On the development of sleep states in the first weeks of life

Tomasz Wielek $^{1,2}$, Renata Del Giudice $^3$, Adelheid Lang $^{1,2}$, Malgorzata Wisiowska $^{1,2}$, Peter Ott $^4$, Manuel Schabus $^{1,2}$

$^1$Laboratory for Sleep, Cognition and Consciousness Research, University of Salzburg, Salzburg, Austria

$^2$Centre for Cognitive Neuroscience (CCNS), University of Salzburg, Salzburg, Austria

$^3$Università degli Studi di Milano-Bicocca

$^4$ITS Informationstechnik & System-Management, Salzburg University of Applied Sciences, Salzburg, Austria

*Corresponding author(s)

E-mail: wielekto@stud.sbg.ac.at

E-mail: manuel.schabus@sbg.ac.at
Abstract

Human newborns spend up to 18 hours sleeping. The organization of their sleep differs immensely from adult sleep, and its quick maturation and fundamental changes correspond to the rapid cortical development at this age. Manual sleep classification is specifically challenging in this population given major body movements and frequent shifts between vigilance states; in addition various staging criteria co-exist. In the present study we utilized a machine learning approach and investigated how EEG complexity and sleep stages evolve during the very first weeks of life. We analyzed 42 full-term infants which were recorded twice (at week two and five after birth) with full polysomnography. For sleep classification EEG signal complexity was estimated using multi-scale permutation entropy and fed into a machine learning classifier. Interestingly the baby’s brain signal complexity (and spectral power) revealed huge developmental changes in sleep in the first 5 weeks of life, and were restricted to NREM ("quiet") and REM ("active sleep") states with little to no changes in state wake. Data demonstrate that our classifier performs well over chance (i.e., >33% for 3-class classification) and reaches almost human scoring accuracy (60% at week-2, 73% at week-5). Altogether, these results demonstrate that characteristics of newborn sleep develop rapidly in the first weeks of life and can be efficiently identified by means of machine learning techniques.
The organization of newborn sleep differs from adult sleep, and its ongoing maturation over time corresponds with cortical development. However, sleep scoring in this population is challenging given frequent artifacts in polysomnography (PSG) and absence of established staging criteria which results in low inter-scorer reliability. To investigate changes in the early brain activity, we analyzed large sample of newborn data at week 2 and 5 after birth. First we evaluated sleep that was previously scored visually, in terms of both entropy and oscillatory power. Next we accessed the performance of automatic sleep scoring based on machine learning compared with conventional, manual scoring. We observed clear developmental changes on the brain-level in the first 5 weeks of life in human newborns. These changes were limited to “quiet” NREM and “active” REM sleep. Also our classifier data demonstrated that we can classify well above chance and similar to human scorers using multi-scale permutation entropy (and just 6 EEG and 5 physiological channels).
Introduction

Sleep of newborns greatly differs from the sleep of kids or adults. Adult-like classification of sleep into classical sleep stages is possible only from the age of 2-3 months onwards, since only then typical NREM patterns, like sleep spindles, K-complexes, or slow waves emerge (AASM; [1]). Until then the EEG landscape is dominated by low-voltage-irregular (REM/Wake), high-voltage slow (NREM/REM), mixed (Wake/NREM/REM) and tracé alternant patterns in NREM. Oscillatory activity of newborns is dominated also during wake by slow oscillations of a very high amplitude up to 100 µV [2]. Another hallmark of early brain activity is bursting activity (known also as spontaneous “activity transients” or “delta brushes”), characterized by slow delta-like waves with superimposed fast, beta range activity [3, 4]. Hence in the first weeks of life the only discrimination of sleep stages possible is between wake, active sleep (REM) and quiet sleep (NREM) [5, 6].

Importantly the neonatal sleep-wake state organization also impacts on later development [7-9]. For instance sleep characteristics during the first postnatal days are related to cognitive development at an age of 6 months [10]. Being able to reliably characterize sleep in newborns has been recognized as crucial for both research and pediatric practice but is even to-date an inherently difficult endevaour.

Traditionally, sleep of infants is staged based on a visual inspection of polysomnographical (PSG) recordings, which often is supplemented by simultaneous observation of overt behavior and respiratory activity. However such manual sleep staging is time consuming, costly, requires high expertise, and can be quite variable due to varying staging criteria and/or noisy data. Crowell and colleagues [11] for example reported moderate inter-scorer reliability for infant sleep staging, with kappa coefficient going below 0.6, when staged according to the modified Anders manual [5]. Reliability can be improved by refining criteria for a specific age group individually, as done for example by Satomaa and colleagues [12], who reached kappa score of 0.73 indicating substantial agreement for one-month old babies. From a practical point of view however such fine tuning of sleep staging criteria might render rather low reproducibility as the results cannot generalize to other
age groups. We here try to address this issue and reached out for a more data-driven and objective analysis of EEG data using machine learning.

In the context of brain oscillations typically frequency-resolved information is exploited by means of spectral methods such as Fast Fourier. More recently however, a stronger focus on irregular dynamics of brain signals gave rise to entropy-based features (for a review see [13]). Entropy quantifies the extent of irregularity in the EEG time signal, where repeating, predictable signal yield low entropy, while irregular, unpredictable signal yields high entropy. In contrast to the power spectrum capturing only linear properties of the brain signal, entropy-based features emphasize also additional characteristics of the EEG that are related to non-linear dynamics [14]. A non-linear behavior of human EEG was for example detected during adult sleep, especially in N2 stage [15, 16]. Also, a big practical advantage is that entropy-based features, such as permutation entropy, are typically more robust against common EEG artifacts as compared to spectral measures [17] which is especially important when analyzing the data of newborns.

Mounting evidence suggests that fluctuations in EEG entropy reflect both transient, state-like changes in human brain activity (e.g. wake and sleep states), as well as slower and more long lasting dynamics across a day or even brain maturation. Sleep studies investigating healthy adults report an overall trend of entropy decrease from WAKE, across transitional (N1) and light (N2) to deep (N3) sleep, with a relative increase during REM sleep [18-20]. A similar pattern was reported in newborns with higher entropy levels in active/REM sleep as compared to quite/NREM sleep [21]. Diurnal changes in EEG entropy between daytime and night time periods – although diminished in size in relation to healthy individuals – were also found in patients following a severe brain-injury [22].

Across the development from childhood to adulthood a permanent increase of EEG entropy has been observed [23]. Also in adults (19–74 years) an age-related increase of EEG entropy was reported [24]. Interestingly, during the first weeks of infancy, this pattern is much less consistent, and may even be accompanied by transient declines in EEG entropy. For instance, Zhang and colleagues [25] reported EEG entropy during sleep to increase across the first month of baby’s life.
Then suddenly this pattern changes, such that entropy remains constant during quiet sleep, but decreases during active sleep. This transient change in entropy evolution is in good agreement with previous results showing a general decrease in high frequency (beta band) power occurring within the first month of life ([26, 27] also for sleep stage specific findings see [28]).

To delve deeper into the dynamically changing landscape of early brain activity, we analyzed a well-sized sample of newborn data (PSG and hdEEG) at week 2 and 5 after birth. First we evaluated sleep that was previously scored visually, in terms of both entropy and oscillatory power. In contrast to Zhang and colleagues, where temporally unspecific entropy was used, we quantify entropy over multiple temporal scales, with the aim to add knowledge, especially in light of classic, ‘frequency-resolved’ findings. Second we aimed at testing the possibility of automatizing sleep staging by using the previously extracted entropy measures in a machine learning approach. This approach could ultimately complement or even replace visual staging and thus make the sleep scoring in newborns more objective and replicable. Last but not least, we tried cross-session classification such that we could assess whether our algorithm can generalize between age groups and reveal the similarity or “dissimilarity” in sleep organization this early in life.
Participants and methods

Participants and EEG recording

Mothers of 42 full-term infants (15 female) were recruited for a study on prenatal learning. Polysomnography (PSG) was recorded from all but one newborn during two separate sessions: first at 2 weeks (14.8±4.3 days) after birth and then 5 weeks (36.7±4.3 days) after birth. Recording took place in the home environment of the mother and infant. EEG were recorded with an ambulatory, high-density (128-EEG channels cap) system using a Geodesic Sensor Net (Geodesic EEG System 400, Electrical Geodesics, Inc, Eugene, OR; US). The signal was recorded continuously over 35min (n=11) or 27min (n=31) with a sampling rate of 500Hz. The study was approved by the ethic committee of the University of Salzburg (EK-GZ 12/2013) and all parents provided written informed consent before participation.

Data preprocessing

Preprocessing was done in Brain Vision Analyzer (Brain Products GmbH, Gilching, Germany, version 2.0) and MNE-python software (version 0.16.1). Data was down-sampled to 125 Hz and re-referenced to average reference of all the 128 EEG channels. EEG was band-pass filtered (FIR filter with hamming window) between 0.5 and 35 Hz. Segments with artifacts were rejected based on simple power spectral density (PSD) thresholds and resulted in an exclusion of 4.5% of segments. Note that this step was performed in addition to the exclusion of segments staged as movement or transitional sleep by the human scorer. The high density EEG montage was then subsampled to the equivalent of a habitual sleep montage with only 6 EEG channels (F3, F4, C3, C4, O1, O2) and 5 peripheral channels: bipolar ECG, bipolar EMG, bipolar VEOG, as well as a HEOG left, and HEOG right both re-referenced to the right ear as recommended in [29]. For the subsequent automatized sleep scoring we used the same reduced PSG setup as we used for the visual sleep scoring. This allows making a fair comparison between the two sleep scoring approached on one hand, and increases applicability of our classifier to other baby PSG datasets on the other hand.
Visual sleep staging

Eighty-four recordings were visually sleep-scored by an expert sleep scorer (Scholle) according to scoring criteria [30] based on 30-second PSG segments. Each segment was assigned to one of the five classes: quiet sleep (NREM), active sleep (REM), wake (WAKE), movement time, and transitional sleep. To account for possible difference in the amount of movement from week-2 to week-5 all epochs scored as movement (as well as transitional sleep) were excluded from further analysis (7.7% of segments). Ten recordings were considered “unscorable” by our expert and removed from further analysis. In the next step we validated the manual scorings by examining the simultaneously recorded videos. We followed established Prechtl staging criteria [31]. Whenever we detected mismatch between PSG-based scorings and video recordings, for example observed infant’s open eyes in epochs staged as sleep, we sought for a consensus score. Due to technical issues (EEG signal corrupted) we excluded one of the recording sessions of one participant. All in all 72 recordings (34 at 2 weeks of age; 38 at 5 weeks of age) were included in the final analyses.

Entropy measure

Entropy quantifies the irregularity or complexity of signal fluctuations, where repeating, highly predictable signal yield low entropy, while irregularity yields high entropy. We used permutation entropy (PE) as a robust entropy measure that first converts EEG time series into a sequence of data-patterns (where each pattern describes the order relations between neighboring EEG voltages), and next quantifies the distribution of these patterns by using the Shannon entropy equation (cf. Fig 1a). Highest PE (maximal information) is attained when all patterns have equal probability. Further generalization of this method, called multi-scale permutation entropy (MSPE), applies coarse-graining to the original broadband signal by averaging data within non-overlapping windows [32]. Like low-pass filtering, coarse-graining eliminates fast fluctuations from the signal biasing the complexity estimates towards increasingly slower time scale (cf. Fig 1b).
MSPE was calculated for non-overlapping 30s segments, for each PSG channel separately, and for 5 different levels of coarse-graining (scales). To maximize the predictive power of the classifier we used all 5 temporal scales as an input. This resulted in 55-dimensional feature vectors (per segment) for each subject. The classifier is later called as MSPE-based. For univariate analysis reported later we used MSPE computed at a scale of 1 (original signal, mixture of fast and slow temporal scales) and at a scale of 5 (fast temporal scales eliminated), later referred to as fast scale and slow time scales respectively. Analysis for intermediate scales was included in supplementary materials for completeness (cf. S6 Fig).

Spectral measure

We also calculated power spectral density (PSD) for the same 30s segments in 1 Hz steps and for frequencies between 1-30Hz. Welch’s method with a Hamming window was used. To facilitate comparison between session, PSDs were transformed to relative powers by dividing PSD by the sum of the PSD from all frequency bins (relative power was expressed in %). For univariate analysis we summed up frequency bins between two frequency ranges: 1-3 Hz and 12-30 Hz as capturing slow and fast activity. Analysis for intermediate frequency band (4-11Hz) was included in supplementary materials for completeness. Similarly to MSPE analysis, we used all frequency bins as an input for the classifier (later called PSD-based) providing 330-dimensional (11 x 30) feature vectors (per segment) for each subject.

Fig 1. Multi-scale permutation entropy as useful feature for neonatal sleep stage classification. (a) Six possible ordinal patterns are identified ($p_i$), their distribution is formed and Shannon entropy is computed. (b) Simulated time series prior (left) and after the coarse-graining procedure (right) are shown. Note that the granulation removes the fast-varying changes, which allows estimating the entropy on a slower temporal scale. (c) PSDs of individual epochs for a single recording (week-5) and corresponding visual/manual sleep staging; note that dashed lines mark three different stages for illustration, log-transformed power is shown for better visibility. (d) MSPE for these three exemplary epochs depicted separately for original signal (left) and coarse-grained signal (right). Note that these three epochs or states are distinguishable on both scales, yet different patterns in the original and coarse-grained computation.
Statistical analysis

Univariate statistical analysis was performed for both entropy measures as well as PSD estimates. Linear mixed models were used as they are more suited to deal with unbalanced datasets than repeated measure ANOVA. This is of a particular benefit as the number of available segments was limited (for example NREM for week 2; cf. Table 1). To provide equal number of observation for each subject and for each sleep stage MSPE (or PSD) values were averaged first across EEG channels, and next across a random sample of 10 epochs. Matrix of MSPE (or PSD) values (sleep stage x participant x session) entered the model as a dependent variable. Two sessions (week-2 and week-5) and three sleep stages (NREM, REM, WAKE) served as fixed effects, and participants as a random effect. We also included a random slope to account for individual amount of changes in complexity of each baby from week-2 to week-5. To select the model with optimal fit the Akaike information criterion was applied. All model parameters were estimated using restricted maximum likelihood estimation. Wald Chi-Squared test was used to test for significance of the model variables. Two independent testings were performed with MSPE at scale of 1 and at scale of 5 as a dependent variable. Analogous testing was performed for the PSD measures. Linear mixed model analysis was performed with the lme4 package [33] by using the statistical software R 3.4.0 [34]. Post hoc, multiple comparison procedure was performed with Tukey test using the glht-method of the multcomp package [35].

Machine Learning

The principle behind supervised machine learning (ML) is to train a predictive model, by automatically encapsulating information from previously labeled dataset (in our case visual sleep staged PSG epochs). We performed an epoch-by-epoch classification into one of three distinct sleep classes: NREM, REM, WAKE. A random forest (RF) classifier was used as a main component of the classification pipeline. We employed repeated (10 repetitions) two-fold cross-validation by indexing half of the subjects as testing and the other half as training subjects. It has been shown that this
procedure has smaller variance than the typical leave-one-subject-out cross validation [36]. Training and testing sets were created by concatenating data of all subjects assigned to one of the two groups. Random under sampling of epochs was performed to equalize the number of epochs across sleep class, both in the training and testing set. Additional cross-validation with the training set was performed to find optimal configuration of the classifier parameters (hyper-parameters). Finally we used two types of the two-fold cross-validation: (1) within sessions such that both training and testing subjects were of the same age, (2) across sessions such that training subjects and testing subjects were of different age (cf. S2 Fig). Importantly both procedures were run as a between-subject classification, such that each participant was assigned to either the training or testing set (i.e., no two sessions of the same subject in training and testing at the same time). Each cross-validation was repeated 20 times, and the median scores (accuracy, F1-score) are reported. Chance level of the accuracy scores was estimated by running each cross-validation on shuffled data (100 repetitions). The machine learning analysis was performed in Python using the scikit-learn package [37]. To evaluate the performance of a classifier standard performance metrics were used. Class-wise performance (for NREM, REM and WAKE classes separately) as well as the overall performance was accessed using the F1 score and the accuracy score respectively (cf. Fig 2).

To compare output of the classifiers based on different data both a MSPE- and PSD-based classification was repeated 20 times on the full dataset (i.e., the week 2 and week 5 data merged). The corresponding accuracy scores of MSPE- and PSD-based classifiers were compared statistically with Mann-Whitney U test. Since the MSPE-based classifier had significantly better performance as compared to the PSD-based classifier (cf. S1 Fig), we restrict our subsequent classification analysis to MSPE classifier results.

Fig 2. Evaluation of the performance (a) The confusion matrix relates the actual (visual) scoring to the classifier predictions. (b) Values from the confusion matrix were used to calculate the overall accuracy and the class-specific F1 scores. Note that diagonal corresponds to agreement between the visual sleep staging and the MSPE-based automatic classification. Whereas accuracy (top panel) reflects relative amount of agreement between predicted and actual scores, F1 accounts for precision (middle panel) and recall (bottom panel). Precision (known also as positive predictive value) takes into account false positives and is defined as the ratio of epochs classified by both the classifier and the human scorer as
given sleep stage to all of the epochs that the classifier assigned as that sleep stage (b, middle panel). Recall (known also as sensitivity) in turn takes into account the amount of false negatives and is defined as the ratio of epochs classified by both the classifier and the human scorer as given sleep stage to all of the epochs that the human scorer classified as the given sleep stage (b, lower panel). (c) The F1 score combines the two measures into a single metric.
Results

To assess sleep rhythm of newborns EEG recordings in a sample of 42 participants were used. Final analysis included 72 visually sleep scored PSG recordings divided into two age groups: 2-weeks-old (N=34) and 5-weeks-old (N=38). We evaluated changes in sleep stage distribution (including NREM, REM and WAKE stages) from the age of 2 weeks to the age of 5 weeks, but found no significant effects (Wilcoxon Signed-Ranks tests; NREM: Z=10.5, p=.15, REM: Z=191.0, p=.56, WAKE: Z=32.5, p=.21). Subsequently, EEG signal complexity was investigated across different sleep stages and recording sessions. Please see Table 1 for a detailed overview with respect to age, distribution of sleep stages and the complexity measure – multi-scale permutation entropy.

Table 1. Sample descriptives.

| Session | Age (in days) | Sleep stages (visually scored 30s epochs) | Subjects | MSPE fast | MSPE fast | MSPE slow | MSPE slow |
|---------|---------------|------------------------------------------|----------|-----------|-----------|-----------|-----------|
|         | Mean  | SD   | Name | n   | Mean  | SD   | Mean  | SD   |
| week-2  | 13.9  | 3.7  | NREM | 213 | 1.413 | 0.045 | 1.61 | 0.036 |
|         |       |      | REM  | 1004 | 1.486 | 0.045 | 1.616 | 0.033 |
|         |       |      | WAKE | 431 | 1.488 | 0.069 | 1.581 | 0.054 |
| week-5  | 36.2  | 3.9  | NREM | 387 | 1.367 | 0.048 | 1.573 | 0.049 |
|         |       |      | REM  | 1056 | 1.442 | 0.04 | 1.603 | 0.026 |
|         |       |      | WAKE | 392 | 1.483 | 0.06 | 1.569 | 0.035 |

The analyzed data set consisted of two recordings from a group of healthy newborns with three sleep stage classes that were identified using visual scoring. EEG signal complexity was estimated using multi-scale permutation entropy (MSPE). The subject count reflects the number of participants who actually showed a specific sleep stage in the respective (week-2 or week-5) recording sessions.
Entropy and spectral measures

We observed that the multi-scale permutation entropy (MSPE) at fast scale (i.e., no coarse-graining, incl. mixture of low and faster frequencies) as well as slow scale (i.e., fast temporal scales eliminated) significantly differed between sleep stages as well as recording sessions. Table 1 and Fig 3 illustrate the results for MSPE averaged over all 6 EEG channels (for individual channels also see S4 Fig).

At a fast time scale (Fig 3, left) there was a main effect of session (week-2 vs week-5: Wald chi-square \( (1, 73) = 22.01, p<.001 \)) indicating higher permutation entropy during week-2 as compared to week-5. Furthermore we observed a main effect of sleep stage (NREM vs REM vs WAKE: Wald chi-square \( (2, 73) = 93.54, p<.001 \)), and a marginally significant interaction between session and sleep stage (Wald chi-square \( (2, 73) = 4.9, p=.086 \)). A post-hoc Tukey test for the factor sleep state confirmed higher entropy during WAKE compared to both REM (mean(SE)=.024(.008); Z=2.98, \( p=.008 \)) as well as NREM (mean(SE)=.098(.009); Z=9.91, \( p<.001 \)), there was also significant difference between REM and NREM (mean(SE)=.073(.009); Z=8.34, \( p<.001 \)). A post-hoc Tukey test on the interaction revealed higher entropy during WAKE compared to REM, but only at week-5 (mean(SE)=.039(.011); Z=3.54, \( p=.005 \)). Note that this effect appears pronounced at frontal recordings sites (cf. S4 Fig). REM in turn has higher entropy than NREM at both week-2 (mean(SE)=.068(.015); Z=4.61, \( p<.001 \)) and week-5 (mean(SE)=.069(.011); Z=6.397, \( p<.001 \)) recordings. Similarly, WAKE has higher entropy than NREM at both week-2: mean (SE) = .074(.016); Z=4.48, \( p<.001 \) and week-5: mean (SE) = .11(.012); Z=9.03, \( p<.001 \).

At a coarse time scale (Fig 3, right) we found a similar main effect of session (week-2 vs week-5: Wald chi-square \( (1, 73) = 7.8, p=.005 \)) indicating an overall decrease in EEG complexity from week-2 to week-5. Additionally we observed a significant main effect of the sleep stage (NREM, REM, WAKE) (Wald chi-square \( (2, 73) = 35.6, p<.001 \)). A post-hoc Tukey test for the factor sleep state revealed higher entropy during REM compared to NREM (mean(SE)=.02(.006); Z=3.59, \( p<.001 \), also higher entropy during REM compared to WAKE (mean(SE)=.03(.006); Z=5.69, \( p<.001 \)). Moreover, we
observed that the pattern of relative entropy levels (across sleep stages) reverses at coarse temporal scale (Fig 3, right) compared to fine temporal scale (Fig 3, left); such that WAKE shows the lowest entropy level for coarse temporal scale.

Fig 3. Entropy at both fast (left) and slow (right) time scale across sleep/wake states and the two recording sessions. MSPE values were averaged across 6 EEG electrodes. Note that at the fast scale all three stages are distinguishable at week 5, but not yet at week 2. At both scales there is a clear difference in signal complexity between week 2 and week 5 with week 2 being generally higher in entropy.

Similarly, we also evaluated changes in power spectral densities (PSD) across sleep stages and recording session. At younger as compared to older age babies revealed increased power of fast oscillations (12-30Hz) (Fig 4, left) (week-2 vs week-5: Wald chi-square (1, 73) = 5.96, p=.015). Also we observed a main effect of sleep stage (NREM vs REM vs WAKE: (2, 73) = 20.3, p < .001). A post-hoc Tukey test for the factor sleep state revealed higher power of fast oscillations during WAKE compared to NREM (mean(SE)=.68(.15); Z=4.56, p<.001), similarly higher power for REM compared to NREM (mean(SE)=.47(.13); Z=3.51, p=.001). In contrast no significant difference was found from week 2 to week 5 for slow oscillations (1-3Hz) (Fig 4, right). However there was a main effect of sleep stage (NREM vs REM vs WAKE: (2, 73) = 7.65, p=.02). A post-hoc Tukey test for the factor sleep state revealed higher power of slow oscillations during NREM compared to REM (mean(SE)=1.85(.67); Z=2.74, p=.02). No significant differences were found for power in the intermediate frequencies (4-11 Hz) (cf. S5 Fig).

Fig 4. Relative power in the fast (left) and in the slow (right) oscillations across sleep/wake states and the two recording sessions. Power was averaged across 6 EEG electrodes. In lower frequencies (1-3Hz) no power difference is evident between week 2 and week 5 data. Relative power in the y-axis is plotted as the ratio between the respective frequency band (1-3 or 12-30Hz) to 1-30Hz total power; as illustrated more than 90% of the spectrum are in the 1-3Hz range.


**Machine learning classification**

**Influence of the feature extraction on the classification performance: PSD vs MSPE**

Multiscale permutation entropy improves the overall classification accuracy as compared to PSD (cf. S1 Fig). The per-class evaluation shows that MSPE improves performance in the discrimination of all classes, but in particular for WAKE (73% instead of 59% median accuracy) wherefore we decided to focus on MSPE in all subsequent classification analyses.

**Separate classification for week-2 and week-5 – within age-group**

To automate sleep staging we used machine learning and previously extracted entropy features. We first performed within age-group classification and evaluated results from the two independent and age-specific classifiers. That is, we computed the performance scores of a classifier that was both train and tested on data from either (1) week-2 or (2) week-5 recording. Accuracy scores across all 3 sleep stages were significantly higher than would be expected by chance. In randomization test across 3 classes, chance level was at ~33%. For both the week-2 and week-5 classification (cf. Fig 5, left upper panel). Moreover, the classifier performed better on the week-5 (Mdn=72.7%) as compared to the week-2 (Mdn=60.1%) babies (U=1, Z=6.5, p<0.001). The per-class evaluation (cf. Fig 5, left lower panel) shows higher F1 classification scores during week-5 (Mdn=63.1%) as compared to week-2 for REM (Mdn=53.2%); (U=24, Z=5.2, p<0.001), as well as for WAKE (Mdn=83.6% vs. Mdn=58.5%; U=1, Z=6.6, p<0.001).

A confusions matrix (cf. S3 Fig) as well as visual inspection of the single subject results (cf. S7 Fig for an exemplary subject) reveals that a very limited proportion of NREM epochs is actually falsely classified as WAKE (on average 3%) or vice versa (8% of the epochs). During week-2 the classifier is worst in distinguishing REM and WAKE, while at week-5 the classification generally increases and NREM and REM are sometimes wrongly assigned.
Cross classification between week-2 and week-5 – across age-groups

In a second and last step we tested whether the classifier can generalize across age groups. That is we trained the classifier on a given age-group and looked at the classification accuracy onto the other age-group (cf. Fig 5, right).

Compared to the within-session classification (week-2 to week-2 and week-5 to week-5), the cross-session (week-5 to week-2 and week-2 to week-5) classification accuracy was decreased by 7.7% and 9.9%, respectively. Classification however remained well above chance level (~33%) and was better when the classifier was trained on week-2 and tested on week-5 data (Mdn=62.8%) as vice versa (Mdn=53.1%) (cf. Fig 5, right). Interestingly, the per-class evaluation shows that detection of NREM deteriorates when trained on week-5 (and tested on week-2) data, and REM classification deteriorates when trained on week-2 (and tested on week-5) data only. The classification of state WAKE remains relatively stable indicating that the way this stage presents itself in the EEG remains widely constant from week-2 to week-5 in the newborn.

Fig 5. Classification of sleep states in the week-2 and week-5 old newborn (left panel) and its generalization across age (right panel). Sleep in older (5 week old) babies can be classified more accurately as compared to 2 weeks (left panel). Right panel shows that the "generalization" or classification across age-groups leads to lower classification accuracy specifically for detecting stage NREM (week-5 training, week-2 test), but also stage REM (week-2 training, week-5 test). Vertical histograms represent the null distribution, with the empirically estimated chance levels in both cases being close to 33% (red dashed line). 95% confidence-intervals are displayed on the basis of bootstrap analysis.

Uncovering the black box of classification—feature importance and decision boundaries

Evaluation of channel importance from a trained random forest classifier revealed that primarily the horizontal EOGs, frontal brain channels (F4, F3) as well ECG contribute to a good classification accuracy (S8 Fig). Visualization of classification boundaries of a trained random forest classifier in addition confirms previous results and reveals more compact and distinct sleep/wake classes at week-5 as compared to week-2 (see S8 Fig for details).
Discussion

Over the first weeks of life the human brain is growing at a rapid rate establishing a complex network that includes trillions of synaptic connections [38]. Accordingly, a continuous increase in brain signal complexity could be expected. Instead, we observed a clear decrease in EEG complexity from week-2 to week-5 in present data. Although similar findings have already been reported for both entropy [25] as well as spectral [26-28, 39] EEG brain measures, the understanding of what might account for this effect remains limited. Most of the aforementioned authors point to a decline in bursting activities (known also as spontaneous “activity transients” or “delta brushes”), which are abundant in premature infants, but which remain detectable until about the end of the first month of life [40]. This corresponds to our second measurement at week 5 and would be in good agreement.

In contrast to prior approaches, where temporally more unspecific entropy measures were used [25], we used multi-scale entropy (MSE) providing more details about temporal scale or frequency band which may contribute to the effect. In our data we found that both EEG entropy at fast temporal scales and spectral power in high frequencies decreases with age, which agrees with the classical characterization of delta brushes as slow delta-like waves with superimposed fast activity [3, 4]. Interestingly, whereas there was also a decrease in entropy at the slow temporal scale, we found no differences in 1-3Hz power. This confirms that not all changes in EEG complexity are reflected by changes in power [14, 15]. Physiologically, this could reflect an ongoing reorganization of brain oscillations where ‘multiple delta waves’ at week-2 gradually settle down, forming a single ‘more mature delta wave’ at week-5. This interpretation is in agreement with the general idea of brain oscillations as being in constant flux and realignment mode [41, 42].

Most of the before mentioned studies focus exclusively on sleep periods (as it constitutes about 70% of the time in newborns from birth to 2 months), and leave out periods of wakefulness. It was therefore rather unclear whether foremost “quiet” NREM and “active” REM sleep states show these extensive changes early in development or if similar changes in the “brain signatures” are found in wakefulness during the first weeks of life. Indeed, data indicate that a massive drop in
entropy at the fast temporal scale is exclusive to NREM and REM sleep states, with no significant changes during wakefulness. Likewise, we observed a decrease in EEG spectral power, in higher frequency ranges (12-30Hz), from week-2 to week-5. These findings support the first line of interpretation, as in both neonatal rodents [43] and human [4] newborns delta brushes (after term age) are observed mainly during immobility periods, that is sleep and feeding. Interestingly, the age-related decrease in entropy at a slow temporal scale is also observed only for NREM which supports the idea of a reorganization of brain oscillations, and the formation of more stable sleep-wake cycles with dominant slow waves during NREM.

Finally, in adults the relative entropy level across sleep stages is strongly time-scale dependent. At fast temporal scale EEG entropy follows a pattern of WAKE > REM > N2 > N3 whereas at slow temporal scale this pattern reverses (N3 > N2 > N1 > REM > WAKE) and is interpreted as a reflection of mainly local information processing during WAKE and increasingly global synchronization during NREM (reaching its culmination during N3) [15]. Our result shows that in neonates this pattern reverses partially. Both NREM and REM sleep show higher entropy than WAKE, however there is no significant difference between NREM and REM at the slow temporal scale which suggests that global connections may not be yet fully functional likely due to incomplete myelination of the newborns brain [44].

In addition to this descriptive analysis of baby EEG data in the first weeks of life, we developed a neonatal sleep classifier by employing machine learning using entropy-based features. We reveal that our classifier performs well over chance (60.1% week-2, 72.7% week-5) and is close to human scoring performance with adapted scoring rules (inter-scorer agreement of 80.6%, kappa score of 0.73 [12]). Crowell and colleagues [11] for example reported moderate inter-scorer reliability for infant sleep staging, with kappa coefficient going below 0.6. However it has to be noted, that visual sleep scoring of neonatal recordings is particularly difficult, even for experienced sleep experts. In order to reach an acceptable inter-scorer agreement intensive training and careful attention to scoring specifications (which even vary in the literature)
are required [11], which undermines the practical applicability of this approach. Note that even in adults the interrater agreement rarely exceeds 80% agreement [45]. As a matter of fact, 12 out 72 recordings in our sample have been classified as “difficult” by our scoring expert (Scholle S.;[30]). Considering the uncertainty in the ‘ground truth’, the obtained results are notably high. In fact after excluding the 12 “difficult” recordings the overall classification raises by about 6% which confirms presence of mislabeling in the dataset. The high classification accuracy of our automatic classifier also demonstrates that complexity measures are able to capture and quantify essential characteristics of vigilance states in newborns. Recently an independent group of researchers has used a similar approach to study sleep in preterm infants finding a significant correlation between the EEG complexity and the infant’s age (ranging from 27 to 42 weeks) [46]. In contrast to the aforementioned study (where sample entropy was used), our approach was based on robust permutation entropy. A limitation of our current approach is certainly that we did not yet include respiration (mainly due to practical reasons, e.g., to ensure a rapid recording-start in the newborns) and that our “ground truth” scoring is lacking a double scoring. Given that we included a pioneer (Scholle S.) in infant sleep staging which carefully reviewed all our baby PSG recordings (including simultaneous video recordings and Prechtl vigilance-scorings from the recordings) we believe that our manual scoring is as good as it can get for the moment. Future studies should still try to add respiration and independent second scorers which potentially would lead to further improvement of classification accuracy.

The significantly higher overall classification performance for week-5 as compared to week-2 recordings indicates that the neonatal sleep/wake states become more distinguishable within a very short time spanning only about 3 weeks of early human development. For across age-groups classification we observed overall decrease in accuracy as compared to within-group classification, which indicates marked "dissimilarity" in the sleep organization between the two sessions. It is worthy of note that although fully consistent with our univariate entropy results, the classification approach models simultaneously all temporal scales and all channels (including EMG, ECG, EOGs),
which provides a more complete picture on the entropy changes across time. Hence, to account for
the rapid developmental changes in newborns and thus further improve the automated sleep
classification, we suggest using ‘transfer learning’ procedure, known from recent application in deep-
learning. In this scenario a classifier could be pre-trained on larger dataset and further fine-tuned on
a specific age-group.

Interestingly testing on week-5 (when trained on week-2) is outperforming testing on week-2
(when trained on week-5) classification. This seems surprising as in case of cross-session (or -
condition) classification it is in general more efficient to train a classifier on data with clear class-
boundaries (high signal-to-noise; in our case week-5) and test it on data with less definite class-
boundaries (low signal-to-noise; in our case week-2) than the other way around [47]. In our case
however, this difference in signal-to-noise ratio has been counterbalanced by the sleep specific
entropy decrease from week-2 to week-5, which manifests itself as decreased performance in
classifying NREM (week-5 training, week-2 test). Interestingly there is no change for stage wake
when applying cross-session classification. This confirms that developmental changes in signal
complexity are more pronounced for NREM and REM states whereas stage wake may be (oscillatory-
wise) already widely developed.

In summarizing we observed massive developmental changes on the brain-level in the first 5 weeks
of life in human newborns. These changes were limited to “quiet” NREM and “active” REM sleep and
showed an unexpected drop of signal complexity from week-2 to week-5. In addition our classifier
data demonstrated that we can classify well above chance and similar to human scorers using multi-
scale permutation entropy (using just 6 EEG and 5 physiological channels). Altogether, these results
highlight the need to perform electrophysiological studies during the first weeks of life where rapid
changes in neuronal development and related brain activity can be observed.
Acknowledgments

We are especially grateful to S. Scholle for manual sleep staging and A. Lang for manual verification of all sleep scorings, with Prechtl scores and simultaneous video files.
References

Uncategorized References

1. Iber C, Ancoli-Israel S, Chesson A, Quan S. American Academy of Sleep Medicine. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Westchester: American Academy of Sleep Medicine; 2007.

2. Grigg-Damberger MM. The Visual Scoring of Sleep in Infants 0 to 2 Months of Age. J Clin Sleep Med. 2016;12(3):429-45.

3. Ellingson RJ. Electroencephalograms of normal, full-term newborns immediately after birth with observations on arousal and visual evoked responses. Electroencephalogr Clin Neurophysiol. 1958;10(1):31-50.

4. Whitehead K, Pressler R, Fabrizi L. Characteristics and clinical significance of delta brushes in the EEG of premature infants. Clin Neurophysiol Pract. 2017;2:12-8.

5. Anders TF, Emde T, Parmelee A. A manual of standardized terminology, techniques and criteria for scoring states of sleep and wakefulness in newborn infants. Los Angeles, CA: UCLA Brain Information Service, NINDS Neurological information Network. 1971.

6. Scholle S, Schäfer T. Atlas of states of sleep and wakefulness in infants and children. Somnologie – Schlafforschung und Schlafmedizin. 1999;3(4):163-241.

7. Anders TF, Keener MA, Kraemer H. Sleep-wake state organization, neonatal assessment and development in premature infants during the first year of life. II. Sleep. 1985;8(3):193-206.

8. Borghese IF, Minard KL, Thoman EB. Sleep rhythmicity in premature infants: implications for development status. Sleep. 1995;18(7):523-30.

9. Gertner S, Greenbaum CW, Sadeh A, Dolfin Z, Sirota L, Ben-Nun Y. Sleep-wake patterns in preterm infants and 6 month’s home environment: implications for early cognitive development. Early Hum Dev. 2002;68(2):93-102.

10. Freudigman KA, Thoman EB. Infant sleep during the first postnatal day: an opportunity for assessment of vulnerability. Pediatrics. 1993;92(3):373-9.

11. Crowell DH, Brooks LJ, Colton T, Corwin MJ, Hoppenbrouwers TT, Hunt CE, et al. Infant polysomnography: reliability. Collaborative Home Infant Monitoring Evaluation (CHIME) Steering Committee. Sleep. 1997;20(7):553-60.

12. Satomaa AL, Saarenpaa-Heikkila O, Paavonen EJ, Himanen SL. The adapted American Academy of Sleep Medicine sleep scoring criteria in one month old infants: A means to improve comparability? Clin Neurophysiol. 2016;127(2):1410-8.

13. Ma Y, Shi W, Peng CK, Yang AC. Nonlinear dynamical analysis of sleep electroencephalography using fractal and entropy approaches. Sleep Med Rev. 2018;37:85-93.

14. Jordan D, Stockmanns G, Kochs EF, Pilge S, Schneider G. Electroencephalographic order pattern analysis for the separation of consciousness and unconsciousness: an analysis of approximate entropy, permutation entropy, recurrence rate, and phase coupling of order recurrence plots. Anesthesiology. 2008;109(6):1014-22.
15. Miskovic V, MacDonald KJ, Rhodes LJ, Cote KA. Changes in EEG multiscale entropy and power-law frequency scaling during the human sleep cycle. Hum Brain Mapp. 2018.

16. Shen Y, Olbrich E, Achermann P, Meier PF. Dimensional complexity and spectral properties of the human sleep EEG. Electroencephalogs. Clin Neurophysiol. 2003;114(2):199-209.

17. Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Phys Rev Lett. 2002;88(17):174102.

18. Burioka N, Miyata M, Cornelissen G, Halberg F, Takeshima T, Kaplan DT, et al. Approximate entropy in the electroencephalogram during wake and sleep. Clin EEG Neurosci. 2005;36(1):21-4.

19. Bruce EN, Bruce MC, Vennelaganti S. Sample entropy tracks changes in electroencephalogram power spectrum with sleep state and aging. J Clin Neurophysiol. 2009;26(4):257-66.

20. Nicolaou N, Georgiou J. The use of permutation entropy to characterize sleep electroencephalograms. Clin EEG Neurosci. 2011;42(1):24-8.

21. Janjarasjitt S, Scher MS, Loparo KA. Nonlinear dynamical analysis of the neonatal EEG time series: the relationship between sleep state and complexity. Clin Neurophysiol. 2008;119(8):1812-23.

22. Wislowska M, Giudice RD, Lechinger J, Wielek T, Heib DP, Pitiot A, et al. Night and day variations of sleep in patients with disorders of consciousness. Sci Rep. 2017;7(1):266.

23. McIntosh AR, Kovacevic N, Itier RJ. Increased brain signal variability accompanies lower behavioral variability in development. PLoS Comput Biol. 2008;4(7):e1000106.

24. Waschke L, Westermann M, Obleser J. States and traits of neural irregularity in the age-varying human brain. Sci Rep. 2017;7(1):17381.

25. Zhang D, Ding H, Liu Y, Zhou C, Ding H, Ye D. Neurodevelopment in newborns: a sample entropy analysis of electroencephalogram. Physiol Meas. 2009;30(5):491-504.

26. Sterman MB, Harper RM, Havens B, Hoppenbrouwers T, McGinty DJ, Hodgman JE. Quantitative analysis of infant EEG development during quiet sleep. Electroencephalogr Clin Neurophysiol. 1977;43(3):371-85.

27. Scher MS. Neurophysiological assessment of brain function and maturation. II. A measure of brain dysmaturity in healthy preterm neonates. Pediatr Neurol. 1997;16(4):287-95.

28. Myers MM, Grieve PG, Izraelit A, Fifer WP, Isler JR, Darnall RA, et al. Developmental profiles of infant EEG: overlap with transient cortical circuits. Clin Neurophysiol. 2012;123(8):1502-11.

29. Scholle S, Zwacka G, Scholle HC. Sleep spindle evolution from infancy to adolescence. Clin Neurophysiol. 2007;118(7):1525-31.

30. Scholle S, Feldmann-Ulrich E. Polysomnographic atlas of sleep-wake states during development from infancy to adolescence. Landsberg, Germany: Ecomed Medizin. 2012.
31. Einspieler C, Prechtl HF. Prechtl’s assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. Ment Retard Dev Disabil Res Rev. 2005;11(1):61-7.

32. Costa M, Goldberger AL, Peng CK. Multiscale entropy analysis of biological signals. Phys Rev E Stat Nonlin Soft Matter Phys. 2005;71(2 Pt 1):021906.

33. Bates D, Maechler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015;67(1):1-48.

34. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria 2013.

35. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biom J. 2008;50(3):346-63.

36. Jamalabadi H, Alizadeh S, Schonauer M, Leibold C, Gais S. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum Brain Mapp. 2016;37(5):1842-55.

37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

38. Khazipov R, Luhmann HJ. Early patterns of electrical activity in the developing cerebral cortex of humans and rodents. Trends Neurosci. 2006;29(7):414-8.

39. Jenni OG, Borbely AA, Achermann P. Development of the nocturnal sleep electroencephalogram in human infants. Am J Physiol Regul Integr Comp Physiol. 2004;286(3):R528-38.

40. Vanhatalo S, Palva JM, Andersson S, Rivera C, Voipio J, Kaila K. Slow endogenous activity transients and developmental expression of K+-Cl-cotransporter 2 in the immature human cortex. Eur J Neurosci. 2005;22(11):2799-804.

41. Klimesch W. The frequency architecture of brain and brain body oscillations: an analysis. Eur J Neurosci. 2018;48(7):2431-53.

42. Klimesch W, Sauseng P, Hanslmayr S, Gruber W, Freunberger R. Event-related phase reorganization may explain evoked neural dynamics. Neurosci Biobehav Rev. 2007;31(7):1003-16.

43. Leinekugel X, Khazipov R, Cannon R, Hirase H, Ben-Ari Y, Buzsaki G. Correlated bursts of activity in the neonatal hippocampus in vivo. Science. 2002;296(5575):2049-52.

44. McArdle CB, Richardson CJ, Nicholas DA, Mirfakhraee M, Hayden CK, Amparo EG. Developmental features of the neonatal brain: MR imaging. Part I. Gray-white matter differentiation and myelination. Radiology. 1987;162(1 Pt 1):223-9.

45. Magalang UJ, Chen NH, Cistulli PA, Fedson AC, Gislason T, Hillman D, et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep. 2013;36(4):591-6.

46. Wel OD, Lavanga M, Dorado A, Jansen K, Dereymaeker A, Naulaers G, et al. Complexity Analysis of Neonatal EEG Using Multiscale Entropy: Applications in Brain Maturation and Sleep Stage Classification. Entropy. 2017;19(10):516.
47. King JR, Dehaene S. Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn Sci. 2014;18(4):203-10

Supporting Information

S1 Fig. Influence of the feature extraction procedure on the classification performance. Multiscale permutation entropy as compared to PSD boosts discrimination of sleep stages. MSPE improves classification especially of WAKE, also REM and slightly NREM class (note the diagonals of the lower panels).

S2 Fig. Cross-validation schemes. Splitting into training (in blue) and testing (in green) sets was performed within sessions (upper row) or across sessions (lower row). Note, that half of the subjects are used to train and half to test (two-fold cross validation), with both sessions of a single subject (week 2 and 5) always being in separate sets.

S3 Fig. Confusion matrices summarizing the classification results for week-2 and week-5 (MSPE-based classifier). Note the off-diagonals showing limited proportion of NREM falsely classified as WAKE (on average 3%) and similarly WAKE falsely classified as NREM (on average 8%).

S4 Fig. Comparison of MSPE at fast time scale between sleep stages and the two recording sessions (week-2 vs week-5 data) — frontal and occipital channels. Note that difference between REM and WAKE at week-5 is more pronounced at the frontal channels.

S5 Fig. PSD delta 4-11Hz. No significant differences were found for power in frequency band of 4-11Hz.

S6 Fig. Comparison of multi-scale permutation entropy values across different scales between sleep stages and sessions. Points represent the averages and error bars show the 95% bootstrap confidence intervals. Note that results for scale=1 and scale=5 (X-axis) correspond to the results presented in the main text. Maximal entropy in NREM and WAKE (also REM) is attained at different temporal scale (X-axis). Also, similarly to the main text there is no statistical difference in WAKE between sessions (week-2 vs week-5) across all temporal scale.

S7 Fig. Exemplary single-subject classification (week-5). Note the limited proportion of NREM epochs being falsely classified as WAKE and similarly WAKE classified as NREM.

S8 Fig. Channel importance extracted from a trained random forest classifier and a decision boundaries of a trained random forest classifier. Horizontal EOGs, frontal channels as well ECG contribute most to the sleep classification (upper panel). For visualization purposes multidimensional scaling was used to reduce dimension of the MSPE data to two (lower panel, X and Y axis). Points represent epochs (N=100 for each class), colors (red, green and blue) represent true class labels and shading (pink, light blue and light grey) shows the decision boundary. Note that in week-5 (right panel) there is more apparent overlap between the true class labels (points) and the predictions (shading) as compared to week-2, which agrees with higher classification accuracies for week-5.
Fig1 main text
F1 = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}
Fig3 main text
Fig 4 main text
Fig 5 main text