Supplementary material: "The Sleep Well Baby project: an automated real-time sleep-wake state prediction algorithm in preterm infants"

Thom Sentner†, Xiaowan Wang‡, Eline R. de Groot, Lieke van Schaijk, Maria Luisa Tataranno, Daniel C. Vijlbrief, Manon J.N.L. Benders, Richard Bartels†, Jeroen Dudink*  

1 Digital Health, University Medical Center Utrecht, Utrecht, The Netherlands  
2 Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, The Netherlands  
3 Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

† Shared Co-first authors  
‡ Shared Co-last authors  
* Corresponding author

1 Inter-patient variability

Figure S1. Parameter distributions of the patients in the training dataset for heart rate, respiratory rate, and oxygen saturation. Distributions highlighted in black illustrate the most extreme differences between patients.  

Figure S1 illustrates that apparent differences exist between parameter distributions of individual patients. To correct for interpatient variability, reference data were collected from a 24-hour reference period, which can then be used to scale data using a standard scalar (Equation S1). This allows the model to look at changes relative to an individual its baseline physiological state rather than consider absolute differences.
Equation S1. Normalization of observation period values ($O$) based on (corrected) reference period values ($\bar{R}$), applied per parameter.

$$\tilde{O}(O, \bar{R}) = \frac{O - \bar{R}}{\sigma(\bar{R})}$$

For a bedside implementation, reference data will always be collected for the 24-hours preceding the prediction. However, for the patients in the train and validation datasets, 24-hours of consecutive data was not always available right before the observation period. As such, reference data nearest to the observation period (before or after) was collected. To explore the effect of this time gap in data collection, the absolute difference in modes between the distributions during the observation and reference period was compared (Equation S2).

Equation S2. Absolute mode ratio.

$$\text{absmoderatio}(A, B) = \begin{cases} \frac{\hat{a}}{\hat{b}} & \hat{b} < \hat{a} \\ \frac{\hat{b}}{\hat{a}} & \hat{a} \leq \hat{b} \end{cases} \text{ where } \hat{x} \text{ is the mode of } X = \{…\}$$

Note that mode was considered, rather than the median or mean, as the parameter distributions are strongly affected by the ratio of wake during any period, to which the mode is relatively robust since it is expected that a patient spends most time in the sleep stages. As such, the mode provides an anchor to compare distributions regardless of period length.

It was found that increasing time differences between the observation period and reference period correlated with increasing differences in the modes of the parameter distributions (Figure S2). Large parameter value differences were especially prevalent at a time delta of twelve hours or more. Patients in the training set where the reference period differed by more than 2.5 days from the observation period were removed (7 patients).

![Figure S2](image-url)

Figure S2. Relation between the time gap between reference and observation period and similarity of reference and observation period, as measured by the ratio of their modes (Equation S2). Black line and gray fill illustrate a linear regression model and confidence interval fit to the relationship. Training-set patients (blue) with an absolute time delta of 2.5 days or larger were excluded (7 right-most dots). Differences for validation-set patients (orange) were much
smaller. Gray dashed lines highlight the intervals of [0, 0.5), [0.5, 2.5) and [2.5, \infty) days between the collection of reference data and the observations. Whereas the first interval mostly includes points with a high similarity between the reference and observation period distributions, the latter interval does not include any points with similar distributions between the periods.

Distributions of heart- and respiratory-rate reference data for all remaining patients in the train set were re-scaled such that its mode matches that of the observation data (Equation S3). This mimics the effect of the data being collected directly prior to the observation period. Validation patients were omitted from this correction, as time deltas between the observation period and reference period were small and as such a correction differs from the envisioned bedside implementation. After application of Equation S3, inter-patient variability can be corrected for by applying a standard-scalar (Equation S1)

Equation S3. Mode correction of reference period parameter values \( R \) using values from the observation period \( O \). Applied for heart- and respiratory rate.

\[
\hat{R}(O, R) = \frac{R}{\text{mode}(R)/\text{mode}(O)}
\]

1.1 Validation mode re-scaling
To confirm the concept of mode re-scaling (Equation S3), synthetic data distributions were used. Synthetic data was produced for hypothetical patients that were awake 0% and 25% of the observation period, and 10% of the 24-hour reference period. An average of about 10% wakefulness during the day corresponds to the mean ratio of wake observed in our training dataset. Reference data values were multiplied by a factor of 1.3 to simulate a patient whose reference data are unsimilar to the observation period data, e.g., due to a time shift (Figure S3). When next applying the mode correction of reference values as was also done for train patients (Equation S3), one may visually confirm that the reference data distributions is similar to that of the observation data, while maintaining the characteristic skew due to wakefulness, as hypothesized above.
Figure S3. Illustration of mode correction on reference data (right-most panel, blue and grey lines) for synthetic data of patients that were awake 0% (upper left panel) and 25% (bottom left panel) of the observation period. Both hypothetical patients were assumed to be awake 10% of the time during the 24-hour reference period (middle panels). Mean and mode values of the observation and reference period are highlighted with solid and dashed vertical lines, respectively.

To confirm similar effects in actual patient data, similar graphs were produced for patients who were awake 31%, 8% and 0% during the observation period (Figure S4). In contrast to the synthetic data, the ratio of wake during the reference period is unknown, as no sleep-state observations were taken. Still, it is evident that the correction steps happen practically equal to the synthetic data, establishing its application.
Figure S4. Illustration of mode correction of reference data (right panels, blue and grey lines) for example patients that were awake 31% (upper panels), 8% (middle panels) and 0% (lower panels) of observation period (left panels). Observed sleep-wake states were unavailable for the reference period (middle panels). Mean and mode values of the observation and reference period are highlighted with solid and dashed vertical lines, respectively. Note that in the middle panels there is a large deviation between observation and reference period distributions. In the bottom right panel the difference is so small that the black and blue lines overlap.

The effect of this correction and the resultant interpatient normalization on model performance is set out in Figure S5. In comparison to no normalization, model performance is clearly improved when applying interpatient variability correction based on the reference period, but only when applied in combination with the mode-rescaling procedure.
Figure S5. Effect of interpatient variability correction and reference period mode correction on model performance, calculated using the nested cross-validation procedure on the training dataset. Error bars highlight standard deviation over outer splits.

2 Hyperparameter grid

The hyperparameters that were explored in the inner-loop of the nested cross-validation procedure are listed in Table S1. Any remaining hyperparameters were set to their default values in scikit-learn.

Table S1. Hyperparameter grids used during model training. Underlined hyperparameters were selected for the final model.

| Hyperparameter        | Random forest | Logistic regression | Decision tree |
|-----------------------|---------------|---------------------|---------------|
| Class weight          | balanced      | balanced            | balanced      |
| Number of estimators  | 250           | -                   | -             |
| Max depth             | 2, 3, 5, 20, 50 | -                   | 2, 3, 5, 20, 50|
| Min samples split     | 2, 4.2%, 8.3%, 16.7%, 33.3% | - | 2, 4.2%, 8.3%, 16.7%, 33.3% |
| Min samples leaf      | 1, 0.4%, 3%, 6% | -                   | 1, 0.4%, 3%, 6% |
| Max features          | 5%, 10%, 20%, 40% | -                   | 2, 3, 5, 20, 50 |
| C                     | -             | $10^{-4}$, $10^{-3.5}$, …, $10^{3.5}$, $10^{4}$ | -             |

The effect of this hyperparameter grid on train and test split is visualized for all models in Figure S6. For the decision tree classifier, it is evident that hyperparameters resulting in a high test score correspond to a relative lack of overfitting, with train scores being comparable to the test scores. For logistic regression, the grid has little effect on model performance. For the random forest, there exists substantial variance in train scores, whilst the test scores are almost completely unaffected by the hyperparameter settings. Despite these minimal differences, models that are most overfit result in the highest test scores. Therefore, conservative and less overfit hyperparameters were selected (Table S1 and Figure S6) using the elbow method, similar to what is done in clustering. This ensures a higher chance of generalization to novel populations.
3 Patient demographics

The results of the statistical tests to compare the train and validation datasets are summarized in Table S1.

Table S2. Statistical analyses on patient demographics observation and parameter data, between train and validation datasets. P-values of post hoc analyses were corrected with Bonferroni correction. (DOF = degrees of freedom; NA = not applicable; ** = p ≤ 0.01; *** = p ≤ 0.001)

| Variable                                      | Test statistic(s) | DOF | p-value |
|-----------------------------------------------|-------------------|-----|---------|
| Gender                                        | $\chi^2$          | 0.02| 1       | 0.88    |
| Delivery method                               | $\chi^2$          | 3.58| 2       | 0.17    |
| Birth weight                                  | t                 | -0.41| 37     | 0.68    |
| Multiple birth                                | $\chi^2$          | 2.48| 1       | 0.12    |
| Apgar score - at 1 min                        | Z                 | 0.25| NA      | 0.80    |
| Apgar score - at 5 min                        | Z                 | 1.81| NA      | 0.07    |
| Apgar score - at 10 min                       | Z                 | 1.09| NA      | 0.27    |
| Gestational age                               | t                 | -0.15| 37     | 0.88    |
| Postmenstrual age at study                    | t                 | 0.49| 38      | 0.63    |
| Observation period length per patient         | Z                 | -0.81| NA     | 0.42    |
| Confidence scores of observed sleep-wake states| Wilk's $\Lambda$   | 0.62| 3; 35   | 0.0007***|
| - Confidence score -1 (post hoc)              | F                 | 7.14| 1       | 0.005** |
| - Confidence score 0 (post hoc)               | F                 | 11.32| 1      | 0.0005***|
| - Confidence score 1 (post hoc)               | F                 | 17.21| 1      | 0.24    |
| Observed sleep-wake states with confidence score ≥0 | Wilk's $\Lambda$ | 3.23| 1       | 0.88    |
| Observed sleep-wake states with confidence score ≥0 | F                 | 1.17| 4; 35   | 0.34    |
4 Additional model performance analyses

4.1 Confusion matrix
Figure S7 shows the confusion matrix for the reference model described in the main text. It illustrates that wake is unlikely to be confused with quiet sleep and that active sleep is predicted with the highest precision of the sleep-wake states, at 75%.

![Confusion matrix](image)

Figure S7. Confusion matrixes of observed sleep-wake states vs predicted sleep-wake states (W = Wake, AS = Active Sleep, QS = Quiet Sleep) calculated for the random forest classifier using (A) the regular cross-validation procedure on the training dataset and (B) the validation dataset. Percentages and respective coloring add to 100% per row.

4.2 Feature importance
Figure S8 explores feature importance by looking at the performance of models trained using only a subset of the available vital parameters. A model based solely on HR-related features approaches the performance of a model including all parameters.

![Feature importance](image)

Figure S8. AUROC (A) and F1 (B) of Wake from nested cross-validation for the random forest model using subsets of the various parameters (OS = oxygen saturation, RR = respiratory rate, HR = heart rate). Error bars highlight standard deviation over outer cross-validation splits.
To assess the added value of each time window over which features are calculated, models using only a single time window were compared to the full model (Figure S9). Of the individual time windows, it is evident that 60 seconds is most predictive. Nevertheless, model performance improves when all time windows are included, indicating that the model can exploit some of the time series information these time windows collectively encode.

Figure S9. AUROC from nested-cross validation for the random forest model for features based on individual time windows, and the full model (All). Error bars highlight standard deviation over outer cross validation splits.

4.3 Dependence of model performance on training-set size

In Figure S10, model performance dependence on the number of training patients is studied. For this purpose, a number $n$ patients are randomly selected from the training data and a random forest is trained following the procedure outlined in the main paper. As can be seen, model performance quickly reaches a plateau at around an AUROC of 0.76. However, variance on the model performance greatly decreases with increased dataset size. This plot indicates that including a much larger sample size will not significantly improve our current model.

Figure S10. Macro-averaged AUROC for different training set sizes. Patients were randomly excluded from the training dataset prior to model training procedure. Boxplots illustrate variance of model performance over 10 iterations.
4.4 Qualitative analysis of model performance

Other than the quantitative results discussed in the paper, it is also worthwhile to look at the qualitative behavior of the model predictions. Of note is that although incorrect predictions may be prevalent for a patient in their observation period, global trends are often captured correctly. An example patient is plotted in Figure S11, whose quiet sleep ranges are correctly predicted if allowing for a few minutes offset. Furthermore, for this patient a single minute of wake was observed at the end of the observation period, which was predicted a minute late by the model. Although this wake prediction was formally incorrect, it would not hamper clinical useability.

Figure S11. Observed (top) and predicted (bottom) sleep-wake states (Active Sleep, red; Quiet Sleep, orange; Wake, green; Missing observations, white) per one-minute window for an example patient from the training dataset. As an example feature, the median heart rate during the last 120 seconds after interpatient correction is shown as a black line (right y-axis).

5 Post-hoc analysis of the validation dataset

Performance on the validation set was slightly worse than what was found from cross-validation of the training data. In particular, the calibration curve (main text, Figure 3) shows a disparity between the training and validation data. Several causes for this are discussed below.

First, it should be noted that only 10 observation runs were included in the validation data. Since one-minute windows of a single observation run are correlated, it is possible that a single or a small number of patients for whom predictions are off negatively affect the calibration results. In Figure S12, the calibration for the validation dataset is shown when leaving out the two worst performing patients according to their individual one-versus-one macro-averaged AUCs. For these patients the total amount of wake was vastly overestimated whereas no wake was observed during the observation period. A larger validation cohort could further improve model calibration, since isotonic calibration is known to improve with larger sample sizes.
Figure S11. Calibration curve (left) and total predicted vs. observed time spent in sleep-wake states for individual patients (right) for the validation data when after leaving out the two worst performing patients based on their one-versus-one macro-averaged AUCs.

Overprediction of wake can be caused by a discrepancy between the reference heart-rate distribution and the heart rate at the time of observation. On the training dataset, the mode correction was applied, as discussed above, mostly because of time gaps between the reference period and the observation period. For the validation dataset these large time gaps were not present and it was assumed that reference data from an adjacent 24-hour time window would be similar enough for accurate predictions to be made. However, for certain patients this turned out not to be the case, negatively affecting predictions. The mode correction enforces some degree of similarity between distributions of the observation and reference period. As such, it may also boost performance in the validation dataset.

The results on the validation dataset demonstrate that using a 24-hour reference period might not be sufficient to accurately capture the expected heart rate and respiratory-rate distributions of a patient at a particular instant. In fact, the heart-rate distribution is expected to vary during the day due to a circadian rhythm (Figure S13). This illustrates that a more accurate reference model of heart rate and respiratory-rate distributions at a specific time of day could potentially improve the performance of our model on the broader population.

5.1 Mode correction of validation patients

To illustrate the potential gains of the discussion above, the validation dataset was re-analyzed, this time with application of the mode correction. A performance enhancement is indeed observed (Figure S13-Figure S16; figures similar to main text). All performance metrics improve with mode correction enabled, with performance very similar to cross validation performance (Table S3). Considering this analysis was performed post hoc, its effect would need to be validated in an independent dataset.
Figure S12. Heart rate values in 24-hour reference period for an example patient, averaged by minute. A clear circadian rhythm can be observed.

Figure S13. AUROCs calculated for the random forest classifier on the validation dataset, with reference period mode correction enabled. Performance mean and 95% Confidence Intervals (solid lines and filled areas, respectively) were calculated using 250-fold bootstrapping. Left, macro-averaged over all sleep-wake states; right, for the individual sleep states.
Figure S14. Confusion matrixes of observed sleep-wake states vs predicted sleep-wake states (W = Wake, AS = Active Sleep, QS = Quiet Sleep) calculated for the random-forest classifier on the validation dataset, with reference period mode correction enabled.

![Confusion Matrix](chart.png)

Figure S15. (Left panel) Model calibration per sleep-wake state (W = Wake, AS = Active Sleep, QS = Quiet Sleep), with data points split over 20 quantiles per sleep-wake state, for the validation dataset, with reference period mode correction enabled. (Right panel) Predicted time spent in a particular sleep-wake state (horizontal axis) versus the actual time observed in that sleep-wake state (vertical axis). Dots represent individual patients and are colored by sleep-wake state (Green/W = Wake, Red/AS = Active Sleep, Orange/QS = Quiet Sleep). Predicted time was calculated by summing all model probability predictions. Graph shows the validation dataset, with reference period mode correction enabled. Note that one patient is included twice as it was observed in two three-hour observation rounds.

![Calibration Curve](chart2.png)

![Time in Sleep-Wake State](chart3.png)
Table S3. Overview of performance metrics on the validation dataset, with reference period mode correction enabled, with 95% CI values between brackets, calculated using 250-fold bootstrapping. Any metrics indicated with a sleep-wake state were calculated for that specific sleep-wake state vs the combination of other sleep-wake states.

| Validation set | Balanced accuracy wake | 70.0% (59.0-82.0%) |
|----------------|------------------------|--------------------|
|                | Sensitivity wake       | 45.0% (21.0-72.0%) |
|                | Specificity wake       | 95.0% (92.0-97.0%) |
|                | F1 score (macro-averaged) | 54.0% (46.0-59.0%) |
|                | F1 score wake          | 46.0% (26.0-58.0%) |
|                | AUROC (macro-averaged) | 75.0% (70.0-80.0%) |
|                | AUROC wake             | 89.0% (83.0-94.0%) |
|                | Cohen’s kappa          | 0.259 (0.135-0.379) |
|                | Cohen’s kappa wake     | 0.4 (0.211-0.521)  |
|                | Brier Active Sleep     | 0.233 (0.207-0.261) |
|                | Brier Quiet Sleep      | 0.188 (0.161-0.211) |
|                | Brier Wake             | 0.068 (0.045-0.089) |

6 References

1. Thorndike RL. Who belongs in the family? *Psychometrika* 1953; **18**: 267–76.

2. Menon AK, Jiang X, Vembu S, Elkan C, Ohno-Machado L. Predicting accurate probabilities with a ranking loss. *Proc 29th Int Conf Mach Learn ICML 2012* 2012; 1: 703–10.