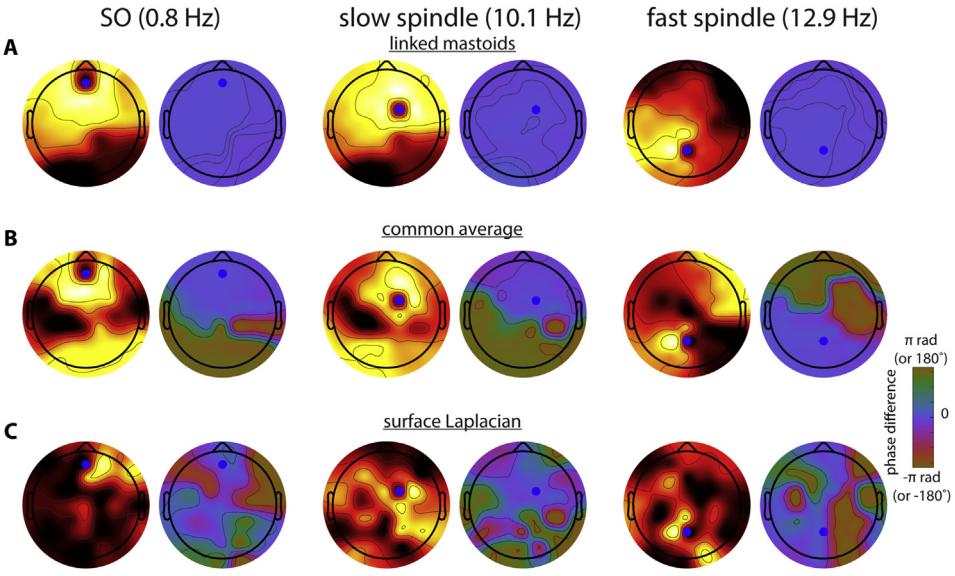

Fig. 3. Effect of montage on phase synchrony. Phase-locking values (left in each topography pair) and phase differences (right in each pair) between seed electrode (indicated in blue) and the rest of the scalp for slow oscillation, slow spindle, and fast spindle frequencies. Phase-locking values calculated across 30 s segments for a total of 10 min of N3 sleep. Self-connectivity (from the seed to itself) has been set to the minimum across the scalp for visualization purposes. (A) Common linked mastoids reference. (B) Common average reference. (C) Surface Laplacian estimated from spherical splines. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

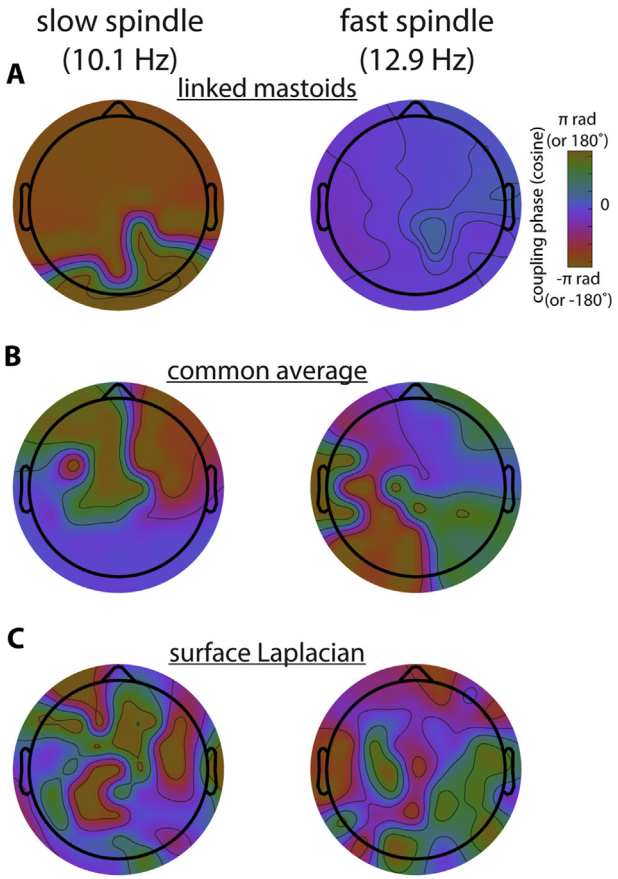

Fig. 4. Effect of montage on slow oscillation-spindle coupling phase. Slow oscillation (SO) phase at which slow spindle (left) and fast spindle (right) activity is maximal, calculated using the debiased phase-amplitude coupling method across 30 s segments for a total of 10 min of N3 sleep. Zero degrees reflects SO peak. (A) Common linked mastoids reference. (B) Common average reference. (C) Surface Laplacian estimated from spherical splines.

suggest much less widespread connectivity, with noisier (and arguably more realistic) phase difference maps.

Although here we considered the PLV metric as it provides an intuitive link to the results from the previous paragraphs, this metric should generally not be used for scalp-level data due to its sensitivity to volume conduction. However, we have observed similarly large montage effects for a variety of phase- and amplitude-based metrics. For example, amplitude envelope correlations [30] for the CA reference show similar long-range connectivity between regions of opposite polarity SOs as PLV.

In sum, while LM-based functional connectivity is largely due to volume conduction (unless this effect is suppressed by employing specialized metrics), the CA montage tends to produce long-range connectivity that is likely wholly due to its requirement that potentials sum to zero. SL-based connectivity looks different yet again, arguably yielding a more plausible pattern. Although theoretical considerations and simulations [31,32] support this argument, caution remains warranted. Code to reproduce Fig. 3 can be found in *referencePhaseSynchrony.m*.

Montage affects cross-frequency coupling

Finally, reference choice affects cross-frequency phase-amplitude coupling (PAC), whereby the amplitude (or power) of a faster oscillation depends on the phase of a slower oscillation. This is thought to enable brain communication across multiple spatiotemporal scales [33]. Fig. 4 shows topographies of the SO phase at which slow (left) and fast spindles (right) are maximally expressed, calculated using the debiased phase-amplitude coupling (dPAC) method [34]. The LM montage maps have very uniform appearances (Fig. 4A), with slow spindles peaking in the SO trough around 180°, and fast spindles in the SO peak around 0°. (The closely spaced color transitions in the slow spindle map occur due to phases transitioning between 180° and −180°.) This result is again consistent with volume conduction ensuring that SOs and spindles, while having maximal amplitudes and power at different electrodes (Figs. 3A and C-1A), are picked up across the scalp. In contrast, the CA montage yields roughly bimodal maps, such that spindle activity in anterior and posterior regions is tied to opposite SO phases, reflecting the polarity reversals intrinsic to this reference choice (Fig. 4B). Finally, SL phase maps appear noisy (Fig. 4C), although this may be expected given that the Laplacian yields relatively focal hotspots of SO and/or spindle activity (Fig. C-1C), making it difficult to extract meaningful phase coupling estimates from all channels. Indeed, the degree (rather than phase) of SO-spindle coupling (or coupling strength) also depends on montage. Generally, one should limit interpretation of coupling phase to channels actually showing such coupling. However, raw coupling strength values are often difficult to interpret, as we will discuss in the section *Surrogate testing*. Code to reproduce Fig. 4 can be found in *referencePhaseAmplitudeCoupling.m*.

Instantaneous phase and amplitude

Accurately extracting instantaneous phase and amplitude information is crucial to answer fundamental and clinical sleep research questions related to functional connectivity and cross-frequency interactions. For instance, previous sleep studies found diminished spindle coherence in schizophrenia [9], altered alpha-band phase synchrony in posttraumatic stress disorder [11], and less precise SO-spindle PAC associated with age-related memory decline [4,5]. While the preceding section already considered phase synchrony and PAC in the context of montage choice, calculating such metrics involves many considerations [35], some of which (more) particular to sleep EEG. Depending on the employed metric, band-limited phase and/or amplitude information needs to be extracted, either for a specific frequency band or a range of frequency bands. For many metrics, these calculations can be performed equivalently in the time or frequency domain. Whereas frequency-domain calculations rely on signals' full PSD and cross-spectral density, time-domain approaches require extraction of sample-wise (or instantaneous) phase and amplitude information. Two popular ways of obtaining these instantaneous estimates are 1) band-pass filtering the EEG in the frequency band(s) of interest and applying the Hilbert transform,[6] and 2) convolving[7] the EEG with a set of complex-valued wavelets of different frequencies. Both techniques provide complex-valued output, from which phase (as the angle) and amplitude (as the magnitude) estimates can be obtained. However, there are numerous ways in which the returned estimates may be inaccurate. In some cases, where phase/amplitude estimates are an intermediate step in the calculation of a metric, this will be of no consequence. For example, systematically

[6] The Hilbert transform adds a phase-shifted copy of the original signal as an imaginary component, creating the complex-valued "analytic signal".

[7] Convolution with a complex wavelet involves sliding the wavelet across the data and multiplying the two signals at every sample, thereby giving a time-resolved indication of how well the data matches the wavelet's frequency.

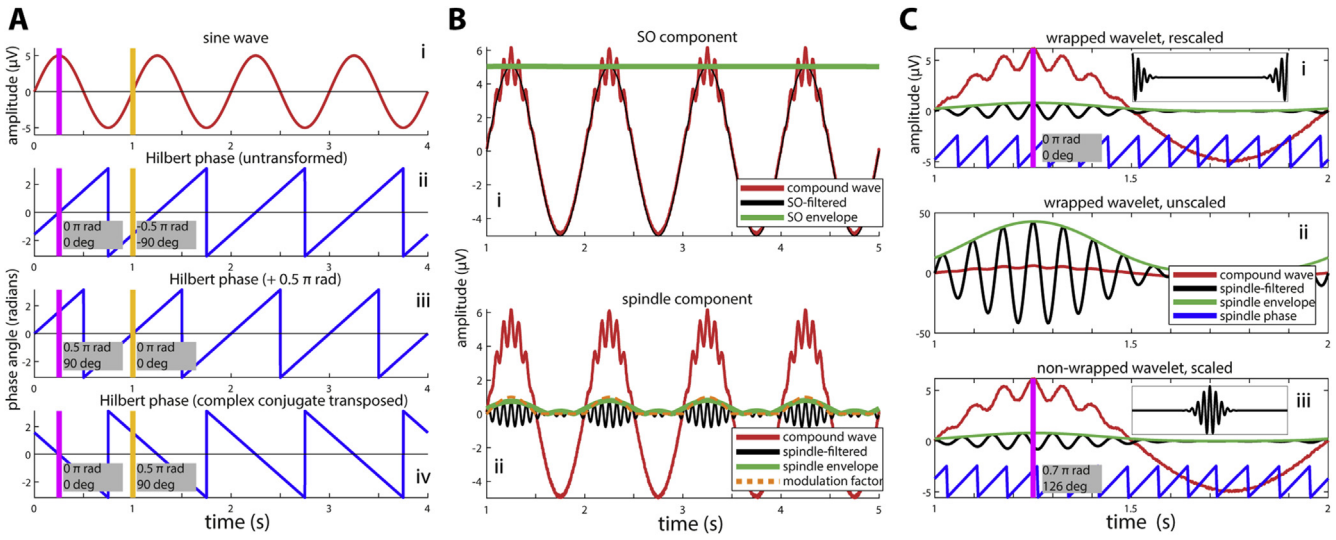

Fig. 5. Extracting phase and amplitude. (A) Sine wave and extracting phase estimates using different approaches. (B) Artificial slow oscillation-spindle coupling and extracting amplitudes with the filter-Hilbert approach. (C) Artificial slow oscillation-spindle coupling and extracting amplitudes with wavelets.

shifted phase estimates do not impact the calculation of between-channel phase differences, and thereby various metrics of phase synchrony. In other cases, accurate phase/amplitude estimates are critically important, as when assessing the phase of SO-spindle coupling, or when amplitude estimates are fed to downstream detection algorithms. We now discuss several issues related to phase/amplitude estimation.

Extracting phase

Phase indicates the relative position in a wave's cycle, expressed as an angle in radians or degrees. Although the assignment of phase angle to position is ultimately arbitrary, it is commonly defined with respect to a cosine or sine reference wave. When extracting phase from the complex-valued results of the (correctly applied) filter-Hilbert or wavelet convolution techniques, it is implicitly defined with respect to a cosine, such that peaks correspond to 0 radians (or 0°) and positive zero-crossings to $-0.5\pi/1.5\pi$ radians (or −90/270°) (Fig. 5A, i and ii). (This was also the convention used in Fig. 4). If sine-relative phase angles are desired, 0.5π radians have to be added manually (Fig. 5A, iii). However, it is not always evident from the literature which convention was adopted. A further issue arises in Matlab when the transpose operator (') is applied to complex values: As this operator formally performs a complex conjugate transpose, imaginary values are sign flipped and phases progress in the reverse direction, leading to faulty phase estimates (Fig. 5A, iv). Instead, the nonconjugate transpose (.') should be used. Finally, it is important that statistical properties (e.g., mean, standard deviation, correlation) of angular variables are calculated using dedicated circular tools (e.g., [36]). Code to reproduce Fig. 5A can be found in *extractPhase.m*.

Extracting Hilbert amplitude

Fig. 5B shows simulated SO-spindle phase-amplitude coupling, with spindle amplitudes maximal in the SO peak (red traces). Once data have been filtered in the SO (Fig. 5B, i) and spindle (Fig. 5B, ii) ranges, and the Hilbert transform has been applied, the filtered signal and its amplitude envelope can be obtained as the real part (black) and magnitude (green) of the Hilbert-transformed data, respectively. However, filtering necessarily introduces both spectral and temporal imprecisions,[8] as evident from a) apparent spindle activity outside the windows actually containing such activity (orange), and b) a somewhat reduced spindle envelope relative to its specified amplitude of 1. Code to reproduce Fig. 5B can be found in *extractHilbertAmplitude.m*.

Extracting wavelet amplitude and phase

Although extracting amplitude/envelope information from wavelets is conceptually similar to the Hilbert approach, the wavelets require rescaling in the frequency domain to obtain values in line with the original data (Fig. 5C, i and ii). Moreover, when wavelet convolution is performed in the frequency domain, phase estimates will be incorrect (phase-shifted) if the time-domain wavelet is centered on its maximum as it typically is (Fig. 5C, iii). The precise offset depends on wavelet properties, including frequency and wavelet length. Instead, the wavelet should be "wrapped around" such that its two time-domain halves are essentially swapped (Fig. 5C, i, inset) [22]. Code explaining these issues in detail can be found in *extractWaveletPhaseAmplitude.m*. As mentioned earlier, whether these concerns affect analyses depends on how exactly phase/amplitude estimates will be used in downstream analyses.

Creating wavelets

The discussion so far has ignored the important question of how to create wavelets with appropriate properties for sleep EEG. Although many types of wavelets exist, here we only consider (complex) Morlet wavelets, which are (complex) sine waves tapered by a Gaussian window. Besides their frequency, they are defined by many parameters (length, number, frequency spacing, temporal resolution, frequency resolution, time-frequency trade-off) that require careful consideration in general, and different considerations for sleep EEG. In particular, wavelet frequencies should be chosen to match the oscillatory components of interest, which may require both very slow wavelets to capture SOs, and

[8] Designing adequate filters is important, but beyond the scope of this article.

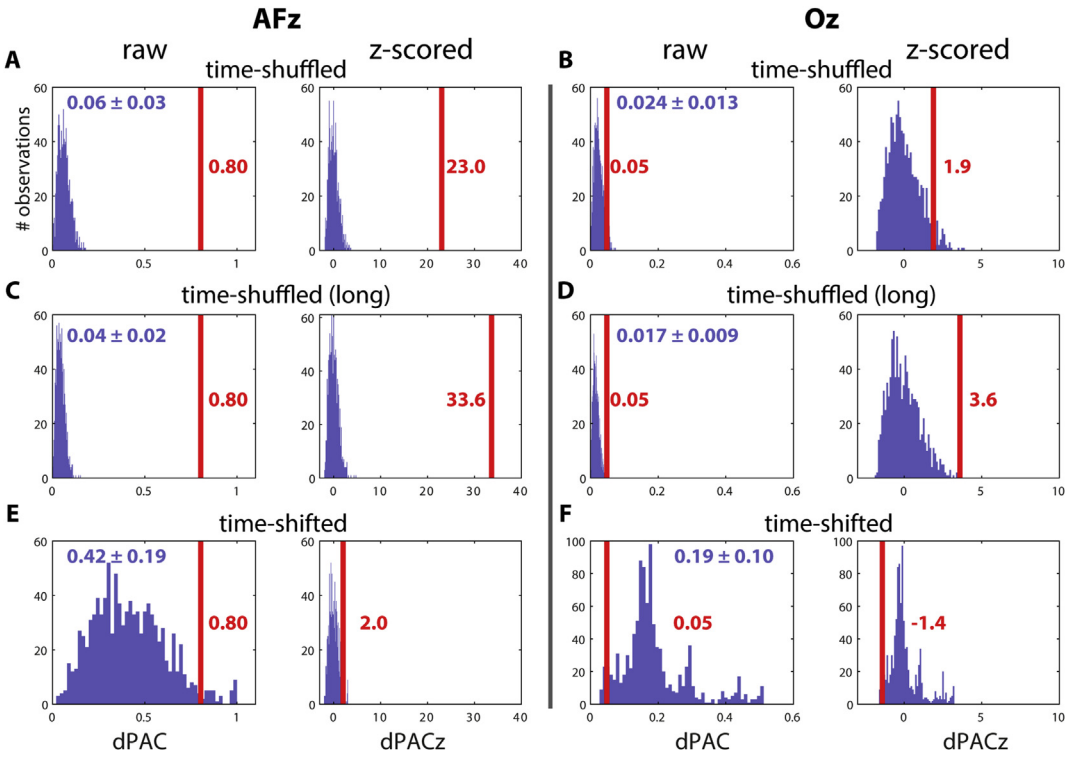

Fig. 6. Effect of surrogate choice on slow oscillation-slow spindle coupling for different channels. All panels based on data from 30 s of N3, wavelet-derived 0.8 Hz phase information, 10.1 Hz power information, and 1000 surrogates. Vertical lines indicate observed coupling strength value before (left) and after (right) z-scoring. Magenta indicates surrogate distributions. Left panels show coupling for frontal channel AFz using time-shuffled phases (A), time-shuffled phases after doubling data length (C), and time-shifted phases (E). Right panels (BDF) show the same, but for posterior channel Oz.

properly spaced wavelets with relatively high spectral precision to discern slow and fast spindles. Moreover, relatively long (e.g., 20 s) wavelets ensure that a Gaussian shape is maintained in the frequency domain at even the slowest frequencies. Hence, defaults used in various toolboxes are often not optimized for sleep EEG. Code to flexibly construct and visualize Morlet wavelets can be found in *createWavelets.m*. We also include code for wavelet creation with properties that we deem reasonable for sleep scalp EEG (*createWavelets_sleepScalpEEG.m*).

Surrogate testing

After calculating metrics of functional connectivity or cross-frequency coupling, raw values are often compared between channels, frequencies, groups, etc. However, raw values are not always straightforward to interpret because it is not known what the expected values are in the absence of connectivity/coupling. While for many metrics theoretical values of zero indicate no connectivity/coupling, values of exactly zero are hardly ever observed and it is unclear how far values should deviate from zero (0.1, 0.01, 0.001…?) to signal meaningful connectivity/coupling. Moreover, expected values in the absence of a true effect often depend on factors like power, frequency, channel, and individual, which should be taken into account if fair comparisons are to be made. Finally, it is often helpful to establish that obtained values differ from noise in the first place, before any comparisons are made. These issues can often be resolved through surrogate testing techniques, which are particularly suited for fundamental and translational sleep EEG studies with their often large amounts of data.

Surrogate construction

The usual approach for testing the presence of connectivity/coupling is to construct a large set of surrogates: altered versions of original data that keep certain properties intact. The metric of interest is recalculated for every surrogate, and the observed value from the actual data is compared to the distribution of surrogate values. When the observed value is sufficiently unlikely given the surrogate distribution (also called null distribution as it reflects the null hypothesis of no effect), connectivity/coupling may be inferred. However, many considerations factor into surrogate testing, some with large effects.

There are different ways to contrast the observed value to the surrogate distribution. First, one may simply evaluate the proportion of surrogate values greater or equal than the observed value, providing a non-parametric one-sided[9] p-value. In the example of Fig. 6A, observed PAC at frontal channel AFz between SO phase and slow spindle amplitude (0.80) is greater than each of 1000 surrogate values ($P < 0.001$). Second, provided the surrogate values are approximately normally distributed, the distribution's mean and standard deviation (0.06 ± 0.03) can be used to z-score the observed value (Fig. 6A, right), here yielding 23.0. As each z-score has an associated p-value, this approach offers another (parametric) way to determine significance (here, $P<<10^{-16}$; note the large difference with the non-parametric approach, which is limited by the number of surrogates). Moreover, as z-scoring indicates how far, in terms of standard deviations, a value is from its

[9] For metrics that go in both directions, like amplitude envelope correlations, slightly more involved two-sided testing is needed.

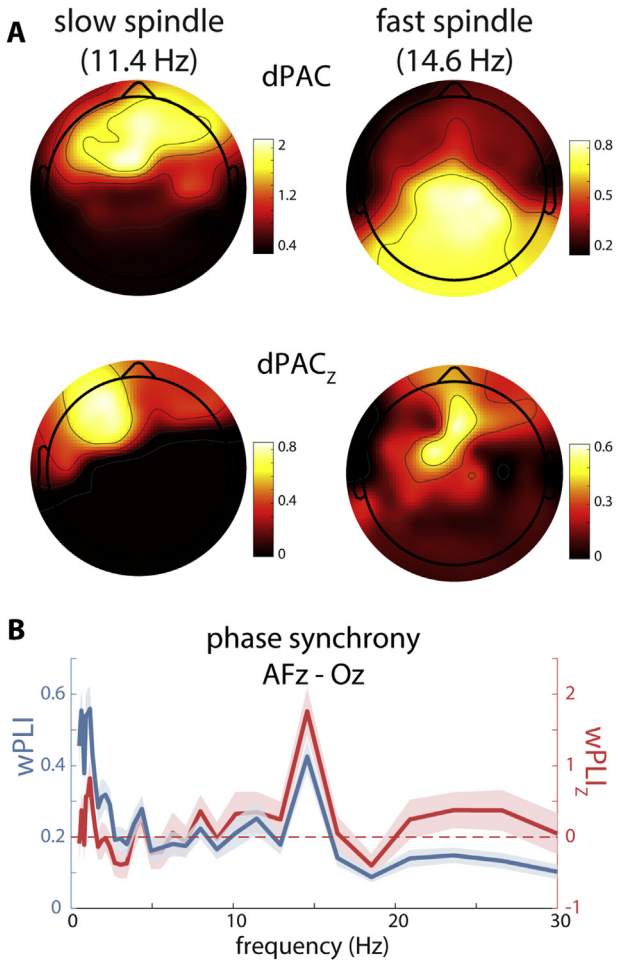

Fig. 7. Effect of surrogate-based normalization on coupling dynamics. (A) Coupling strength (debiased phase-amplitude coupling (dPAC) metric) between slow oscillations and slow (left) and fast spindles (right), and for raw (top) and surrogate-based z-scored values (bottom). (B) Phase synchrony (weighted phase lag index (wPLI)) between AFz and Oz, for raw and surrogate-based z-scored values. Metrics calculated per 30 s segment, using 200 time-shifted surrogates, for a total of 10 min of N3 sleep. Error band in (B): standard error of the mean across segments.

mean, it offers normalized scores that are more easily compared when surrogate distributions vary in these respects. This is evident when considering posterior channel Oz (Fig. 6B). While raw PAC is reduced considerably (0.05) compared to AFz, the raw surrogate distribution is also shifted, still yielding a modest positive z-score of 1.9.

Importantly, surrogate choice has a very large effect on results. Fig. 6A, B makes use of an often applied, but non-recommended, time-shuffling approach, where on each iteration the observed SO phase estimates are fully shuffled (while leaving spindle amplitude intact), before PAC is recalculated[10]. (Alternatively, one could shuffle amplitudes while leaving phases intact, and similar shuffling approaches can be devised for other metrics as applicable.) Although the positive z-scores of Fig. 6A, B seem highly indicative of coupled SO-spindle activity, this is misleading for several reasons. First, the time-shuffling approach is very sensitive to data length. This is shown in Fig. 6C, D, where the 30 s of data underlying Fig. 6A, B was concatenated to itself, doubling the amount of data while leaving intrinsic coupling intact. Although observed PAC values are left unchanged, repeated phase shuffling results in much higher z-scores than before (33.6 vs. 23.0; 3.6 vs. 1.9). Hence, arbitrarily large z-scores may be obtained by including more data. Second, and more fundamentally, shuffling phases (or amplitudes) destroys many properties of the original signal besides those relevant for the tested metric. In essence, the time-shuffling approach merely assesses whether a signal is oscillatory in nature, and it should usually be avoided [37,38].

A more appropriate time-shifting approach, where on each iteration the SO phase time series is shifted by a random amount with respect to the spindle amplitude time series, yields much wider surrogate distributions, and correspondingly, much smaller z-scores (Fig. 6E, F). Critically, while frontal SO-spindle coupling remains present with this approach (z-score: 2.0), posterior coupling is completely abolished (z-score: −1.4).[11] Moreover, doubling data length does not lead to noticeable changes (not shown). Still, it is often computationally convenient to calculate raw and surrogate-normalized metrics on shorter segments of fixed length (e.g., 30 or 60 s), which can be averaged and additionally used to estimate variability. Of course, numerous other surrogate approaches may be devised, each with particular strengths and weaknesses in relation to analysis goals.

Surrogates affect coupling dynamics

Fig. 7A shows topographies of SO-spindle phase-amplitude coupling strength (dPAC metric), for slow and fast spindles, and for raw and time-shifted surrogate-normalized metrics. Whereas raw and normalized coupling strengths show similar frontal topographies for slow spindles, fast spindles appear most strongly modulated by the SO phase over either parieto-occipital (raw) or frontal (normalized) regions. Which is considered the more appropriate depiction depends on how one conceptualizes coupling strength. The raw dPAC metric is a complex function of 1) the distribution of spindle envelope values (reflecting both the presence and amplitude of spindles), 2) the distribution and accuracy of SO phase estimates (partly depending on the presence and amplitude of SOs), and 3) the association between the two. As such, a channel difference in raw coupling strengths could partially reflect differences in SO/spindle density/amplitude. In contrast, surrogate-normalized dPAC indicates the degree of coupling given the distributions of spindle envelopes and SO phases present. While both are valid representations of phase-amplitude coupling, they may lead to substantially different ideas of where strongest coupling appears.

Fig. 7B shows long-range AFz-Oz phase synchrony across frequencies when considering raw or normalized weighted phase lag index values (wPLI; a measure that suppresses the effects of volume conduction; [39]). Raw wPLI (blue) shows highest connectivity values (~0.55) in the SO band, and another peak (~0.4) in the spindle range, which might lead to the conclusion that SOs are more synchronized than spindles.[12] In contrast, time-shifted surrogate-normalized wPLI$_Z$ (red) indicate strongest connectivity in the spindle range (~1.75) with a much smaller effect for SOs (~0.75). Moreover, whereas raw connectivity for other frequencies (values between ~0.1 and ~0.2), might suggest meaningful oscillatory interactions, normalized scores hover mostly around zero, indicating no greater synchrony than expected by chance. We stress that there

[10] We have previously used this approach inappropriately ourselves [19].

[11] In fact, the negative z-scores suggest a form of "decoupling", such that parts of the signal where the spindle envelope fluctuates at the SO frequency are temporally separated from parts where SOs are present. Time-shifting will tend to align these signal components, yielding higher coupling for surrogates.

[12] In general, raw connectivity values tend to decrease with frequency [40], somewhat reminiscent of 1/*f* PSD scaling.

are numerous situations where comparing raw metrics is appropriate (e.g., contrasting connectivity between within-subject conditions with similar spectral power). However, careful thought and consideration is required to ensure that the employed metric matches the particular research goal.

Minimizing false positives

While many statistical considerations are worth discussing [41], false positives pose a particular threat to the validity of sleep studies. The issue begins with the vast number of potentially interesting outcome variables that may be extracted from sleep EEG. For example, our spindle detector routinely provides channel-by-channel information on eight primary spindle properties, further separated by (or pooled across) sleep stage and spindle type (slow, fast), yielding a total of 4320 unique spindle variables for 60-channel data. This number may be arbitrarily increased by considering these properties by, for example, sleep cycle, descent to vs. ascent from N3, or circadian time [42]. Similar multiplicative paths exist for SO characteristics, PSD, time-frequency power, cross-frequency coupling, functional connectivity, and so on. This is exacerbated further by the flexibility in upstream analysis choices, as discussed in the preceding sections on e.g., power normalization and wavelet creation. Similarly, event detection parameters are highly flexible (e.g., varying lower and upper frequency, amplitude, and duration thresholds for spindle detection), altogether making the number of unique outcome measures truly staggering.

Here, the multiple comparisons problem comes into full effect. Considering traditional null hypothesis significance testing,[13] the null hypothesis in most sleep EEG studies states that there is no group difference in EEG measures, or no association between electrophysiology and cognitive measures. An often-neglected point is that, under this null hypothesis, p-values exhibit a uniform distribution, such that values between 0 and 0.05 are as likely as values between e.g., 0.95 and 1 [43]. Hence, considering twenty independent statistical tests or random data sets is (on average) sufficient to obtain a significant p-value at a conventional alpha level of 0.05.[14] Of course, outcome variables in sleep EEG are often highly correlated between nearby channels and frequencies, between different normalizations, or between different metrics (e.g., different indicators of SO activity [44]). However, this reduction in the number of independent outcome measures is easily offset by the number of measures that are routinely available for testing. Stated differently, continued statistical testing makes it exceedingly improbable to *not* yield a p-value below 0.05, and the researcher who does not observe any "significant" effects is highly unlucky.[15]

Unfortunately, many sleep EEG studies ignore this problem in one of several ways. First, there are instances where a number of equivalently treated outcome measures are compared between groups without any form of multiple comparison correction, and the measures attaining traditional statistical significance are interpreted as reflecting a population effect. Second, it is not uncommon to see significant effects reported exclusively for highly specific outcome metrics. As a fictitious example, consider a correlation between N2 slow spindle duration at channel P6 in the third sleep cycle, and overnight reaction time differences to one of four stimulus categories for one of three behavioral paradigms. While in some instances a highly specific hypothesis may indeed have been held prior to data analysis, such results are often suggestive of selective reporting, whereby many more analyses were performed than described. As a side note, whenever experimental specificity is claimed (e.g., an effect in N2 but not N3, or for slow but not fast spindles), this should be supported by a direct contrast between these effects rather than only observing significance in one condition but not the other [45]. Third, and perhaps most insidious, is the aforementioned possibility to repeat analyses using a multitude of different parameter settings until significance is reached, while reporting results as if only the final analysis pipeline has ever been used. For example, numerous PSD normalization options may be tried until PSD topographies look "correct" (e.g., in relation to a pre-sleep task hypothesized to engage specific regions), or spindle detection thresholds may be adjusted to maximize group differences.[16] On the other hand, there are good reasons for assessing the effects of analysis parameters, as will be discussed in the next paragraph.

While one should guard against ignoring results from otherwise sensible analysis tracks simply because they do not yield desired results, strategically varying analysis parameters allows checking the robustness of the analysis pipeline, and ensures that effects are not simply due to a highly specific set of parameter values. Similarly, checking that a non-sensible parameter choice removes a previously present effect may instill confidence in the correctness of the analysis procedure. Moreover, an important message from the preceding sections is that developing any analysis pipeline requires a considerable back-and-forth to ensure that outcome measures actually capture the phenomenon of interest. A potential strategy to balance these opposing forces is to postpone group-level analyses until all aspects of the analysis pipeline have been deemed satisfactory in one or two pilot subjects. That said, sometimes problems only become apparent when considering the full data set. Weighing these issues against each other is difficult, and it is up to individual researchers to strike a reasonable compromise.

More generally, it is difficult to offer clear guidelines on how to manage false positives. The best way is to limit the number of analyses by having clear hypotheses. But even then, hypotheses are often rather loose and conceptual in nature (e.g., coordinated SO-spindle "activity" predicts memory performance), still offering a large search space of related analysis procedures and metrics, with some combination invariably resulting in significance by chance. As per our previous suggestion, this problem may be avoided by having the complete analysis pipeline ready prior to analysis (and ideally even pre-registering it). In practice, however, the large investment of time and resources needed to acquire sleep EEG data offers an understandable incentive to continue analyses until there are "results". We see no problem with reusing datasets to tackle different research questions, or to perform exploratory analyses with limited or even no statistical control. Rather, it is reporting exploratory results *as if* they were confirmatory, and thereby misrepresenting the degree of evidence for an effect, that should be avoided.

It is outside this article's scope to cover the strengths and weaknesses of different approaches for multiple comparisons correction, of which there are many (e.g., Bonferroni, False

[13] Analogous concerns exist when considering multivariate approaches.

[14] While using arbitrary cutoffs (0.05 or otherwise) to make binary decisions as to whether effects exist is not ideal, the reality of running dozens to millions (e.g., time-frequency-channel data) of tests in a given study often makes this simplification unavoidable.

[15] It is sometimes argued that uncorrected significance *could* still be reflective of a population effect, suggesting no need for multiple comparison correction. While this is strictly speaking true, the point is that such significance becomes meaningless when the same number of below-threshold p-values are expected when analyzing (similarly correlated) random data (again acknowledging that metrics from real data are not independent).

[16] Of course, one can envision situations where such approaches are appropriate because they are independent of the statistical tests to be performed.

Discovery Rate [46], cluster-based approaches [47]). However, it is important to consider that the control of false positives typically also increases false negatives (i.e., non-detection of true effects). Most fundamentally, different statistical philosophies exist as to what exactly constitutes the family of tests that should undergo statistical correction [48]. Moreover, none of these recommendations are tailored to the practical realities of sleep EEG data, with it sometimes millions of tests that can be hierarchically organized by metric, channel, time, frequency, condition, and so on. Importantly, while deciding which tests should be combined for correction is a subjective choice, it should be made prior to statistical testing to avoid choosing whatever grouping maximizes post-correction significance. Again, it is up to individual researchers to balance all these interacting factors, to determine the strength of evidence for results, and to accurately convey their confidence in these findings to the broader community.

Discussion

We have reviewed various methodological concerns in the analysis of human sleep EEG, particularly in relation to spectral analyses, montage choice, extraction of phase/amplitude information, surrogate construction, and false positives. While it is hoped that the issues raised will be of practical use, they are intended to illustrate some broader points, which extend to analytic approaches not covered here.

First, as illustrated by the sections on PSD normalizations, montage choice, and surrogate-normalized vs. raw values, factors that are often deemed conceptually uninteresting can have a massive impact on results and conclusions. Because of this, it is critical to report methodology in sufficient detail for others repeat the analyses. In addition, while it is clearly unrealistic to examine the full effect of each potentially relevant factor and their interactions on all outcome measures, it is worth considering to what extent central study findings depend on the chosen analysis approach and metrics. In our view, observing similar fronto-parietal spindle phase synchrony with two metrics and two reference schemes provides fundamentally different support for this phenomenon compared to a situation where it is evident for only one of these four scenarios.

Second, as indicated by the discussions on phase/amplitude extraction and surrogate choice, small calculation details can cause resulting metrics to provide poor or meaningless estimates of the phenomenon of interest (e.g., oscillatory phase through non-wrapped wavelets, or indications of coupling through time-shuffled surrogates). Even with theoretical knowledge of these matters, such issues are difficult to diagnose based on code review or inspection of raw values alone. While such practices are certainly important, we believe visualizing output of intermediate processing steps is critical to ensure code operates as intended. Often, this can be done by applying code to simple examples, either a small segment of single-channel real data (e.g., Fig. 6), or simulated data that offers a ground truth (e.g., Fig. 5). Similarly, developing automated detection algorithms without visualization makes it impossible to assess whether and how well detected events correspond to the EEG elements of interest (e.g., spindles, SOs).

Finally, we relate our statistical considerations to some reflections on the sleep and cognition field more broadly, where emerging work suggests poor replicability (for some noteworthy null results, see [49–58]). While non-replication can stem from a host of factors related to both the original and replication studies, it is our impression that many published studies are of insufficient methodological quality and will turn out to be false positives.[17] We suggest that part of the reason for poor practices is that for many study designs, *any* group difference or *any* correlation, in any direction, and for any metric, frequency band, channel and so forth, is considered interesting. Indeed, the highly variable literature facilitates embedding almost any finding in a well-referenced and compelling narrative, offering an incentive to interpret and publish positive results that would not withstand proper statistical control. While negative findings are gradually beginning to take their rightful place in the broader clinical, psychology, and neuroscience literatures, the sleep field appears relatively slow to adapt and less willing to embrace null results. However, a deeper understanding of the sleeping brain requires both uncovering exciting potential new paths, and determining and acknowledging when these paths are likely to be dead ends.

To conclude, we have highlighted several methodological issues in the analysis of sleep EEG that can alter or even invalidate study conclusions. Relating sleep EEG to cognition and disease is important, but it is also difficult. While progress has certainly been made, methodological quality can and should be improved if a genuine understanding of what sleep does, and does not do, is desired. We hope that the current paper makes a small contribution towards that goal.

Practice points

- Consider that signal processing choices may have large effects on results, affecting interpretation of fundamental and clinical sleep EEG studies.
- Evaluate effects of analysis choices to determine robustness of results.
- Inspect and visualize intermediate results, not only when outcomes suggest errors, but also when such indications are absent.
- Consider how many related analyses have been run, how false positives can be mitigated, and where results fall on the exploration-confirmation spectrum.

Research agenda

- Formulate precise hypotheses when possible, including the sleep EEG metrics used as (clinical) outcome measures, and their exact calculation.
- Complete analysis pipeline, optionally with preregistration, before application to the full dataset.
- Conduct large-sample, replication, and meta-analytic studies to determine evidence for previously reported effects.

Conflicts of interest

The authors do not have any conflicts of interest to disclose.

[17] Of course, false positives are expected even with the most rigorous methodology.

Acknowledgements

This work was supported by the German Research Foundation, Germany (FE366/9-1 to JF). We thank Thorsten Rings for valuable input.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.smrv.2020.101353.

References

[1] Manoach DS, Stickgold R. Abnormal sleep spindles, memory consolidation, and schizophrenia. Annu Rev Clin Psychol 2019;29.
[2] Rasch B, Born J. About sleep's role in memory. Physiol Rev 2013;93:681–766. https://doi.org/10.1152/physrev.00032.2012.
[3] Stickgold R, Walker MP. Sleep-dependent memory triage: evolving generalization through selective processing. Nat Neurosci 2013;16:139–45. https://doi.org/10.1038/nn.3303.
*[4] Muehlroth BE, Sander MC, Fandakova Y, Grandy TH, Rasch B, Shing YL, et al. Precise slow oscillation–spindle coupling promotes memory consolidation in younger and older adults. Sci Rep 2019;9. https://doi.org/10.1038/s41598-018-36557-z.
[5] Helfrich RF, Mander BA, Jagust WJ, Knight RT, Walker MP. Old brains come uncoupled in sleep: slow wave-spindle synchrony, brain atrophy, and forgetting. Neuron 2018;97:221–30. https://doi.org/10.1016/j.neuron.2017.11.020. e4.
[6] Mander BA, Marks SM, Vogel JW, Rao V, Lu B, Saletin JM, et al. β-amyloid disrupts human NREM slow waves and related hippocampus-dependent memory consolidation. Nat Neurosci 2015;18:1051–7. https://doi.org/10.1038/nn.4035.
[7] Demanuele C, Bartsch U, Baran B, Khan S, Vangel MG, Cox R, et al. Coordination of slow waves with sleep spindles predicts sleep-dependent memory consolidation in schizophrenia. Sleep 2017;40. https://doi.org/10.1093/sleep/zsw013.
[8] Ferrarelli F, Huber R, Peterson MJ, Massimini M, Murphy M, Riedner BA, et al. Reduced sleep spindle activity in schizophrenia patients. Am J Psychiatry 2007;164:483–92. https://doi.org/10.1176/ajp.2007.164.3.483.
[9] Wamsley EJ, Tucker MA, Shinn AK, Ono KE, McKinley SK, Ely AV, et al. Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation? Biol Psychiatr 2012;71:154–61. https://doi.org/10.1016/j.biopsych.2011.08.008.
[10] Wang C, Ramakrishnan S, Laxminarayan S, Dovzhenok A, Cashmere JD, Germain A, et al. An attempt to identify reproducible high-density EEG markers of PTSD during sleep. Sleep 2020;43. https://doi.org/10.1093/sleep/zsz207.
[11] Laxminarayan S, Wang C, Ramakrishnan S, Oyama T, Cashmere JD, Germain A, et al. Alterations in sleep electroencephalography synchrony in combat-exposed veterans with post-traumatic stress disorder. Sleep 2020. https://doi.org/10.1093/sleep/zsaa006.
[12] Limoges É, Mottron L, Bolduc C, Berthiaume C, Godbout R. Atypical sleep architecture and the autism phenotype. Brain 2005;128:1049–61. https://doi.org/10.1093/brain/awh425.
[13] Lambert A, Tessier S, Chevrier é, Scherzer P, Mottron L, Godbout R. Sleep in children with high functioning autism: polysomnography, questionnaires and diaries in a non-complaining sample. Sleep Med 2013;14:e137–8. https://doi.org/10.1016/j.sleep.2013.11.310.
[14] Tempesta D, Socci V, De Gennaro L, Ferrara M. Sleep and emotional processing. Sleep Med Rev 2018;40:183–95. https://doi.org/10.1016/j.smrv.2017.12.005.
[15] Epstein CM. Analog signal recording principles. In: Schomer DL, Lopes da Silva FH, editors. Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields. 6th ed. 2011. p. 111–8.
[16] Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004;134:9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.
[17] Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, et al. The visual scoring of sleep in adults. J Clin Sleep Med 2007;3:121–31.
[18] Cox R, Schapiro AC, Manoach DS, Stickgold R. Individual differences in frequency and topography of slow and fast sleep spindles. Front Hum Neurosci 2017;11:433. https://doi.org/10.3389/fnhum.2017.00433.
[19] Cox R, Mylonas DS, Manoach DS, Stickgold R. Large-scale structure and individual fingerprints of locally coupled sleep oscillations. Sleep 2018;41. https://doi.org/10.1093/sleep/zsy175.
[20] Dumermuth G, Walz W, Scollo-Lavizzari G, Kleiner B. Spectral analysis of EEG activity in different sleep stages in normal adults. ENE 1972;7:265–96. https://doi.org/10.1159/000114432.
*[21] Welch P. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust 1967;15:70–3. https://doi.org/10.1109/TAU.1967.1161901.
*[22] Press WH, Teukolsky SA, Flannery BP, Vetterling WT. Numerical recipes in FORTRAN 77: volume 1, volume 1 of Fortran numerical recipes: the art of scientific computing. Cambridge University Press; 1992.
[23] Epstein CM, Brickley GP. Interelectrode distance and amplitude of the scalp EEG. Electroencephalogr Clin Neurophysiol 1985;60:287–92. https://doi.org/10.1016/0013-4694(85)90001-X.
[24] Sleigh JW, Steyn-Ross DA, Steyn-Ross ML, Williams ML, Smith P. Comparison of changes in electroencephalographic measures during induction of general anaesthesia: influence of the gamma frequency band and electromyogram signal. Br J Anaesth 2001;86:50–8.
[25] Vijayan S, Lepage KQ, Kopell NJ, Cash SS. Frontal beta-theta network during REM sleep. ELife 2017;6. https://doi.org/10.7554/eLife.18894.
[26] Cox R, Rüber T, Staresina BP, Fell J. Heterogeneous profiles of coupled sleep oscillations in human hippocampus. Neuroimage 2019;202:116178. https://doi.org/10.1016/j.neuroimage.2019.116178.
[27] Bastos AM, Schoffelen J-M. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. Front Syst Neurosci 2015;9:175. https://doi.org/10.3389/fnsys.2015.00175.
[28] Lachaux J-P, Rodriguez E, Martinerie J, Varela FJ. Measuring phase synchrony in brain signals. Hum Brain Mapp 1999;8:194–208. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C.
[29] Fell J, Axmacher N. The role of phase synchronization in memory processes. Nat Rev Neurosci 2011;12:105–18. https://doi.org/10.1038/nrn2979.
[30] Bruns A, Eckhorn R, Jokeit H, Ebner A. Amplitude envelope correlation detects coupling among incoherent brain signals. Neuroreport 2000;11:1509–14. https://doi.org/10.1097/00001756-200005150-00029.
*[31] Kayser J, Tenke CE. Issues and considerations for using the scalp surface Laplacian in EEG/ERP research: a tutorial review. Int J Psychophysiol 2015;97:189–209. https://doi.org/10.1016/j.ijpsycho.2015.04.012.
[32] Tenke CE, Kayser J. Surface Laplacians (SL) and phase properties of EEG rhythms: simulated generators in a volume-conduction model. Int J Psychophysiol 2015;97:285–98. https://doi.org/10.1016/j.ijpsycho.2015.05.008.
[33] Canolty RT, Knight RT. The functional role of cross-frequency coupling. Trends Cognit Sci 2010;14:506–15. https://doi.org/10.1016/j.tics.2010.09.001.
[34] van Driel J, Cox R, Cohen MX. Phase-clustering bias in phase–amplitude cross-frequency coupling and its removal. J Neurosci Methods 2015;254:60–72. https://doi.org/10.1016/j.jneumeth.2015.07.014.
*[35] Cohen M. Analyzing neural time series data. MIT Press; 2014.
[36] Berens P. CircStat: a MATLAB toolbox for circular Statistics. J Stat Software 2009;31. https://doi.org/10.18637/jss.v031.i10.
[37] Rings T, Cox R, Rüber T, Lehnertz K, Fell J. No evidence for spontaneous cross-frequency phase–phase coupling in the human hippocampus. Eur J Neurosci 2019. https://doi.org/10.1111/ejn.14608. ejn.14608.
[38] Scheffer-Teixeira R, Tort AB. On cross-frequency phase-phase coupling between theta and gamma oscillations in the hippocampus. ELife 2016;5. https://doi.org/10.7554/eLife.20515.
[39] Vinck M, Oostenveld R, van Wingerden M, Battaglia F, Pennartz CMA. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. Neuroimage 2011;55:1548–65. https://doi.org/10.1016/j.neuroimage.2011.01.055.
[40] Cox R, Rüber T, Staresina BP, Fell J. Phase-based coordination of hippocampal and neocortical oscillations during human sleep. Commun Biol 2020;3. https://doi.org/10.1038/s42003-020-0913-5.
*[41] Cohen MX. Rigor and replication in time-frequency analyses of cognitive electrophysiology data. Int J Psychophysiol 2017;111:80–7. https://doi.org/10.1016/j.ijpsycho.2016.02.001.
[42] Purcell SM, Manoach DS, Demanuele C, Cade BE, Mariani S, Cox R, et al. Characterizing sleep spindles in 11,630 individuals from the National sleep research resource. Nat Commun 2017;8:15930. https://doi.org/10.1038/ncomms15930.
[43] Murdoch DJ, Tsai Y-L, Adcock J. P-values are random variables. Am Stat 2008;62:242–5. https://doi.org/10.1198/000313008X332421.
[44] Muehlroth BE, Werkle-Bergner M. Understanding the interplay of sleep and aging: methodological challenges. Psychophysiology 2020. https://doi.org/10.1111/psyp.13523.
*[45] Nieuwenhuis S, Forstmann BU, Wagenmakers E-J. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat Neurosci 2011;14:1105–7. https://doi.org/10.1038/nn.2886.
*[46] Benjamini Y, Hochberg Y. Controlling the false Discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B (Methodol) 1995;57:289–300.
*[47] Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 2007;164:177–90. https://doi.org/10.1016/j.jneumeth.2007.03.024.

* The most important references are denoted by an asterisk.

*[48] Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–9. https://doi.org/10.1016/S0895-4356(00)00314-0.
[49] Ackermann S, Hartmann F, Papassotiropoulos A, de Quervain DJ-F, Rasch B. No associations between interindividual differences in sleep parameters and episodic memory consolidation. Sleep 2015;38:951–9. https://doi.org/10.5665/sleep.4748.
[50] Bailes C, Caldwell M, Wamsley EJ, Tucker MA. Does sleep protect memories against interference? A failure to replicate. PloS One 2020:14.
[51] Cordi M, Ackermann S, Bes FW, Hartmann F, Konrad BN, Genzel L, et al. Lunar cycle effects on sleep and the file drawer problem. Curr Biol 2014;24:R549–50. https://doi.org/10.1016/j.cub.2014.05.017.
[52] Henin S, Borges H, Shankar A, Sarac C, Melloni L, Friedman D, et al. Closed-loop acoustic stimulation enhances sleep oscillations but not memory performance. Eneuro 2019;6. https://doi.org/10.1523/ENEURO.0306-19.2019. ENEURO.0306-19.2019.
[53] Humiston GB, Wamsley EJ. Unlearning implicit social biases during sleep: a failure to replicate. PloS One 2019;14:e0211416.
[54] Lafon B, Henin S, Huang Y, Friedman D, Melloni L, Thesen T, et al. Low frequency transcranial electrical stimulation does not entrain sleep rhythms measured by human intracranial recordings. Nat Commun 2017;8:1199. https://doi.org/10.1038/s41467-017-01045-x.
[55] Sahlem GL, Badran BW, Halford JJ, Williams NR, Korte JE, Leslie K, et al. Oscillating square wave transcranial direct current stimulation (tDCS) delivered during slow wave sleep does not improve declarative memory more than sham: a randomized sham controlled crossover study. Brain Stimul 2015;8:528–34. https://doi.org/10.1016/j.brs.2015.01.414.
[56] Schäfer SK, Wirth BE, Staginnus M, Becker N, Michael T, Sopp MR. Sleep's impact on emotional recognition memory: a meta-analysis of whole-night, nap, and REM sleep effects. Sleep Med Rev 2020;51:101280. https://doi.org/10.1016/j.smrv.2020.101280.
[57] Haba-Rubio J, Marques-Vidal P, Tobback N, Andries D, Preisig M, Kuehner C, et al. Bad sleep? Don't blame the moon! A population-based study. Sleep Med 2015;16:1321–6. https://doi.org/10.1016/j.sleep.2015.08.002.
[58] Pesonen A-K, Ujma P, Halonen R, Räikkönen K, Kuula L. The associations between spindle characteristics and cognitive ability in a large adolescent birth cohort. Intelligence 2019;72:13–9. https://doi.org/10.1016/j.intell.2018.11.004.