ESTIMATION OF HUMAN CIRCADIAN PHASE VIA A MULTI-CHANNEL AMBULATORY MONITORING SYSTEM AND A MULTIPLE REGRESSION MODEL

Vitaliy Kolodyazhniy, Jakub Späti, Sylvia Frey, Thomas Götz,
Anna Wirz-Justice, Kurt Kräuchi, Christian Cajochen, and Frank H. Wilhelm

SUPPLEMENTARY ONLINE MATERIAL

Sleep logs

The subjects were instructed to fill out the sleep logs carefully and to keep their pre-defined bed times which they planned themselves and communicated to the study supervisor several days before the experiment, separately for each day of the ambulatory phase taking into account both their habitual wake times on weekends and workdays and their concrete plans for the ambulatory week. For the analysis presented in this paper, the times of switching the lights off before sleep and the wake-up times were used. Filling out of the sleep log was scheduled 10 minutes after the planned wake time for each day. Subjects were notified to fill out the sleep log by a beep tone from the electronic diary, repeating itself every five minutes until the sleep log was filled out. compliance in filling out the electronic sleep log was checked by the time stamps of the data entries, as well as by rest/activity and light data. Subjects were informed of these compliance checks and did not significantly deviate from their self-planned bed times.

Checking and transferring data

For the ClockWatcher and VarioPort devices, the batteries needed to be replaced in the middle of the ambulatory part (3 or 4 days after the start) and at its end. The replacement of the

batteries was done by the study supervisor. At the same time, the data were downloaded for inspection and backup, although the memory cards had enough capacity to store the data for the entire protocol (about 1 GB). At the same time the data from the light sensor were checked. The entire procedure including battery replacement took about 20 minutes. The data were transferred and inspected using a portable computer in the presence of the test subject.

The sleep diary was checked before the CR in the laboratory and used for scheduling the CR according to the sleep midpoint during the ambulatory week.

Import of data from all devices including the skin temperature sensors into the database for analysis was done after the end of the entire protocol for each subject.

Circadian software toolbox and database

For multiple regression modeling based on data from multiple multi-channel ambulatory monitoring devices it is essential that the data from heterogeneous sources are synchronized and properly pre-processed. For storing the pre-processed data, a database running on MySQL Server v. 5.0 was created.

Import of the study data into the database, preparation of the data for the analysis, and the analysis including the multiple regression modeling approach were done using a specially developed circadian software toolbox running under the general-purpose numerical computing environment MATLAB v. 6.5.

Recordings from 'slow' data sources with sampling periods of one minute or more were imported into the database without any pre-processing (iButtons, LightWatcher, E-diary, melatonin levels). The data from the ClockWatcher and VarioPort devices with sampling rates up to 512 Hz were pre-processed prior to importing them into the database. In the pre-processing

step, the sampling periods for all channels were reduced to 30 seconds after some transformation of the recorded signals, primarily the following: heart rate and respiratory rate were derived from the respective raw signals (VarioPort recordings already contained a heart rate channel) and accelerations along axes X, Y, and Z were divided into the respective motion and posture components, and leg movement was computed from the respective 1D accelerometer.

Motion and posture variables were computed as follows. First, the low-frequency components X_{post}, Y_{post}, and Z_{post} corresponding to posture were computed from the respective acceleration channels X_{acc}, Y_{acc}, and Z_{acc} using a second-order IIR filter with a 3 dB cutoff at 0.15 Hz. Then the high frequency components corresponding to motion were then extracted via rectifying the difference of the acceleration and posture:

$$X_{mov} = abs(X_{acc} - X_{post})/2,\ Y_{mov} = abs(Y_{acc} - Y_{post})/2,\ Z_{mov} = abs(Z_{acc} - Z_{post})/2.$$

Leg movement was computed in the same way.

An integrated variable for motion was also computed as square root of sum of squared values of motion along the respective coordinates: $M = \sqrt{X_{mov}^2 + Y_{mov}^2 + Z_{mov}^2}$.

Via downsampling the original high sampling rate to the sampling period of 30 seconds, the overall amount of data was dramatically reduced from more than 1 gigabyte per subject to less than 10 megabytes. At the same time, the sampling period of 30 seconds allowed us to visually detect CBT probe slips as a fast decline in temperature and edit them using the graphical user interface of our software.

The software uses real time and date for all data records in the database such that all time stamps from multiple heterogeneous data sources (including multi-channel physiological recordings, electronic diaries, and melatonin assays) are synchronized. The data imported into the database are retrieved by the software and resampled ‘on the fly’ to an arbitrarily common

sampling period that can be chosen from 30 sec to 6 h independent of the original sampling rate of the raw data or the preprocessed data in the database. This provides convenient means for multivariate modeling and visualization.

Data preparation for regression modeling

The following steps of data preparation were performed with the data previously imported into the database to form the data set for multiple regression modeling:

1) *Skin temperatures.* For skin temperatures measured with iButtons, six variables were produced from the eleven sensors: averaging of the left- and right-hand side sensors (except for thorax) resulted in skin temperature variables for *shoulders*, *hands*, *feet*, *thorax*, *upper legs*, and *lower legs*. Temperature data were detrended by subtracting from each temperature channel its moving average with a window of 24h. The moving average was computed by averaging the data in the interval of ±12 hours from the current point (i.e. from the center of the moving window) so that no phase distortion was introduced. Detrended data were z-transformed for each participant. In the z-transformed data, all values outside the range of [-2, 2], i.e. ± two standard deviations, were discarded as outliers and the gaps were interpolated using cubic polynomials. For details on interpolation, also for other measured variables mentioned in items 3) through 8) below, see subsection '*Interpolation of missing data*'. Additionally, the distal-proximal skin temperature gradient (DPG) was computed.
2) *Ambient light.* Irradiance values below 0.01 in the five spectral bands were replaced with 0.01, and the variable was log10-transformed to accommodate the very broad range of irradiance between darkness and bright sunlight. Additionally, average irradiance of all

spectral bands and of the three visual spectral bands was computed for every data point of the irradiance data.

3) *Motion.* Only the integrated variable for motion along axes X, Y, and Z was used. Missing values were interpolated using cubic polynomials.
4) *Leg movement.* Missing values were interpolated using cubic polynomials.
5) *Posture.* The vertical component (axis Y) was used to allow for differentiation between the supine and standing position. Missing values were interpolated using cubic polynomials.
6) *Heart rate.* Any values outside the physiologically plausible range of 40...200 beats per minute were discarded and all missing values were interpolated using cubic polynomials.
7) *Respiratory rate*. Data outside the physiologically plausible range of 6...60 breaths per minute were discarded and missing values were interpolated using cubic polynomials.
8) *CBT.* This variable was edited for probe slips. Missing values were interpolated using cubic polynomials. Probe slips were determined as a decline in the temperature within several minutes below the CBT minimum of the entire recording.
9) *Melatonin.* A periodic BSBCF waveform was extracted from the CR-derived melatonin rhythm for each subject and extrapolated backwards onto the 7 days of the ambulatory part preceding the in-lab experiment. Then, after scaling to the range of [0, 1] and adding normally distributed random noise with a standard deviation of 0.01, this variable was used as the target for fitting a multiple regression model with input variables from the multi-channel recordings of both the ambulatory and laboratory parts of the experiment that were 9 days long for each subject. The scaling of the BSBCF waveform was performed in order to eliminate individual differences in levels of melatonin secretion

which complicate the development of a subject-independent circadian rhythm model, and the random noise improved model identification.

To model a realistic situation of real-world application, no recordings that were used in the analysis were manually edited, i.e. the data that were included in the analysis were automatically processed raw data. The only editing was of the CBT recordings which were used for comparison of prediction of circadian phase with the regression model and were used neither as a predictor nor as the dependent variable in the model itself.

All channels for each subject were aligned based on the respective time stamps and the data were further resampled to 30 minute bins. To avoid phase distortion in resampling, the time stamp of all bins on the time axis corresponded to the centre of the bin. From each subject, data from both the ambulatory and laboratory parts were used for fitting a prediction model using multiple regression techniques and the least squares method.

Interpolation of missing data

For interpolation of missing points in data from the monitoring devices, piecewise cubic Hermite interpolating polynomials were used (MATLAB function 'interp1' with method 'cubic'). The polynomials use 3 datapoints on each side of the gap to fit the interpolating curve. In our software, we set the threshold for the maximum length of interpolated gaps to 12h. However, most of the gaps were only several minutes long and occurred when the subjects took shower or during replacement of the batteries and memory cards in the ClockWatcher and VarioPort devices.

Waveform analysis

The bimodal skewed baseline cosine function (BSBCF, Van Someren & Nagtegaal, 2007) is defined as follows:

$$f(t) = b + \frac{H}{2(1-c)}(\cos(t-\phi+\nu\cos(t-\phi)) + m\cos(2t-2\phi-\pi) - c + |\cos(t-\phi+\nu\cos(t-\phi)) + m\cos(2t-2\phi-\pi) - c|),$$

where t is time (in radians, $2\pi = 24$ hours), b is the baseline, H is the peak level (>0), c is the peak width (≥–1 and <1), ϕ is phase (in radians, 0 to 2π), ν is skewness (between ≥–0.5 and ≤0.5), and m is bimodality (≥0 and <1).

Fitting of the BSBCF curve was done using the MATLAB optimization toolbox using data with a sampling period of 30 minutes. The fitted data were either melatonin data or the predicted circadian rhythm from the regression models. Hourly melatonin data were resampled to match the timestamps of other data channels first by upsampling to one minute sampling rate using linear interpolation and then by downsampling to 30 minute bins. An example of the curve fitting is shown in Fig. S1.

Calculation of circadian phase

Circadian phase was determined as time corresponding to the centre of gravity (COG) of area (Wetterberg, 1998) under the periodic BSBCF curve for one period of 24h

$$COG = \frac{\sum_{t=t_{start}}^{t_{start}+24} t \cdot (f(t) - b)}{\sum_{t=t_{start}}^{t_{start}+24} (f(t) - b)},$$

where t is time in hours, b is the baseline found from fitting the BSBCF function, and t_{start} is the time point when the 24h period starts (Fig. S1).

The start of the 24h period (t_{start}) was found as the first time point where the BSBCF curve differed from the baseline. For precise computing of the circadian phase, the fitted BSBCF curve was resampled with a sampling period of one minute. The combination of the BSBCF and COG approaches can be interpreted as average circadian phase for each subject, i.e. for 32h of the reference melatonin data (1.33 days with two offsets and one onset) or for the 6-day long prediction of the circadian rhythm from the multiple regression model.

Cross-validation

Prediction of circadian phase with multiple regression models was based on the subject-independent cross-validation approach, where data from 15 subjects out of 16 were used for identifying a model, and data from another subject was used for validation. In the first iteration subjects N2-N15 were used for model identification, and subject N1 for validation, in the second iteration subjects N1, N3-N15 for model identification, and subject N2 for validation, and so on (see Fig. S2). Thus, in each of the iterations one model was identified using data from all but one subject and validated with the data from that subject "unseen" during model identification, i.e. *there were 16 fits with 15 subjects each validated against an unknown subject*. The approach to subject-independent cross-validation was similar to that widely accepted in the literature, e.g. in (Ho et al., 2009; Howard et al., 2009; Zhao & Lu, 2005).

For model identification, data with both ambulatory and laboratory parts was used, whereas validation was done only with ambulatory data of the respective subject in order to provide a realistic estimate of circadian phase in real-life conditions. Special care was taken to ensure that no data from the CR and the adaptation night was included into the data set for validation, also taking into account the detrending method for skin temperatures described above. Thus, the

predictions of the ambulatory circadian phase started 24h after the protocol start due to the maximum lag of 24h in the prediction model, and ended 12h before the adaptation night in the laboratory due to the detrending method for skin temperatures.

In each iteration of cross-validation the data were re-standardized to simulate a real situation where unknown observations with unknown means and standard deviations arrive and are used for prediction of circadian phase based on the statistical characteristics of the known data that were used for fitting the regression model. The data standardization in each of cross-validation iterations was done as follows: first, means and standard deviations were computed for the respective variables in the entire data set used for fitting the model comprising data from fifteen subjects. Then these means and standard deviations were used to perform z-transformation of both the data used for fitting the model (fifteen subjects) and of the validation data from another subject.

Variable selection

Usually, the cross-validation approach is combined with a so-called feature selection algorithm which iteratively finds a combination of input variables providing the best performance. However, the number of variables involved in our regression modeling multiplied by the number of their possible lags made the time requirements for a feature selection procedure prohibitively large: our final model contained 164 predictors (8 variables with their respective lags). Finding such a subset of predictors from an even larger set of measured variables with their possible lags using a feature selection algorithm was not feasible taking into account the time that is required for the model to be evaluated with every possible combination of predictors.

Instead, we first used an initial set of handpicked variables with pre-defined lags, and modified that selection based on iterative tests guided by the accuracy of prediction of circadian phase (standard deviation of prediction error). Initially, we included skin temperatures, blue light, motion, and posture. For the skin temperatures, there was no lag initially, and for light, motion, and posture the lags were up to 24h. The reasons for this choice were as follows:

1) skin temperatures were meant to provide the information of the endogenous circadian phase and were the physiological parameters that were the most easy to measure with the non-invasive tiny sensors (iButtons®) – this would be an advantage for the final model; zero lag would help us reveal the current levels of the skin temperatures with respect to the daily range (the temperatures were detrended as described in the subsection '*Data preparation for regression modelling*');

2) light is the external factor that determines entrainment of the circadian clock; we chose the blue spectral band as the most important for entrainment of the circadian clock (Lockley et al. 2003; Cajochen et al., 2005; Smith et al., 2009) and the light history equal to the length of external day/night cycle, i.e. lags up to 24h;

3) motion provides information about the phase of entrainment and at the same time helps unmask the skin temperatures; we chose lags of up to 24h as a usual period of rest/activity;

4) posture – same reasoning as for motion, lags up to 24h.

Based on standard deviation of prediction error in the cross-validation setting, we found out that the lags for skin temperatures should be 0 to 5h (this can be interpreted as introducing filtering properties to smooth out the noise in the data and irrelevant fluctuations due to masking) and that the posture variable should be excluded from the model (the relevant information on rest/activity information is contained in the motion variable). Thus, the final model structure

presented in the paper was obtained, that included skin temperatures, blue light, and motion as predictors (see subsection '*Final model structure*' of 'Materials and methods').

Inclusion of any other variables in the multiple regression models did not result in improvement of prediction accuracy. Contrary to that, the prediction was worse. A possible reason for the worsening of the prediction accuracy with a larger number of predictor variables is the nonlinear relation of the prediction accuracy used as the criterion for variable selection to the goodness of fit of the least squares regression: the best fit with the smallest sum of squared residuals does not necessarily correspond to the best prediction accuracy which is decided based on the BSBCF and center of gravity approaches. All the other variables measured by the wearable ambulatory devices and derived from the measurements were iteratively tried one by one with lags of 0 to 24h: distal-proximal gradient, heart rate, respiratory rate, leg movement, light from the spectral bands other than blue, as well as light summed up from all spectral bands and the three visible bands (red, green, blue). The lack of contribution of heart rate and respiratory rate to the final model was likely due to the severe masking of these variables, primarily by the sleep-wake cycle and physical activity.

Testing the model with a reduction in the number of temperature variables like feet, shoulder, hands, etc, showed that all of them appeared to be important. However, given that six temperature variables were used and eleven sensors, the doubled right-left measurements could probably be eliminated.

All these attempts confirmed the optimal fit of the model described in the paper (subsection '*Final model structure*' of 'Materials and methods').

References

Ho TC, Brown S, Serences JT (2009) Domain general mechanisms of perceptual decision making in human cortex. J Neurosci 29:8675-87.

Howard J, Plailly J, Grueschow M, Haynes JD, Gottfried JA (2009) Odor quality coding and categorization in human posterior piriform cortex. Nature Neurosci 12:932-8.

Van Someren EJW, Nagtegaal E (2007). Improving melatonin circadian phase estimates. Sleep Med 8:590-601.

Wetterberg L (1998) Melatonin in adult depression. In *Melatonin in Psychiatric and Neoplastic Disorders*, M Safii, SL Safii (eds), 1st ed., pp. 43-80, American Psychiatric Press.

Zhao Q., Lu H (2005) GA-driven LDA in KPCA space for facial expression recognition. In *Lecture Notes in Computer Science, vol. 3611*, L Wang, K Chen, YS Ong, (eds), pp. 28-36, Springer-Verlag, Berlin.

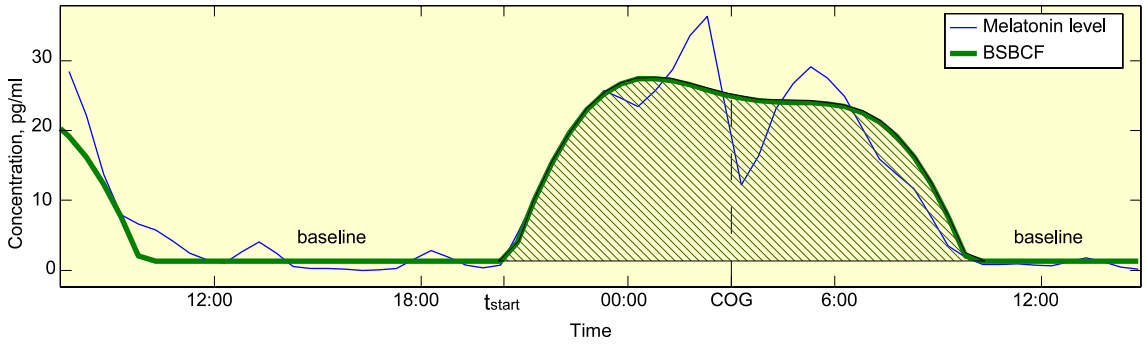

Figure S1. Determining circadian phase from CR melatonin using bimodal skewed baseline cosine function (BSBCF) and center of gravity (COG) of area under BSBCF curve for one period (shaded area in the plot): t_{start} is the first time point where the BSBCF curve differs from the baseline and marks the beginning of the interval for computing the COG, baseline is found from fitting the BSBCF curve.

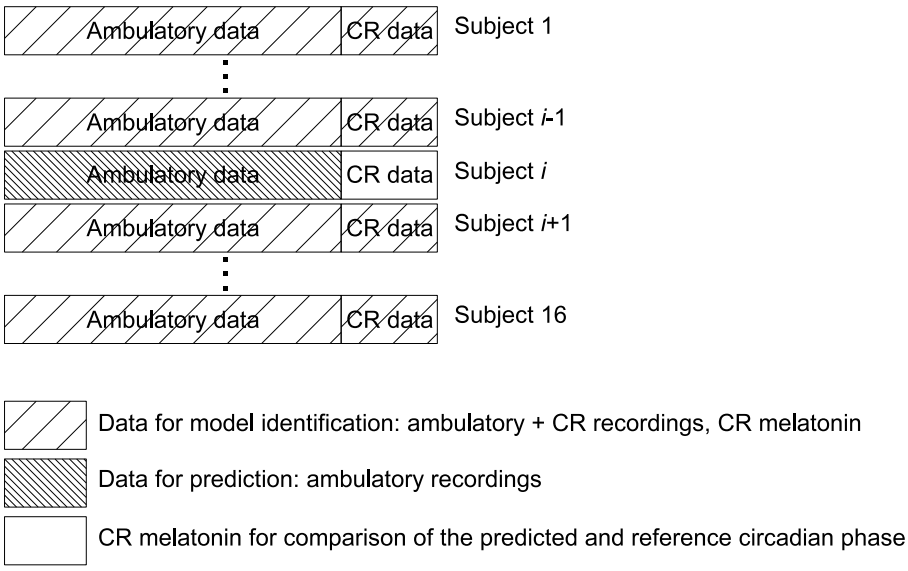

Figure S2. Data split for 16-fold cross-validation with 16 subjects: for subject number *i* (*i*=1...16), complete ambulatory and CR data from the other 15 subjects are used for identification of the *i*-th prediction model, and the *i*-th subject's own ambulatory data and CR melatonin are used for validation of the *i*-th model.